池州学院学报
Journal of Chizhou University
2017年12月 第31卷 第6期
Dec. 2017 Vol.31 No.6
DOI:10.13420/j.cnki.jczu.2017.06.013

基于PageRank改进算法的在线社交网络意见领袖挖掘研究

刘纯丽

(池州学院 文学与传媒学院，安徽 池州 247000)

[摘要]Web3.0为在线社交网络方面提供了更加完善的互动和个性体验，形成的海量数据在商业和研究领域价值很大，尤其是在线社交网络中的意见领袖在广告投放、舆情监控与引导方面具有重要意义。在PageRank算法基础上综合考虑用户关系、用户关注行为和时间因素对意见领袖的挖掘展开研究，从而形成了新TBLP(Time-Behavior-Leader PageRank)算法。采集新浪微博的数据进行实验表明，算法在意见领袖挖掘方面更加有效且识别精度高。

[关键词]在线社交网络；意见领袖；用户影响力；关注行为；PageRank

[中图分类号]TP311 [文献标识码]A [文章编号]1674-1102(2017)06-0048-04

1 引 言

伴随着数字媒体技术、计算机网络技术、移动通信技术的迅猛发展，人们使用互联网的方式不再只有传统的网页浏览和信息检索，还实现了社交网络关系的构建维护以及信息的交流与共享。据中国互联网信息中心（China Internet Network Information Center，简称CNNIC）发布的《第39次中国互联网发展状况统计报告》[1]显示：截止2016年12月底，国内的微博网民数量达27143万人，较2015年增长了4098万，增长率达到17.8%；其中手机微博网民规模达到24086万人，较2015年增加了5396万，增长率达到28.9%。

在线社交网络中，影响力大的用户在信息的交流和分享过程中越能施加影响，称之为“意见领袖”。由于意见领袖具有众多的粉丝，其发布的新闻评论或消息会得到广泛的传播，形成热门话题，在舆情监控及引导和产品的社会化营销等方面得到广泛的应用。因此，在线社交网络意见领袖的挖掘已经成为社交网络研究的热点。

2 在线社交网络意见领袖相关理论

本节主要阐述在线社交网络的概念及在线社交网络结构特征，介绍意见领袖的定义的基础上分析意见领袖的挖掘方法。

2.1 在线社交网络及分析

在线社交网络[2]是一种由用户以及用户间交互关系形成的网络，其中网络中的节点就是用户，网络的边界[3]由用户之间关联关系和交互行为形成。社交网络服务是基于社会学的相关理论——六度分隔理论、邓巴数理论、强关系理论、弱关系理论、小世界关系理论等，以互动交友为基础，为分享兴趣爱好、参加活动、交流学习提供社会关系网络服务[4]。对于社交网络的研究与社交网络的发展过程[5]也是同步的：从最早期的概念化阶段对于E-mail的挖掘，经历了结交陌生人阶段的BBS数据的挖掘研究，娱乐化阶段典型的应用是人人网、开心网，目前所处的社交图阶段实现了线上线下的映射——国内研究的热点是新浪微博、QQ和微信等的数据挖掘研究。

对于在线社交网络的研究分析是集合计算机科学和社会学的交叉学科，同时为了更好的分析网络用户行为，通常采用网络图论和数据库相结合的数据挖掘技术。将在线的社交网络抽象成一个有/无向图G(V,E,W)，其中，V代表网络中的节点（用户）集合，E代表网络中边（用户之间关系）的集合，W代表各边之间的权重值。E有向边根据用户的出入度确定，W值可以根据用户之间的关注关系、互动关系（点赞、评论、转发）确定。如图2所示，通过关系图直接反映了用户行为和及其之间形成的相

收稿日期：2017-10-18
基金项目：池州学院自然科学研究重点项目（2016ZRZ013）。
作者简介：刘纯丽（1983-），女，安徽石台人，池州学院文学与传媒学院讲师，硕士，研究方向为社会计算及计算机网络。

