专题研究

微博舆情社会网络关键节点识别与应用研究*

王曰芬 杭伟梁 丁 洁 (南京理工大学经济管理学院信息管理系 江苏 210094)

摘要 文章在对国内外相关文献调查与梳理的基础上,从社会网络结构角度出发,设计舆情数据爬取系统,运用社会网络分析法来探索舆情网络中关键节点的识别与应用。具体应用时,以新浪微博为研究平台,通过爬虫系统抓取舆情数据并解析,构建微博舆情社会网络,运用改进 PageRank 算法对舆情网络中的关键节点进行识别与评价,实验结果证明本文所用算法是有效的。

关键词 新浪微博 网络舆情 社会网络 关键节点 改进 PageRank 算法

Identification and Application of Microblog Public Opinion Social Network Critical Node

Wang Yuefen Hang Weiliang Ding Jie
(Department of Information Management, School of Economics & Management, of Nanjing University of Science & Technology, Jiangsu, 210094)

Abstracts Based on investigating and reviewing the related literature at home and abroad, this paper embarks from the perspective of social network structure to design a crawler system of public opinion data, then uses social network analysis to explore the identification and application of the key nodes in the network of public opinion. In practical application, using the Sina Weibo as a research platform to grab public opinion data through the crawler system. By analyzing the data, the authors build a social network of microblog public opinion, then, use the improved PageRank algorithm to identify and evaluate the key nodes in the network of public opinion. The experimental results show that the algorithm used in this paper is effective.

Keywords Sina Weibo, network public opinion, social network, key nodes, improved PageRank algorithm

1 引言

在以网络平台与新媒体为数据源的研究中，将社会网络分析方法导入分析信息传播的特点与规律等渐成一种主要的研究趋势。例如：在国外研究方面,Akshay Java 等[1]以 Twitter 为研究对象,借用社会网络分析对用户使用 Twitter 的目的进行了总结分类;Teutle[2]构建了基于 Twitter 的社会网络,并利用密度、介数、出入度等特征指标的演化来描述 Twitter 社会网络的演化特征;Haewoon Kwak 等[3]通过对 Twitter 样本网络的分析探讨了微博社会网络的结构与特点，并与传统的在线社区进行比较;Krishnamurthy 等[4]从粉丝和关注数的角度分析了 Twitter 用户的特征;Huberman 和 Ping Liang 等[5-6]分别以 Twitter 微博和新浪微博为研究对象，并分析其微博社会网络结构的中心性;Karen Stepanyan 等[7]基于社会网络分析方法,对学生在 Twitter 中的交互行为进行研究，发现水平相当的学生更容易实现互动。在国内研究方面,平亮、宗利永[8]以新浪微博为研究对象,运用社会网络分析中的点度中心性、中间中心性和接近中心性测度方法对微博社会网络进行分析,发现微博中拥有“被关注数”的节点在舆情发展中发挥着更大的作用;金鑫[9]以真实的舆情事件为研究案

* 本文系国家社会科学基金重点项目“大数据环境下社会舆情与决策支持方法体系研究”(编号:14AZD084)、江苏高校哲学社会科学重点研究基地“社会计算与舆情分析”(培育点)的研究成果之一。

例,构建其在微博平台中形成的舆情社会网络,并研究了社会网络拓扑结构对舆情传播演化的影响;邱均平、李威[10]以“科学网博客”为研究对象,构建博主与评论者的2-模网络关系网络,并对2-模网络关系网络中节点中心度和小团体特征进行了分析。王学东等[11]利用社会网络分析方法,对学术博客的知识交流网络进行了分析;通过在网络上实时采集的数据,对之前的分析进行了验证,得到了可以表征该学术博客知识交流网络的分析数据。

在上述研究中,无论是国外的基于在线社会网络的分析还是国内的对微博或者博客社会网络结构与关系的研究,都需要以社会网络的关键节点识别与特征指标描述为基础。因此,识别出社会网络的关键节点并根据社会网络指标加以描述,是基于社会网络分析方法对数据进行挖掘的前提条件。那么,在舆情的社会网络分析中,如何抓取数据、获取关系并识别出关键节点,是值得深入研究的。

