第41卷 第6A期
2014年6月

计 算 机 科 学
Computer Science

Vol. 41 No. 6A
June 2014

微博社会网络重要用户节点筛选及舆情引导

谢天保[1] 张晓雯[1] 仵凯博[2]

(西安理工大学经济与管理学院 西安 710054)[1] (西安理工大学计算机科学与工程学院 西安 710054)[2]

摘 要 首先通过研究网络爬虫以及新浪微博的开放平台,设计实现新浪微博专用爬虫,获取研究数据。其次,通过实验得到重要用户节点指标,提出贝叶斯-PageRank算法筛选重要用户节点,并实验验证重要用户节点的有效性。最后通过对重要用户节点的监测实现网络舆情发现并给出相关舆情引导策略。

关键词 微博社会网络,专用网络爬虫,重要用户节点筛选,舆情引导

中图法分类号 TP391 **文献标识码** A

Important User Node Screening and Public Opinion Guiding for Microblogging Social Network

XIE Tian-bao[1] ZHANG Xiao-wen[1] WU Kai-bo[2]

(School of Economy and Management, Xi'an University of Technology, Xi'an 710054, China)[1]

(School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710054, China)[2]

Abstract Firstly, studied the theory of web crawler and open platform Sina microblogging, designed Sina microblogging special reptile, access to research data. Secondly, got indicators of important user nodes through experiments, proposed Bayesian-PageRank algorithm screening important user nodes, and verified through experiment. Finally achieved internet public opinion through the monitoring of important user node, and gave the relevant public opinion guidance strategy.

Keywords Micro-blog social network, Private network crawler, Important user node screening, Public opinion guiding

1 引言

微博社会网络实时展现微博用户的言论,形成的网络舆论对社会安全构成了一定的威胁。但海量的微博用户中只有极少数的"意见领袖"——重要用户节点对网络舆论的发展起到决定性的作用。筛选微博社会网络中重要用户节点,对其言论进行有效的监测、引导,能在一定程度上保证网络、社会生活的安全、和谐。然而,目前国内外针对微博社会网络中重要用户节点的筛选以及网络舆情的有效监控的研究仍处于起步阶段,监测、控制网络舆情安全迫在眉睫。本文通过新浪微博专用爬虫,获取网络结构数据和微博消息数据,通过实验得到重要用户节点筛选指标,提出贝叶斯-PageRanks算法筛选重要用户节点,在此基础上给出舆情引导的有效策略,为维护网络安全提供一定的支持。

2 国内外相关研究

目前国内外学者对微博在线社会网络重要用户节点筛选的研究大致归为微博用户行为模式、微博重要用户节点筛选。

.微博用户行为模式

当前国内外学者大量分析并验证了微博用户行为模式与信息传播的关系。Kwak等发现用户对信息的转发是微博网络信息传播的主要途径[1],Yang、Cha等认为网络信息传播的速度、规模与用户的被提及率、登录频度等行为模式有较高相关性[2,3],Tang等通过研究Digg社会网络,认为信息传播的关键包括用户行为同步性[4],Xu等研究Twitter社会网络,认为影响用户行为的因素包括用户兴趣、好友信息以及突发新闻等[5],Leavitt等以用户粉丝数评估社会网络用户影响力大小[6]。

上述研究对微博用户行为模式进行了分析,但针对微博社会网络中的重要用户节点的筛选以及网络舆情有效引导研究不够深入。

.微博重要用户筛选

Brown等基于K-shell分解提出重要用户识别方法[7],Wang等提出一种可调节参数的重要用户识别方法[8],Heidemann等结合PageRank算法与centrality评估用户重要性[9],Zhang等基于用户转发与评论行为评估用户重要性[10],Hallberg等通过微博转发量、转发者与作者关系来分析用户重要性[11]。

上述研究没有深入分析微博社会网络中舆情信息高冗余、传播速度快等特性,也没有研究微博社会网络信息的动态性以及舆情信息对筛选重要用户的影响。

3 新浪微博专用爬虫设计与实现

本文结合新浪开放平台,深入分析新浪微博数据,研究其网络拓扑结构与微博数据,设计实现新浪微博专用网络爬虫。其中网络数据爬虫离不开高频度访问,新浪对高频度访问有限制:新浪微博服务器会暂停来自同一IP的高频率访问服务;控制开放平台中同一应用每小时的服务次数。为高效爬

谢天保(1966—),男,博士,副教授,主要研究方向为物流与供应链管理,E-mail:312259015@qq.com;张晓雯(1990—),女,硕士,主要研究方向为物流与供应链管理,E-mail:wwkb03@sina.com(通信作者);仵凯博(1986—),男,硕士,主要研究方向为计算机应用技术。

取研究数据，本文采取的解决方案是：采用多个免费网络代理避免同一 IP 地址高频访问；采用多线程方式规避开放平台同一应用的访问次数限制[12]。

3.1 新浪微博结构数据的获取与存储

本文采用 Python 语言设计实现网络爬虫，已经获取新浪用户节点数据 1725671 条。微博用户的“关注”行为是微博拓扑结构的表现形式，因此，准确存储微博网络用户间的拓扑关系才能保证数据的真实性与完整性，将微博网络用户间的关系用有向图表示，如图 1 所示。

