基于传染病动力学的博客舆情话题传播模型研究

丁学君[1,2]　梁昌勇[1]

(1. 合肥工业大学 管理学院,安徽 合肥 230009;

2. 东北财经大学 管理科学与工程学院，辽宁 大连 116025)

摘　要 以群组动力学为视角，本文将博客用户群划分为未知组群、易感组群、传播组群及免疫群组，建立了基于 USIR 的博客舆情话题传播模型。与之前的研究相比，该模型并不依赖于任何个体实证参数。仿真结果表明，该模型可以真实地描述博客网络中的舆情话题传播规律，并具有较好的预测性能。

关键词 博客，舆情话题，传播模型，传染病动力学

中图分类号 C931.6

1 引言

博客（Blog）是一个以网络为载体的个人信息发布系统。网络用户在提供博客服务的网站上注册成为博主（Blogger）后，在其个人博客空间中所作的日记称为博客帖子（Post）或“网络日志”[1]。据 CNNIC 发布的数据显示，截至 2014 年 6 月，我国博客用户规模为 1.22 亿，博客使用率为 19.3%。博客逐渐由早期的平民化、草根化向专业化、精英化转变，普通用户开始倾向于使用互动性更强的微博、微信等社交平台进行沟通和交互，而一些“超级博主”的博客、技术类博客却依旧保持着较高的点击率和影响力[2]。

然而，博客作为一种开放性的网络平台，允许公众自由地表达和传播思想及意见，使其与传统媒体相比，“把关人”的作用明显弱化。这就导致了舆情事件发生时，一些毫无根据的话题甚至谣言的滋生与蔓延。因此研究博客网络中舆情话题的传播规律，并对其传播趋势进行有效地预测，将为舆情分析人员对博客网络中舆情的监控和分析提供一定的理论依据。

2 相关研究

当前经典的信息扩散模型主要有独立级联模型（Independent Cascade Model，IC 模型）、线性阈值模型（Linear Threshold Model，LT 模型）、传染病模型（Epidemics Model）等。Kempe 等人将社会网络信息传播问题形式化为一种离散的优化问题，并在此基础上分别提出了 IC 模型、LT 模型以及加权级联模型（Weighted Cascade Model，WC 模型）[3]。Chen 等人

*基金项目：国家自然科学基金项目(71503033)、辽宁省社会科学规划基金项目(L14DGL045)、东北财经大学青年科研人才培育项目(DUFE2014Q56)

通信作者：丁学君，合肥工业大学管理学院，博士后；东北财经大学管理科学与工程学院，讲师，E-mail: dingxj812@163.com。

通过分析大量的博客数据，给出了影响博客信息传播的五个因素，进而提出了一个博客信息传播模型[4]。Kimura 等人则提出了一个基于社区结构的社会网络传播模型[5]。王萍在 LT 模型的基础上，基于选择性注意理论提出了一个改进的社会网络扩散模型[6]。郑蕾对 LT 模型进行了改进，提出了面向社会网络的多信息并行竞争传播模型[7]。Seung 等人则在 IC 模型的基础上，分析了博客网络中的信息扩散规律和博主的传播行为[8]。

随着在线社交网络的兴起，较多学者都把研究重心从传统社会网络转移到了在线社交网络，随之涌现了一些基于在线社交网络的信息扩散模型[9-19]。Budak 等人提出一个新的信息扩散模型 GLCM（Gaussian Logit Curve Model）来模拟用户的传播行为[9]。Jaewon 等人建立一个 LIM（Linear Influence Model）扩散模型来预测当前节点对其他节点的影响[10]。Bakshy[11]和 Mark[12]的研究则强调了弱关系在信息传播中担任更重要的角色。史文国等人提出了一个移动网络中基于交互式马尔科夫链的信息传播模型，并在模型基础上挖掘 top-k 节点[13]。Zeng 等采用隐性马尔科夫链模型（HMM），对网络舆情话题的生命周期进行建模并预测了其传播趋势[14]。此外，部分研究者通过构建 Agent-Based 或 Multi-Agent-Based 模型来研究社会网络中的信息扩散规律[15-19]。

由于信息在社会网络中的扩散，在某种程度上类似于传染病在人群中的传播。因此，部分研究者借鉴传染病扩散模型中的思想和方法，来构建社会网络中的信息扩散模型[20-34]。赵文兵基于 SIR 模型和 LT 模型的设计思想，给出了一个在线社交网络中的信息接受—浏览—分享模型（Accept-Browse-Share，ABS Model）[20]。Leskovec 等人基于传染病动力学中的 SIS 模型，构建了博客网络中的话题传播模型[21]。Gruhl 等人则以传染病动力学中的另一个模型—SIRS 模型为基础，建立了博客网络中的话题传播模型，同时给出了一种获取节点间阅读概率和复制概率的算法[22]。Zhou 等人提出了一个预测博客网络中话题讨论趋势的动态概率模型，并利用最大似然估计和主观经验来确定模型中的参数[23]。Zhao 等人提出了一个博客网络中突发性话题的传播模型，该模型以传染病动力学中的另一个仓室模型—SI 模型为基础，并在模型中引入了一个实证参数—个体适应度，但是精确地估计出个体适应度却十分困难[24]。赵丽等则在其构建的博客网络突发性话题传播模型中，考虑到了节点知名度、节点活跃度、话题传播源以及外部媒体对舆情话题传播速度的影响[25]。

以上模型虽然体现了突发性话题在博客网络中的传播规律，但却均需要依赖于博客用户的个体行为参数，而博客用户的个体行为信息通常是嵌入到海量的 Web 信息之中，例如新浪博客每天会产生几百万 MB 的数据量，且每天的内容都在不断更新，要从规模如此庞大的 Web 信息中挖掘出个体用户的行为特征，显然是十分耗时且难以实现的。Zhang 等人以博客及 BBS 网络中的热点话题传播作为研究对象，将博客用户群划分为目标热点话题讨论群组、相关话题讨论群组以及话题讨论退出群组，并分析了上述三类群组的状态转变过程[35]。然而，舆情话题通常是由突发性新闻事件所引发，其传播特点与一般性的热点话题相比存在明显差异，且博客网络作为在线社会网络的一种，其用户间的信息分享是信息传播的重要渠道之一。因此，Zhang 等人构建的模型无法更为真实地描述舆情话题在博客网络中的传播规律。本文基于传染病动力学中的 SIR 模型，构建了基于 USIR 的博客舆情话题传播模型。实验结果表明，该模型不仅可以很好地描述舆情话题的传播规律，且具有较好的预测性能。