2 国内外社会网络的关键节点识别研究现状

社会网络分析需要两个基本要素,一是节点,二是节点之间的连线(关系)。在舆情研究中,拉扎斯菲尔德(Lazarsfeld)最早提出“意见领袖”这一概念,并将其定义为:在将媒介信息传给社会群体的过程中,那些扮演某种有影响力的中介角色[12]。在网络平台与新媒体下的社会网络中,意见领袖往往承担着重要的信息转达者和信息把关者的工作,并能够对周边用户产生重要的影响。由于意见领袖与社会网络中关键节点的概念相对应,两者代表的含义及意义在本文中完全一致,因此,下文将统一以“关键节点”的概念进行论述。

传统环境下的关键节点识别方法主要包括特征识别方法和调查量表方法[13]。但由于评价方法的过于主观性,不再适合网络环境下的关键节点识别研究。近几年研究中,国内外关于社会网络中的关键节点识别方法主要有层次分析法、聚类分析法、社会网络分析法、HITS算法及PageRank算法等。

层次分析法的关键节点识别是通过对节点的多属性特征分析来实现的,即依据意见领袖的特征和定义分析,构建关键节点的影响力指标体系,将综合影响力较高的节点识别为网络中的关键节点。如Darus等[14]利用AHP方法研究选择团队领袖的团队构建模型;刘志明等[15]利用层次分析法,基于用户影响力和用户活跃度两个维度,对微博意见领袖的评价指标进行量化,从数据结果上证明意见领袖是主题依赖的;管飘[16]基于AHP模型识别高校BBS中的意见领袖,并通过分析意见领袖的构成,提出切实可行的校园舆论引导策略。

聚类分析法的关键节点识别是将网络结构中相似度较大的节点进行归类,并依据以往经验中提取的意见领袖特征,从各子类中挑选出最符合关键节点特征的子类作为识别出的关键节点结果。如王珏等[17]识别出网络论坛用户的特征值向量,并在此基础上设计基于EM算法的用户聚类算法,并在其中筛选出最符合意见领袖特征的子类作为论坛中的关键节点;夏霖[18]在对话题聚合的基础上,识别出BBS中较为热门的话题事件,在此基础上,识别出对不同话题事件感兴趣的“兴趣团体”,并通过实验证明了聚类分析方法应用于意见领袖挖掘的可行性。

社会网络分析法来源于图论的理论,图论中众多描述节点属性的特征成为关键节点识别的重要依据。如Hon Wai Lam等[19]以在线拍卖网站eBay为研究对象,构建基于社会网络分析的BuyerRank模型识别商家所有购买者中的关键用户;丁雪峰等[20]赋予相对点度中心度、相对CPN级数、CPN点度中心势等各SNA参数不同的权重,构造识别网络舆论意见领袖综合指数及其算法;朱义生[21]根据意见领袖的特点,针对特定主题利用社会网络分析方法中的点度中心度、中间中心度和接近中心度以及网民的活跃性、支持度和传染性构建意见领袖指标体系。

HITS算法及PageRank算法是目前非常重要的意见领袖识别算法,这两种算法均基于网络超链接结构分析实现,部分学者同样将上述两种算法归类于社会网络分析。如Jun Zhang等[22]运用HITS等多种算法对Jave论坛中用户的权威度进行评价;Zhou Hengmin等[12]在PageRank算法基础上,提出改进的算法,并利用实验数据证明在PageRank算法基础上考虑情感因素,可以得到更准确的意见领袖识别结果;宁连举等[23]构建基于改进PageRank的网络意见领袖识别模型,并以北邮人论坛团购版块数据为样本进行实证研究;肖宇等[24]在传统PageRank算法基础上,利用用户回帖倾向性对用户间链接的权重重新赋值,构建新的基于倾向性分析的LeaderRank意见领袖发现算法,并利用实验分析验证模型的准确性和合理性。

为了支持社会网络分析研究,需要构建全面可靠的社会关系网络,使得基于该网络的分析结果可靠并且符合实际。但是,由于意见领袖大都具有话题依赖性,只有很少用户可以在不同主题内都成为意见领袖[25],有效地获取某个话题的数据与关系,然后根据数据所反映的关系构建相应的社会关系网络,再根据社会网络分析方法识别出关键节点,是目前常用的关键节点识别方法,也是研究过程中需要解决较多问题并加以实际应用的方法之一。

