第47卷 第2期 西 安 交 通 大 学 学 报 Vol. 47 No. 2
2013年2月 JOURNAL OF XI'AN JIAOTONG UNIVERSITY Feb. 2013

DOI：10.7652/xjtuxb201302021

微博类社交网络中信息传播的测量与分析

张赛[1]，徐恪[1]，李海涛[2]

（1. 清华大学计算机科学与技术系，100084，北京；2. 西蒙弗雷泽大学计算科学学院，V5A 1S6，加拿大温哥华）

摘要：为了更好地掌握在线社交网络中信息传播的特征规律和用户行为，以新浪微博为代表对社交网络中的信息传播进行了较大规模的测量、统计和分析，提出了一种三角和算法用于探测用户粉丝数阈值。该算法根据散点分布的统计规律来估计使微博热度达到某一值的粉丝数的临界值，发现为使微博热度大于10，用户粉丝数应大于150。其他测量分析结果表明：新浪微博具有很强的“名人效应”，用户频繁地发帖并不能引起较大的关注，热门微博的热度几乎都以激增方式增长。这些结论对网络营销和网络监管具有参考价值。

关键词：在线社交网络；信息传播；微博热度；新浪微博

中图分类号：TP393.4 **文献标志码**：A **文章编号**：0253-987X(2013)02-0124-07

Measurement and Analysis of Information Propagation in Online Social Networks Like Microblog

ZHANG Sai[1], XU Ke[1], LI Haitao[2]

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. School of Computing Science, Simon Fraser University, Vancouver V5A 1S6, Canada)

Abstract: A large scale measurement and analysis of information propagation in online social networks with Sina microblog as a representative is performed to better grasp the characteristics of information propagation and users' behaviors in online social networks. A trigonometric sum algorithm is proposed to detect a threshold to the number of fans. The algorithm bases on the statistic law of the scatter distribution to estimate the threshold for the number of fans, and to get a given mircoblog popularity. It is found that one should have at least 150 fans to make microblog popularity more than 10. Other analytical results show that: the Sina microblog possesses a strong "celebrity effect", and users' frequent posting fails to arouse much attention. Most hot messages gain their popularity through surge-increase. Conclusions obtained from analyzing will be good references for network marketing and network supervision.

Keywords: online social networks; information propagation; popularity of microblog; Sina microblog

如今，在线社交网络（online social networks，OSNs）如新浪微博、人人网、Facebook、Flickr等，已经成为人们网络交流、获取信息和发布信息的主要平台之一。OSNs上的用户通过建立单向或者双向的好友关系交流、分享信息资源，他们不再是被动接受信息的媒体受众，而是积极地参与到网络活动中来，成为

收稿日期：2012-07-17。 作者简介：张赛（1988—），男，硕士生；徐恪（通作作者），男，教授。 基金项目：国家科技支撑计划资助项目（2011BAK08B05-02）；国家科技重大专项基金资助项目（2012ZX03005001）；国家自然科学基金资助项目（61170292，60970104）；国家“973计划”资助项目（2012CB315803）。

网络出版时间：2012-10-31 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20121031.1100.003.html

信息的制作者、分享者和传播者。由于 OSNs 具有巨大的用户基数、良好的传播时效性、相对自由的传播机制和廉价的使用成本，因此已经成为一个非常好的交友、商业营销（病毒式营销）等以信息分享和传播为目的的活动平台。这就有必要对 OSNs 的信息传播进行较为全面和深入的测量和分析。

学术界对 OSNs 的研究由来已久。最初对 OSNs 的研究主要着眼于其静态拓扑图[1]。在用户行为特征方面，Schneider 等人利用网络层追溯的方法，通过定义会话考察用户的交互行为特征[2]。

OSNs 的一个重要特征就是信息在好友关系连接上的传播，这种传播方式是用户对某一好友的某一条状态的转发。针对不同类型的 OSNs，信息传播的建模工作有较多的结果，比如在博客[3-5]、电子邮件[6]、类似于 Twitter 的微博网站[7-9]等方面。