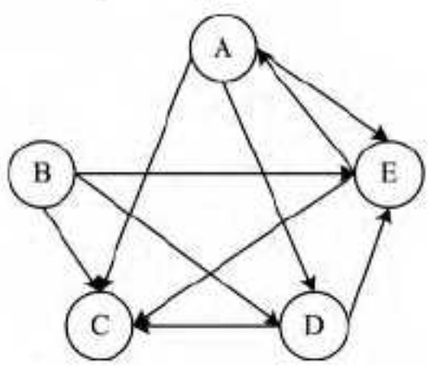


图 1 微博结构图

图 1 中 A、B、C、D、E 为微博网络中的用户节点，连接用户的边表示用户之间的关系，边的入度表示该用户的粉丝，边的出度表示该用户的关注。传统图论存储微博结构图一般采用邻接表或邻接矩阵，本文采用 XML 字段存储微博社会网络中粉丝、关注的关系。微博结构数据量巨大，存储时采用分表机制，其中用户表结构如表 1 所列。

表 1 用户表结构

序号	字段名	字段类型	是否可以为空	备注
1	UID	Varchar	否	UID
2	Nickname	Varchar	否	昵称
3	Fanscount	Int	否	粉丝数
4	Followcount	Int	否	关注数
5	Profilecount	Int	否	发微博数
6	URL	Varchar	否	用户微博 URL
7	Followlist	XML	是	关注者列表
8	Fanslist	XML	是	粉丝列表
9	Sex	Varchar	否	性别
10	Registime	Datetime	否	微博注册时间
11	Location	Varchar	是	所在地
12	Certifmark	Varchar	否	认证标志
13	Authentictype	Varchar	否	认证类型
14	Concerncount	Int	否	互粉数

3.2 新浪微博消息数据的获取与存储

微博消息数据通过爬虫采集 1 万新浪用户 3 个月发布的微博数据，共 1305679 条。本文采用分表机制存储微博消息数据以确保数据的真实性与完整性。其中，微博消息表反映了用户节点和消息之间的映射，微博 ID 和发微者 ID 实现了微博消息和发布微博用户之间的映射，如表 2 所列。

表 2 微博消息表结构

序号	字段名	字段类型	是否可以为空	备注
1	ID	Varchar	否	消息 ID
2	UserID	Varchar	否	发微者 ID
3	MID	Varchar	否	消息 MID
4	Createdtime	Datetime	否	微博发布时间
5	Weibotext	Varchar	否	微博内容
6	Weibofrom	Varchar	否	微博来源
7	RepliesID	Varchar	否	回复人 ID
8	GeograInf	Varchar	否	地理信息
9	Forwarcount	Int	否	转发次数
10	Commcount	Int	否	评论次数

新浪微博结构数据能够在一定程度上还原新浪微博网络结构，在此基础上的微博信息流能够真实地还原微博社会网络中信息在具体网络拓扑结构环境下的消息流动趋势。故此上述新浪微博数据的获取对本论文实验结果、结论具有极其重要的支撑作用。

4 微博社会网络重要用户节点筛选方法

4.1 重要用户节点与舆情信息的关系

通过实验发现微博社会网络中，微博用户粉丝规模呈现“重尾分布”[13]。为了深入研究重要用户节点与舆情信息的关系，追踪实验集新浪微博的用户节点一个月，每天在固定时间提取被检测用户信息，并对同一用户粉丝数量、关注者数量、发布微博数量做纵向分析，得到微博网络用户每日状态变化表示如下：ΔUserInfo（UID，Nickname，ΔFansCount，ΔFollowCount，ΔProfileCount，URL）。

通过上述数据进一步研究用户粉丝增长量，分析实验集合用户日粉丝增长量，并进行用户规模统计，得到 96.58% 的用户粉丝数日均增长 3.42%，3.42% 的重要用户节点却获得了 96.58% 的粉丝总量，表明在微博社会网络中用户每日新增连接（关注）数也符合“重尾分布”。

