微博用户的影响力指数模型*

原福永[1]　冯　静[1]　符茜茜[1,2]
[1](燕山大学信息科学与工程学院　秦皇岛 066004)
[2](秦皇岛职业技术学院经济系　秦皇岛 066100)

【摘要】以新浪微博为研究对象,提出微博用户的用户影响力指数模型。首先将得到的用户被关注度取代当前存在虚假的用户粉丝数,通过较为合理的用户被关注度计算得到微博用户的用户活跃度与微博影响力,最后将用户活跃度与微博影响力作为用户影响力的影响因子合成微博用户的用户影响力。模型考察用户与用户微博两个角度的多个活动因子。实验结果表明,用户影响力指数模型在降低微博僵尸粉影响的同时,能够较合理地体现微博用户的实际影响力。

【关键词】新浪微博　用户影响力　用户被关注度　活跃指数　微博影响力
【分类号】TP302

Influence Index Model of Micro – blog User

Yuan Fuyong[1]　Feng Jing[1]　Fu Qianqian[1,2]
[1](College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)
[2](Economics Department, Qinhuangdao Institute of Technology, Qinhuangdao 066100, China)

【Abstract】The paper proposes users influence index model based on Sina micro – blog platform. Firstly, new users' attention is used to replace the current false fans value, then the model uses this reasonable value to calculate users' active index and micro – blog influence, at last it obtains the user influence by combining users' active index and micro – blog influence reasonably. The model studies many factors of two aspects of users and users' micro – blogs. The experiment result shows that the model can reduce the false fans' interference and reflect the users' real influence ability.

【Keywords】Sina micro – blog　User influence　User attention　Active index　Micro – blog influence

1　引　言

微博的兴起使很多网络用户开始习惯于从虚拟社区获取信息和表达情感[1],也使微博成为热点事件产生和谈论的重要场所[2]。140 字的信息成为社会舆论热点,信息本身的吸引力和信息在网络中的传播同样重要。用户和用户关系作为网络中的节点和边是信息传播的基础,转发是信息传播的途径。许多微博信息在发表的初始没有引起关注,而在几个或一些网络中心节点用户转发后受到围观,这些中心节点用户即高影响力用户。微博用户影响力作为影响力研究在微博领域的延伸吸引了大量的研究者[3]。高影响力用户的存在和转发是引起信息持续传播和形成更大传播规模的关键因素[4]。网络拓扑结构中的中心节点用户只是潜在的高影响力用户,如果用户不进行活动行为,用户的实际影响力则难以得到体现。当前新浪微博平台使用粉丝数表示用户的受关注度,但严重的僵尸粉问题使得此用户的被关注度并不真实。本文在现有研究成果的基础上以新浪微博为研究对象,提出

收稿日期: 2012 – 05 – 03
收修改稿日期: 2012 – 06 – 10
* 本文系河北省自然科学基金项目"面向协同过滤的可信推荐算法研究"(项目编号: F2011203219)的研究成果之一。

了新的微博用户影响力模型，模型使用基于 PageRank 和用户行为权值相结合的评价方法。从链接分析的角度定义了用户被关注度的概念，分析用户受关注程度的同时也判断了是否为网络中的中心节点用户，即得到用户的潜在影响力；将用户被关注度融入用户活跃度和用户微博影响力，结合用户的行为实现对微博用户实际影响力的研究，模型通过对多个影响指标的综合考察，最终得到微博用户的影响力指数。

2　相关研究

微博用户影响力作为影响力研究在微博领域的延伸始于链接分析[5]，Weng 等[5] 根据 PageRank 设计了基于用户和链接结构的话题相似性的影响力排序算法，算法设计简单易行，但同时存在客观性和准确性的问题；Cha 等[6] 从用户行为的角度出发，通过对微博用户的粉丝、微博被转发和用户被提及（@ 用户名）对用户影响力进行了考察，同时也指出链接分析更多体现了用户之间的一种关系，对于用户影响力的体现只有极少部分。该文认为对用户影响力的考察存在其合理性，但其中僵尸粉问题不容忽视，此外对用户影响力的影响因子的选取与考察不够全面，用户影响力不仅仅体现在微博被转发和用户被提及，还有很多如被评论等。石磊等[7] 针对单一制排名机制的不足，综合考虑了用户的几个主要活动行为，提出了用户活跃度模型，较为合理地展现了微博用户的活跃度，但用户活跃度只考察了用户活跃状态，并不能充分地体现微博用户的用户影响力。Ye 等[8] 将用户粉丝数量影响力、回复影响力、转发影响力、粉丝数、微博数、回复和转发数作为排序的准则进行了计算和比较，认为从回复最多的角度得出的用户影响力值最稳定，并按此进行影响力排序作为标准。李军等[3] 针对微博影响力的评价方法进行了较全面的论述，将微博用户影响力的评价方法归结为不同的 4 类，并根据中文微博的特点对 Yamaguchi 等[9] 的 TURank 算法进行了改进，但对于改进后方法的效果并没有做出验证和评价。

