基于多特征融合的微博用户权威度定量评价方法

张仰森,郑 佳,唐安杰
(北京信息科技大学智能信息处理研究所,北京 100101)

摘 要: 微博用户权威度是评价微博信息可靠性的重要因素之一.本文针对微博用户权威度的定量计算提出了一种基于多特征融合的微博用户权威度定量评价模型.首先,提出了用户权威度的概念,将其定义为用户影响力和被信服度两部分组成;在暂不考虑用户领域影响因子的情况下,基于新浪微博数据,抽取出微博用户信息传播影响力、用户信息完整度、用户活跃度以及用户平台认证指数 4 项评价特征,以构建了用户权威度定量计算模型;然后,采用层次分析法对所构建模型的 4 项评价特征的权值进行确定,并分别给出了 4 项评价特征的提取算法.同时,在用户关注关系网络的基础上,提出了一种基于用户被关注价值的用户信息传播影响力模型 UIRank,并通过实验验证了其比 PageRank 算法更加有效.实验结果表明,本文提出的微博用户权威度定量计算模型比较合理,为用户权威度的定量评价提供了一种可行的解决方案.

关键词: 微博;用户权威度;用户影响力;UIRank;层次分析法
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)11-2800-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.11.030

A Quantitative Evaluation Method of Micro-blog User Authority Based on Multi-Feature Fusion

ZHANG Yang-sen, ZHENG Jia, TANG An-jie
(*Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing* 100101, *China*)

Abstract: Micro-blog user authority is one of the important factors to evaluate the reliability of micro-blog information. In this paper, a quantitative evaluation model of micro-blog user authority is proposed based on multi-feature fusion. Firstly, the concept of user authority is proposed, which is defined by two parts. the user influence and the convinced degree. In the case that the user domain influence factor is not considered, we extracted four user characteristics which include the user information spread influence, the user information integrity degree, the user activity degree and the user platform authentication index based on the Sina Weibo data and construct the user authority quantitative calculation model. Then, we determine the weight of the four user characteristics based on the analytical hierarchy process and the extraction algorithms of them are given respectively. At the same time, we put forward the UIRank model based on the followed value and the follow relationship network between the users which is used to calculate the user information spread influence and proved to be more effective than the famous PageRank algorithm through experiments. The experimental results show that the method proposed in this paper is more reasonable to calculate the user authority of micro-blog user, and it provide a feasible scheme for the quantitative evaluation of the user authority.

Key words: micro-blog; user authority; user influence; UIRank; analytical hierarchy process

收稿日期:2016-11-06;修回日期:2016-12-16 日;责任编辑:梅志强
基金项目:国家自然科学基金(No. 61070119, No. 61370139, No. 61772081, No. 61602044);北京市属高等学校创新团队建设与教师职业发展计划(No. IDHT20130519)

1 引言

微博,即微型博客的简称,是一种基于有线或无线互联网终端向平台发布精短共享信息的即时社交网络,可通过计算机、手机、掌上电脑等多种终端进行微博的浏览、发布和评论.微博以其独特的开放性、实时性、互动性和低门槛性已然发展成为人们日常生活中不可或缺的信息传播媒介,以极快的速度影响着社会的信息传播格局[1].微博用户不仅包括普通的平民用户,还包括许多知名人士以及一些权威机构或政府职能部门,其用户层次跨度较大,也使其成为社会热门话题产生的聚集地[2].因此,对微博传播的特点及影响因素的研究已经成为重要的研究课题.其中,微博用户作为信息发布和传播的主体,其权威度是微博信息传播影响因素中的一个极为重要的评判指标.

目前对于微博用户研究的大多数文献都集中在可信度方面,对于权威度的研究鲜有涉及,即使有涉及的也都是和可信度的概念混为一谈.但是权威度与可信度是两个相互独立的概念,在现代汉语词典中,对于权威的解释是:①使人信从的力量和威望;②是在某种范围里最有地位的人或事物.而对于可信的解释是:可以相信或者可以信赖.权威和可信两者所作用的对象是完全不同的.权威强调的是某个人、某种思想体系或某种组织,由于其活动内容的价值、功绩或品德被社会所公认,产生了一种使人能够信服的力量并被公众自愿认可、服从和支持.而可信仅仅是指某条消息是可以相信的,并不涉及消息发出者的可信程度,或许消息发出者发出的这一条信息是可信的,但是发出的另一条消息可能却是不可信的.因此,权威度高的一般可信度较高,但是可信度高的不一定具有权威度.

一般认为[3],在微博体系中,从微博用户类型角度来看,组织机构的权威性要高于个人用户,认证用户的权威性要高于非认证用户;从微博社会传播网络角度来看,影响力越大的用户,权威性越高.但是,这些评价方式都过于泛化,没有构成一定的评价体系,对于以用户权威度为基础而开展其它研究来说,目前缺乏一种定量的用户权威度评价模型.因此,本文将微博用户权威度的定量评价作为研究课题,希望为微博用户权威度的定量评价提供一个可行的研究思路和解决方案.

为方便后面的研究,本文对于微博用户权威度做如下定义:

定义 1(用户权威度 UA)　微博用户权威度 UA(User Authority)为微博用户在微博关系网络中具有的影响力与公众对其信服的程度.

其中,对微博用户在微博关系网络中的影响力与公众对其信服程度定义如下:

定义 2(用户影响力 UI)　微博用户影响力 UI(User Influence)由微博用户在微博关系网络中的领域影响因子及其发布的信息被传播的程度所组成.在本文中,主要是构建一种通用的用户权威度定量评价模型,暂不考虑微博用户的领域影响因子,只考虑用户所发信息被传播程度,我们将此称为用户信息传播影响力.

定义 3(被信服度 CD)　微博用户被信服度 CD(Convinced Degree)为微博用户由其信息的完整程度、一段时间内的活跃程度以及微博官方的评价等因素所构成的威望程度,也是被其它微博用户所认可的程度.

