◎数据库、数据挖掘、机器学习◎

改进LeaderRank算法的意见领袖挖掘

徐郡明,朱福喜,刘世超,朱碧颖
XU Junming, ZHU Fuxi, LIU Shichao, ZHU Biying

武汉大学 计算机学院,武汉 430072
School of Computer, Wuhan University, Wuhan 430072, China

XU Junming, ZHU Fuxi, LIU Shichao, et al. Identifying opinion leaders by improved algorithm based on LeaderRank. Computer Engineering and Applications, 2015, 51(1):110-114.

Abstract: Opinion leader mining is an important topic of social network research which is of great significance in terms of internet public opinion control and information dissemination. LeaderRank is an effective opinion leader mining algorithm. This paper proposes a new algorithm based on LeaderRank with sentiment analysis and user's liveness to identify the opinion leaders in comment network to achieve a better performance. Experiments based on SIR model show that both accuracy and anti-jamming capability of the proposed algorithm have been effectively improved.
Key words: social network; opinion leader mining; LeaderRank; sentiment analysis; user's liveness

摘 要: 意见领袖挖掘是社会网络研究的重要课题,对于舆情控制、信息传播等方面具有重要意义。LeaderRank算法是一个有效的意见领袖挖掘算法。为提高LeaderRank算法的准确性和抗干扰能力,在LeaderRank算法基础上,加入用户之间的情感倾向、用户活跃程度,提出了改进的LeaderRank算法。基于SIR模型的实验验证表明,改进算法的准确性和抗干扰能力均得到了有效提升。
关键词: 社会网络; 意见领袖挖掘; LeaderRank算法; 情感倾向; 用户活跃程度
文献标志码: A **中图分类号:** TP393 doi:10.3778/j.issn.1002-8331.1403-0032

1 引言

在复杂网络社区中,各用户因其环境、性格不同,所起到的作用也有所不同。其中,部分用户会积极的接受并传播某些信息和观点,并且对其他用户有着重要的影响力,这些用户即为"意见领袖"。"意见领袖"传播的信息会影响网络中另一部分用户,引导这部分用户的行为。因此,若使用某些信息影响"意见领袖",就可能高效地影响到网络中的大多数用户。所以,进行"意见领袖"的挖掘,对于利用已成型、稳定的社区网络进行信息推广、舆情监测等具有重要的理论及现实意义。

LeaderRank算法是一个有效的意见领袖挖掘算法,本文在该算法的基础上,结合用户之间的情感倾向信息和用户活跃度因素,以提高意见领袖挖掘的准确度。

2 相关研究

新兴媒体(如微博、微信等)逐渐渗透到人们的日常生活中,在用户之间形成了一个"用户-用户"的复杂网络。越来越多的研究人员开始关注信息如何在这个复杂网络中进行传播[1]以及如何控制传播速度[2-4],从而找到合适的方式促进正确的社会价值信息的传播,抑制谣言的流行,提高广告投放的精确度。

Aral等[5]的研究证明意见领袖在复杂网络的信息传

基金项目: 国家自然科学基金(No.61272277)。
作者简介: 徐郡明(1991—),男,硕士研究生,研究方向为Web数据挖掘;朱福喜(1957—),男,教授,博士生导师,研究领域为智能计算,Web数据挖掘,自然语言处理;刘世超(1989—),男,博士研究生,研究方向为社会网络,Web数据挖掘;朱碧颖(1990—),女,硕士研究生,研究方向为智能计算,数据挖掘。E-mail:fxzhu@whu.edu.cn
收稿日期: 2014-03-06 **修回日期:** 2014-05-13 **文章编号:** 1002-8331(2015)01-0110-05
CNKI **网络优先出版:** 2014-07-11, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1403-0032.html

播中起着关键性作用。Bai等[6-7]认为可以将复杂网络中度最大的节点看作“意见领袖”，通过影响这些节点来控制信息在网络中传播，这一方法被应用于多种复杂网络的意见领袖挖掘[8-10]。

