ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.5, May 2009, pp.1384−1392
doi: 10.3724/SP.J.1001.2009.03512

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

博客网络中具有突发性的话题传播模型*

赵 丽[1+], 袁睿翕[1], 管晓宏[1,2], 贾庆山[1]

[1](清华大学 自动化系 智能与网络化系统研究中心,北京 100084)

[2](西安交通大学 智能网络与网络安全教育部重点实验室,陕西 西安 710049)

Bursty Propagation Model for Incidental Events in Blog Networks

ZHAO Li[1+], YUAN Rui-Xi[1], GUAN Xiao-Hong[1,2], JIA Qing-Shan[1]

[1](Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing 100084, China)

[2](Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China)

+ Corresponding author: E-mail: zhaoli04@mails.tsinghua.edu.cn, http://cfins.au.tsinghua.edu.cn/personalhg/zhaoli/

Zhao L, Yuan RX, Guan XH, Jia QS. Bursty propagation model for incidental events in blog networks. 2009, 20(5):1384−1392. http://www.jos.org.cn/1000-9825/3512.htm

Abstract: A discrete time dynamic model is proposed for bursty propagation of incidental events based on the node popularity and activeness in blog networks. The parameters of this model are clearly associated with the actual propagation and can reflect the characteristics of the dynamic propagation process. The model can provide a basis for predicting the trend of social events propagation in blog networks. Numerical testing is performed with the data from widely discussed events in Sina Blog, one of the most popular blogospheres in China in several months, and the results show that this model can emulate the actual event propagation and reflect the heavy tail phenomena of the decreasing propagation rate.

Key words: blog network; node popularity; node activeness; topic field strength; topic propagation

摘 要: 提出了一个基于节点知名度和活跃度的离散时间话题传播模型.该模型的参数具有明确的物理意义,可体现话题动态传播过程的特征,并可为话题传播趋势的预测研究提供依据.通过统计和分析中国最大的博客站点——新浪博客在几个月中若干具有突发性的事件引起的热门话题数据,结果表明,所提出的模型可以较为精确地再现话题的实际传播过程并体现传播速率的重尾现象.

关键词: 博客网络;节点知名度;节点活跃度;话题场强;话题传播

中图法分类号: TP393 **文献标识码**: A

博客(blog)是一种新型的具有开放性的互联网应用,最早出现于美国,2002 年进入中国,其用户数近几年迅

* Supported by the National Natural Science Foundation of China under Grant Nos.60574087, 60736027, 60704008 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.2007AA01Z480, 2007AA01Z475, 2007AA01Z464 (国家高技术研究发展计划(863)); the Program of Introducing Talents of Discipline to Universities of China under Grant No.B06002 (高等学校学科创新引智计划); the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No.20070003110 (高等学校博士学科点专项科研基金)

Received 2008-03-10; Accepted 2008-10-27

速增长[1].中国互联网络信息中心在 2007 年 12 月发布的《2007 年中国博客市场调查报告》[2]中指出,中国的博客作者数约有 4 698.2 万人,比 2006 年 8 月增长了 2 948.2 万人[3].作为 Web2.0(第二代互联网)时代的产物,博客以现代网络技术和通信技术为支撑,比传统媒体和第一代互联网在新闻传播的速度和空间上以及报道的广度和深度上更具优越性.特别是在对突发性事件的报道中,博客的时效性和现场感是传统媒体所不能比拟的[4].博客网络作为一种网络媒体,群众意见可以较自由地表达和传播,但"把关人"作用弱化[5],无根据的话题甚至谣言更容易产生和蔓延.因此,对博客网络中舆论传播规律和传播趋势预测的研究非常重要,它将有利于设计相应的机制对传播过程进行引导和控制等方面.

与博客网络中话题传播规律相关的研究主要包括分析博客网络拓扑结构和构建博客网络中话题传播模型,此类研究在国内外受到高度关注,已取得一定的成果,但均处于起步阶段.已有的研究主要分为 3 类:第 1 类研究主要针对博客网络中由好友链接形成的人际关系网络的拓扑结构特征,例如文献[6],研究结果表明,博客网络是一个度分布呈幂律分布的网络.此类文献没有涉及博客网络中信息的传播.第 2 类研究通过提出的某种合理算法推测隐式的信息传播网络拓扑结构.文献[7,8]基于页面内容相似度分析,得出信息在网页之间隐式的传播路径,以最终推断信息源.此类研究没有提出博客网络中信息传播的模型.第 3 类研究关注信息在博客网络中的级联传播特性[9,10].文献[9]参考传染病模型中的 SIS(susceptible-infected-susceptible)模型,构建了博客网络中信息的级联传播模型,通过仿真生成话题传播网络,与真实的日志间链接关系形成的网络特征较为吻合.文献[10]在传染病模型中的 SIRS(susceptible-infected-removed-susceptible)的基础上提出了话题在博客网络中的传播模型,给出一种计算节点间阅读概率和复制概率的算法,并考察了仿真网络中信息流量排名靠前的节点或边是否与真实情况吻合.此类文献分析了传染病模型与信息在博客网络中的传播的相似性,这对我们认识信息在博客网络中的传播规律有很大的帮助.但是上述文献中提出的模型并没有讨论单个话题的传播速率随时间变化的统计特性,研究结果不能应用于话题传播状态估计和发展趋势预测.

