·人工智能及识别技术· 文章编号:1000-3428(2018)02-0203-07 文献标志码:A 中图分类号:TP301.6

# 基于PageRank的网络社区意见领袖发现算法

周 飞,高茂庭

(上海海事大学 信息工程学院,上海 201306)

**摘 要:** 意见领袖是网络社区中积极的信息传播者和信息引导者,对其影响力的评估是社交网络分析的一项重要内容。针对现有算法对用户动态行为分析和动态内容影响考虑欠缺而不能客观反映真实情况的问题,提出一种基于用户影响力和PageRank的意见领袖发现算法,综合考虑用户自身影响力、用户动态行为影响度和用户行为给动态内容带来的真实影响。通过从知乎网络社区收集的大规模数据实验结果表明,该算法更具合理性并能有效地提高网络社区意见领袖的识别准确度。

**关键词:** 网络社区;意见领袖;影响力;PageRank算法;用户行为

**中文引用格式:** 周 飞,高茂庭. 基于PageRank的网络社区意见领袖发现算法[J]. 计算机工程,2018,44(2):203-209,219.

**英文引用格式:** ZHOU Fei,GAO Maoting. Discovery Algorithm of Opinion Leaders for Network Community Based on PageRank[J]. Computer Engineering,2018,44(2):203-209,219.

# Discovery Algorithm of Opinion Leaders for Network Community Based on PageRank

ZHOU Fei,GAO Maoting

(College of Information Engineering,Shanghai Maritime University,Shanghai 201306,China)

**【Abstract】** Opinion leader is active information promulgator in the network community and the guide of information transmission. The assessment for opinion leaders' influence is an important content of social network analysis. To improve the considerations about the analysis of user dynamic behavior and the influence of dynamic content,and reflect the real situation objectively,a discovery algorithm of opinion leaders based on user's influence and PageRank is proposed. It combines user's own influence,the influence degree of user dynamic behaviors and the real impact of dynamic content brought by the dynamic behaviors. Experimental results on the large-scale data collected from Zhihu network community demonstrate that the algorithm is more reasonable and can effectively improve the recognition accuracy of the network community opinion leaders.

**【Key words】** network community;opinion leader;influence;PageRank algorithm;user activity

**DOI:** 10.3969/j.issn.1000-3428.2018.02.036

## 0 概述

随着互联网技术的迅猛发展,社交网络服务(Social Network Service,SNS)作为互联网应用发展的必备要素,不再局限于信息传递,而是与沟通交流、商务交易类应用融合,借助其他应用的用户基础,形成更强大的关系链,从而实现信息的广泛、快速传播。网络社区是具有相同兴趣爱好的网民相互交流、共享资源的虚拟社区,越来越多的人通过网络社区分享信息、图片,表达意见、观点或参与话题讨论。与线下社区一样,网络社区同样存在社会分层,不同的是网络社区更多依据思想和观点的影响力进行划分,影响力较高的成员就成了群体中的重要角色,即意见领袖。意见领袖通常是网络社区中的活跃分子,是信息的积极传播者,能够提供大量信息、意见,引起大量关注并影响社区中的舆论导向,对网络信息传播、网络营销、广告投放、舆论引导等方面起着极其重要的作用[1]。因此,对网络社区中意见领袖的发现进行研究具有重要意义。

"知乎"是社区氛围友好与理性、连接各行各业精英的一个网络问答社区。用户利用各自的专业知识、经验和见解,为互联网源源不断地提供高质量的信息。知乎不同于微博与传统社区,社会身份并非是知乎社区意见领袖的决定因素,知乎特有的投票

**基金项目:** 国家自然科学基金(61202022)。

**作者简介:** 周 飞(1993—),男,硕士研究生,主研方向为数据挖掘、数据分析;高茂庭,教授、博士、CCF高级会员。

**收稿日期:** 2016-12-12 **修回日期:** 2017-02-12 **E-mail:** phil_chow@outlook.com

机制和关注模式催生了大批草根意见领袖[2]。鉴于现有意见领袖发现算法中对用户动态行为分析和动态行为所带来的真实浏览量考虑不足的问题[3-6]，本文对用户自身影响力、用户动态行为及其给动态内容带来的真实影响等 3 个方面进行研究，提出一种基于用户自身影响力、影响力传播度和 PageRank 的意见领袖识别算法。

## 1 相关研究

文献[7]提出的二级传播理论是关于意见领袖的最早研究，该理论指出意见领袖在主要以广播和报纸为信息传播媒介的当时占有不可或缺的地位，媒介信息必须经由某些意见领袖才能到达其他人群。随着互联网的发展，网络社交媒体成为人们日常生活的重要工具，同时也吸引了众多学者对其信息传播、社会影响力、意见领袖发现等方面进行研究。文献[8]指出在网络社区中人们通常通过用户发布信息的数量来认定意见领袖。文献[9]通过 Twitter 网络证实了信息传播过程中两级传播理论的存在。文献[4]用关注用户数量、粉丝数量、是否被验证身份和发布的微博数量等 4 项数据构建微博客意见领袖识别多维模型，对微博客用户重要性进行评分。文献[5]利用从网络中采集到的基本数据，构造网络话题参与者的“属性矩阵”，提出意见领袖综合评价算法。文献[6]选取 7 个用户特征，采用聚类分析方法筛选出具有意见领袖特点的群体。文献[4-6]都是通过提取意见领袖属性特征进行归纳分析，提出意见领袖发现算法。但是这些算法都没有考虑到用户与用户之间的关注关系，因此，可能存在用户大量发帖但并没有人对其回复却被误认为是意见领袖的情况，与客观事实存在一定偏差。文献[10]通过考虑用户的兴趣空间和回复关系，提出基于兴趣领域的意见领袖识别算法。文献[3,11]将情感倾向性作为用户之间评价的指标，并作为网络权重分别提出 OpinionRank 算法和 LeaderRank 算法。文献[12]基于话题相似度和用户间关注关系提出 TwitterRank 算法。文献[13]发现消息在微博网络中的传播过程可近似分解为各个意见领袖所驱动的子过程的特性，提出基于消息传播的微博意见领袖影响力建模方法，并得出影响力衰减指数的大小以及影响力持续时间的长短与粉丝数量几乎无关的结论。文献[14]综合考虑用户自身影响力和用户之间的链接关系，提出基于用户影响力的 PageRank 意见领袖识别算法，简称 UilRank 算法。该算法虽然考虑了网络论坛中的发帖数、回帖数和被回复数、被浏览数，但是缺乏用户动态行为分析，以及存在使动态内容阅读量增长的来源指向不明确的现象。

