Science Discovery
2017; 5(7): 509-514
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20170507.16
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

Study on the Monitoring and Guidance of Public Opinion of Micro-Blog Based on Complex Network in Colleges and Universities

Gao Weimin*, Chang Yunjie

Computer and Information Science Department, Hunan Institute of Technology, Hengyang, China

Email address:
gwmhy@163.com (Gao Weimin), 305182059@qq.com (Chang Yunjie)
*Corresponding author

To cite this article:
Gao Weimin, Chang Yunjie. Study on the Monitoring and Guidance of Public Opinion of Micro-Blog Based on Complex Network in Colleges and Universities. *Science Discovery*. Vol. 5, No. 7, 2017, pp. 509-514. doi: 10.11648/j.sd.20170507.16

Received: November 3, 2017; **Accepted**: November 14, 2017; **Published**: December 28, 2017

Abstract: With the advent of the information age, the Internet has become the most important platform for people to express their opinions and attitudes toward public affairs or hot emergencies. Micro-blog public opinion has become the most influential network of public opinion of the means of transmission, with a short time, a wide range of communication features. Based on the analysis of complex network characteristics and network public opinion, this paper analyzes the complex network topological properties of complex network using the statistical properties of complex networks, the number of nodes, the point intensity and the average weighted clustering coefficient. Choosing "heat" and "situation" "And other indicators and node number, degree distribution and average weighted clustering coefficient of these three complex network parameters to achieve one-to-one mapping, and network life cycle of public opinion four-stage characteristics of building an emergency network public opinion monitoring and guidance model, public opinion real-time Monitoring, to determine the appropriate public opinion to guide the strategy. The actual public opinion monitoring and guidance through actual cases and real data verify the feasibility of the network public opinion monitoring and guidance model.

Keywords: University Micro-Blog Public Opinion, Complex Network, Monitoring and Guidance

基于复杂网络的高校微博舆情监测与引导研究

高为民*，常赟杰

计算机与信息科学学院，湖南工学院，衡阳，中国

邮箱
gwmhy@163.com（高为民），305182059@qq.com（常赟杰）

摘要：随着信息时代的到来，互联网已成为民众对公共事务或热点突发事件发表意见、态度的最重要平台。微博舆情已成为网络舆情中最具影响力的传播途径，具有时间短、传播范围广的特点。本文在对复杂网络特征和网络舆情分析的基础上，利用复杂网络的统计特性，分析其节点数、点强度和平均加权集聚系数的复杂网络拓扑性质；选择跟网络舆情有关“热度”和“态势”等指标与节点数、度分布和平均加权集聚系数这三个复杂网络参数实现一一映射，并结合网络舆情生命周期四阶段特征，构建突发事件网络舆情监测与引导模型，对舆情进行实时监测，确定适宜的舆情引导策略。通过实际案例和真实数据进行实际的舆情监测与引导，验证了该网络舆情监测与引导模型的可行性。

关键词：高校微博舆情，复杂网络，监控与引导

1. 引言

中国互联网信息中心（CNNIC）第39次《中国互联网发展状况统计报告》显示，截至2016年12月，中国网民规模达到了7.31亿，互联网普及率为53.2%，手机网民规模达到6.95亿，增长率连续三年超过10%。对于大多数网民而言，互联网已经成为他们获取信息的主要渠道，随着微信、微博和QQ即时通信软件的发展，人们从互联网上获取和发布信息变得非常简单，因而存在网络舆情的不可控性。如果不能在舆情前期准确获取和发现舆情，并及时地控制舆情的导向，那么就会使舆情发展偏离轨道，出现淆乱视听、颠覆黑白、以讹传讹等恶意误导情况，造成恶劣的社会影响。文献[1]从系统结构决定系统功能的角度,利用复杂网络的研究方法对网络舆情的传播规律进行了实证分析和理论研究；文献[2]通过正反案例深入分析，探索高校网络舆情引导有效方本，为高校网络舆情引导研究提供考。采取大学生实时连线、社所连线、思想连线开展“从高校BBS观察大学生社会主义核心价值观”实证研究，从学生思想脉动上，完善舆情引导机制；文献[3]根据复杂网络的理论分析，综合运用管理、计算机等多学科知识，根据观点动力学和传播动力学建模范式，揭示网络舆情动态演进的影响机理；在此基础上，分别从网络成员关系和网络结构等两个维度，深入分析基于复杂网络的网络舆情动态演进机制。文献[4]利用复杂网络方法对突发事件网络舆情进行分析，重点分析事件发生后，公众对突发事件的舆情反馈，并结合“生命周期理论”，构建突发事件网络舆情监测与引导模型，然后，对公众反馈的网络数据进行监测，以确定适宜的舆情引导策略。文献[10]在分析自媒体时代高校网络舆情危机内涵及特点的基础上，认为在应对网络舆情危机时，应把握以人为本、责任担当、快速反应、信息公开、统一口径的基本原则，同时注重预防、强化网络舆情管理、建立健全舆论引导机制以及完善善后恢复机制。

