Humanities and Social Sciences
2015; 3(5): 234-239
Published online November 17, 2015 (http://www.sciencepublishinggroup.com/j/hss)
doi: 10.11648/j.hss.20150305.21
ISSN: 2330-8176 (Print); ISSN: 2330-8184 (Online)

The Evaluation of the Social Network's Nodes Influence Based on Users' Behavior

Fudong Wang[1, *], Pei Wang[1], Sunzeng Yao[2]

[1]Gloria Sun Business School, Donghua University, Shanghai, China
[2]Postgraduate Department, Donghua University, Shanghai, China

Email address:
wfd@dhu.edu.cn (Fudong Wang), xiaopei0602@163.com (Pei Wang), Sunzengyao@dhu.edu.cn (Sunzeng Yao)

To cite this article:
Fudong Wang, Pei Wang, Sunzeng Yao. The Evaluation of the Social Network's Nodes Influence Based on Users' Behavior. *Humanities and Social Sciences.* Vol. 3, No. 5, 2015, pp. 234-239. doi: 10.11648/j.hss.20150305.21

Abstract: How to effectively identifying opinion leaders has become a hot research point. It's essence is to identify the nodes with the strong influence in social network. This paper analyzes the behaviors of users in social networks and proposes the impact evaluation algorithm based on multi-angle user behaviors which is User-Activity Rank algorithm. The algorithm uses the basic idea of PageRank algorithm, considers user's creativity, interactivity and content quality and designs the uneven distribution mechanism between the users' UA Rank value, making the computation nodes more accurate. This paper uses Car Home Forum Case for analysis and finds that this algorithm can be more accurate and objective. User-Activity Rank algorithm can help companies improve the accuracy of identifying opinion leaders and promote network marketing with their influence.

Keywords: Social Network, Opinion Leaders, PageRank, Users' Behavior

基于用户行为的社交网络节点影响力评价

王扶东[1, *]，王佩[1]，孙增耀[2]

[1]旭日工商管理学院，东华大学，上海，中国
[2]研究生部，东华大学，上海，中国

邮箱
wfd@dhu.edu.cn（王扶东），xiaopei0602@163.com（王佩），Sunzengyao@dhu.edu.cn（孙增耀）

摘要：有效地识别意见领袖是社交网络研究中备受关注的焦点问题，其本质即为识别社交网络中影响力较强的节点。本文深入分析了用户在社交网络中的各项行为，提出了基于多角度用户行为的社交网络节点影响力评价算法即User-Activity Rank算法。该算法借鉴PageRank算法的基本思想，综合考虑用户创造力、互动性及发布内容质量等，设计了用户间UA Rank值的不均匀分配机制，使社交网络节点影响力的计算更加准确。最后以汽车之家论坛为样本进行了实例分析，发现该算法能更加准确、客观地评估社交网络节点影响力。User-Activity Rank算法可帮助企业提高识别意见领袖的准确性，并借助其影响力进行网络营销。

关键词：社交网络，意见领袖，PageRank，用户行为

1．引言

随着互联网的发展，社交网络逐渐成为用户信息发布、分享、沟通交流的平台。在这个虚拟在线网络中，人们如同在现实社会中一样，从事着大量的社交活动，在这个过程中逐渐形成一些中心节点，他们的行为会对网络中的其他用户行为认知和决策起到一定的导向作用，可以将这些中心节点称为网络中的意见领袖。在营销领域快速识别出社交网络中的意见领袖，可以帮助企业通过这些意见领袖

进行品牌宣传与产品推广，这既降低了用户的搜寻成本也帮助企业提高了广告投入效率从而增加企业收益。因此，意见领袖的识别对于社会管理、商业营销等方面都具有广泛的应用和意义。