新浪微博是目前国内最大的微型博客网站。本文对新浪微博上的信息传播进行了较大规模的测量和分析，想回答以下几个问题：

（1）新浪微博上“名人效应”的程度有多大？这种“名人效应”是否会减弱普通用户的影响力？

（2）微博热度与用户粉丝数有何种关系？是否可以找到限制微博传播的粉丝数阈值？

（3）微博热度与用户发帖时间有何种关系？哪一个时间段发帖可使微博流行的可能性最大？

（4）用户喜欢什么内容的微博？

（5）微博一般的传播规律如何？热门微博的热度增长是否遵循某种模式？

1　测量方法

1.1　数据采集方法

新浪微博是国内最大的微型博客网站，也是全球使用最多的微博客提供商之一。新浪微博于 2009 年 8 月 14 日开始内测，目前注册用户已经突破 3 亿大关，用户每日发博量超过 1 亿条，日活跃用户比例为 9%[10]。

微博转发是一种典型的信息传播行为。本文主要研究微博消息（信息）在新浪微博上的转发（传播）特征，因此在下文中，“微博”与“信息”、“转发”与“传播”不加区分，微博热度是指一条微博的转发数，以微博转发数来度量一条微博的流行度。

由于微博用户数据的隐私性，直接从微博运营商获取数据非常困难。本文采取网络爬虫的方法采集数据，主要利用美国麻省理工学院（MIT）开发的一个 Firefox 组件 Chickenfoot 进行数据采集。

1.2　数据描述

数据采集工作尽量基于新浪微博自身的功能设计。数据采集的时间为 2012 年 2 月 12 日至 2012 年 4 月 5 日。新浪微博有一个“随便看看”页面，主要功能是随机地刷新用户最近发布的微博。我们从这个页面共采集了 2 085 个随机用户，同时采集每个用户所发微博的所有相关信息（发帖时间、转发数），一共约 1 836 468 条微博。新浪微博还有一个“风云人气榜”，记录了它所统计的名人微博信息。我们从这个页面共采集了 523 个名人用户，同时采集每个用户所发微博的所有相关信息（发帖时间和转发数），一共约 2 745 473 条微博。

2　新浪微博的“名人效应”

本文选取了 1 103 个随机用户和 100 个名人用户。经计算，1 103 个随机用户中，共有 870 个用户（78.88%）关注了名人，其中有一个用户关注了 98 个名人（最大值），这体现了新浪微博的名人效应（见图 1）。大量用户（523 个）关注了不超过 10 个名人，用户关注名人数量的平均值为 9.907 5，而关注大量名人的用户则很少（见图 2）。

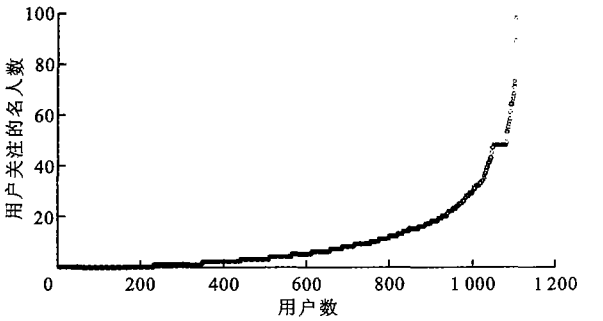


图 1　新浪微博用户关注名人的数量分布图

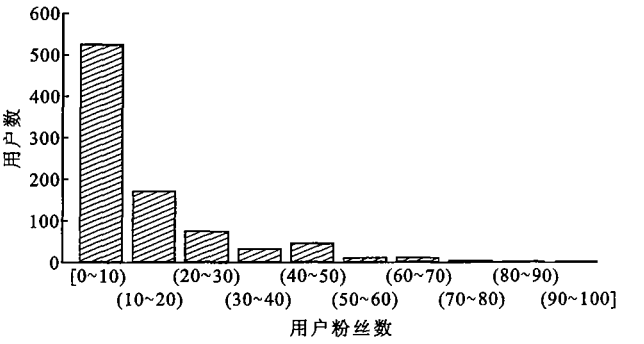


图 2　新浪微博用户关注名人的数量统计图

图 3 显示了名人的粉丝数情况，并与随机采集的情况（即统计名人在随机采集的数据集中的粉丝数量）作对比。在所统计的名人用户中，最大粉丝数

是15 900 262，平均粉丝数是 6 560 705；而在随机采集中，最大粉丝数是 280，平均粉丝数是 95。图 3 表明，随机采集的数据集中的名人的粉丝数与实际这些名人的粉丝数的分布情况一致，这验证了本实验所采集数据的随机性。

