Journal of Computer Applications
计算机应用,2013,33(8):2091 - 2094
ISSN 1001-9081
CODEN JYIIDU
2013-08-01
http://www.joca.cn

文章编号:1001-9081(2013)08-2091-04
doi:10.11772/j.issn.1001-9081.2013.08.2091

基于微博网络的影响力最大化算法

吴　凯*,季新生,郭进时,刘彩霞
(国家数字交换系统工程技术研究中心,郑州 450002)
(*通信作者电子邮箱 wukai10@gmail.com)

摘　要:由于影响范围的重叠效应,单纯的影响力度量算法并不能解决微博网络中的影响力最大化问题,针对这一研究现状,提出一种用于微博网络中 Top-*K* 节点挖掘的算法 GABE。通过归纳决定微博用户影响力的关键因素,提出了节点间影响率的概念,进而建立了用于用户影响力度量的 WIR 算法;根据得到的 WIR 值提出了符合微博特性的影响力传播模型,运用贪婪算法挖掘出微博网络中的 Top-*K* 节点。以爬取到的新浪微博数据进行了模拟验证,结果发现 GABE 在影响范围上与传统的最大化算法和影响力度量算法相比分别提高了 7.7% 和 20%。这表明通过引入微博特性和贪婪思想,GABE 较好地解决了微博网络中的影响力最大化问题。

关键词:微博;影响力度量;PageRank 算法;影响力最大化;贪婪算法
中图分类号:TP393.094　　**文献标志码**:A

Influence maximization algorithm for micro-blog network

WU Kai*, JI Xinsheng, GUO Jinshi, LIU Caixia
(*China National Digital Switching System Engineering and Technological R&D Center, Zhengzhou Henan* 450002, *China*)

Abstract: Influence maximization problem in micro-blog cannot be solved by simple user rank algorithm. To solve this problem, a greedy algorithm based on Extended Linear Threshold Model (ELTM) was proposed to solve Top-*K* problem in microblog. A concept of influence rate and a WIR (Weibo Influence Rank) algorithm were established to determine the user's influence by summarizing the key factors. Then, based on WIR values, an influence propagation model was proposed. After using greedy algorithm, the Top-*K* nodes were excavated. A simulation test based on Sina micro-blog was performed to validate the effectiveness of the proposed method. The result shows that the method outperforms the traditional algorithm.

Key words: micro-blog; influence measure; PageRank algorithm; influence maximization; greedy algorithm

0　引言

微博作为一种迷你型博客,在近年来得到了广泛的应用。据统计,Twitter 当前在全球多个国家拥有超过 18 种语言的用户近 2 亿,中国现有的四大微博即新浪微博、腾讯微博、搜狐微博、网易微博的发展呈爆炸式状态,据 CNNIC 统计,2012 年微博注册人数已超过 3 亿。与传统社会网络中通过互相认证的好友关系建立拓扑结构不同,微博是通过"关注"行为构成了具有广播性质的信息扩散网络,其信息传播的速度、广度和效率都得到了极大的提高。微博已经成为消息扩散和舆论传播的主要平台。因此,在微博中具有影响力的少数用户非常值得关注,这部分用户在信息传播、舆论形成中起到关键作用。挖掘微博网络中的影响力节点,解决微博网络中的影响力最大化问题在市场营销、舆情管控等方面具有重要意义。