本文将研究博客网络中具有突发性话题的动态传播过程,构建可反映传播速率变化的离散时间话题传播模型,该模型的参数可以根据传播初期的数据拟合获得,进而预测话题未来的传播趋势,为舆情分析和预测打下基础.该问题的研究难点在于影响话题传播的因素复杂和真实数据的验证较为困难.首先,信息在博客网络中的传播受其他媒体的影响很大.本文中,我们引入外部话题场强表示发布话题信息的网页和电视等传统媒体对话题传播的影响;其次,我国注册的博客有 7 282.2 万个[2],如果跟踪这些博客每日更新的日志,则是难以实现的.本文中,我们通过搜索引擎得到话题相关日志的链接,进而得到日志的其他信息.

博客网络上话题的传播特性与话题内容关系很大:有些话题具有很强的突发性,例如新闻类话题;有些话题则会被持续地讨论甚至内容发生改变,例如技术类话题和生活类话题.本文仅关注具有突发性的单一事件话题的传播特性,此类话题内容比较单一,能够体现话题在网络中的传播特性,且往往时效性强,在短时间内可以形成激烈的讨论,之后受关注程度迅速衰减,传播过程中传播速率曲线在传播初期出现一次明显的尖峰.

1 具有突发性的话题传播模型

1.1 话题传播过程

博客的基本含义是"网络日志",相当于在网上所作的日记[4].博客主要包括按照时间顺序排列的日志(亦称帖子).通常,网络用户可以方便地通过在提供博客服务的网站注册成为博主,并在其个人博客上表达思想和转载信息.

本文中的话题是指与某事件相关的内容.突发性话题一般是指由新闻事件引起的话题.此类话题的信息源可能是博客,也可能是其他媒体.事件发生以后,一些博主在其博客中发布相关日志,成为话题的传播源.然后,一些博主通过阅读这些博客得知该话题并在自己的博客中以撰写日志的形式转载或发表评论,成为新的传播源.在下一时刻,话题又会传播到更多的博客.我们将具有事件关键词的日志称为话题的相关日志.在本文的研究中,我们不考虑未知话题信息或已知话题信息但不发表相关日志的博主,因为他们不影响话题的传播.

上述话题的传播过程是一个典型的级联传播过程[11].但是,由于博主有可能从其他网络媒体得知话题信息,

我们用外部话题场强来表示外部信息源对话题传播的影响,详细的建模将在第 1.2 节中叙述.

1.2 话题传播模型

首先引入封闭世界假设,即话题从博客网络中产生,并仅在博客网络中传播.这样我们可以定义传播模型.

定义无向图 $G=\{V,E,W\}$描述博客网络中某个话题的实际传播网络,其中,V 是博客节点集合.我们将博主及其博客站点统一地看作博客网络的节点.该节点可以发布新话题的日志,也可以通过访问其他节点的日志获得话题信息,并发布相关日志供其他节点访问.研究中,我们假设所有博客节点都是公开的,不考虑限制访问的节点.对于一个特定的话题,博客网络节点中仅有一部分参与讨论,而其他节点或者不知道该话题,或者知道了但不参与讨论,这些节点对话题的传播不起作用,因此本文的模型中不考虑这些节点.

E 是由所有连接博客节点的边组成的集合,代表话题可能的传播路径.由于博客网络相对于其他网络有自己的特点,不同于互联网等物理网络,博客网络是一个关系网络.由于搜索引擎、首页推荐的存在,不同博客站点之间可以方便地结识并互相访问,这使其区别于好友网络、电子邮件网络等一般关系网络.我们可以认为博客网络中任意两个节点之间都可以相互访问,博客网络是一个全连通的无向图.