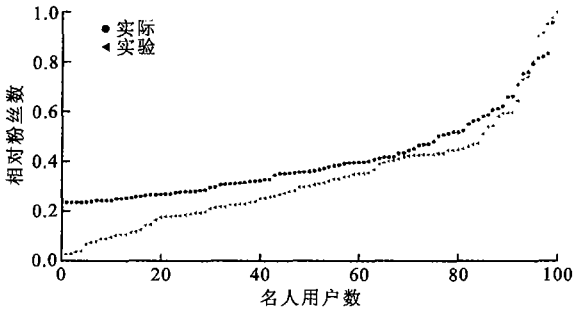


图 3 名人用户的粉丝数

3 微博及微博的转发特征

3.1 微博热度与粉丝数

一条微博的热度会与众多因素有关，比如用户粉丝数、微博内容等等。这一节主要考察微博热度和粉丝数之间的关系。

本文统计了 1 089 个随机用户的信息，包括用户粉丝数(λ_i)、微博数和微博转发数，同时计算了每个用户的微博最大转发数(ω)。

图 4 显示了粉丝数与微博最大转发数之间的正相关性。为了定量考察它们的相关性，本文利用下式计算二者的相关系数

$$R(i,j)=\frac{C(i,j)}{(C(i,i)C(j,j))^{1/2}} \tag{1}$$

式中：i、j 分别代表第 i、j 列向量；$C=\mathrm{cov}(i,j)$。

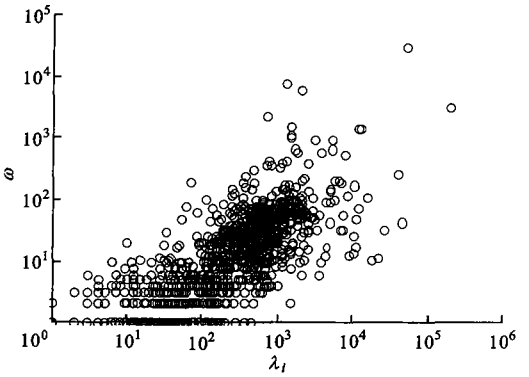


图 4 用户粉丝数与微博最大转发数的对数关系图

经计算，得粉丝数和微博最大转发数的相关系数为 0.305 4。可见，虽然二者具有某种正相关性，但这种相关性并不很强。从图 4 还可以看出，有较多拥有 100 左右粉丝数的用户的微博最多仅被转发 1 到 2 次，这在一定程度上说明粉丝数并不是微博热度的决定性条件。

微博热度和用户粉丝数之间是否有更深入的关系呢？本文提出一种三角和算法来考察这一问题。

3.2 三角和算法

设样本空间 Ω 是含有 2 085 个(ω,λ_i)的二元组集合。

(1)输入：微博最大转发数 ω。

(2)输出：粉丝数临界值 λ。

(3)算法描述：一个数值 a 的临界值 b 的严格定义是，当 $x\geqslant b$ 时，$f(x)\geqslant a$，且当 $x\leqslant b$ 时，$f(x)\leqslant a$。具体算法的解释可由图 5 说明。此时，落入第 1 区间的点 λ_i 是严格临界值的正贡献量，而 2、4 区间的点 λ_i 是严格临界值的负贡献量。因此，λ_i 是 ω 的严格临界值的概率为