互关系，并通过出入度、节点中心度等数学概念对社交网络进行测量，对用户之间的关系进行深度挖掘。

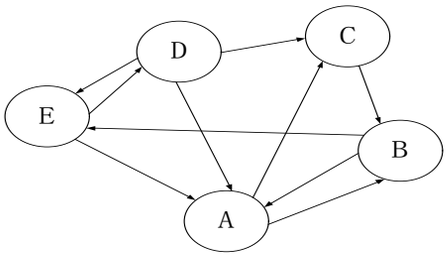

图1 小型社交网络关系图

2.2 意见领袖定义及挖掘方法

上个世纪40年代拉扎斯菲尔德[6]等人通过对的美国总统大选的民调数据进行分析，提出了大众传播的两级模式：大众传播-意见领袖-一般受众[7]，此后又通过一些研究证实了意见领袖在购物、流行、时事等领域的存在。在互联网技术不断深入到人们生活工作的方方面面时，网络意见领袖[8]承担网络平台信息传播中的桥梁任务，其对相关事件发表的观点被其他用户赞同、采纳以及转发。网络传播是更为复杂的多级传播过程，多级意见领袖发挥了其“较强意见”特性，影响了其它用户的决策，从而实现了消息的指数级传播。

社交网络意见领袖挖掘过程包括：采集数据→确定用户影响力的影响因子→确定意见领袖挖掘算法公式→实验结果对比分析→结果。目前最广泛的用户影响力衡量因子是用户的入度，也就是用户粉丝数越多，影响力越大。本文采用基于PageRank算法综合考虑用户交互行为的意见领袖挖掘算法。

3 PageRank算法及其不足

在互联网中用户的影响力水平受到物理结构特性和行为特性影响，物理特性包括相互的关注关系，即形成社交网络的拓扑结构；行为特性包括自身行为（发表微博数）、相互的点赞、评论及转发微博数量。因此，为了更好的计算用户影响力，本节首先介绍PageRank算法并借鉴来计算用户影响力，在此基础上阐述PageRank算法的不足及改进之处。

3.1 PageRank算法

3.1.1 PageRank算法背景　1997年谷歌创始人拉里·佩奇和谢尔盖·布林在搭建搜索引擎系统初期时发现的基于链接分析的网页排序算法——PageRank算法[9]，从此，该算法成为谷歌用来衡量网页等级的核心算法。谷歌将网页的级别划分为0到10级，综合主题、关键字等各项特征，利用PageRank算法计算网页的PR值，值越大网页就越重要，那么在搜索结果中的排名也越靠前，谷歌也因此取得了巨大的成功。随着社交网络研究的深入，很多学者开始将PageRank算法应用到意见领袖的挖掘中，使其成为学术界研究的焦点。

3.1.2 PageRank算法分析　PageRank算法的基本思想是网页的重要程度由链接到它的页面的数量及重要程度决定。如图2所示，假设在这个网页链接结构图中，网页B、C、D、E均建立到网页A的链接，网页A的PageRank值就可以简单的表述为：

$$PR(A)=\frac{PR(B)}{OL(B)}+\frac{PR(C)}{OL(C)}+\frac{PR(D)}{OL(D)}+\frac{PR(E)}{OL(E)} \quad (1)$$

将上述公式进行迭代得到如下公式：

$$PR(A)=\sum_{W\in WS}\frac{PR(W_i)}{OL(W_i)} \quad (2)$$

考虑到实际情况有些网页没有出链，网页迭代遍历到此就无法继续，那么该网页的PR值就会不断增加，由此设置一个跳转概率p。当一个网页没有出链时，那么其跳转到网络中其它网页的概率是一样的，由此得到完整的PageRank公式[10]：

$$PR(A)=\frac{1-p}{N}+p*\sum_{W_i\in WS}\frac{PR(W_i)}{OL(W_i)} \quad (3)$$

公式中W_i表示网络中的网页，$PR(W_i)$表示网页W_i的PageRank值，N为网络中网页总数量，p为阻尼系数，$OL(W_i)$表示网页W_i的出度，WS表示链入网页的集合。当p=1时，回到公式（3），且p取值越小，收敛越快，有效性越低，当p=0时，所有页面的PR值相同。研究表明，p=0.85[9]最为合适。