基于现有的研究与需求,本研究选取新浪微博作为数据来源,首先编写可靠的网络爬虫工具并抓取某个话题的相关微博数据,然后根据这些微博数据所反映的用户之间的关系构建社会网络,并对PageRank算法进行改进以识别出微博关键节点,以达到支撑微博数据进一步深化分析的目的。

3 微博舆情数据爬取与社会网络构建

微博(MicroBlog)是Web2.0时代新兴起的互联网社交应用,一条微博可以包含文本、图片、表情、话题标

签、视频、音频等多种形式的信息内容,字数一般限制在 170 字以内,若用户想发布更多字数的微博,则可以使用长微博工具,最多可支持 10000 字的文章。微博用户之间通过关注、转发、评论、回复、点赞、私信、搜索和 @ 提醒等行为来交流并建立联系。任何用户都可以对某一条微博进行转发,转发后会生成一条新的微博,并推送给转发者的粉丝,因此微博有原创微博和转发微博之分(为方便描述,本文把原创微博统称为“微博”,把转发微博统称为“转发”),每一条微博的转发量与转发内容都可以在该微博的主页上显示出来。同样,微博的评论也可以在微博的主页上看到,不过微博用户对某一条微博进行评论时,并不会像转发那样会产生一条新的微博,并推送给自己的粉丝。所以,在分析用户之间通过转发建立起来的关系往往比通过评论建立起来的关系更加牢靠,更加能说明转发者确实受到了被转发者发布微博的影响。因此,本文拟在爬去相关数据基础上通过转发关系来建立微博用户之间的社会网络。

3.1 微博舆情数据爬取研究

本研究需要某一事件从发生到结束或到当前时间的尽可能多的相关微博和这些微博的转发、评论信息,以及微博、转发和评论的发布者的基本信息。鉴于新浪微博平台的开放性与社会影响力,研究中常采用从该平台获取的数据,但是,新浪微博对搜索接口有着严苛的限制,不能进行关键词全文匹配,一次搜索只能返回最新的 200 条微博,不能满足大规模研究的需求。因此,需要通过开发系统利用爬虫技术来获取某主题相关的微博数据。而想要使用爬虫获取包含微博信息的网页源代码,必须在爬虫程序中执行 JavaScript 脚本,这给爬虫程序的编写带来很大的难度,同时还降低了爬虫程序的抓取效率。本研究借助 Selenium WebDriver[26] 执行 JavaScript 脚本,并获取脚本执行过后的网页源代码,通过使用解析网页源代码获取网页中包含的微博数据,包含微博的发布者编号、微博转发数、微博评论数、微博获赞次数、微博内容、微博 URL、微博 mid,其中 mid 为该条微博或转发或评论或私信在 web 系统中的 id 值[27]。通过微博 mid 可以得到该微博的转发和评论信息,同时利用微博发布者编号或昵称可以调用非开放平台的 JSON API 获取该微博用户的基本信息。

在具体设计时,主要考虑设计的需求、原则、功能与运行效果四个方面,并从需求分析、概念设计、数据库设计、技术选型、系统详细设计和实现五个环节入手。其中,在需求分析基础上的系统设计原则是为保证抓取数据的可用性与全面性,系统的功能需要考虑在逐页抓取微博、评论和转发时,能实时记录当前的抓取状态(抓取到第几页和当前的抓取时间等),使该系统在抓取微博、转发和评论时具有断点保存以及断点恢复的能力。在运行时需要考虑到网络中往往存在各种不稳定因素可能导致的各种异常,如不能连接服务器、连接服务器时间太长、请求失败以及服务器拒绝访问等,因此要求在系统运行时具有很完善的异常处理机制。在概念设计时,在以微博高级搜索功能为基础之上定义了抓取任务与搜索任务,一个抓取任务可以拆分为多个搜索任务,搜索任务的开始时间为抓取任务的开始时间,搜索任务的结束时间为抓取任务的当前时间,抓取任务由用户来设置,而搜索任务则由爬虫程序根据当前抓取状态动态生成,用户不必干预,若根据当前抓取状态生成的搜索任务执行时,搜索结果只有一页,则说明搜索已基本结束,抓取任务完成;在数据库设计时,根据需求抽象出微博数据、微博用户关系数据和系统运行日志数据三种类型的数据,其中:微博数据包括微博、转发、评论和微博用户等,微博用户关系类数据用于存储转发关系、评论关系和 @ 提醒等用户之间的各种关系等,系统运行日志类数据用于存储系统过去和现在的运行状态以保证系统能够稳定地运行以及数据能够全面抓取;在技术选型上,本研究采用 HttpClinet 的最新版本 4.3 来发送 HTTP 请求,调用 JSON API,并获取数据,而后选取 Fastjson 解析 JSON 数据,选用 jsoup 工具包解析 HTML 文本,借助 Selenium WebDrive 实现浏览器自动化操作。另外,选择 MySQL 数据库存储数据,使用 Spring JDBC 工具包来简化 JDBC 操作,同时使用 c3p0 数据库连接池;在系统详细设计与实现上,将功能设计为抓取模块、解析模块、数据存储模块和任务调度模块四个部分。抓取模块主要负责从新浪微博服务器抓取数据,包括微博、转发、评论、用户基本信息等,解析模块负责从 HTML 文本中和 JSON 格式数据中解析出需要的信息,存储模块负责保存数据到数据库中。任务调度模块为系统总调度模块,是系统的“大脑”,封装所有的业务流程操作。