3 模型建立

3.1 博客网络中的信息互动模式

博客网络的信息互动模式如图 1 所示，其中 A、B、C、D 分别表示博客网络中的不同用户节点。由于舆情话题通常是由某个突发性的新闻事件所引发，且媒体会在该事件的发生、发展以及演变过程中对其进行持续地报道，因此，博客用户往往会通过各种形式的外部媒体获知该话题。在图 1 中，用户 A 通过外部媒体获知了某一突发性新闻事件，并因为对该事件产生兴趣，而在其博客空间中发布相关日志，成为该舆情话题在博客网络中的一个传播源；用户 B 和用户 C 分别与用户 A 建立了双向“好友”关系和单向的“关注”关系，因此可以随时关注用户 A 的日志更新；如果用户 B 在浏览到用户 A 所发布的舆情话题相关日志后，对该话题产生兴趣，则其将会对该话题进行转载或评论，成为继 A 之后的下一个传播源；用户 D 虽然未与用户 A 建立“好友”关系或“关注”关系，但是作为一种开放式的个人网络出版平台，博客网站允许用户公开发表博客日志，因此用户 D 仍有机会以“访客”身份阅读到 A 所发表的相关日志，并在自己的博客空间转载或发表评论，成为该舆情话题的一个新传播源；此外，用户 A 还可以加入其感兴趣的博客圈，并与圈内好友就该舆情话题进行分享和讨论。上述过程描述了基于个体行为的博客网络舆情话题传播过程，而实际的博客网络是由成千上万个博客节点构成，各博客节点之间以上述模式进行信息的传递和分享，进而使信息在整个博客网络上以级联的方式进行扩散。

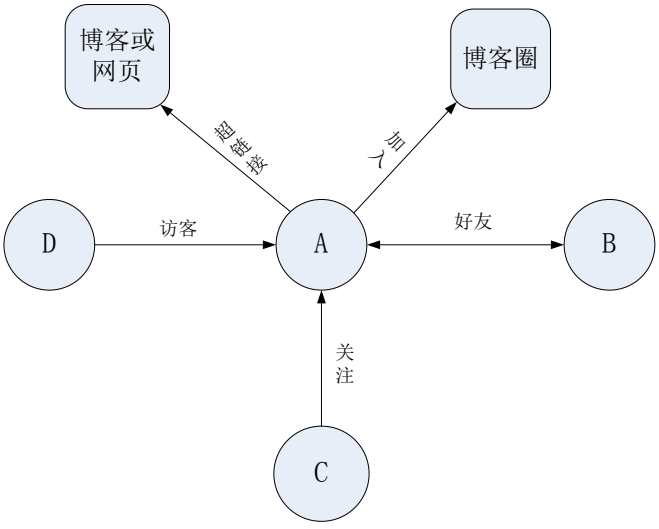


图 1 博客网络中的信息互动模式

3.2 SIR 模型

传染病动力学即是根据传染病的发生、发展及环境变化等情况，建立能反应其变化规律的数学模型，通过模型动力学性态的研究来显示传染病的发展过程，预测其流行规律和发展趋势，分析其流行的原因和关键因素，从而为人们制定传染病预防和控制决策，提供理论基础和数量依据。

在传染病动力学研究中，SIR模型是最为经典、应用最广泛的“仓室”模型之一，由Kermack和Mckendrik在研究黑死病的传播规律时所提出[36]。所谓仓室模型，就是针对某种

传染病将研究对象分成若干类，即若干个仓室。SIR模型的演化规则可以描述为：

（1）人群划分为以下三个仓室：易感类（Susceptible），由未染病者但有可能被传染的个体所组成的仓室；感染类（Infective），由已染病并具有传染力的个体所组成的仓室；免疫移出类（Removed），由疾病治愈后具有免疫力的个体所组成的仓室。

（2）在疾病传播期内，人群总人口数 N 不变，既不考虑生死，也不考虑人员迁移。

（3）设 $S(t)$，$I(t)$，$R(t)$ 分别为 t 时刻易感者、感染者以及免疫者在人群总人口中所占的比例，即三类不同状态个体的密度，则 $S(t)+I(t)+R(t)=1$。

（4）时间以天为计量单位，每个感染个体（病人）每天有效接触的平均人数为常数 λ，称为日接触率。当感染个体与易感个体（健康者）接触时，会使易感个体受感染变为感染个体。

（5）每天被治愈的感染者人数占感染者总数的比例为常数 γ，称为日治愈率，且被治愈的个体会获得免疫能力。

根据以上假设，每个感染个体每天可使 $\lambda S(t)$ 个易感个体转变为感染状态，因为人群中的感染人数为 $NI(t)$，所以每天共有 $\lambda NS(t)I(t)$ 个易感个体被感染，即易感人数 $NS(t)$ 的减少率为 $\lambda NS(t)I(t)$；每天共有 $\gamma I(t)$ 个感染个体被治愈并获得免疫能力，即免疫人数 $NR(t)$ 的增加率为 $\gamma I(t)$；则每天感染人数的变化率为 $\lambda NS(t)I(t)-\gamma I(t)$。SIR模型可以用以下微分方程组进行描述：

$$\begin{cases} \dfrac{dS(t)}{dt} = -\lambda I(t)S(t) \\ \dfrac{dI(t)}{dt} = \lambda I(t)S(t) - \gamma I(t) \\ \dfrac{dR(t)}{dt} = \gamma I(t) \end{cases} \quad (1)$$

3.3 基于 USIR 的博客舆情话题传播模型

研究表明，博客网络上的话题传播机制与传染病的传播机制类似[37]，因此本文以传染病动力学中的SIR模型为基础，来构建博客网络中的舆情话题传播模型。由博客网络中的信息互动模式可知，博客用户之间可以通过“好友”关系或“关注”关系，进行交流互动；用户可以利用“访客”身份对某篇公开发表的博客日志进行浏览、转载或评论；具有共同兴趣和爱好的博客用户可以加入“博客圈”，并与圈内好友进行信息分享。此外，由于舆情话题往往是由突发性的新闻事件所引发，使得博客用户可以通过外部媒体的报道获知该事件，并对相关新闻报道进行转载或评论，进而使相关舆情话题在博客网络中进行传播。因此，为了更为真实地反映舆情话题在博客网络中的传播规律，在构建博客网络舆情话题传播模型时，需要考虑舆情话题获取的两种主要渠道：外部媒体报道以及节点间的信息分享，对舆情话题传播过程的影响。

基于以上分析，本文将博客用户群划分为未知组群（Unknown group）、易感组群（Susceptible group）、传播组群（Infected group）以及免疫组群（Recovered group），构建了基于USIR的博客网络舆情话题传播模型。未知组群表示从未阅读过目标舆情话题的相关信息，即处于“未知”状态的用户集合；易感组群表示以好友、关注者或者访客身份阅读过目标舆情话题相关博客日志，即处于“易感”状态的用户集合；传播组群表示对目标舆情话题