上述研究表明：重要用户节点的日均粉丝增长数明显高于非重要用户节点。其月均微博发布量柱状图及拟合曲线如图 2 所示。

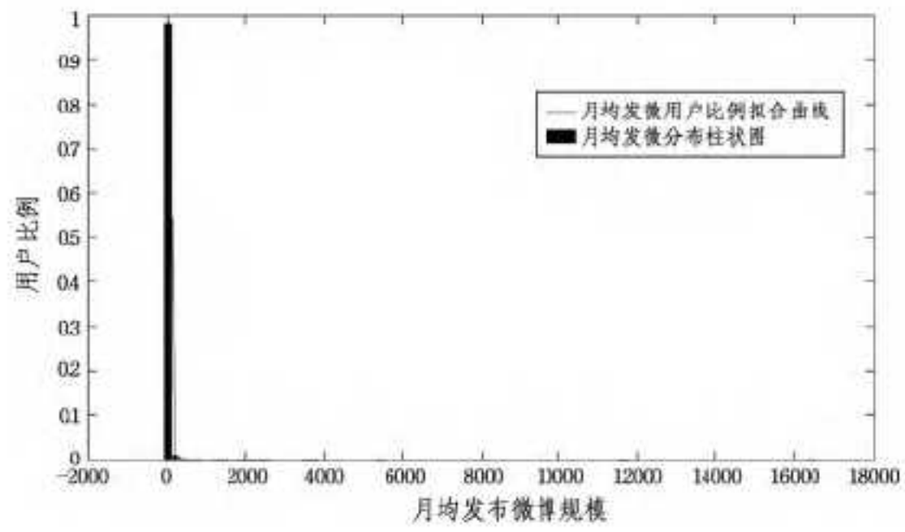


图 2 月均发布微博量_用户比例分布图

由上述分析可得，极少量的微博用户表现活跃，月均发布微博量远远超过了普通用户月均发布微博量。由此可见微博信息发布数量也存在重要用户节点，其对微博信息传播有极大影响力。

通过对微博社会网络中实验集用户粉丝数量、日均粉丝增长量和月均发布微博量的全面分析，表明有极少数用户表现活跃，且存在相似性，在一方面活跃的用户在其他两方面也表现突出。即有较大粉丝数量、日均粉丝增长量和月均发布微博量用户存在一定的交集，即为微博社会网络的重要用户节点。

4.2 舆情信息监测方法的核心策略

通过上述实验，可以明显地看出重要用户节点在微博社会网络中具有重要地位。首先，重要用户节点具有数量众多的粉丝群，所以其在微博社会网络的结构中具有极其重要的地位；其次，重要用户节点在每日微博网络新产生的链接中同样也能获取到较大数量的微博网络链接，即呈现出“富者愈富”的状况；最后，重要用户节点在微博社会网络中的消息贡献量也较大。由此可见重要用户节点在微博社会网络信息传播的过程中具有重要的作用。

微博社会网络的信息具有高冗余性，故在对其中的网络舆情进行监测时，传统方法往往不能满足时效性。结合上述实验结果，针对微博社会网络中的舆情信息监测，本论文提出通过基于重要用户的节点监测，实现对微博社会网络中舆情信息的监测。充分利用微博社会网络中重要用户节点在粉丝规模、粉丝增长数量、微博发布等方面的突出表现，对微博社会网络中的重要用户节点进行有效监测，可有效地降低微博社会网络舆情信息监测的复杂度，同时对微博网络中高冗余的信息也具有一定的过滤作用。

本文对微博社会网络舆情信息的监测通过对重要用户节点的监测来实现，对此本论文将按照下述步骤展开：首先，对微博社会网络中的重要用户节点进行筛选，得到微博社会网络中重要用户节点的集合；在此基础上，通过对比试验分析重要用户节点筛选方法的有效性；最后，在实际网络中验证基于重要用户节点的微博社会网络舆情信息监测方法的有效性。

4.3 微博社会网络重要用户节点的筛选

上述主要研究微博社会网络中重要用户节点在微博社会的网络拓扑结构和舆情信息传播过程中的重要特征，同时也对通过重要用户节点进行网络舆情检测的有效性进行了简单的论证。本节将论述基于贝叶斯-PageRank 的重要用户节点的筛选算法。

4.3.1 贝叶斯网络模型

贝叶斯网络模型的先验概率学习，主要有领域专家知识或数据学习构造两种方法。在本论文中采用领域专家知识获得先验概率，实验中首先人工标定新浪微博网络中有影响力的核心用户，然后，对具有影响力的重要用户节点的表 3 所列的属性值进行学习。

表 3 重要用户节点影响力考察表

属性值	备注
用户粉丝总数	
用户微博总数	
微博累计被转发数	1 个月内发布微被转发总数
微博累计被评论数	1 个月内发布微被评论总数
微博平均被转发数	1 个月内微博被转均数
微博平均被评论数	1 个月内微博被评均数

在确定重要用户节点的影响力考察属性之后，我们采用基于舆情事件的重要用户节点选取策略，选取新浪微博在 2013 年 1 月 20 日至 2013 年 1 月 26 日内舆情事件传播过程中的具有影响力的重要用户节点集合，其中以当时新浪微博内的事件：“腊八”、春运抢票、奥巴马就职、李娜闯入澳网四强、年会、赵本山退出春晚、处女座、201314、切糕、台湾保钓船、空姐代购被判刑、我是歌手、赵红霞、光盘行动、柴静、闯黄灯扣分、沈阳地震等作为舆情事件，并从中选取在舆情事件传播过程中具有一定影响力的微博用户作为重要用户节点，累计得到 94 位有效的重要用户节点。