庞大的大学生微博、微信用户群体，让微博微信成为校园网舆情发展和舆情监控的重点所在。在这种背景下，研究高校网络舆情的监测与引导是非常有必要的。

2．复杂网络的统计特性

根据复杂网络理论研究，度是用于描述复杂网络整体系统中每一节点总体网络连接状态的统计学特性，它能够在一定程度上体现复杂网络的总体演进过程及其演进特征。

（1）节点数

用矩阵法表示的复杂网络，是由一个节点集合和一个邻接矩阵分别表示节点和边。节点集合为V={V1, V2,... Vn}，其中，n的值表示节点数。在突发事件时空复杂网络中，节点代表的是评论信息中的空间(即:网民评论此条新闻时所在的地理位置)，因此节点数的多少也即不同的地理位置的多少。

（2）度分布

一个节点度分布在复杂网络中，度是一个非常重要的统计概念，它是指与一个节点i连接的边的总数，体现了该节点i的重要程度，见公式1。

$$k_i = \sum_j a_{ij} = \sum_j a_{ji} \quad (1)$$

复杂网络中节点度的平均值和最大值分别用公式2和公式3表示。

$$\bar{k} = k_i = \frac{1}{N}\sum_i k_i = \frac{1}{N}\sum_{ij} a_{ij} \quad (2)$$

$$k_{\max} = \max_i k_i \quad (3)$$

（3）度的相关性

假设任意节点i和节点j的度分别为k和k’，节点i和节点j连接的概率为p(k’,k)，则节点i和节点j相连接时的条件概率为:

$$p(k'|k) = \frac{\bar{k}p(k,k')}{kp(k)} \quad (4)$$

它表示某个度值为K的节点与另一个值为K’的节点相连接时的条件概率。其中p(k’,k)表示两个度值分别为K和K’的节点相连接的综合概率。

定义：

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) \quad (5)$$

3．网络舆情传播基本原理

近年来学者们纷纷将SIR模型应用到复杂网络中，本文在Kermack与McKendrick于1927年建立了经典的传染病SIR模型基础上，综合考虑部分未接收到舆情信息的网民在接收信息后不参与传播和部分接收信息但不传播的网民在舆情衍生话题的影响下变为传播者这两种情况，建立突发事件网络舆情演化传播的SIR（Susceptible-Infected-Recovered）模型。[5, 6]

网络舆情发生时，基于经典SIR模型，S也可能以概率d直接变为R;当原始网络舆情衍生新话题时，R会因为衍生话题舆情信息选择传播话题或制造新话题变为I的概率为。易感染者S,感染者I和移出者R的转换关系如图1所示。

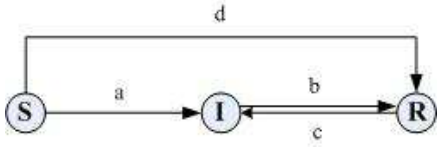

图1 改进SIR模型。

改进后的SIR模型，网络舆情传播模型的具体形式为：

$$\frac{ds}{dt} = -\beta s(t)I(t)$$
$$\frac{dI}{dt} = \beta s(t)I(t) - \gamma I(t)$$
$$\frac{dR}{dt} = \gamma I(t)$$
$$s(t) + I(t) + R(t) = N(t)$$

其中：I(t)表示t时刻仍具有感染性的主机数；R(t)表示t时刻已经从被感染的机器中免疫的主机数；S(t)表示到t时刻所有被感染过的主机数；N(t)是总人数。a为感染率，b为衰退率，c衍生影响率，d为转化率。本文舆情传播模型的易感染者S为未接收到舆情信息的用户，感染者I为接收到舆情信息并传播的用户，移出者R为接收到舆情但不传播或未接收到舆情且不传播的用户。