发表日志（评论或转载），即处于“传播”状态的用户集合；免疫组群表示对目标舆情话题失去兴趣，再也不会发表任何相关博客日志，即处于“免疫”状态的用户集合。在舆情话题传播过程中，博客用户会以一定的概率在未知状态、易感状态、传播状态和免疫状态之间进行转换，从而使组群规模发生变化。

博客网络中的舆情话题传播规则定义如下：

（1）假设博客用户总数为N，且不考虑博客网络中用户的迁入与迁出，即N的值保持不变；

（2）设$U(t)$、$S(t)$、$I(t)$、$R(t)$分别表示t时刻博客网络中未知组群、易感组群、传播组群以及免疫组群的用户数在博客用户总数中所占的比例，则$U(t)+S(t)+I(t)+R(t)=1$；

（3）λ为接触率，对应于传染病动力学中的有效接触率，表示t时刻访问I中任意一个用户发表的舆情话题相关日志的平均用户数，则$\lambda NU(t)$表示t时刻访问I中的任意一个用户所发表的舆情话题相关日志的U中用户数。由于以上访问行为的发生，使得之前处于对舆情话题未知状态的用户获知该话题，即由话题“未知状态”转变为“易感状态”，从而移入易感组群。

（4）α为外部影响概率，即t时刻U中的用户通过外部媒体获知目标舆情话题，因对该话题产生兴趣，而就该话题发表相关日志的平均概率，即因为外部媒体的驱动，使得部分用户由话题“未知状态”转变成为“传播状态”，则$\alpha U(t)$表示t时刻传播组群的外部输入率。

（5）β为传播率，表示S中的用户因阅读过I中用户发表的与目标舆情话题相关的博客日志，对该话题产生兴趣，并在t时刻针对该话题发表日志的平均概率，即因为博客网络内部的关注关系，使得部分用户由话题“易感状态”转变成为“传播状态”，则$\beta S(t)$表示t时刻传播组群的内部输入率；

（6）γ为免疫率，表示I中的用户因为对目标舆情话题失去兴趣，而在t时刻终止对该话题发表相关日志的平均概率，即t时刻之后这些用户将再不参与话题的传播与讨论，则$\gamma I(t)$表示t时刻传播组群的输出率。

博客网络中四类用户组群$U(t)$、$S(t)$、$I(t)$和$R(t)$的状态转变过程如图2所示。

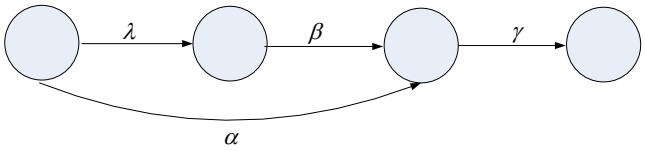


图2 博客网络中不同用户组群的状态转变图

根据以上假设可知，t时刻未知组群$U(t)$中共有$\lambda NU(t)I(t)$个用户进入易感组群$S(t)$，$\alpha NU(t)$个用户进入传播组群$I(t)$，即未知组群$U(t)$的用户变化率为$-\lambda NU(t)I(t)-\alpha NU(t)$；易感组群$S(t)$中共有$\beta NS(t)$个用户进入传播组群$I(t)$，则易感组群$S(t)$的用户变化率为$\lambda NU(t)I(t)-\beta NS(t)$；传播组群$I(t)$中共有$\gamma NI(t)$个用户进入免疫组群$R(t)$，则传播组群$I(t)$的用户变化率为$\alpha NU(t)+\beta NS(t)-\gamma NI(t)$；免疫组群$R(t)$的用户变化率为$\gamma NI(t)$。

根据图2所描述的四类不同用户组群的状态转变过程，构建基于USIR的博客网络舆情话题传播模型，如式（2）所示：

$$\begin{cases} \dfrac{dU(t)}{dt} = -\lambda U(t)I(t) - \alpha U(t) \\ \dfrac{dS(t)}{dt} = \lambda U(t)I(t) - \beta S(t) \\ \dfrac{dI(t)}{dt} = \alpha U(t) + \beta S(t) - \gamma I(t) \\ \dfrac{dR(t)}{dt} = \lambda I(t) \end{cases} \quad (2)$$

4 实验仿真

4.1 实验数据集

本文利用从新浪博客中实际抓取到的数据构建实验数据集，其构建过程如下：

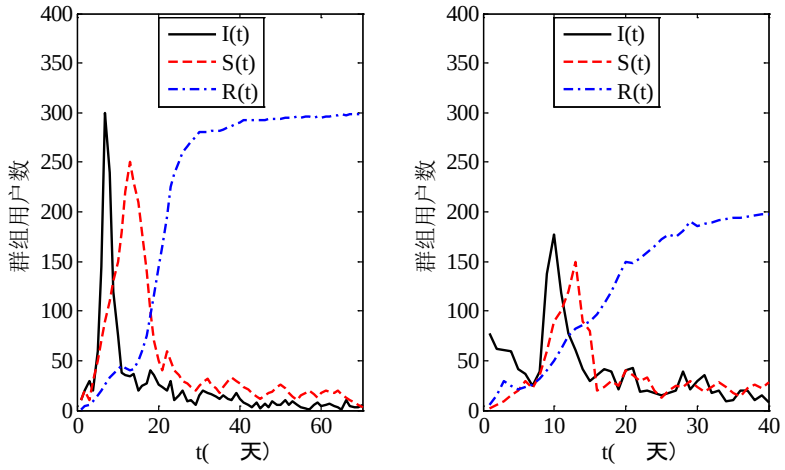
（1）选取 2011 年 3 月~2012 年 2 月期间，由突发性新闻事件所引发的 1 个政治类话题（话题 1）和 1 个社会类话题（话题 2）作为目标舆情话题，并分别提取出相应的话题关键词，如表 1 所示。

表 1 实验数据集描述

编号	话题	采样区间	话题关键词
1	本·拉登被击毙	2011 年 5 月 1 日~2011 年 8 月 1 日	本·拉登被击毙； 本·拉登死亡； 基地组织
2	“小悦悦”事件	2011 年 10 月 18 日~2012 年 1 月 18 日	小悦悦； 王悦； 被碾压女童

（2）根据表 1 给出的话题关键词，利用新浪博客的日志搜索功能，得到采样区间内 10249 条与话题 1 相关的博客日志；6508 条与话题 2 相关的博客日志。此外，本文所构建的博客数据集还包括：1）每天就目标话题发表日志的用户数量及 ID；2）每天浏览目标话题相关日志的访客数量、ID 以及所访问的博主 ID；3）曾在采样区间内发表目标话题相关日志，但之后终止对目标话题发表日志的用户数量、ID 以及最后一次发表日志的时间；4）在目标舆情话题采样区间内，新浪博客中用户总数的平均值。采样点设置为每天 22:00，采样周期 $T = 24$ 小时。

