基于任务队列的新闻报道模型

尤志强，朱燕燕，韩筱璞，吕琳媛

(杭州师范大学阿里巴巴复杂性科学研究中心 杭州 311121)

【摘要】基于新浪新闻数据，对热点新闻的连续发表事件时间间隔序列进行了统计分析，以探究新闻内容的选择机制。实证发现该时间间隔分布在个类与总体层面上都遵循带指数截断的幂律分布，由此提出一种考虑时效性的，并基于严格优先及偏好优先选择混合机制的队列模型来揭示新闻选择背后的机制。该模型的数值模拟结果与实证统计数据较好地吻合，表明该模型规则在一定程度上可用于解释新闻报道中出现的非泊松时间特性。

关 键 词 爆发性; 新闻选择; 幂律分布; 任务队列模型; 时间间隔分布

中图分类号 N94 **文献标志码** A doi:10.3969/j.issn.1001-0548.2016.03.023

Queuing Model for News Reports

YOU Zhi-qiang, ZHU Yan-yan, HAN Xiao-pu, and LÜ Lin-yuan

(Alibaba Research Center for Complexity Sciences, Hangzhou Normal University Hangzhou 311121)

Abstract In this paper, based on the news data of Sina website, inter-event time interval sequences of hot news publication are analyzed to reveal the hidden rules of news selection. Empirical analysis shows that the distributions of the inter-event time intervals between two consecutive news with common keywords follow power-law-like distribution with exponential cutoff both on individual level and aggregated level. Focusing on this finding, we propose a timeliness-based queuing model with mixed mechanisms of strict and preferential priority selections to reveal the hidden principle of news selection. The model results are generally in agreement with the empirical findings, indicating that the proposed model can explain the emergence of non-Poisson properties in news reports.

Key words burstness; news selection; power-law distribution; queuing model; time interval distribution

信息传播目前是学术界炙手可热的研究领域。学者对不同网络上的信息传播及动力学进行了大量的研究[1-5]，有助于人们理解信息扩散的机制及对舆论控制的研究。然而，对于信息内容的产生机制的研究却鲜有报道。目前学界关注的重点集中在社交网络层面，如谣言传播[6-7]、创新扩散[8]、人类行为对传播的影响[9-10]等，而对新闻这类主流信息传播主体的研究十分缺乏，特别是针对新闻内容产生机制的研究更是难觅踪迹。文献[11]虽关注的是新闻，不过其研究的是新闻的密集报道产生的影响力。新闻媒体在现代信息传播中扮演着重要的角色。至今，人们对于各类新闻的发表规律知之甚少。因此，对新闻内容产生机制的研究，将有助于更好地理解新闻的性质特点以及加深对信息传播的理解。

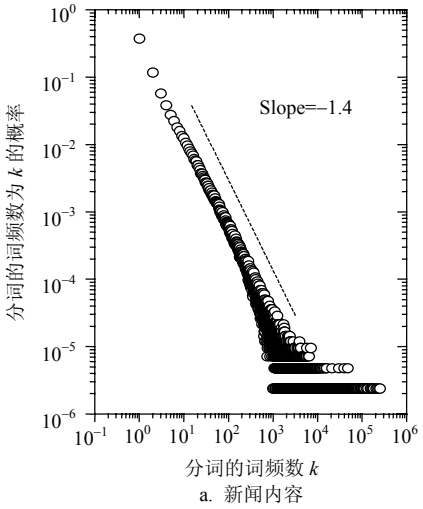
新闻，顾名思义，是一种新近发生的事件，通常人们会认为新闻的选择是基于时间及重要性的绝对优先原则，那是否意味着只有最新最重要的事件才会被报道，或存在其他的新闻产生机制？目前，针对新闻的相关性质特点的研究主要集中在社会科学领域，但随着网络科学在复杂系统中的应用日趋成熟，使用复杂网络领域知识来研究新闻的选择机制值得尝试。当前复杂网络研究在包括人类任务处理[12]、地理活动[13-15]、邮件[16-17]、短信[18]、通话[19]等方面都取得了相当丰硕的成果。文献[17]通过研究用户从接收信件到回复信件之间的间隔反应时间序列发现该反应时间间隔分布存在幂律现象。文献[18]通过研究用户连续进行短信发送事件的时间间隔序列，发现在个体用户层面的连续事件时间间隔分布遵循幂律分布。受此类研究方法的启发，本文从时间统计特性的角度对新闻数据进行研究，分析新闻