2 研究现状分析

虽然对于微博用户权威度的研究很少有文献涉及,但是其研究方法和思路与目前研究较多的用户可信度和排名有一定相似之处,它们的研究方法和思路对于用户权威度的研究有一定的借鉴意义.目前,国内外在用户可信度和排名方法方面的研究已经取得了很多成果,国外的研究大多集中在 Twitter 用户上[4~6],而国内则集中在新浪微博用户上[7~9],其目的都是为了构建一种合理的用户可信度和排名模型,为微博信息的获取、舆情的分析、社会关系网络的挖掘、社交网络垃圾邮件的识别以及品牌推广等方面的研究[10~14]提供支持和帮助.

目前,用户可信度和排名的方法主要有以下几类:第一类方法,也是最简单的方法,就是通过用户的粉丝数量直接衡量用户的可信度和排名.目前,国内的很多在线网站就是直接利用这种简单的方法提供用户排名的,但是这种方法是不够严谨的,因为在微博平台中存在一种称为“僵尸粉”(即虚假粉丝,一般不发表或很少发表微博,也不参与或很少参与话题讨论,花钱就可以买到关注)的用户.第二类方法,也是目前比较主流的用户可信度和排名方法,就是参考搜索引擎中常用于网页排名的 PageRank 算法,构建微博用户可信度和排名评价体系.如:程传鹏等[15]在分析了微博用户的领域集中度、微博用户之间的兴趣相似度和关注度对微博用户领域权威度的影响之后,依据用户关注与被关注的关系,参考 PageRank 算法,提出了一种面向领域的用户权威度的计算方法.值得说明的是,文献[15]中所指的权威度的概念实际与可信度的概念是基本相似的,在文献中并没有区分,与本文提出的权威度是截然不同的两个概念.Weng 等[16]通过对 Twitter 的传播进行分析,发现 Twitter 中存在一种同质性的现象,即用户因为某种共同的局部利益而存

在一种严重的“跟随”现象. 基于此,文献[16]采用LDA(Latent Dirichlet Allocation)主题模型识别用户所发Tweets的主题,并提取出用户所感兴趣的主题,然后针对特定的主题,构造Twitter用户的关系网络,最后参考PageRank算法,构建了一种TwitterRank算法,用于计算Twitter用户话题相似性排名,从而确定用户的影响力. Yamaguchi等[17]在考虑Twitter中实际信息流的基础上,基于链接分析,提出了用于计算用户可信度和排名的评分算法TURank(Twitter User Rank). 第三类方法是考虑用户自身信息,包括用户发布的微博、转发的微博、评论的微博、关注与被关注的情况以及其活跃度等方面的因素,综合评价用户的排名. 如:王君泽等[18]根据用户关注数量、粉丝数量、是否验证身份和发布的微博数量四个方面,构建了微博意见领袖识别的多维模型,提出微博用户重要性评分公式. Bakshy等[19]在研究用户关注关系影响力时,指出用户之间的链接关系对用户影响力的体现不大,需要结合被转发及被评论的数据进行全面考虑.

综上所述,用户可信度和排名的衡量方法无非集中在两个方面,一个方面是利用用户的主动行为信息,包括用户关注、发布微博、转发微博、评论微博以及用户个人基本信息等;另一个方面就是利用用户被动行为信息,包括用户被关注、微博被转发、微博被评论等. 本文综合考虑这两个方面的信息,同时,将微博官方的认证体系引入到用户的权威度评价之中,将用户的权威度评价转化为用户影响力和被信服度的定量计算,在暂不考虑用户领域影响因子的情况下,基于用户信息传播影响力、用户信息完整度、用户活跃度、用户平台认证指数四个方面,采用层次分析法,构建了基于多特征融合的微博用户权威度定量评价体系.

3 用户权威度定量评价模型的构建

3.1 微博用户特征的分析

本文以新浪微博为研究对象,根据用户权威度评价的要求,通过分析新浪微博的用户信息平台和微博用户的行为,提取出了微博用户的个人信息和行为数据,另外,新浪微博的官方用户认证平台属于微博官方认证,经过了微博官方的审查,具有较高权威性,也将其作为本文用户权威度评价的一个重要指标. 通过对微博用户权威度影响因素的分析与整理,在暂不考虑用户领域影响因子的情况下,构建了基于多特征融合的微博用户权威度定量评价的四个评价维度:用户信息传播影响力、用户信息完整度、用户活跃度以及用户平台认证指数,各维度包含的用户信息如表1所示.

表1 用户权威度评价特征

特征类别		微博用户信息
用户影响力	用户信息传播影响力	用户粉丝数
		微博被转发次数
		微博被评论次数
	用户领域影响因子	暂不考虑
被信服程度	用户信息完整度	基本资料
		联系方式
		职业信息
		教育信息
		标签信息
	用户活跃度	用户关注数目
		用户发布微博总数
		用户评论微博总数
		用户注册时间
	用户平台认证指数	是否为认证用户
		是否为VIP用户
		账号等级

3.2 基于多特征融合的微博用户权威度评价体系的构建

基于以上分析,我们构建了用户权威度定量评价体系,其层次结构如图1所示:

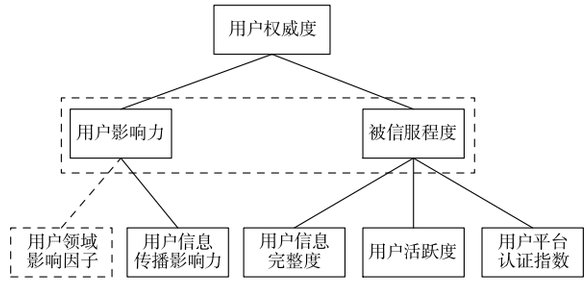

图1 用户权威度评价体系层次结构

按照图1所示的用户权威度评价体系层次结构并结合定义1、定义2和定义3,在暂不考虑用户领域影响因子的情况下,我们构建了用户权威度(User Authority,简称UA)评价特征四元组E(UIS,UII,UAD,UPI),其中,UIS为用户信息传播影响力,UII为用户信息完整度,UAD为用户活跃度,UPI为用户平台认证指数. 将微博用户权威度的计算转化为评价特征四元组E(UIS,UII,UAD,UPI)中各个评价特征的线性加权和,对于用户u_i,其用户权威度$UA(u_i)$计算如式(1)所示.

$$UA(u_i)=w_1\cdot UIS(u_i)+w_2\cdot UII(u_i)+w_3\cdot UAD(u_i)+w_4\cdot UPI(u_i) \quad (1)$$

其中,$w_i(i=1,2,3,4)$为各个评价特征的权值系数,满

足 $w_i>0$ 且 $\sum_{i=1}^{4} w_i = 1$. 式(1)构建的用户权威度评价体系将用户权威度的评价转化为定量的计算过程,实现了用户权威度的定量评价.

4　用户权威度评价特征权值的确定

美国运筹学家 Saaty[20] 提出了一种解决复杂的多因素决策问题的层次分析法(Analytical Hierarchy Process,简称 AHP). 它将目标分解为多个子目标,建立多要素、多层次的评价系统,采用定性与定量相结合的方法,通过定性信息定量化的途径,使复杂的评价问题变得可定量计算. 本文构建的微博用户权威度定量评价体系,采用层次化的分层结构,将用户权威度的评价问题分解为用户信息传播影响力、用户信息完整度、用户活跃度以及用户平台认证指数四个维度的定量计算问题,其结构与思想正好与层次分析法相吻合. 因此,本文将采用层次分析法确定用户权威度评价特征的权值.

本文在层次分析法的基础上,将利用特征向量法对用户权威度评价特征的权值进行计算,具体的计算过程如下:

第 1 步　根据用户权威度评价特征四元组构造用户权威度评价特征的判断矩阵.

设用户评价特征的判断矩阵为 $\boldsymbol{A}_{\mathrm{UA}}$,其中的元素 a_{ij} 表示特征 i 比特征 j 对评价目标影响的重要程度的倍数. 本文认为,最能体现用户权威的首要因素是用户信息传播影响力,其次是用户平台认证指数,因为它是经过了官方的审查,第三是用户活跃度,第四是用户信息完整度,之所以将用户信息完整度列为第四,是因为用户信息中有可能含有虚假的成分,对用户权威度的度量造成干扰,因此,将其重要程度设置为最小. 为了表征判断矩阵中各个特征的重要程度,一般引入数字 1 ~ 9 及其倒数作为度量[21]. 基于以上分析,我们按照四元组 E(UIS,UII,UAD,UPI)中评价特征的顺序,为判断矩阵设置初值,并经过多次迭代计算及一致性检验后,得到式(2)所示的用户权威度评价特征的判断矩阵 $\boldsymbol{A}_{\mathrm{UA}}$. 具体的迭代计算及一致性检验过程在下面第 2 步中进行论述.

$$\boldsymbol{A}_{\mathrm{UA}}=\begin{bmatrix} 1 & 9 & 3 & 2 \\ 1/9 & 1 & 1/3 & 1/5 \\ 1/3 & 3 & 1 & 1/2 \\ 1/2 & 5 & 2 & 1 \end{bmatrix} \tag{2}$$

第 2 步　求解判断矩阵最大特征值的特征向量,并进行一致性检验.

$$\boldsymbol{D}=\begin{bmatrix} 4.0080 & 0 & 0 & 0 \\ 0 & -0.0040-0.1785\mathrm{j} & 0 & 0 \\ 0 & 0 & 0.0040+0.1785\mathrm{j} & 0 \\ 0 & 0 & 0 & 1.9067\mathrm{e}-16 \end{bmatrix} \tag{3}$$

$$\boldsymbol{V}=\begin{bmatrix} 0.8413 & -0.7942+0.0000\mathrm{j} & -0.7942+0.0000\mathrm{j} & 0.9733 \\ 0.0909 & -0.0267-0.0266\mathrm{j} & -0.0267-0.0266\mathrm{j} & -0.1622 \\ 0.2611 & 0.1965+0.1992\mathrm{j} & 0.1965+0.1992\mathrm{j} & 0.1622 \\ 0.4646 & 0.2243-0.4892\mathrm{j} & 0.2243-0.4892\mathrm{j} & -0.0000 \end{bmatrix} \tag{4}$$

对上述式(2)的用户权威度评价特征的判断矩阵 $\boldsymbol{A}_{\mathrm{UA}}$ 求取其全部特征值,构成对角阵 $\boldsymbol{D}$ 如式(3)所示,并求取 $\boldsymbol{A}_{\mathrm{UA}}$ 的所有特征值的特征向量,将其构成的列向量矩阵 $\boldsymbol{V}$ 如式(4)所示.

通过式(3)我们可以得到 $\boldsymbol{A}_{\mathrm{UA}}$ 的最大特征值 $\lambda_{\max}=4.0080$,通过式(4)可以得到与 $\lambda_{\max}$ 对应的特征向量为 $\boldsymbol{w}=[0.8413,0.0909,0.2611,0.4646]^{\mathrm{T}}$. 接下来需要对 $\boldsymbol{A}_{\mathrm{UA}}$ 进行一致性检验,具体过程如下[21]:

①计算一致性指标 CI(Consistency Index)

判断矩阵的一致性指标 CI 度量了判断矩阵的平均偏离一致性,CI = 0 是判断矩阵一致性的充要条件,而且 CI 越小,判断矩阵偏离一致性程度就越小[22]. 对阶数为 n 的判断矩阵,其 CI 的计算公式如式(5)所示.

$$\mathrm{CI}=\frac{\lambda_{\max}-n}{n-1} \tag{5}$$

②确定平均随机一致性指标 RI(Random Index)

平均随机一致性指标 RI 是一致性指标 CI 的期望,表示 CI 的集中程度[22]. Saaty 通过实验计算出了各阶判断矩阵的 RI 值[21],当 $n=4$ 时,RI = 0.89.