针对以上算法中用户动态行为分析缺失和动态内容阅读数增长不明确等问题，本文以网络社区“知乎”为研究对象，综合分析意见领袖影响力因子，在 UilRank 算法的基础上，又从用户动态行为影响传播度和用户行为对动态内容带来的真实影响两个方面考虑，提出一种基于 PageRank 的知乎意见领袖影响力发现算法。其中用户自身影响力来源于诸如用户粉丝数、获得赞同数、回答问题数等用户自身属性。用户动态行为及其对问题的真实影响将通过对用户动态行为信息和问题动态变化信息分析得出，两者共同决定用户影响力传播度的大小，继而作用于改进的 PageRank 算法中。

## 2 意见领袖发现

### 2.1 算法基础

PageRank 的初衷指的是计算某个人在任意次点击链接之后到达某一网页的可能性，在网络社区意见领袖发现中可把用户之间的关注关系看作是用户之间的网络拓扑结构，通过分析网络拓扑结构可获得用户影响力排名。因此，用户影响力可以通过 PageRank 算法得出，如式(1)所示。

$$PR(u)=(1-d)+d\times\sum_{v\in L_u}\frac{PR(v)}{N(O_v)} \tag{1}$$

其中，$PR(u)$表示网页 $u$ 的 PageRank 值，$L_u$ 表示指向网页 $u$ 的网页集合，$N(O_v)$表示网页 $v$ 指向其他网页的总个数，$d$ 为阻尼因子，表示某页面被访问的概率，一般设为 0.85。

文献[14]在 PageRank 算法的基础上，提取用户属性特征并给出权重，将用户间的回复次数作为影响力占比分配原则，提出基于用户影响力的意见领袖发现算法，简称 UilRank 算法，如式(2)、式(3)所示。

$$R(u)=(1-d)+d\sum_{v\in T_u}W_{uv}\times R(v) \tag{2}$$

$$W_{uv}=\frac{I_u\times k_{uv}}{\sum_{p\in B_v}I_p\times k_{vp}} \tag{3}$$

其中，$R(u)$表示用户 $u$ 的影响值，$T_u$ 为回复 $u$ 的用户集合，$W_{uv}$表示用户 $u$ 在所有影响用户 $v$ 的节点中所占比例，$I_u$ 代表用户 $u$ 的初始影响值，$k_{uv}$表示用户 $u$ 和 $v$ 之间的回复次数，$B_v$ 表示用户 $v$ 回复的用户集合，通过数次迭代直至达到收敛状态，得到用户影响值。

### 2.2 问题及解决思路

在用户影响力传播度计算上，现有意见领袖识别算法往往采用均分原则平均分配，与实际网络中意见领袖对不同用户影响程度不同的这一情形不相符。

意见领袖在对某一提问做出回答行为或者对某一答案做出点赞行为时，他的行为动态就产生了，继而将影响他的部分粉丝也对该提问或回答产生行为动态。然而，在这个过程中，有部分粉丝看到意见领袖的行为动态后，只是浏览了这个动态内容，并没有发出回答或点赞的行为动态，因此，不能确定意见领袖是否对该用户产生了影响。

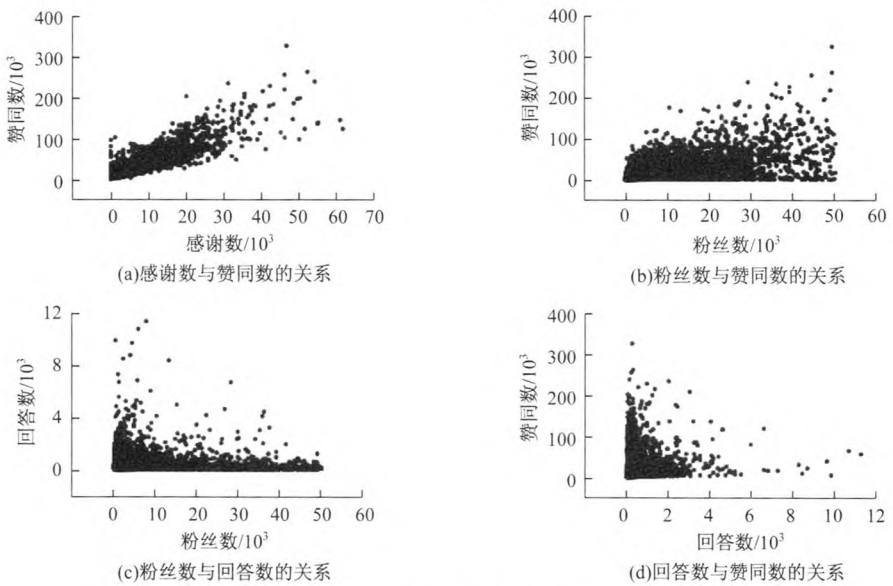
为了解决这些问题,对这些属性值量化处理后得到用户的初始影响力值,即用户自身影响力。还将往往被研究者遗忘的动态内容被阅读数作为一个考核标准,考量用户在对某动态内容产生行为动态后一段时间内该内容阅读量的真实变化情况,阅读量变化情况考量是对用户行为动态影响度量存在缺漏现象的补充。然而在实际网络中又存在普通用户紧随意见领袖产生行为动态而将意见领袖对阅读量变化产生的影响据为己有的现象,用户行为动态影响度量又反过来制约了此现象,防止普通用户被认为拥有高影响力传播度。2 种度量方法相辅相成,构成用户影响力传播度。最后将用户自身影响力和用户影响力传播度引入到改进的 PageRank 算法中得到每个用户的最终影响力,排名靠前者即为网络社区意见领袖。