收稿日期：2014－11－10；修回日期：2015－03－24
基金项目：国家自然科学基金(11205040,11205042)；浙江省新苗人才计划(ZX13005002062)
作者简介：尤志强(1990－)，男，硕士生，从事复杂网络与数据挖掘方面的研究

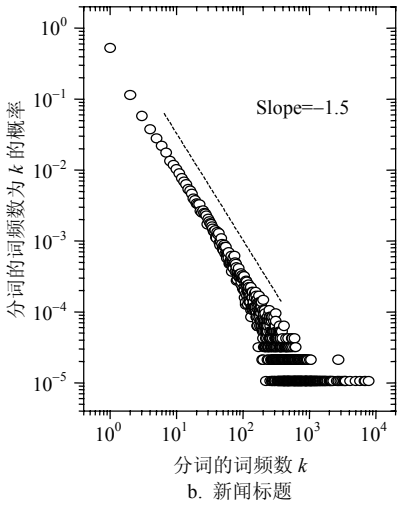
选择的潜在机制。

本文使用新闻标题关键词表征新闻类别，如关键词“暴雨”表示一类新闻。根据新闻的发表时间信息可以刻画出每一类新闻的连续发表事件时间间隔序列，该时间间隔表示同类新闻连续两次发表之间的时间差，实证分析发现新闻的连续发表事件时间间隔分布在个类层面和总体层面上都呈现为带指数截断的幂律分布。基于该实证发现，本文提出一种考虑时效性的混合机制队列模型来研究新闻选择机制的动力学过程，模型所得结果与实际数据较好地吻合，表明对新闻内容的选择在新闻时间统计特性产生中可能起了重要作用。

1 数 据



a. 新闻内容



b. 新闻标题

图1 新闻内容与新闻标题分词词频概率统计分布

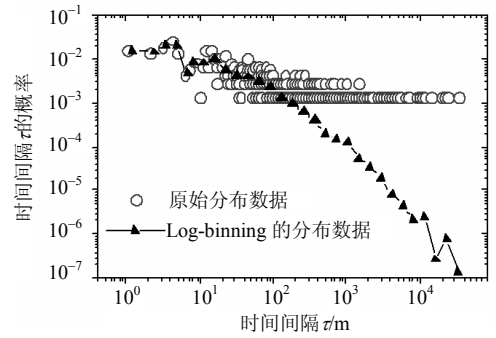
新浪新闻是中国重要的新闻内容提供方，内容涵盖了社会、体育、娱乐、财经等领域。本文采用了新浪新闻2012年1月1日—2012年12月31日的新闻数据[20]。该新闻数据以季度划分，每个季度为一个文件，共包含25万条新闻，约2.5亿字。每条新闻包含以下内容：新闻的URL、使用的字符编码、标题、关键字、描述、报道媒体以及新闻内容等，格式为XML。另外，新闻URL信息中包含了每一条新闻发布的具体时间信息，精确到分钟。

标题是对具体内容的高度浓缩，对标题的关键词提取，可以便捷地得到新闻的主题内容信息。因此，本文重点对标题关键词进行了提取和统计分析。为了验证使用标题关键词的合理性，需要确保标题关键词的词频与新闻文本内容关键词词频具有相似的分布。本文使用中文分词工具盘古分词软件[21]对新闻标题及文本内容进行分词。为了排除虚词的影响，本文过滤掉长度小于2的词语，分别对内容以及标题的关键词词频进行统计。图1a表示新闻内容分词词频概率分布，图1b表示新闻标题分词词频概率分布，可以看到两者具有相似的幂律分布特性。此外，针对标题文本的分词，选取长度不小于2且词频数不少于500的关键词，作为后面研究分析的对象。限制词频数不小于500，是为了确保可以得到足够长的同类新闻连续发表事件时间间隔序列以利于分析相关性质。通过以上数据预处理，得到新闻关键词331个，每个关键词表征了一类新闻。