3.1.3 PageRank算法实现

首先假设用户从一个网页跳转到其它与其链接的网页的概率均为1，由图2建立表1所示的N维邻接矩阵——其中i行j列表示从网页j跳转到网页i的概率，再规范化得到转移矩阵如表2所示。

表1 邻接矩阵

结点	A	B	C	D	E
A	0	1	0	1	1
B	1	0	1	0	0
C	1	0	0	1	0
D	0	0	0	0	1
E	0	1	0	1	0

表2 规范化转移矩阵

结点	A	B	C	D	E
A	0	1/2	0	1/3	1/2
B	1/2	0	1	0	0
C	1/2	0	0	1/3	0
D	0	0	0	0	1/2
E	0	1/2	0	1/3	0

设置好初始非负向量$(1/5,1/5,1/5,1/5,1/5)^T$，用表2所示的转移矩阵去乘以这个向量，得到的值再用转移矩阵去乘。经过多次迭代计算，最终趋于收敛值，于是得到如表3所示的PageRank值，根据数值可得到网页B在网络中的重要性最大。

表3 PageRank值

A	B	C	D	E
0.27	0.3	0.17	0.1	0.17

3.2 不足与改进

PageRank算法设定一个网页的链入网页数多,该网页就能得到较高的PageRank值。在社交网络中,用户的粉丝数量多,其发表的言论被传播的概率也会更大,因此,可以利用PageRank算法基于用户间关注关系计算用户影响力。

事实上,不管是网页还是社交网络中,由于其链接的网页或者关注的用户不同,也就会采用不同的交互行为。PageRank算法假设网络结构中每个页面对其它页面的影响程度是相同的,这样计算出来的结果在某种程度上可靠性受到影响。同时,有些网页或者社交网络中用户,由于注册时间长而积累了很多的链入用户而得到很高的PageRank值,这不利于挖掘那些活跃度高、关注度跟进速度快的新注册的网页或用户。

3.2.2 Brin和Page[10]的博士论文在回顾了网页排序算法的基础上,深入探讨了PageRank算法如何实现大规模搜索。作为计算网络结点影响力的经典算法,有学者借鉴了PageRank算法来计算社交网络的用户影响力。Ming-Feng Tsai等学者考虑到用户行为及兴趣领域两方面来发现社区意见领袖[11], Jianshu Weng考虑用户在不同话题中的影响力不同问题提出了TwitterRank算法[12],Diego等学者考虑到话题被转发时间长短在PageRank算法基础上引入了时间因素[13]来计算话题影响力,Jiangjiao Duan等学者考虑到用户发表文本情感因素进而挖掘社区中的意见领袖[14]。

4 意见领袖挖掘改进算法

通过对社交网络意见领袖挖掘和PageRank算法的研究,本文综合考虑用户关系、用户关注行为和时间因素对意见领袖的挖掘的影响,从而形成了新TBLP(Time-Behavior-Leader PageRank)算法。

4.1 用户影响力因子分析

根据度中心性思想,社交网络中用户的粉丝数目越多,其号召力也就越大,这在PageRank算法中通过将其粉丝的影响力迭加计算已经有所体现。同时用户自身属性除了其入度(粉丝数目)之外,还有一个很重要的因素就是自身的活跃程度。在特定时间内用户在社交网络中,自身活跃程度越大,对网络上其他用户的影响也就可能越大。用户在社交网络中进行的行为主要包括发布微博、转发微博、回复评论、评论、点赞和发表评论,从信息传播的角度考虑用户的活跃程度主要考虑其发布微博和转发微博的数量两个因子。