（3）对舆情话题原始数据进行筛选，去除其中的无关日志及重复条目，分别得到每个采样时刻未知组群 U、易感组群 S、传播组群 I 及免疫组群 R 的用户数量。图 3（a）、（b）分别描述了话题 1 和话题 2 的不同用户组群的实际用户规模随时间的变化关系。



(a) 话题 1　　(b) 话题 2

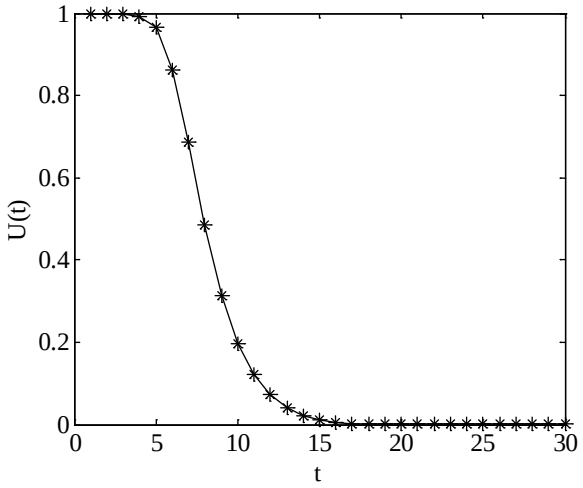
图 3 不同组群的实际用户规模随时间的变化关系

4.2 博客网络舆情话题传播模型的仿真实现

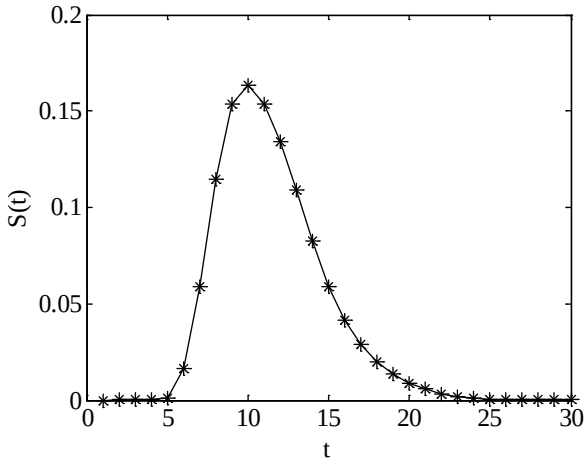
本文利用 Matlab 仿真软件对式（2）给出的基于 USIR 的博客网络舆情话题传播模型进行了实验仿真，并对仿真结果进行了分析。

（1）不同组群的用户规模随时间的变化关系

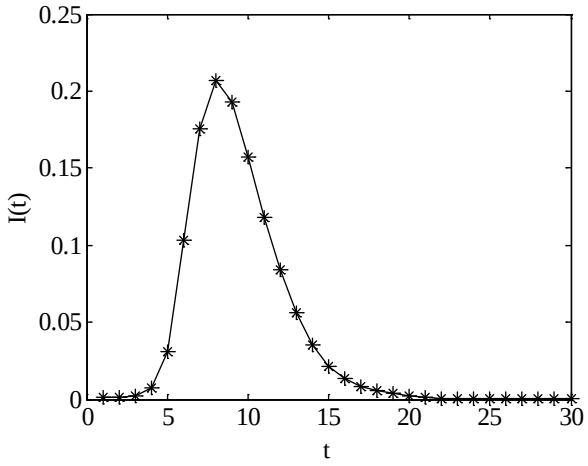
设博客网络的用户总数 $N=1000$，初始时刻博客网络中仅有一个用户节点处于传播状态，即传播组群规模 $I(0)=0.001$，其余用户全部为未知状态。式（2）中的各参数设置如下：接触率 $\lambda=0.5$，外部影响概率 $\alpha=0.2$，传播率 $\beta=0.1$，免疫率 $\gamma=0.5$。基于式（2）得到的未知组群 $U(t)$、易感组群 $S(t)$、传播组群 $I(t)$ 以及免疫组群 $R(t)$ 的用户规模随时间变化的关系如图 4 所示，实验迭代次数为 100 次。



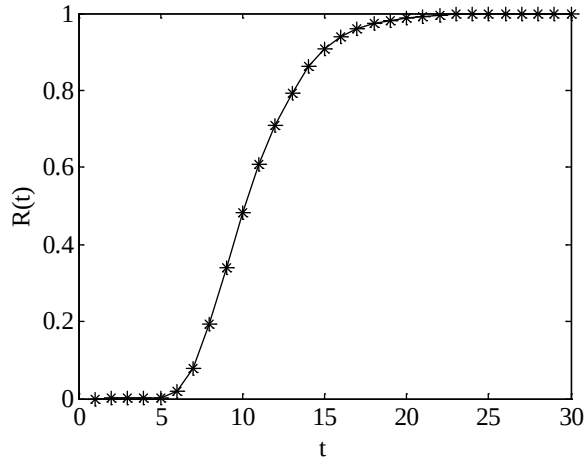
(a) 未知组群



(b) 易感组群



(c) 传播组群



(d) 免疫组群

图 4 不同组群的用户规模随时间的变化关系

由图 4 可知：在话题传播初期，网络中未知组群规模$U(t)$迅速减少，并最终趋近于 0；易感组群规模$S(t)$在话题传播初期增长缓慢，其后呈快速增加趋势，在其达到峰值后迅速递减，最终趋近于 0；传播组群规模$I(t)$与易感组群规模$S(t)$具有相似的变化趋势，即在传

播过程中也出现了一个明显的峰值，但 $S(t)$ 峰值的出现时刻（$t=10$）滞后于 $I(t)$ 的峰值出现时刻（$t=8$）；免疫组群规模 $R(t)$ 在话题传播初期增长缓慢，之后则呈现较快的增长态势，并在其达到峰值后趋于稳定，最终趋向于 1，即免疫状态为网络中的吸收状态。

由上述仿真结果可知，式（2）所描述的舆情话题传播过程与现实社会中舆情话题的演化规律相符，即与其他生活类和技术类话题相比，舆情话题通常由某个突发性的新闻事件所引发，其会在短时间内引发某些用户的强烈兴趣，从而形成激烈的讨论，并以“裂变式”的传播模式形成传播峰值；而随着时间的推移，由于话题的退化机制及用户的遗忘机制的共同作用，使得用户对该话题的兴趣度逐渐减退，并最终成为对该话题免疫的用户。