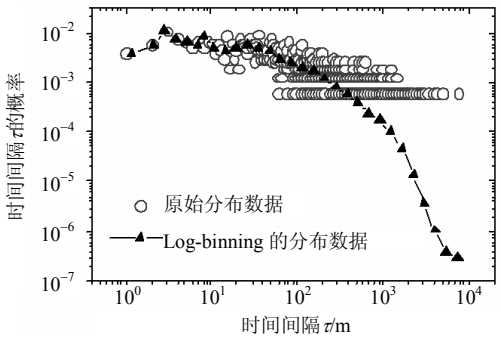
2 统计分析

本文以分钟为基本时间单位，对所有热点新闻分别提取其相应的新闻发表事件的时间数据，并依此得到每一类热点新闻的每连续两次发表事件的时间间隔序列，该时间间隔使用τ表示，并进一步对热点新闻在个类及总体层面(即综合所有新闻类别)上统计了新闻连续发表事件的时间间隔分布。图2展示了其中的4类高频热点新闻的时间间隔分布(其余热点新闻均表现出相似分布特征)，这4个关键词分别为暴雨、爆炸、官员、枪击，其中空心圆表示实际数据的分布，实心三角形表示logarithmic binning处理后的结果。从图2可以看到同类新闻连续发表事件的时间间隔分布具有带指数截断的幂律分布特性，如表1所示，本文使用带指数截断的幂律分布函数$y=Ax^{-B}e^{-Cx}$对分布做了拟合，即同类新闻会在短时间内频繁发布，而较少出现长时间静默的情形。此外，不同类别的新闻虽然总体趋势相似，但指数截断强度存在差别，表明不同新闻在长时间静默表现上具

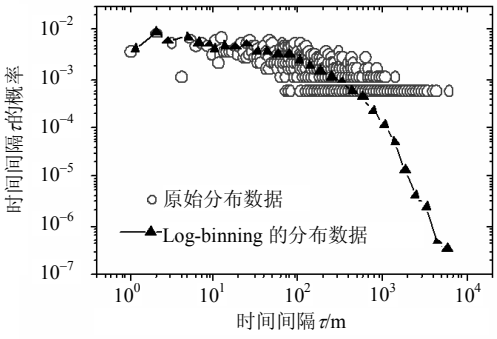
有不同的倾向，如“暴雨”，指数截断现象较弱，尾部分布近似于幂律，而爆炸、官员、枪击等新闻则表现出较强的指数截断，表明这些类别新闻更偏好于短时间的集中发布。