W 代表博客节点的知名度集合,本文中假设信息的传播只与节点的知名度和活跃度有关.节点的知名度和活跃度分别表示节点在话题传播中所起的作用和节点参与讨论的可能性.节点的知名度是在话题传播过程中描述博客节点对其他博客节点影响力的非负实数.节点的知名度受多种因素的影响,如博客日常的访问量、是否被博客首页推荐、是否被其他博客站点链接、是否被搜索引擎检索等.节点的活跃度是描述博客节点是否经常访问其他节点并发文的活跃程度的非负实数.在博客网络中,节点的知名度与活跃度之间一般有关联.在话题传播网络中,由于知名度高的节点往往热衷于关注新鲜话题,常常在话题出现早期就在自己的博客中发布了该话题的信息供其他节点访问,说明在相同条件下,活跃度高的节点得知话题信息并参与讨论的概率也高.因此,为了简化模型,本文假设节点的知名度和活跃度相等.

基于上面的分析,可建立话题传播模型,描述博客网络人群对某个话题关心和参与的程度随时间变化的趋势.因此,我们希望建立发帖节点随时间变化的动态关系.本文讨论离散时间模型,t_0 表示初始时刻,$t_1,t_2,\ldots,t_n,\ldots$表示经过 1,2,…,$n$ 个单位时间后的时刻.定义 $I(t_n)$为 t_n 时刻处于已发帖状态的节点数,传播速率 $r(t_n)$表示$(t_{n-1},t_n]$时间段内由未发帖状态转变为发帖状态的节点数,则

$$I(t_n)=I(t_{n-1})+r(t_n) \tag{1}$$

显然,传播速率 $r(t_n)$具有不确定性.我们希望通过仿真得到 $r(t_n)$的统计特性,从而得到发帖节点数 $I(t_n)$的统计特性.

假设博客网络中参与某个话题讨论的节点总数为 N_E.节点 i 的知名度和活跃度均用 w_i 表示,其分布为 $p(w)$,服从广义的帕累托分布[12]:

$$p(w)=(1+\beta w)^{-1-\frac{1}{\beta}} \tag{2}$$

其中,β是分布的形状参数.广义的帕累托分布是幂律分布的一种形式,自变量的定义域是[0,∞].

节点可取两种状态:未发表话题相关日志的状态(未发帖状态)和已发表话题相关日志的状态(已发帖状态).处于未发帖状态的节点获知话题的信息后,很快发表相关日志变为已发帖状态,成为一个新的“传播源”.在 t_n 时刻,如果节点 i 处于已发帖状态,则状态不再变化;如果节点 i 处于未发帖状态,则以一定的概率由未发帖状态转变为已发帖状态.用示性函数$\delta_i(t_n)$表示节点 i 在 t_n 时刻所处的状态:

$$\delta_i(t_n)=\begin{cases}0, & \text{在时刻}t_n\text{,节点}i\text{处于未发帖状态}\\ 1, & \text{在时刻}t_n\text{,节点}i\text{处于已发帖状态}\end{cases} \tag{3}$$

假设 t_{n-1} 时刻节点 i 在处于未发帖状态,节点 j 处于已发帖状态,那么在(t_{n-1},t_n)时间段内,节点 i 从节点 j 获知话题信息并发帖的概率为

$$P_{ij}(t_n)=1-(1-\lambda)^{w_iw_j} \tag{4}$$

其中,w_i 表示节点 i 的活跃度,w_j 表示节点 j 的知名度,w_iw_j 为节点 i 受节点 j 影响的强度,也可视为单位时间 i 节

点访问 j 节点的次数.λ表示每次访问时获知话题信息并发帖的概率,是话题本身的特征,表示了话题的传播概率.由于话题在网络中节点之间传播是相互独立的事件,容易计算节点 i 从网络中处于已发帖状态的节点获知信息并发帖的概率为

$$\tilde{P}_i(t_n)=1-(1-\lambda)^{w_i\left(\sum_{j=1}^{N_E}\delta_j(t_{n-1})w_j\right)} \tag{5}$$

我们用 $B_1(t_n)=\sum_{j=1}^{N_E}\delta_j(t_{n-1})w_j$ 表示网络内部形成的话题场强,并由公式(3)可知,t_n 时刻网络内部的话题场强是网络中处于已发帖状态节点的知名度之和.

然后,我们放松封闭世界假设的约束.从上面的分析可以总结出,每一时刻的话题场强是所有已知话题节点的知名度之和.类似地,定义常数 B_2 表示网络外部节点形成的话题场强,并假设这个场强在传播过程中不变.