近年来，如何有效地识别意见领袖已逐渐成为社交网络中的研究热点。假如用户在社交网络中拥有领袖地位那么表示与该用户相对应的节点在网络结构中具有相同重要地位，因此可通过评估社交网络中节点的影响力来科学地识别意见领袖。许多学者[1-2]通过经典的PageRank算法对网络中的节点进行度量，将PageRank算法中分数最高的1%作为意见领袖。Song[4]等学者提出InfluenceRank算法，其在博客数据集上考虑了博文新颖度对网络的贡献，研究基于新颖度的影响力个体发现方法。Li[5]等学者依靠微博中的历史消息和社会交互记录，利用统计学习过程构造历史意见和意见影响力，提出话题级的意见影响力模型，并合并了话题因素和社会影响力。Weng等学者[6]提出了TwitterRank算法，该算法测量了网络中用户的节点地位并对研究了用户之间关注主题的相似度来对意见领袖的影响力进行度量。Agarwal等学者[7]对用户发布博客的引用数量、评论数量、内容长度和新颖程度进行影响力分析，进一步对意见领袖进行识别。以上学者们的研究使意见领袖的识别在理论和方法上更高效，不足的是社交网络上用户从事着大量的社交活动，触发各种动作，单一地抓住用户的某个行为，从单个角度对节点影响力进行评价会降低意见领袖识别的准确性。

针对上述问题，本文借鉴PageRank算法的思想，深入分析了用户在社交网络中的各项行为，针对用户行为特性，构建用户活跃度综合评价模型，进而设计用户间相对关注程度，即UA Rank值分配的机制，最终提出新的意见领袖识别算法——User-Activity Rank算法（简称UAR算法）。该算法根据社交网络上用户之间的关注与被关注关系，综合考虑了用户的创造力、互动性与发布内容质量，考虑了用户触发的各类动作对其地位的影响，让算法更好地反映客观实际。为了使研究有具体的应用背景，有真实的数据支持，本文还以汽车之家论坛为样本进行了实例分析，发现该算法能更加准确、客观地评估社交网络节点影响力。

2．研究模型

2.1．PageRank算法思想

PageRank算法是由Larry Page和Sergey Brin提出的衡量网页重要性的算法，该算法通过计算网页的PageRank值（简称PR值）来量化网站所获得内部链接和外部链接的重要价值，即链接重要度。在PageRank算法的思想中Web上的每个页面都有属于自己的职能与地位，页面之间的链接关系构成了一个完整的生态网络，算法中将每个页面的PR值均匀地分配到该页面指向的网页，通过反复迭代，网络中的页面PR值达到稳定、收敛状态，最终通过对页面PR值进行排序从而对网页的链接重要度进行排序。

将$V=\{v_1,v_2,\cdots,v_3\}$表示页面集合，v_i为任意一个页面，$E(v_i)$为链入页面v_i的链接页面集合，v_j为链入到v_i的其中一个页面，$N(v_j)$为页面v_j链接到其他页面的链接数量，d为阻尼系数，表示用户在浏览某个页面后以d的概率继续浏览其中一个链出的页面，以$\frac{1-d}{n}$的概率重新选择一个随机的页面浏览，n为网络中网页的总数，随机浏览模型更加接近于用户的浏览行为，一定程度上解决了Rank Leak和Rank Sink的问题，并保证PageRank具有唯一值。则页面v_i的PR值$PR(v_i)$可用公式（1）表示：

$$PR(v_i)=\frac{1-d}{n}+d\sum_{v_j\in E(v_i)}\frac{PR(v_j)}{N(v_j)} \qquad (1)$$

本文选择借鉴PageRank算法的原因有三点：1、网络结构类似。PageRank算法的思想源于社会网络分析，社交网络中用户所构成的网络是一个有向图，用户的关注行为与被关注行为就如web页面中链出和链入；2、评价思想类似。用户影响力的评估本质上就是对用户在社交网络上的地位进行排序，这与PageRank算法的思想是一致的；3、算法有效性。PageRank算法在Google搜索引擎上取得巨大成功，证明其在网页重要性评价方面的有效性。