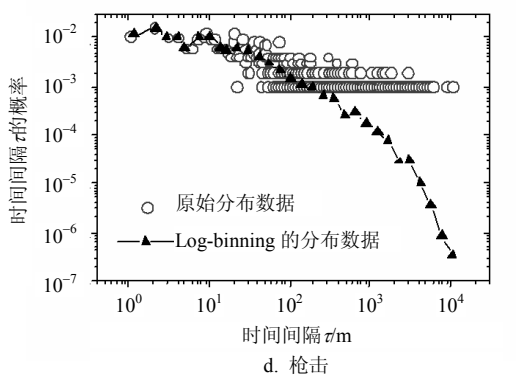
a. 暴雨



b. 爆炸



c. 官员



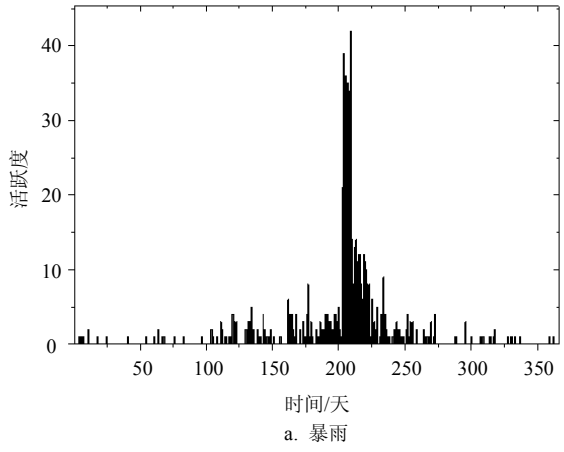
d. 枪击

图2　单个新闻关键词的发表时间间隔分布

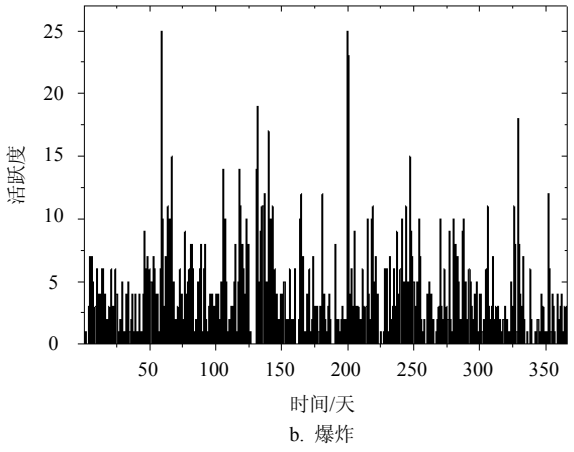
表1　新闻热词连续发表事件时间间隔分布拟合结果

热词	幂指数B	指数$C\times10^{-3}$
美国	0.45	3.54
中国	0.45	4.25
日本	0.4	3.25
男子	0.4	4.50
叙利亚	0.25	2.20
北京	0.25	2.20
暴雨	0.3	5.00
爆炸	0.35	2.00
官员	0.25	2.00
枪击	0.3	3.00

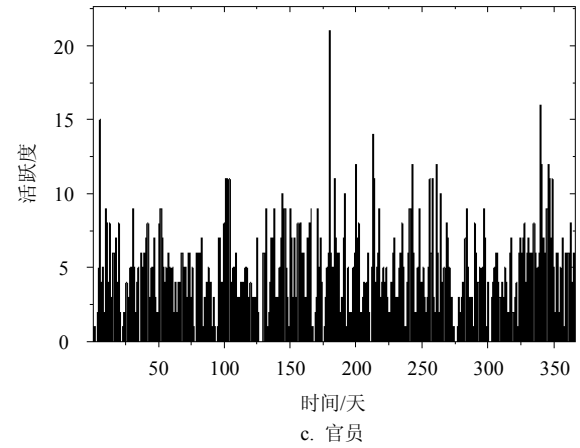
为了更好理解该现象，本文分别画出了这4个词的活跃性分布图，以天为单位，一天内该新闻发表的次数为活跃性，如图3所示，可以看到“暴雨”在6、7月份异常活跃，短时间内发表非常频繁，而其他时候基本处于长时间静默，导致其指数尾不明显。“枪击”活跃性分布呈现出明显的周期性，在80、200、350天左右呈现高频爆发，而其他时间相对静默，导致其指数尾也较弱。然而相对暴雨和枪击，爆炸和官员新闻则没有表现出明显的阵发现象，呈现出一定的随机性，导致产生较明显的指数尾。