3.2 社会网络的构建研究与特征指标描述

社会网络这一概念本身包括两层含义:一是本体论含义,表征现实存在的社会关系网络结构;二是方法论含义,表征为社会网络中涉及的主要研究方法。从本体论角度来说,所有满足有社会关系或社会结构的网状结构都可以称为社会网络。从方法论角度来说,社会网络是对社会关系结构及其属性加以分析的一套规范和方法[28]。在社会网络中,最基本的元素是节点和连线,其中节点代表社会行为者或用户,连线代表社会行为者之间的关系或者联系[29]。社会网络的构建首先需要确定节点 node(社会行动者)及节点间的连线 link(社会行动者间的关系)。

在微博舆情社会网络构建中,目前最主要的方式包括:(1)以舆情事件中的热点词汇为节点,以共现频次为连线。基于文本主题词共现来构建社会网络,可以获取整个舆论事件的核心主题词、主题词之间的关联,同时可以发现被边缘及解构化的主题词。(2)以舆情平台用户为节点,以关注和被关注为连线。关注和被关注情况可以分析表征微博网络的结构,但是无法表征特定舆情事件传播过程中的连接关系。(3)以舆情平台用户为节点,以评论或转发为连线。转发和评论关系是微博平台中用户交互的最重要途径,并且具有明确的事

件性和主题性。部分学者在转发评论频次的基础上,综合考虑回复及评论文本本身,在节点关联中增加权重,构建更能表征舆情传播的社会网络。

社会网络的特征指标从描述对象区分,可分为描述网络指标和描述节点指标两类。在综述的基础上,本研究汇总目前较为常用的社会网络指标,如表 1 所示。

表 1 社会网络特征指标汇总

类别	拓扑结构指标	说明
网络指标	网络节点及边数	网络中所有的节点数和边数,表征舆情热度
	网络密度	网络中各个节点彼此信息交流传播的紧密程度,表征互动紧密程度
	网络平均最短距离	网络中所有节点最短路径长度的平均值,表征沟通难易程度
	网络平均度	络中所有节点的度数中心度平均值,表征互动频繁程度
	网络聚类系数	网络集团化的程度,表征中心趋势
节点指标	节点出现频次	网络中节点出现次数,表征节点参与度
	节点度数中心度	网络中节点连接其他节点次数,表征是节点权力
	节点结构洞 Effsize	网络中节点控制结构洞数量,表征节点权力

在社会网络构建时,考虑到转发相比评论更能表征舆情事件中信息的传播方式,本研究选取 2014 年 2 月“官员夫妇殴打护士”这一特定案例为研究对象,利用上述自行开发的微博舆情数据爬取系统抓取该事件的数据,在进行处理的基础上构建以微博平台用户为节点、以转发关系为连线的舆情社会网络。