Kitsak等[11]人指出，网络中节点的传播能力与节点所处的位置有重要关系，处于网络核心位置的节点，即使度较小，也会有较高的影响力；反之亦然。与此同时，意见领袖挖掘方面的其他研究人员也陆续提出了大量其他指标，例如接近中心性[12]、特征向量中心性[13]、路由中心性[14]和环中心性[15]。这些指标主要是用来衡量一个节点在网络中的传播能力。

Song等[16]借鉴PageRank算法对用户进行排名，并结合用户之间情感倾向分析、评论隐含关系分析、评论的时间衰减等因素挖掘复杂网络中的意见领袖。但使用PageRank算法不适合在结构快速变化的复杂网络中挖掘意见领袖。

鉴于PageRank算法在意见领袖挖掘方面表现出来的种种问题，Lv等人[17]提出了一种意见领袖挖掘的新算法——LeaderRank算法。在意见领袖挖掘方面，LeaderRank算法比PageRank算法的准确性更高，面对噪音和恶意攻击时的稳定性更强。

Li等[18]针对LeaderRank算法进行改进，提出加权的LeaderRank算法，并通过实验验证了加权的LeaderRank算法具有更好的准确性和稳定性。

因此，国内外学者在意见领袖挖掘研究工作中非常注重LeaderRank算法。但LeaderRank算法及其加权改进算法也存在一些可改进之处，例如：算法没有考虑网络中“用户之间的情感倾向”、“用户活跃程度”等因素对意见领袖挖掘的影响，这会制约算法性能的进一步提升。本文旨在这些方面做出改进。

3　意见领袖发现算法

3.1　PageRank算法分析

PageRank算法将对页面的链接看成投票，实现将互联网中“链接价值”概念作为排名因素。PageRank算法将计算得到的网页在不同时刻的得分作为衡量网页重要性的标准。PageRank算法的核心公式如式(1)所示：

$$P_i(t+1)=c+(1-c)\times\sum_{j=1}^{N}\left[\frac{\alpha_{ij}}{k_j^{\text{out}}}\left(1-\varphi_{k_j^{\text{out}},0}\right)+\frac{1}{N}\varphi_{k_j^{\text{out}},0}\right]P_j(t) \tag{1}$$

其中，N 表示页面总数；$P_i(t)$ 表示在 t 时刻，i 页面的得分；α_{ji} 表示当页面 j 中是否存在指向页面 i 的超链接，存在时，$\alpha_{ji}=1$，否则，$\alpha_{ji}=0$；k_j^{out} 表示 j 页面中指向其他页面的超链接的个数；当 $k_j^{\text{out}}=0$ 时，$\varphi_{k_j^{\text{out}},0}=1$，当 $k_j^{\text{out}}\neq 0$ 时，$\varphi_{k_j^{\text{out}},0}=0$；$c$ 是“跳转概率”，当一个用户访问一个页面时，以概率 c 通过地址栏，随机跳到其他网页，以概率 $1-c$ 通过网页中的超链接跳转到其他网页。

在网络结构快速变化的社会网络中，PageRank算法的“跳转概率” c 的最佳值，需要随着网络结构的变化而不断调整。针对某一特定结构的网络，要经过多次训练才能得到参数 c 的最佳取值，这对于结构变化频繁的网络并不适用。除此之外，社交网络是一个非连通图，而PageRank算法并不保证在非连通图上具有收敛性。上述两原因导致PageRank算法在意见领袖挖掘方面有很大限制。

3.2　LeaderRank算法分析

Lv等[17]提出的LeaderRank算法，对PageRank算法的改进主要是在网络中添加一个公共节点“Ground Node”，并让此节点连接网络中的全部节点，这样一个 N 节点 M 边的网络 $G(N,M)$ 就变成 $N+1$ 节点 $M+2N$ 边的网络 $G(N+1,M+2N)$。

添加“Ground Node”的作用如下：

(1)由于“Ground Node”连接了网络中其余各节点，从而形成一个连通图，这就保证了LeaderRank算法的收敛性。不仅如此，“Ground Node”的加入还减小了整个网络的半径，增加收敛速度。

