doi:10.3969/j.issn.1000-436x.2013.11.019

面向话题的微博网络测量研究

刘玮[1,2,3]，王丽宏[3]，李锐光[3]
（1.中国科学院 计算技术研究所，北京 100190；2. 中国科学院大学，北京 100049；
3. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要：针对话题生成网络的动态时序特性，设计定量计算方法，从微博内容、网络结构、用户行为角度开展面向话题的新浪微博网络测量研究，结果发现：少数微博被大量转发，转发次数与对应微博数呈现近似的幂率分布；话题热度呈现明显的突发性和变化趋势，局部波动率能够有效地在大量背景微博中发现突发话题；基于话题生成的转发网络的小世界特性并不明显，且密集的关注关系不一定引发频繁的转发行为；传播能力强的话题中含有较大比例的持续参与用户，用户行为的话题相关性能够有效检测潜在关键用户。测量结果有助于了解话题生成网络的内容传播特点、网络结构特性及用户行为模式，测量指标能够有效应用于微博话题影响力分析等相关研究。
关键词：内容特征；网络结构；用户行为；小世界特性；局部波动率；影响力分析
中图分类号：TP393.08 **文献标识码：**A **文章编号：**1000-436X(2013)11-0171-08

Topic-oriented measurement of microblogging network

LIU Wei[1,2,3], WANG Li-hong[3], LI Rui-guang[3]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Graduate University of Chinese Academy of Science, Beijing 100049, China;
3. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: According to the dynamic and temporal characteristics of the topic-generated network, a method of quantitative calculation was designed, and then the topic-oriented research on the network measurement technology from many aspects such as the features of the content was conducted, as well as the network topology and the characteristics of the user behavior. The experiments on the SINA microblog showed four new results. The first is that only a small portion of tweets has been forwarded broadly and the number of retweets follows the power-law distribution. The second is that the tweets' number of one topic is episodic and changing frequently, and the burst topic can be detected by the local volatility feature found in the massive background microblog data. The third is that the small-world feature in the topic-generated retweeting network is not obvious, and the dense relationship doesn't necessarily induce the frequent retweeting behavior. The fourth is that the topic which has been propagated broadly usually has a portion of the consistently participating users, and the correlation of the user behavior can be used to detect the potential and important users. The experimental results are helpful for understanding the propagating mode, the structural chara and the pattern of the user behavior in a topic-generated network, and the indicators measured in the experiment can also be effectively applied in the future analyses.
Key words: content features; network topology; user behavior; small-world property; local volatility; influence analysis

1 引言

我国微博应用于 2009 年正式发布，经过 3 年的蓬勃发展，迅速以其内容简洁、交互便捷和快速传播等特点，发展成为人们表达观点、抒发情绪、传递信息的重要社会媒体。截至 2012 年 12 月底，我国微博用户规模为 3.09 亿[1]。微博 140 字的内容限制、“//@、#”符号标记语言以及单向认证的“弱

收稿日期：2013-05-02；**修回日期：**2013-09-03
基金项目：国家自然科学基金资助项目（61170230）；国家科技支撑计划基金资助项目（2012BAH46B01）
Foundation Items: The National Natural Science Foundation of China(61170230); The National Key Technology R&D Program (2012BAH46B01)

关系”等机制，使得微博在内容生成方式、用户参与的广泛性和即时性、信息扩散模式和速度等方面均不同于新闻、论坛、博客等传播网络媒体，微博能够更真实、全面、快捷地表达现实世界的社会关系和社会生活。在“黄岩岛事件”、“北京特大暴雨灾害”等2012年典型互联网舆情事件的发酵和爆发过程中，微博都起到了重要的推动甚至导向性作用，是网络舆情分析的重要观测对象。特别是手机微博用户规模的大幅提高[1]，使微博表现出了强大的信息即时分享和井喷式扩散能力[2]，给热点敏感话题的快速发现、话题影响力的评估预测以及话题传播的及时引导和有效处置等提出了严峻挑战。

开展网络测量是分析微博网络特性、研究微博网络拓扑结构和信息传播机制的重要方法。网络测量是指按照某种规律，用数据表示观测现象，对微博网络结构或信息传播规律特性进行量化描述。微博网络可以形式化表示为由大量节点和连边组成的大规模网络，复杂程度位于规则与随机之间，因此，复杂网络相关理论和研究方法被广泛应用于解决微博网络拓扑结构分析和社区发现等问题。Kwak 等人[3]针对Twitter数据开展的网络测量研究发现Twitter网络中互粉比例低、转发层数少、被转发率高的用户重要程度高等特性，Chew 等人[4]通过测量2009年H1N1爆发期间Twitter采样数据，验证了微博能够反映公众关注热点、关键内容及情感，Ren 等人[5]通过Twitter用户兴趣研究，发现用户兴趣会随时间发生变化，Patil 等人[6]研究了游戏网络和合作网络中的社区演变和消亡现象及预测方法，Grabowicz 等人[7]研究了Flickr网站中社区形成的原因，并提出了基于兴趣形成和基于社会关系形成的社交网络识别方法，Java等人[8]通过测量Twitter网络数据的拓扑结构和用户地理位置属性，发现社区内的用户兴趣相似，樊鹏翼等人[9]基于新浪微博采集数据开展的测量分析表明，新浪微博具有小世界特性、网络出入度不具有相关性、用户转发微博多于回复。Kempe 等人[10~13]研究了社交关系网络、节点度分布、最短路径、社区结构等网络结构特性，并用于分析社交网络影响力最大化问题，Barbieri 等人[14]在传统基于节点度的节点重要度计算方法中引入话题内容，但没有考虑用户行为特征。

