基于自激点过程的网络热点话题传播模型

韩忠明 张 梦 谭旭升 段大高 司慧琳

(北京工商大学计算机与信息工程学院 北京 100048)

摘 要 建模互联网中热点话题的传播过程具有重要的意义和价值,该文以网络热点话题为研究对象,基于自激霍克斯过程提出一个话题传播模型(Self-Exciting Point Process Model,SEPPM). SEPPM 利用用户参与话题的自激效应,将话题传播过程建模为一个随机点过程. 同时,SEPPM 也考虑了话题传播的外部因素,综合形成话题传播模型. 为了验证该模型的有效性,该文从仿真和实证两个角度分别进行了大量的实验比较,提出话题仿真算法,仿真结果说明 SEPPM 可以生成多种符合热点话题传播特征的模式. 实际数据上的结果说明 SEPPM 不仅能够很好地拟合真实话题的传播过程,还能够有效地预测话题传播趋势.

关键词 社会媒体;热点话题;传播模型;霍克斯过程;自激点过程;社交网络;数据挖掘

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2016.00704

An Efficient Topic Propagation Model Based on Self-Exciting Point Process

HAN Zhong-Ming ZHANG Meng TAN Xu-Sheng DUAN Da-Gao SI Hui-Lin

(*School of Computer and Information Engineering, Beijing Technology and Business University, Beijing* 100048)

Abstract Modeling propagation processes of hot topics on Internet has significant meaning and value. This paper focuses on modeling hot topics on Internet and proposes a topic propagation model (Self-Exciting Point Process Model, SEPPM) based on self-exciting Hawkes process. SEPPM models the propagation process of one topic as a random point process by using self-exciting effect of user participation. At the same time, SEPPM also takes external factors for propagation into account, thus puts forward a formal topic propagation model. To evaluate effectiveness of SEPPM, comprehensive simulation and empirical experiment are conducted. A simulation algorithm for SEPPM is proposed and the simulated results show that SEPPM can generate a variety of patterns of hot topics with different propagation characteristics. The experimental results on real datasets show that SEPPM can not only splendidly fit real topic propagation process, but also can effectively forecast spreading trend.

Keywords social media; hot topics; propagation model; Hawkes process; self-exciting point process; social networks; data mining

---

收稿日期:2014-10-08;在线出版日期:2015-06-21. 本课题得到国家自然科学基金(61170112)、教育部人文社会科学研究青年基金(13YJC860006)及中央财政支持地方高校发展专项资金人才培养和创新团队建设项目(19005323132)资助.**韩忠明**,男,1972 年生,博士后,副教授,中国计算机学会(CCF)会员,主要研究方向为互联网数据挖掘、大数据分析与信息检索. E-mail: hanzm@th.btbu.edu.cn. **张 梦**,女,1991 年生,硕士研究生,中国计算机学会(CCF)会员,主要研究方向为互联网数据挖掘、信息检索. **谭旭升**,男,1989 年生,硕士研究生,主要研究方向为互联网数据挖掘、社区发现. **段大高**,男,1976 年生,博士,副教授,主要研究方向为社会计算、多媒体信息处理. **司慧琳**,女,1972 年生,博士,讲师,主要研究方向为大数据与多媒体信息处理.

1 引　言

交互式社会媒体是用户传播各类消息最主要的平台，其发展的速度已经超越了人们对它的预想，越来越多的用户都参与到交互式网络中，并直接性引发话题的产生，成为话题发展的推动力. 分析和建模热点话题的传播过程，可以了解热点话题的传播规律，监控热点话题的发展趋势，实现早期预测和检测热点话题，并对控制话题的传播起着重要的意义.

热点话题多数是基于 TDT 技术演变发展的. 话题检测与跟踪(Topic Detection and Tracking，TDT)的研究最早起源于 1996 年，美国国防高级研究计划署 DARPA(Defence Advanced Research Projects Agency)根据个人需求，提出研发一种新技术，能在没有人工干预的情况下自动判断新闻数据流的话题，这样就诞生了 TDT 的研究课题[1]. 随着网络用户规模的急剧扩大，信息传播量日益剧增，一些热点事件很容易在交互式媒体平台中展现出来. 因此，话题检测研究也受到学者们广泛地关注，成为研究重点. 文献[2]通过 TDT 技术先对信息进行分类整理，然后对每个类别使用聚类技术进行话题跟踪，可以快速了解各热点事件. 文献[3]提出基于相关性的方法，根据用户浏览行为和话题受关注的程度来发现热点话题，并使用复杂网络理论分析追踪模型.

相比较发现话题，人们有时候更关注的是某一个话题的传播过程. 由于热点话题发展趋势与话题背后的人类行为有关，而人类行为难以检测与量化，除此之外还与媒体的推动、话题本身的吸引力等因素有关，因此理解和建模热点话题的传播过程存在较大困难. 本文根据自激点过程的基本思想，融合热点话题传播特征，提出了自激点过程话题模型(Self-Exciting Point Process Model，SEPPM)，采用随机过程的建模思想对话题事件进行时间序列分析，并且进行模拟仿真以及话题预测实验.

本文第 2 节回顾相关研究工作；第 3 节介绍模型的相关定义与假设条件；第 4 节详细描述模型；第 5 节进行实验与结果分析；第 6 节是总结与后续工作的展望.

2 相关工作

随着复杂网络的产生，传染病模型也受到了广泛的研究与分析. 在传染病动力学中主要沿用 SIR (Susceptible-Infective-Removal)模型进行建模，它一共包含 3 种划分的固定人群，其中 S 表示易受感染者，I 表示被感染者，R 表示恢复者. 在 t 时刻单位时间内，易感染者以 β 的概率受到病毒感染，已感染者以 γ 的概率从传染病中恢复健康. 通过研究分析发现，网络中的用户群体与人类社会中的群体相似，网络中的话题传播与人类社会中的疾病传播也具有高度的一致性，因此采用传染病模型理解与分析网络中话题消息传播具有广泛的应用. 采用传染病模型理解消息传播时一般将消息作为一种传染病，而全体网络用户作为感染源，话题的传播就转化成疾病感染问题. 许多学者通过改进传染病模型，应用到复杂网络话题传播研究与分析中. 例如，根据 SIR 模型检测网络中多重信息源[4]，在真实的信息来源未知的情况下，提出一种适用于普遍网络的启发式算法估计信息来源的数量. 通过变换经典的 SIS (Susceptible-Infective-Susceptible)和 SIR 模型，提出一种 SII(Susceptible-Infectious-Immunized)模型来建立例如邮件蠕虫或者 Facebook 蠕虫的动态传播[5]，分析并预测其对网络的潜在破坏力. Satoshi[6] 聚焦于网络上的谣言传播现象，基于 SIR 模型建立了多动机信息传播模型，使人们明确区分正确信息与错误谣言. Edward 等人[7] 基于 SIS 模型，考虑到用户在面对话题时产生对其持有怀疑观点或延迟一段时间再参与讨论的情况，提出了 SEIZ(Susceptible-Exposed-Infected-Skeptical)模型来表达网络中话题与谣言的传播过程. SpikeM 模型为了保证模型服从幂律下降的规律，假设已感染节点的感染力度随着时间长久而降低的函数服从指数为 -1.5 的幂律分布[8]. 但是 SpikeM 也存在一些缺点：(1) 假设用户最多参与同一个话题一次，在实际情况中一个用户可以参与相同或者不同话题多次；(2) 传播链上每个节点的影响力没有区分开，每个节点的影响力假设相同. 但是在实际传播链上影响力高的节点远远比影响力低的要重要.

信息级联模型[9] 表示将一个网络中的事件看成由一系列级联活动组成. 每一个用户通过观察其他用户的行为来决定自身的行为. Gruhl 等人[10] 对独立级联模型进行了改进，提出一种计算节点间阅读概率和复制概率的算法，给每一条边定义了一个行为发生概率，使此模型更加适用于传播过程中可能发生延迟的真实社会网络. Saito 等人[11] 为了深入考虑时间延迟的影响，进一步提出了持续时间延迟

独立级联模型和持续时间延迟线性阈值模型,模型中使用了一个连续时间轴,为图中每条边添加了时间延迟参数.冀进朝等人[12]根据节点之间不仅存在衰减影响,也存在增强或无变化影响,在独立级联模型上提出了完全级联传播模型.通过分析信息级联模型的规则发现,其传播模式必然呈现出一个上升和下降的过程,不会出现波动现象,通常用于分析节点影响力.

