基于 PageRank 的微博用户影响力算法研究

作者	孙红，左腾
机构	上海理工大学；上海现代光学系统重点实验室
发表期刊	《计算机应用研究》
预排期卷	2018 年第 35 卷第 4 期
访问地址	http://www.arocmag.com/article/02-2018-04-041.html
发布日期	2017-04-01 17:22:31
引用格式	孙红，左腾. 基于 PageRank 的微博用户影响力算法研究[J/OL]. [2017-04-01]. http://www.arocmag.com/article/02-2018-04-041.html.
摘要	伴随着互联网的高速发展与普及，微博作为信息交流与传播的载体，已成为新型社会化媒体的代表。在中国，微博用户规模已经达到了 2.42 亿。微博用户影响力计算对社会日常信息在微博里面有效传播，正确传播，健康传播有着非常重要的意义。本文以新浪微博数据为实验的对象，通过改进传统的 PageRank 模型，提出了的新的微博用户影响力排名算法---MBUI-Rank（Micro-Blog User Influence Rank）算法。该算法在传统的 PageRank 算法模型上，加入了微博用户自身在微博里面行为活动，同时…
关键词	PageRank，新浪微博，用户影响力，用户自身行为
中图分类号	TP391
基金项目	国家自然科学基金资助项目（61170277，61472256）；上海市教委科研创新重点项目（12zz137）；沪江基金资助项目（C14002）

基于 PageRank 的微博用户影响力算法研究 *

孙 红[1,2]，左 腾[1]

(1. 上海理工大学，上海 200093；2. 上海现代光学系统重点实验室，上海 200093)

摘 要：伴随着互联网的高速发展与普及，微博作为信息交流与传播的载体，已成为新型社会化媒体的代表。在中国，微博用户规模已经达到了 2.42 亿。微博用户影响力计算对社会日常信息在微博里面有效传播，正确传播，健康传播有着非常重要的意义。本文以新浪微博数据为实验的对象，通过改进传统的 PageRank 模型，提出了的新的微博用户影响力排名算法---MBUI-Rank（Micro-Blog User Influence Rank）算法。该算法在传统的 PageRank 算法模型上，加入了微博用户自身在微博里面行为活动，同时考虑到了微博用户的自身行为，结合用户权值得到最终影响力。实验结果表明，MBUI-Rank 算法与传统的 PageRank 算法相比，可以更加真实有效地反映微博用户的实际影响力。

关键词：PageRank；新浪微博；用户影响力；用户自身行为

中图分类号：TP391

Research on algorithm of micro-blog user influence based on PageRank

Sun Hong[1,2], Zuo Teng[1]

(1. University of Shanghai for Science & Technology, shanghai 200093, China; 2. Shanghai Key Lab of Modern Optical System, Shanghai 200093, China)

Abstract: With the rapid development and popularization of the Internet, micro-blog as a carrier of information exchange and dissemination, has become a new type of social media representatives. In China, micro-blog user size has reached 242 million. Microblogging user influence calculation of the daily information in the micro-blog inside the effective dissemination, the correct dissemination of health communication has a very important significance. This paper takes Sina Weibo data as the object of experiment, and improves the traditional PageRank model, and puts forward the new micro-blog user influence rank algorithm. The algorithm in the traditional PageRank algorithm model, joined the micro-blog users themselves in the micro-blog inside the behavior of activities, taking into account the micro-blog user's own behavior, combined with the user's weight to get the ultimate influence. The experimental results show that the MBUI-Rank algorithm can reflect the actual influence of micro-blog users more effectively and effectively than the traditional PageRank algorithm.

Key Words: PageRank ;Sina Weibo;user influence; user’s behavior

0 引言

现如今随着互联网的快速发展，大数据时代也随之到来。社交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，从而实现对信息的广泛、快速传播。微博（Micro-Blog）作为当前最为流行的社交网络之一，它有着诸多优点。如用户可以随时随地发布消息，用户之间互动性强，操作简单简。据 CNNIC 发布的第 38 次《中国互联网络发展状况统计报告》数据显示[1]，截至 2016 年 6 月， 微博用户规模为 2.42 亿，逐渐回升，使用率为 34%，与 2015 年底相比略有上涨。在微博里面，新浪微博发布的消息种类丰富，可以是 140 个字符以内的简短微博，也可以是超过 140 个字符的长微博，可以配上图片，可添加地点，可以添加短视频可以“@”(艾特)其他微博用户等。因为有着诸多特点，新浪微博占有着微博界的主导地位。