上述研究都是针对整体微博网络开展，没有考虑微博网络的话题相关性和动态性，而在实际的微博话题相关分析中，对话题生成的关注网络和转发网络相关特性的研究具有重要意义。其次，现有微博网络测量研究大多集中在节点度分布、最短路径等网络拓扑结构特性分析，以及以24 h作息时间为周期、以7天自然周为周期的宏观行为模式等方面，没有充分考虑微博作为话题传播载体所具有的动态时序特性和突发特性，即微博话题的传播规律特性和用户行为的话题相关性。

针对上述问题，本文面向话题开展微博网络测量研究，构建基于话题生成的关注网络和转发网络，从突发话题传播规律、网络拓扑结构、用户行为的话题相关性3个维度分析了话题生成网络的动态时序特性，设计了微博倾向性热度及传播规律、微博话题的突发特性、基于话题的微博网络结构特性、用户转发行为规律，以及用户兴趣和行为的话题相关性等测量方法。提出了话题热度局部波动率、用户转发惯性、用户兴趣的话题相似度、用户行为的话题相关性等定量测量指标。经过在真实新浪微博数据上的测量研究和现象分析表明，所提出的方法和指标能够有效测量文中微博话题传播的内容特点、话题生成网络的结构特性及用户行为的话题相关性等，能够较好地应用于微博突发话题检测、话题影响力分析、关键人物发现等相关研究。

2 测量对象与数据描述

本文选取新浪微博作为目标观测网络，考虑到微博中含有大量账号粉丝数少、活跃度低，容易造成网络过于庞大和稀疏，为了提高网络连通度，选取新浪微博中粉丝数最多的70万账号，采用“滚雪球”策略，采集了这些账号截至2012年8月8日所发的微博消息、Profile信息、关注关系等信息，以及被这些账号转发过消息的账号和对应信息，经过垃圾微博去除等预处理过程后，构建了包括600万个微博账号、约12亿条微博消息的基本观测网络。

本文研究微博话题生成网络，设计定量计算方法，通过测量结果的描述和分析，发现其内容传播规律、网络拓扑结构、用户行为的话题相关性等特性，进而表明测量方法和指标的有效性。因此，首先构建微博话题生成网络，确定研究对象。本文选取2012年具有代表性的互联网舆情事件——“黄岩岛事件”作为本文的观测话题对象，开展面向话题的微博网络测量。通过话题标签、关键词提取，从基本观测网络中抽取参与2012年“黄岩岛事件”话题的微博账号及个人信息、关注关系、相关微博，最终抽取得到的话题数据集描述如表1所示。

表 1　面向话题的微博数据集

账号数/个	关注关系/条	微博数/条	时间范围
26 755	733 606	59 992	2012.4.11-8.8

通过对话题数据集中的微博热度统计分析，绘制出话题传播热度趋势（如图 1 所示），其中深色曲线表示参与话题的用户在 2012 年 4 月 11 日到 2012 年 8 月 8 日期间发表的背景微博的热度统计趋势图，浅色曲线表示话题相关微博的热度统计趋势图（为了便于对比展示，本文对热度值、节点数等统计量进行了归一化处理）。从图 1 可以看出，整体微博热度随时间呈现较为规律的周期性，每周的星期一到星期五微博热度较高，星期六和星期天微博热度较低，这与人们微博行为经常发生在工作日的行为规律相符，而话题热度呈现明显的突发性和变化趋势，未观测到周期性特征。另一方面，通过与同一时间段内该话题在“百度指数”的用户关注度趋势图（如图 2 所示）的对比分析，可以表明本文所构建的话题数据集能够反映实际的话题传播情况。