目前，针对网络舆情预测问题，研究方法可分为基于数理统计模型的预测方法、基于概率计算的预测方法与启发式预测方法。

4．网络舆情监测系统的关键技术

4.1．网络爬虫链接去重技术

链接(URL)去重，指的就是在网络爬虫的操作中，不去爬行那些在之前已经被爬行过的页面。舆情系统对实时性提出了很高的要求，而网络爬虫是系统中一个比较耗时的操作，因而必须要对其进行优化。本文引用了BF（Bloom Filter）算法，[7]它的基本思想是：在判断一个元素x是否在集合中的时候，一个哈希函数会导致大量冲突，但若能多引入几个hash函数，那么这些哈希函数同时冲突的概率将会大大降低。BF算法有一个m位的位数组和K个独立的哈希函数，初始时位数组中每一位都为0。当要增加一个元素x时，先通过K个哈希函数，得到K个哈希值，然后将位数组中这K个位置的元素置为1。如下图2所示。

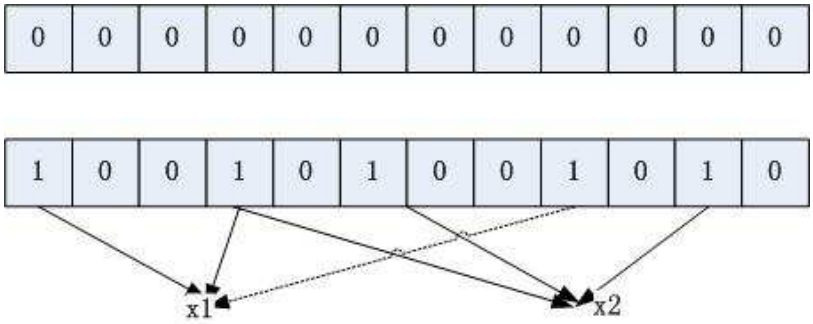

图2 BF算法。

4.2．信息抽取技术

网页信息抽取技术通过对网络舆情网站的页面进行处理，从自然语言中抽取出预定好的实体、关系、事件的集合，并用结构化的表示来记录并保存到数据库，方便用户获取和利用这些信息。它的关键是保证算法的准确性和健壮性。[8, 9, 11]本文提出的网络信息自动抽取算法主要包括URL模板过滤网页、网页信息结构化、网页解析模板匹配和数据库存储等。算法流程如图3所示。

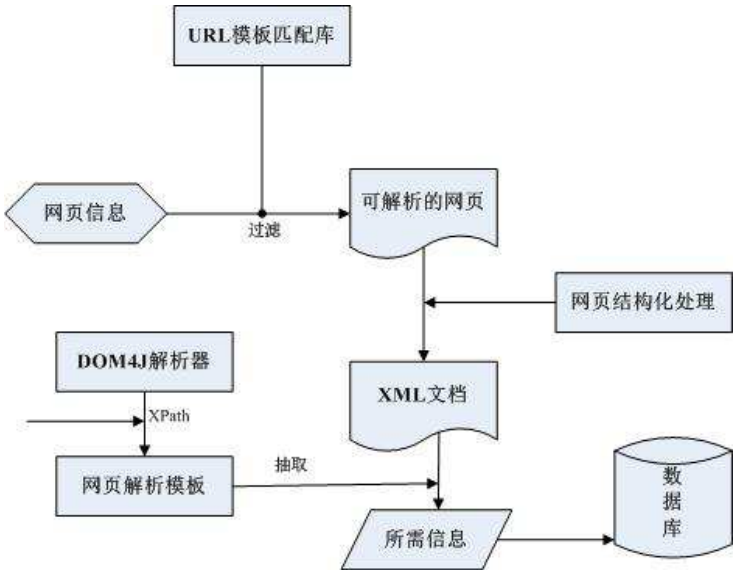

图3 网络信息自动抽取算法流程。

网页信息自动抽取首先通过URL模板匹配过滤出可以解析的网页，然后将可解析的Html文档进行网页结构化处理，生成xml文档。最后结合DOM4J和Xpath语言建立页面解析模板，从xml文档中抽取指定节点信息，并将其存储进入数据库。