每一个微博用户可以通过微博获取最新的实时新闻资讯，了解朋友及其他社会名人，社会媒体，公众媒体等等其他一些本文关心的实时动态。本文可以给别人的微博点赞，评论，转发，这样从而提高了新闻信息的传播速度及其影响的范围。如果一则微博消息在微博里面迅速传播开来，如有很多人转播，评论，点赞，或着其他一些微博用户也发布相关主题的微博，那么这则微博相关的话题就会上微博热搜或者微博头条，从而会有更多的人知道这则消息，从而影响到整个社会舆论情况。

基金项目：国家自然科学基金资助项目（61170277，61472256）；上海市教委科研创新重点项目（12zz137）；沪江基金资助项目（C14002）

作者简介：孙红（1964-），女，北京人，副教授，硕导，博士研究生，主要研究方向为计算机网络通信与云计算、管理科学与工程、计算机科学与技术、控制科学与工程、模式识别与智能系统（823372873@qq.com）；左腾（1992-），男，湖北人，硕士研究生，主要研究方向为大数据、云计算.

如果某个微博用户有众多的粉丝和关注度，那么他发布的微博消息就会得到广泛的关注，那么该用户就影响到了信息的在微博里面的传播。所以，现在就会有很多人通过购买粉丝（也称水军），来提高自己微博信息传播的影响力，然而这些所谓的粉丝是不能够当作正常微博用户来看的，他们会照成谣言和不良信息的的散布，给社会带来不利的影响和舆论恐慌。

1 研究背景

微博最早起源于国外的，所以相关的研究算法也起源于国外学者的研究，也取得了相应成果。因为 Twitter 作为微博的鼻祖，国外对微博的研究主要集中在对 Twitter 的研究。目前有关微博影响力分析算法主要参考 Google 的 PageRank[2]算法及其改进后的算法[3,4]和 HITS[5]算法及其改进后算法[6,7]。PageRank 算法模型是为了实现网页排名，该算法也是 Google 使用的搜索结果排名中的重要组成部分。事实上，PageRank 模型本质上是用于有向图的节点级的计算技术，因此应用于用户对微博的影响是自然的。 2009 年，Tunkelang 等人 构建了一个基于链接的有向图，并使用 PageRank 模型来实现 Twitter 用户的影响力排名。借鉴 PageRank 算法的思想，文献[8]提出了 TwitterRank 算法，该算法主要是衡量一个用户在某一话题内的影响力，主要思想是给定一个话题，用户的影响力定义为他的所有粉丝的影响力之和[9]。但是该算法仅考虑拥有相似话题的用户间相互的影响力，不具有一般性。文献[10]中，对 Twitter 的传播特性进行了分析，使用粉丝数量和微博转发数量比对研究用户在话题传播过程中的影响，结果发现拥有众多粉丝数量的微博用户所发的微博不一定会得到很多的转发或者评论，这说明粉丝数量和影响力之间并没有必然的联系，但是那些有影响力的用户能够在微博里面产生显著的影响。国内也有诸多学者对社交网络进行影响力分析，如丁兆云[11]等，综合考虑转发关系、回复关系、复制关系、阅读关系四种网络关系，对话题层次影响力进行分析。李国良[12]等人提出了多社交网络上的影响计算模型来建模节点间的影响力，然后扩展了基于树的算法模型以适应多社交网络上的影响力最大化问题。毛佳昕[13]等人利用微博用户发布微博的时间和数量来分析用户影响力，该方法并没有引入实际的话题来分析用户对社会问题的影响。师亚凯[14]等利用微博用户内容建立词共现矩阵，继而运用 LDA 模型进行潜在主题的识别并进行降维，通过 KL 散度的方法得到用户之间的相似性，最后结合用户影响力权值得到用户的影响力。在文献[15]里面，Xun Chen 等人提出来 Personal Rank 算法，该算法也可以用于计算微博用户的影响力，但是在计算时还是需要依赖 PageRank 算法。在文献[16]里面，Jun Zhou 等人使用回归模型来预测每个用户的影响力分数分析个人财产及其内容消息，而且他们的研究揭示了个体大多数用户的影响随时间而变化。在文献[17]里面，Guo-Jun Mao 等人对用户的活跃度进行了分析，但是该文章里面只对用户的评论数进行了分析，并没有深入分析也没有剔除“僵尸粉”的干扰。