a. 暴雨



b. 爆炸



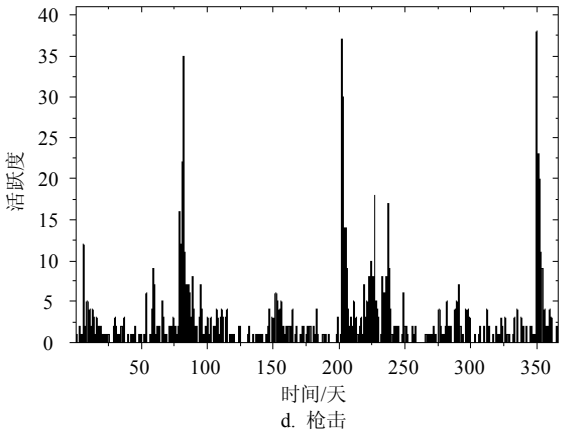


图3 单个新闻关键词每天的活跃性变化

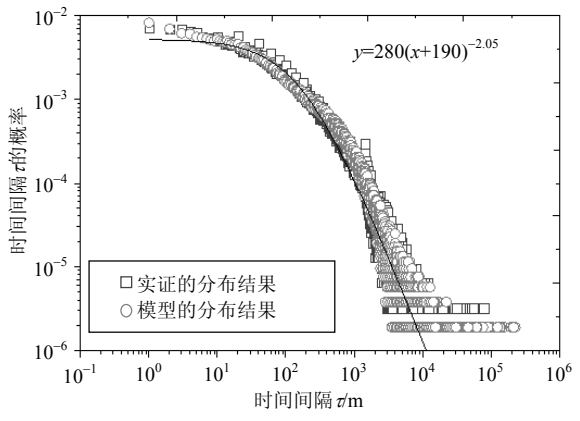


图4 总体层面上模型结果与实证数据的连续发表事件时间间隔分布对比

图4比较了模型结果与实证数据在总体层面上的新闻连续发表事件时间间隔分布，其中，空心框表示在总体层面上的实证数据分布结果，空心圆表示模型数据分布结果，其中C为新闻类别数量，L为新闻候选列表长度，α为机制选择概率，此时模型中各参数分别为C=600，L=200，α=0.22，黑线为函数拟合结果，可以看到模型结果与实证结果两者基本符合，可以用漂移幂律很好拟合，特别是在分布尾部模型数据可以与实证数据分布保持一致，都存在明显的非泊松特性。

3 模 型

新闻工作者从候选新闻素材中选择正式发表的新闻内容与人们处理任务队列中任务的行为相似，因此，在文献[12]提出的反映人类行为的队列模型基础上，本文提出了一种考虑时效性的混合机制队列模型。该模型的核心机制主要考虑了如下3点：1) 绝对优先机制。该机制严格依据新闻的重要性从新闻素材候选队列中选择重要性权重最大的新闻，该机制强调新闻本身的重要性；2) 偏好优先机制。按照新闻的重要性权重成比例地从新闻队列中随机选择新闻，权重值大的新闻更有可能被选中，但权重小的新闻依然有机会被选择；3) 新闻信息冗余和强时效性，即可供发布的新闻远远多于能够发布的新闻，而且选择的新闻一般为近期的新闻素材，未能及时发布的新闻随时间推移迅速丧失其意义。

考虑上述因素后，该模型首先定义C个类别新闻，每一类新闻赋予固定的权重值ω以表征其重要程度，ω在0～1之间随机选取。固定新闻的权重值，是因为各个类别的新闻的重要程度不会出现较大波动，如“枪击”“总统”等类别新闻重要程度一直很高。模型更新规则如下：

1) 在t=0时刻，初始化长度为L的新闻列表，该列表可视为新闻的备选库。这L条新闻的类别是从C种类别中随机选择。由于各个新闻类别的重要性ω值已经固定，因此选入队列的新闻的ω值也由其类别确定。

2) 进行新闻选择过程，如图5所示。图5a表示有5条新闻的待选队列，圆圈表示新闻，圆圈大小正比于新闻的重要性ω，灰度深浅用以区分新闻类别。模型以概率α使用绝对优先机制选择新闻，即直接选取队列中ω最大的发表，如图5b示；或者以概率1–α按照偏好优先机制进行新闻选择，即某新闻i被发表的概率$\Omega_i=\omega_i/\Sigma\omega$，如图5c所示。

3) 选择完成后，从队列中删除被选中的新闻，并往队列中添加一条新的新闻，这条新的新闻的类别也是从C类新闻中随机选择。

4) 如果超过了L/20时步，一条新闻还没被选过，那么就从候选列表中将它删除，并从C类新闻中随机选择一条新闻添加到新闻候选列表。选择L/20，是考虑到新闻的时效性，新闻在L/20时步后仍未被发布则视为失去发布意义。

在模拟过程中，迭代时间总步数设置为366×24×60步，即模拟2012年全年的分钟数。待完成迭代，提取同类新闻连续发表事件的时间间隔序列，并综合所有新闻类别，统计总体水平上该连续事件时间间隔分布。

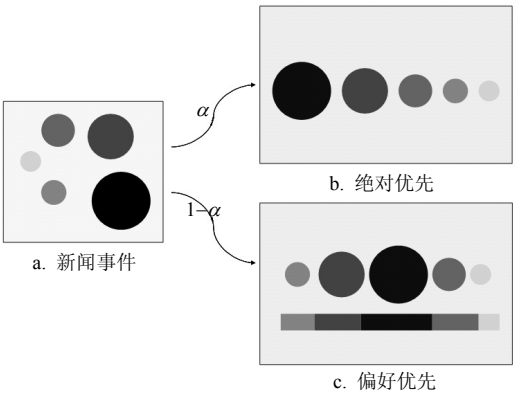
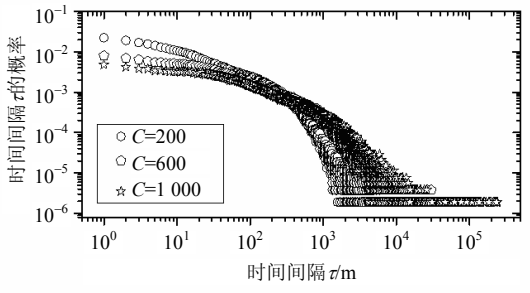