因此，本文借鉴PageRank算法的思想，将网页间的链出与链入关系转化为用户间的关注与被关注关系，针对用户行为特性，构建用户活跃度综合评价模型，进而设计用户间相对关注程度，即UA Rank值分配的机制，更加客观地评估节点影响力。

2.2．用户活跃度综合评价模型建立

传统的PageRank算法中，网页根据链接关系，将PR值均匀地传递给每一个链出的页面，但在真实的社交网络中，用户不可能对所有的粉丝投入一样的关注程度。实际上，用户在社交网络上存在丰富的动作，例如浏览、评论、回复、发布内容等，这些行为对其地位均有影响。因此，为了避免了传统PageRank算法的局限性，本文在构建User-Activity Rank算法过程中，运用多指标综合评价方法，考虑了用户的行为特征，选择有代表性的多个系数综合成一个指数，从而对用户的在网络结构中的活跃度做出综合的评价。

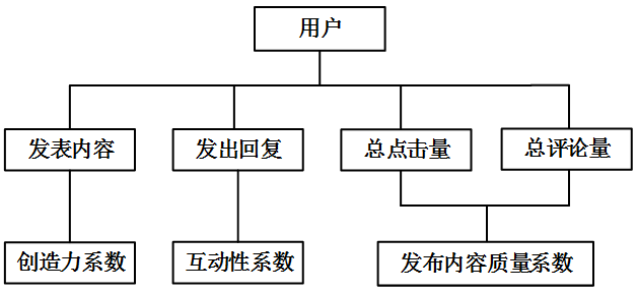

图1 多指标评价方法。

以汽车之家（http://www.autohome.com.cn）论坛为例，论坛里的用户可以发表主题、浏览主题、评论及回复主题，通过研究用户的行为，可将用户触发的行为转化为

三个系数（如图1所示）：用户发布内容情况可以转化为其创造力系数；用户与其他用户之间的评论交流情况可以转化为其互动性系数；用户发布内容后获得的总点击量和总评论量可以转化为其发布内容质量系数，综合以上三个系数，即可转换为用户在社交网络上的活跃度，活跃度是分配UA Rank值的重要依据。

2.2.1. 用户创造力系数

在社交网络中，用户最基本的行为是浏览和发表内容，用户在社交网络中的影响力也与其发表内容有关，如果该用户在某段时间内发表的内容较多，说明该用户在这段时间内具有创造力，其在信息传播中的影响力应比那些缺乏创造力的用户高。用户创造力系数定义为在统一的时间范围内用户发表内容的频率，将时间进行统一是为了避免有些用户只在某一小段时间内发表大量内容，但在其他大部分时间未表现出创造力，从而更客观地刻画用户的创造力。

设社交网络中的一个节点为用户 i，N_i 指用户 i 在统一的时间 T 内发表的内容数，则用户 i 的创造力系数 C_i 为可用公式（2）表示：

$$C_i = \frac{N_i}{T} \tag{2}$$

2.2.2. 用户互动性系数

由于意见领袖需要制造话题，引导用户参与其中，进行平等的对话、交流，在潜移默化中达到广告营销、企业品牌宣传等目的，因此在节点影响力评估过程中，需要评估节点的互动性系数，互动性系数代表节点与其他节点发生的互动性情况，即用户对社交网络上其他用户发出的评论情况。

设社交网络中的一个节点为用户 i，M_i 为用户 i 在统一的时间 T 内对其他用户发出的评论数，则用户 i 互动性系数 I_i 可用公式（3）表示：

$$I_i = \frac{M_i}{T} \tag{3}$$

2.2.3. 用户发布内容质量系数

内容质量系数表示为用户平均每篇主题获得的点击与评论的情况，该系数的设定是为了避免有些用户通过高频率发布一些没有价值的内容来提高网络排名的情况，一般来说，内容质量系数越高，代表该用户发布的内容引起更多的阅读者关注，带来的影响力更广。