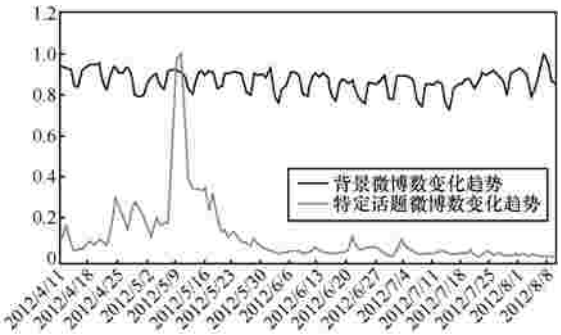


图 1　话题微博热度变化趋势

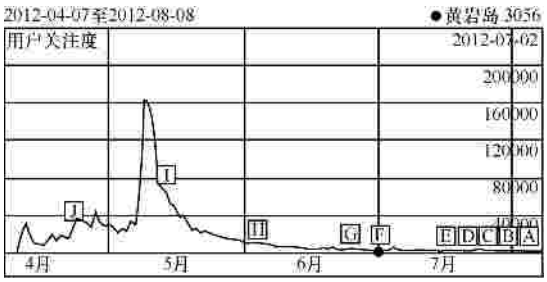


图 2　“百度指数”的用户关注度趋势图

3　面向话题的新浪微博内容测量及分析

3.1　微博倾向性热度及传播规律测量

为了观测不同倾向性微博消息的热度、被转发概率等传播规律，本文利用知网倾向性词典[15]、微博消息符号标记等特征，对“黄岩岛事件”期间的相关微博消息进行了倾向性分类，并分别统计了不同倾向性微博的消息数、被转发数，进而计算出其被转发率。由于微博内容长度限制及情感表达较为鲜明的特点，本文通过计算微博消息中所包含正负倾向性词和情感符号标记的数量来计算微博倾向性，为了转发数统计结果的准确性，本文采用微博消息数据中表示转发次数的字段值来计算该条微博的实际转发次数。测量结果如表 2 所示，在“黄岩岛事件中”，负倾向性微博数是正倾向性微博数的 4 倍以上，平均每条负倾向性微博被转发的次数为 42.78 次，是正倾向性微博的 2 倍以上。这表明在“黄岩岛事件”的相关微博中，负倾向性微博数量占较大比例，且更易于得到转发。这是因为负倾向性微博往往携带了更多新的衍生信息，情感表达方式也较为激烈，易于使得话题得到传递、发展和演化。因此，针对微博话题，赋予消息内容一定倾向性，有助于提高其传播影响力。

表 2　正负倾向性微博热度及被转发次数

倾向性	微博数/条	转发次数/次	平均转发次数/次
正面	11 145	217 685	19.53
负面	46 117	1 972 946	42.78
中立	2 730	20 899	7.66

那么，是否所有负倾向性微博都能够得到同样的高转发率呢？针对这个问题，对微博被转发次数及对应的微博数进行了测量分析，结果如图 3 所示，被转发次数与对应的微博数呈现近似的幂率分布，表明微博被转发的能力极不均匀，转发次数最多的前 341 条微博所达到的累积转发达到 155 098 次之多，是平均被转发次数的 10 倍。这种少量微博得到大规模转发的现象，使得快速定位不良信息并有效控制信息的扩散成为可能。

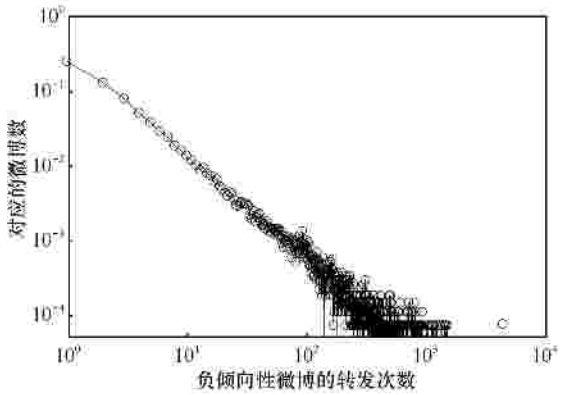


图 3　负倾向性微博被转发次数分布曲线

3.2 微博话题的突发特性测量

在图 1 所示的话题微博热度变化趋势曲线中，可以观测到话题微博的热度趋势与当前背景微博的热度趋势具有显著差异。那么，如何测量这种差异性，这种差异性能否量化并作为检测微博话题突发特性及其变化趋势的有益指标？如果直接采用热度值作为测量标准，会遇到 2 个问题，首先不同话题所引发的关注程度不同，即使是相同话题在不同时期所呈现的热度也不尽相同，阈值设定过高会导致突发话题的漏检率高，而阈值设定过低又将导致突发话题的误报率升高。其次，相比于海量背景微博消息，话题热度值的区分度并不明显（本文数据集的话题观测周期内，背景微博高达 2 亿条、而话题相关微博约 6 万条），所以仅凭经验判断阈值的方法难以在实际中应用。