(2)某 i 节点的信息来源(即由 i 节点发出的指向其他节点的边)的多少反比于节点 i 流向“Ground Node”节点的个数。随着网络结构的变化，不同的节点会有不同的“跳转概率”。LeaderRank算法不再需要参数“跳转概率” c。

Lü等[17]还通过实验证明了LeaderRank算法具有更高的准确性和更强的稳定性。LeaderRank算法的核心公式如式(2)、(3)所示：

$$s_i(t+1)=\sum_{j=1}^{N+1}\frac{\alpha_{ji}}{k_j^{\text{out}}}s_j(t) \tag{2}$$

$$S_i=s_i(t_c)+\frac{s_g(t_c)}{N} \tag{3}$$

其中，$s_i(t)$ 表示 i 节点在 t 时刻的得分；t_c 表示 $s_i(t)$ 收敛的时刻；$s_g(t_c)$ 表示在 t_c 时刻“Ground Node”节点的得分；S_i 表示 i 节点最终的得分，其他变量含义同式(1)。

3.3　LeaderRank算法的缺陷分析

在现实网络中，用户之间存在明显的观点差异或是有相当数量的恶意注册用户，都会导致意见领袖排名出现偏差，而LeaderRank算法会受到这两方面因素的影响。因此，改进的LeaderRank算法增加了用户的情感倾向和恶意注册用户两个方面考察。

3.3.1　用户的情感倾向

用户可以接受其他信息源的某个观点并加以传播，

也可以反对其他信息源的观点并阻碍其传播。这取决于用户的情感倾向。例如在微博中,粉丝中有所谓的“红粉”(支持博主观点)和“黑粉”(反对博主观点)之分,如图1所示,第一、三条是支持博主的评论,而第二条反对博主的评论。而之前的LeaderRank算法都没有考虑到用户之间的情感倾向。

图1　某微博评论图

在微博中,存在一些特殊符号,例如@、#等,在图1中的第三条微博回复中,此用户(用户 A)就使用了@符号,指向其他用户(用户 B、用户C)。虽然用户 A 不是直接被用户 B、C 的微博内容影响,但用户 B、C 的其他行为显然也影响到了用户 A(比如,这里的影视剧作品)。将这种关系定义为成“隐式关联”,LeaderRank算法也都没有考虑数据集中存在的“隐式关联”。为保留“隐式关联”对意见领袖排名的影响,会在实验前的数据整理过程中,将其显式表示出来(即显式的添加边 $A \rightarrow B$ 和边 $A \rightarrow C$,并在计算 A 与 B、C 的情感倾向时,将此评论计算进去)。

3.3.2　恶意注册用户

除了考虑用户间的情感倾向外,需注意到网络中存在的一定数量的恶意注册用户,例如微博中出现的“僵尸粉”。“僵尸粉”通常由第三方系统自动产生。如图2所示,第二、三、四条评论即是恶意注册用户对某微博进行恶意评论的情况。

图2　某微博恶意评论图

这种恶意注册用户在现实网络中,并不对信息进行传播或阻碍,因此在挖掘意见领袖时,将这种恶意注册用户也计算进去是不恰当的。虽然在LeaderRank算法以及加权LeaderRank算法中,通过添加“Ground Node”可以让恶意注册用户的更多分数流入“Ground Node”,但是恶意注册用户始终还是存在分数流入其他信息源的情况,当这种恶意注册用户足够多时,会影响到意见领袖的排名,从而影响意见领袖挖掘的准确性。

3.4　改进的LeaderRank算法

针对上述LeaderRank算法存在的缺陷,本文做出了如下改进。

3.4.1　添加用户情感倾向

从图1中可以得到,微博内容主要是由文本、表情符合以及“赞”标志个数构成,微博内容的情感倾向也主要由这三部分的情感倾向组成。为简单起见,本文使用“赞”标志个数表示微博的情感倾向,其计算公式如式(4)所示:

$$\mu_{ij} = \lg \delta \tag{4}$$

其中,μ_{ij} 表示节点 i 对节点 j 发布的某条回复(或微博)的感情倾向得分;δ 表示此回复的“赞”的个数;$\lg \delta$ 是此条回复的“赞”标志的得分。