## 3　改进的意见领袖发现算法

### 3.1　用户自身影响力

#### 3.1.1　用户属性特征提取

文献[15-17]指出在 Twitter 网络环境中,粉丝数量在信息传播过程中和用户影响力呈弱相关性。本文通过获取到的知乎真实数据,对用户粉丝数、获得赞同数、回答问题数、获得感谢数 4 个属性特征两两刻画相关性散点图,如图 1 所示,从图 1(a)发现获得赞同数和获得感谢数存在一定的线性相关性,因此,将赞同数和感谢数看作是相同的影响因子。另外,从图 1(b)~图 1(d)可以看出用户粉丝数、获得赞同数和回答问题数这 3 个属性特征不存在线性相关性,因此,使用这 3 个属性特征对用户自身影响力进行评估,其中回答问题数是对意见领袖活跃度的一种肯定。




(a)感谢数与赞同数的关系　(b)粉丝数与赞同数的关系

(c)粉丝数与回答数的关系　(d)回答数与赞同数的关系

图 1　用户各属性特征散点图

#### 3.1.2　用户属性特征计算

从图 1 可看出各个特征数据不具备一致性参考标准,因此不能直接比较,需要对每个特征指标进行归一化处理。由于特征数据跨度较大,如用户粉丝数高的用户可以达到百万级别,低的用户甚至一个粉丝都没有,因此本文采用对数归一化处理方法。这种处理方式计算简单、运算速度快、处理后数据跨度小,如式(4)所示。

$$F_u=\begin{cases}0, f_u=0\\ \dfrac{\lg f_u}{\lg f_{\max}}, f_u>0\end{cases} \tag{4}$$

其中,$F_u$ 表示对用户 $u$ 实际粉丝数做归一化处理后得到的值,$f_u$ 表示用户 $u$ 的实际粉丝数,$f_{\max}$ 代表所有用户粉丝数的最大值。同理,对用户 $u$ 获得赞同数和回答问题数用式(4)归一化处理后分别表示为 $S_u$、$A_u$。

本文定义以下公式计算用户 $u$ 的自身影响力值。

$$SI(u)=\omega_1F_u+\omega_2S_u+\omega_3A_u \tag{5}$$

其中,$SI(u)$ 代表用户自身影响力值,对应 UilRank 算法中的 $I_u$,$F_u$、$S_u$、$A_u$ 分别是用户粉丝数、获得赞同数和回答问题数归一化处理过后的值,$\omega_1$、$\omega_2$、$\omega_3$ 代表不同特征的权重值。为了将各属性重要程度数学化、系统化,本文采用层次分析法确定每个属性特征的权重值,该方法对于多准则、多目标的系统有较好的判定效果[18]。构建以下判断矩阵:

$$\boldsymbol{M}=\begin{bmatrix}1 & 2 & 6\\ 1/2 & 1 & 4\\ 1/6 & 1/4 & 1\end{bmatrix} \tag{6}$$

通过计算,得到各个属性特征权值,一致性检验结果为 0.079 33 < 0.1,满足一致性检验,各属性特征权值 $\omega_i$ 如表 1 所示。

表 1 属性特征权重

| 属性特征 | 粉丝数 | 获得赞同数 | 回答数 |
|---|---|---|---|
| 权重 | 0.588 | 0.323 | 0.089 |

### 3.2 用户影响力传播度

在实际的网络传播中,存在以下 2 种现象:

1)在意见领袖发出回答、点赞等行为动态后,部分粉丝接收到意见领袖的动态,阅读了相关动态内容并对此动态内容发出行为动态。那么意见领袖对于这部分粉丝的影响是显而易见并且可以通过收集动态行为数据得到。但是仍然会存在一些粉丝在阅读了动态内容后,不发出任何行为动态,对于这部分粉丝则无法通过动态行为数据知晓意见领袖是否对其产生了影响。

2)粉丝们会通过意见领袖发出的行为动态浏览这一动态内容,那么该动态内容在该意见领袖发出回答、点赞等行为之后某个时间段内的浏览数增长便可在一定程度上反映该意见领袖的影响力。但是当 2 个意见领袖 A 和意见领袖 B 相近时间发出同样的行为动态时,就无法确定给动态内容带来的影响是意见领袖 A 还是意见领袖 B,或者是他们分别带来了多少影响。