为了得到先验概率，作者对此 94 位重要用户节点的属性（见表 3）进行了实验分析，并与本次实验集用户节点进行比对，标定的重要用户节点和实验集用户节点在粉丝规模方面具有较明显的区分度，故通过用户粉丝规模，我们拟定粉丝量-影响力概率公式如下：

$$P(Inf|Fans)=\begin{cases}1.0, & Fans\geqslant 100000\\ 1-\dfrac{100000-Fans}{100000}, & Fans<100000\end{cases} \tag{1}$$

其中，$P(Inf|Fans)\in[0,1]$表示在已知粉丝规模时用户成为重要用户节点的概率，$Fans$ 表示用户节点的粉丝数，Inf 表示重要用户节点。式(1)表示，针对用户节点若其粉丝量超过 10 万，则认为其成为重要用户节点的概率为 1，否则按照其粉丝规模与 10 万的差值量化其影响力。此公式能够有效地通过粉丝规模对用户重要性进行量化，其中对标定用户的准确率为 69.89%，通过公式可将实验集合 4.94% 的用户划归为重要用户节点，此结果与之前实验集合粉丝规模所呈现出的“重尾分布”情况较为符合，故式(1)可作为粉丝规模的影响力计算公式。

下面将对用户节点的月发布微博规模属性进行考察，标定的重要用户节点和实验集用户节点在每月发布微博总数规模的属性方面具有较明显的区分度，通过对用户每月发布微博总数规模的分析，我们拟定每月发布微博总数_影响力概率公式如下：

$$P(Inf|PostWeiboPerMonth)=\begin{cases}1.0, & PostWeiboPerMonth\geqslant 100\\ 1-\dfrac{100-PostWeiboPerMonth}{100}, & PostWeiboPerMonth<100\end{cases} \tag{2}$$

其中，$P(Inf|PostWeiboPerMonth)\in[0,1]$。$P(Inf|PostWeiboPerMonth)$表示在已知用户节点每月发布微博总数规模时用户成为重要用户节点的概率，$PostWeiboPerMonth$ 表示用户节点每月发布微博总数规模，Inf 表示重要用户节点。式(2)说明，用户每月发布微博数量在 100 条以上的即可认为是重要用户节点，否则根据其每月发布微博数量与 100 的差值，量化其影响力值。该公式有效地通过每月用户发布的微博数量规模对用户影响力进行了量化，其中对标定用户的准确率为 66.67%，通过此公式可将实验集 8.05% 的用户划归为重要用户节点，此结果较为符合微博社会网络所呈现的“重尾分布”特征，故式(2)可作为每月发布微博数量规模_影响力概率计算公式。

下面将对用户节点的月发布微博累计被转发规模属性进行考察，标定的重要用户节点和实验集合用户节点在每月发布微博累计被转发的规模属性方面具有较明显的区分度，故针对用户节点每月发布微博累计被转发的规模，拟定概率公式如下：

$$P(Inf|PostWeiboRepostPerMonth)=\begin{cases}1.0, & PostWeiboRepostPerMonth\geqslant 1000\\ 1-\dfrac{1000-PostWeiboRepostPerMonth}{1000}, & PostWeiboRepostPerMonth<1000\end{cases} \tag{3}$$

其中，$P(Inf|PostWeiboRepostPerMonth)\in[0,1]$。$P(Inf|PostWeiboRepostPerMonth)$表示在已知用户节点每月发布微博累计被转发规模时用户成为重要用户节点的概率，$PostWeiboRepostPerMonth$表示用户每月发布微博累计被转发的规模，Inf 表示重要用户节点。式(3)能够有效地通过每月发布微博累计被转发的规模对用户重要性进行量化，其中对标定用户的准确率为 81.72%，通过此公式可将实验集 1.53% 的用户划归为影响力用户，此结果较为符合微博社会网络所呈现的“重尾分布”特征。公式说明，用户节点每月发布微博累计被转发规模超过 1000 条，即可认为是重要用户，否则通

过其微博转发量量化其产生的影响力。故式(3)可作为用户节点每月发布微博累计被转发规模属性_影响力计算公式。

下面将对用户节点的每月发布微博累计被评论规模属性进行考察，标定的重要用户节点和实验集用户节点在每月发布微博被评论的规模属性方面具有较明显的区分度，故通过用户每月发布微博累计被评论的规模，我们拟定概率公式如下：

$$P(Inf|PostWeiboComsMonth)=\begin{cases}1.0, & PostWeiboComsMonth\geqslant 5000\\ 1-\dfrac{5000-PostWeiboComsMonth}{5000}, & PostWeiboComsMonth<5000\end{cases} \tag{4}$$

其中，$P(Inf|PostWeiboComsMonth)\in[0,1]$。$P(Inf|PostWeiboComsMonth)$表示在已知用户节点每月发布微博累计被评论规模时用户成为重要用户节点的概率，$PostWeiboComsMonth$表示用户节点每月发布微博累计被评论的规模，Inf表示重要用户节点。式(4)能够有效地通过月发布微博累计被评论的规模对用户重要性进行量化，其中对标定用户的准确率为76.34%，通过此公式可将实验集1.10%的用户划归为影响力用户，此结果较为符合微博社会网络所呈现的“重尾分布”特征。公式说明，用户节点每月发布微博累计被评论规模超过5000条，即可认为是重要用户，否则通过其微博评论量量化其产生的影响力。故式(4)可作为用户节点月发布微博累计被评论规模属性_影响力计算公式。