$$p(\lambda_i)=\frac{1}{1+2+4} \tag{2}$$

当 ω 固定时，使得 p 取最大值的 λ 即为 ω 的临界值。

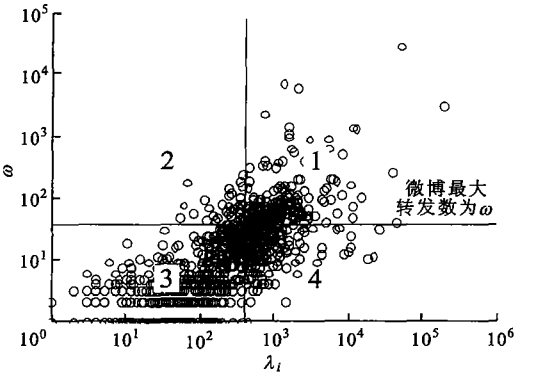


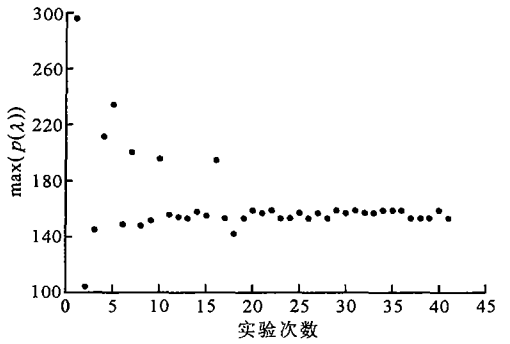
图 5 三角和算法描述示意图

本文以 50 为步长，每次从 2 085 个样本中随机选择 $i\times 50$ 个样本计算其对应的 λ，这样共计算 41 次，即 $i=41$。算法伪代码如下：

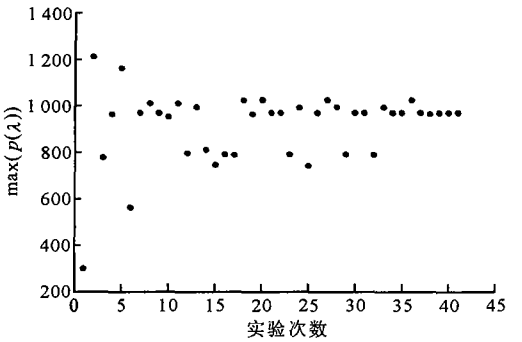
```
For each i from 1 to 41
    randomSelecting(dataset,i×50);
    For each j from 1 to max(fansCnt)
      For each k from 1 to i×50
        If (x(k)≥j&&y(k)≥maxforwarding)
            C1=C1+1;
        End
        If((x(k)<j&&y(k)≥maxforwarding) || (x(k)
          ≥j&&y(k)<maxforwarding))
            C2=C2+1;
```

```
            End
        End
        L(j)=C1/(C1+C2);
    End
    [value,location]=max(L);
    fannum(i)=location;
End
```

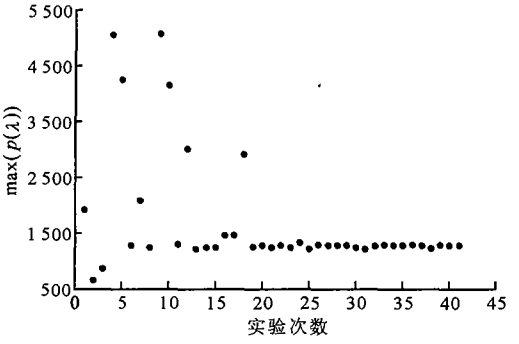
图 6 是在增加样本数量的情况下，取 max{$p(\lambda)$}时对应的粉丝数 λ 的变化规律。当样本数量逐渐增加，λ 也趋于稳定，当 $\omega=10,50,100$ 时，λ 的稳定区间分别为(150,160)、(950,1 000)、(1 200,1 300)，这也验证了该算法的合理性。



(a)$\omega=10$



(b)$\omega=50$



(c)$\omega=100$

图 6　不同 ω 对应的粉丝数趋势图

3.3　微博热度与发帖频率

这一节考察微博热度与用户发帖频率的关系。本文统计了 523 个名人微博信息，并计算每个用户的发帖频率(平均每天所发微博数)、平均转发数。经计算，平均发帖频率为 7.12 d^{-1}，最大值为 212.25 d^{-1}，最小值为 1.04 d^{-1}。