（2）接触率 λ 对传播过程的影响

图 5 给出了在其他初始条件保持不变的情况下，接触率 λ 取不同值时，传播组群用户规模 $I(t)$ 随时间的变化关系曲线。由图中可知：不管接触率 λ 取何值，传播组群用户规模 $I(t)$ 均在 $t=8$ 时出现峰值，且在 $t=23$ 时达到稳定状态，即接触率 λ 的值对传播组群规模的影响并不显著。这是因为 λ 表示未知组群中的用户对传播组群用户的平均接触率，当未知用户在浏览一篇博客日志后，虽然由未知状态转变成为了易感状态，但该用户是否会对话题进行传播，还受到其对舆情话题的兴趣度、发帖用户的知名度、用户间的亲密程度以及个体行为倾向等多方面因素的影响，是一个极其复杂的心理和行为决策过程。因此，λ 的改变对舆情话题传播过程不具有决定性的影响。

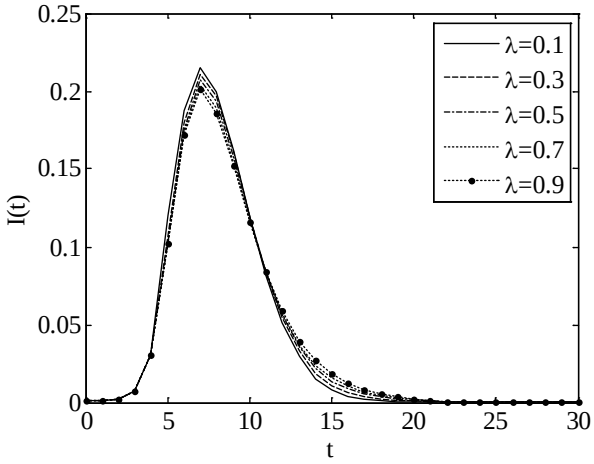


图 5 λ 取不同值时，传播组群用户规模随时间的变化关系

（3）外部影响概率 α 对传播过程的影响

图 6 给出了在其他初始条件保持不变的情况下，α 取不同值时，式（2）中传播组群的用户规模 $I(t)$ 随时间的变化曲线。由图中可知，α 取值越大，传播组群的用户规模就越大，且其达到稳定状态的时间（即弛豫时间）越长，即外部媒体对舆情事件的关注度会显著影响博客用户的舆情话题传播行为，这与赵丽等人的研究结论一致[25]。

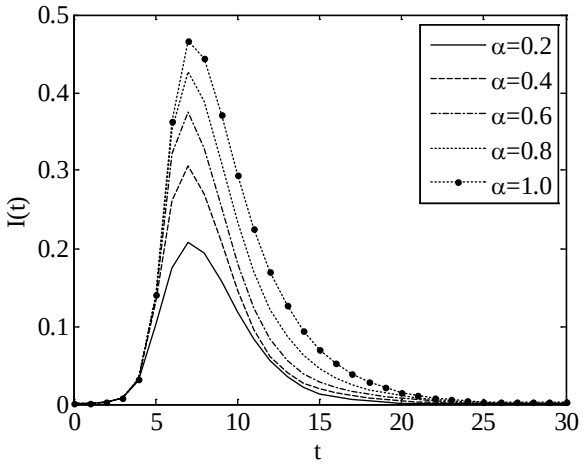


图 6 α 取不同值时，传播组群用户规模随时间的变化关系

（4）传播率 β 对传播过程的影响

图 7 给出了在其他初始条件保持不变的情况下，传播率 β 取不同值时，传播组群的用户规模 $I(t)$ 随时间的变化关系曲线。β 表示了易感用户因阅读传播用户发表的相关日志，而针对目标舆情话题发帖的概率，β 的取值越大，易感用户的发帖概率就越大，传播的次级联效应也因此增强，从而导致传播组群的规模及网络的弛豫时间随之增加。因此，图 7 给出的仿真结果与舆情话题的传播规律相符。

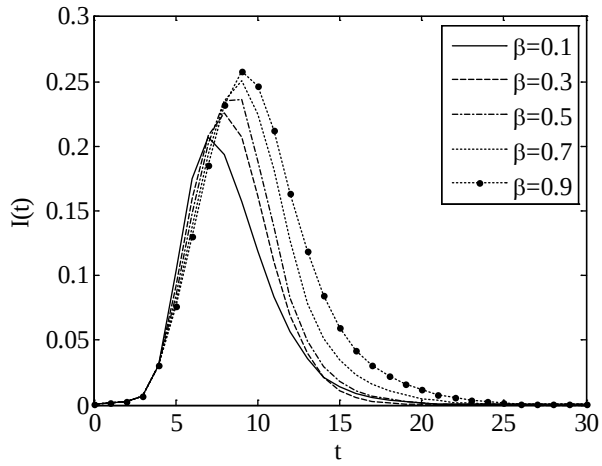


图 7 β 取不同值时，传播组群用户规模随时间的变化关系

（5）免疫率 γ 对传播过程的影响

图 8 给出了在其他初始条件保持不变的情况下，免疫率 γ 取不同值时，传播组群的用户规模 $I(t)$ 随时间的变化关系曲线。γ 表示传播用户因为对目标舆情话题失去兴趣，而终止发帖的概率。通过分析可知，γ 的取值越大，传播用户终止发帖的概率就越大，传播组群的用户规模也会随之减少，而网络的弛豫时间则越短。因此，图 8 给出的仿真结果与舆情话题的传播规律相符。

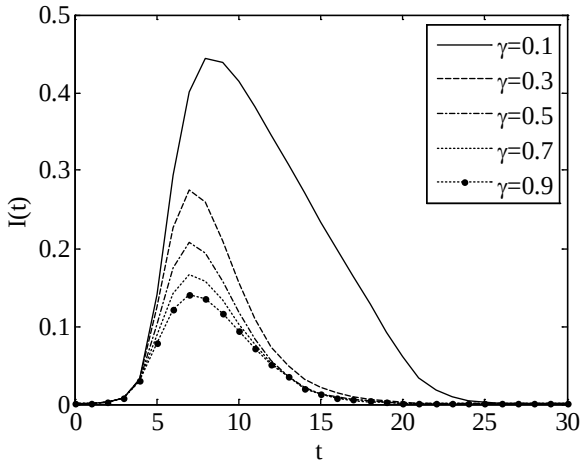


图 8 γ 取不同值时，传播组群用户规模随时间的变化关系

由图 4-8 可知：1）接触率 λ 取值的变化对传播组群用户规模的影响并不显著；2）外部影响概率 α 及传播率 β 的取值越大，传播组群的用户规模就越大，模型的收敛速度越慢，网络的弛豫时间越长，而传播的次级联效应也因此增强；3）免疫率 γ 的取值越大，传播组群的用户规模会随之减少，模型的收敛速度越快，网络的弛豫时间则越短，而传播的次级联效也会随之减弱；4）不管 λ 、 α 、 β 及 γ 取值如何，舆情话题传播网络都将最终演化到稳定状态，即式（2）给出的模型具有较好的收敛性和稳定性。