分析这 2 种现象可以发现,其实现象 2 就是对现象 1 中意见领袖影响缺失的一个补充,现象 1 则是对现象 2 中给动态内容带来影响重叠的一个制约。对于现象 1,将采用高行为动态数据(即参与者人数较多的动态内容数据)根据时间节点建立有向无环图计算用户行为动态信息下的影响力传播度。对于现象 2,采用低行为动态数据(即参与者人数较少的动态内容数据)计算行为动态后的问题被浏览增长率,确定基于动态内容浏览数增长下的用户影响力传播度。最后将两者加权累加得到用户影响力传播度。

#### 3.2.1 用户行为动态信息下的影响力传播度

在分析以时间线为基准的用户行为动态后发现,该动态行为序列构成一个有向无环图,如图 2 所示。

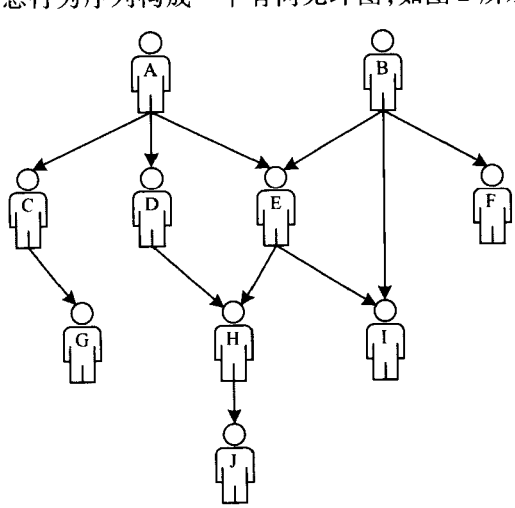



图 2 用户行为动态结构

图 2 描述了影响力传播的 3 种情况:

1)在用户 A 发出某一行为动态后,他的粉丝用户中 C、D、E 也对该内容发出行为动态,则可认为 A 对 C、D、E 产生了影响。

2)用户 H 是用户 D、E 的共同粉丝,且用户 D、E 在用户 H 前发出行为动态,那么认定用户 H 同时受到用户 D、E 的影响。

3)用户 I 是用户 B、E 的共同粉丝,且用户 E 在用户 B 之后发出行为动态,用户 I 在用户 E 之后发出行为动态,那么认定用户 I 同时受到用户 B、E 的影响,用户 E 受到用户 B 的影响。

由此,根据用户动态行为数据建立用户动态行为结构图(在图中认定出度为 0 的节点为叶子节点),并依次从叶子节点向根节点遍历,统计每个用户的用户动态行为影响力值。用户行为动态信息下的影响力传播度算法描述如下,其中 *qid* 表示动态内容编号。

**输入** *qid*

**输出** 用户行为动态影响度 *Degree*

执行步骤:

1) $Userlist \leftarrow$ 动态内同编号为 *qid* 并以时间节点排序的用户列表;

2) for 用户 $u \in UserList$:{

用户 $u$ 的孩子集 $ChildSet_u \leftarrow UserList$ 中排在 $u$ 之后的用户集和用户 $u$ 粉丝集的交集;

将添加到 $ChildrenSet_u$ 中每个用户 $c$ 的父集合 $ParentSet_c$ 中;}

3) $LeafSet \leftarrow UserList$ 中孩子集合为空的用户集合;

4) for 用户 $u \in LeafSet$:{

标记 $v$ 已经被查找过;

将 $v$ 父集合 $ParentSet_v$ 中每个父节点 $p$ 的深度 $Degree_p$ 自加 1;

如果 $p$ 没有被查找过且不在 *LeafList* 中,将 $p$ 添加到 *LeafSet* 中;}

5) if *LeafSet* 不为空,转到 4),否则转到 6);

6)用对数归一化法对 *Degree* 进行归一化处理;

7) Return *Degree*。

该算法中步骤 2)和步骤 3)的时间复杂度都是 $O(N)$,步骤 4)、步骤 5)为二层循环,时间复杂度为 $O(N^2)$,步骤 6)为归一化处理,时间复杂度为 $O(N)$。因此,该算法时间复杂度为 $O(N^2)$。另外,需要 $3N$ 的额外空间存储结果和中间变量。因此,该算法空间复杂度为 $O(N)$。

在使用用户行为动态影响力算法对每一个行为动态计算之后,得到每个动态下的用户影响力值,返回归一化后的用户动态行为影响力传播度。

$$Uar(u) = \sum_{qid \in Qlist}^{N} Degree(qid)[u]/N \tag{7}$$

其中,$Degree(qid)[u]$ 为对第 *qid* 号的动态行为做

用户动态行为影响力算法后用户 $u$ 的影响力传播度,$Qlist$ 为用户行为动态编号列表。

3.2.2　动态内容浏览数增长下的影响力传播度

文献[13]指出在微博网络环境中,在意见领袖发出一条消息后 300 min 内,消息以激增的态势传播,随后逐渐减弱,第二天会有所增长但影响将逐渐消失。由于微博信息繁杂且动态内容更新速度快,表现出快速增长和快速消亡的特性。但对于知乎而言,这个过程就相对缓慢一些,因此,以 2 天为一个行为动态的影响周期,计算这段时间内的最快增长,把增长率作为用户动态行为给问题带来实际影响的考量标准。