某个用户被其他用户关注,在PageRank算法中每个用户所获得的关注度是相同的,而实际在社交网络中并非同等对待的,因此在迭代计算用户影响力时不应该采用平均值而应该按比例分配。在社交网络中,用户发布的微博得到点赞、评论、转发获得不同比例的关注度,因此在进行意见领袖挖掘时进行影响力计算时将获得不同比例的影响因子。

4.2 TBLP算法提出

根据用户发布微博和转发微博的数量定义其基于时间特性(一年内)的活跃程度,用于计算用户u_i的活跃程度:

$$T(u_i)=\lg(M_{u_i}*N_{u_i}) \tag{4}$$

其中$T(u_i)$表示用户u_i的归一化之后的活跃程度,M_{u_i}表示用户发布微博的数量,N_{u_i}表示用户转发微博的数量。

根据粉丝用户点赞、评论和转发的交互行为定义其从粉丝出获得影响力:

$$B_{ji}=\frac{l*0.2+c*0.3+f*0.5}{l+c+f} \tag{5}$$

其中B_{ji}表示用户u_j对用户u_i的归一化之后的行为权重值,其中l表示点赞数目,c表示评论数目,f表示转发数目。公式中根据行为本身的影响力赋予不同的交互行为的不同权重。

本文综合考虑用户基于时间的活跃程度和交互行为提出了在线社交网络意见领袖挖掘算法TBLP(Time-Behavior-Leader PageRank),用于计算社交网络用户影响力值:

$$TBLP(u_i)=(1-d)T(u_i)+d*\sum_{j\in Fl(i)}\frac{B_{ji}}{\sum_{k_i\in Fr(j)}B_{kj}}TBLP(u_j) \tag{6}$$

其中$TBLP(u_i)$表示用户u_i的影响力值,Fl(i)表示用户u_i的所有粉丝集合,Fr(j)表示用户u_i粉丝的所有出度,d为阻尼系数,取值0.85。

TBLP算法具体描述如下:

算法1 在线社交网络意见领袖挖掘算法(TBLP)

输入:网络拓扑关系(节点集合U和边关系集合E及交互权重B)

输出:节点影响力排名(TOP-10)

1 计算用户u_i的自身活跃程度权重值$S(u_i)$:

$$T(u_i)=\lg(M_{u_i}*N_{u_i})$$

$$S(u_i)=(1-d)*T(u_i)$$

2 初始化ε//用于计算两次迭代的差值,当近似相等时,TBLP趋于稳定值

3 while($\varepsilon>10^{-7}$)

4 do

5 $\varepsilon=0$

6 计算粉丝用户u_j对u_i的影响力贡献：

$$TBLP(u_i)=TBLP_{old}(u_i)+d*\frac{B_{ji}}{S(u_i)}TBLP_{old}(u_i)$$

7 计算前后两次迭代的差值之和：

$$\varepsilon+=|TBLP(u_i)-TBLP_{old}(u_i)|$$

8 end while

9 输出影响力较大的10个节点。

5 对比实验

实验数据包括100个粉丝数超过200万的新浪微博用户信息，以及他们的微博相应被点赞、转发和评论信息表。数据为sql脚本文件，能够直接导入到数据库。对这些数据进行整理统计并分类存储，每张表的数据字段分布为：

（1）用户信息表：计数，id，昵称，是否加V，是否达人，是否会员，等级，性别，地区，自我介绍，关注数，粉丝数，微博数，头像，标签，关注列表：

（2）微博信息表：计数，id，发表时间，来源微博id，转发数，评论数，点赞数。

使用PageRank算法和本文提出的TBLP算法迭代计算用户的影响力，收敛后得到影响力值最大的用户为意见领袖，如表4所示。

表4 影响力排名前十用户

	PageRank		TBLP	
排名	得分	昵称	得分	昵称
1	0.072607	新手指南	0.081326	思想聚焦
2	0.071365	微博管理员	0.081273	全球健身中心
3	0.032409	谢娜	0.081263	人民日报
4	0.032327	陈坤	0.081201	新浪新闻
5	0.032315	姚晨	0.079422	回忆专用小马甲
6	0.032204	赵薇	0.078564	中国新闻网
7	0.032005	何炅	0.078021	央视新闻
8	0.031989	angelababy	0.073291	最神奇的视频
9	0.030326	林心如	0.071273	当时我就震惊了
10	0.030265	郭德纲	0.070756	人民网