图 7 为发帖频率与平均转发数的对数关系，可以看出图中并没有显示用户发帖频率和微博热度(平均转发数)之间的关系，可见二者之间的相关性较差，经计算相关系数仅为 −0.016 1。但是，这并不能说明二者呈现弱的负相关性，因为频率越高，基数越大，所得的平均值也会越小，而当使用最大转发数作为微博热度的评估时，相关系数仅为 0.003 4。因此，频繁发帖并不能给微博热度带来增益。

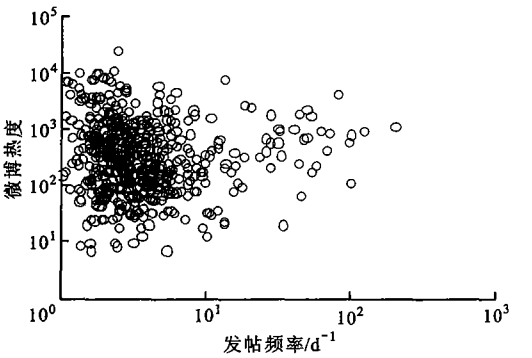
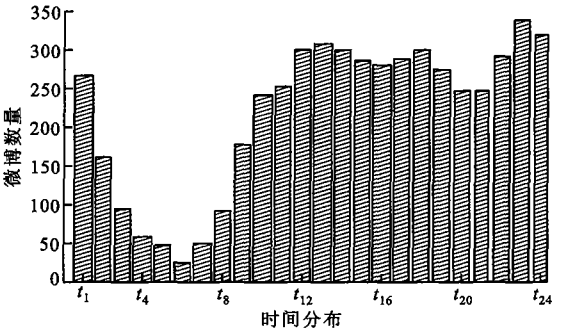


图 7　发帖频率与平均转发数的对数关系图

3.4　微博热度与发帖时间

当发帖时能够被更多的人看到，该微博流行的可能性就会增加。同样以 523 个名人为对象，采集每个用户的前 10 条具有最大转发数的微博的发布时间，一共是 5 230 条微博信息，并统计每个时间段内的微博数量。

图 8 和图 9 分别是不同时间间隔下微博数量的分布情况。近似地把不同区间内微博的数量看作该



t_1 为 00:00～01:00；…；t_4 为 03:00～04:00；…；t_8 为 07:00～08:00；…；t_{24} 为 23:00～00:00

图 8　微博热度与发帖时间的统计分布图(时间间隔 1 h)

时间段内微博的热度，可以看出：一天中 03:00 到 06:00 之间的微博数量最少；09:00 以后微博数量均较多，12:00～13:00、17:00～18:00 和 22:00～23:00 是之后的 3 个高潮；21:00 到 00:00 是一天中微博数量最多的时间段。

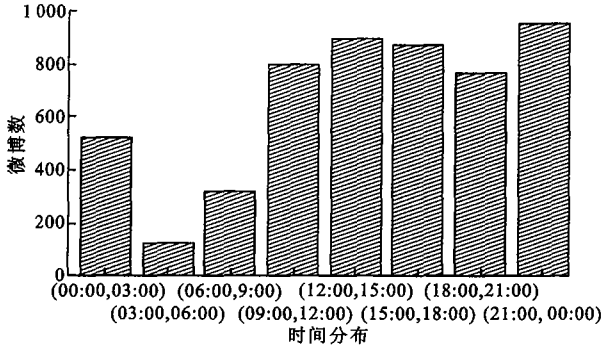


图 9 微博热度与发帖时间的统计分布图
（时间间隔 3 h）

3.5 微博的转发类型

从 523 个名人微博中随机挑选了 50 条微博，按照其转发数分为 2 类：取 25 条小于 1 000 次；25 条大于 1 000 次。同时，挑选那些已经衰亡的微博，以一天内少于 3 人转发作为衰亡的标志。通过以下 4 个较典型的微博，对这些微博的转发过程进行对比。