下面将对用户节点的每月发布微博平均被转发规模属性进行考察，标定的重要用户节点和实验集合用户在每月发布微博平均被转发的规模属性方面具有较明显的区分度，故通过用户每月发微平均被转发的规模，拟定概率公式如下：

$$P(Inf|PstWbAvgRePerMon)=\begin{cases}1.0, & PstWbAvgRePerMon\geqslant 120\\ 1-\dfrac{120-PstWbAvgRePerMon}{120}, & PstWbAvgRePerMon<120\end{cases} \tag{5}$$

其中，$P(Inf|PstWbAvgRePerMon)\in[0,1]$。$P(Inf|PstWbAvgRePerMon)$表示在已知用户节点每月发布微博平均被转发规模时用户成为重要用户节点的概率，$PstWbAvgRePerMon$表示用户节点每月发布微博平均被转发的规模，Inf表示重要用户节点。式(5)能够有效地通过用户每月发布微博平均被转发的规模对用户重要性进行量化，其中对标定用户的准确率为67.74%，通过此公式可将实验集1.80%的用户划归为影响力用户节点，此结果较为符合微博社会网络所呈现的“重尾分布”特征。公式说明，用户节点每月发布微博平均被转发规模超过120条，即可认为是重要用户，否则通过其微博平均转发量量化其产生的影响力。故式(5)可作为用户每月发布微博平均被转发规模_影响力计算公式。

下面将对用户节点的每月发布微博平均被评论规模属性进行考察，标定的重要用户节点和实验集合用户节点在每月发布微博平均被评论的规模属性方面具有较明显的区分度，故通过用户每月发布微博平均被评论的规模，拟定概率公式如下：

$$P(Inf|PstWbAvgCmmPerMon)=\begin{cases}1.0, & PstWbAvgCmmPerMon\geqslant 30\\ 1-\dfrac{30-PstWbAvgCmmPerMon}{30}, & PstWbAvgCmmPerMon<30\end{cases} \tag{6}$$

其中，$P(Inf|PstWbAvgCmmPerMon)\in[0,1]$。$P(Inf|PstWbAvgCmmPerMon)$表示在已知用户节点每月发布微博平均被评论规模时用户成为重要用户节点的概率，$PstWbAvgCmmPerMon$表示用户节点每月发布微博平均被评论的规模，Inf表示重要用户节点。式(6)能够有效地通过每月发布微博平均被评论的规模对用户重要性进行量化，其中对标定用户的准确率为80.64%，通过此公式可将实验集2.64%的用户划归为具有影响力用户节点，此结果较为符合微博社会网络所呈现的“重尾分布”特征。公式说明，用户节点每月发布微博平均被评论规模超过30条，即可认为是重要用户，否则通过其微博平均评论量量化其产生的影响力。故式(6)可作为月发布微博平均被评论规模_影响力计算公式。

上述实验通过对用户的粉丝总数、用户微博总数、用户发布微博累计被转发数、用户发布微博累计被评论数、用户发布微博平均被转发数、用户发布微博平均被评论数方面的实验分析和研究，建立了用户属性_影响力的贝叶斯概率公式。在此基础上，利用乘法原理将各属性计算得到的重要概率相乘，从而计算获得用户节点的影响力值，并对用户节点的影响力值进行排序得到用户影响力结果，如表4所列。

表4　贝叶斯概率计算得到用户节点影响力量化结果

ID编号	昵称	粉丝影响力	月发微影响力	月转发影响力	月评论影响力	月均发微影响力	月均评论影响力	贝叶斯影响力
647263235	当当网	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1643971635	凤凰卫视	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1676751317	搜狗浏览器	1.000	1.000	1.000	1.000	1.000	1.000	1.000
…	…	…	…	…	…	…	…	…
1756648217	Style-Notes时装笔记	1.000	0.243	1.000	0.369	1.000	1.000	0.090
1251204491	叶千荣	0.723	1.000	0.928	0.377	0.652	0.529	0.087
1682883335	李桢航2012BAR	1.000	0.207	0.421	1.000	1.000	1.000	0.087
…	…	…	…	…	…	…	…	…
2376786091	SerenaWilliams小威	0.555	0.007	0.006	0.014	0.747	1.000	0.000
1303642135	916刘烨	0.278	0.253	0.007	0.057	0.022	0.378	0.000
1245369991	李立君	0.897	1.000	0.016	0.032	0.011	0.045	0.000

如表4所列，我们对实验集合用户节点分别从用户节点粉丝状况、用户节点月累计微博发布状况、用户节点月累计微博转发状况、用户节点月累计微博评论状况等方面进行影响力评估。最后得到用户节点贝叶斯影响力量化值。从上述实