社会网络中的影响力研究由来已久,Richardson 等[1]将影响力最大化问题定义为如何选择 *K* 个初始节点使最终的影响力扩散范围最大化。Kempe 等[2]在线性阈值模型(Linear Threshold Model,LTM)的基础上提出了一种自然的爬山贪心算法,它在每一步都选择当前"最具影响力"的节点作为初始传播对象进行传播。所谓"最具影响力"的节点,即是当前能够激活最多节点的节点。但是贪心算法也存在着明显的缺陷,在数据规模较大的情况下,贪心算法的时间复杂度极高。针对这个问题,Leskovec 等[3]进行大量工作后提出 CELF 改进算法将算法执行效率提高了数百倍,Chen 等[4]也提出自己的改进算法可以以较高的运算效率在大规模数据集上进行计算。Narayanam 等[5]提出了基于合作博弈的 Shapley 值解概念的 SPIN 算法,大幅度提高了计算影响力最大节点集合的效率。田家堂等[6]提出了一种两步骤的启发式算法,以此提高运行效率。近年来,随着微博应用的兴起,挖掘微博网络中的影响力用户成为研究者关注的热点。初期的研究集中在对微博影响力的定性分析及定义上[7-8]。Cha 等[9]使用粉丝数量和微博转发数量对用户影响力进行了衡量,结果表明粉丝数量多的用户微博不一定会得到很多的转发或者评论。郭浩等[10]基于用户消息传播范围对用户影响力进行量化定义,并给出用户影响力的计算方法。随着研究的深入,当前的研究大多借鉴了 PageRank 算法的思想,对微博中的用户影响力进行排名。Weng 等[11]利用 PageRank 算法的思想,设计了 TwitterRank 算法来衡量一个用户在某一主题内的影响力。杨长春等[12]引入了博主传播能力的概念,提出 InfluenceRank 算法来评估博主影响力。

目前关于微博的影响力研究主要集中在对微博用户的影响力度量排序上,这种度量方法由于在结果上的聚合特性以及传播范围上的重叠性,忽视了微博网络中的弱连接结构,无法挖掘出使影响范围最大化的节点,因此不能解决微博网络

收稿日期:2013-02-04;**修回日期**:2013-05-08。　**基金项目**:国家 863 计划项目(2011AA7116031,2011AA010604)。
作者简介:吴凯(1988 -),男,河北武安人,硕士研究生,主要研究方向:社会网络分析;　季新生(1968 -),男,江苏如东人,教授,博士,主要研究方向:移动通信;　郭进时(1987 -),女,吉林四平人,硕士研究生,主要研究方向:社会网络分析;　刘彩霞(1974 -),女,山东烟台人,副教授,博士,主要研究方向:移动通信、社会网络分析。

中的影响力最大化问题。而社会网络中的影响力最大化算法由于没有建立可以体现微博特征的影响力传播模型,并不适用于微博网络。针对这一问题,本文的解决思路是将微博中的影响力度量与 Top-K 节点挖掘算法相结合,具体为:1)通过引入影响率的概念建立一种 WIR(Weibo Influence Rank)算法对微博用户影响力进行度量;2)利用 WIR 值建立一种新的扩展的线性阈值传播模型,并在此基础上运用贪婪算法,最终形成基于微博网络的影响力最大化算法。

1 微博用户影响力度量

用户的影响力本质上是用户之间的相互作用,一个用户能够对其他用户发生的作用越大,该用户的影响力也越大。社会网络中的用户影响力定义多以节点度数为依据,度数大的节点发布的信息将被更多的用户接收到,因此具有更大的影响力。这种定义方式无法体现出微博平台的应用特点,本文将微博网络中的影响力定义如下:

定义 微博用户影响力。微博网络中的用户影响力体现为一个用户通过发布微博行为激发另一用户发生评论或转发行为的潜力。

依据微博用户影响力的定义,本章将首先分析决定影响力的关键因素,并在此基础上综合考虑用户之间的影响程度和用户活跃程度,借鉴 PageRank 算法的思想,提出一种 WIR 影响力度量算法。

1.1 微博用户影响力关键因素分析

决定微博用户的影响力的关键因素有:

1)用户间亲密程度。两用户之间的历史转发及评论数体现了两个用户之间的亲密程度和影响能力,越高的历史转发及评论数说明未来用户之间产生信息行为的可能性也越大。

2)用户活跃程度。用户的活跃度可以表示为用户单位时间内发表的微博数量。有的用户经常发表微博或者发表评论,有的用户则很少发微博。因此,用户的活跃度反映了用户的参与程度、积极程度。活跃程度越高,发布的微博被转发或评论的机会越大,影响其他用户的可能性也越大。

3)用户粉丝数及粉丝的影响力。粉丝数量是微博用户影响力的一个重要因素。一般来说,一个用户的粉丝越多,那么该用户的影响力也越大,用户的粉丝越多,激发另一用户产生信息行为的潜力就越大,越容易影响到其他人,且所花费的代价也越小,反之亦然。同时,粉丝与粉丝之间也不能同等对待,粉丝自身的影响力以及网络的用户规模也是衡量用户影响力的重要指标。