在确定网络中任意两个用户 α 和 β 之间的情感倾向时,并不是取决于 α 和 β 之间的某次回帖的情感倾向,而是由特定时间段内,α 和 β 之间全部回帖的情感倾向的平均值决定的。ρ_{ij} 表示节点 i 与节点 j 之间的“平均情感倾向”,其计算公式如式(5)所示:

$$\rho_{ij} = \sum_{t=0}^{n} \mu_{ij} \Big/ n \tag{5}$$

其中,μ_{ij} 表示在 t 时刻节点 i 对节点 j 回复的情感倾向得分,共计 n 个时刻。

3.4.2　减弱恶意注册用户的影响

通过对微博等社交网络的观察,发现恶意注册用户与非活跃用户之间存在很大交集,因此本文使用时间衰减的方式来限制恶意注册用户对意见领袖挖掘的影响。衰减比例公式如式(6)所示:

$$\theta_i(t_c) = E \cdot D^{|t_l - t_c|} \tag{6}$$

其中,t_l 表示用户 i 最后一次有效操作的时间(例如,发帖、发表评论、转帖等操作);D 是衰减指数;E 是阻尼系数;$\theta_i(t_c)$ 表示的是节点 i 在 t_c 时刻的衰减比例。

3.4.3　算法的改进

本文将用户之间的情感倾向、时间衰减因素与加权LeaderRank算法结合,改进的LeaderRank算法核心公式如式(7)所示:

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{\theta_j(t) \cdot \rho_{ji} \cdot \omega_{ji}}{k_j^{\text{out}}} s_j(t) \tag{7}$$

其中,ω_{ij} 为一条边的权重值,取值规则为:若 j 为“Ground Node”,$\omega_{ig} = 1$,$\omega_{gi} = k_i^{\text{in}}$;若不存在 i 指向 j 的边,$\omega_{ij} = 0$;其余情况 $\omega_{ij} = 1$。ρ_{ji} 为式(4)中得到的节

点 i 与节点 j 之间的情感倾向值。$\theta_j(t)$ 为式(5)式中得到的节点 j 在 t 时刻的衰减比例；其他变量含义同公式(1)、(2)。

4　实验及分析

实验模型：实验采用 Susceptible-Infected-Removed (SIR)模型[19]，该模型由 Kermack 与 McKendrick 在1927年提出，直到现在SIR模型仍被广泛使用。在SIR模型中，用 $N(t)$ 表示 t 时刻的总人数，总人口分为以下三类：易感者(Susceptibles)，记为 $S(t)$，表示 t 时刻未染病但有可能被该类疾病传染的人数；染病者(Infectives)，记为 $I(t)$，表示 t 时刻已被感染成为病人而且具有传染力的人数；移出者(Removal)，记为 $R(t)$，表示 t 时刻已从染病者中移出的人数。在实验过程中发现，将SIR模型中的康复免疫概率 γ 设置成1，感染概率 β 设置成0.02时，效果较好。

数据准备：实验使用的数据集为新浪微博在2013年8月到12月份中对某一热门话题的讨论数据。由于新浪微博API存在接口访问频次等限制，所以实验并没有使用新浪API获取实验数据，而是使用课题组实现的页面爬虫工具进行抓取，再将得到的原始数据进行整理，除去新浪微博活动、广告等非用户发布的信息，并且将上文提到的“隐式关联”显式表示出来。数据整理后，共计67 031个节点，130 122条边。

参数分析：在上述公式(6)中，参数 D 的取值对减弱恶意用户的影响有重要作用。若 D 取值过大，则减弱恶意用户影响力的效果不明显；若 D 取值过小，则会影响正常用户中，短时间未活动用户的影响力。

下面通过实验确定参数 D 较合适的取值。考察对于 D 的不同取值，改进LeaderRank得到前5名的意见领袖的所影响的用户个数的平均值。从图3中可以得到，当 D 取值在区间[0.8，0.9]之间较合适。