③计算一致性比率 CR(Consistency Ratio)

一致性比率 CR 为 CI 与 RI 的比率,如式(6)所示. 当 CR < 0.1 时[21],所构建的判断矩阵的一致性是可以接受的,否则,需要对判断矩阵做相应的调整,再进行迭代计算.

$$\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}} \tag{6}$$

通过一致性检验,本文构建的判断矩阵 $\boldsymbol{A}_{\mathrm{UA}}$ 的一致性比率 CR = 0.0030,远小于 0.1,符合一致性检验结果.

第 3 步　对上面求的特征向量 $\boldsymbol{w}$ 进行归一化处理,即可得到式(1)中各个评价特征的权值为 $(w_1,w_2,w_3,w_4)=(0.5075,0.0548,0.1575,0.2802)$.

5　用户权威度评价特征的提取计算

5.1　用户信息传播影响力计算模型——UIRank

微博用户信息的传播主要依靠用户之间的关注与被关注的关系,根据微博用户之间的关注关系可以构建一种用户关注关系网络有向图如图 2 所示,其中圆点代表微博用户,箭头代表关注与被关注的关系. 用户间的这种网络关系与网页之间的链接结构非常相似,在微博中,一个微博用户关注其它的微博用户相当于网

页中从当前页面链接到其它页面,同样,一个微博用户被其他微博用户关注相当于网页中当前页面被其它页面链接.

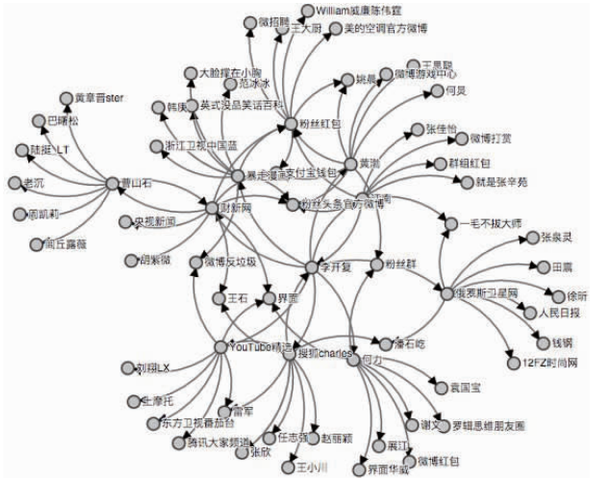

图2 用户关注关系网络

在网页搜索排名计算中常常使用 PageRank[23] 算法,它是目前很多重要的链接分析算法的基础,也是很多计算微博用户信息传播影响力的文献中常引用的算法[24,25]. 具体计算方法如式(7)所示.

$$PR(a_i) = q\sum_{a_j \in B}\frac{PR(a_j)}{L(a_j)} + \frac{1-q}{n} \tag{7}$$

对于式中各参数的实际意义参阅文献[23~26].

但是,在利用 PageRank 算法计算微博用户信息传播影响力时,只考虑了微博用户之间的关注与被关注的关系,没有考虑到"僵尸粉"的影响. 过多的"僵尸粉"会让用户的关注数目增多,从而增加了通过式(7)计算的用户信息传播影响力的结果. 因此,本文在 PageRank 算法和用户关注关系网络的基础之上,引入了用户被关注价值的概念,将微博用户之间的互动关系,即微博用户之间微博转发与评论的次数引入到用户关注关系网络中,构建了一种基于用户被关注价值的用户信息传播影响力计算模型——UIRank.

5.1.1 用户被关注价值

设 $FU(u_j,u_i)$ 表示用户 u_j 转发用户 u_i 微博的次数,$CU(u_j,u_i)$ 表示用户 u_j 评论用户 u_i 微博的次数,则 u_j 对 u_i 的关注度定义为如式(8)所示.

$$W(u_j,u_i) = \eta \cdot FU(u_j,u_i) + \lambda \cdot CU(u_j,u_i) \tag{8}$$

其中,$W(u_j,u_i)$ 表示 u_j 对 u_i 的关注度,η 和 λ 分别代表 u_j 转发与评论 u_i 微博的权值系数. 相较于评论微博,转发微博对用户 u_i 的信息传播作用更大,表明 u_j 对 u_i 所发布的微博比较认同,对其关注度更高,而评论微博尽管也表明 u_j 对 u_i 所发布的微博有关注,但对 u_i 所发微博并未进行传播,没有扩大 u_i 所发微博的影响力,因此,本文认为,转发微博的行为比评论微博的行为更能体现用户 u_j 对 u_i 的关注度,其比值设为 3∶2 比较合适,因此,在本文中,设置 $\eta = 0.6$,$\lambda = 0.4$. 采用微博用户之间的关注度定义用户被关注价值如定义 4 所示.

定义 4(用户被关注价值 FV) 若存在用户 u_i,其粉丝的集合为 B,则用户 u_i 的被关注价值 $FV(u_i)$(Followed Value)为其粉丝集合 B 中所有粉丝对其关注度之和,计算如式(9)所示.

$$FV(u_i) = \sum_{u_j \in B} W(u_j,u_i) \tag{9}$$

用户被关注价值越高,则当其关注别的用户时,对这些用户的影响力也就越高.