由于“官员夫妇殴打护士”舆情事件属于近年来非常普遍的“医患关系”和“仇官”类舆情事件,选取案例对象具有普遍性和合理性。

具体地,本研究利用新浪微博,以“护士被打”、“口腔医院打人”、“官员殴打护士”、“袁亚平”、“陈星羽”、“董安庆”等作为关键词,限定微博发布时间跨度为 2 月 15 日至 4 月 15 日,最后共获取 360 549 名微博用户,33 079 条原创微博信息及 424 898 条转发微博信息。由于新浪微博目前提供的接口中无法直接获取转发路径中涉及的所有用户,需要进行预处理操作,从爬取的微博数据中分离出转发路径上的中间用户[30]。微博信息中涉及的字符串包括“//@”“@”及“回复 @”等,其中“//@”字符表征微博用户之间的转发关系。由于转发信息中往往存在多个用户的传递转发关系,为防止转发关系的重复累加,设定单条转发微博信息只提取最近层的转发传递关系。从抓取的转发微博信息中,共获取 410 418 条转发关系,构建出以微博用户为节点,以转发关系为连接的微博转发社会网络。网络基本拓扑指标如表 2 所示。其中|V|表示节点数,|E|为边关系数,d 为整体网络密度,$\bar{L}$为网络平均最短距离,C 为网络聚类系数(Clustering Coefficient),kmax 为最大节点度,$\bar{k}$为平均节点度。

表 2 “官员夫妇殴打护士”舆情社会网络基本统计特征

特征值	|V|	|E|	d	$\bar{L}$	C	kmax	$\bar{k}$
	360549	410418	0.0011	1.819	0.074	26281	8.196

4 微博舆情关键节点识别及评价

社会网络 G 是由非空的节点(或顶点)V 和有限的边集 E 构成,在图论中的表征方式为 G=(V,E)。社会网络不仅研究边集 E 的特征变化,同时关注点集 V 中重要的行动者及其在社会网络中的等级和优势[28]。本研究在构建舆情社会网络构建的基础上,基于改进的 PageRank 算法实现微博平台的关键节点识别,以为进一步研究关键节点在舆情社会网络演化过程中的作用提供研究基础。

4.1 改进的 PageRank 微博关键节点识别算法

PageRank 算法是基于网页链接分析的经典网页排名算法之一。PageRank 的核心排序算法借鉴了传统引文分析中的思想:所有页面中的链接指向都是一次“投票”,当网页 $T_i(i=1,2,\cdots,n)$ 有链接指向页面 A,则页面 A 便获得了 T_i 对它“投票”的分值。但是分值的大小还取决于 T_i 本身的重要程度,即网页 T_i 的重要性越大,网页 A 获得的 PageRank 值(以下简称 PR 值)就越高。由于网络中网页链接的相互指向,一个网页的 PR 值总是递归地由其他网页的 PR 值决定,任何网页都对其他网页的 PR 值产生影响,该分值的计算为一个迭代过程,最终网页根据所得分值进行检索排序[31]。上述 PR 值分配过程的数学表达式如下:

$$PR(A)=(1-d)+d^{*}\left(\frac{PR(T_1)}{C(T_1)}+\cdots+\frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

式中,PR(A)代表网页 A 的 PR 值,$\frac{PR(T_n)}{C(T_n)}$代表有链接到网页 A 的网页 T_n 的 PR 值,$C(T_n)$代表网页 T_n 中所有的正向链接数,d 为阻尼因子,默认值 0.85。阻尼因子 d 的引入是为了有效解决 Rank Sink 现象,同时可以有效保证最后计算结果的收敛性。

微博用户转发网络与网络链接网络存在着类似的拓扑结构,转发行为同样存在链接的指向,可以利用 PageRank 算法对微博用户的权威 PR 值进行计算排序,其中 PR 值可以作为微博用户是否为有影响力的关键节点的重要指标。用户的 PR 值大小不仅取决于其被转发和被评论的简单出入度频次,同时取决于对该用户微博进行转发的用户重要性程度(即 PR 值)。

但是,微博用户转发网络与网页链接也存在一定差异,需要依据微博用户转发网络的特性对传统 PageRank 算法进行改进。网页 A 与 T_i 间的链接关系在单方向中只存在一条,但是微博用户的转发行为的链接关系更为复杂,除考虑链接关系的方向性外,还应有体现交互频次的权重数据 Weight。本研究定义 $Retweet\_times(T'_i,T'_j)$ 表示用户 T'_i 转发用户 T'_j 微博的频次,$Retweet_total(T'_i)$表示用户 T'_i 转发微博的总频次。若用户 T'_i 对用户 A' 的微博进行转发,则 $Weight(T'_i,A')$ 的计算公式表示如下:

$$Weight(T'_i,A)=\frac{Retweet\_total(T'_i)}{\sum_{j=1,j\neq 1}^{N} Retweet\_times(T'_i,T'_j)} \quad (2)$$