4.3 博客舆情话题传播趋势预测

（1）预测步骤

步骤 1：分别从新浪博客中提取出两个目标舆情话题的原始数据集。

步骤 2：对原始数据集进行去噪处理，并利用前文给出的用户数据集构造方法，分别构建话题 1 和话题 2 的实验数据集。根据本文给出的四类用户组群的定义，将原始数据集中的博客用户群划分为未知组群、易感组群、传播组群及免疫组群。以话题 1 为例，未知组群 U 表示对话题 1 毫不知情的用户集合；易感组群 S 表示访问过与话题 1 相关的博客日志的用户集合；传播组群 I 表示针对话题 1 发表日志的用户集合；免疫组群 R 表示在 $t-1$ 时刻针对话题 1 发表日志，却在 t 时刻终止发表相关日志的用户集合。本文首先从实验数据集中，提取出前 30 天的博客数据，构建训练数据集，对式（2）给出的博客网络舆情话题传播模型进行训练，并对模型参数进行拟合；利用第 31 天~第 90 天的博客数据，构建测试数据集，以对训练得到的模型进行检验。

步骤 3：利用样本数据集，计算式（2）中传播率 β 的值。 β 表示 S 中的某个用户因阅读过 I 中用户发表的与目标舆情话题相关的博客日志，从而在 t 时刻针对该话题发帖的平均概率。由于博客用户的发帖行为受到个体心理及行为特征、外部环境以及话题特征等多方面因素的影响，因此直接估计出 β 的值十分困难，但是却可以利用样本数据集，计算出传播组群的内部输入率 $\beta S(t)$ 的值，进而计算得到 β 的值。

步骤 4：利用样本数据集，计算式（2）免疫率 γ 的值。 γ 表示 I 中的用户在 $t-1$ 时刻针对目标舆情话题发表日志，却从 t 时刻开始不再发表相关日志的平均概率。由传染病动力学可知，$1/\gamma$ 表示感染者的平均患病期[38]，此处则表示舆情话题传播者的平均传播周期。与传

播率 β 一样，免疫率 γ 也受到个体心理及行为特征、外部环境以及话题特征等多方面因素的影响，因此直接估计出 γ 值也是极其困难的，但是却可以利用样本数据集，计算出传播组群的输出率 $\gamma I(t)$ 的值，进而计算得到 γ 的值。

步骤 5：利用样本数据集，计算式（2）中接触率 λ 的值。λ 表示 t 时刻访问 I 中任意一个用户发表的舆情话题相关日志的平均用户数。设 $\lambda_k(t)(k=1,2,...,NI(t))$ 为 t 时刻传播组群中的第 k 个用户发表的舆情话题相关日志的访客数，则利用式（3）可以计算出 t 时刻的接触率 $\lambda(t)$：

$$\lambda(t)=\frac{\sum_{k=1}^{NI(t)}\lambda_k}{NI(t)} \tag{3}$$

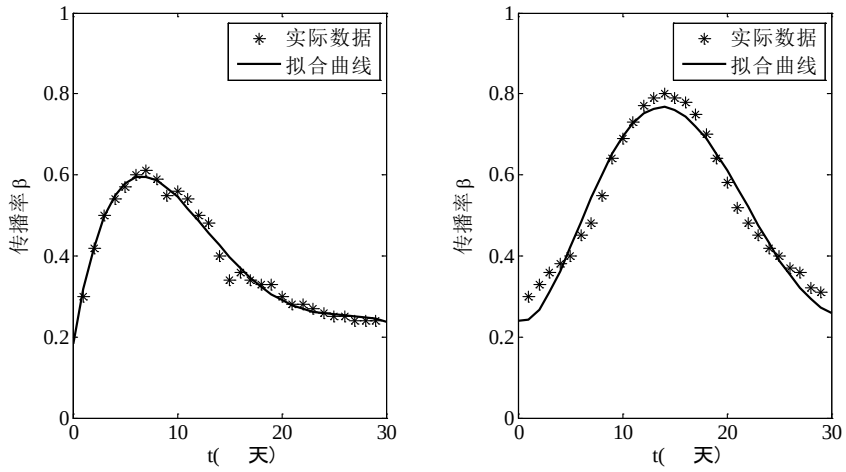
步骤 6：利用样本数据集，计算式（2）中外部影响概率 α 的值。α 表示 t 时刻 U 中的用户通过外部媒体获知目标舆情话题，并就该话题发表相关日志的平均概率。由于突发性新闻事件在演变过程中，会受到媒体的持续关注，大量的新闻报道和评论以多种形式出现在不同媒体平台上。因此，丰富的信息获取渠道使得直接估计出 α 的值变得十分困难，但却可以利用样本数据集，计算出传播组群的外部输入率 $\alpha U(t)$ 的值，进而得到 α 的值。

步骤 7：利用最小二乘法以及步骤 3~6 的计算结果，分别对话题 1 和话题 2 的传播率 β、免疫率 γ、接触率 λ 以及外部影响概率 α 进行拟合，得到相应的拟合函数 $\beta(t)$、$\gamma(t)$、$\lambda(t)$ 以及 $\alpha(t)$。

步骤 8：将步骤 7 中得到的拟合函数 $\beta(t)$、$\gamma(t)$、$\lambda(t)$ 及 $\alpha(t)$ 作为模型参数，代入式（2）给出的博客舆情话题传播模型，并利用该模型分别对话题 1 和话题 2 的传播组群规模进行预测，通过实际数据与预测数据的比较，检验模型的预测性能。

（2）预测结果及讨论

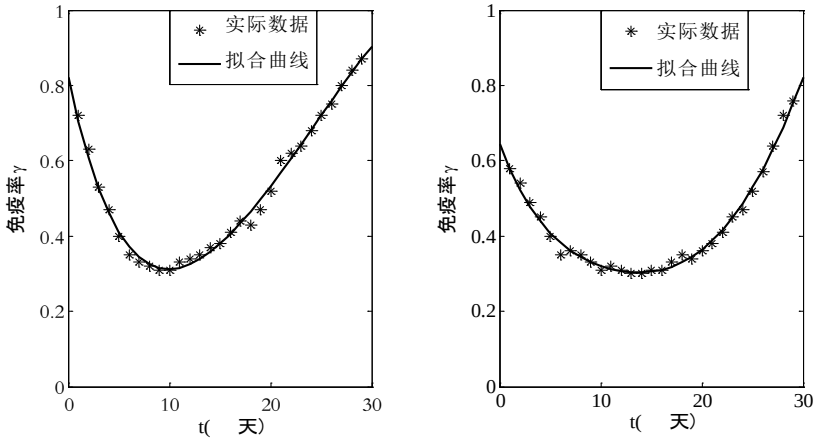
利用最小二乘法得到话题 1 和话题 2 的传播率 $\beta(t)$ 的拟合曲线，如图 9 所示。