微博用户之间的关系就好像网页页面之间的关系，因此，利用他们的相似性，PageRank 模型可以计算出微博用户的影响。本文综合考虑微博用户的关系网络特性和微博用户实际行为特征，提出了名为 MBUI-Rank 的微博用户影响力排名算法。

2 MBUI-Rank 微博用户影响力模型

2.1 PageRank 模型

最初的 PageRank 模型，是 Google 在搜索引擎结果中对网站排名的核心算法。PageRank 通过计算页面链接的数量和质量来确定网站的重要性的粗略估计，PageRank 是基于从许多优质的网页链接过来的网页，必定还是优质网页的回归关系，来判定所有网页的重要性。

为了方便研究，现给出与本文相关的公式定义。

定义 1 PageRank 算法。对所有研究的网页给定一个有向图 G=<V,A>，它是由顶点的集合 V 和边的集合 A 组成。网页 $P_i \in (G)$的排名可以用下面的公式来计算。

$$\mathrm{PageRank}(p_i) = (1-\alpha)/N + \alpha \sum_{p_j \in M_{p_i}} \mathrm{PageRank}(p_j)/L(p_j) \quad (1)$$

其中 M_{pi} 是所有对 P_i 网页有出链的网页集合，$L(P_i)$是网页 P_i 的出链数目，N 是研究网页总数，α 代表的是阻尼因子，取值范围是 0-1。根据上面的公式，本文可以计算每个网页的 PR(下文的所有的 PR 代表 PageRank)值，在不断迭代趋于平稳的时候，即为最终结果。图 1 显示了运用 PageRank 算法时，网页 P_i 与网页 P_j 之间的关系。

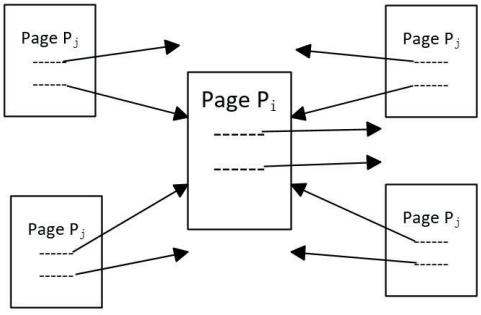

图 1 网页 P_i 和网页 P_j 的关系

在表现网页之间链接关系时，Google 使用了矩阵，即下面的定义 2。

定义 2 Google 初始矩阵。本文可以用一个矩阵来表示这张图的出链入链关系，初始矩阵 $S=(s_{i,j})$，其中 $s_{i,j}$ 代表网页 j 跳转到页面 i 的概率。 也就是说，对于 $i,j \in G$，与有向图 G 相关的 Google 矩阵可以被设置如下：

$$s_{i,j} = \begin{cases} \dfrac{1}{L(j)}, & \text{当} \langle j, i \rangle \in A \text{时} \\ 0. & \text{其他情形} \end{cases} \quad (2)$$

其中 L（j）是网页 j 的出站链接总数。

现在给出 N 为 4 的一个例子（共有 A、B、C、D 四张网页）帮助说明这个矩阵。对于图 2 所示的有向图，其 Google 初始矩阵可以通过公式（2）获得，图 3 给出了 Google 初始矩阵 S 计算结果。

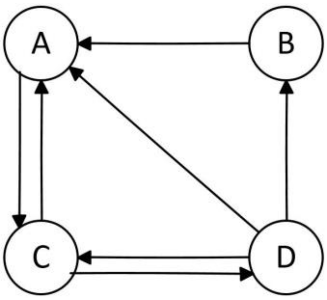

图 2 页面有向图

$$S=\begin{bmatrix}0 & 1 & 1/2 & 1/3\\ 0 & 0 & 0 & 1/3\\ 0 & 0 & 0 & 1/3\\ 0 & 0 & 1/2 & 0\end{bmatrix}$$