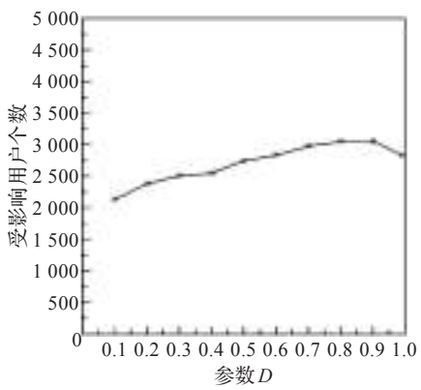


图3　参数 D 取值测试图

4.1　准确性比较

以PageRank算法和LeaderRank算法为对照组，实验验证改进LeaderRank算法的准确性。在上述数据集中，分别应用上述3种算法找到的排名前20的意见领袖，实验结果数据如表1所示。

表1　排名前20的意见领袖对比表

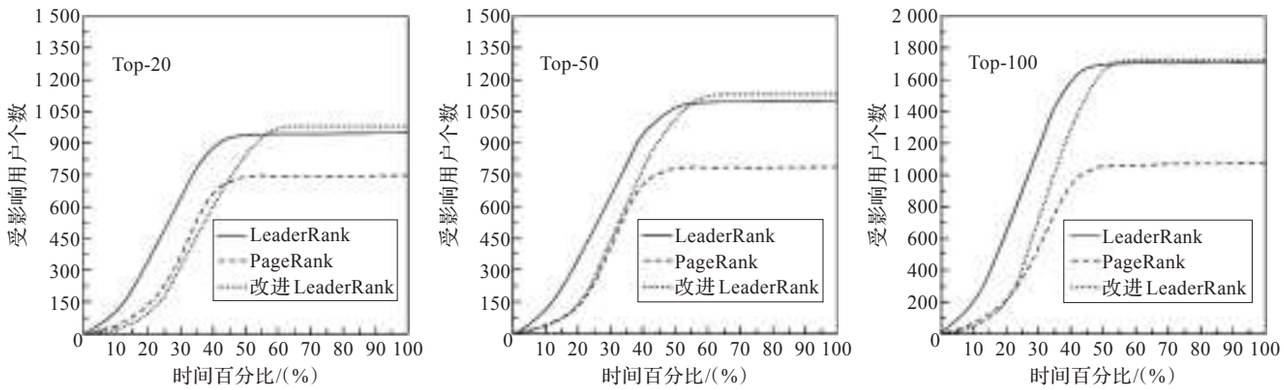
排名	LeaderRank	PageRank	改进LeaderRank
1	13423	13423	13423
2	314	314	314
3	1207	37	1207
4	1203	517	1203
5	517	1203	544
…	…	…	…
16	7021	9315	512
17	9315	512	7021
18	26792	143	9315
19	31241	26792	31241
20	512	7021	26792

表1第一行第二列表示应用LeaderRank算法得到的排名第一的意见领袖为“id=13423”号用户，其他列以此类推。由实验结果整理可得，LeaderRank算法得到的前20名中的{1207，517，31241}号意见领袖无法由其他两种算法得到。同理，PageRank算法得到的{37，517，143}和改进LeaderRank算法得到的{1207，544，31241}无法由其他两种算法得到。针对表1中的数据，专门查看了造成此差异的具体原因，以“id=517”号用户为例，在LeaderRank算法的结果中排名为第5，在改进LeaderRank算法的结果中排名为第151。这里发现其微博的粉丝和评论很多，但是相当一部分都是反对其观点的评论，并且其微博“赞”个数几乎没有，如图4所示。

图4　反对评论示例图

接着，给这些只由单一算法得到的意见领袖施加影响，观察一段时间内，受这些意见领袖影响的其他用户的人数，由此判定这些意见领袖的影响力，作为衡量“意见领袖挖掘”算法准确性的度量标准。实验得到的前20位意见领袖，即Top-20曲线如图5(a)所示。出于实验的严谨考虑，同时实验Top-50曲线和Top-100曲线，分别如图5(b)和5(c)所示。横坐标表示迭代时间 t，纵坐标表示被影响的节点个数 N，PageRank算法的 c 取0.15；以30次实验平均值度量实验结果。