本文以新浪微博为研究对象，采用链接分析与用户行为相结合的方法，提出了微博用户的影响力指数模型，将链接分析得到的用户被关注度取代当前存在的虚假粉丝数，进而对用户活跃度与用户微博受关注度进行考察，最终合成微博用户的用户影响力。实验结果表明该模型较合理地展现了微博用户的用户影响力。

3　微博用户影响力指数模型

微博用户影响力是用户在微博平台的传播和影响能力，主要与用户被关注度、活跃度和微博影响力相关。其中用户被关注度从链接关系的角度展现了用户的潜在影响力，有效地降低了僵尸粉的影响，但并不能从根本上解决僵尸粉问题。本模型将用户的潜在影响力，即用户关注度分别融入微博用户活跃度和用户微博影响力，实现对用户实际影响力的考察，而没有使其作为用户影响力的直接影响因子；用户活跃度是用户影响力产生的动力，意在考察用户在微博平台的活跃状态，主要与用户被关注度、用户关注的增加频率和用户发布微博的频率相关；微博影响力是用户影响力的主要体现方式之一，旨在考察用户所发布微博的传播范围和影响能力，与用户被关注度、微博被转发和评论次数等相关。

3.1　用户被关注度

目前微博平台使用用户粉丝数体现微博用户的受关注程度，用户粉丝作为用户主页的入链，可以体现用户的潜在影响力。但僵尸粉问题使得当前用户关注度并不真实。本文从链接分析的角度给出用户被关注度的概念和计算方法，较为合理地展现用户的受关注程度和用户潜在影响力。

微博用户以个人主页形式存在，每位粉丝都是用户主页的一条入链，表示对用户的支持，用户自身的关注则是用户主页的出链。链接分析更多地体现了一种用户关系[6]，本文从链接分析的角度引入改进的 PR 算法[10-12]，使用户之间的关注关系体现用户的受关注程度，即微博用户的用户被关注度。模型中用户被关注度将取代用户粉丝数，作为微博用户活跃度与微博影响力的影响因子。

改进后的 PR 的基本思想是：重要页是被多个网页引用或者是能被其他重要页引用的网页，同时被引用的次数大大超过其本身引用其他网页的次数。通过计算页面的入链和出链数的比，得到其获得 PR 值的能力；再根据其在作为入链时在所有入链中所占的比重，得到其带给所关注不同用户时的不同 PR 值。算法较好地区分了不同页面在作为入链页面时的重要程度。算法描述为：

$$PR(u) = (1-d) + d\sum_{v_i \in P(u)} m_{v_i} PR(v_i) \quad (1)$$

其中，$d = 0.85$ 为经验值，$P(u)$ 为指向页面 u 的页面 v_i 的集合，$m_{v_i} = \frac{IO_i}{\sum_{i=0}^{n} IO_i}$，为页面 u 从其链入页面 v_i 得到的 PR 值的比重。$IO_i = \frac{InL_i}{OutL_i}$ 表示该网页 v_i 从其入链页面获取 PR 值的能力，IO 值越大，则越容易获得大的 PR 值。考虑到有出链数为 0 的时候，IO 值将出现无穷大的情况，将出链数为 0 的网页的 IO 值设置为一个较小的常数 m_0，如新浪微博的加“V”用户郎咸平的主页，只有数以百万计的入链，而没有一个出链。InL_i 表示网页 v_i 的入链数，相当于输入了 PR 值；$OutL_i$ 表示网页 v_i 的出链数，相当于输出了 PR 值。在指向页面 u 的所有页面 $v_i \in P(u)$ 中，网页 v_i 的全部 n 个出链 O_1，O_2，…，O_n 的入链总数和出链总数的比值 IO_i，分别为 IO_1，IO_2，…，IO_n 对应其从链入页面中获得 PR 值的能力。