图 10～图 13 为 A、B、C、D 4 个典型微博的生命周期曲线。可以很明显地看出，微博转发过程中的特征因素：拐点和增长趋势。从增长趋势上看，热门微博常以一种阶梯函数似的激增方式增长，经过一段很短时间的弧形拐点后达到潜伏状态或者衰亡。因此，我们把热门微博的转发生命周期粗略地分为潜伏期、激增期和衰亡期，见图 10。潜伏期和衰亡期的划分依赖于是否存在二次增长，见图 13，这种二次增长往往是由博主或者其他有影响力的用户转发造成的。热门微博这种激烈而单一的增长方式说明：一条微博要么快速受到关注，要么受到冷落，或者就此衰亡，或者等待新一轮的激增。

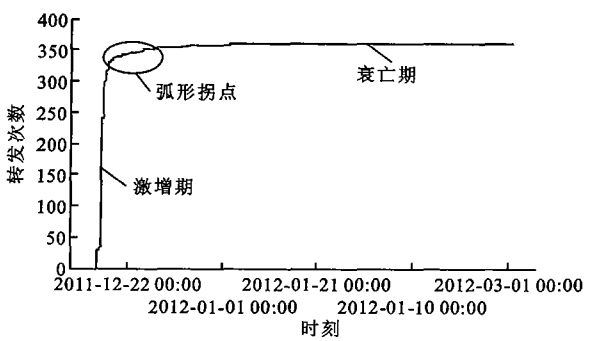


图 10 微博 A 的生命周期曲线

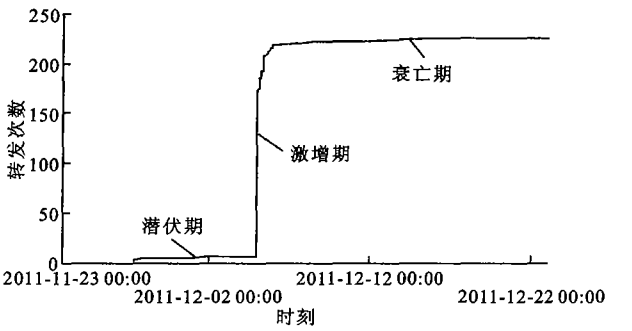


图 11 微博 B 的生命周期曲线

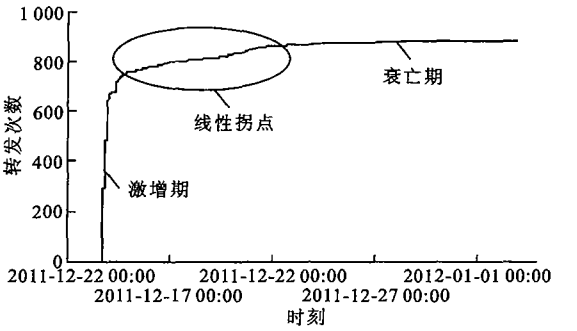


图 12 微博 C 的生命周期曲线

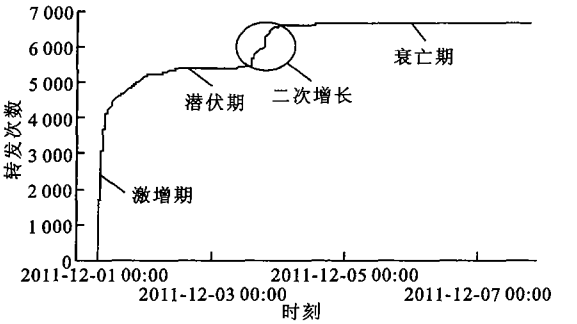


图 13 微博 D 的生命周期曲线

并不是说线性增长在热门微博转发中不存在，例如在微博 C 中出现了线性拐点（见图 12），它出现在激增期结束之后，给微博转发带来了 11.1%的增益。在我们考察的 50 条微博中，没有出现开始就呈线性增长趋势的微博，50 条微博中仅有一条微博开始即出现潜伏期，见图 11。

本文的工作可以与 Cha 等人的工作[11]做对比。他们通过考察 Flickr 上照片粉丝数的增长过程，对照片的热度增长模式进行分类，也得到了 4 种类型。但照片的热度增长与微博明显不同的是，Flickr 上照片热度的线性增长占很大比例。