随机点过程可以有效描述随机事件的产生过程.很多随机现象发生的时刻、地点、状态等可以用某一空间上的点来表示.自激点过程[13]描述的是当前事件的发生概率依赖于以往事件的发生情况.该过程可以看做一个满足典型随机过程的泊松过程,它的特点是话题传播趋势呈幂律上升状态,某个时刻的话题强度不是一个常数,而是依赖于自身和之前的传播趋势.该过程的自激效应表示此刻话题受到的关注度(参与数)不仅依赖于话题影响力和用户特性,还依赖于自身以及过去话题传播趋势的影响,也可以去影响后续话题事件的传播.也就是说,如果之前有较多的评论、转发者,那么此刻也会有很多的评论、转发者,继而影响之后有更多的用户关注话题.自激点过程的核心思想就是一次事件的产生将提升后续事件的产生概率,导致的结果就是事件的发生在时间序列上呈现簇状分布.该过程被提出后便广泛应用于建模自然和物理现象.Crane 等人[14]用带参数的自激霍克斯泊松过程对 YouTube 视频评论进行建模,模型结果显示评论的热度呈现幂律上升,幂律下降的特征.文献[15]分析了部分国家的恐怖袭击事件过程,假设恐怖袭击事件时间分布情况服从一个自激霍克斯过程,由于事件过程中有很多的缺失信息,所以在自激霍克斯过程中加入了约束条件,这样实现网络中复杂事件的自激霍克斯过程参数估计以及事件预测.

3 符号定义和假设

本文主要研究对象为交互式社会媒体中的热点话题.任意一个用户在网络上都可以参与话题讨论,使其成为一个新的传播源,吸引更多的用户发现话题并参与讨论.每个热点话题由很多事件形成,一个话题事件的产生从初始时刻开始,形成一个生命周期.关注的用户人数越多,话题受到的关注度就越高.

首先定义模型中的基本概念:

(1)节点.参与话题讨论并产生事件的用户称为节点;

(2)话题.在微博、论坛等上被用户讨论的内容或主题,比如"郭美美事件"就是一个讨论的话题;

(3)事件.每个用户参与一次话题的评论、转发形成一次事件,对于一个话题,在各个时刻所有参与该话题讨论的用户,会产生很多事件;

(4)热度.对于一个话题,在一定时间间隔内产生事件的数量称为该话题在此时间间隔内的话题热度,热度越高表示话题受到用户的关注度越高;

(5)热度序列.一定时间范围内的热度值构成该话题热度时间序列.根据话题的热度序列,可以画出热度值随着时间变化的热度时间序列图,反映热点话题受到关注的发展趋势;

(6)强度函数.一个话题在每个时刻瞬时的热度值比率,用 $\lambda(t)$ 表示.强度函数与热度呈正比关系.

影响每个用户参与话题讨论的因素:

(1)外部因素 μ.话题背景发生率,即话题本身对用户的吸引度,话题背景发生率越高,该话题对用户的吸引力就越大.为了模型分析的简单化,假定针对一个话题而言,该话题的背景发生率是一个不变的常数,不同的话题背景发生率不一样;

(2)自激效应.如果之前有较多的用户参与话题讨论,会吸引更多的用户参与讨论:

① k_0 表示过去时刻的话题热度对当前时刻话题热度影响的缩放因子;

② ω 表示话题热度的衰减率,过去时刻节点产生的事件量影响当前时刻产生事件量的衰减速度,随着 ω 的增加,过去时刻的事件数量并没有导致当前时刻有更多的用户参与话题讨论;

③ p 为衡量背景发生率和自激效应的相对大小,若背景发生率在一个话题传播过程中占有 p 的影响,那么该话题受到 $(1-p)$ 的自激效应影响.

4 SEPPM 话题传播模型

热点话题的传播过程可以看成用户参与话题讨论的随机过程.一个用户初始在某一时刻发布了一个消息,随后一些用户开始关注此消息并进行评论或转发(产生事件),这些新的评论和转发成为新的传播源,其他用户会受到这些已参与用户的影响而对话题产生了兴趣进行讨论,导致产生了激励效应,当参与的事件越来越多,就形成了热点话题.该过程反映一个话题产生的事件数目是怎样随着时间的推移而发生变化的.本节我们首先进行模型描述,然后

引入模型的参数求解和仿真算法.

4.1 SEPPM 模型描述

不同主题的热点话题本身对不同用户的吸引力度是不一致的，用户本身也会受到其他用户行为的影响来改变自身的决策状态. 因此，本文根据热点话题的特征以及用户行为特点，综合构建热点话题传播模型，描述交互式热点话题中话题受到用户的关注度随着时间推移的变化趋势. 每个时刻内产生话题事件数目越多，该话题受到用户的关注度就越高. 假设在 t_0 时刻产生一个话题，话题初始时刻没有用户关注讨论，那么在 t_1 时刻就会有用户发现此消息并且对其进行评论、转发，参与话题的讨论，产生事件 $\{P_1,P_2,P_3\}$，以此类推，在 t_2 时刻产生事件 $\{P_4,P_5\}$，在 t_3 时刻产生事件 $\{P_6,P_7,P_8,P_9\}$. 为了得到话题受到的关注度随着时间的变化趋势，在每个时刻都计算该时刻产生事件的总数量 N_t. 然后，可以得出 t_1,t_2,t_3 时刻分别产生 3,2,4 个事件，这样就形成了一个事件在各个时刻的随机计数过程，该过程简单描述如图 1 所示. 通过该计数过程，可以得出各个时刻产生的事件数量. 如果此刻的事件数量大小不仅依赖于过去时刻产生的事件数量，还与话题本身的吸引度以及外部影响因素有关，那么这样一个过程就是本文提出的自激点过程.

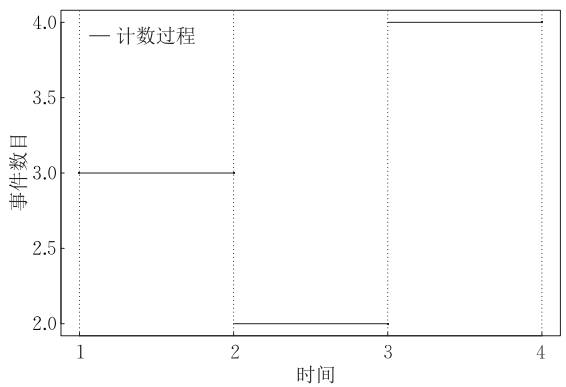


图 1　点的计数过程

假设 $\{N_t,t\geq 0\}$ 是一个随机点计数过程，它表示在 $(0,t]$ 的时间区间内，在每个时间 $t_1,t_2,\cdots,t_{N_t}$ 上产生事件的数量 N_t. 以新浪微博为例，N_t 表示一个主题名为“微博实名制”的热点话题在时间 $(0,t]$ 内产生事件的个数，其中 $t\in[0,+\infty)$，$S_1,S_2,\cdots,S_{N(t)}$ 表示用户参与该主题的讨论而产生的话题事件，即微博用户对该热点话题进行评论、转发等讨论. 假设在 t_0 时刻产生了一个特定主题的热点话题，那么此刻就会有一些用户关注，对话题进行评论或转发，产生话题事件数目为 N_0. 以此类推，得到产生的话题事件满足一个随机计数过程 $\{N_t,t\geq 0\}$，该过程描述的是在每个时刻用户产生事件的数量，它满足如下性质[13]：

$$P\{N_0=0\}=1;$$

对任意实数 $t\geq 0$ 和 $h\geq 0$，有

$$P\{N_{t,t+h}=1\mid N_t,S_1,S_2,\cdots,S_{N(t)}\}=\lambda(t,N_t,S_1,S_2,\cdots,S_{N(t)})h+o(h)\quad(1)$$

$$P\{N_{t,t+h}\geq 2\mid N_t,S_1,S_2,\cdots,S_{N(t)}\}=o(h)\quad(2)$$

由式(1)、(2)得出强度函数

$$\lambda(t,N_t,S_1,S_2,\cdots,S_{N(t)})=\lim_{h\to 0}h^{-1}P\{N_{t,t+h}\geq 1\mid N_t,S_1,S_2,\cdots,S_{N(t)}\}\quad(3)$$

称式(3)为此过程 $\{N_t,t\geq 0\}$ 的过程强度，一个不仅依赖于事件数目，还依赖于时间 t，并且满足式(1)和(2)的随机计数过程为自激点过程，由于此过程由霍克斯(Hawkes)等人[13]提出，也称为自激霍克斯过程. 其中，构造强度函数 $\lambda(t,N_t,S_1,S_2,\cdots,S_{N(t)})$ 的目的是为了表示该过程 $\{N_t,t\geq 0\}$ 在不同时刻中话题受到用户关注的强度大小，它随着时间的推移与产生的事件数目的不同而变换，通常简写为 $\lambda(t)$. 从式(3)中可以看出 $\lambda(t)$：它和无穷小时间增量的乘积给出当给定过程在 $(0,t]$ 中的产生的事件数和发生时间时，在无穷小区间 $[t,t+h)$ 有事件发生的条件概率. 因此给定了自激点过程的强度，也就规定了它在无穷小区间中有事件产生的条件概率. 在此过程上，强度函数越高，表示话题受到用户的关注度就越高，越来越多的人都参与话题讨论，进行评论、转发，进而提高话题事件数量.