图 3 Google 初始矩阵

定义 3　Google 矩阵。得到初始矩阵后，本文就可以计算 PR 值了，当只有 α 概率的用户会点击网页链接，剩下(1-α)概率的用户会跳到无关的页面上去，而访问的页面恰好是这 4 个页面中 A 的概率只有(1-α)/4（α 是阻尼系数，Google 在计算网页排名的时候取 α 等于 0.85，所以本文在这里也取 0.85），所以真正的 Google 矩阵由公式 3 计算，Google 矩阵 GM 如下图 4 所示。

$$gm_{i,j}=\alpha\times s_{i,j}+\frac{(1-\alpha)}{N} \tag{3}$$

$$GM=\alpha\times\begin{bmatrix}0 & 1 & 1/2 & 1/3\\ 0 & 0 & 0 & 1/3\\ 0 & 0 & 0 & 1/3\\ 0 & 0 & 1/2 & 0\end{bmatrix}+\frac{1-\alpha}{4}\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\end{bmatrix}$$

图 3 Google 初始矩阵

定义 4　PR 值计算。在有向图 G 及其 Google 矩阵 GM 里面，其中 n 是 G 中的节点数然后. 于是得到 P_n=GM×P_{n-1},可以通过以下公式（3）逐级地迭代更新秩向量，直到得到 P_n=P_{n-1} 时，才算迭代完成，这时的 P_n 就是 PR 的值。于是计算 PR 值的过程就变成了一个 Markov 过程。

$$\boldsymbol{P}_{\mathrm{n}}=\boldsymbol{GM}\times\boldsymbol{P}_{n-1} \tag{4}$$

2.2　MBUI-Rank 模型

传统的 PageRank 模型可以帮助评估微博用户的影响力，但其有效性并不被大多数人认可。传统的 PageRank 模型仅考虑链接关系，即用户与用户之间的关注与被关注关系。把微博用户的粉丝看做网站的入站链接，微博用户关注的人看做网站的出站链接，这样就可以把 PageRank 模型应用到计算微博用户影响力里面来[18]。但是，在计算影响力的时候，那些拥有众多“僵死粉”的微博用户的影响力就会被虚假提高，从而计算出来的影响力并不准确，其实那些拥有众多“水军”的用户影响力并没有那么高。所以，首先要剔除微博用户的僵尸粉，在进行粉丝筛选的时候，需要选取高质量的粉丝。需要选取长期关注博主的粉丝，而不是近期突然增加的粉丝。

定义 5　粉丝选取。认定选取的粉丝集合为 F,其中选取的粉丝是关注博主三个月以上的用户。而且这些用户在三个礼拜以内有和别人互动，且评论或者转发过该博主的微博等这样的交互行为。

$$\mathrm{F(i)}=\{\mathrm{j}|(\mathrm{j,i})\in \mathrm{A}\cup FT>3m\cup AT>3w\} \tag{5}$$

F(i)代表用户 i 的粉丝集合，j 代表其中一个粉丝，FT 代表的是粉丝 j 关注博主的时间，大于三个月，AT 代表粉丝 j 活跃的天数是三周。这一步只是对粉丝进行了筛选，剔除了一些僵死粉的干扰。选取完粉丝后就需要计算用户的活跃度了，下面的定义 6 代表粉丝活跃度计算公式。

定义 6　用户活跃度计算。用户的活跃度包括很多条件，如用户发布的微博数，转发微博数，点赞微博数，评论微博数，@别人的微博数，收藏微博数量以及在微博中的活跃的天数。综合这些条件，下面给定微博用户 i 其活跃度的计算公式：

$$Activity(i)=\frac{nb_i}{NB}+\frac{nr_i}{NR}+\frac{nc_i}{NC}+\frac{nl_i}{NL}+\frac{na_i}{NA}+\frac{d_i}{D} \tag{6}$$

其中 nb_i 是微博用户 i 发布的微博数量，NB 是整个微博里面所有微博用户发布的微博数量；nr_i 是微博用户 i 转发的微博数量，NR_i 是整个微博里面所有微博用户转发微博数量；nc_i 是微博用户 i 评论微博数量，NC_i 是整个微博里面所有微博用户评论微博数量；nl_i 是微博用户 i 点赞微博数量，NL_i 是整个微博里面所有微博用户点赞微博数量；na_i 是微博用户 i“@”其他用户微博数量，NA_i 是整个微博里面所有微博用户“@”其他用户微博数量；d_i 是微博用户 i 的活跃天数，D_i 是整个微博里面所有微博用户平均活跃天数。