由于话题源节点的知名度在传播过程中是不变的,并且形成的话题场强存在于话题传播始终,我们将话题源节点看作外部节点.这样,初始时刻网络内部节点全部处于未发帖状态,信息从话题源和其他媒体组成的外部传入网络内部.容易计算,如果节点 i 在 t_{n-1} 时刻处于未发帖状态,则在 t_n 时刻变为已发帖状态的概率为

$$P_i(t_n)=1-(1-\lambda)^{w_i(B_1(t_n)+B_2)} \tag{6}$$

在每个单位时间内,处于未发帖状态的节点根据当前的内部场强 $B_1(t_n)$、外部场强 B_2 以及节点自身的活跃度 w_i 可计算出变为已发贴状态的概率 $P_i(t_n)$,并以此概率确定其状态是否发生改变.

已知网络中各节点 t_{n-1} 时刻的状态,可计算传播速率 $r(t_n)$的数学期望如下:

$$E(r(t_n))=\sum_{i=1}^{N_E}(1-\delta_i(t_{n-1}))P_i(t_n) \tag{7}$$

用 $r_1(t_n)$表示归一化的传播速率.$r_1(t_n)$与 $r(t_n)$的关系可用下面的公式来表示:

$$r_1(t_n)=\frac{r(t_n)}{N_E} \tag{8}$$

通过仿真计算可得 $r_1(t_n)$的均值和其他统计特性.用 N 表示仿真时间长度,即仿真从 t_0 时刻开始 t_N 时刻结束.一次传播过程的仿真步骤如下:

(1) 生成 N_E 个节点,并为节点标号.按照广义帕累托分布 $P(w)$为每个节点 i 生成知名度 w_i.设置外部话题场强 B_2,仿真时间长度 N 和初始时刻 n=1.

(2) 按照公式(6)计算每一个未发帖节点的发帖概率 $P_i(t_n)$,并以此概率判定节点 i 是否在第 n 个时间段,得知话题并发帖.统计从未发帖状态变为已发帖状态的节点数 $r(t_n)$.计算归一化的传播速率 $r_1(t_n)=r(t_n)/N_E$.

(3) $n=n+1$.如果 $n>N$,则仿真结束;否则,重复步骤(2)和步骤(3).

2 仿真结果与真实数据的比较和分析

2.1 博客中话题数据的获取

通过获取含有事件关键词的日志,可以得到几乎所有的对话题感兴趣的博客节点.我们选取 2006 年 8 月~2007 年 8 月参与讨论人数较多的两个娱乐事件(事件 1、事件 2)和两个政治事件(事件 3、事件 4)的数据作为研究样本.样本数据是从中国最大中文博客网站新浪博客(http://blog.sina.com.cn)获取的.为了获取新浪博客中讨论某话题的所有日志,我们用 Python 语言编写了一个网络爬虫脚本.该爬虫利用中文搜索服务全面、返回条目质量高的百度搜索引擎得到相关话题的日志地址,再通过日志地址进而得到日志发表时间和博主 ID 的信息.

由于百度搜索引擎的限制对每次搜索只返回页面排序前 760 的条目(Google,Yahoo 等搜索引擎也有类似的限制),我们采用多重搜索取并集的方法,尽可能多地抓取数据.该方法的数学表达式如下:

$$A=\bigcup_{T_i\in\Gamma}(\text{“}Topic+T_i\text{”的查询结果}) \tag{9}$$

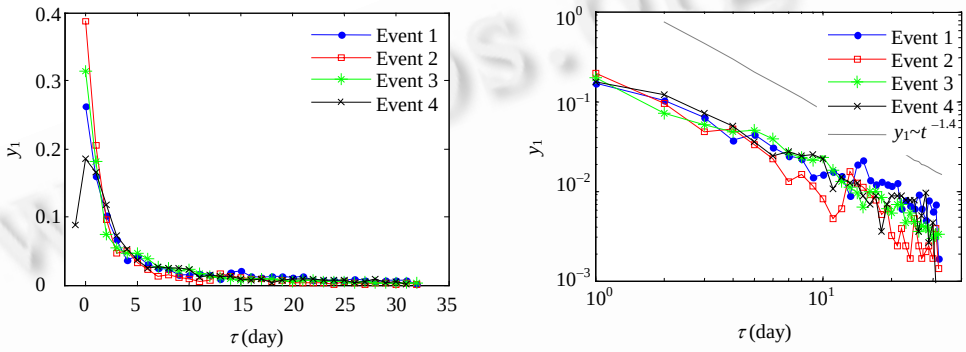
其中,A 是所有文章条目的集合;$Topic$ 是话题关键词,我们通过这个关键词搜索相关日志;T_i 是日志发表时间的区

间关键词;I是我们感兴趣的时间区间集合.虽然加入时间区间关键词后每次搜索返回的条目仍被限制在760条以内,但对所有使用时间关键词所得搜索结果取并集,即可得到几乎所有的查询结果.同时,我们在搜索关键词中加入搜索条件 site:blog.sina.com.cn,将搜索结果限制在新浪博客.通过上述方法,我们得到了很好的搜索结果,本文研究的4个话题的相关日志数分别是事件1:3 634,事件2:3 152,事件3:12 793,事件4:1 244.对返回结果抽样检测,有效条目占总返回条目数的比例分别是93%,99%,95%,99%.