该过程本质上是一个满足典型随机过程的非齐次泊松过程，它在各个时刻的话题强度函数并不是一个常数，同时该强度函数不仅是时间的函数，还是生成事件过程整个过去的函数，即在时刻 t 以前出现的所有话题事件的数目 N_t 以及产生事件的时间 $t_1,t_2,\cdots,t_{N_t}$，能够对继 t 之后的所有事件的数目及其发生时刻产生影响. 也就是说，话题被用户参与讨论产生事件的数量在此刻的状态和之前一段时间内该话题被参与讨论产生事件的数量是有依赖关系的，之前有很多用户都关注了该话题，并对其进行了评论、转发，那么此刻关注到该话题的新用户会看到了因为有许多人讨论它，所以也会评论、转发该话题，前导事件的产生会影响后续事件的改变，已发表过的评论的影响力也会随着时间推移呈现幂率下降[16]的状态，即用户对话题的关注度会变弱，参与讨论话题的用户数目会减少，话题的事件数目也会降低. 自激点过程模型受两个因素影响.

(1)外界因素,该热点话题对于用户的吸引度的强弱,如果吸引更多用户参与该话题的讨论,那么外界因素的影响就比较大.为了模型的简单化,对于同一个话题而言,话题的吸引度是一个固定不变的常量,不相同的话题对用户的吸引度不一样.

(2)前导事件的影响,即以往话题的评论、转发数量.如果在过去的时间里有很多的用户关注了该话题,并且参与了讨论,那么此刻的用户就会产生激励效应,也就是此刻的用户受到之前用户的影响,对该话题产生了较高的兴趣度,也参与了讨论.

构造自激点过程需要定义其强度函数,表示用户参与话题讨论的热度随着时间的变化趋势.假设当前事件线性依赖于已往事件,则该话题线性强度函数为

$$\lambda(t)=\mu+k_0\int_{-\infty}^{t}\nu(t-s)\mathrm{d}N_s=\mu+k_0\sum_{t_i<t}\nu(t-t_i,\omega) \tag{4}$$

其中,霍克斯提出了一种指数函数,用来表示话题事件数目随着时间推移的衰减函数:

$$\nu(t)=\sum_{j=1}^{Q}\omega_j\mathrm{e}^{-\omega_j t}1_{R_+} \tag{5}$$

式(5)表示话题事件数量是随着时间 t 指数增加或减少的.其中 1_{R_+} 表示在正实数的时间序列上存在一个点过程,也就是有事件产生;ω 表示话题在过去时刻产生事件的数量对当前时刻产生事件量的衰减速度;Q 代表维数,使模型简单考虑,本文取 $Q=1$,只考虑一个话题在时间序列上事件数目的变化情况.把式(5)代入到式(4)中可以变换成:

$$\lambda(t)=\mu+k_0\sum_{t>t_i}\omega\mathrm{e}^{-\omega(t-t_i)} \tag{6}$$

式(6)中 μ 为话题对参与讨论的用户的吸引强度,强度越大表示更大量节点对话题进行转发、评论,从而产生更多的事件,本文中假设随着时间的变化针对不同话题对用户的吸引强度是一个符合指数函数的随机数;k_0 是一个缩放因子,表示已受到影响的节点对当前时刻节点的影响,在同一个话题中是一个常数;ω 为衰减率,随着 ω 的增加,前导事件对后续事件的影响会迅速降低;$\nu(t-t_i,\omega)$ 为衰减函数,表示过去时间$(t_0,t_1,\cdots,t_{i-1})$讨论话题生成的事件对当前热度过程上事件的影响力度.话题的强度过程主要由两部分组成,一部分是话题事件不受任何影响的传播过程,此部分在整体事件传播过程中以概率 p 存在,另一部分表示事件是受到自激过程的影响,导致话题是有影响的传播,以概率$(1-p)$存在.那么式(6)可以写为

$$\lambda(t)=p\mu+(1-p)k_0\sum_{t>t_i}\omega\mathrm{e}^{-\omega(t-t_i)} \tag{7}$$

其中 p 表示背景发生率与自激效应相对大小.对于该话题的传播过程中,用户此刻关注话题的因素,有概率 p 是因为事件本身对用户的吸引率的影响,有$(1-p)$的概率是因为自激效应的产生,即看到之前有许多用户关注了该话题,也对这个话题产生了兴趣,并参与话题的讨论.式(7)表达了 SEPPM 基本过程,模型通过 p,μ,k_0,ω 这 4 个参数控制话题传播过程.

4.2 SEPPM 参数求解

SEPPM 是一个随机过程,受到 4 个参数的控制.在拟合和预测实际话题传播过程中,需要学习 4 个参数的值,观察一个话题在一段时间范围内的传播过程,得到话题样本数据,本文采用极大似然估计法进行参数求解.

通过式(6)得到自激点过程的一般形式:

$$\lambda(t)=\mu+k_0\sum_{t>t_i}\omega\mathrm{e}^{-\omega(t-t_i)} \tag{8}$$

该强度函数的对数似然估计函数为

$$\log(L)=\sum_{i=1}^{n}\log(\lambda(t_i))-\int_0^{\mathrm{T}}\lambda(t)\mathrm{d}t \tag{9}$$

求解参数为 μ,k_0 和 ω,使似然函数最大化:

$$\max_{\mu,k_0,\omega}\sum_{i=1}^{n}\log(\lambda(t_i))-\int_0^{\mathrm{T}}\lambda(t)\mathrm{d}t \tag{10}$$

求解强度函数的完整模型:

$$\lambda(t)=p\mu+(1-p)k_0\sum_{t>t_i}\omega\mathrm{e}^{-\omega(t-t_i)} \tag{11}$$

因为该模型由两个过程组成,一个是基于背景发生率的过程,另一个是受自激效应影响的过程,因此需要存在一个概率 p,表示背景发生率与自激效应相对大小,n 表示发生事件的数目.

分别对 p,μ,k_0 和 ω 求偏导,使得

$$\begin{cases}\dfrac{\mathrm{d}\lambda(t)}{\mathrm{d}p}=0,\ \dfrac{\mathrm{d}\lambda(t)}{\mathrm{d}\mu}=0\\ \dfrac{\mathrm{d}\lambda(t)}{\mathrm{d}k_0}=0,\ \dfrac{\mathrm{d}\lambda(t)}{\mathrm{d}\omega}=0\end{cases} \tag{12}$$

求解出参数 p,μ,k_0,ω 相对应的值.

4.3 仿真算法

SEPPM 分析实际话题时需要学习 4 个参数的值,为了验证该模型能否产生出与实际话题传播一致的过程,本文提出一个仿真算法.仿真算法目的是为了在没有实际话题传播数据时,人工提供参数值,构造虚拟的传播过程,来评价模型能否呈现出话题不同的传播形态.

本文采用细化过程来构造仿真算法，模拟出在各个时刻话题事件数量变化的过程，从而实现 SEPPM 模型过程. 细化过程(Thinning Procedure)[17]是一种可以拟合有界强度上的随机点过程算法，其基本思想是将热点话题中的事件作为点过程中各个时刻随机产生的点，每个时刻记录点的数量就是话题事件的数量，通过细化过程仿真对于任意连续的遵循泊松分布标准的点过程. 本文模型进行话题仿真，可以呈现出与实际话题相似的传播过程，通过调节参数便可呈现话题只有一个高峰或者有多个完整高峰的不同形态.