在用户被关注度的加权计算和矩阵循环迭代过程中会使得高 PR 值的权值扩大，而低 PR 值的权重缩小，僵尸粉在作为带入 PR 值的入链时，由于正常的用户不会关注没有意义的僵尸用户，所以其关注度较低，即带入 PR 值的能力较小。实验结果证明用户被关注度不仅可以削弱僵尸粉的影响，同时也将较大数量级的用户粉丝差异调整到一个适当的范围，合理地体现用户受关注程度，如图 1 所示：

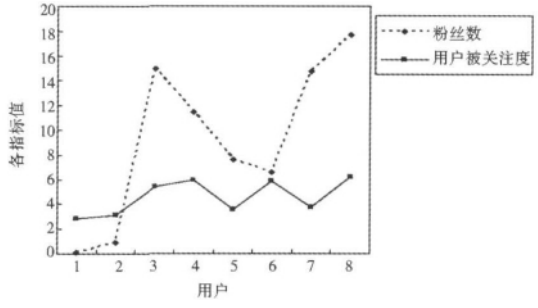


图 1　用户粉丝数与被关注度对比

模型将得到的用户潜在影响力，即用户被关注度与模型中的用户活跃度与用户的微博影响力这两个影响因子结合，实现对微博用户实际影响力的合成计算。

3.2　模型的建立

(1) 用户活跃度

用户活跃度体现了用户在微博平台的活动状态。用户的活动行为是用户影响力扩大的动力，如发布微博、添加关注等；用户被关注度的提升则是用户影响力扩大的表现。石磊等[7]给出用户活跃指数模型，将用户粉丝、用户添加关注的频率与用户发布微博的频率结合，提出了用户的活跃指数的计算方法。本文中的用户活跃度与其不同的是将用户被关注度取代了其中的用户粉丝数，将用户添加关注和发布微博的行为作为用户活跃度的主动影响因子与用户被关注度结合考虑，依照不同的权重计算用户活跃度，公式如下：

$$UV_i = \sum_{j=1}^{k} \gamma_{ij} x_{ij} + UA_i \quad (2)$$

其中，UV_i 为用户 i 的用户活跃度，UA_i 为用户 i 的用户被关注度，x_{ij} 是用户 i 的活跃度的第 j 个影响因子，k 为影响因子的个数，γ_{ij} 为用户 i 的第 j 个影响因子的权重。

$$x_{ij} = \frac{X_{ij}}{T_{first,j} - T_{last,j}} \quad (3)$$

其中，$T_{first,j}$ 为用户 i 第 j 个影响因子最近发生的时间，最新发布的微博消息或添加关注用户的时间，$T_{last,j}$ 为用户最早发布的微博消息或添加关注用户的时间，两者的差则为用户某影响因子发生的总天数。X_{ij} 为第 j 个影响因子的总数，x_{ij} 即是用户 i 的第 j 个影响因子发生的频率。

(2) 微博影响力

微博影响力主要通过微博被转发和评论数体现。微博转发和评论次数越多，说明微博内容越受关注，同时影响力也越大。根据用户被关注度、微博转发数和评论数对微博影响力贡献大小的不同，提出计算用户 i 每一条微博影响力的公式如下：

$$MI_{i,j} = UA_i + \sqrt[3]{MRt_{i,j}} + \sqrt{MC_{i,j}} \quad (4)$$

其中，$MI_{i,j}$ 为用户 i 的第 j 条微博的影响力，UA_i 为用户 i 的关注度，$MRt_{i,j}$ 为用户 i 第 j 条微博被转发数，$MC_{i,j}$ 为用户 i 第 j 条微博的评论数。