1.2 WIR 算法描述

定义微博网络为有向网络 $G = \langle V, E \rangle$,其中:V 表示用户节点集合;E 表示通过用户间的关注关系形成的边集合,边的方向为关注者指向被关注者。定义节点 i 所关注的节点集合为

$$A(i) = \{j \mid (i,j) \in E\}$$

定义节点 j 的粉丝节点集合为

$$N(j) = \{i \mid (i,j) \in E\}$$

为了衡量邻居节点 $j \in A(i)$ 对 i 的影响能力,本文提出节点对之间影响率的概念,如下所示:

$$I(i,j) = act(i,j) * fri(i,j) \tag{1}$$

其中:$act(i,j)$ 表示节点 j 在活跃程度上对 i 的影响程度,令 $|T_j|$ 表示用户 j 在一段时间所发布的微博总数,$\sum_{a \in A(i)} |T_a|$ 表示节点 i 所关注的节点发布微博数量之和,则

$$act(i,j) = \frac{|T_j|}{\sum_{a \in A(i)} |T_a|} \tag{2}$$

$fri(i,j)$ 表示节点 j 对 i 在转发及评论行为上的影响程度,定义如下:

$$fri(i,j) = \frac{RC(i,j) + 1}{\sum_{a \in A(i)} RC(i,a) + 1} \tag{3}$$

其中 $RC(i,j)$ 表示用户 i 在一段时间内对用户 j 转发及评论的数量。式(2)体现了在 i 所关注的所有用户中,j 对 i 在信息行为上的影响程度。

影响率 $I(i,j)$ 从用户活跃度和历史转发评论数两个方面建立了用户间的影响力度量标准,$I(i,j)$ 的值越大,反映了节点 j 对 i 的影响力越大,然而为了对微博用户的影响力进行准确地度量,还应考虑到节点拓扑结构的因素。PageRank 算法是衡量网络中节点重要程度的经典算法,算法的核心思想是每个节点的值根据反向链接的数量均匀流向所有的关系节点,每个节点的 PR 值为所有邻居好友对其贡献值的综合,如果网页 X 链接到网页 Y,则认为网页 X 给网页 Y 投了一票。但是 PageRank 不仅仅只统计网页的得票数,它也会参考那些投票的网页的重要性,那些重要网页投出的选票要比一般网页投出的选票高,这与决定用户影响力的粉丝因素相似。PageRank 算法可以很好地体现网络拓扑结构对在节点影响力上的作用。因此,借鉴 PageRank 算法的核心思想,本文提出了 WIR 影响力度量算法,定义如下:

$$WIR(j) = d + (1-d) \sum_{i \in N(j)} [S_{ij} * WIR(i)] \tag{4}$$

其中:$WIR(j)$ 是用户 j 的 WIR 值;$N(j)$ 是节点 j 的粉丝集合;S_{ij} 是节点 i 的影响力分配给节点 j 的比例因子,体现了节点 j 在所有影响节点 i 的节点中所占的比重,表达式如式(5),

$$S_{ij} = \frac{I(i,j)}{\sum_{a \in A(i)} I(i,a)} \tag{5}$$

d 是阻尼系数,可以设定在(0,1)范围。将所有节点的初始 WIR 值设为 0.1,通过迭代可以得到所有用户的 WIR 值。

2 微博网络影响力最大化算法

影响力最大化问题定义为如何选择 K 个初始节点使得最终的传播影响范围最大化,由上一章的分析可知,通过计算用户的 WIR 值,可以得到用户的影响力排序,然而若直接选择 WIR 值靠前的 K 个节点作为初始节点,并不能保证最终的影响范围最大化,这是由于微博用户的粉丝影响力越大,则用户影响力也就越大,导致 K 个节点有很大的可能性聚集在同一簇内,忽视了网络结构中的弱连接节点。如图 1 所示,节点[1,2,3,4,5,6]构成了一个社团,其社团中的节点度数相对较高,因此影响力度量算法更容易将此社团内的节点排序到前几位,然而这样得到的结果并不能使影响范围最大化,因为其忽视了节点[8,9,10]所构成的一些较小社团。