得到相关话题的日志地址后,“爬虫”自动抓取博客页面,并通过解析页面的html文本得到日志发表时间、博主ID等信息.

2.2 数据表述及初步分析

通过上述方法得到的数据可以统计每日新帖数,即真实的话题传播速率 $y(t_n)$.类似于 $r(t_n)$的定义,用 $y(t_n)$表示(t_{n-1},t_n)时间段内新发表的帖数,$y_1(t_n)$表示归一化的真实话题传播速率.以天为单位时间,我们可以将4个事件的相关日志统计得到 $y_1(t_n)$绘制在一张图中,如图1所示.



(a) Normalized propagation rate of incidental events 1~4
(a) 事件1~事件4的归一化的传播速率

(b) Log-Log plots of normalized propagation rate
(b) 对数坐标下的传播速率

Fig.1 Normalized propagation rate of incidental events 1~4
图1 事件1~事件4的归一化话题传播速率

图1(a)中,4条曲线分别代表事件1~事件4归一化的传播速率.横坐标τ表示时间,以天为单位;纵坐标 y_1 表示归一化的话题传播速率.为了便于观察和比较,我们将图线平移,使得横坐标τ=0 的点表示曲线的峰值时刻,因此,图中的 $y_1(\tau)=y_1(t_n-t_{峰值})$.对于每一个事件,统计了事件发生后33天的日志数据.图1(b)与图1(a)的横、纵坐标及图例的意义均相同,但采用双对数坐标.

从图1(a)中可以明显地看出,4个事件虽然传播规模差别很大,但传播特征却非常相似:在事件发生后,短期内达到峰值,峰值之后又很快衰落下去.从图1(b)可以看出,话题中峰值后的传播速率随时间幂律下降,即 $y_1 \sim t^{-1.4}$,幂律下降的指数约为–1.4,这与文献[9]中统计的日志的URL传播速率的下降特征基本一致.

2.3 仿真与分析

模型中的未知参数分别是公式(4)中的传染概率λ,公式(2)中的帕累托分布参数β和公式(6)中的外部话题场强 B_2.定义均方误差准则函数为

$$J(\boldsymbol{a})=\frac{1}{N}\sum_{n=1}^{N}E(r_1(\boldsymbol{a},t_n)-y_1(t_n))^2 \tag{10}$$

其中,$J(\boldsymbol{a})$是要最小化的目标函数,$\boldsymbol{a}=(\lambda,\beta,B_2)$.$r_1(\boldsymbol{a},t_n)$是在参数 $\boldsymbol{a}$ 给定的情况下仿真系统的输出,$y_1(t_n)$是真实系统的输出,即归一化的每日新帖数.$J(\boldsymbol{a})$中的数学期望,理论上需要通过无穷次仿真取平均才可精确计算.实验中,我们用20次仿真的平均值作为 $E(r_1(\boldsymbol{a},t_n)-y(t_n))^2$ 的估计.对于每个事件,使得目标函数 $J(\boldsymbol{a})$最小的参数 $\hat{\boldsymbol{a}}$ 不能直接通过真实数据求得,需要先在各参数分量的合理取值范围内各任意取一个值,对模型进行求解,然后将解出的数