3.6 热门微博的内容分布

上面的讨论都不涉及微博的内容，然而微博内

容对其热度显然有非常重要的影响。比如最近一段时间社会上发生了某个热门事件，关于此话题的微博就会很热门。具体对微博内容的信息挖掘会有更深入的研究，这已超出了本文的范围，这里只粗略考察热门微博的内容分类。

注意到不同类型的名人往往代表不同的话题，不同类型名人微博的热度在一定程度上体现了微博用户对不同话题的喜好。以下为新浪微博对名人的分类，本文只考察一级分类和若干二级分类。

名人：
- 体育、娱乐、财经、文学、科技
- 草根：内容类、个人草根（达人）
- 名人 媒体：报纸、杂志、电视、电台
- 政府：公安、交通、司法、医疗卫生
- 网站：生活、电商、科技、娱乐

本文选取 8 个较典型的分类：娱乐、体育、财经/房产、文化/时尚、科技、草根、媒体、网站，分别考察了每个类别中用户微博的平均转发数，见图 14。

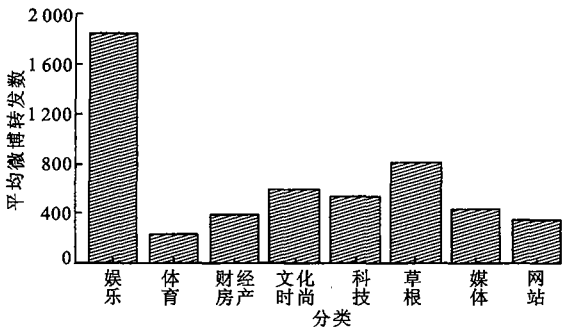


图 14　微博转发数的内容分布

从图 14 可以看出微博用户的兴趣分布：娱乐类微博占很大的优势；草根类微博屈居第二。这说明，虽然微博具有前述的名人效应，但作为“非名人”的草根用户仍会在微博中拥有非常大的关注度。另外，一些新闻媒体、网站等非个人用户在新浪开博的效果也很好，受到了较大的关注。

4　总　结

本文主要考察了与微博转发有关的若干因素，并分析了这些因素对微博转发的影响，借此来理解和认识信息在在线社交网络中的传播规律。现在可以回答引言中提出的问题：

（1）新浪微博是一个以名人为聚集中心的在线社交网络，这些名人通常来自于现实中的知名人士，同时也包括现实中作为普通人的草根用户，因此新浪微博的名人效应并没有限制普通用户的影响力；

（2）微博热度与用户粉丝数呈正相关性，即粉丝数越多，微博越易被转发，但这种相关性并不很大，这与一条微博复杂的传播机制有关；

（3）微博热度与粉丝数之间存在一定的阈值关系，即可以计算出为了使一条微博转发次数达到某一值的粉丝数的最佳值；

（4）发帖频率与转发数之间没有直接的相关性；

（5）发帖时间对微博热度有着较明显的影响，每天的 21:00 到 0:00 是微博的最佳转发时间，而 3:00 到 6:00 属于转发“静默期”；

（6）不同内容的微博的热度也有着较大差距，其中娱乐类微博占了很大一部分份额，草根用户的微博也不容小视，他们通常受到用户更多的关注；

（7）热门微博的热度增长模式主要是激增方式，开始没有受到关注的微博很难成为热门微博，二次转发是给微博带来转发增益的重要方式。

影响微博消息传播的因素很多，本文只是管中窥豹。如何用数学语言严格刻画 OSNs 中的信息传播，将是我们下一步的工作方向。

参考文献：

[1] MISLOVE A, MARCON M, GUMMADI K P, et al. Measurement and analysis of online social networks [C]// Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM, 2007: 29-42.

[2] SCHNEIDER F, FELDMANN A, KRISHNAMURTHY B, et al. Understanding online social network usage from a network perspective [C]// Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM, 2009: 35-48.

[3] ADAR E, ADAMIC L A. Tracking information epidemics in blogspace [C] // Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC, USA: IEEE Computer Society, 2005: 207-214.