本文通过式(8)~式(10)计算用户 $u$ 给动态内容带来的平均真实影响度。

$$Gn(q,t)=\max(B(q,t+1)-B(q,t)),\forall t\in[t,t+2] \tag{8}$$

$$Gr(q,u)=\frac{Gn(q,utime)}{\max(Gn(q,T))},\forall T\in[ST,ET] \tag{9}$$

$$Qir(u)=\sum_{q\in Qlist}^{N}Gr(q,u)/N \tag{10}$$

其中,$B(q,t)$ 表示问题 $q$ 在 $t$ 时刻的被浏览次数,$Gn(q,t)$ 表示问题 $q$ 在 $[t,t+2]$ 时间区间内被浏览次数增长最大值,$utime$ 表示用户 $u$ 对问题 $q$ 产生行为动态的时刻,$\max(Gn(q,T))$ 表示在整个数据集时间段中问题 $q$ 的被浏览次数增长最大值,$Gr(q,u)$ 表示用户 $u$ 在问题 $q$ 下的影响力比率,$Qir(u)$ 表示用户 $u$ 在众多问题动态中给问题带来的平均真实影响度。

综合用户行为动态信息下的影响力传播度和行为动态给动态内容带来的实际影响度得到用户影响力传播度(User Influence Transfer Degree,UITD)。

$$UITD(u)=\frac{Uar(u)+Qir(u)}{2} \tag{11}$$

### 3.3　ZhihuRank 意见领袖发现算法

本文在 PageRank 算法的思想基础上提出了基于用户自身影响力、用户影响力传播度和 PageRank 的意见领袖发现算法,简称 ZhihuRank 算法,如式(12)、式(13)所示。

$$ZR(u)=(1-d)+d\sum_{v\in FR_u}W(u,v)\times ZR(v) \tag{12}$$

$$W(u,v)=\frac{SI(u)\times UITD(u)}{\sum_{k\in FE_v}SI(k)\times UITD(k)} \tag{13}$$

其中,$ZR(u)$ 表示用户的影响力值。$d$ 为阻尼因子,表示用户受到影响的概率,通常在(0,1)之间,本文设为 0.85。$FR_u$ 表示用户 $u$ 的粉丝集合,对应于 UilRank 算法中的 $T_u$ 集合。$W(u,v)$ 表示用户 $u$ 在用户 $v$ 关注的人集合中影响力传播度的占比。$FE_v$ 表示用户 $v$ 关注的人的集合,对应于 UilRank 算法中的 $B_v$ 集合。$SI(u)$ 表示用户自身初始影响力值。$UITD(u)$ 代表用户 $u$ 的影响力传播度。

假设网络社区个体数为 $N$,设定 2 个结束标志,一个为网络循环迭代次数 $iterations$,另一个为 $\alpha$,表示每个个体当前 $ZR$ 值和上一次迭代结果 $ZR_{old}$ 值的差值的阈值。算法结束后 $ZR$ 为最终用户影响力值,$SORT()$ 是以 $ZR$ 为基准的逆排序函数。ZhihuRank 算法描述如下:

**输入**　$N$,$iteration$,$a$

**输出**　用户影响力排名

执行步骤:

1)对 $ZR$ 进行初始化,将所有节点 $ZR$ 值设为 1。

2)使用式(12)、式(13)计算每个节点的 $ZR$ 值。

3)如果 $iteration\leqslant 0$ 或者对于任意用户 $u$ 都有 $|ZR(u)-ZR_{old}(u)|<a$,则转到步骤 5);否则,转到步骤 4)。

4)$iteration\leftarrow iteration-1$,转到步骤 2)。

5)Return $SORT(ZR)$。

该算法中步骤 1)为初始化赋值,时间复杂度为 $O(N)$。步骤 2)~步骤 4)有 3 层循环,时间复杂度为 $iteration\times O(N^2)$,但是在实际操作中,可以对式(13)中的 $W(u,v)$ 进行预处理,利用额外空间换取时间,将时间复杂度降低为 $iteration\times 2\times O(N)$。$iteration$ 为常数,因此,该算法的时间复杂度为 $O(N)$,空间复杂度为 $O(N)$。

## 4　实验设置与结果分析

### 4.1　实验数据收集与软硬件环境

本文以知乎问答社区为数据来源,通过爬虫以作者的知乎账号为种子,收集作者关注的人的信息并存入到数据库中,再迭代循环爬取数据库中没有被爬取过的用户,用户信息数据包含用户 ID、用户名、粉丝数、关注数、获得赞同数、获得感谢数、回答问题数等。

在分析了用户属性特征散点图后,为了方便爬取用户动态行为信息,将可能是意见领袖的用户抽取出建立小型爬取源,减少不必要的网络流量。随后抓取此爬取源中用户行为动态信息和动态内容变化数据,分别存入数据库中。用户行为动态数据包含用户 ID、问题唯一标识 qid、动态产生时间、动态类型等。动态内容变化数据包含问题唯一标识 qid、问题标题、爬取时间、当前阅读量等。本文共收集了约 14 万知乎用户信息、近 8.5 万条用户行为动态信息、将近 22 万条问题变化信息及用户间的关注关系信息作为实验数据集。

实验软件环境为 Python,版本 2.7.10;数据库使用非关系型数据库 MongoDB,版本 3.2.9;硬件环境为 macOS,内存 8 GB,处理器 1.6 GHz Intel Core i5。

### 4.2　算法评测标准

目前对于影响力模型评测还没有一个统一的评

测标准,大多研究者采用覆盖度[19]和核心率[20]作为评价指标,或使用人工评价的方法。

覆盖度是从用户间网络拓扑结构的角度考虑,通过意见领袖发出的行为动态,在直接或者间接影响的用户数占全部用户数的比重来衡量意见领袖的影响力。覆盖度分为单步覆盖度和全路径覆盖度。本文采用单步覆盖度和带阻尼因子的全路径覆盖度对算法进行定性评测。