5.1.2 用户信息传播影响力模型——UIRank

在微博用户关注关系网络中引入用户被关注价值,对 PageRank 算法进行改进,构建基于用户被关注价值的用户信息传播影响力模型——UIRank,实现对用户信息传播影响力的量化,如式(10)所示.

$$UIS(u_i) = q\sum_{u_j \in B}\frac{UIS(u_j)\cdot W(u_j,u_i)}{FV(u_j)} + \frac{1-q}{n} \tag{10}$$

其中,$UIS(u_i)$ 表示用户 u_i 的信息传播影响力,u_j 为用户 u_i 的关注者集合 B 中的用户,q 与 PageRank 算法中一样,为阻尼系数. 由式(10)可以看出,用户 u_i 的关注者 u_j 的 UIS 值与两者之间的关注度的乘积除以 u_j 的被关注价值可以看作是用户 u_j 对 u_i 影响力,u_i 关注者集合 B 中所有用户的综合影响力,就决定了 u_i 的信息传播影响力. 因此,用户信息传播影响力与用户的关注者数目、关注者的 UIS 值以及其被关注价值有关.

5.2 基于微博信息平台的用户信息完整度的计算

微博用户的信息是其在微博社交网络中的名片,是用户真实性的一种反映,完整、全面、真实的用户信息会给人一种可以信赖的感觉,具有更高的权威度. 新浪微博平台本身具有一套完整的个人信息体系,各项个人信息在填写时,微博平台都设计了严格的格式校正,保证了信息的真实、有效. 其涉及的用户信息主要包括个人资料、联系方式、职业信息、教育信息以及标签信息五大类,十多个小类. 本文根据微博用户权威度评价的需要,构建微博用户信息完整度评价体系如表 2 所示,并定义用户信息完整度如定义 5 所示.

定义 5(用户信息完整度 UII) 用户信息完整度 UII(User Information Integrity Degree)为微博用户愿意向公众公开的个人信息标签数目占微博用户信息完整度评价体系标签数目的比例. 其计算如式(11)所示.

表 2　微博用户信息完整度评价体系

序号	标签	
1	昵称	个人信息
2	性别	
3	地区	
4	生日	
5	血型	
6	博客	
7	简介	
8	邮箱	联系方式
9	QQ	
10	职业信息	职业信息
11	教育信息	教育信息
12	标签信息	职业信息

$$\mathrm{UII}(u_i) = \frac{\sum_{j=0}^{n} \mathrm{IP}_j(u_i)}{n} \tag{11}$$

其中,$\mathrm{UII}(u_i)$ 表示用户 u_i 的信息完整度,n 为微博用户信息完整度评价体系中标签的总数目,根据表 2 所示,本文中 n 取值为 12,$\mathrm{IP}_j(u_i)$ 的定义如式(12)所示.

$$\mathrm{IP}_j(u_i) = \begin{cases} 1, & \text{序号为 } j \text{ 的信息标签对公众公开} \\ 0, & \text{否则} \end{cases} \tag{12}$$

5.3　基于用户行为频率的用户活跃度的计算

在微博传播体系中,微博用户自发的、主动的行为往往会增加微博用户的权威度. 在本文中,我们引入用户活跃度的概念来衡量用户在微博平台中的主动行为发生的频率. 微博用户的主动行为主要包括:微博用户关注其他用户、发布微博、浏览微博、转发微博、评论微博等. 简单来说,用户活跃度就是微博用户在平台中与其他用户进行互动的行为频率. 由于微博用户关注其它用户和浏览微博的行为很难搜集到时间节点的信息,因此本文中并没有将其纳入用户活跃度的度量体系之中. 根据微博用户发布微博的数目和评论其它用户微博的数目,定义微博用户活跃度如定义 6 所示.

定义 6(用户活跃度 UAD)　用户活跃度 UAD(User Activity Degree)为微博用户在一定时间内发布微博数目与评论其它用户微博数目的线性加权求和对时间的均值,计算如式(13)所示.

$$\mathrm{UAD}(u_i) = \frac{\sum_{j=1}^{n} (\alpha \cdot w_j(u_i) + \beta \cdot c_j(u_i))}{n} \tag{13}$$

其中,$\mathrm{UAD}(u_i)$ 表示用户 u_i 在时间段 n 内的活跃度,n 可以取近期的一个月或者一年,也可以定义为从注册日起到当前时间为止. $w_j(u_i)$ 与 $c_j(u_i)$ 分别表示用户 u_i 在第 j 天发布的微博的数目与评论其它用户微博的数目. α 和 β 分别表示用户发布微博数目与评论微博数目的权值,由于用户发布微博往往是一种有感而发的主动行为,而评论微博则是在看到别人的微博后所产生的一种有感而发的随动行为,相较于评论微博,发布微博更具有主动性,更能体现用户的主观思想,而评论微博则随发布微博而动,因此,本文认为发布微博的行为和评论微博的行为对用户活跃度的贡献比应设为 3:2 比较合适,这样我们可设置 $\alpha = 0.6, \beta = 0.4$.

5.4　基于新浪微博认证平台的用户平台认证指数的计算

平台认证为微博官方认证平台给出的评价,经过了官方的审查,具有较高的权威性. 根据前文表 1 提取的用户权威度评价特征,用户平台认证指数包含 3 项认证因素,分别为:是否为认证用户、是否为 VIP 用户以及微博账号等级,以此构建用户平台认证因素三元组 $\mathrm{UP}(A,G,V)$. 采用 $\mathrm{UP}(A,G,V)$ 构建用户平台认证指数的计算方法如式(14)所示.

$$\mathrm{UPI}(u_i) = \delta \cdot A(u_i) + \tau \cdot G(u_i) + \gamma \cdot V(u_i) \tag{14}$$

其中,$\mathrm{UPI}(u_i)$ 表示用户 u_i 的平台认证指数,$A(u_i)$ 表示用户 u_i 是否为平台认证用户对 $\mathrm{UPI}(u_i)$ 的贡献值,其计算方法如式(15)所示.