在研究了用户的活跃度之后，但还需要对用户的微博质量进行研究。微博的质量体现在微博的内容是否健康，真实，是否对本文整个社会有一定的影响。例如那些在微博里面散布广告，发布虚假消息的微博用户，那么他们的这些行为对于整个微博环境，对于整个社会带来了负面的影响，应当予以否定。所以，在选取微博内容时，就需要剔除那些质量较差的微博，选取能够对于社会产生正面影响的微博。因此，在筛选微博数据的时候，本文就选取了和社会话题相关的微博作为研究对象。同时对用户微博被评论的数量，被转发的数量，被点赞的数量，和被收藏数量进行研究。如果微博被众多人评论，转发，点赞和收藏自然说明该用户的影响力要高。根据上面提出的思想，下面的定义 7 给出了在用户微博质量的简单评估计算方法。

定义 7　用户微博质量计算。对于微博用户 i，选取的研究话题为社会话题为 SI(Social Issues)，用户微博质量计算公式如下：

$$QSI(i)=\frac{nsi_i}{N_i} \tag{7}$$

$$QR(i)=\frac{rsi_i}{R_i} \tag{8}$$

$$QC(i)=\frac{csi_i}{C_i} \tag{9}$$

$$QL(i) = \frac{lsi_i}{L_i} \tag{10}$$

$$QF(i) = \frac{fsi_i}{F_i} \tag{11}$$

$$Quality(i) = QSI(i) + QR(i) + QC(i) + QL(i) + QF(i) \tag{12}$$

其中 nsi_i 是用户 i 发布与主题 SI 相关的微博数量，N_i 是用户 i 在微博中发布的所有微博数量；rsi_i 是用户 i 发布与主题 SI 相关的微博被转发的数量，R_i 是在微博里面所有被转发的微博数量；csi_i 是用户 i 发布与主题 SI 相关的微博被评论的数量，C_i 是在微博里面所有被评论的微博数量；lsi_i 是用户 i 发布与主题 SI 相关的微博被点赞的数量，L_i 是在微博里面所有被点赞的微博数量。fsi_i 是用户 i 发布与主题 SI 相关的微博被收藏的数量，F_i 是在微博里面所有被收藏的微博数量。

此外，还有另一个因素也可以在很大程度上反映用户的影响力，就是微博用户的可信度。比如微博里面的那些大 V,一些经过官方认证博主，那么他们的发言就会很有影响力，人们愿意相信这些人所发布的微博。目前在新浪微博平台上已经提供认证机制，共有四种认证方式。第一种是兴趣认证，第二种是自媒体认证，第三种是身份认证，最后一种是官方认证。在此，本文利用微博上不同的认证，给出不同的用户信誉度。定义 8 给出了在微博上用户可信度的简单计算方法。

定义 8　微博用户可信度。对于微博用户 i，其可信度可以计算公式如下：

$$Cerdibility(i) = \begin{cases} 0.5, & \text{i为兴趣认证用户时} \\ 1, & \text{i为自媒体认证用户时} \\ 1, & \text{i为身份认证用户时} \\ 1.5, & \text{i为官方认证用户时} \\ 0. & \text{i为非认证用户时} \end{cases} \tag{13}$$

在进行运算 PageRank 算法之前还需要修改微博用户的权重，由定义 9 给出。

定义 9　微博用户权重。对于微博用户 i，它的权重 w(weiht)计算公式如下：

$$w(i) = Activity(i) + Quality(i) + Credibility(i) \tag{14}$$

然后将用户权重加入进原始 PageRank 模型计算。

定义 10　加权 Google 矩阵。对于有向图 G=<V,A>，令其加权的 Google 矩阵 Z=($z_{i,j}$)，其中 $z_{i,j}$ 计算如下：

$$z_{i,j} = \begin{cases} gm_{i,j} \times w(i), & \text{当<j, i>}\in A \text{ 时} \\ 0. & \text{其他情形} \end{cases} \tag{15}$$

公式(15)中的 $z_{i,j}$ 是公式 3 中的值 $gm_{i,j}$ 与用户权重 w(i)的乘积，所以加权的 Google 矩阵考虑的用户的链接结构包括用户活跃度，用户微博质量和用户可信度等因素。 假设已经获得图 1 中每个节点的评估参数，如令 w(A)=w1 ，w(B)= w2，w(C)= w3 和 w(D)=w4 ，那么通过计算，就可以获得加权后的的 Google 矩阵 Z。