根据观测结果，背景微博热度随时间呈现较为规律的周期性，话题热度呈现明显的突发性和变化趋势，本文提出波动率指标 s_t^S 来测量微博话题的突发特性，计算方法如式(1)所示。

$$s^S(t)=\sqrt{\sum_{i=1}^{t}\left(N_i^S-\hat{N}^S\right)^2\Big/t-1} \tag{1}$$

其中，S 表示所属话题，t 表示话题天数，N_i^S 表示距离观测起始时间第 i 天的与话题 S 相关的微博消息数，$\hat{N}^S$ 表示 N_i^S 的平均值。根据热度波动率指标分别对背景微博和话题微博的热度趋势进行了统计，结果如图 4 所示，背景微博的热度波动率稳定在 0~0.07 范围内，因混合了多个话题的综合热度变化效应，而使得整体变化趋势较平稳。而话题微博的热度波动率在 0~0.35 范围内变化，在 5 月 10 日左右出现了陡升现象，表示话题热度在该时间前急速增长。然而在高峰期后，由于热度的累积效应，该算法对局部热度的变化程度反应不敏感。

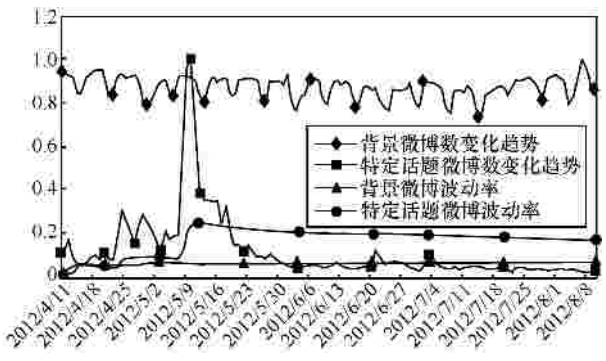


图 4 话题微博热度累积波动率

因此，本文在波动率指标基础上提出基于时间窗的局部波动率指标 $s_t^S(\Delta T)$，ΔT 表示计算波动率的时间窗口大小，$s_t^S(\Delta T)$ 的计算方法如式(2)所示。

$$s_t^S(\Delta T)=\sqrt{\sum_{i=t-\Delta T}^{t}\left(N_i^S-\hat{N}^S\right)^2\Big/\Delta T}$$

其中，$\Delta T \geq 1, t-\Delta T \geq 1$ (2)

图 5 所展示的是 $\Delta T=7$ 时统计测量得到的背景微博热度和话题微博热度的局部波动率曲线。可以看出，局部波动率能够更准确地反应热度变化过程，能够有效检测出热度绝对值不大但局部变化较大的现象，该方法与基于热度阈值的热点话题检测方法相结合，能够更好地应用于突发话题的早期检测和话题演化过程的跟踪分析。

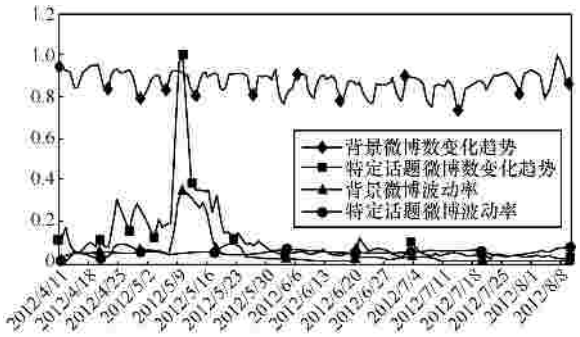


图 5 话题微博热度局部波动率

4 面向话题的新浪微博网络结构特性测量及分析

4.1 基于话题的微博网络拓扑结构特性测量

“小世界性”和“无标度性”是复杂网络的 2 个基本特性，小世界性是指实际网络中的平均节点距离比规则网络小得多，平均集群系数比随机网络高的多。无标度性是指网络中节点度概率分布呈现幂函数分布，表明节点间相互作用的能力极度不均匀。

真实社会网络几乎都具有上述复杂网络特性，所以作为真实社会网络在互联网空间的虚拟映射，社交网络也被纳入复杂网络范畴进行研究，社交网络影响力传播、链路预测、社区发现等问题建模也通常建立在复杂网络相关理论和研究方法之上，拓扑结构分析往往是认识和利用社交网络结构特性的基础性工作。

如果把这种直接以微博用户为节点、关注关系为连边所形成的相对稳定的微博网络称为基础微博网络，那么，以在某个时期内参与同一话题而聚

集的相关用户为节点、关注关系为连边所形成的话题相关的微博网络就称为话题关注网络，在此基础上，建立的以话题相关用户为节点、消息转发关系为连边所形成的基于话题和用户行为的动态微博网络就称之为话题转发网络。

针对后两类微博网络的网络结构特性的测量分析对微博网络中的话题影响力分析、链路预测、特定群体行为分析等都具有重要意义。那么，话题关注网络和话题转发网络是否也符合小世界特性？具有什么新结构特性呢？为此，首先从背景微博数据集，构建面向话题的关注网络和转发网络，结构说明如表 3 所示。