设社交网络中的一个节点为用户 i，R_i 为用户 i 在同一时间 T 内发布的内容获得的点击总数量，C_i 为用户 i 在同一时间 T 内发布的内容获得的评论总数量，N_i 为用户 i 在统一的时间 T 内发表的内容数，则用户 i 的发布内容的质量系数 Q_i 可用公式（4）表示：

$$Q_i = \frac{R_i + C_i}{N_i} \tag{4}$$

2.2.4. 用户活跃度综合评价指数

综上无论是用户创造性系数、用户互动性系数、用户发布内容的质量系数，都能很好地反应用户在社交网络上的影响力，将三个指标综合成一个指数需要完成两步：无量纲化的指标转化与指标权数的构造，前者是因为现实情况中各指标往往具有不同的计量单位，所以第一步需要消除量纲与量纲单位的影响，无量纲化的指标转化，是对数据的标准化处理；后者是因为不同的指标对最终评价的贡献度是不同的，需赋予不同的权重。因此将三个指标作归一化处理，取值均为(0,1]，且三个指标对UA Rank值的分配同等重要，权重各设定为1。

设社交网络中的一个节点为用户 i，$\max C_i$ 用户中创造力系数的最大值，$\max I_i$ 用户中互动性系数的最大值，$\max Q_i$ 用户中内容的质量系数的最大值，则用户 i 的用户活跃度 A_i 可由公式（5）表示：

$$A_i = \frac{C_i}{\max C_i} + \frac{I_i}{\max I_i} + \frac{Q_i}{\max Q_i} \tag{5}$$

2.3. User-Activity Rank算法构建

在UAR算法中认为，用户创造力越高、互动性越强、发布的内容质量越好，则该节点在社交网络上的活跃度越大，影响力越广，那么他的粉丝对他的关注度也较高，分配给该用户的UA Rank值也相对较高；反之，如果一个用户的活跃度较低，则他的粉丝对他的关注程序也较弱，那么该粉丝分配相对较少的UA Rank值给该用户。

现假设用户 v 关注 m 个用户，u 是其中一个，设 $A(v,u)$ 是用户 v 分配给用户 u 的UA Rank值的比例，该值由用户 u 的活跃度 A_u 来决定，即公式（6）所表示：

$$A(v,u) = \frac{A_u}{\sum_{i=1}^{m} A_i} \tag{6}$$

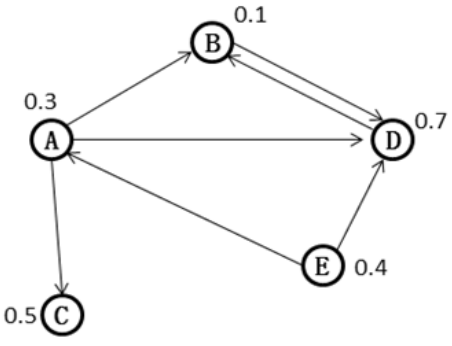

图2 UAR值分配比例示意图。

使用图2所示的网络结构来说明UA Rank值的分配具体是如何计算，假设该网络中有A、B、C、D、E五个用户，根据用户之间的关注与被关注情况，构建如下网络结构图，利用公式3-7计算每个用户的活跃度 A_u，每个用户的活跃

度已在图中标志出。已知A的好友分别是B、C、D，则A分配给B的UAR值的比例应该为B的活跃度在A所有好友的活跃度总值中所占的比例，即 $\frac{A_B}{A_B+A_C+A_d}$，为 $\frac{0.1}{0.1+0.7+0.5}=\frac{1}{13}$，相应的C、D分配的比例 $\frac{0.5}{0.1+0.7+0.5}=\frac{5}{13}$，$\frac{0.7}{0.1+0.7+0.5}=\frac{7}{13}$。