仿真算法的基本思想如下.

考虑一个拥有强度函数为 $\lambda^*(t)$ 的一维非齐次 Poisson 过程，在时间区间 $[0,T)$ 上存在事件的数量为 $N^*(T_0)$，并且满足参数是 $\mu_0^* = \int_0^{T_0} \lambda^*(s)\mathrm{d}s$ 的 Poisson 分布. 假设 $0 \leq t < T_0$，$\lambda(t) \leq \lambda^*(t)$，循环 i 次，使得 $i=1,2,\cdots,N^*(T_0)$，删除那些概率为 $1-[\lambda(t_i^*)/\lambda^*(t_i^*)]$ 的点，然后保持所有点遵循一个在区间 $[0,T)$ 上拥有强度函数为 $\lambda(t)$ 的非齐次 Poisson 过程.

仿真过程中的参数物理意义如下.

lambdas：产生一个参数为 λ 的泊松过程；

T_i：表示截取的时间区间为 $[0,T_i)$，研究在此时间区间上的话题传播趋势；

k：把时间区间一共划分为 k 个小区间，然后在每个 a/k 的小区间上产生非齐次泊松过程来计算在当前区间上的强度函数 $\lambda(t)$；

n：为了计算在划分完 k 个区间后每个小区间上的强度函数，把当前第 j 个区间再划分成 n 份，在每一份小区间上，都满足参数为 λ 的泊松过程，选择强度最大的值作为当前区间上的强度函数.

仿真算法的具体过程如算法 1 所示. 首先进行初始化，假设当前过程强度并没有受到自激效应的影响，完全受到话题背景发生率的影响；然后在时间 $[0,T)$ 内产生均匀分布的随机事件；每次循环都计算每个当前小区间上的强度值大小，删除不符合要求的点，然后取最大的值作为当前区间上的强度函数，通过这样一个细化过程，求出每个时刻对应的强度函数. 最后生成这样一个在时间 $[0,T)$ 内的话题仿真过程.

算法 1.　SEPPM 仿真算法.

输出：在时间 $[0,t)$ 内的仿真过程

1. INITIALIZATION: SET $\lambda_j \leftarrow \mu, t=0, J=1, I=0$;
2. FIRST EVENT: $U \rightarrow u_{[0,1]}, X \leftarrow -1/\lambda_i \log(U)$
3. IF $t+X>t$, GO TO STEP 8;
4. ELSE $t \leftarrow t+X$;
5. NEW EVENT: $U \rightarrow u_{[0,1]}, X \leftarrow -1/\lambda_i \log(U)$
6. IF $U \leq \lambda(t)/\lambda_j$, THEN $I=I+1$, $S(I)=t$;
7. ELSE RETURN STEP 2.
8. IF $J=K+1$, STOP;
9. $X=(X-t_j+t)\lambda_j/\lambda_{j+1}$, $t=t_j$, $J=J+1$;
10. RETURN STEP 3.
11. OUTPUT: RETRIEVE THE SIMULATED PROCESS $\{t_n\}$ ON $[0,T)$

5　实验结果与分析

5.1　实验设计与环境

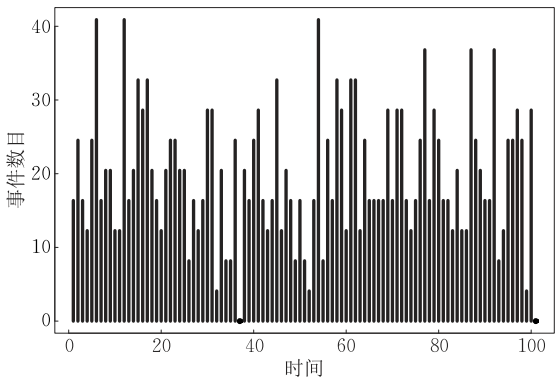
本文从仿真实验和真实数据实验两个角度评估模型效果. 仿真实验目的是研究模型参数和话题传播模式之间的关系；真实数据实验则可以对传播模式进行拟合，以拟合结果来检验与矫正模型，并对话题传播趋势进行预测. 本文的实验均在 Windows 7 操作系统下进行，采用了 R 语言实现了模型，在 RStudio3.0.1 版本下进行实验操作，模型参数采用极大似然估计和非线性最小二乘法对测试数据进行求解.

5.2　仿真实验

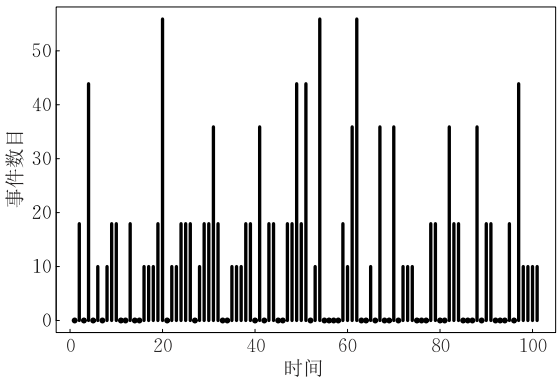
根据国内外文献[18-19]对社会媒体网络上热点话题传播相关研究，利用聚类技术，结论得出真实话题的传播可以呈现出不同特征的形态，如文献[18]中图 8(f)中话题受到大量节点影响的多峰形态，表现出了在不同时刻上产生话题事件的不同状态，这类话题对用户的吸引力度比较高，时刻都有用户关注并参与讨论；文献[19]中图 13(b)中话题受到少量节点关注的簇状形态，这类话题因为对用户的吸引力度不大，所以不能保证时刻都有用户关注，因此有些时刻有用户参与讨论，有些时刻话题受到噪音小浮动波动；或者如文献[19]中图 8(b)中话题在某一时刻事件数量极速上升或下降的单峰形态，该话题只有一个高峰，高峰前后只具有小浮动噪音形成的波动，这一类话题一般表现为明星事件，大量用户一般只在一个时刻集中参与讨论，过了此刻，只有极少的用户会偶然再次参与讨论. 因此，本文模型根据相关文献结论得出的真实话题不同的传播形态，利用仿真算法构造虚拟话题传播过程，呈现出话题传播的不同模式. 仿真实验试图评价 SEPPM 的仿真算法能否产生出与实际话题传播一致的各种传

播形态.本文采用不同的参数,利用仿真算法模拟产生话题不同传播模式,来验证仿真算法的有效性.

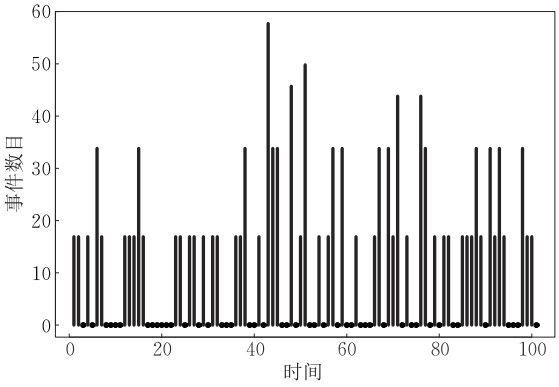
图 2 是 SEPPM 仿真 8 个话题的传播模式.其中横坐标表示[0,100)时间区间内,单位时间为 1h,