由图5可以看出，改进的LeaderRank算法挖掘出的意见领袖的影响力略高于LeaderRank算法，但收敛速度



(a)算法Top-20实验对比　(b)算法Top-50实验对比　(c)算法Top-100实验对比

图5　3种算法Top-*N*实验对比图

比LeaderRank算法慢一些,故而适用于对精度要求更高而对时间消耗要求相对较宽松的场合。

4.2　抗干扰能力比较

下面设计实验,比较3种算法的抗干扰能力。在网络中添加一定数量的恶意注册用户,并将这些恶意注册用户随机与图中节点相连,并随机对相连节点的微博进行回复操作和“点赞”操作。比较添加恶意用户之后各节点排名与添加恶意用户之前各节点排名,变化较小的即被认为是抗干扰能力强的。抗干扰能力公式,如公式(8)所示:

$$I_{\text{Robust}} = \left| S'_i - S_i \right| \tag{8}$$

其中,S'_i 表示添加恶意注册用户前,i 节点的得分;S'_i 表示添加恶意注册用户后,j 节点的得分。

实验对排名前100的意见领袖施加干扰,3种算法下的抗干扰能力结果如图6所示,越接近对角线则认为其抗干扰能力越强。实验结果证明,改进的LeaderRank算法的抗干扰性优于LeaderRank算法和PageRank算法。

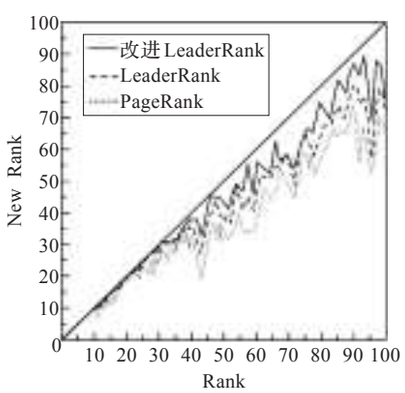


图6　3种算法稳定性对比图

5　结束语

“意见领袖挖掘”在社会网络研究领域是一个重要的研究课题。本文的主要贡献在于:通过将网络中用户之间的情感倾向、用户活跃程度与LeaderRank算法相结合,对LeaderRank算法进行改进,提高了LeaderRank算法的准确性和抗干扰能力。但是,改进后的LeaderRank依然存在不足,例如微博内容的情感倾向没有考虑文本以及表情符号的情感,这将是下一阶段的研究重点。

参考文献:

[1] Zhou T,Fu Z Q,Wang B H.Epidemic dynamics on complex networks[J].Progress in Natural Science,2006,16(5):452-457.

[2] Lv L,Chen D B,Zhou T.The small world yields the most effective information spreading[J].New Journal of Physics,2011,13(12).

[3] Doerr B,Fouz M,Friedrich T.Why rumors spread so quickly in social networks[J].Communications of the ACM,2012,55(6):70-75.

[4] Schläpfer M,Buzna L.Decelerated spreading in degree-correlated networks[J].Physical Review E,2012,85(1).

[5] Aral S,Walker D.Identifying influential and susceptible members of social networks[J].Science,2012,337:337-341.

[6] Bai W J,Zhou T,Wang B H.Immunization of susceptible-infected model on scale-free networks[J].Statistical Mechanics and its Applications:Physica A,2007,384(2):656-662.

[7] Hébert-Dufresne L,Allard A,Young J G,et al.Global efficiency of local immunization on complex networks[R].Scientific Reports,2013.

[8] Zhou Y B,Lv L,Li M.Quantifying the influence of scientists and their publications:distinguishing between prestige and popularity[J].New Journal of Physics,2012,14(3).

[9] Park J,Newman M E J.A network-based ranking system for US college football[J].Journal of Statistical Mechanics:Theory and Experiment,2005(10):10014.

[10] Huang X,Vodenska I,Wang F,et al.Identifying influential directors in the United States corporate governance network[J].Physical Review E,2011,84(4).

(下转166页)