根据以上的UA Rank值分配模型，将该社会网络图定义为邻接矩阵G，该矩阵中每一个顶点代表一个用户，假设其中用户总数为 n。若用户 v 关注了用户 u 则矩阵中 $g_{vu}=1$，否则 $g_{vu}=0$。从而矩阵G变为一个 $n*n$ 的0,1方阵，设 $B(u)$ 为关注用户 u 的粉丝集合，$A(v,u)$ 为用户 v 分配给用户 u 的UA Rank值的比例，n 为网络中网页的总数，d 为阻尼系数，则User-Activity Rank算法可由公式（7）表示：

$$UAR(u)=\frac{1-d}{n}+d\sum_{u\in B(u)}A(v,u)UAR(v) \qquad (7)$$

UAR算法表明用户之间的关注程度不是一致的，且该关注程度由用户的活跃度决定。这个分配机制符合现实生活中的状态：一个人不可能对所有人都一视同仁，我们会对团体中活跃的个体投以更多的关注。

3．实例分析

3.1．实验数据收集与处理

本文选取汽车之家网站的论坛栏目作为研究对象。由于意见领袖具有天然的领域性，其影响力和权威性受限于其个人的兴趣领域，因此，选取垂直型的汽车论坛，可提高意见领袖识别的准确性。目前在社交网络上获得数据的方式主要为利用网络爬虫程序（Crawler）通过程序模拟用户登录页面的操作，直接访问Web页面，得到HTML格式的数据，将HTML文本读取到内存中，然后通过正则表达式匹配等方式进行信息抽取，获得指定的数据。

笔者通过爬虫程序（Python），采用雪球采样的方法执行广度优先搜索（BFS）算法，以在福克斯论坛上的版主“自由的风”为初始用户，获得其好友及跟随者列表，再以该列表为操作对象，分别获取他们的好友及跟随者列表，按照逐层爬取的方法，共爬取了论坛中7893名用户，剔除掉从未有动作的僵尸用户后，共获得5387名有效用户。对这部分用户的出度和入度进行分析，发现用户的关注数量和粉丝数量服从幂率分布，且具有胖尾特性（如图3和图4所示），说明论坛中绝大部分用户的关注和粉丝数量很小，而拥有较大粉丝数量和关注较多用户数量的用户只占所有用户当中的少部分。由于其分布服从 $p(\tau)\propto\tau^{-\gamma}$，表明该论坛具有复杂网络的无标度的网络特性，可以使用社会网络分析方法。

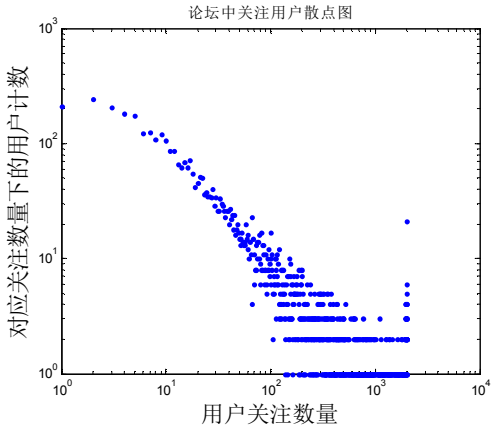

图3 论坛中用户关注散点图

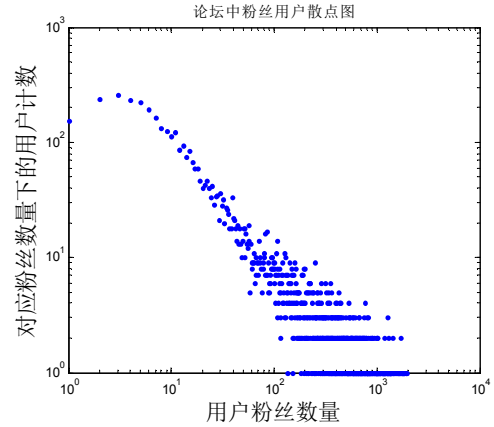

图4 论坛中用户粉丝散点图.