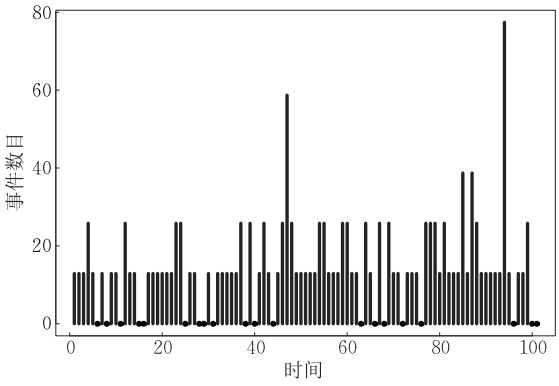
(a) 仿真话题1的传播模式



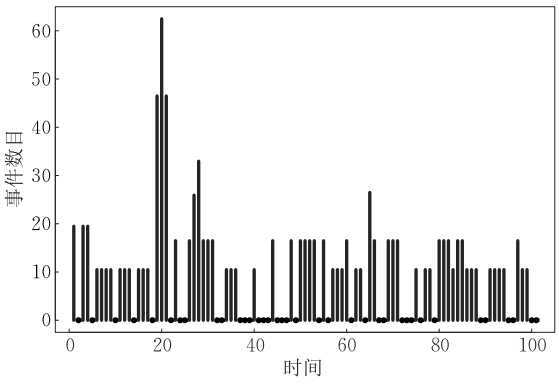
(b) 仿真话题2的传播模式



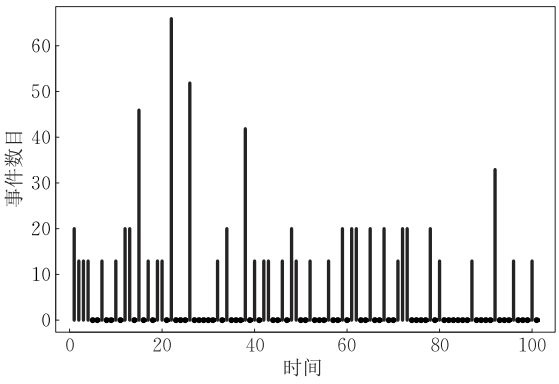
(c) 仿真话题3的传播模式



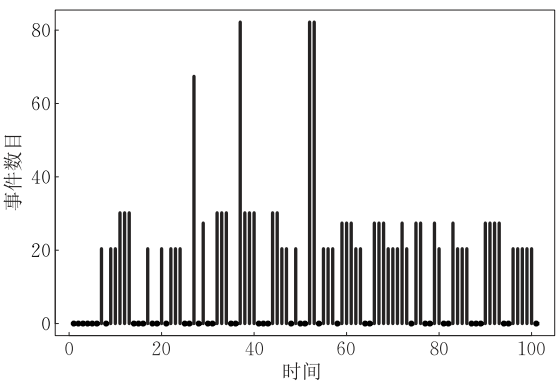
(d) 仿真话题4的传播模式



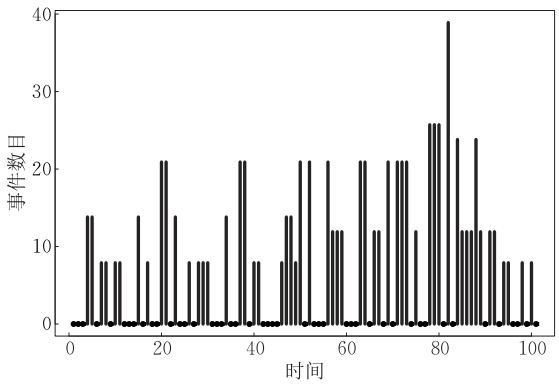
(e) 仿真话题5的传播模式



(f) 仿真话题6的传播模式



(g) 仿真话题7的传播模式



(h) 仿真话题8的传播模式

图 2 仿真话题的传播模式

纵坐标表示在对应时间段内话题的事件数量. 假设参与话题讨论的网络节点总数为 $N=10\,000$. 其中 p,μ,k_0,ω 是模型参数. 表 1 给出 8 个仿真话题对应的模型参数值.

表 1　模型仿真参数值

μ	ω	k_0	p	仿真话题
0.0100	0.2000	0.1000	0.5000	1
0.0010	0.2000	0.1000	0.5000	2
0.0100	0.7008	0.1083	0.0446	3
0.0100	0.5931	0.0924	0.1000	4
0.0010	0.5931	0.0924	0.1000	5
0.0008	0.7008	0.1083	0.0446	6
0.0100	0.7121	0.1165	0.0393	7
0.0010	0.7121	0.1165	0.4000	8

图 2(a)与图 2(b)的传播模式都具有多峰特征，但图 2(a)传播峰值较密集，能表示具有持续高峰传播的话题，而图 2(b)的多峰相对稀疏，能表示具有一定周期性高峰传播的话题. 图 2(a)和图 2(b)所代表的模型参数只有话题背景发生率 μ 不同，图 2(b)相对图 2(a)参数 μ 的值缩小了 10 倍. 因为话题背景发生率降低了，话题对节点的吸引强度也变弱了，一些非热门时刻较少节点参与话题讨论，因此与在一个时间序列上话题整体传播过程相比较为松散. 反之，如果话题背景发生率较高，话题参与讨论的节点数会增加，关注度也较高，那么参与到该话题讨论中的节点产生事件的数目会增多，话题传播过程较为紧密，几乎时刻都有节点关注话题.

图 2(c)的传播模式呈现出缓慢上升，缓慢下降的特征，图 2(d)则反应了典型的快速上升，快速下降的传播形态. 根据对应的参数分析，图 2(c)话题的衰减率 ω 相对较高，当衰减率 ω 增加的时候，过去产生的话题事件对后续产生的话题事件的影响就减弱，当 $\omega\to\infty$ 的时候，前导事件对后续事件就几乎没有影响了，因此很难出现一种快速上升或瞬时下降的传播形态，整个话题呈现比较平缓的发展趋势. 图 2(d)的衰减率 ω 较小，那么前导事件对后续事件的影响相比之下会非常大，话题受到的衰减率增加，就会出现一种快速上升或者下降的传播模式.

图 2(e)和图 2(f)仿真话题的参数 p 都较小，表示话题主要受自激效应的影响，传播过程在多数时间内呈簇状形态，体现出具有稳定用户参与传播的性质，到达某一时间点，突然受到较多节点的参与讨论，事件数目快速增加，随着时间推移，用户参与兴趣下降，事件数目下降. 图 2(e)话题的参数 μ 较大，所以与图 2(f)明显的区别是图 2(e)话题在时间序列上的传播相对紧密，话题对节点的吸引度较高，受到的关注度较大.

从表 1 的参数可以看出，图 2(g)仿真话题 7 受到的背景发生率比图 2(h)话题 8 大，自激效应产生的影响也相对较高，话题在时间序列上传播的紧密程度较高，随着话题的传播，话题热度在某一时刻达到了一个峰值，随着话题时间的推移，用户对话题的关注度也逐渐减少，话题呈下降趋势传播，然后小浮动波动传播. 比较符合一般网络热点话题的传播趋势. 话题 8 参数值 p 较大，话题受事件本身对节点的吸引度较高，受到自激效应影响较弱，所以整体趋势呈现较稳定，符合具有较长时间热度的话题传播，如热播影视话题等.

为了评价 SEPPM 仿真算法的有效性，本文把仿真结果与文献[18]中结论得出的话题典型传播形态作对比，结果发现.

(1) 热点话题的传播模式可以呈现不同特征的形态，如图 2(d)、(e)、(f)、(g)所示，有的话题只有一个高峰，高峰前后只具有小幅噪声形成的波动，与文献[18]中图 8(a)、(b)、(c)的 3 种曲线形态相似；

(2) 有的话题表现出用户一般只参与话题讨论一次，但是少量用户偶然参与多次讨论，如图 2(c)、(h)所示，这一类话题经常表现为公众事件，与文献[18]中图 8(d)、(e)的 2 种曲线形态相似；

(3) 有的话题会形成用户之间的交互辩论，如图 2(a)、(b)所示，话题事件数目会在不同的时刻具有不同的状态，从而会形成多个完整的高峰形态，与文献[18]中图 8(f)曲线形态相似.

通过对比仿真结果与实际传播模式，SEPPM 仿真模拟出来的话题传播过程与实际话题的各类形态都非常相似. 说明仿真算法能够有效性地模拟话题的传播过程，也说明 SEPPM 能够刻画出话题的传播模式.

5.3　真实话题实验

本文采用文献[19]中的话题数据集，从经济类、军事类、社会类、娱乐类和运动类 5 种不同类型的话题中一共抽取了 15 个话题，以 1 h 作为单位时间计算每个热点话题的事件数量，截取 100 个时间点上的事件进行实验. 为了客观衡量 SEPPM 的性能，本文选择 SpikeM 和 SEIZ 模型进行比较性实验. 文献[19]综合比较了主要的话题传播模型，试验结果表明 SpikeM 模型能较好地拟合话题传播模式. 文献[7]提出了近年来比较新颖的传染病建模方法，它改进了 SIS 模型的缺点，可以更好地表达话题传播过

程.所以本文采用 SpikeM 和 SEIZ 模型作比较性方法.