下面是 MBUI-Rank 算法的主要处理过程。

输入：微博用户社交网络图 G； 阻尼系数 α；迭代终止条件 Ɛ。

输出：用户节点 P 的影响向量。

处理过程:

a) 计算 G 的 google 矩阵 :GM=($gm_{i,j}$);

b) for i∈G 执行

c) 通过定义 6，7，8 计算 Activity(i),Quality(i), Credibility(i);

d) w(i)= Activity(i)+Quality(i)+ Credibility(i);

e) 结束 for

f) for i∈G 执行

g) forj∈G 执行

h) $z_{i,j}=w(i)\times gm_{i,j}$ ；

i) Z=($z_{i,j}$);

j) P_0=I;

k) 重复

l) P=Z×P_0;

m) $P_n= P_{n-1}$

n) 直到$|P-P_0|\leq Ɛ$

o) 将 P 作为最终的影响向量返回。

在算法 MBUI-Rank 中，首先给每个页面赋予随机的 PR 值，特征向量 P 的初始值为值为 1 的 N×1 矩阵,通过 $P_n=GM\times P_{n-1}$ 不断地迭代 ，得到收敛的 PR 值，即当满足下面的不等式$|P_n-P_{n-1}|\leq Ɛ$ 后迭代结束，获得用户的影响力。

3　实验与分析

3.1　实验数据

在获取新浪微博数据时，首先要先注册新浪微博账号，然后利用该账号在新浪微博开放平台完成开发者的注册，注册身份为学生，然后在开放平台上创建一个应用，创建完应用后，开发者会得到获取两个非常重要的参数 App Key 和 Secret Key。在创建应用过程中需要填写一个授权回调页。应用创建完成后，就可以利用新浪官方 API 进行开发了。首先进入开放平台，在文档里面资源下载和 API 选项，首先需要在资源下载下面下载相关的 SDK，本文用到的是 JAVA SDK。下载完了 JAVA SDK 后就需要导入到 eclipse 里面，然后是配置下载下来的 JAVA SDK。主要改一个文件，src 文件夹下面的 config.properties，配置如下参数。

1.client_ID ：appkey 创建应用获取到的 appkey（App Key）

2.client_SERCRET ： app_secret 创建应用获取到的 appsecret（Secret Key）

3.redirect_URI：回调地址 OAuth2 的回调地址（就是在高级信息里面填写的授权回调页）。在配置好配置文件后就需要进行 Oath2.0 认证，这是在调用所有 API 之前都需要进行的操作。

调用 example 下面 weibo4j.examples.oauth2 包里面的 OAuth4Code.java。如果这个步骤完成了就可以任意调用微博 API 了。然后再微博 API 选项下面查阅相关的 API 文档，就可以在 eclipse 里面下载到需要用到的数据。然后利用微博 API 获取了 59528 个微博用户，其中包含用户的基本信息和用户关系网络和发布的微博信息等一些数据，然后从 59528 个用户里面选择 2958 个微博用户作为实验对象。 虽然 2958 个用户只是新

浪微博用户总数的一小部分，但这些用户的关系是相对完整的，所以本文可以使用它们来测试本文的算法在本文中的有效性。

3.2 实验环境

本次实验使用 1 台计算机，配置为:CPU 为 8 核 Intel 酷睿 i7 6700HQ 主频为 3.4GHz，内存 DDR4 2133MHz 8GB，硬盘 1TB，操作系统为 win10。

3.3 实验结果

本文对实验结果进行比较分析选取社会话题作为PageRank 和 MBUI-Rank 算法计算时的主题，分别对传统的 PageRank 算法和本文 MBUI-Rank 算法计算出的用户影响力进行排序，列出 PageRank 算法和 MBUI-Rank 算法影响力排名前 10 的用户，计算结果如表 1 和表 2 所示。

表 1 PageRank 算法计算结果

排名	微博用户名	粉丝数	微博数	认证
1	新手指南	172567565	10002	官方
2	微博管理员	156722636	978	官方
3	谢娜	88124987	9010	身份
4	陈坤	80757130	5036	身份
5	姚晨	80321437	9058	身份
6	赵薇	78988286	4286	身份
7	何炅	82099269	7725	身份
8	angelababy	78024696	2423	身份
9	人民日报	50976411	71985	官方
10	央视新闻	48576165	81213	官方