通过爬虫程序（Python）将抓取的数据存入到后台数据库MYSQL中，构建关注矩阵确定用户之间的链接关系，分别计算用户的创造性、互动性、发布内容质量这三个指标系数，得到UA Rank值分配比例，利用Matlab实现UAR算法，其中阻尼系数定为经典的取值0.85，如图5所示，经过77次迭代运行后，最后的结果收敛，每个用户得到趋于稳定的UA Rank值。

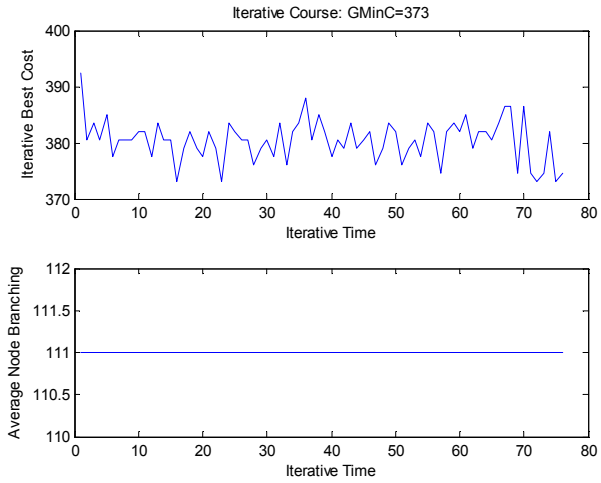

图5 UAR算法计算迭代。

3.2. 结果分析

对收敛后的节点根据UA Rank值进行排序，可得到用户的影响力排名情况，UA Rank值越大，说明其在社交网络上的影响力越大，具体排序如下表1所示：

表1 UA Rank值排名前十的用户列表。

UA Rank值排名	用户ID	用户昵称	关注数	粉丝数	活跃度
1	940342	lmh200801	530	1723	2.18
2	215457	hehe6666	313	960	1.94
3	8318066	兰色多瑙河	980	2030	1.47
4	9021112	fyyd001	792	1172	1.86
5	11598409	欢动新手	574	1161	1.65
6	275921	鬼城来客	606	1661	0.99
7	5064429	云W飞扬	713	1084	1.23
8	5945913	农夫三拳14	334	1334	0.92
9	1833163	雨轩听风	769	1805	0.55
10	4076968	独惟远志	550	1016	0.6

UA Rank值的排序中，对节点影响力的评价不仅限于用户间的链入链出关系，排名第一的lmh200801用户，虽然其粉丝数不是最高，但其UA Rank值远高于其他用户，是因为该用户在实验时间内发布内容较频繁，且发布的内容质量系数较高，因此他的活跃度值较高，为2.18。用户欢动新手虽然粉丝数较少，但是由于其表现相对较为活跃，排名比粉丝数更多的鬼城来客更高。

为了验证算法的有效性，将UAR算法的结果与传统PageRank算法进行对比，如下表2所示。排名第一的仍旧是lmh200801用户，因为其在社交网络上的链入与链出占相对优势，因此排名未发生变化，hehe6666用户从第二名降到十名外，是因为传统PageRank算法中未考虑活跃度，PR值是根据关注行为均匀分布，这也是UAR算法的意义，综合考虑用户在社交网络上的行为，使意见领袖识别算法更加客观、合理。