5.3.1 话题传播拟合实验

话题传播拟合实验的目的是评价不同模型是否能够拟合真实话题传播过程.本文拟合了 15 个话题的传播模式.相关系数是可以用来反映变数之间相关关系密切程度的统计指标,按照积差方法计算,通过两个离差相乘来反映两变数之间相关程度,反映曲线间趋势相似度.为了衡量 SEPPM 与 SpikeM 拟合话题趋势的差异,本文分别计算了 SEPPM 与 SpikeM 拟合 15 个话题的相关系数,结果如图 3 所示.从图 3 可以看出.

(1)在 13 个话题上,SEPPM 的相关系数明显高于 SpikeM 的相关系数,两个模型在 2 个话题(话题 3 和 15)上相关系数接近.说明 SEPPM 拟合效果要优于 SpikeM;

(2)SEPPM 拟合不同类型话题的相关系数比较稳定,都在 0.6 左右,而 SpikeM 拟合不同类型话题的相关系数差异很大,值范围从 $-0.2\sim0.8$.说明 SEPPM 在拟合不同类型的话题上具有较好的适应性,而 SpikeM 受到模型假设条件的限制,难以有效拟合一些具有复杂传播模式的话题.SEPPM 的平均相关系数值为 0.69,SpikeM 的平均相关系数值为 0.37,这说明 SEPPM 在建模中能够有效建模不同类型的话题.

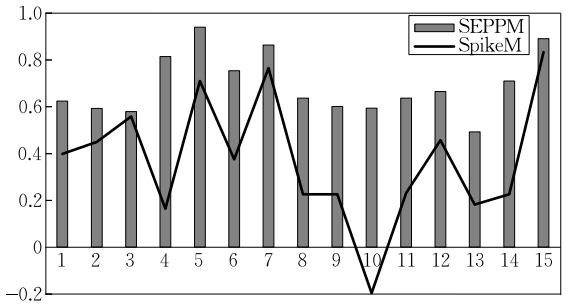


图 3 SEPPM 与 SpikeM 模型拟合相关系数比较

SEIZ 模型具有传染病模型的特点,呈现出先快速上升,然后下降的过程.拟合真实话题趋势的相关度较低.

本文从 15 个话题中选取了 6 个具有不同传播模式的话题.话题 1 关于"神舟八号飞船发射成功",发射前具有较长时间的平稳和一些小高峰,发射后快速爆发;话题 2 是"郭美美事件",呈现出 2 个不同的高峰;话题 3 是"韩国海警殴打中国船员"消息,刚开始话题没有传开因此一直小幅传播,当消息传播一定范围后,该话题立刻上升到较高热度并反复讨论;话题 4 是"刘翔亚运会三连冠"消息,呈现出高峰维持时间长、慢速下降的特点.话题 5 是"叶诗文游泳世锦赛 200 混合泳夺冠"消息,最开始关注度不够密集,到达某一个时刻迅速受到用户反复讨论.话题 6 是"谢霆锋张柏芝复合"消息,符合娱乐新闻最开始受到高关注状态.

表 2 模型学习参数

μ	ω	k_0	p	话题
0.001	0.7	0.010	0.20	1
0.003	0.5	0.090	0.10	2
0.007	0.6	0.100	0.04	3
0.010	0.7	0.100	0.03	4
0.080	0.1	0.001	0.20	5
0.006	0.3	0.010	0.20	6

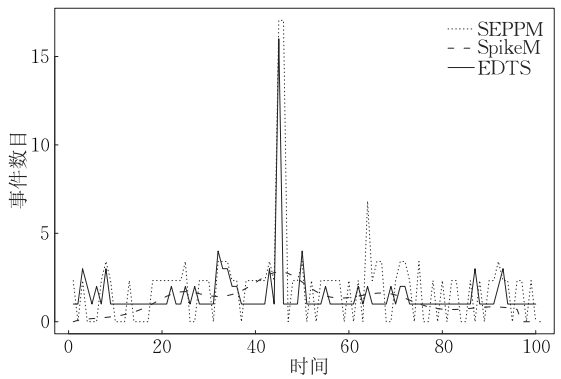
图 4 的(a)、(b)、(c)、(d)给出 4 个典型话题真实传播模式与 SEPPM 和 SpikeM 模型的拟合结果;图 4 的(e)、(f)是真实传播模式与 SEPPM 和 SEIZ 模型的拟合结果.图中 EDTS(Exprimental Data Time Series)表示试验话题序列,即真实话题数据;SEPPM 表示本文提出的模型拟合序列;SpikeM 表示 SpikeM 模型拟合序列;SEIZ 表示 SEIZ 模型拟合序列.图 4 中横坐标为时间点,以 1h 为单位时间,纵坐标表示的是在时间区间上事件的数量,数量越高,参与话题讨论的节点越多,话题在某个时刻强度函数越大.

模型学习得到的参数如表 2 所示.对比图 4 每个子图的 SEPPM 拟合结果,可以发现:

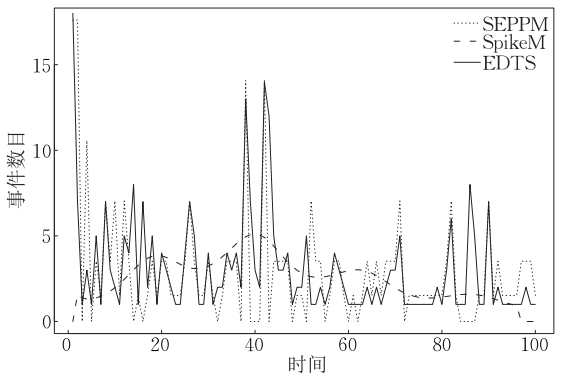
(1)SEPPM 能够拟合出话题的不同传播模式,例如图 4(a)对应的话题先后一直小浮动传播,在中间时刻事件数量爆发式上升;图 4(b)、(e)对应的话题具有 2 个高峰且间隔一定时间;图 4(c)话题的传播模式虽然刚开始事件数目较少,但是随着时间发展关注话题的节点逐渐增加,之后一段时间都呈现多峰形态;图 4(d)、(f)的话题也在一段时间中呈现高峰形态,节点参与话题讨论,并且持续一段时间.

(2)与仿真结果一致,图 4(a)所示话题的参数 p 值较大,也就是话题传播时,节点之间除了受到自激效应之外,还受到较大的外部影响,因为该话题属于公众激励消息,所以有众多节点自我参与.对于图 4(d)所示的话题,参数 μ 和 ω 的值较大,而参数 p 较小,说明话题不仅具有较高的背景发生率,也就是主动参与的节点较多,同时也受到群体影响,说明在话题发展过程中,节点参与存在相互间的激励效应.

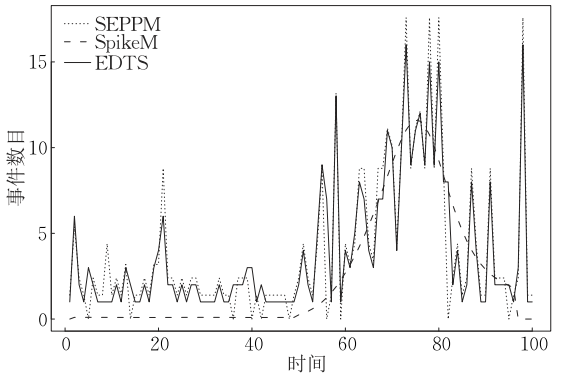
(3)根据 6 个真实话题拟合效果分析,SpikeM 能够刻画出话题的总体发展趋势;SEPPM 不仅能