5．复杂网络的舆情监测与引导模型

图4 Web数据挖掘过程。

对网络舆情的监测和引导的过程，从本质上来说即是对Web数据的挖掘过程。即从网页监测数据中搜集和发现有用的知识的过程。Web数据挖掘的过程分为四个阶段:数据收集、数据预处理、模式发现和模式分析。具体如图4所示。

数据收集。数据是一切分析的源头，搜集到的数据的质量直接决定了网络舆情监测的质量。通过一定的搜索引擎或其他计算机数据抓取方法，将HTML格式或XML格式的网页信息进行搜集，以获取目标数据，本文所取数据即是具有用户评论的Web网页。

预处理。原始的Web网页信息，包含了太多冗余、不完整信息，比如：Html标签等。需要对其进行数据清洗，进行数据分解、合并和格式转换等，以供进一步的数据分析。

模式发现。利用各种数据分析、挖掘算法对处理后的数据进行挖掘，传统的分析、挖掘算法有统计分析、关联规则、序列分析、路径分析、聚类和偏差分析等等，在此基础上获取各种模式，突发事件网络舆情的数据分析，主要运用复杂网络分析方法。

模式分析。模式分析是指从已经发掘到的各种模式中，筛选出有用的、有意义的模式，以满足客户的特定需求，本文发掘的模式主要是用于舆情引导。

根据Web数据挖掘过程，我们建立基于复杂网络的网络舆情监测与引导模型，如图5所示。

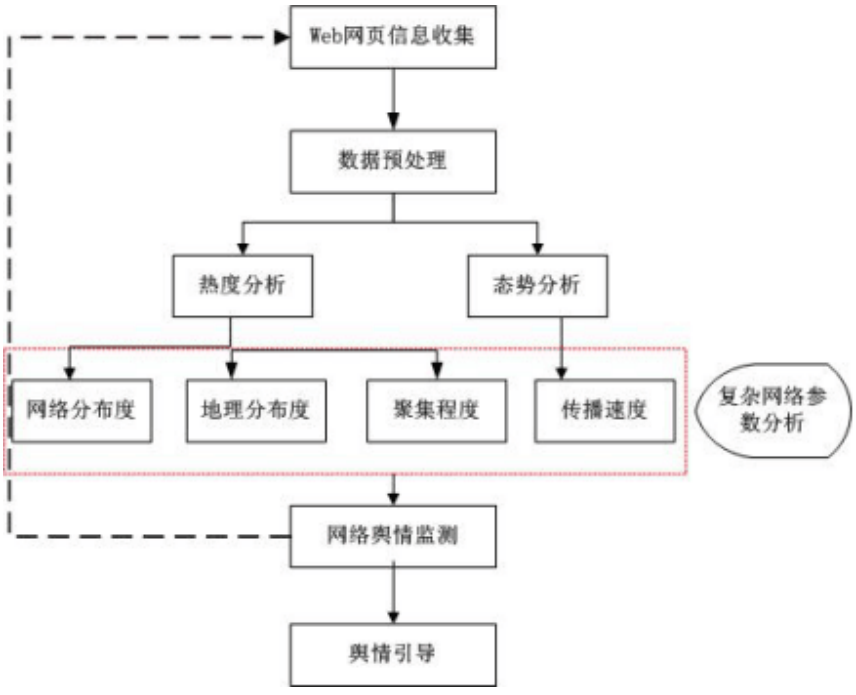

图5 网络舆情监测与引导模型。

从图5中可以看出，突发事件网络舆情监测与引导模型大体由四部分组成：

(1) Web网页信息收集。根据所要监测的舆情主题，选择合适网站，确定所要抓取的网页范围，并将抓取的数据存储到数据库中；

(2) 数据预处理。将获取到的原始数据，进行数据清洗。对于新闻评论来说，需要过滤无关的信息，保留供后续分析所用的有用信息，比如:新闻标题、发布时间、新闻内容、评论人、评论时间、评论地点、评论数量等；

(3) 舆情数据分析部分。根据高校管理者和研究文献对网络舆情指标的选取结果来看，重点关注的是舆情的“热度”和“态势”，在本文中，主要关注的是具有时间和空间属性的舆情发展情况，因此，所选取的