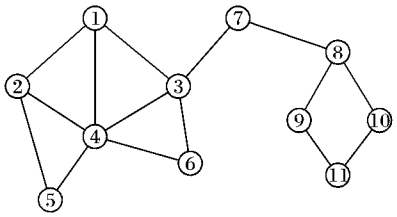

图 1 微博网络结构示意图

本章将首先利用 WIR 值,建立一种扩展的线性阈值模型

(Extended Linear Threshold Model,ELTM),并在此基础上提出引入贪婪算法来解决微博网络中的影响力最大化问题。

2.1　影响力传播模型

线性阈值模型是所有基于节点特异性阈值模型的核心。给定一个社会网络 $G=\langle V,E\rangle$,定义 $N(v)$ 为节点 v 的邻居节点集合。被激活的节点 u 对邻居节点 v 存在影响 b_{uv},一个节点 v 的所有邻居节点对 v 的影响力总和小于等于1。定义 $A(v)$ 为节点 v 的邻居节点中已激活的节点集合。每个节点 v 有一个特异性阈值 θ_v,如果满足 $\sum_{u\in A(v)} b_{uv}\geqslant\theta_v$,则 v 被激活。

线性阈值模型中的 b_{uv} 体现了激活的节点 u 对邻接节点 v 产生的影响,在社会网络的分析中,b_{uv} 通常用 $1/d(v)$ 来表示,$d(v)$ 为节点 v 的度数。这种表示方式意味着 v 的邻居节点对它的影响力都相同,这并不符合现实情况,也没有体现出微博网络的特性,结合微博用户的影响力度量算法,本文给出一种新的 b_{uv} 计算方法,表达式如下:

$$b_{uv}=\frac{WIR(u)}{\sum_{a\in N(v)} WIR(a)} \tag{6}$$

其中:$N(v)$ 表示节点 v 所关注的节点集合;b_{uv} 体现了用户 u 的影响力在集合 $N(v)$ 中所占的比重大小;v 被激活的概率由其邻居节点中所有已激活节点的影响力大小所决定,影响力越大,v 被激活的概率就越大。

2.2　基于 ELTM 的微博网络影响力最大化算法

改进后的线性阈值模型体现了微博特征,节点之间的信息传递概率取决于节点的传播影响力,在 ELTM 的基础上运用贪婪算法可以实现针对微博网络的影响力最大化算法,基于此,本文提出了基于 ELTM 的贪婪算法(Greedy Algorithm Based on ELTM,GABE)。算法的核心思想包括三个阶段:

1)利用式(4)通过迭代计算网络中每个节点的 WIR 值,建立微博网络的影响力度量模型;

2)利用得到的 WIR 值,通过式(6)计算网络中每条边的影响力权值 b_{uv},构建扩展的线性阈值模型;

3)贪心阶段,在 ELTM 的基础上运用贪婪算法,每一步都选取使传播影响范围增量最大的节点,最终挖掘出种子节点集合。

定义微博网络为 $G=\langle V,E\rangle$,S 为包含 K 个节点的种子集合,s_v 为节点 v 一次扩散得到的传播范围,$IS(S)$ 为种子集 S 最终的影响范围,则基于 ELTM 的贪婪算法如下所示。

算法　GABE。

输入:Weibo network $G(V,E)$,K;
输出:Top-K nodes set S。

```
1)  Initialize S = ∅, R = 1000
2)  for each Vertex v in V do
3)    Calculate WIR(v)
4)  end for
5)  for each edge (u,v) in E do
6)    Calculate b_uv
7)  end for
8)  for i = 1 to K do
9)    for each Vertex v in V
10)     s_v = 0
11)     for j = 1 to R
12)       s_v += |IS{v}|
13)     end for
14)     s_v = s_v/R
15)   end for
16)   S = S ∪ { arg max_v {s_v} }
17) end for
18) end
```

3　实验仿真

3.1　数据集描述

为了验证本文提出的微博影响力最大化算法的有效性,本文选取了新浪微博中的“微群”数据进行实验验证。微群是微博群的简称,能够聚合具有相同爱好或者相同标签的用户,将所有与之相应的话题全部聚拢在微群里面。因此,同一微群里的微博用户具有较高的聚合度和活跃程度,适合作为影响力分析的数据源。