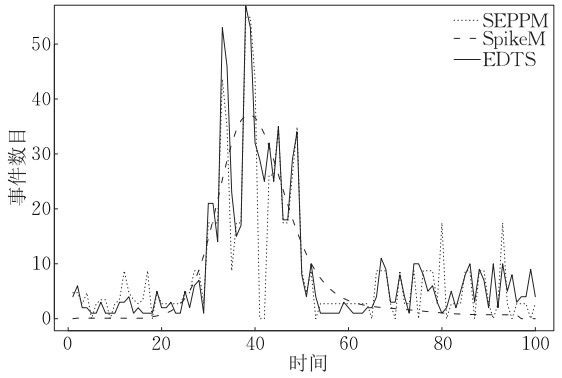
(a) 话题1传播趋势图



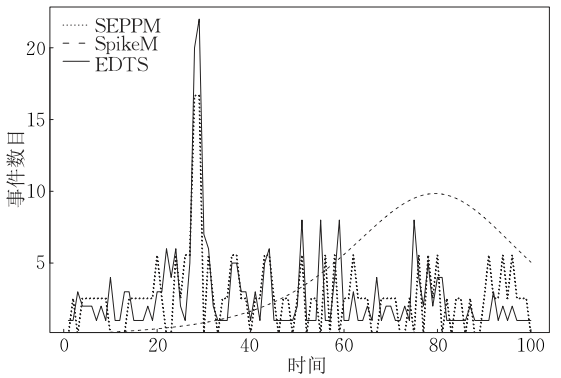
(b) 话题2传播趋势图



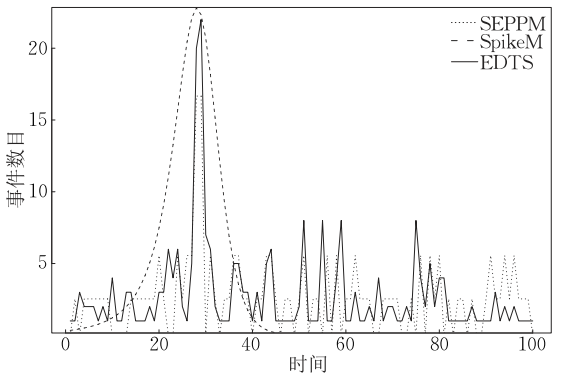
(c) 话题3传播趋势图



(d) 话题4传播趋势图



(e) 话题5传播趋势图



(f) 话题6传播趋势图

图 4　话题拟合结果对比图

够较好地刻画话题的总体趋势，还能够表示话题发展过程中的波动；SEIZ 则只能近似表示话题的总体发展过程，由于 SEIZ 依赖于传染病模型，对具有快速上升和快速下降的单一传播模式的话题具有较好的效果. 真实话题的传播受到很多因素的影响，具有很多波动，其传播模式较为复杂，导致 SpikeM 和 SEIZ 模型刻画真实话题传播的效果较差.

SEPPM 拟合出来的话题传播模式与真实数据下的传播模式存在一些误差，例如图 4(a)、(f)后期的波动. 产生误差的主要原因是原始数据中包含有大量噪声，话题中存在一些不属于本话题的事件，此外数据采集、网络等原因，也使话题中的部分事件存在没有采集完备的情况，这是导致模型拟合存在一些误差的主要原因.

5.3.2　模型预测实验

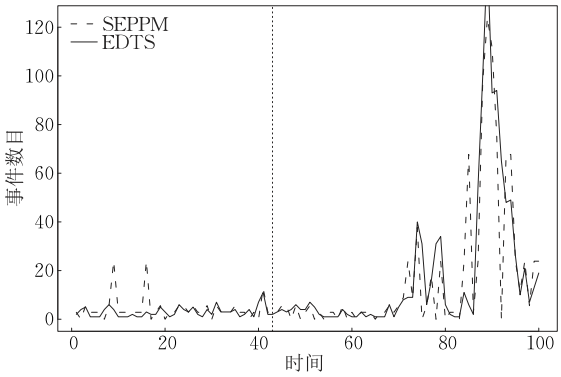
为了评估 SEPPM 的预测能力，检验模型在实际应用中的价值. 本文在真实数据上进行预测实验. 选择 4 个代表性话题作为实验对象，这 4 个话题呈现出不同的传播特征，例如多个峰值和单个峰值等不同的形态. 在[0,100)的时间段内，采用前 43 个时间点的数据作为学习阶段，后 57 个时间点作为预测，通过最小二乘法学习样本参数值，将求解出来参数代入模型中，得到预测模型传播过程. 图 5 分别给出 4 个话题的真实传播模式和预测结果，每个图上

垂直虚线标出的时间点为学习、预测分割点.从图 5 各个子图可以看出:

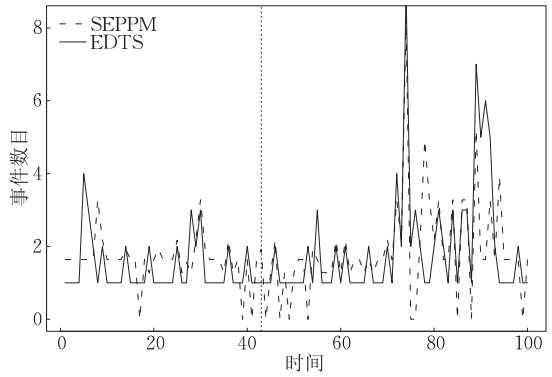
(1)学习阶段,SEPPM 能够在较短的时间段内拟合 4 个话题的传播趋势.图 5(a)对应的话题 1 与图 5(d)对应的话题 4 在学习阶段已经出现峰值,拟合的峰值形态非常明显;图 5(b)和图 5(c)对应的话题在学习阶段没有出现峰值,SEPPM 拟合出的传播模式也是呈现小幅波动;

(2)预测阶段. 对于话题 1,传播高峰出现在学习阶段,SEPPM 可以预测出话题传播后期的局部波动.对于话题 2,传播高峰的后期并且具有一些缓慢的局部上升过程,SEPPM 同样可以预测出高峰和上升过程.对于话题 3,高峰在学习阶段的前期,后期具有一些波动,SEPPM 仍然可以预测出最近的高峰和后期的波动.对于话题 4,学习阶段就可以拟合出峰值不同变化的形态,SEPPM 同样也可以预测出之后峰值不规则波动的情况.

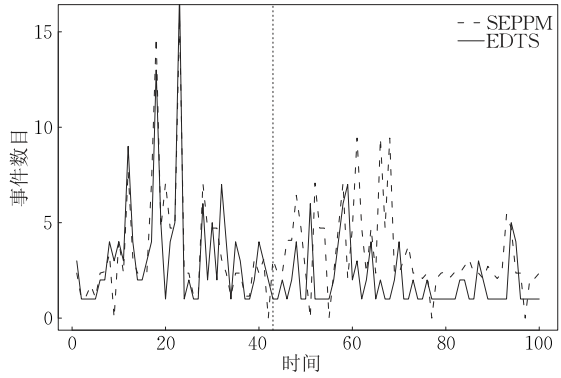
(3)SEPPM 不仅可以预测话题传播过程的一般波动,也能预测话题传播过程的高峰,而 SpikeM 则无法预测话题传播的高峰.但 SEPPM 预测也存在一些误差,例如话题 2 和话题 3 中的局部波动中,SEPPM 在局部波峰下降过程中存在一些没有精准预测的时间点.



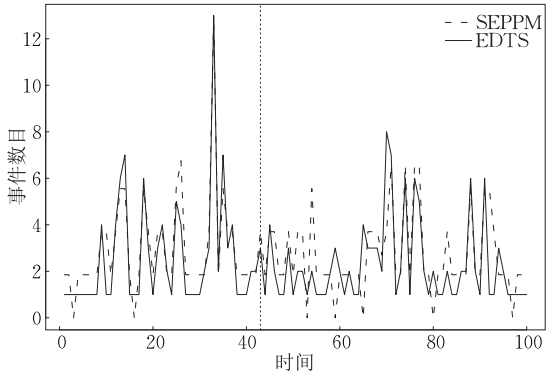
(a) 话题1预测结果图



(b) 话题2预测结果图



(c) 话题3预测结果图



(d) 话题4预测结果图

图 5 话题预测结果图

为了更好地分析模型预测效果,分别采用均方根误差(*RMSE*),判定系数 R^2 和相关系数 r 方法比较 SEPPM 与 SpikeM 的差异.因为计算 *RMSE* 误差值,可以比较模拟数据与真实数据之间的误差大小,误差越小表示模型效果越优;通过计算判定系数 R^2,可以分析随着时间推移反映对应事件数目的影响程度,建模的目的是预测应变量的值,所以判定系数较高则比较好;计算相关系数 r 可以表达预测结果与真实数据相关程度,系数值较高则相关性越高,预测结果较好.通过这 3 个指标,可以全面衡量本文模型预测效果.