指标更侧重“时空”属性。用“网络分布度”、“地理分布度”和“聚集程度”三个指标测量舆情“热度”，同时，用舆情的“传播速度”测量舆情“态势”。而对这三个指标“地区关注度”、地区的“关注度”和“传播速度”的定量化测度需要用到复杂网络的参数分析，同时结合生命周期理论论述的舆情的阶段性特点，以确定所分析和监测的舆情数据大致处在生命周期的哪个阶段，在此过程中，重点关注的是舆情处于快速增长期到成熟期这两个阶段；

(4) 舆情引导。在前述舆情分析和监测的基础上，确定舆情引导的时机及可行的舆情引导策略，化解可能产生的网络舆情危机。

6．案例研究及性能评估

6.1．案例设计

本文以高校微博舆情的监测与引导为例，通过登录新浪微博，采用爬虫软件爬取所需要的数据，主要从以下几个方面进行分析：

（1）关键字出现频率

对于数据库中的关键字进行读取，对比同一时间段内全国高校随机抽取的2所高校汇总关键字出现的频率，得出该时间段内高校重点关注的舆情情况以及预测该时间段内高校折线图的走势。

（2）南北关键字出现频率

通过对比抽取的南北高校中出现的高校微博关键字汇总，对比同一时间段内南北高校中是否因地域不同而出现舆情情况不同，进而得出该时间段内南北高校舆情内容的重视以及范围的热度走势。

（3）东中西关键字出现频率

通过对比抽取的东中西部高校微博中出现的关键字汇总，可以得出东中西部高校在同一时间段对于关键字内容的重视以及范围的热度走势，从而判断东中西舆情侧重点趋势。

（4）微博发布来源

通过汇总全国高校微博中随机抽取的两所高校提取微博关键字数据，得到高校总体微博发布来源，进而判断使用微博发布是移动端占据优势还是PC端占据优势。

（5）点赞，评论数，转发数

通过汇总全国高校微博中随机抽取的两所高校提取出微博关键字数据，总体关键字汇总中的点赞数以及总体评论数，转发数，可以得到关键字微博内容产生的舆论意义，以及传播范围影响。

6.2．测试结果

案例测试数据主要来源于新浪微博，抽取各高校学生比较感兴趣的10大关键词如：工作、学习、面试、科研、创新等，统计出关键词出现的频率、微博发布来源、舆情地理分布、热度等，再通过可视化方式展示出来。测试结果如下图6所示。