表 2 MBUI-Rank 算法计算结果

排名	微博用户名	粉丝数	微博数	认证
1	中国新闻网	31089110	77884	官方
2	人民日报	50976411	71985	官方
3	央视新闻	48576165	81213	官方
4	新浪新闻	10687917	12276	官方
5	人民网	37137367	107630	官方
6	法制晚报	16163311	82710	官方
7	环球时报	7106371	116736	官方
8	中国经营报	2336498	90318	官方
9	新京报	27433032	36902	官方
10	头条新闻	51435942	11932	官方

PageRank 和 MBUI-Rank 影响力排名前 10 的用户与粉丝数、关注数和微博数的相关性如图 5 和图 6 所示。

3.4 结果分析

对比表 1 和 2 可以清楚地看到在相同的数据集上面，不同算法得到的结果排序是截然不同的。

通过对表 1 的分析可以看出：影视大咖等这一些知名度较高的微博用户的影响力较高，他们的粉丝众多而且他们发布的微博数量高。由此可看出如今很多微博用户习惯于在微博中查看相关娱乐名人的生活动态，因此该类用户对微博里面的信息传播有着重的引导能力。而且大多数影响力大的用户都是微博认证用户，这样就增加了微博用户对他们的信任感。结合表 1 和图 5 来看，用户粉丝数量、用户活跃度高的微博用户，他们的影响力也比较高。但是，本文发现排名第一和第二微博用户是新手指南和微博管理员，但是他们实际的影响力并没有计算结果那么高。他们的排名之所以这么高是应为他们拥有众多粉丝，他们拥有粉丝数量是排名第三微博用户（谢娜）的将近两倍，因为他们拥有这么多的粉丝数量，在传统的 PageRank 算法计算影响力的时候就把他粉丝的影响力加权起来了，这样影响力就比其他用户的高了许多。拥有这么多粉丝数量的原因是在用户最开始注册微博的时候就系统就帮微博用户就自动关注了他们，而且用户也没有去取消关注，还有就是当微博用户不去用他们的微博账号后，之前的微博账号不能注销，所以这样在无形之中就产生了“僵尸用户”，而且这些“僵尸用户”也不会被系统清除，所以他们的粉丝才会这么多，但其活跃度相对而言不是很高。所以在目前微博里面就会有明星花钱去买粉丝，来提高自己的想象力，提高自己的知名度。在表 1 里面，本文还可以看到，人民日报和央视新闻也挤入了前十，虽然他们得粉丝数量和前面的微博用户相差很大，但是他们是社会话题的发布者，所以这样就提升了他们的影响力。

图 5 Pagerank 排名

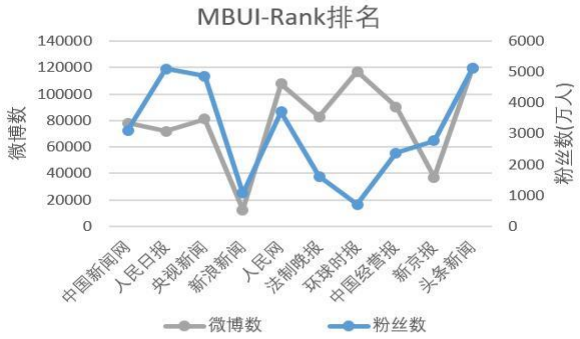

图 6 MBUI-Rank 排名

结合表 2 和图 6 可以看出， MBUI-Rank 算法得到的结果和粉丝数量没有呈现出正相关的关系。而微博用户的自身的活动，他的微博数量，所发布微博的内容等一些他的动态行为成了评论影响力强有力的标准。如微博用户头条新闻他的粉丝数量和微博数量是最高的，但是他的排名却不是靠前的，只是勉

强进入前十。究其原因可以知道，因为头条新闻他发布的微博很是杂乱，不管什么新闻他都会发布，比如娱乐新闻，社会新闻，体育新闻等等一些事实，所以在社会问题上面就不是很突出。排名靠前的是中国最具权威的报刊和新闻媒体人民日报和央视新闻，可以看到，他们的粉丝数量和微博数量都是非常高的。本文还可以看出，MBUI-Rank 算法的前 10 个用户他们都是官方认证用户，而且拥有足够的微博，关注者和粉丝。此外，为了评估文章的质量，本文把主题放在社会问题上，所以 MBUI-Rank 算法得到的前 10 个用户，大多是官方认证的新闻媒体，他们具有更大的权威性，而且这 10 个用户里面没有一个娱乐明星，这也说明这些娱乐明星发布的微博很少与社会话题有关，通过查看他们的微博，他们发布的微博话题大多数都是他们的日常生活或者一些宣传，因此在 MBUI-Rank 算法里面他们都排不上名。PageRank 算法在评估用户影响力时没有考虑到用户自身的动态行为，而是依赖于粉丝数量，过于简单片面，所以在一定程度上受到“僵尸粉的”干扰。对比 PageRank 算法与 MBUI-Rank 算法排名发现，MBUI-Rank 算法在交互式用户排名之前考虑用户的活动，质量和可信度，这使得用户以综合的和面向主题的方式影响评估。