在PageRank算法中排在前十位的用户粉丝量巨大，其中排在第一位的昵称为“新手指南”的粉丝数为174985532，微博数为10903，但是其微博的转发、评论和点赞数都偏低，其他绝大部分账号用户均是娱乐明星，拥有大量的粉丝，但是由于其发布的消息转发、评论量相对不高，这也就影响了其影响力值。

本文提出的TBLP算法排名前十位的用户虽然粉丝量虽没有达到那些娱乐明星，但是其发布的微博内容具有很强的互动性，被广泛的转发、评论和点赞，使其影响力值大大提升。通过分析发现影响力排在前十的账户主要是一些传统媒体的微博帐号或者是自媒体知名博主，是在线社交网络的意见领袖。

6 总结与展望

本文基于时间因素对在线社交网络的意见领袖挖掘展开研究。从用户发表微博以及微博的影响力出发，在考虑用户的活跃程度基础上，更加准确的挖掘那些新注册的具有很高活跃程度的意见领袖。通过实验分析比较，本文提出的TBLP算法能有效的挖掘出那些潜在的新注册的意见领袖。

同时，本文的不足之处是实验数据量小，随着规模的扩大，用户的增加，挖掘的效率会受到严重的制约影响。在以后的研究中，将对云计算平台的意见领袖分布式识别算法进行探究，提高计算效率，从而使在线社交网络意见领袖挖掘达到企业应用价值。

参考文献：

[1] 中国互联网络信息中心(CNNIC). 第39次中国互联网络发展状况统计报告[EB/OL].[2017-01-01] http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201701/t20170122_66437.htm.

[2] TANG L，LIU H.Community Detection and Mining in Social Media[J]. Synthesis Lectures on Data Mining and Knowledge Discovery，2010，2(1):1-137.

[3] Charu C. Aggarwal. Social Network Data Analytics[M]. New York: Springer，2011:1-15.

[4] 李立耀.社交网络研究综述[J].计算机科学，2015(11)：8-21.

[5] Kumar R，Novak J，Tomkins A. Structure and evolution of online social networks[M]//Link Mining: Models，Algorithms，and Applications. Springer New York，2010:337-357.

[6]P.Lazarsfeld，Bernard Berelson，Hazel Guadet. The People’s Choice[M]. New York: Deull Pearce and Sloan Press，1948.

[7] Katz E，Lazarsfeld P F. Personal influence，The part played by people in the flow of mass communications[M].Transaction Publishers,1970.

[8]丁雪峰，胡勇，网络舆论意见领袖特征研究[J].四川大学学报(工程科学版). 2010，42(2):145-149.

[9] PAGE L，BRIN S，MOTWANIR，etal The PageRank Citation Ranking:Bringing Order to the Web[R]. Technicalreport: Stanford University，1999.

[10] Brin S，PAGE L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[J]. Computer Networks and ISND Systems，1998，30(1):107-117.

[11] Tsai M F，Tzeng C W，Lin Z L，et al. Discovering leaders from social network by action cascade[J]. Social Network Analysis and Mining，2014，4(1):1-10.

[12] Weng J，Lim EP，Jiang J，et al. Twitterrank : finding topic-sensitive influential twitterers[C] //Proceedings of the third ACM international conference on Web search and data mining. ACM，2010: 261-270.

[13] Saez-Trumper，D.，Comarela，G.，Almeida，V.etc. (2012). Finding trendsetters in information networks. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2012: 1014－1022.

[14] Duan Jiangjiao，Zeng Jianping，Luo Banghui. Identification of opinion leaders based on user clustering and sentiment analysis[C]// 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence(WI) and Intelligent Agent Technologies(IAT) . IEEE，2014.

[责任编辑：桂传友]