验结果可明显看出，贝叶斯概率模型能够较好地对用户影响力进行量化，并且排序的结果较为符合实际情况。

4.3.2 PageRank 算法思想

PageRank 算法的基本原理是：将一个网页的级别或者重要性的排序问题转化成一个公共参与、以群体民主投票的方式求解的问题。网页之间的链接即被认为是投票行为。同时，各个站点投票的权重不同，重要的网站投票具有较大的权重，而该网站是否重要的标准还需要参照其 PageRank 值。这看似是一个矛盾的过程：即我们需要用节点 PageRank 值来计算节点 PageRank 值。这既像是递归，又像是迭代，似乎不能收敛，谷歌公司的创始人佩奇和布林证明了这个过程最终收敛值与初始值无关，此处不再赘述此过程的证明过程。

在微博社会网络中，用户之间的关注与被关注关系完全可以用图论的理论来解释，其中用户被关注的关系即为图论中节点的入度，入度的总数即对应用户的粉丝总数；另一方面用户的关注即为图论中节点的出度，出度的总数即对应用户的关注总数，如图 3 所示。

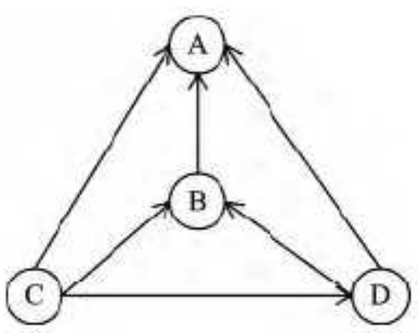


图 3 微博社会网络结构示意图

微博社会网络中的关系可描述为：

$$G=(V,E)$$

$$V=\{A,B,C,D\}$$

$$E=\{\overrightarrow{CA},\overrightarrow{CB},\overrightarrow{CD},\overrightarrow{BA},\overrightarrow{DA},\overrightarrow{DB}\}$$

$$Fans(A)=\{C,B,D\},Follow(A)=\phi$$

$$Fans(B)=\{C,D\},Follow(B)=\{A\}$$

$$Fans(C)=\phi,Follow(C)=\{A.B,D\}$$

$$Fans(D)=\{C\},Follow(D)=\{A,B\} \tag{7}$$

其中，G 表示微博社会网络拓扑图；V 表示微博社会网络中用户节点集合；E 表示微博社会网络中节点之间的关系集合，如 $\overrightarrow{CA}$ 表示用户 C 关注了用户 A，C 为 A 的粉丝；$Fans(A)$ 表示用户 A 的粉丝列表；$Follow(A)$ 表示 A 所关注的用户列表。其他的公式也是类似的解释，不再赘述。上述公式反映了微博社会网络用户之间的关系。

在充分研究微博社会网络用户节点的关注关系后，我们采用 PageRank 算法对用户节点间的关注关系所产生的影响力进行量化，进一步筛选微博社会网络之中的重要节点。

4.3.3 贝叶斯-PageRank 重要节点筛选方法

前两小节分别介绍了贝叶斯概率模型和 PageRank 算法的核心思想，本小节将着重介绍微博社会网络舆情信息的监测，对重要节点筛选所使用的贝叶斯-PageRank 算法，进而将此算法运用于实验集合验证该算法的有效性。

下面详细介绍本文所采用的核心算法：贝叶斯-PageRank。微博数据较高的冗余性和较强的突发性，导致经典的数据挖掘算法较难在保证准确性和时效性的情况下对微博社会网络进行舆情信息监测。因此本文提出采用基于重要用户节点的微博社会网络舆情信息监测方法，在该方法中，对重要用户节点的筛选是舆情信息监测的基础保证。对此我们在分析了微博社会网络用户的重要特性之后提出贝叶斯-PageRank 方法，其首先运用贝叶斯概率对微博社会网络中用户自身属性已知的情况下用户的影响力进行估算，得到微博社会网络中用户影响力量化表；其次，进一步运用 PageRank 算法对微博社会网络中用户节点之间的关系所导致的影响力传递进行估算。该方法在筛选微博社会网络中重要用户节点时不仅考虑了微博用户自身的属性，而且充分考虑到微博社会网络中用户之间的关注关系。下面将逐步介绍该方法的实现。

步骤 1 运用微博网络爬虫，爬取本次试验中所有被考察的用户节点的属性（包括用户粉丝规模量、用户微博总数、微博累计被转发数、微博累计被评论数）；在此基础上计算微博平均转发量和微博平均评论量。

步骤 2 通过手动标定的方式，从新浪微博社会网络中标定基于舆情事件的重要用户，并对标定出的重要用户节点通过微博网络爬虫获取其属性信息。在得到手动标定的重要用户节点的属性信息之后，通过对重要用户属性和实验集合用户属性的比对，得到在用户属性已知条件下用户影响力概率公式。