[4] GOMEZ-RODRIGUEZ M, LESKOVEC J, KRAUSE A. Inferring networks of diffusion and influence [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2010: 1019-1028.

[5] GRUHL D, GUHA R, LIBEN-NOWELL D, et al. Information diffusion through blogspace [C] // Proceedings of the 13th International Conference on World Wide Web. New York, USA: ACM, 2004: 491-501.

[6] LIBEN-NOWELL D, KLEINBERG J. Tracing information flow on a global scale using Internet chain-

letter data [J]. Proceedings of the National Academy of Sciences, 2008, 105(12): 4633-4638.

[7] GOYAL A, BONCHI F, LAKSHMANAN L V. Learning influence probabilities in social networks [C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2010: 241-250.

[8] LERMAN K, GHOSH R. Information contagion: an empirical study of the spread of news on Digg and Twitter social networks [C]// Proceedings of 4th International Conference on Weblogs and Social Media. Washington, DC, USA: AAAI, 2010: 66-75.

[9] BAKSHY E, HOFMAN J M, MASON W A, et al. Everyone's an influencer: quantifying influence on twitter [C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2011: 65-74.

[10] 新浪科技有限公司. 2011年第四季度及全年财报 [EB/OL]. (2012-02-28) [2012-06-10]. http://tech.sina.com.cn/i/2012-02-28/05296776965.shtml.

[11] CHA M, MISLOVE A, GUMMADI K P. A measurement-driven analysis of information propagation in the Flickr social network [C]// Proceedings of the 18th International Conference on World Wide Web. New York, USA: ACM, 2009: 721-730.

(编辑 刘杨 赵大良)

(上接第32页)

[3] BARHAM P, DRAGOVIC B, FRASER K, et al. Xen and the art of virtualization [C]// Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York, USA: ACM, 2003: 164-177.

[4] VMWare Inc. Resource management with VMWare DRS [EB/OL]. (2006-06-05) [2012-04-10]. http://www.Vmware.com/vmtn/resources/401.

[5] ZHUANG Wei, GUI Xiaolin, HUANG Ruwei, et al. TCP DDOS attack detection on the host in the KVM virtual machine environment [C]// Proceedings of the 11th IEEE/ACIS International Conference on Computer and Information Science. Piscataway, NJ, USA: IEEE Computer Society, 2012: 62-67.

[6] MATTHEWS J N, DOW E M, DESHANE T, et al. Running Xen: a hands-on guide to the art of virtualization [M]. Boston, USA: Prentice Hall, 2008.

[7] CLARK C, FRASER K, HAND S, et al. Live migration of virtual machines [C]// Proceedings of the Second Conference on Symposium on Networked Systems Design and Implementation. New York, USA: ACM, 2005:273-286.

[8] HU Liting, JIN Hai, LIAO Xiaofei, et al. Magnet: a novel scheduling policy for power reduction in cluster with virtual machines [C]// Proceedings of the 2008 IEEE International Conference on Cluster Computing. Piscataway, NJ, USA: IEEE Computer Society, 2008: 13-22.

[9] 周文煜, 陈华平, 杨寿保, 等. 基于虚拟机迁移的虚拟机集群资源调度 [J]. 华中科技大学学报:自然科学版. 2011, 39(S1): 130-133.
ZHOU Wenyu, CHEN Huaping, YANG Shoubao, et al. Resource scheduling in virtual machine cluster based on live migration of virtual machine [J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2011, 39(S1): 130-133.

[10] FADWA G, MEDIA A. Web based multi criteria decision making using AHP method [C]// Proceedings of the 2010 International Conference on Information and Communication Technology for the Muslim World. Piscataway, NJ, USA: IEEE Computer Society, 2010: 6-12.

[11] CAO Jian, YE Feng, ZHOU Gengui, et al. A new method for VE partner selection and evaluation based on AHP and Fuzzy theory [C]// Proceedings of The 8th International Conference on Computer Supported Cooperative Work in Design. Ontario, Canada: NRC Research Press, 2004: 563-566.

(编辑 武红江)