将用户 i 所有微博的微博影响力的平均值作为用户 i 的微博影响力，公式如下：

$$MI_i = \overline{MI_i} = \frac{\sum_{j=0}^{n} MI_{i,j}}{n} \quad (5)$$

其中，MI_i 为用户 i 的微博影响力，n 为用户 i 所发布微博的总条数。

(3) 用户影响力

将得到的用户活跃度与微博影响力按照适当的权重结合，得到用户影响力的计算公式如下：

$$UI = \alpha \cdot UV + (1 - \alpha)\, MI \qquad (6)$$

其中，α 为调节因子，用来调整 UV 与 MI 之间的权重。

4　运算过程与结果分析

4.1　数据的采集与准备

本文选取国内新浪微博作为数据来源，通过新浪开放 API 得到相关数据，将数据及数据关系从以下几个方面进行概括：

(1) 用户属性：用户 ID、用户类型、关注数、粉丝数、微博数；

(2) 微博属性：微博数、发布时间、转发次数、评论条数；

(3) 用户关系：用户 ID、关注用户 ID、粉丝 ID。

研究所用的数据采集工作于 2011 年 11 月开始到 2012 年 3 月 5 日结束，按照数据模型的要求，选取可用的用户进行实验，其中实验数据为 2012 年 2 月 5 日到 2012 年 3 月 5 日所跟踪检测的 5 位用户。

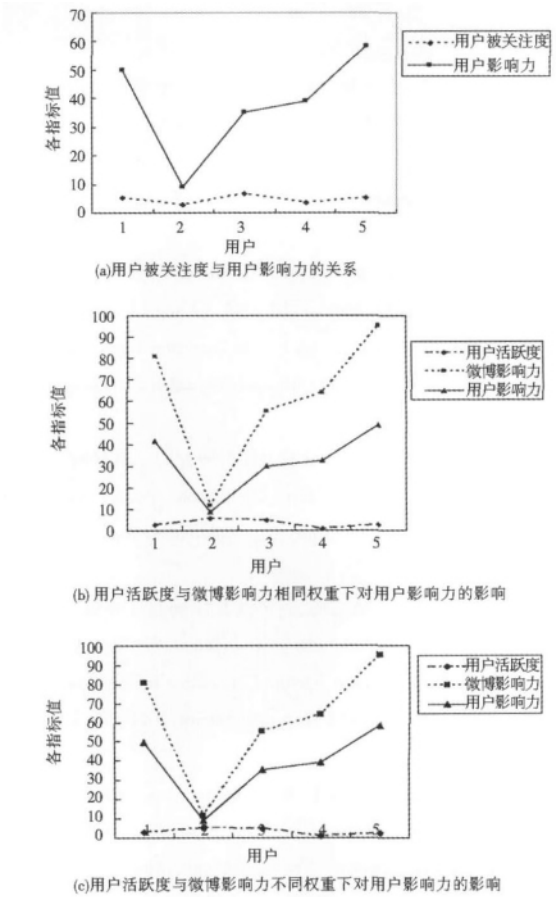
4.2　实验结果分析

不同影响因子的权重问题通过层次分析法计算得到，其中用户活跃度中 $\gamma = \{0.202\ 5, 0.177\ 8, 0.619\ 7\}$，用户影响力中 $\alpha = 0.397$。实验中的 5 位用户被关注度、用户活跃度、微博影响力和用户影响力的计算结果如表 1 所示：

表 1　各指标值

用户	被关注度	活跃度	微博影响力	影响力
1	5.122 1	2.674 9	80.8	49.55
2	2.874	5.412 4	11.66	9.16
3	6.616 2	4.800 0	55.37	35.14
4	3.574	0.891 3	64.13	38.83
5	5.433	2.495 8	95.15	58.09

实验结果如图 2 所示。图 2(a) 是用户被关注度，即用户潜在影响力与用户实际影响力之间的关系，结果表明，通常状况下用户潜在影响力变化与用户实际影响力变化一致，而用户 4 出现的异常则说明，用户被关注度只是用户的潜在影响力而非用户实际影响力。这一结果符合 Cha 等[6] 的结论。图 2(b) 是用户活跃度和微博影响力在相同权重下对用户影响力的影响状况，用户活跃度与微博影响力同为用户影响力的影响因子，用户活跃度作为影响用户影响力的影响因子的能力是微弱的，而用户的微博影响力则在改变用户影响力中起到了主导作用。这一结果符合微博用户通过



(a)用户被关注度与用户影响力的关系

(b) 用户活跃度与微博影响力相同权重下对用户影响力的影响

(c)用户活跃度与微博影响力不同权重下对用户影响力的影响

图 2　各指标对比

发布微博内容起到号召力的事实。图 2(c) 则是本文中赋予不同权重的结果，微博影响力作为改变用户影响力的主导因素，获得的权重较用户活跃度要高，所以用户微博影响力较高的用户，用户影响力相对提高率也较大。