$$A(u_i) = \begin{cases} 1, & \text{用户 } u_i \text{ 为平台认证的用户} \\ 0, & \text{否则} \end{cases} \tag{15}$$

$G(u_i)$ 表示用户 u_i 在微博平台中的账号等级对 $\mathrm{UPI}(u_i)$ 的贡献值,其计算方法如式(16)所示.

$$G(u_i) = \frac{e^{g(u_i)} - e^{\min g}}{e^{\max g} - e^{\min g}} \tag{16}$$

其中,e 为自然常数,$g(u_i)$ 为用户 u_i 的账号等级,$\max g$ 为新浪微博平台中账号的最大等级数,取值为 48,$\min g$ 为新浪微博平台中账号的最小等级数,取值为 1,$G(u_i)$ 的取值为归一化后的结果,其取值范围为[0,1].

$V(u_i)$ 表示用户 u_i 是否为平台 VIP 用户对 $\mathrm{UPI}(u_i)$ 的贡献值,其计算方法如式(17)所示.

$$V(u_i) = \begin{cases} 1 & \text{用户 } u_i \text{ 为 VIP 会员} \\ 0 & \text{否则} \end{cases} \tag{17}$$

其中,δ、τ 和 γ 为 $\mathrm{UP}(A,G,V)$ 三者的权值,由于这三项特征之间相对的重要性并不相同,因此采用第 4 节中的特征向量的权值确定方法确定各自的权值. 本文认为,决定用户平台指数的首要因素是用户是否为平台认证用户,其次是用户是否为 VIP 用户,第三是用户的等级,之所以将用户等级列为第三,是因为微博平台中存在一些“刷等级”的用户,会对用户平台指数的评价造成干扰.

基于以上分析,按照用户平台认证因素三元组 UP(A,G,V)中认证因素的顺序,为判断矩阵设置初值,经过多次迭代计算和一致性检验,最终构建用户平台认证因素的判断矩阵 $\boldsymbol{A}_{\mathrm{UPI}}$ 如式(18)所示.

$$\boldsymbol{A}_{\mathrm{UPI}}=\begin{bmatrix}1 & 9 & 4\\ 1/9 & 1 & 1/2\\ 1/4 & 2 & 1\end{bmatrix} \tag{18}$$

通过计算,得到 A_{UPI} 的最大特征值为 $\lambda_{\max}=3.0015$,且一致性比率 CR = 0.0014,远远小于 0.1,符合一致性检验的相关要求,说明 A_{UPI} 是合理的. 最后,将 $\lambda_{\max}=3.0015$ 的特征向量进行归一化处理,得到了用户平台指数的各项认证因素的权值为(δ,τ,γ) = (0.7373, 0.0853,0.1773).

6 模型验证及实验结果分析

6.1 实验数据

本文的实验数据采用网络爬虫从新浪微博平台获取,获取到的数据包括:微博用户的基本公开信息、关注的微博用户、所有发布的微博、每条微博的获赞数目、每条微博的评论数目、每条微博的转发数目、微博的发布时间、用户获赞次数总和、微博被转发次数总和、微博被评论次数总和以及粉丝数目等用户数据,总共爬取微博用户 1209 个. 采用这些数据我们的实验主要包括两个部分:

(1)UIRank 模型合理性验证.

(2)微博用户权威度的定量评价模型验证.

6.2 UIRank 模型合理性验证

从爬取的微博用户数据集中抽取了五种比较有代表性的微博用户类型,构成微博用户集合{A,B,C,D,E},其关注关系网络的结构如图 3 所示.

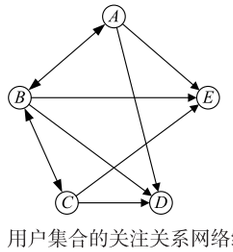

图3 用户集合的关注关系网络结构

在用户关注关系网络结构图中,微博用户彼此之间的转发与评论数目在爬取的数据集中并没有包含,同时,由于新浪微博平台的限制,要想大批量的获取这两个方面精确的数据也是很困难的. 在此,仅为验证模型的合理性,我们在新浪微博平台采用人工统计的方法,初步统计了五位用户间的互动数据,并且为了计算的方便,对数据进行了适当的处理,处理后的五位用户的互动数据如表 3 所示.

表3 用户转发评论数目

主动	被动	转发数	评论数
A	B	2	3
A	D	30	20
A	E	20	25
B	A	10	4
B	C	2	1
B	D	40	80
B	E	30	50
C	B	3	0
C	D	8	2
C	E	5	3

从表 3 中的数据可以看出,微博用户 A 与 B 都是较为活跃的用户,而微博用户 D 与 E 虽然没有关注别人,但获得的微博转发数目与微博评论数目很多,充分体现了知名人士这类微博用户的特点,而微博用户 D 与 E 在粉丝数目相同的情况下,D 的微博被转发与被评论的数目明显超过 E,因此,可以认为 D 的信息传播影响力大于 E.

本文分别采用传统的 PageRank 算法与本文提出的 UIRank 算法,计算微博用户的信息传播影响力,其计算结果如表 4 所示.

表4 UIRank 与 PageRank 比较结果

用户	PageRank	UIRank
A	0.1556	0.1371
B	0.2010	0.1506
C	0.1556	0.1307
D	0.2438	0.3108
E	0.2438	0.2708

从表 4 中实验结果可以看出,由 PageRank 算法得出的微博用户 D 与 E 的用户信息传播影响力相同,这与实际的分析是不相符的. 而本文提出的 UIRank 算法,考虑了用户的被关注价值对用户信息传播影响力的影响,由于用户 D 的微博被转发与被评论的数目明显超过 E,则用户 D 的被关注价值大于 E,因此用户 D 的信息传播影响力高于 E,更加接近于实际的情况,因此,可以认为本文提出的 UIRank 算法计算微博用户的信息传播影响力是合理的.