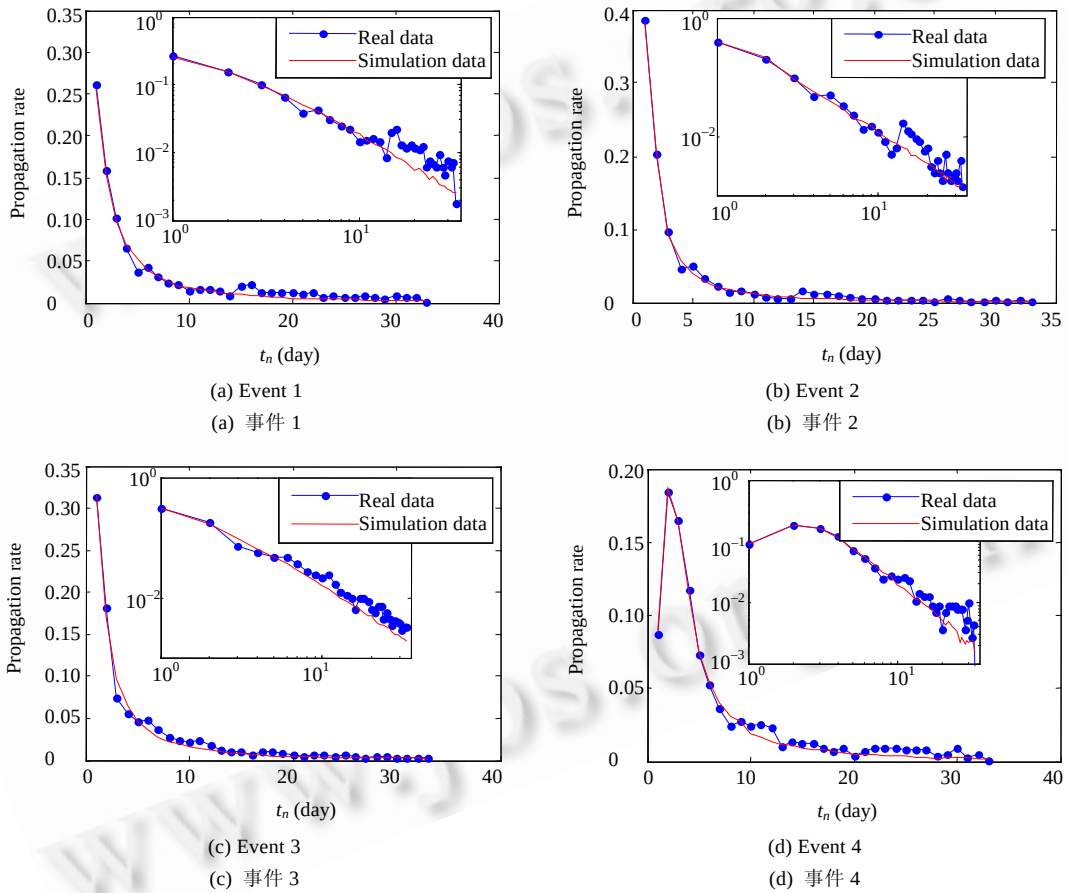
值解与实际统计数据进行比较,若两者存在差异,则调整这两个参数的值.重复上述工作,直至理论值与实际值趋于一致.我们对事件 1~事件 4 的真实数据在合理的参数取值范围内利用计算机程序估计的最优参数,结果见表 1.

Table 1 Parameter table

表 1 参数列表

Events	$\hat{\lambda}$	$\hat{\beta}$	$\hat{B}_2$	$J(\hat{\boldsymbol{a}})$
1	9.0×10^{-6}	0.66	22000	2.6×10^{-5}
2	1.0×10^{-5}	1.00	33000	1.6×10^{-5}
3	1.6×10^{-6}	0.86	151200	3.4×10^{-5}
4	1.6×10^{-4}	0.55	300	1.7×10^{-5}

4 个事件的数据拟合结果如图 2 所示.



(a) Event 1 (b) Event 2

(a) 事件 1 (b) 事件 2

(c) Event 3 (d) Event 4

(c) 事件 3 (d) 事件 4

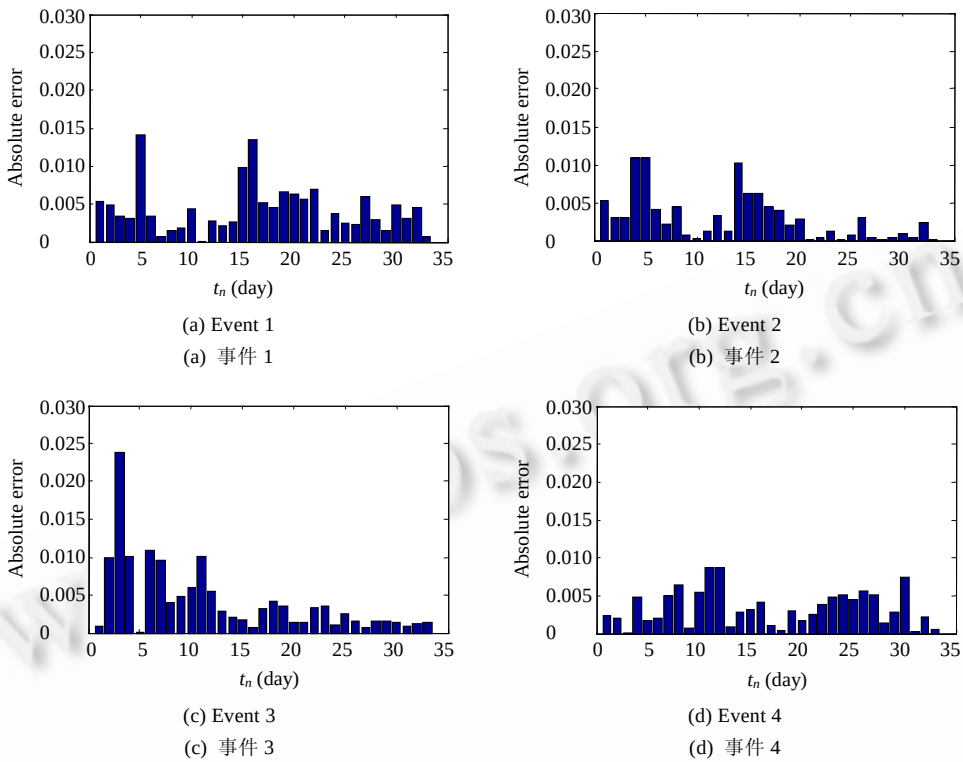
Fig.2 Comparison of simulation data and real data of events 1~4