表 3　话题关注网络和话题转发网络结构描述

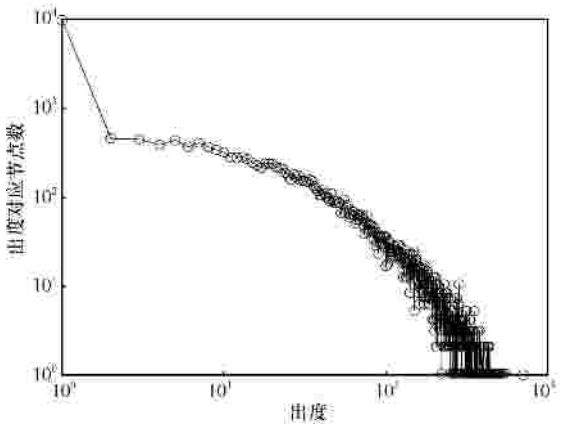
网络名称	节点数/个	连边数/条	图类型
话题关注网络	26 755	733 606	有向无权图
话题转发网络	26 755	5 781	有向带权图

话题转发网络的连边数较为稀疏，在具有关注关系的用户中，只有少量用户发生了消息转发行为。本文具体对两类网络的平均度、平均路径长度、网络直径、聚集系数指标开展了测量。平均度指所有节点度的平均值，聚集系数采用平均聚集系数计算，平均度和聚集系数是网络中节点与邻居节点的紧密程度指标；平均路径长度指所有节点对之间最短路径的平均值，网络直径指任意节点对间最短路径的最大值，是网络连通性指标。

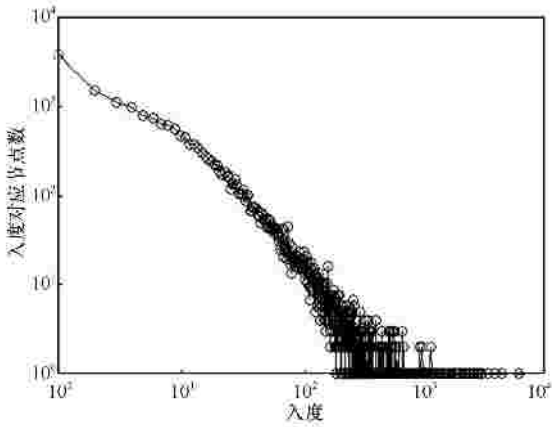
如表 4 所示，话题关注网络的平均度为 30.337，其分布曲线如图 6(a)和图 6(b)所示，节点度取某一定值的概率与节点个数服从幂率函数分布，表明话题关注网络具有无标度特性。平均路径长度为 3.664，网络直径为 13，参与该话题的微博用户较为均匀分布在基础网络中，聚集系数仅为 0.077，表明参与该话题的用户之间关注关系密集，但没有明显的社区结构，如图 7(a)所示。

表 4　话题关注网络和话题转发网络拓扑结构分析

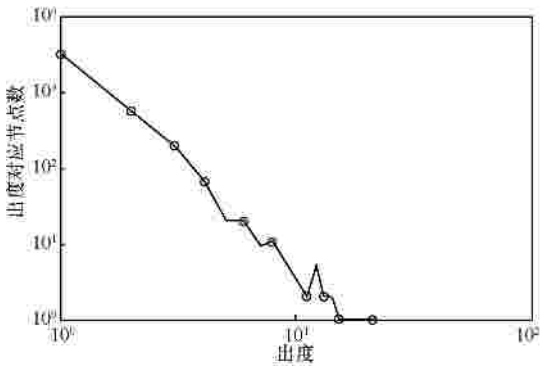
测量指标	话题关注网络	话题转发网络
平均度	30.337	0.216
网络直径	13	24
平均路径长度	3.664	8.729
聚集系数	0.077	0.003



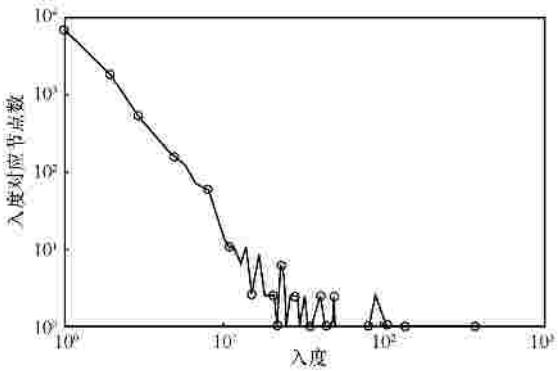
(a) 关注网络出度分布



(b) 关注网络入度分布



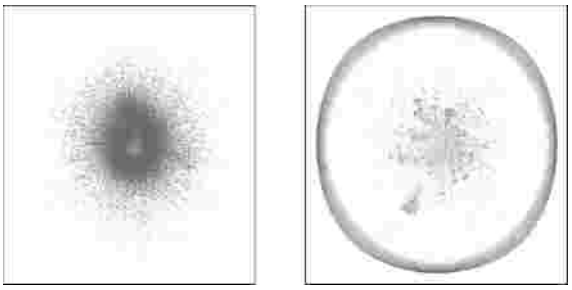
(c) 转发网络出度分布



(d) 转发网络入度分布

图 6　话题关注与话题转发网络节点出入度分布

话题转发网络的平均节点度数仅为 0.219，节点间的连边关系非常稀疏，表明密集的关注关系并不一定能引发频繁的消息转发行为。根据图 6(c)和图 6(d)所示的节点度分布曲线，话题转发网络的节点度取定值的概率与节点个数呈现近似的幂率分布。节点间平均路径长度为 8.729，网络直径为 24，表明话题转发网络存在大量离散节点，如图 7(b)所示，且聚集系数仅为 0.003，所以话题转发网络的小世界性并不明显。