5　结　语

本文针对微博用户的排名机制进行研究，从链接分析和用户行为两个角度对用户影响力进行衡量。首先将链接分析法计算的用户被关注度取代目前使用粉丝数表示的用户被关注度，较为合理地体现用户的受关注程度和微博用户的潜在影响力；再将用户被关注度分别与微博用户的用户活跃度和微博影响力相结合，实现对微博用户的影响力指数的评价。实验结果表明，用户影响力较合理地体现了微博用户的影响力。

参考文献:

[1] 许晓东,肖银涛,朱士瑞. 微博社区的谣言传播仿真研究[J]. *计算机工程*,2011,37(10):272-274.(Xu Xiaodong, Xiao Yintao, Zhu Shirui. Simulation Investigation of Rumor Propagation in Microblogging Community [J]. *Computer Engineering*, 2011, 37(10):272-274.).

[2] 杨亮,林原,林鸿飞. 基于情感分布的微博热点事件发现[J]. *中文信息学报*,2012,26(1):84-90.(Yang Liang, Lin Yuan, Lin Hongfei. Micro-blog Hot Events Detection Based on Emotion Distribution [J]. *Journal of Chinese Information Processing*, 2012, 26(1):84-90.).

[3] 李军,陈震,黄霁崴. 微博影响力评价研究[J]. *信息网络安全*,2012(3):10-13,27.(Li Jun, Chen Zhen, Huang Jiwei. Micro-blog Impact Evaluation Study [J]. *Information Network Security*, 2012(3):10-13,27.)

[4] 袁毅. 微博客信息传播结构、路径及其影响因素分析[J]. *图书情报工作*,2011, 55(12):26-30.(Yuan Yi. The Analysis of Structure, Path and Impact Factor of Microblog Information Communication [J]. *Library and Information Service*, 2011, 55(12):26-30.).

[5] Weng J S, Lim E P, Jiang J, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers [C]. In: *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*(*WSDM* 2010). New York: ACM, 2010: 261-270.

[6] Cha M Y, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy [C]. In: *Proceedings of International AAAI Conference on Weblogs and Social Media*(*ICWSM'10*), Washington. Menlo Park: The AAAI Press, 2010.

[7] 石磊,张聪,卫琳. 引入活跃指数的微博用户排名机制[J]. *小型微型计算机系统*,2012(1):110-114.(Shi Lei, Zhang Cong, Wei Lin. Introducing Active Index to the Microbloggers' Ranking [J]. *Journal of Chinese Computer Systems*, 2012(1):110-114.)

[8] Ye S Z, Wu S F. Measuring Message Propagation and Social Influence on Twitter.com [C]. In: *Proceedings of the 2nd International Conference on Social Informatics*(*SocInfo'10*). Heidelberg: Springer-Verlag, 2010: 216-231.

[9] Yamaguchi Y, Takahashi T, Amagasa T, et al. TURank: Twitter User Ranking Based on User-Tweet Graph Analysis [C]. In: *Proceedings of the 11th International Conference on Web Information Systems Engineering*(*WISE'10*). Heidelberg, Berlin: Springer-Verlag, 2010: 240-253.

[10] 王冬,雷景生. 一种基于 PageRank 的页面排序改进算法[J]. *微电子学与计算机*,2009,26(4): 210-213.(Wang Dong, Lei Jingsheng. An Improved Ranking Algorithms Based on PageRank [J]. *Microelectronics & Computer*, 2009, 26(4):210-213.)

[11] 王德广,周志刚,梁旭. PageRank 算法的分析及其改进[J]. *计算机工程*,2010,36(22):291-292, F0003.(Wang Deguang, Zhou Zhigang, Liang Xu. Analysis of PageRank Algorithm and Its Improvement [J]. *Computer Engineering*, 2010, 36(22):291-292, F0003.)

[12] 金迪,马衍民. PageRank 算法的分析及实现[J]. *经济技术协作信息*,2009(18):118.(Jin Di, Ma Yanmin. The Analysis and Realism of the PageRank Algorithms [J]. *Economic and Technological Cooperation Information*, 2009(18):118.)

(作者 E-mail: kefeng510@163.com)