### 4.3 实验结果及分析

#### 4.3.1 意见领袖影响力排名

为了验证算法的合理性,将本文提出的 ZhihuRank 意见领袖发现算法与 PageRank 算法、UilRank 算法进行对比实验,3 种算法 Top-10 排名的结果如表 2 所示。从表 2 可以看出,3 种算法排名在意见领袖用户选取方面还是比较接近的,由于篇幅所限,不能显示更多的意见领袖用户。在使用 UilRank 算法计算用户初始影响力时数据集有稍许差别,通过分析,把本文中的用户获得赞同数当作 UilRank 的被回复数,用于计算用户的初始影响力。

**表 2 3 种算法结果对比**

| 排名 | PageRank 算法 | | UilRank 算法 | | ZhihuRank 算法 | |
|---|---|---|---|---|---|---|
| | 用户 ID | 影响值 | 用户 ID | 影响值 | 用户 ID | 影响值 |
| 1 | zhang-jia-wei | 20.37 | jixin | 28.76 | zhang-jia-wei | 30.98 |
| 2 | jixin | 18.56 | imike | 21.69 | gejinyuban | 26.79 |
| 3 | ma-bo-yong | 17.64 | zhouyuan | 19.44 | ma-bo-yong | 26.60 |
| 4 | gejinyuban | 17.48 | excited-vczh | 17.80 | jixin | 25.65 |
| 5 | zhouyuan | 13.68 | zhang-jia-wei | 17.46 | zhouyuan | 19.70 |
| 6 | liangbianyao | 13.29 | yolfilm | 17.42 | raymond-wang | 19.15 |
| 7 | imike | 13.21 | raymond-wang | 17.32 | amuro1230 | 18.95 |
| 8 | raymond-wang | 12.54 | fu-er | 16.79 | liangbianyao | 17.29 |
| 9 | yolfilm | 12.36 | amuro1230 | 16.26 | yolfilm | 17.28 |
| 10 | amuro1230 | 12.14 | ma-bo-yong | 15.05 | zhu-xuan-86 | 17.19 |

从表 2 还可以发现,位于 ZhihuRank 排名前 10 位的用户大多可以在 PageRank 或 UilRank 排名中找到,不同的是他们的排序位置存在一定的差异。图 3 为表 2 中所有用户的属性特性值在这些用户总特征值的加权占比情况。

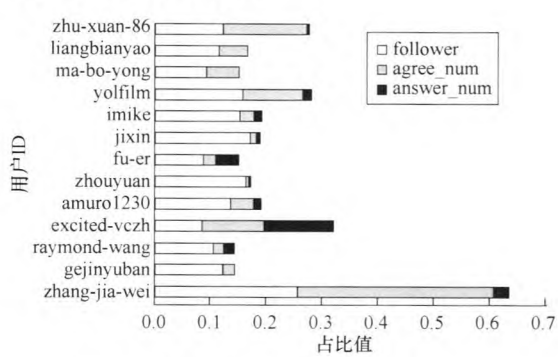



图 3 用户加权属性值占比堆积柱状图

用户“zhang-jia-wei”在 PageRank 和 ZhihuRank 算法排名中都处于第一位,这得益于他的高粉丝和高获赞以及有着较高的影响力传播度,且该用户是知乎网络上公认的意见领袖级用户。用户“excited-vczh”在 UilRank 算法排名中居于第 4 位但是在 ZhihuRank 中却未进前 10(第 11 位),而且他的属性占比情况也不差,原因在于 UilRank 算法中回帖数拥有较大属性特征权重,在本数据集中表现为回答问题数。但是在本文算法中,弱化了回答问题数这个属性,认为粉丝数和获得赞同数更为重要,相比于回答问题数更能体现用户的影响力,因此在 ZhihuRank 算法排名前 10 位中没有用户“excited-vczh”。

从图 3 中还发现 ZhihuRank 算法排名结果中的第 2 位、第 3 位用户“gejinyuban”和用户“ma-bo-yong”在特征数据占比中表现平平,原因有二。其一,从表 3 中可知他们有着高影响力传播度,而且虽然在数据占比上没有突出表现,但是基数很大,计算出的自身影响力实际并不低。其二,很多高质量用户(包括 Top-10 中的用户)是他们的粉丝,表 3 中给出了拥有 Top-30 的粉丝数,由于 Top-30 用户关注关系过于复杂,图 4 给出 Top-10 用户间的关注关系,从图中可以看出这 2 位用户的入度都非常高,说明了他们位于前列的合理性。

**表 3 用户影响力信息和 Top-30 粉丝拥有量**

| 排名 | 用户 ID | 初始自身影响力 | 影响力传播度 | Top-30 中粉丝数 |
|---|---|---|---|---|
| 1 | zhang-jia-wei | 0.986 | 0.866 | 18 |
| 2 | gejinyuban | 0.848 | 0.956 | 18 |
| 3 | ma-bo-yong | 0.872 | 0.987 | 16 |
| 4 | jixin | 0.879 | 0.828 | 20 |
| 5 | zhouyuan | 0.855 | 0.855 | 17 |
| 6 | raymond-wang | 0.88 | 0.88 | 16 |
| 7 | amuro1230 | 0.904 | 0.893 | 17 |
| 8 | liangbianyao | 0.87 | 0.869 | 14 |
| 9 | yolfilm | 0.934 | 0.817 | 15 |
| 10 | zhu-xuan-86 | 0.911 | 0.87 | 14 |