图6 关键字汇总趋势。

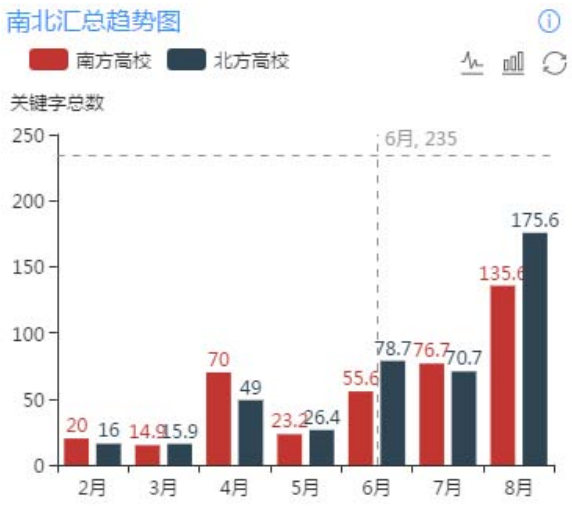

图7 南北高校关键字分类汇总。

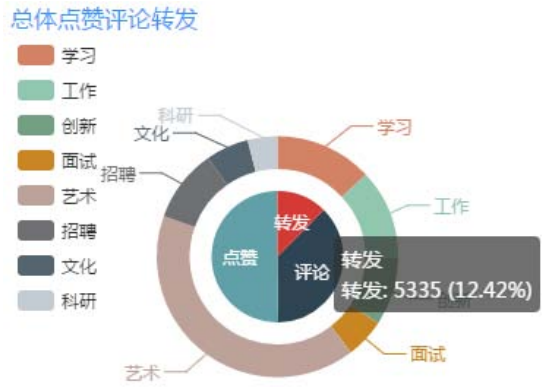

图8 总体点赞及评论转发测试。

6.3．舆情预警测试

以“比特币勒索病毒”为搜索关键词，以每两个小时为搜索区间，抓取新浪微博上2017-05-12—2017-05-21十天

的微博，针对每条微博爬取微博发布者的昵称、发布时间、转发数、评论数、赞数等。每天汇总的微博数量可反映舆情传播中传播者的舆情传播水平。

表1 “比特币勒索病毒”事件预测结果。

日期	实际值	预测值	误差率	日期	实际值	预测值	误差率
5-12	36	41	13.9%	5-17	47	60	27.6%
5-13	132	135	2.3%	5-18	35	37	5.7%
5-14	321	360	12.1%	5-19	23	28	21.7%
5-15	235	254	8.1%	5-20	20	25	25%
5-16	134	141	5.2%	5-21	17	20	17.6%

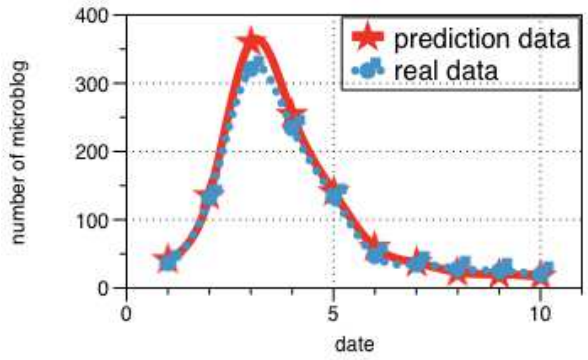

图9 预测结果对比分析。

由上述仿真结果可以看出，利用基于舆情传播模型的信息自动抽取算法，结合实际网络中的高校舆情大数据，可挖掘感染率、衰退率等舆情传播模型参数，预测该事件后续的舆情传播过程。仿真预测结果与实际数据变化趋势基本吻合，拟合效果较好，说明可以用基于复杂网络舆情传播模型来仿真模拟高校微博舆情的监测。

7．结论

突发事件由于其瞬间性、偶然性、危机性和危害性等特点，一旦爆发，如果没能及时监测跟引导，将有可能使舆情急剧扩散、放大，导致网络舆情危机的产生。本文力求用复杂网络方法对高校中某一热点事件的新闻评论进行分析，结合网络舆情生命周期四个阶段的划分一一潜伏期、成长期、成熟期、衰退期，构建突发高校网络舆情监测与引导模型，确定舆情引导时机以及相关引导策略。最终以“新浪微博舆情监测”为实例，验证了模型的有效性。

致谢

本文为教育部人文社科青年基金项目《手机媒体视角下高校微博舆情的监控与引导研究》(15YJC880017)的阶段性成果之一；湖南工学院“三个一批”人才支持项目资助。

参考文献

[1] 潘新. 基于复杂网络的舆情传播模型研究[D]. 大连：大连理工大学博士论文，2010。

[2] 赵杨. 高校网络舆情引导研究[D]. 长春：东北师范大学，2017。

[3] 董靖巍. 基于复杂网络的网络舆情动态演进影响机制研究[D]. 哈尔滨：哈尔滨工业大学，2016。

[4] 吕娜. 基于复杂网络的突发事件网络舆情监测与引导模型研究[D]. 北京：中国地质大学，2014。

[5] 李国佳，高校微博平台的网络舆情引导研究[D]. 郑州：华北水利水电大学，2015。

[6] 王国华，冯伟，王雅蕾. 基于网络舆情分类的舆情应对研究[J]. 情报杂志，2013, 32(5):1-4。

[7] 曾振东. 基于灰色支持向量机的网络舆情预测模[J]. 计算机应用与软件，2014, 02:300-302+311。

[8] 张华. 基于优化BP神经网络的微博舆情预测模型研究[D]. 华中师范大学，2014。

[9] Mitzenmacher, Michael. Compressed bloom filters [J]. IEEE/ACM Transactions 2003, 10(5):604-612.

[10] 王柳梅. 自媒体时代高校网络舆情危机管理研究[D]. 长春：长春工业大学，2017。

[11] 王杨，尤科本，王梦瑶，等. 基于博弈论的网络社区舆情传播模型[J].计算机应用研究，2013, 30(8):2480-2482。