上式中 N 表征参与微博交互的用户总数。最终基于微博转发网络改进的 PageRank 算法(为与 PageR-

ank 区分,以下称为 WeiboRank 算法)表示如下:

$$WR(A')=(1-d)+d*\sum_{i=1,T'_i\neq A}^{N} Weight(T'_i,A)*WR(T'_i) \quad (3)$$

式子中 WR（A′）表示微博用户 A′的重要程度值,WR(T'_i)表示转发用户 T'_i的重要程度值。上式表明转发用户 T'_i的重要程度值不再均匀分配给被转发用户,而是依据 T'_i转发不同用户的微博数量按比例分配出去。

4.2 微博关键节点识别结果评价

本研究利用改进的 PageRank 算法对上文构建的“官员夫妇殴打护士”舆情社会网络中的关键节点进行识别。为验证关键节点识别结果的正确性和合理性,首先利用目前常用的关键节点识别算法评价指标——黄金准则(Golden Standard)及节点核心率(Coverage Ratio,公式表征如下)[18]对关键节点识别结果进行评价。

$$Coverage\ Ratio=\frac{前\ n\%关键节点覆盖的节点数}{总节点数} \quad (4)$$

黄金准则计算中，本研究选取目前基于社会网络的关键节点识别中最常用的节点中心度 Degree（以下简称 DE)为黄金指标,以 DE 的用户排序为基准,并以用户粉丝数 Fans_num(以下简称 FN)及用户出现频次 Occurrence Frequency(以下简称 OF)指标作为相关对比衡量指标,获取 WeiboRank、FN 及 OF 得到用户排序与黄金指标排序的 Kendall's tau-b 相关性指标，具体如图 1 所示;节点核心率计算中,仍以 DE、FN 及 OF 为对比衡量指标,计算各评价指标下的前 n%关键节点团体所覆盖的数值总比例,具体结果如图 2 所示。

由图 1 可知,WeiboRank 及 OF 与 DE 的 Kendall's tau-b 相关性指标明显高于 FN 的相关性指标系数;WeiboRank 相关性指标稳定波动在 0.6~0.7 之间,与 DE 指标相关性较高且保持稳定;OF 在 10%以内关键节点识别中与 DE 保持着高度的相关性,但随着关键节点比例的增加相关性指标不断下降。由图 2 可知,FN 算法关键节点覆盖率最高,前 1%的关键节点已经近乎完全覆盖全体用户数据;WeiboRank 及 DE 算法关键节点覆盖率相近，分别是由前 20%及前 30%的关键节点近乎覆盖全体用户数据;OF 无法实现关键节点的有效覆盖,Coverage Ratio 指标过低。总结来看,本研究构建的 WeiboRank 算法能够达到近乎 DE 指标的意见领袖节点核心率,并且能够规避 FN 识别意见领袖主题无关性及 OF 无法有效覆盖关键节点的缺陷,故本文的微博关键节点识别结果具有正确性和合理性。

4.3 微博关键节点识别结果分析

本研究进一步对关键节点的身份区分为政府、媒体及草根三种类别。由于总体用户数量巨大,下面首先对识别 Top10、Top20、Top50 及 Top100 关键节点身份及认证类型分布情况进行统计,具体如表 3 所示。

由表 3 可知,微博平台“官员夫妇殴打护士”事件中,Top10、Top20 及 Top50 关键节点以媒体类节点为主。媒体类关键节点在整体舆情事件中承担着重要的传播作用。Top100 关键节点中草根类关键节点的比例