6.3 微博用户权威度的定量评价模型验证

在前面 6.2 节已经提到,微博用户彼此之间的转发与评论数目是很难大批量、精确的获取的,因此在利用 UIRank 算法计算大量用户信息传播影响力时,利用数据集中所有用户微博的被转发数目与评论数目的均值

来近似衡量. 同时,微博用户每天的评论数目也很难获得,在计算用户活跃度时,以爬取数据集中的微博用户发布微博数目的均值进行近似计算. 在计算了四项评价特征的具体值之后为了缩小波动性,还对各项评价特征的值进行归一化处理使得所有的评价特征的数值将被约束在[0,1]的区间内.

在进行实验时,首先从爬取的微博数据集中设定若干个起始种子用户,种子用户的类型包括知名人士、机构以及普通用户,然后以种子用户为起点,对其关注关系网络进行拓展,从数据集中抽取了包括种子用户在内的 1000 个微博用户的数据.

通过对实验结果进行分析,发现计算出的微博用户权威度基本符合实际情况,部分用户的权威度如表 5 所示.

表 5　用户权威值计算结果

权威值排名前 10 位		
用户昵称	用户类型	权威值
李开复	知名人物	0.9666
姚晨	知名人物	0.9654
网易云阅读	官方微博	0.9654
高晓松	名企高管	0.9588
青春如歌- 乡村教师代言人-马云	名企高管	0.9453
韩寒	知名人物	0.9433
新浪科技	官方微博	0.9331
中国移动	官方微博	0.9331
央视新闻	官方微博	0.9298
权威值排名后 10 位		
用户昵称	用户类型	权威值
爱吃牛奶的饼干	个人微博	0.0340
Dark _ Magi	个人微博	0.0313
治愈系情绪	个人微博	0.0297
晚秋 46712501	个人微博	0.0263
玥闻传媒	个人机构	0.0233
东篱一哥 _ 944	个人微博	0.0233
奇奇怪怪的 2570220155 _ 383	个人微博	0.0183
为你披白纱	个人微博	0.0168
水瓶天气真好	个人微博	0.0150

从表 5 可以发现,知名人士(如名企高管、娱乐明星以及官方认证机构)的用户权威度一般都较高,而那些普通的个人微博或者个人机构(广告性质)则处在一个较低的位置. 这是由于普通的个人微博如果没有经过实名认证,或者说是在微博关注关系网络中活跃性较低,那么这个微博用户被关注、被认可、被信服的可能性就较小,其用户权威度必然较低. 同时,根据实验结果统计发现权威度较低的微博用户大部分近期都没有发布微博的行为,且关注量与粉丝量也并不高. 因此,本文提出的基于多特征融合的微博用户权威度定量评价模型的计算结果是符合实际中人们对不同类别的用户群体权威度的认知[3]. 但是,模型的构建均是针对微博用户的历史数据进行的,因此本文提出的基于多特征融合的微博用户权威度评价模型计算的数值与 PageRank 算法一样,只能在一定的时间段内是有效的,随着时间的推移,微博用户行为的变化,其权威度也将会变化.

7　结束语

本文在将权威度与可信度比较的基础上提出了微博用户权威度的定义,指出微博用户权威度是由用户影响力和被信服度共同决定的. 然后在对新浪微博用户信息体系和微博用户行为分析的基础上,对微博用户特征进行提取,在暂不考虑微博用户领域影响力的基础上,总结出了用户信息传播影响力、用户信息完整度、用户活跃度以及用户平台认证指数 4 项影响用户权威度的评价特征,并分别给出了提取与计算方法,然后利用层次分析法确定各个评价特征的权值,将 4 个评价特征进行融合,构建了基于多特征融合的微博用户权威度定量评价模型. 同时,在对微博用户信息传播影响力的计算中,在用户关注关系网络的基础上,基于用户的被关注价值提出了一种 UIRank 用户信息传播影响力计算模型. 最后,通过相关实验验证了本文提出的基于多特征融合的微博用户权威度定量评价模型和 UIRank 用户信息传播影响力计算模型的合理性. 本文提出的多特征融合的用户权威度定量计算模型为微博用户权威度的定量评价提供了一种可行的计算方法.

在后续的工作中,我们将爬取更加丰富的微博数据,对微博用户的行为特征和影响力进行更加深层的量化计算,探求各个影响要素之间的相关性,对模型进行补充和优化,将用户权威度评价层次做的更加深入、细化,进一步提升模型的合理性与实用性. 同时,我们还将研究微博领域影响因子对于微博用户权威度的影响,构建基于领域的微博用户权威度定量评价模型.

参考文献

[1] 丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展,2014,51(4):691 - 706.

DING Z Y,JIA Y,ZHOU B. Survey of data mining for microblogs[J]. Journal of Computer Research and Development,2014,51(4):691 – 706. (in Chinese)

[2] 曹玖新,吴江林,石伟,等. 新浪微博网信息传播分析与预测[J]. 计算机学报,2014,37(4):779 – 790.
CAO J X,WU J L,SHI W,et al. Sina microblog information diffusion analysis and prediction[J]. Chinese Journal of Computers,2014,37(4):779 – 790. (in Chinese)

[3] 蒋盛益,陈东沂,庞观松,等. 微博信息可信度分析研究综述[J]. 图书情报工作,2013,57(12):136 – 142.
JIANG S Y,CHEN D Y,PANG G S,et al. Research review of information credibility analysis on microblog[J]. Library and Information Service,2013,57(12):136 – 142. (in Chinese)

[4] Pal A,Counts S. Identifying topical authorities in microblogs[A]. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining [C]. Hong Kong:ACM,2011. 45 – 54.