本文首先利用 Web 爬虫技术采集了某一微群内所有成员的用户 ID,之后利用新浪提供的 API 接口采集对应 ID 的相关数据,具体包括:

1)各用户 ID 对应的用户信息,包括用户名称、关注数、粉丝数、发布微博数等;

2)用户的关注关系,包括用户所关注的其他用户 ID,仅限制在收集该微群内的用户;

3)用户的转发及评论信息,包括被转发的消息 ID,被转发及评论的用户 ID 仅限制在收集该微群内的用户。

基于采集到的数据,构建了微群中的关注网络,数据集中包含了 3 694 个用户节点以及 14 624 条关注关系组成的边。

3.2　对比算法

为了验证本文提出的 GABE 在微博网络影响力最大化问题上的有效性,采用以下三种常用影响力最大化算法或微博影响力度量方法作为对照:

1)KKT 算法:Kemple 和 Kleinberg 提出的一种自然的爬山贪心算法,算法的每一步都选择当前最有影响力的节点放入种子集合中,将这种算法作用到线性阈值模型中就形成了当前在社会网络最大化问题中常用到的 KKT 算法。

2)PageRank 算法:PageRank 算法是常用的影响力度量算法,其影响力的分配依据节点的度数大小。

3)粉丝数排名(Followers):依据用户的粉丝数目对用户影响力进行排序。

3.3　实验结果

在传统的影响力最大化算法研究中,挖掘到的 Top-K 节点只在影响力传播模型上仿真其覆盖效果,本文将评估 GABE 及其他对比算法挖掘到的 Top-K 节点在真实网络中的传播覆盖效果,以此验证 GABE 的有效性。

定义信息传播的实际影响覆盖人数 M 为节点传播能力的评价指标,如下所示:

$$M_i=\frac{1}{N}\sum_{j=1}^{N} RT_j \tag{7}$$

其中:RT_j 为第 i 个用户所发布的第 j 条微博的传播覆盖用户数,N 为统计得到的用户 i 在一段时间内发布的微博数。

将四种算法挖掘到的 Top-K 节点在数据集中的影响力传播覆盖范围进行对比,结果如图 2 所示。

从图 2 中可以看出,利用 GABE 选取的初始节点在影响范围上优于常用的 KKT 算法,节点数量高出了 7.7%。这说明 GABE 通过引入改进的线性阈值模型有效地模拟了微博网络中的信息传播过程,进而挖掘出了在微博社区中具有更高传播影响力的节点。此外,PageRank 算法与依据粉丝数排序

的方法在效果上非常接近,这是因为两者都是以节点度数作为核心衡量指标。

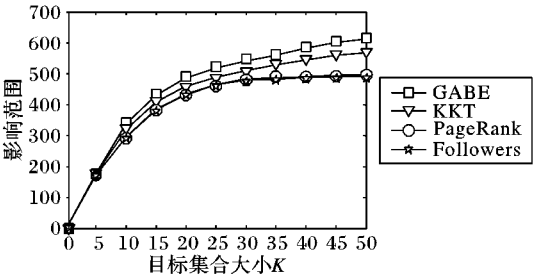

图 2 四种算法的影响范围对比

由图 2 的仿真结果还可发现,PageRank 算法的效果要明显差于 GABE,这说明简单的影响力度量方法由于没有考虑到社团之间的弱纽带链接节点并不能解决影响力最大化的问题。为了更好地说明影响力最大化问题与影响力度量的不同,本文定义了相似性比较函数 $F(K)$,如下所示:

$$F(K) = |N(K) \cap N'(K)| / K \tag{8}$$

其中:$N(K)$ 与 $N'(K)$ 分别表示所比较的两个方法的 Top-K 节点集合,$|N(K) \cap N'(K)|$ 表示 $N(K)$ 与 $N'(K)$ 中相同节点的个数。比较本文提出的 GABE 与 WIR 算法挖掘到的 K 个节点,得到结果如图 3 所示。