(a)话题关注网络图 (b)话题转发网络图

图 7 话题关注与话题转发网络对比

4.2 用户转发惯性特性测量

为了进一步测量参与话题用户的转发关系规律，针对用户的消息转发行为是否存在惯性开展测量，用户惯性是指用户从同一用户多次转发话题相关微博的倾向程度。本文通过计算在观测期内，一个用户从同一用户转发话题相关微博的次数来表示用户惯性指标，并绘制出用户转发惯性分布图。

根据图 8 中的观测结果分析得出，在微博话题下，从同一用户转发 2 次以上微博的行为仅占所有转发行为的 13%，用户定向获取与话题相关信息的倾向性（即用户惯性）并不明显，所以，用户转发某条话题相关微博的行为具有一定随机性，这对话题影响力分析与话题引导策略研究具有一定的启发作用。

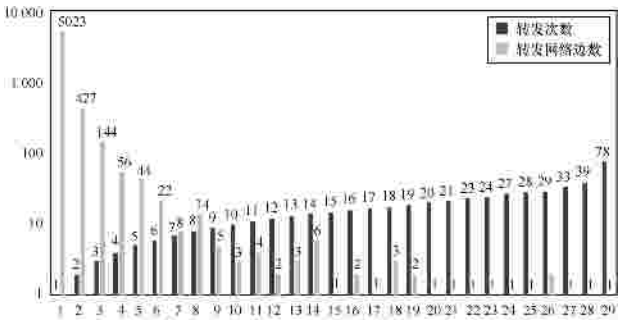


图 8 用户转发惯性

5 面向话题的新浪微博用户行为相关性测量及分析

用户行为模式研究是社交网络领域研究的重要方向，比如基于宏观微博热度数据统计得出的以 24 h 作息时间为周期的行为模式、以及以 7 天自然周为周期的行为模式等。更进一步的研究则涉及用户兴趣和偏好等微博内容分析层面，这里含有一个基本假设：用户倾向于参与到与其兴趣相似的话题中，并且参与的人数越多，话题就越热。这种方法在基于兴趣的推荐和检索应用中取得了较好效果，但在面向话题影响力分析过程中，参与用户是否都具有与话题内容相似的兴趣？除参与用户的绝对数之外，有没有能够从用户参与的过程性和话题传播的全局性角度，度量什么样的用户行为模式对话题的快速传播和持续演化具有关键性作用？针对上述问题，本节提出用户兴趣的话题相似度和用户行为的话题相关性 2 个指标，对参与用户的兴趣与话题的相似度，以及用户行为模式进行了测量和分析。

5.1 用户兴趣与话题内容的相似度测量

本节通过用户历史兴趣与话题内容的相似度来测量用户兴趣与话题内容的相似度。首先，将话题相关微博内容表示为带权重的特征词向量，然后根据特征选择算法提取该话题最具代表性的 Top10 000 个关键词。然后，根据话题用户在话题起始点前的历史微博，将每个用户的兴趣表示为布尔值特征向量。由于微博内容非常短，特征词是否出现能够更好地描述用户兴趣特征，所以，本文采用 Jaccard 距离计算每个用户兴趣与话题内容的相似度，相似度对应用户数的分布以及曲线拟合结果如图 9 所示。

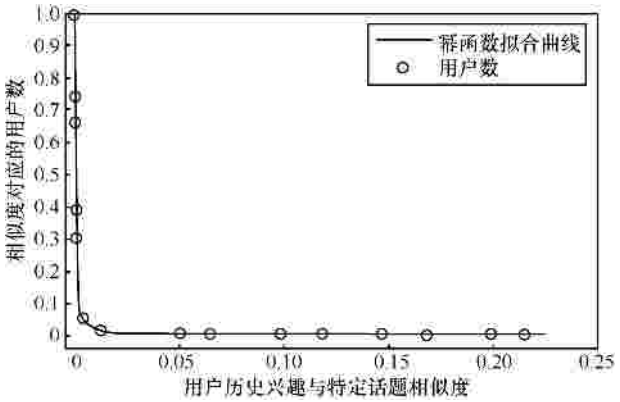


图 9 用户历史兴趣与话题的相似度

从图 9 可以看出，参与话题的用户中，历史兴趣与话题相关度集中在 0.02 至 0.06 之间，99.9%的用户历史兴趣相关度都小于 0.1。这表明，参与话题的用户并不一定具有或者是表现出与话题相关的历史兴趣。这是因为微博网络很大程度上反映的是用户某一方面或者其愿意呈现在网络上的兴趣内容，通过微博内容学习得到的用户历史兴趣难以真实、全面、即时地表示出用户的实际兴趣。该测量结果提出了运用基于内容的用户兴趣建模方法进行话题影响力分析时可能遇到的问题。