步骤 3 根据用户影响力概率公式，对实验集合所有用户采用概率公式得到其属性产生的影响力量化值。采用如下公式：

$$P(Inf)=\prod P(Inf|Attr) \tag{8}$$

其中，$P(Inf)$ 表示用户影响力值，$P(Inf|Attr)$ 表示用户属性为某一具体值时用户影响力的值，按照乘法原理，逐个考察用户的属性之后最终确定用户影响力的值

步骤 4 通过上述步骤，利用贝叶斯概率计算获得用户影响力值之后，仅考虑了用户的自身属性特征。紧接着利用 PageRank 算法对微博社会网络中的用户之间的关注关系进行量化。具体按照如下公式展开：

$$P_i(BPInf)=P(Inf)+\sum_{j=1}^{fansconut} P_j(trf) \tag{9}$$

$$P_j(trf)=P_j(BPInf)/followcount \tag{10}$$

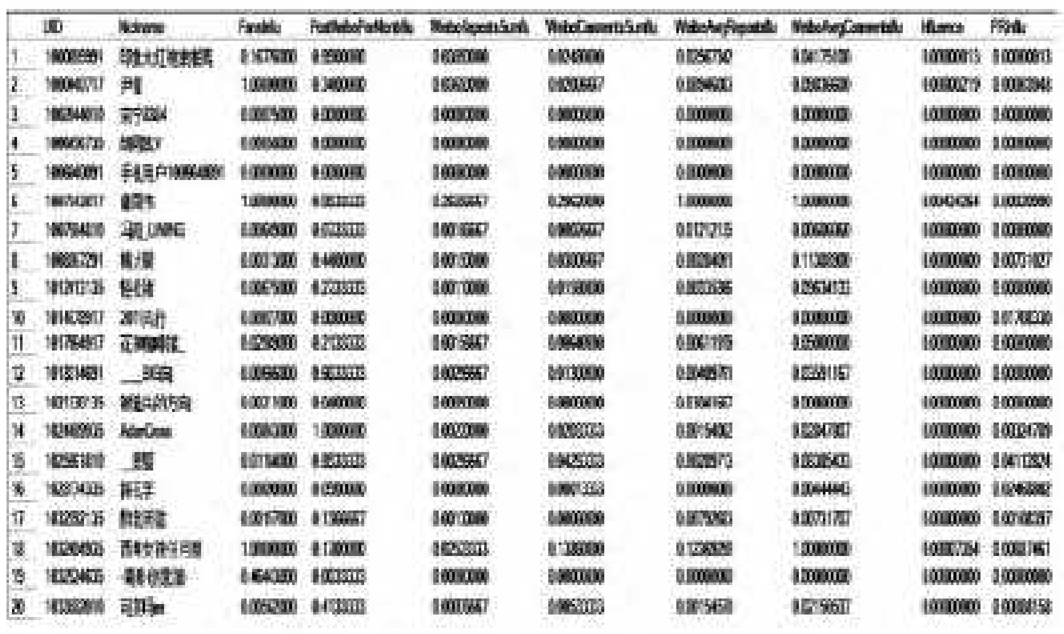
在式(9)和式(10)中，$P_i(BPInf)$ 表示用户节点 i 运用贝叶斯-PageRank 计算得到的影响力值；$P(Inf)$ 表示从节点属性方面评估节点影响力的量化值；$fansconut$ 表示用户节点 i 粉丝数目；$P_j(trf)$ 表示用户节点 i 的粉丝 j 通过关注关系，传递给被关注用户节点 i 的影响力；在式(10)中，$P_j(BPInf)$ 表示用户 j 运用贝叶斯-PageRank 计算得到的影响力值；$followcount$ 表示用户 j 关注数量。

步骤 5 按照步骤 4 即可求得所有实验用户集节点的影响力值。对此进行排序即可得到微博社会网络之中影响力较大的用户节点集合，即重要用户节点。

上述过程即为本文所提出的贝叶斯-PageRank 算法，下节将按照以上所述的算法步骤进行实验。

4.4 贝叶斯-PageRank 实验及结果分析

在上一节中，对本文重要用户节点筛选算法——贝叶斯-PageRank 的实现步骤做了详细描述，本节将在真实的实验集合上验证此方法。以微博用户节点的属性为条件，计算微博用户节点属性产生的影响力值，在此基础上量化微博用户之间的关注关系，最终得到用户的影响力，如图 4 所示。



	UID	Nickname	FansIdu	PostWeiboPerMonthIdu	WeiboRepostsSumIdu	WeiboCommentsSumIdu	WeiboAvgRepostIdu	WeiboAvgCommentIdu	Influence	PRIdu
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	尹雄	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	宋宁8384	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	胡明凯V	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	曹国伟	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

图 4 贝叶斯-PageRank 实验结果

图 4 为运用贝叶斯-PageRank 算法对实验集合数据进行处理后，用户影响力排名的最终结果。从结果可以明显地看出，将该算法运用于实验集合节点后得到用户节点“印象大红袍衷柏夷”、“尹雄”、“宋宁 8384”、“胡明凯 V”、“曹国伟”等排名较高。