表2 UAR算法与传统PR算法排序对比。

UAR算法				PR算法			
昵称	关注数	粉丝数	活跃度	昵称	关注数	粉丝数	活跃度
lmh200801	530	1723	2.18	lmh200801	530	1723	2.18
hehe6666	313	960	1.84	兰色多瑙河	980	2030	1.47
兰色多瑙河	980	2030	1.47	fyyd001	792	1172	1.86
fyyd001	792	1172	1.86	鬼城来客	606	1661	0.99
欢动新手	574	1161	1.65	浅色	977	1064	0.68
鬼城来客	606	1661	0.99	欢动新手	574	1161	1.65
云W飞扬	713	1084	1.23	农夫三拳14	334	1334	0.92
农夫三拳14	334	1334	0.92	火焰山88	716	971	0.73
雨轩听风	769	1805	0.55	云W飞扬	713	1084	1.23
独惟远志	550	1016	0.6	快乐人生77	880	998	0.63

4. 结束语

本文根据用户之间的关注与被关注关系，综合考虑用户各种行为对其地位的影响，通过节点本身的特性来分配UA Rank值，从而使UA Rank值的传递是非均匀的，让模型更好地反映客观实际。实证分析中，以汽车之间的论坛为例，综合考虑用户的创造力、互动性、博文发布质量三方面因素，并与传统的PageRank算法进行对比分析，验证新算法的有效性。

通过本文的研究，一方面可以帮用户准确、有效地找到社交网络中的意见领袖，在垂直型网络社区中获取有价值的信息；另一方面，可以使企业通过“意见领袖”制造话题，进行品牌宣传与产品推广，提高企业收益。

同时，本研究也存在一定的局限性，本文虽然综合考虑了用户的浏览、评论、回复、发布内容等动作，但没有考虑到用户之间丰富的交互文本，在社交网站上，这些主观色彩浓厚的评论信息蕴含着大众舆论对目标实体的看法。在下一步的工作中，需对用户之间主观的情感因素进行分析，从而识别出真正得到其他用户支撑的意见领袖。

致谢

本文为中央高校基本科研业务费专项资金项目（项目编号：15D110801）的阶段性成果之一。

参考文献

[1] Hajian B,White T. Modelling influence in a social network: Metrics and evaluation[C]. Proceedings of the 3rd IEEE International Conference on Social Computing. Boston, USA, 2011:497-500.

[2] Tang J, Lou T, Kleinberg J. Inferring social ties across heterogenous networks [C].Proceeding of the 5th ACM International Conference On Web Search And Data Mining. Seattle,USA,2012:743-752.

[3] Hon Wai Lam, Chen Wu. Finding Influential eBay Buyers for Viral Marketing－A Conceptual Model of Buyer Rankl[C]. Proceedings of IEEE Conference on Commerce and Enterprise Computing. IEEE, 2009: 778－785.

[4] Song X,Chi Y,Hino K,Tseng B. Identifying opinion leaders in the blogosphere[C].Proceedings of the 16th ACM International Conference On Information And Knowledge Management .lisbon, Portugal, 2007: 971-974.

[5] Ding Z Y,Jia Y,Zhou B,et al.Mining topical influencers based on the multi-relational network in micro-blogging sites [J]. China Communication, 2013,10(1):93-104.

[6] WENG, Jianshu; LIM, Ee Peng; JIANG, Jing; and He, Qi. Twitterrank: Finding Topic-Sensitive Influential Twitterers [C]. ACM International Conference on Web Search and Data Mining (WSDM 2010).

[7] Agarwal N, Liu H, Tang l, Yu PS. Identifying the influential bloggers in a community[C]. Proceedings of the 1st International Conference on Web Search And Data Mining.Palo Alto,USA,2008: 207-21.

[8] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks [C]. In WSDM'10, 207–217, 2010.

[9] Goyal A,Bonchi F,Lakshmanan L V S.Discovering leaders from community actions [C]. Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Vally, USA,2008:499-508.

[10] Michael M. A brief history of generative models for power law and lognormal distributions [J]. Internet Mathematics, 2004, 1(2): 226-251.

[11] 吴信东、李毅、李磊.在线社交网络影响力分析[J].计算机学报.2014.4。

[12] 韩筱璞，汪秉宏，周涛.人类行为动力学研究[J]. 复杂系统与复杂性科学.2010。