5.2　用户行为与话题传播趋势的相关性测量

微博话题传播过程中关键节点分析的实质是发现对话题快速传播和持续演化具有关键性作用的用户，常用的节点重要度计算方法主要基于节点度、PageRank 等算法，节点度方法倾向于发现粉丝数多的节点，PageRank 方法涉及反复迭代过程，在大规模微博网络中非常耗时，且要求网络连通度较好，但如 4.1 节测量结果所示，话题转发网络中存在大量离散节点，削弱了 PageRank 的总体传递效益。此类基于网络结构特性的计算方法，能够发现粉丝数多、关注或转发关系多的用户，但因没有充分考虑话题传播趋势的全局性特征和用户参与行为的过程性特征，难以有效发现粉丝数少、关注和转发关系少，但是在微博话题传播过程中保持持续关注和参与的用户，此类用户是微博话题的潜在关键人物，对话题影响力预测、关键人物分析具有重要意义。因此，针对微博话题中潜在关键人物的测量问题，提出用户行为的话题相关性指标。

定义用户在话题期间参与话题的行为分布与话题热度分布的相关性指标 r_u^S 来测量用户 u 行为与话题 S 传播趋势的共变关系大小，t 表示话题天数，N_i^S 表示距离观测起始时间第 i 天的与话题 S 相关的微博总数，$\hat{N}^S$ 表示 N_i^S 的平均值。$N_u^S(i)$ 表示距离观测起始时间第 i 天用户 u 所发表的与话题 S 相关的微博数，$\hat{N}_u^S$ 表示 $N_u^S(i)$ 的平均值。用户行为的话题相关性计算公式如式(3)所示。

$$r_u^S(t)=\frac{\sum_{i=1}^{t}\left\{\left[N_u^S(i)-\hat{N}_u^S\right]\times\left[N_i^S-\hat{N}^S\right]\right\}}{\sqrt{\sum_{i=1}^{t}\left[N_u^S(i)-\hat{N}_u^S\right]^2}\times\sqrt{\sum_{i=1}^{t}\left[N_i^S-\hat{N}^S\right]^2}} \tag{3}$$

相关系数对应用户数的累积分布函数曲线如图 10 所示，设定相关性阈值为 0.4 时，可计算出 $P\left(\left|r_{\text{user}}^{\text{Topic}}\right|\geq 0.4\right)\approx 20\%$，表明在参与话题的用户中，约 20%的话题用户即 5 350 人持续关注和跟踪了该话题发展过程。以昵称为“蓝色心语 LGZ”的用户为例，该用户在 4 月下旬、5 月中下旬、6 月下旬等话题传播过程中的多个关键时间点均有发表微博行为，且发帖频率和发布时间行为模式与话题传播趋势具有较强的一致性，是持续关注和参与话题的潜在关键用户。但在包含 26 755 个用户的话题关注网络中，该用户的出入度排名为第 3 568 位，PageRank 排名仅为第 11 854 位，而用户行为的话题相关性 r_u^S 值为 0.73，排名第 52 位。上述结果表明，用户行为的话题相关性指标能够有效检测出潜在关键用户。该指标综合考虑了话题微博热度趋势的全局性因素和用户参与话题行为的过程性因素，且时间复杂度低，不依赖于网络连通性，是判断话题影响力扩散能力和持续演化能力的有效指标，可以作为基于节点度分布、PageRank 等结构特性的关键人物发现方法的有益补充。

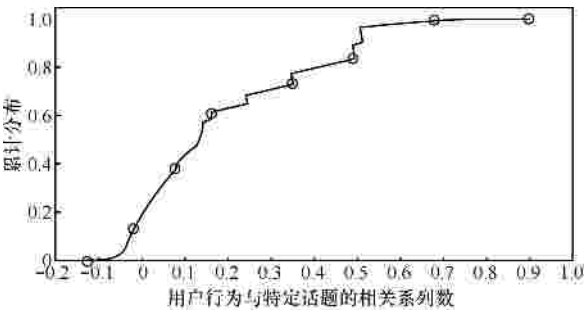


图 10　用户行为与话题的相关性

6　结束语

本文在充分分析了被测网络的话题相关性和动态生成网络的结构特性基础上，从突发话题传播规律、用户行为模式与话题的相关性角度，围绕话题传播的影响因素，开展了面向话题的微博网络测量研究，提出了相关指标的定量计算方法，经过在真实数据上的测量和分析，主要得出了以下结论：1）用户更倾向于转发负面微博，且只有少数微博被大量转发，转发次数与相应微博数呈现近似的幂率分布；2）整体微博热度随时间呈现较为规律的周期性，话题热度呈现明显的突发性和变化趋势，局部波动率能够有效地在大量背景微博中发现突发话题；3）话题转发网络的小世界特性并不明显，密集的关注关系不一定引发频繁的消息转发行为，