(a) 话题 1　　(b) 话题 2

图 9 传播率的估计结果

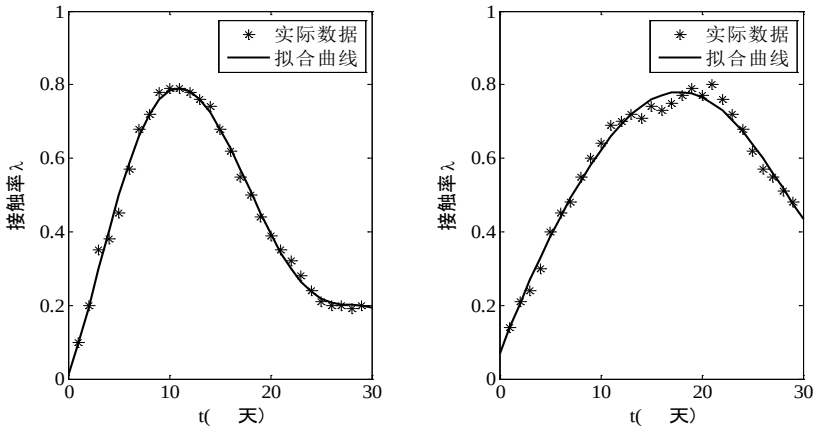
利用最小二乘法分别得到话题 1 和话题 2 的免疫率 $\gamma(t)$ 的拟合曲线，结果如图 10 所示。



(a)话题 1　　(b)话题 2

图 10 免疫率的估计结果

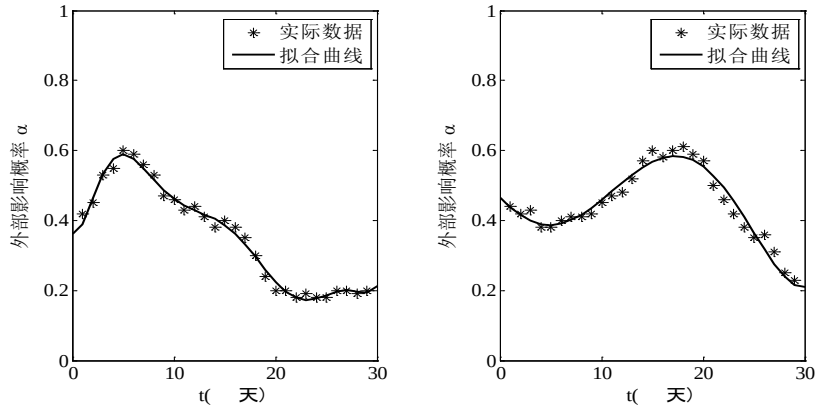
利用最小二乘法得到话题 1 和话题 2 的接触率 $\lambda(t)$ 的拟合曲线，如图 11 所示。



(a)话题 1　　(b)话题 2

图 11 接触率的估计结果

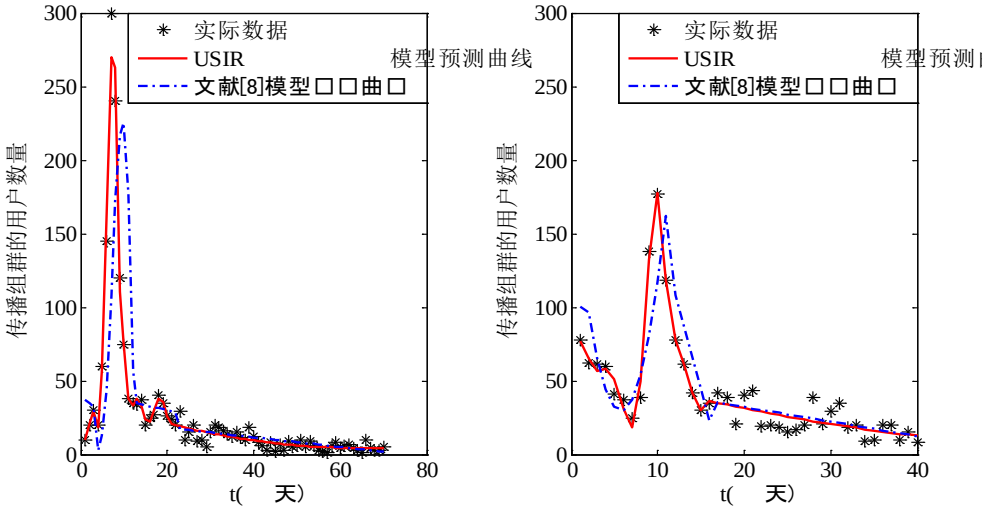
图 12 分别给出了利用最小二乘法拟合得到的话题 1 和话题 2 的外部影响概率 α 的估计结果。



(a)话题 1　　(b)话题 2

图 12 外部影响概率的估计结果

本文分别利用基于USIR的博客网络舆情话题传播模型及文献[8]给出的模型，对话题1和话题2的传播组群规模进行预测，其预测结果如图13所示。由图中可知，与文献[8]给出的模型相比，基于USIR的博客网络舆情话题传播模型，可以更为有效地预测舆情话题的传播趋势，且传播组群规模越大，预测的精度就越高。



(a) 话题 1　　(b) 话题 2

图 13 传播组群规模预测结果

5 结论

本文从博客网络中的舆情话题传播机理出发，基于传染病动力学中的 SIR 模型，构建了基于 USIR 的博客网络舆情话题传播模型。该模型根据博客用户获知舆情话题的两个主要渠道：用户间的信息分享及外部媒体报道，将博客用户群划分为未知组群、易感组群、传播组群及免疫组群，并分析了以上四类组群中的用户状态转变过程。与之前的研究相比，该模型以博客用户的组群行为作为研究背景，因此并不依赖于任何个体实证参数。

本文利用 Matlab 对提出的博客网络舆情话题传播模型进行了实验仿真，并分析了模型中各参数的变化对舆情话题传播组群规模的影响，仿真结果表明，该模型可以较为真实地描述博客网络中舆情话题的传播规律。此外，本文通过从新浪博客上抓取实际数据构建实验数据集，基于给出的 USIR 模型对两个目标舆情话题在博客网络中的传播趋势进行了预测，并与现有模型的预测效果进行了比较。实验结果表明，本文提出的基于 USIR 的博客网络舆情话题传播模型具有更好的预测性能。

参考文献

[1] Bandari R, Asur S, Huberman B. The pulse of news in social media: forecasting popularity[C]. The 6th International AAAI Conference on Weblogs and Social Media, Dublin: AAAI Press, 2012: 26-33.

[2] 赵志立. 博客“热”的“冷”思考—对新闻博客的传播学解读[J]. 南京邮电大学学报(社会科学版), 2006,

8(2): 23-26.