图5　模型机制示意图

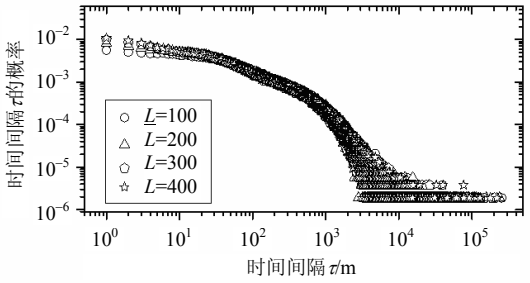
4　结果与分析

此外，本文进一步探究了模型中各参数对结果的影响，如图6所示。图6a展示了新闻类别数量C对结果的影响，固定L=200，α=0.22，分别选取C在200、600、1 000时模型的结果进行比较。可以看到随着C的增加，新闻连续发表事件时间间隔分布在τ小于300的区间出现下降趋势，300～500为过渡区间，大于500区间，随着C的增加，分布出现右移趋势。图6b显示了新闻候选列表长度L对结果的影响，参数固定C=600，α=0.22，分别选取L为100、200、300、400进行实验。可以看到随着L的增大，分布只在τ为[1,10]区间部分发生较明显变化，L越大，该部分分布抬升越显著，而尾部变化不明显。图6c展示机制选择概率α对结果的影响，固定C=600，L=200，分别选取α=0.1、0.2、0.3时的模型结果进行研究。可以看到，α只对模型分布结果在τ为[1,100]区间产生影响，值越大，分布抬升越明显。

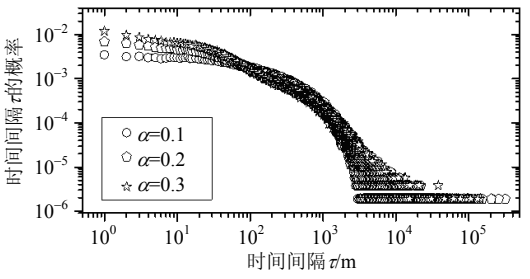
图6显示了随着C的增大，可被挑选到新闻候选列表中的新闻种类变多。由于模型规则规定在补充候选新闻列表时采用随机从C类新闻中抽取的方式，客观上导致每一类新闻被抽到的概率变小，进而造成候选列表中存在同类新闻的可能性降低，最终使得同类新闻短时间内被重复选择的机率降低。在总体层面上，新闻连续发表时间间隔τ及其比例都表现出变大的趋势。而候选列表长度L的变大，使得候选列表中同类新闻的存在可能性增加，提高了同类新闻短时间内被多次发表的概率，不过受此影响最大的应该是具有较高权重值的新闻，因为模型偏好高权重值的新闻发表，因此可以看到较小τ的比值有所增加但幅度较弱，长度L对结果的影响没有新闻种类C变化带来的影响大。另外，选择概率α是控制偏好选择的比例，偏好选择比重增加，导致更多的具有较高权重值的同类新闻被选择的概率增大，但对于低权重值的新闻的影响不大。



a. 新闻类别数 C 变化的影响



b. 新闻候选列表长度 L 变化的影响



c. 机制选择概率α变化的影响

图6　模型各参数对结果的影响

从结果分析来看，高权重新闻的连续发表时间间隔τ易受到规则参数的影响，而低权重新闻对于参数L、α的变化不敏感，但会受到新闻种类C的明显作用。

5　结 束 语

本文通过实证统计分析发现，热点新闻连续发表事件时间间隔分布在个类及总体层面上呈现带指数截断的幂律分布现象。为了揭示新闻选择背后的规律，本文提出了考虑时效性并基于严格优先及偏好优先混合机制的队列模型。通过数值模拟，该模型结果显示了丰富的非泊松时间间隔特性，可以得