且用户并不倾向于多次从同一用户转发话题消息，转发行为表现出一定的随机性；4）参与话题的用户不一定具有或表现出与话题相关的历史兴趣，传播能力强的话题中含有较大比例的持续参与用户，用户行为的话题相关性能够有效检测此类潜在关键用户。

本文所提出的微博倾向性热度及传播规律、微博话题的突发特性、基于话题的微博网络结构特性、用户转发行为规律，以及用户兴趣和行为的话题相关性等测量方法，为展现和认识面向话题的微博网络话题传播内容特点、网络结构特性及用户行为模式规律提供了有效途径。提出的话题热度局部波动率、用户转发惯性、用户兴趣的话题相似度、用户行为的话题相关性等测量指标能够有效应用于微博话题影响力分析、关键人物发现、链路预测等相关研究，测量结果能够体现测量指标的有效性，并展现了许多有趣的结果。

本研究是在微博话题背景下针对微博网络测量开展的探索性工作，下一步可以尝试在不同类型话题数据中开展测量工作，分析测量指标的适用性，检测观测结果的普遍性；也可以考虑如何将本文测量指标与现有话题发现、影响力分析、链路预测等相关模型进行融合和改进，提升算法效果。

参考文献：

[1] http://www.cnnic.cn/gywm/xwzx/rdxw/rdxx/201302/t20130222_38842,htm[EB/OL].

[2] 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]. 2011. 65-74.

[3] KWAK H, LEE C, PARK H, *et al*. What is twitter, a social network or a news media[A]. Proc of the 19th Int Conf on World Wide Web[C]. New York, 2010. 591-600.

[4] CHEW C, EYSENBACH G. Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak[J]. PLoS One, 2010, 5(11):1-13.

[5] REN Z, LIANG S, MEIJ E, *et al*. Personalized Time-Aware Tweets summarization[A]. Proceedings of the 36th International ACM Sigir Conference on Research and Development in Information[C]. 2013.

[6] PATIL A, LIU J, GAO J. Predicting group stability in online social networks[A]. Proceedings of the 22nd International Conference on World Wide Web[C]. 2013.1021-1030.

[7] GRABOWICZ P A, AIELLO L M, EGUÍLUZ V M, *et al*. Distinguishing topical and social groups based on common ide ity and bond theory[A]. Proceedings of the Sixth ACM International Conference on Web Search and Data mining[C]. 2013.627-636.

[8] JAVA A, SONG X, FININ T, *et al*. Why we twitter: understanding microblogging usage and communities[A]. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis[C]. 2007.56-65.

[9] **樊鹏翼，王晖，姜志宏等**. **微博网络测量研究**[J]. **计算机研究与发展**, 2012, 49(4):691-699.
FAN P Y, WANG H, JIANG Z H, *et al*. Measurement of microblogging network[J]. Journal of Computer Research and Development, 2012, 49(4):691-699.

[10] KEMPE D, KLEINBERG J, TARDOS E. Maximizing the spread of influence through a social network[A]. KDD[C]. New York, USA, 2003. 137-146.

[11] **田家堂，王轶彤，冯小军**. **一种新型的社会网络影响最大化算法**[J]. **计算机学报**, 2011, 34(10): 1956-1965.
TIAN J T, WANG Y T, FENG X J. A new hybrid algorithm for influence maximization in social networks[J]. Chinese Journal of Computers, 2011, 34(10):1956-1965.

[12] HE X, SONG G, CHEN W, *et al*. Influence Blocking Maximization in Social Networks Under the Competitive Linear Threshold Model[R]. Technical Report, 2012.

[13] **陈浩，王轶彤**. **基于阈值的社交网络影响力最大化算法**[J]. **计算机研究与发展**, 2012, 49(10): 2181-2188.
CHEN H, WANG Y T. Threshold-based heuristic algorithm for influence maximization[J]. Journal of Computer Research and Development, 2012, 49(10): 2181-2188.

[14] BARBIERI N, BONCHI F, MANCO G. Topic-aware social influence propagation models[A]. Data Mining (ICDM), 2012 IEEE 12th International Conference on[C]. 2012.81-90.

[15] http://www.keenage.com/html/c_index.html[EB/OL].

作者简介：

刘玮（1984-），女，湖北武汉人，中国科学院博士生，主要研究方向为 Web 数据挖掘、智能信息处理等。

王丽宏（1967-），女，辽宁沈阳人，国家计算机网络应急技术处理协调中心副总工程师、研究员、博士生导师，主要研究方向为网络信息安全、智能信息处理等。

李锐光（1979-），男，山西阳泉人，硕士，国家计算机网络应急技术处理协调中心工程师，主要研究方向为网络与信息安全。