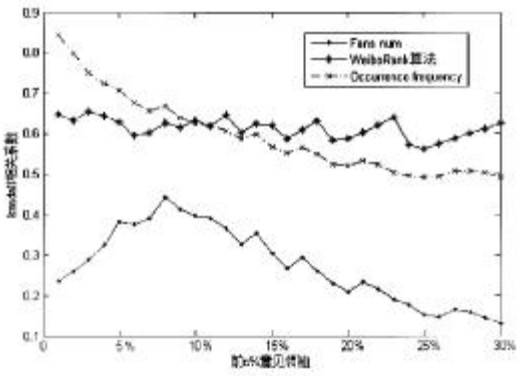

图 1 各算法与 degree 的相关程度曲线图

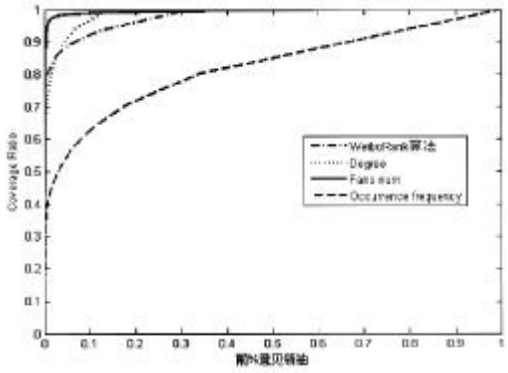

图 2 各算法 Coverage Ratio 指标曲线图

达到 57%,草根类关键节点的出现频次较高,但平均影响力 WR 值偏低。Top100 关键节点中政府类关键节点数量仅为四个(南京鼓楼医院、南京发布、南京玄武警方在线和江宁公安),且影响力 WR 偏低,在微博平台中对舆情演化影响力非常有限。

表 3 Top10、Top20、Top50 及 Top100 节点及其所属类别

	Top10	Top20	Top50	Top100
政府类	2(10%)	3(15%)	3(6%)	4(4%)
媒体类	8(80%)	14(70%)	27(54%)	39(39%)
草根类	0(0%)	3(15%)	20(40%)	57(57%)

为准确表征舆情事件中各类别关键节点在舆情社会网络演化过程中的作用，本文依次选取 TOP5 政府类、媒体类及草根类代表身份的核心关键节点作为研究对象(政府类仅 4 个),具体如表 4 所示。

表 4 TOP5 政府类、媒体类及草根类核心关键节点

	政府类关键节点	媒体类关键节点	草根类关键节点
1	南京发布	新浪江苏	马伯庸
2	南京玄武警方在线	央视新闻	刑法韩友谊
3	江宁公安在线	江苏身边事	摆古论今
4	南京鼓楼医院	人民日报	八卦_我实在是太 CJ 了
5	——	南京日报	孙海

5 结语

本文在对相关研究背景分析与国内外研究归纳总结基础上,提炼出如何抓取数据、获取关系并识别关键

节点的三个研究问题，并借助于现有的理论与方法提出具体的研究方案。

在微博数据爬取研究中，按照需求分析、概念设计、数据库设计、技术选型、系统详细设计和实现五个环节，设计开发了包含抓取模块、解析模块、数据存储模块和任务调度模块四个部分的爬取系统。

在舆情社会网络构建中，本文以舆情事件“官员夫妇殴打护士”为实证案例，以参与舆情讨论的微博用户为节点，以用户之间的转发关系为连线构建舆情社会网络。针对现有研究对舆情网络中的节点特征关注度不足的局限，将微博舆情网络的特征指标分为网络测度指标和节点测度指标两个部分，进一步扩充了现有社会网络特征演化的研究内容。

在网络关键节点识别研究中，本文利用基于微博改进的 PagerRank 算法实现了“官员夫妇殴打护士”舆情事件中的关键节点识别与验证，并具体得到如下的识别结果：(1)政府类关键节点的数量和平均 WR 值均偏低，在微博平台中对舆情演化影响力非常有限；(2)媒体类关键节点的平均 WR 值最高，且排序最为靠前，在整体舆情事件中承担着重要的传播作用；(3)草根类关键节点的平均 WR 值偏低，但出现频次较高，其在舆情事件中传播作用不可小觑。

参考文献

[1] Java A, Song X, Finin T, et al.Why we twitter: understanding microblogging usage and communities[C].Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis.ACM, 2007:56-65.

[2] Teutle A R M.Twitter:network properties analysis[C].Electronics, Communications and Computer (CONIELECOMP), 2010 20th International Conference on.IEEE, 2010:180-186.

[3] 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.ACM, 2010:591-600.

[4] Krishnamurthy B, Gill P, Arlitt M.A few chirps about twitter[C]. Proceedings of the First Workshop on Online Social Networks. ACM, 2008:19-24.

[5] Huberman B A, Romero D M, Wu F.Social networks that matter:Twitter under the microscope [J].arXiv Preprint arXiv: 0812.1045, 2008.