[5] Cha M,Haddadi H,Benevenuto F,et al. Measuring user influence in twitter. the million follower fallacy[A]. Proceedings of the Fourth International Conference on Weblogs and Social Media[C]. Washington,DC:AAAI,2010. 10 – 17.

[6] Achananuparp P,Lim E P,Jiang J,et al. Who is retweeting the tweeters? modeling,originating,and promoting behaviors in the twitter network[J]. ACM Transactions,2012,3(3):1 – 30.

[7] 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,2014,37(4):791 – 800.
MAO J X,LIU Y Q,ZHANG M,et al. Social influence analysis for micro-blog user based on user behavior[J]. Chinese Journal of Computers,2014,37(4):791 – 800. (in Chinese)

[8] 张绍武,尹杰,林鸿飞,等. 基于用户分析的微博用户影响力度量模型[J]. 中文信息学报,2015,29(4):59 – 66.
ZHANG S W,YIN J,LIN H F,et al. A micro-blog user influential model based on user analysis[J]. Journal of Chinese Information Processing,2015,29(4):59 – 66. (in Chinese)

[9] Wang N,Sun Q,Zhou Y,et al. A Study on influential user identification in online social networks[J]. Chinese Journal of Electronics,2016,25(3):467 – 473.

[10] Wang H,Lei K,Xu K. Profiling the followers of the most influential and verified users on sina weibo[A]. Proceedings of the IEEE International Conference on Communications[C]. London:IEEE,2015. 1158 – 1163.

[11] Lin C,He J,Zhou Y,et al. Analysis and identification of spamming behaviors in sina weibo microblog[A]. Proceedings of the 7th Workshop on Social Network Mining and Analysis[C]. New York:ACM,2013. 1 – 9.

[12] Deng Q,Liu Y,Deng X,et al. Semantic analysis on microblog data for emergency response in typhoon Chan-hom [A]. Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management[C]. New York:ACM,2015. 1 – 5.

[13] Lim K W,Buntine W. Twitter opinion topic model. extracting product opinions from tweets by leveraging hashtags and sentiment lexicon[A]. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management [C]. New York: ACM,2014. 1319 – 1328.

[14] 朱江,王柏,吴斌,等. 一种微博用户情感影响者发现模型[J]. 电子学报,2015,43(12):2497 – 2504.
ZHU J,WANG B,WU B,et al. A model for finding emotional influencers in microblog[J]. Acta Electronica Sinica,2015,43(12):2497 – 2504. (in Chinese)

[15] 程传鹏,夏敏捷,胡恩良,等. 一种面向领域的微博用户权威排名方法[J]. 小型微型计算机系统,2014,35(7):1538 – 1542.
CHENG C P,XIA M J,HU E L,et al. Ranking method for domain authority of micro-blog user[J]. Journal of Chinese Computer Systems,2014,35(7):1538 – 1542. (in Chinese)

[16] Weng J,Lim E P,Jiang J,et al. Twitterrank:finding topic-sensitive influential twitterers[A]. Proceedings of the Third ACM International Conference on Web Search and Data Mining[C]. New York:ACM,2010. 261 – 270.

[17] Yamaguchi Y,Takahashi T,Amagasa T,et al. Turank. twitter user ranking based on user-tweet graph analysis [A]. Proceedings of the 11th International Conference on Web Information Systems Engineering[C]. Berlin Heidelberg:Springer,2010. 240 – 253.

[18] 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.
WANG J Z,WANG Y L,YU H,et al. Research on micro blog opinion leader identification model[J]. Journalism & Communication,2011,(6):81 – 88. (in Chinese)

[19] Bakshy E,Hofman J M,Mason W A,et al. Everyone's an influencer. quantifying influence on twitter[A]. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining[C]. Hong Kong:ACM,2011. 65 – 74.

[20] Saaty T L. Decision making with the analytic hierarchy process[J]. International Journal of Services Sciences,2008,1(1):83 – 98.

[21] 汪应洛. 系统工程 [M]. 北京:机器工业出版社,2008. 120 – 130.
WANG Y L. Systems Engineering[M]. Beijing. Machine Industry Press,2008. 120 – 130. (in Chinese)

[22] 闫威,陈长怀,陈燕. 层次分析法一致性指标的临界值研究[J]. 数理统计与管理,2011,30(3):414 – 423.
YAN W, CHEN C H, CHEN Y. The threshold value of consistency index for analytic hierarchy process[J]. Journal of Applied Statistics and Management, 2011, 30(3): 414 – 423. (in Chinese)
[23] Page L, Brin S, Motwani R, et al. The pagerank citation ranking. bringing order to the web[OL]. http://ilpubs.stanford.edu:8090/422/,1999.
[24] Liang H, Lu G, Xu N. Analyzing user influence of microblog[A]. Proceedings of 2012 IEEE Fifth International Conference on Advanced Computational Intelligence[C]. Nanjing: IEEE, 2012. 15 – 22.
[25] Sun Q, Wang N, Zhou Y, et al. Modeling for user interaction by influence transfer effect in online social networks [A]. Proceedings of 39th Annual IEEE Conference on Local Computer Networks[C]. Edmonton: IEEE, 2014. 486 – 489.
[26] Boldi P, Santini M, Vigna S. Pagerank: functional dependencies[J]. ACM Transactions, 2009, 27(4): 1139 – 1141.

作者简介

张仰森　男,1962 年 6 月出生于山西临猗,博士后,教授,研究方向为中文信息处理、人工智能.
E-mail. zhangyangsen@163.com

郑　佳　男,1991 年 10 月出生于湖北松滋,硕士研究生,研究方向为中文信息处理、情感分析.
E-mail. zhengjia0826@163.com

唐安杰　男,1990 年 11 月出生于江苏盐城,硕士,研究方向为中文信息处理.
E-mail: t_anjie@qq.com