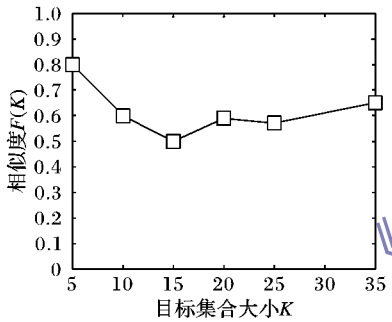

图 3 相似度比较

图 3 的结果显示,GABE 与 WIR 算法结果之间的相似度随着目标集合数目 K 的增加呈现先下降后平稳增加的趋势。这说明当目标集合 K 值较小时,影响力度量的 WIR 算法得到的节点基本可以实现最大化的问题,然而随着 K 值的增加,WIR 算法的前 K 个节点更加趋于聚合,而 GABE 得到的 K 个节点则包含了更多的社团间的弱纽带节点;当 K 值继续增加时,未挖掘到的弱纽带节点逐步减少,相似度呈现平稳增加的趋势。

本文还对数据集中所有节点的 WIR 值进行了统计,得到的结果如图 4 所示,可以看出 WIR 值的分布呈现出幂率特性,这也印证了微博网络中只有少数用户具有较高影响力的结论。

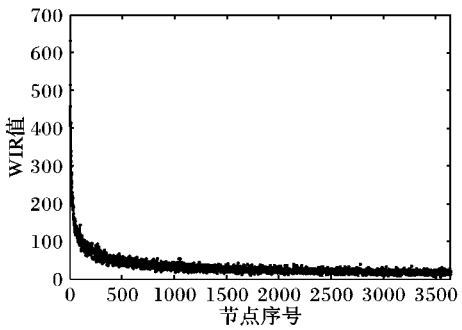

图 4 微博用户 WIR 值统计

4 结语

为解决微博网络中影响力最大化这一问题,本文首先提出微博用户影响力度量的 WIR 算法,构建了符合微博影响力传播特征的扩展的线性阈值模型,进而建立了基于微博网络的影响力最大化算法 GABE。在真实微博数据集中的实验结果表明,GABE 可以较好地解决现有微博影响力排序结果中的范围重叠问题,并且相比常用 Top-K 节点挖掘算法在微博网络上的影响范围上有很好的扩大。后续将在 GABE 的时间复杂度优化方面做进一步研究,以实现效率更高的微博网络影响力最大化算法。

参考文献:

[1] RICHARDSON M, DOMINGOS P. Mining knowledge-sharing sites for viral marketings [C]// KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discoveryand Data Mining. New York: ACM, 2002: 61 - 70.

[2] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social networks [C]// KDD '03: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003: 137 - 146.

[3] LESKOVEC J, KRAUSE A, GUESTRIN C, *et al*. Cost-effective outbreak detection in networkss [C]// KDD '07: Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 420 - 429.

[4] CHEN W, WANG C, WANG Y J. Scalable influence maximization for prevalent viral marketing in large-scale social network [C]// KDD '10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010: 807 - 816.

[5] NARAYANAM R, NARAHARI Y. A shapley value-based approach to discover influential nodes in social networks[J]. IEEE Transactions on Automation Science and Engineering, 2011, 8(1): 130 - 147.

[6] 田家堂, 王轶彤, 冯小军. 一种新型的社会网络影响最大化算法[J]. 计算机学报, 2011, 34(10): 1956 - 1964.

[7] BAKSHY E, HOFMAN J M, MASON W A, *et al*. Everyone's an influencer: quantifying influence on twitter [C]// WSDM '11: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. New York: ACM, 2011: 65 - 74.

[8] KWAK H, LEE C, PARK H, *et al*. What is twitter, a social network or a news media? [C]// WWW'10: Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 591 - 600.

[9] CHA M, HADDADI H, BENEVENUTO F, *et al*. Measuring user influence in twitter: the million follower fallacy [C]// Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, DC: AAAI, 2010: 10 - 17.

[10] 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): 1 - 6.

[11] WENG J, LIM E-P, JIANG J, *et al*. TwitterRank: finding topic sensitive influential twitterers [C]// WSDM '10: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261 - 270.

[12] 杨长春, 俞克非, 叶施仁, 等. 一种新的中文微博社区博主影响力的评估方法[J]. 计算机工程与应用, 2012, 38(25): 229 - 233.