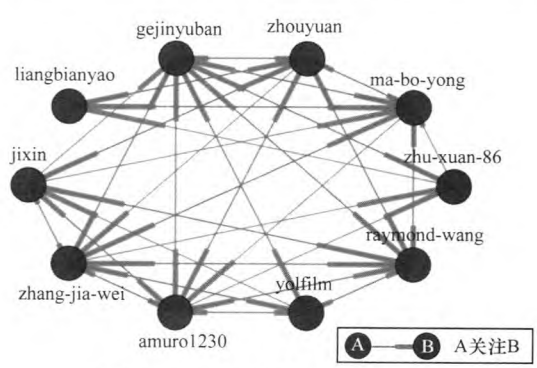



图 4 Top-10 用户间关注关系

综合以上实验结果和分析,ZhihuRank 算法考虑了用户的自身属性特征和用户动态行为影响力传播度,能够更准确、有效地识别网络社区中的意见领袖,且结果更具合理性。

4.3.2　算法评测结果

为了定性评估本文算法的有效性,分别使用单步覆盖度评价方法和传播影响力覆盖度评价算法对 Top-$k$ 意见领袖的影响力覆盖度进行测评,实验结果如图 5、图 6 所示。

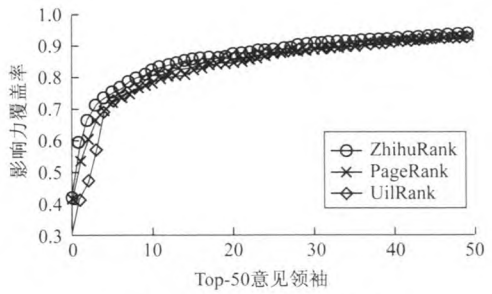



图 5　单步覆盖度评测结果

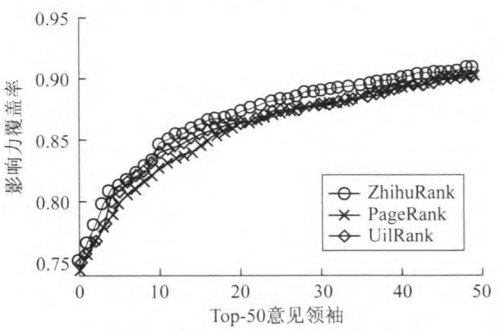



**图 6　传播覆盖度评测结果**

从图 5、图 6 可知,ZhihuRank 算法比其他 2 个算法不管在单步覆盖还是传播覆盖上都具有较大的影响力覆盖度,且有较明显优势,可见算法的有效性。从图 5 中还可发现在单步覆盖度评价中 UilRank 算法起初比 PageRank 的算法覆盖度还要低,说明在单步影响力传播环境下的 UilRank 算法并不能发挥其效果。PageRank 算法是完全基于拓扑环境下循环迭代传播的,所以在一定程度上比基于用户属性得到的影响力覆盖度更有优势。此外,还发现前 50 名意见领袖的影响力可以覆盖大约 90% 的用户。

## 5　结束语

在线网络社区的兴起给意见领袖发现研究提供了理想的实验平台。同时,这方面的研究也反过来影响着网络社区和人们的真实生活。

本文通过收集的知乎网络社区数据,对用户的影响力进行分析,从用户自身影响力、用户动态行为及其给动态内容带来的真实影响这 3 个方面进行研究,根据用户自身属性特征值计算出用户自身影响力,再通过用户动态行为及其对动态内容产生的影响计算出用户影响力的传播力度,最后利用改进后的 ZhihuRank 算法计算用户的最终影响力,发现知乎网络中的意见领袖。虽然本文的研究对象是知乎网络,但是提出的算法同样适用于诸如微博、论坛等类似网络社区。

但是,本文算法还有不足之处,比如在获取用户动态行为信息时可以更细致地将用户行为进行区分。在接下来的工作中,将进一步改进算法,收集更细致的用户动态行为和问题动态变化信息。此外,将对问题文本内容、内容主题、个人情感这些因素加入到影响力分析上,使研究更加精准、全面。

## 参考文献

[1] FREEMANL C. Centrality in Social Networks Conceptual Clarification[J]. Social Networks,1979,1(3):215-239.

[2] 王秀丽. 网络社区意见领袖影响机制研究——以社会化问答社区“知乎”为例[J]. 国际新闻界,2014,36(9):47-57.

[3] YU Xiao, LIN Xia. Understanding Opinion Leaders in Bulletin Board Systems: Structures and Algorithms[C]// Proceedings of the 35th Conference on Local Computer Networks. Washington D. C., USA: IEEE Press, 2010: 1062-1067.

[4] 王君泽,王雅蕾,禹　航,等. 微博客意见领袖识别模型研究[J]. 新闻与传播研究,2011(6):81-88.

[5] 丁雪峰,胡　勇,赵　文,等. 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.

[6] 王　珏,曾剑平,周葆华,等. 基于聚类分析的网络论坛意见领袖发现方法[J]. 计算机工程,2011,37(5):44-46.

[7] LAZARSFELD P F, BERELSON B, GAUDET H. The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign[M]. New York, USA: [s. n.], 2007: 229-233.

[8] YOO Y, ALAVI M. Emergent Leadership in Virtual Teams: What Do Emergent Leaders Do? [J]. Information & Organization, 2004, 14(1): 27-58.