为了比较 SEPPM 和 SpikeM 模型预测能力的差异,本文计算了 SEPPM 和 SpikeM 与真实数据值之间的 *RMSE* 值,R^2 值,以及 r 值,结果如表 3 所示.

从表 3 中可以看出:(1) SEPPM 的误差值较小,预测效果优于 SpikeM;(2) SEPPM 的判定系数明显高于 SpikeM 的判定系数,与真实数据的拟合优度相对较高;(3) SEPPM 的相关系数也比 SpikeM 的相关系数较高. SEPPM 可以很好地对话题传播

模式进行拟合，并且能够预测话题在一定时间内的传播趋势.

表 3　模型预测 *RMSE*、R^2、*r* 值

模型	话题	*RMSE*	R^2	*r*
SEPPM	1	1.139569	0.7271812	0.8593996
SpikeM		1.519055	0.5152252	0.7186066
SEPPM	2	1.300341	0.9295147	0.9686667
SpikeM		3.159360	0.6661686	0.8164472
SEPPM	3	12.187720	0.7878600	0.8909202
SpikeM		15.435870	0.6597173	0.7122940
SEPPM	4	1.278575	0.8340647	0.9260627
SpikeM		2.098495	0.5507354	0.7425772

6　总　结

本文针对建模和预测热点话题传播开展研究，提出了一个 SEPPM 模型，进行了仿真和实际话题数据上的丰富实验. 模型理论分析和实验结果表明：

(1) SEPPM 能够仿真出不同类型的话题传播模式，只需要简单地改变参数值就能实现模拟复杂的传播模式.

(2) SEPPM 能够较好地拟合和预测真实话题的传播过程，能够预测话题的高峰和波动. SEPPM 的拟合精度和预测精度都优于 SpikeM.

SEPPM 预测真实话题趋势时，还存在一些问题，例如峰值位置偏移，峰值绝对大小的误差等，如何利用参与节点间的关系，改进模型的预测效果是未来工作需要研究的内容. 本文研究成果有助于进行网络热点话题发现、流行病学研究、舆情监测等工作.

致　谢　评审专家提出了宝贵意见与建议，在此表示感谢！

参　考　文　献

[1] Allan J, Papka R, Lavrenko V. Online new event detection and tracking//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 1998: 37-45

[2] He Yan, Song Li-Li. On discovery and tracking of hotspots in key areas. Journal of Southwest China Normal University (Natural Science Edition), 2014, 39(7): 67-72(in Chinese)
(何焱，宋丽丽. 关键领域热点发现与跟踪. 西南师范大学学报(自然科学版)，2014，39(7)：67-72)

[3] Deng Jing-Wei, Deng Kai-Ying, Li Ying-Xing. Topic detection and tracking method for Tibetan network. Advanced Materials Research, 2014, 860-863: 2914-2917

[4] Chen Zhen, Zhu Kai, Ying Lei. Detecting multiple information sources in networks under the SIR model//Proceedings of the 48th CISS Annual Conference. New York, USA, 2014: 1-4

[5] Wen Sheng, Zhou Wei, Zhang Jun, Xiang Yang. Modeling propagation dynamics of social network worms. Parallel and Distributed Systems, 2013, 24(8): 1633-1643

[6] Kurihara S. The multi agent based information diffusion model for false rumor diffusion analysis//Proceedings of the 14th WWW Companion International Conference on World Wide Web Companion. Seoul, Korea, 2014: 1319-1320

[7] Fang Jin, Dougherty E, Saraf P, et al. Epidemiological modeling of news and rumors on Twitter//Proceedings of the 7th Workshop on Social Network Mining and Analysis. Chicago, USA, 2013

[8] Stomakhin A, Short M, Bertozzi A. Reconstruction of missing data in social networks based on temporal patterns of Interactions. Inverse Problems, 2011, 27(11): 1-15

[9] Zhao Li, Yuan Rui-Xi, Guan Xiao-Hong, Jia Qing-Shan. Bursty propagation model for incidental events in blog networks. Journal of Software, 2009, 20(5): 1384-1392(in Chinese)
(赵丽，袁睿翕，管晓宏，贾庆山. 博客网络中具有突发性的话题传播模型. 软件学报，2009，20(5)：1384-1392)

[10] Gruhl D, Guha R, Liben-Nowell D. Information diffusion through blogspace//Proceedings of the 13th International Conference on World Wide Web. NewYork, USA, 2004: 491-501

[11] Saitok K, Kimuram M, Oharak K, Motoda H. Selecting information diffusion models over social networks for behavioral analysis//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: PartIII. Barcelona, Spain, 2010: 180-195

[12] Ji Jin-Chao, Han Xiao, Wang Zhe. Community influence maximizing based on comprehensive cascade diffuse model. Journal of Jilin University (Science Edition), 2009, 47(5): 1032-1034(in Chinese)
(冀进朝，韩笑，王喆. 基于完全级联传播模型的社区影响最大化. 吉林大学学报(理学版)，2009，47(5)：1032-1034)

[13] Hawkes A G, Oakes D. A cluster process representation of a self-exciting process. Journal of Applied Probability, 1974, 11(3): 493-503

[14] Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649-15653

[15] Xu Zhi-Jing, Zu Zheng-Hu, Xu Qing, et al. Modeling terrorist activities with self-exciting point process. Military Medical Science Letters, 2012, 36(10): 750-753(in Chinese)
(徐致靖，祖正虎，许晴等. 基于自激点过程的恐怖活动建模研究. 军事医学，2012，36(10)：750-753)

[16] Yang J, Leskovec J. Patterns of temporal variation in online media//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA, 2011: 177-186

[17] Ogata Y. On lewis' simulation method for point processes. Browse Journals and Magazines, 1981, 27(1): 23-31

[18] Yasuko M, Yasushi S, Prakash A B. Rise and fall patterns of information diffusion: Model and Implications//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 6-14

[19] Han Zhong-Ming, Chen Ni, Le Jia-Jin, et al. An efficient and effective clustering algorithm for time series of hot topics. Chinese Journal of Computers, 2012, 35(11): 2337-2347(in Chinese)
(韩忠明, 陈妮, 乐嘉锦等. 面向热点话题时间序列的有效聚类算法研究. 计算机学报, 2012, 35(11): 2337-2347)

HAN Zhong-Ming, born in 1972, Ph. D., associate professor. His current research interests include web data mining, big data and information retrieval.

ZHANG Meng, born in 1991, M. S. candidate. Her current research interests include web data mining and information retrieval.

TAN Xu-Sheng, born in 1989, M. S. candidate. His current research interests include web data mining and community detection.

DUAN Da-Gao, born in 1976, Ph. D., associate professor. His current research interests include social computing and multimedia information processing.

SI Hui-Lin, born in 1972, Ph. D., lecturer. Her current research interests include big data and multimedia information processing.

Background

With the rapid growth of online social media for communicating, sharing and managing significant information and topics, Interactive social media has becoming the main platform for users to spread all kinds of topics. By analyzing and modeling propagation processes of hot topics, one can understand the spreading patterns of hot topics, early detect and predict hot topics. As a result, modeling propagation processes of hot topics on Internet has significant meaning and value.

This paper focuses on modeling hot topics on Internet and proposes a topic propagation model (SEPPM) based on self-exciting Hawkes process. SEPPM models the propagation process of one topic as a random point process by using self-exciting effect of user participation. On the other hand, SEPPM also takes external factors for propagation into account, thus puts forward a formal topic propagation model. To evaluate effectiveness of SEPPM, comprehensive simulation and empirical experiment are conducted. A simulation algorithm for SEPPM is proposed and the simulated results show that SEPPM can generate a variety of patterns of hot topics with different propagation characteristics. The experimental results on real datasets show that SEPPM can not only splendidly fit real topic propagation process, but also can effectively forecast spreading trend.

Our team has done a lot of work on topic detection and tracking. We has investigated the classification of hot topics. All of our related work has provided the good foundation of modeling propagation processes of hot topics. All its previous research results could speed our research. Its research result is very meaningful.

This work is supported by the National Natural Science Foundation of China under Grant No. 61170112, the MOE (Ministry of Education in China) Project of Humanities and Social Sciences under Grant No. 13YJC860006 and the Central Financial Support Local Colleges and Universities to Develop Special Funds for Personnel Training and Innovation Team Building Project under Grant No. 19005323132.