[6] Ping L, Zong L Y.Research on microblog information dissemination based on SNA centrality analysis—A case study with Sina microblog [J].Intelligence, Information & Sharing, 2010(8): 71-72.

[7] Stepanyan K, Borau K, Ullrich C.A social network analysis perspective on student interaction within the Twitter microblogging environment [C] .Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on.IEEE, 2010:70-72.

[8] 平 亮,宗利永.基于社会网络中心性分析的微博信息传播研究——以 Sina 微博为例[J].图书情报知识,2010(6):92-97.

[9] 金 鑫.基于复杂网络分析的微博网络舆情传播[J].吉林大学学报(工学版),2012,42(1):271-275.

[10] 邱均平, 李 威. 基于社会网络分析的博主与评论者关系研究——以“科学网博客”为例[J].情报科学, 2012, 30(7):959-963.

[11] 王学东,杜晓曦,石自更.面向学术博客知识交流的社会网络中心性分析[J].情报科学,2013,31(3):3-9.

[12] Lazarsfeld Paul F, Katz Elihu.Personal Influence [M].New York: Free Press,1957.

[13] 梦 非.社会化商务环境下意见领袖对购买意愿的影响研究[D].南京:南京大学,2012.

[14] Darus N M, Yasin A, Omar M,et al.Team formation model of selecting team leader:an Analytic Hierarchy Process (AHP) approach [J].ARPN Journal of Engineering and Applied Sciences, 2015,10(3):1060-1067.

[15] 刘志明, 刘 鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程, 2011, 29(6):8-16.

[16] 管 飘.高校 BBS 中意见领袖的识别与构成分析[J].新闻传播, 2013(5):225.

[17] 王 珏,曾剑平.基于聚类分析的网络论坛意见领袖发现方法[J].计算机工程,2011,37(5):44-49.

[18] 夏 霖.BBS 中组织拓扑结构研究和意见领袖识别[D].武汉:华中科技大学, 2011.

[19] Hon Wai Lam,Chen Wu.Finding influential eBay buyers for viral marketing -A conceptual model of BuyerRank[C].Proceedings of IEEE Conference on Commerce and Enterprise Computing. IEEE,2009:778-785.

[20] 丁雪峰, 刘嘉勇, 吴 越,等.基于 SNA 的网络舆论意见领袖识别研究[J].高技术通讯,2011, 21(2):167-172.

[21] 朱义生.基于 SNA 面向特定主题的意见领袖发现研究[D].合肥:合肥工业大学, 2012.

[22] Zhang J , Ackerman M, Adamic L.Expertise networks in online communities:structure and algorithms[C].Proceeding of the 16th Conference on World Wide Web, 2007:221-230.

[23] 宁连举, 万志超. 基于团购商品评论的网络意见领袖识别[J]. 情报杂志, 2013, 32(8):204-206.

[24] 肖 宇, 许 炜, 夏 霖.一种基于情感倾向分析的网络团体意见领袖识别算法[J].计算机科学, 2012, 39(2):34-37.

[25] 王来华.舆情研究概论:理论、方法和现实热点[M].天津:天津社会科学院出版社,2003:5-8.

[26] http://docs.seleniumhq.org/projects/webdriver/.

[27] Sznajd-Weron K, Sznajd J.Opinion evolution in closed community[J].International Journal of Modern Physics C, 2000, 11(6): 1157-1165.

[28] 林聚任.社会网络分析：理论、方法与应用[M].北京:北京师范大学出版社,2010.

[29] 侯万友.群体性突发事件微博舆情演化分析[D].哈尔滨:哈尔滨工业大学,2013.

[30] 熊 涛, 何 跃. 微博转发网络中意见领袖的识别与分析[J]. 现代图书情报技术, 2013(6):55-62.

[31] 李稚楹, 杨 武, 谢治军.PageRank 算法研究综述[J].计算机科学, 2011, 38(B10):185-188.

[作者简介]王曰芬,女,1963 年生,南京理工大学经济管理学院信息管理系教授,博士生导师。

杭伟梁,男,1991 年生,南京理工大学经济管理学院信息管理系硕士研究生。

丁 洁,女,1990 年生,南京理工大学经济管理学院信息管理系硕士研究生。

收稿日期：2016-01-15