[3] Kempe D, Kleinberg J, Tardos é. Maximizing the spread of influence through a social network[C]. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC: ACM, 2003: 137-146.

[4] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris: ACM, 2009: 199-208.

[5] Kimura M, Saito K. Tractable models for information diffusion in social networks[C]. The 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin: Springer Berlin Heidelberg, 2006: 259-271.

[6] 王萍. 社会化网络的信息扩散研究[J]. 情报杂志, 2009, 28(10): 39-42.

[7] 郑蕾. 面向社会网络的信息传播模型研究[D]. 上海: 上海交通大学, 2011.

[8] Lim S H, Kim S W, Kim S.Construction of a blog network based on information diffusion[C]. Proceedings of the 2011 ACM Symposium on Applied Computing, New York: ACM, 2011: 931-941.

[9] Budak C, Agrawal D, EI Abbadi A. Diffusion of information in social networks: Is it all local? 2012 IEEE 12th international conference on data mining (ICDM). Brussels: IEEE Computer Society, 2012: 121-130.

[10] Yang J, Leskovec J. Modeling information diffusion in implicit networks. Proceedings of the 2010 IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2010: 599-608.

[11] Bakshy E, Rosenn I. The role of social networks in information diffusion[C]. Proceedings of the 21st International Conference on World Wide Web, New York: ACM, 2012: 519-528.

[12] Granovetter M S. The strength of weak ties[J]. American Journal of Sociology, 1973, 78(6): 1360-1380.

[13] 史文国, 王瑜. 移动社会网昂罗信息传播模型构建于 top-k 节点挖掘[J]. 计算机应用研究, 2012, 29(8): 2830-1832.

[14] Zeng J, Zhang S, Wu C, et al. Predictive model for internet public opinion[C]. IEEE 4th International Conference on Fuzzy Systems and Knowledge Discovery, Haikou: IEEE, 2007, 3: 7-11.

[15] 刘常昱, 胡晓峰, 罗批, 等. 基于 Agent 的网络舆论传播模型研究[J]. 计算机仿真. 2009, (1): 20-23.

[16] 梅珊. 基于复杂 Agent 网络的病毒传播建模和仿真研究[D]. 国防科学技术大学. 2010.

[17] 王战平. 基于 Agent 的网络传播危机信息识别与评价研究[J]. 图书情报工作. 2006, (12): 51-53.

[18] 蒋帅. 基于多 Agent 仿真的在线口碑传播网络形成机制研究[D]. 杭州: 浙江大学, 2010.

[19] 贺筱媛, 胡晓峰, 罗批. 基于 Agent 和 CPN 的 Web 信息传播系统建模研究[J].系统仿真学报. 2010, (03): 715-719.

[20] 赵文兵. Web2.0 环境下在线社交网络信息传播仿真研究[D]. 南京: 南京大学, 2013.

[21] Leskovec J, McGlohon M, Faloutsos C, et al. Cascading behavior in large blog graphs[C]. Proceedings of the SIAM International Conference on Data Mining, New York: ACM Press, 2007.

[22] 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: ACM, 2004: 491-501.

[23] Zhou Y. News Spreading Model Based on Micro-Blogging Platform in Network Era[M]. Informatics and Management Science VI, Springer London, 2013.

[24] Zhao L, Wang J, Chen Y, et al. SIHR rumor spreading model in social networks[J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(7): 2444-2453.

[25] 赵丽, 袁睿翕, 管晓宏, 等. 博客网络中具有突发性的话题传播模型[J]. 软件学报, 2009, 20(5): 1384-1392.

[26] 张彦超, 刘云, 张海峰, 等. 基于在线社交网络的信息传播模型[J]. 物理学报, 2011, 60(5):

50501-050501.
[27] 丁学君. 基于 SIR 的 SNS 网络舆情话题传播模型研究[J]. 计算机仿真, 2015, (1): 241-247.
[28] Xiong F, Liu Y, Zhang Z, et al. An information diffusion model based on retweeting mechanism for online social media[J]. Physics Letters A, 2012, 376(30): 2103-2108.
[29] 苑卫国, 刘云, 程军军, 等. 微博双向“关注”网络节点中心性及传播影响力的分析[J]. 物理学报, 2013, 62(3): 038901.
[30] 李青, 朱恒民, 杨东超. 微博网络中舆情话题传播演化模型[J]. 现代图书情报技术, 2013, 12: 74-80.
[31] 朱恒民, 刘凯, 卢子芳. 媒体作用下互联网舆情话题传播模型研究[J]. 现代图书情报技术, 2013, 29: 45-50.
[32] 钱颖, 张楠, 赵来军, 等. 微博舆情传播规律研究[J]. 情报学报, 2013, 31(12):1299-1304.
[33] 陈静. 复杂网络上基于流行病学的舆情传播模型及其规律研究[D]. 长春: 吉林大学, 2013.
[34] 丁学君. 基于 SCIR 的微博舆情话题传播模型研究[J]. 计算机工程与应用, 2015, 51(8): 20-26.
[35] Zhang B B, Guan X H, Khan M J, et al. A time-varying propagation model of hot topic on BBS sites and Blog networks[J]. Information Sciences, 2012, 187: 15-32.
[36] 王亚奇, 蒋国平. 复杂网络中考虑不完全免疫的病毒传播研究[J]. 物理学报, 2010, 59(10): 6734-6743.
[37] Zhao L, Wang Q, Cheng J, et al. Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(13): 2619-2625.
[38] Tang Y, Li W. Global analysis of an epidemic model with a constant removal rate[J]. Mathematical and Computer Modeling, 2007, 45(7): 834-843.

Research on Propagation Model of Public Opinion Topics in Blogs Based on Infectious Disease Dynamics

DING Xuejun, LIANG Changyong

(1. School of Management, Hefei University of Technology, Hefei 230009, China;
2. School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian 116025, China)

Abstract In the perspective of group dynamics, this paper divides all of the blog users into unknown group, susceptible group, infected group and recovered group, and then a public opinion topic propagation model based on USIR is built up in blogs. Compared with other topic propagation models in blogs, the model given by this paper does not depend on any individual empirical parameters. The simulation results show that the model established in this paper can well describe the propagation law of public opinion topics in blogs, and can predict the trend of the public opinion topics efficiently.

Key words Blog, Public opinion topics, Propagation mode, Infectious disease dynamics

作者简介

丁学君（1978- ），女，合肥工业大学管理学院博士后，东北财经大学管理科学与工程学院讲师，博士，辽宁辽阳人，研究方向包括信息系统、社会计算、物联网等. E-mail: dingxj812@163.com。

梁昌勇（1965- ），合肥工业大学管理学院教授、博士生导师，安徽肥西人，研究方向包括信息系统、企业信息化、决策理论和方法等. E-mail: cyliang@163.com。