图 2 事件 1~事件 4 的仿真数据与真实数据的比较

在图 2 中,对于真实数据,横坐标表示话题传播时间,纵坐标代表相关话题归一化的每日新帖数.仿真曲线横坐标 t 代表离散的时间点,纵坐标 y 代表在该时刻从未发帖状态的节点转变为已发帖状态的节点数目与总节点数的比值.各子图中的内嵌图的横、纵坐标含义与主图相同,但取对数坐标.

通过对比仿真结果与真实数据可以看出,本文提出的模型很好地拟合了真实数据,特别是真实数据的重尾现象.仿真数据与实际数据的绝对误差如图 3 所示.由比较结果可知,我们建立的模型可以很好地模拟话题传播

发展的趋势,是符合实际情况的.



(a) Event 1 (b) Event 2

(a) 事件 1 (b) 事件 2

(c) Event 3 (d) Event 4

(c) 事件 3 (d) 事件 4

Fig.3 Absolute error between simulation results and real data

图 3 模型仿真结果与真实数据的绝对误差

通过对 4 个事件参数的分析可以看出,β的取值比较稳定,说明新浪博客网络中节点知名度的分布与话题相对独立,其分布函数差别不大.4 个事件的λ和 B_2有明显差异,说明话题传播概率和外部话题场强不同.事件 3 的λ取值为 1.6×10^{-6}表明该话题的感染力度较弱,事件 4 的λ取值为 1.6×10^{-4}表明该话题的感染力度较强.事件 4 的 B_2值较小,说明该事件的外部话题场强较小.

该模型可以通过历史数据计算传播速率的变化趋势.为此,我们引入预测误差函数:

$$Er(m)=\frac{1}{N-m}\sum_{n=m+1}^{N}E(r_1(\hat{\boldsymbol{a}}(m),t_n)-y(t_n))^2 \tag{11}$$

其中,$\hat{\boldsymbol{a}}(m)$ 是根据前 m 个真实数据估计出的参数值,其理论值为

$$\hat{\boldsymbol{a}}(m)=\arg\min_{\boldsymbol{a}(m)}\left(\frac{1}{m}\sum_{n=1}^{m}E(r_1(\boldsymbol{a}(m),t_n)-y(t_n))^2\right) \tag{12}$$

实验中,我们通过与计算 $\hat{\boldsymbol{a}}$ 相同的方法计算 $\hat{\boldsymbol{a}}(m)$ 的估计值.$Er(m)$是通过 $\hat{\boldsymbol{a}}(m)$ 预测出的 t_{m+1} 之后的传播速率的均方误差.可绘制出通过不同时间长度的历史数据得到的预测结果的均方误差曲线,如图 4 所示.

从图 4 可以出,$Er(m)$大体上随 m 值的增大而减小,这说明使用的历史数据越多,预测越准确.通过 2 天的历史数据得到的预测均方误差 Er 值已达到 10^{-5} 的数量级,这与使用 33 天数据估计出的最小均方误差 $J(\hat{\boldsymbol{a}})$ 在一个数量级上,说明该模型具有较强的预测能力.