[9] WU Shaomei, HOFMAN J M, MASON W A, et al. Who Says What to Whom on Twitter[C]//Proceedings of International Conference on World Wide Web. New York, USA: ACM Press, 2011: 705-714.

[10] ZHAI Zhongwu, XU Hua, JIA Peifa. Identifying Opinion Leaders in BBS [C]//Proceedings of International Conference on Intelligent Agent Technology. Sydney, Australia: [s. n.], 2008: 398-401.

[11] ZHOU Hengmin, ZENG D. Finding Leaders from Opinion Networks [C]//Proceedings of IEEE International Conference on Intelligence and Security Informatics. Washington D. C., USA: IEEE Press, 2009: 266-268.

(下转第 219 页)

[13] AISOPOS F, PAPADAKIS G, VARVARIGOU T. Sentiment Analysis of Social Media Content Using N-Gram Graphs [C]//Proceedings of the 3rd ACM SIGMM International Workshop on Social Media. New York, USA: ACM Press, 2011: 9-14.

[14] GHIASSI M, SKINNER J, ZIMBRA D. Twitter Brand Sentiment Analysis: A Hybrid System Using N-gram Analysis and Dynamic Artificial Neural Network[J]. Expert Systems with Applications, 2013, 40(16): 6266-6282.

[15] JIANG Long, YU Mo, ZHOU Ming, et al. Target-dependent Twitter Sentiment Classification [C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. Oregon, USA: Association for Computational Linguistics, 2011: 151-160.

[16] WANG Hao, CAN Dogan, KAZEMZADEH A, et al. A System for Real-time Twitter Sentiment Analysis of 2012 U. S. Presidential Election Cycle[C]//Proceedings of ACL 2012 System Demonstrations. Jeju Island, Korea: Association for Computational Linguistics, 2012: 115-120.

[17] LECUN Y, KAVUKCUOGLU K, FARABET C C. Convolutional Networks and Applications in Vision [C]// Proceedings of IEEE International Symposium on Circuits and Systems. Washington D. C., USA: IEEE Press, 2010: 253-256.

[18] 吴　轲. 基于深度学习的中文自然语言处理[D]. 南京:东南大学, 2014.

[19] KIM Y. Convolutional Neural Networks for Sentence Classification[EB/OL]. (2014-09-03). https://arxiv.org/abs/1408.5882.

[20] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[EB/OL]. (2014-09-08). https://arxiv.org/abs/1404.2188.

[21] JOHNSON R, ZHANG T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[EB/OL]. (2015-03-26). https://arxiv.org/abs/1412.1058.

[22] DOS SANTOS C N, GATTI M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts [C]// Proceedings of International Conference on Computational Linguistics. New York, USA: ACM Press, 2014: 69-78.

[23] HU Baotian, LU Zhengdong, LI Hang, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[C]//Proceedings of International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2015: 2042-2050.

[24] HE Hua, KEVIN G, LIN Jimmy. Multi-perspective Sentence Similarity Modeling with Convolutional Neural Networks[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Berlin, Germany: Springer, 2015: 1576-1586.

[25] YIN Wenpeng, HINRICH S. Convolutional Neural Network for Paraphrase Identification[C]//Proceedings of International Conference on Neural Information Processing. Berlin, Germany: Springer, 2015: 901-911.

[26] SEVERYN A, MOSCHITTI A. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification[C]//Proceedings of the 9th International Workshop on Semantic Evaluation. Berlin, Germany: Springer, 2015: 464-469.

[27] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM Press, 2011: 151-161.

[28] 梁　军,柴玉梅,原慧斌,等. 基于深度学习的微博情感分析[J]. 中文信息学报, 2014, 28(5): 155-161.

[29] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global Vectors for Word Representation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Berlin, Germany: Springer, 2014: 1532-1543.

编辑　顾逸斐

(上接第 209 页)

[12] WENG Jianshu, LIM E P, JIANG Jing, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers [C]//Proceedings of ACM International Conference on Web Search & Data Mining. New York, USA: ACM Press, 2010: 261-270.

[13] 王晨旭,管晓宏,秦　涛,等. 微博消息传播中意见领袖影响力建模研究[J]. 软件学报, 2015, 26(6): 1473-1485.

[14] 吴　渝,马璐璐,林　茂,等. 基于用户影响力的意见领袖发现算法[J]. 小型微型计算机系统, 2015, 36(3): 561-565.

[15] ASUR S, HUBERMAN B A, SZABO G, et al. Trends in Social Media: Persistence and Decay[EB/OL]. (2011-02-05). http://www.hpl.hp.com/research/scl/papers/trends/trends_web.pdf.

[16] CHA M, HADDADI H, BENEVENUTOF, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[C]//Proceedings of the 4th International Conference on Weblogs and Social Media. Washington D. C., USA: [s. n.], 2010.

[17] KWAK H, LEE C, PARK H, et al. What is Twitter, A Social Network or A News Media? [C]//Proceedings of International Conference on World Wide Web. New York, USA: ACM Press, 2010: 591-600.

[18] 许树柏. 层次分析法原理[M]. 天津:天津大学出版社, 1988.

[19] SONG Xiaodan, CHI Yun, HINO K, et al. Identifying Opinion Leaders in the Blogosphere [C]//Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York, USA: ACM Press, 2007: 971-974.

[20] MIAO Qingliang, ZHANG Shu, MENG Yao, et al. Domain-sensitive Opinion Leader Mining from Online Review Communities [C]//Proceedings of the 22nd International Conference on World Wide Web Companion. New York, USA: ACM Press, 2013: 187-188.

编辑　顾逸斐