到与实际数据在总体层面上新闻连续发表事件时间间隔分布较一致的结果。

需要注意的是，模型实际上假设了各个类别新闻的出现间隔是均质的，但是对新闻的选择使得发布的新闻的时间间隔出现了爆发性。这一机制揭示出这种人为选择的影响在新闻统计特性中扮演着重要角色，这对于理解各类媒体的行为特性有着重要的意义。该研究成果有助于深入理解新闻背后的选择机制，同时该工作能够被拓展到其他媒体的内容选择规则的研究上，如杂志、电影等，这将为进一步理解人类行为及信息传播提供契机。

本文的研究工作得到杭州师范大学科研启动经费项目(2015QDL005)的资助，在此表示感谢。

参 考 文 献

[1] LÜ L, CHEN D B, ZHOU T. The small world yields the most effective information spreading[J]. New Journal of Physics, 2011, 13(12): 123005.

[2] YANG J, COUNTS S. Predicting the speed, scale, and range of information diffusion in twitter[J]. ICWSM, 2010(10): 355-358.

[3] IRIBARREN J L, MORO E. Impact of human activity patterns on the dynamics of information diffusion[J]. Physical Review Letters, 2009, 103(3): 038702.

[4] DUTTA C, PANDURANGAN G, RAJARAMAN R, et al. Information spreading in dynamic networks[EB/OL]. (2011-12-02). http://arXiv.org/abs/1112.0384.

[5] IRIBARREN J L, MORO E. Branching dynamics of viral information spreading[J]. Physical Review E, 2011, 84(4): 046116.

[6] DOERR B, FOUZ M, FRIEDRICH T. Why rumors spread so quickly in social networks[J]. Communications of the ACM, 2012, 55(6): 70-75.

[7] LIND P G, DA SILVA L R, ANDRADE J J S, et al. Spreading gossip in social networks[J]. Physical Review E, 2007, 76(3): 036117.

[8] MONTANARI A, SABERI A. The spread of innovations in social networks[J]. Proceedings of the National Academy of Sciences, 2010, 107(47): 20196-20201.

[9] MIRITELLO G, MORO E, LARA R. Dynamical strength of social ties in information spreading[J]. Physical Review E, 2011, 83(4): 045102.

[10] PFITZNER R, GARAS A, SCHWEITZER F. Emotional divergence influences information spreading in twitter[J]. ICWSM, 2012(12): 2-5.

[11] CHEN Y Y, CHEN F, GUNNELL D, et al. The impact of media reporting on the emergence of charcoal burning suicide in Taiwan[J]. PloS One, 2013, 8(1): e55000.

[12] BARABÁSI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211.

[13] GONZALEZ M C, HIDALGO C A, BARABASI A L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196): 779-782.

[14] SONG C, QU Z, BLUMM N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018-1021.

[15] BROCKMANN D, HUFNAGEL L, GEISEL T. The scaling laws of human travel[J]. Nature, 2006, 439(7075): 462-465.

[16] MALMGREN R D, STOUFFER D B, MOTTER A E, et al. A Poissonian explanation for heavy tails in e-mail communication[J]. Proceedings of the National Academy of Sciences, 2008, 105(47): 18153-18158.

[17] OLIVEIRA J G, BARABÁSI A L. Human dynamics: Darwin and Einstein correspondence patterns[J]. Nature, 2005, 437(7063): 1251-1251.

[18] HONG W, HAN X P, ZHOU T, et al. Heavy-tailed statistics in short-message communication[J]. Chinese Physics Letters, 2009, 26(2): 028902.

[19] CANDIA J, GONZÁLEZ M C, WANG P, et al. Uncovering individual and collective human dynamics from mobile phone records[J]. Journal of Physics A: Mathematical and Theoretical, 2008, 41(22): 224015.

[20] 张开旭. 2012年新浪新闻语料[DB/OL]. [2013-01-10]. http://pan.baidu.com/s/1pJqrfPh.
ZHANG Kai-xu. The news corpus of Sina.com in 2012. [DB/OL].[2013-1-10]. http://pan.baidu.com/s/1pJqrfPh.

[21] EAGLET. Pan Gu Segment[EB/OL]. [2010-08-18]. http://pangusegment.codeplex.com.

编 辑 蒋 晓