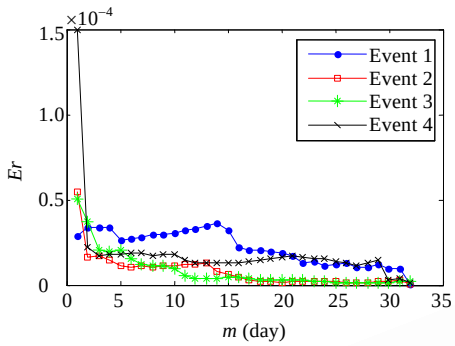


Fig.4 Prediction error $Er(m)$

图 4 预测误差 $Er(m)$

3 结 论

本文从博客网络中热门话题传播的机理出发,提出了一种传播模型.该模型通过引入节点知名度和活跃度刻画话题传播过程中节点起到的作用,通过外部话题场强表示话题传播源和其他传统媒体对话题传播的影响.对真实话题数据的统计发现,话题的传播速率曲线具有重尾现象.本文提出的模型不仅可以较好地拟合真实话题传播速率曲线,而且能够体现真实数据的重尾现象.该模型仅根据话题传播速率估计话题传播特征参数,可以通过数据传播初期的数据估计话题传播参数并预测话题传播趋势,为舆情分析和预测提供依据.

本文提出的模型作为话题在博客网络中传播规律的初步探索,其进一步的研究主要包括以下两方面:首先,可通过深入研究模型参数之间的相互关系和参数与博客节点的其他属性的关系来设计更有效的参数赋值方法;其次,可在本文模型的基础上建立可描述更复杂话题的传播模型.本文的模型仅适用于突发性话题的传播.现实中存在许多更复杂的事件,例如,由多个子事件组成的话题、外部话题场强是时变的等,这些话题不能直接采用本文的模型表示,但可以在本文模型的基础上建立更为恰当的话题传播模型.

References:

[1] Kumar R, Novak J, Raghavan P, Tomkins A. On the bursty evolution of blogspace. In: Proc. of the 12th Int'l Conf. on World Wide Web. New York: ACM Press, 2003. 159−178.

[2] China Internet Network Information Center. Research report of 2007 China blog market. Statistical Report, 2007 (in Chinese). http://www.cnnic.net.cn/html/Dir/2007/12/26/4948.htm

[3] China Internet Network Information Center. Research report of 2006 China blog. Statistical Report, 2006 (in Chinese). http://www.cnnic.net.cn/html/Dir/2006/09/25/4176.htm

[4] Zhao ZL. A clear thinking toward the prosperity of blog: A disseminative understanding of news blog. Journal of Nanjing University of Posts and Telecommunications (Social Science), 2006,8(2):23−26 (in Chinese with English abstract).

[5] Qu H. "Alienation" of online community and media. Southeast Communication, 2006,(3):45−47 (in Chinese with English abstract).

[6] Kumar R, Novak J, Raghavan P, Tomkins A. Structure and evolution of blogspace. Communications of the ACM, 2004,47(12): 35−39.

[7] Adar E, Zhang L, Adamic LA, Lukose RM. Implicit structure and the dynamics of blogspace. In: Proc. of the Workshop on the Weblogging Ecosystem, the 13th Int'l World Wide Web Conf. New York: ACM Press, 2004.

[8] Adar E, Adamic LA. Tracking information epidemics in blogspace. In: Proc. of the 2005 IEEE/WIC/ACM Int'l Conf. on Web Intelligence. Compiegne: IEEE Computer Society Press, 2005. 207−214.

[9] Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M. Patterns of cascading behavior in large blog graphs. In: Proc. of the SIAM Int'l Conf. on Data Mining. New York: ACM Press, 2007. 551−556.

[10] Gruhl D, Guha R, Liben-Nowell D, Tomkins A. Information diffusion through blogspace. In: Proc. of the 13th Int'l Conf. on World Wide Web. New York: ACM Press, 2004. 491−501.

[11] Nisan N, Roughgarden T, Tardos E, Vazirani V. Algorithmic Game Theory. Cambridge: Cambridge University Press, 2007. 613−614.

[12] Hosking JRM, Wallis JR. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 1987,29(3): 339−349.

附中文参考文献:

[2] 中国互联网络信息中心.2007 年中国博客市场调查报告.统计报告,2007.

[3] 中国互联网络信息中心.2006 年中国博客调查报告.统计报告,2006.

[4] 赵志立.博客"热"的"冷"思考——对新闻博客的传播学解读.南京邮电大学学报(社会科学版),2006,8(2):23−26.

[5] 瞿辉."网络社区"与媒介的"异化"现象.东南传播,2006,(3):45−47.

赵丽(1982－),女,山西太原人,博士生,主要研究领域为复杂系统,社会网络.

管晓宏(1955－),男,博士,教授,博士生导师,主要研究领域为网络化系统经济与安全性,网络信息安全.

袁睿翕(1965－),男,博士,教授,博士生导师,主要研究领域为复杂网络,无线通讯.

贾庆山(1980－),男,博士,讲师,主要研究领域为复杂网络化系统性能评价与优化,随机优化.