4 结束语

本文分析了微博社交网络的用户的关系网络和自身的行为，结合传统 Pagerank 算法，新的用户影响力算法，它比传统的 PageRank 模型更适合社交网络用户影响力数据挖掘。MBUI-Rank 算法考虑用户的实际微博活动行为以及用户的链接，使得挖掘结果更客观和准确。通过计算微博用户的活跃度，微博质量和用户可信度来估算用户影响力。从实验结果可以看出，与传统的 PageRank 算法相比，MBUI-Rank 算法能更好的反映微博用户影响力。

在下一步研究工作中，将会研究在不同话题领域的影响力。在算法方面多加考虑约束条件，比如选取那些有代表性的评论和转发，选取有代表性的微博用户，加大这些用户的权重，这样会使得最后得到的结果更加准确。

参考文献：

[1] 第38次中国互联网络发展状况统计报告[R]. 北京: 中国互联 网络信息中心, http: //www. cnnic. net. cn/ 2016 July.

[2] Page, Lawrence, Brin, *et al*. The PageRank citation ranking[C]// Bringing Order to the Web. Stanford InfoLab. 1998: 1-14.

[3] Lamberti F, Sanna A, Demartini C. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines[J]. IEEE Trans on Knowledge & Data Engineering, 2009, 21(1): 123-136

[4] Jing Yushi, S Baluja. Pagerank for product image search[C]//Proc of International Conference on World Wide Web. 2008: 307-316.

[5] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[6] Liu Ying, Lin Y. Supervised HITS algorithm for MEDLINE citation ranking[C]//Proc of IEEE International Conference on Bioinformatics and Bioengineering. 2007: 1323-1327.

[7] Asano Y, Yu T, Nishizeki T. Improvements of HITS Algorithms for Spam Links[J]. Ieice Trans on Information & Systems. 2008, E91D(2): 200-208.

[8] Jianshu Weng, Ee-Peng Lim, Jing Jiang, *et al*. TwitterRank: finding topic-sensitive influential twitterers. [J]. Wsdm, 2010: 261-270.

[9] 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, 48(25): 229-233.

[10] Cha M, Haddadi H, Benevenuto F, *et al*. Measuring User Influence in Twitter: The Million Follower Fallacy[J]. ICWSM, 2010, 10(10-17): 30

[11] 丁兆云, 周斌, 贾焰, 等. 微博中基于多关系网络的话题层次影响力分析[J]. 计算机研究与发展, 2013, 50(10): 2155-2175.

[12] 李国良, 楚娅萍, 冯建华, 等. 多社交网络的影响力最大化分析[J]. 计算机学报, 2016, 39(4): 643-656.

[13] 毛佳昕, 刘奕群, 张敏, 等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014(4): 791-800.

[14] 师亚凯, 马慧芳, 张迪, 等. 融合用户行为和内容的微博用户影响力方法[J]. 计算机应用研究, 2016, 33(10): 2906-2909.

[15] Chen Xu, Wang Pengfei, Qin Zheng, *et al*. HLBPR: a hybrid local bayesian personal ranking method[C]//Proc of International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016: 21-22.

[16] Jun Zhou, Yan Zhang, Bing Wang, *et al*. Predicting user influence in microblogs[C]//Proc of IEEE International Conference on Computer Communication and the Internet . 2016: 292-295.

[17] Mao G J, Zhang J. A PAGERANK-BASED MINING ALGORITHM FOR USER INFLUENCES ON MICRO-BLOGS[J]. 2016.

[18] 徐文涛, 刘锋, 朱二周. 基于 MapReduce 的新型微博用户影响力排名算法研究[J]. 计算机科学, 2016, 43(9): 66-70.