对排名靠前的用户节点，我们逐一核实了其微博网站的属性及粉丝的状况。结果发现，重要节点用户集合的选取结果较为符合微博网络实际情况，另一方面重要节点之间的相对次序同样具有一定的准确率。故此算法针对微博社会网络中重要用户节点的筛选切实有效。

5 引导网络舆情

微博社会网络发展至今，给人们生活带来便捷的同时，也对网络安全造成了威胁。微博社会网络中的重要用户节点对维护网络安全至关重要。本文对微博社会网络重要用户节点进行筛选，通过分析重要用户节点，对网络舆情引导提出如下 3 个对策：

. 微博社会网络中的重要用户教育

对上述研究得到的微博社会网络中的重要用户进行持续教育，树立其维护网络安全的意识，加强其作为重要用户节点的社会责任感。开展网络道德教育的同时，加强网络法制教育。教育重要用户不散步谣言和反动言论，不诽谤他人，不传播有害信息和病毒，维护网络资源建设的安全。

. 加强舆情信息传播应急处理能力

舆情信息的传播速度快、传播范围广、影响力大，加强舆情信息传播应急处理能力，是为了避免失真舆情信息的传播对国民经济及生活造成危害。一方面要识别舆情信息中的虚假信息，及时辟谣，维护社会网络稳定；另一方面要正确面对舆情信息反映的问题，公正公平公开进行处理。

结束语 本文首先设计并实现了新浪微博专用爬虫，获得研究数据。其次通过实验得到重要用户节点指标，提出贝叶斯-PageRank 算法并利用其筛选重要用户节点，实验验证了重要用户节点的有效性。最后，通过对重要用户节点的监测实现网络舆情发现并给出相关舆情引导策略。

参 考 文 献

[1] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media[C] // Proceedings of the 19th International Conference on World Wide Web. 2010: 591-600

[2] Yang Jiang, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[C] // Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010: 355-358

[3] Cha M, Benevenutob F, Ahn Y-Y, et al. Delayed information cascades in Flickr: Measurement, analysis, and modeling[J]. Computer Networks, 2012, 56(3): 1066-1076

[4] Tang Si-yu, Blenn N, Doerr C, et al. Digging in the Digg Social News Website[J]. IEEE Transactions on Multimedia, 2011, 13(5): 1163-1175

[5] Xu Zhi-heng, Zhang Yang, Wu Yao. Modeling User Posting Behavior on Social Media[C] // Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2012: 545-554

[6] Leavitt A, Burchard E, Fisher D, et al. The influentials: New approaches for analyzing influence on twitter[R]. Web Ecology Project, 2009

[7] Brown P, Feng Jun-lan. Measuring User Influence on Twitter Using Modified K-Shell Decomposition[C] // Fifth International AAAI Conference on Weblogs and Social Media Workshop on the Social Mobile Web Brown. 2011: 18-23

[8] Wang Jian-wei, Rong Li-li, Guo Tian-zhu. A new measure of node importance in complex networks with tunable parameters [C] // 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008(WiCOM' 08). 2008: 1-4

[9] Heidemann J, Klier M, Probst F. Identifying Key Users in Online Social Networks: A PageRank Based Approach[C] // ICIS 2010 Proceedings. 2010: 79-100

[10] Zhang Meng, Sun Cai-hong, Liu Wen-hui. Identifying Influential Users of Micro-blog Services: A Dynamic Action-Based Network Approach[C] // PACIS 2011 Proceedings. 2011: 223-236

[11] Hallberg V, Hjalmarsson A, Puigcerver J. TweetRank: An adaptation of the PageRank algorithm to Twitter world[C] // Search Engines and Information Retrieval. 2012: 1-11

[12] 刘华. 网页信息抽取及建库系统 C# 实现[J]. 计算机工程, 2006, 32(16): 213-216

[13] 王一姝，郭海峰. 网络 IP 流量的自相似性分析与研究[J]. 科技信息, 2013(3): 127

(上接第 390 页)

[11] He X, King O, Ma W Y, et al. Learning a semantic space from user's relevance feedback for image retrieval[J]. Circuits and Systems for Video Technology, IEEE Transactions on, 2003, 13(1): 39-48

[12] Peng J, Heisterkamp D R, Dai H K. Adaptive kernel metric nearest neighbor classification[C] // 16th International Conference on Pattern Recognition. IEEE, 2002, 3: 33-36

[13] Bar-Hillel A, Hertz T, Shental N, et al. Learning distance functions using equivalence relations[C] // Proc. International Conference on Machine Learning, 2003

[14] Bar-Hillel A, Hertz T, Shental N, et al. Learning distance functions using equivalence relations[C] // ICML. 2003, 3: 11-18

[15] http://zh.wikipedia.org/wiki/%E5%BA%A6%E9%87%8F%E7%A9%BA%E9%97%B4 [OL]

[16] http://zh.wikipedia.org/wiki/%E8%B7%9D%E7%A6%BB [OL]

[17] Duda R O, Hart P E, Stock D G. Pattern Classification (Second Edition)[M]. 北京：机械工业出版社, 2010: 143-155