物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

一种基于社交影响力和平均场理论的信息传播动力学模型

肖云鹏　李松阳　刘宴兵

An information diffusion dynamic model based on social influence and mean-field theory

Xiao Yun-Peng　Li Song-Yang　Liu Yan-Bing

引用信息 Citation: Acta Physica Sinica, 66, 030501 (2017)　DOI: 10.7498/aps.66.030501
在线阅读 View online: http://dx.doi.org/10.7498/aps.66.030501
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2017/V66/I3

一种基于社交影响力和平均场理论的信息传播动力学模型*

肖云鹏† 李松阳 刘宴兵

(重庆邮电大学, 网络与信息安全技术重庆市工程实验室, 重庆 400065)

(2016 年 6 月 7 日收到; 2016 年 10 月 18 日收到修改稿)

在线社会网络中, 信息传播蕴含着复杂的动力学成因. 本文将传染病模型与社交影响力要素相结合, 并针对影响力度量中主要研究静态网络拓扑结构、忽略个体行为特征的问题, 提出一种基于动态节点行为和用户影响力的信息传播动力学模型, 旨在量化影响力强度, 为研究信息扩散过程中不同用户群体状态转变提供理论依据. 首先, 在网络拓扑结构和用户行为两方面, 提取个人记忆和用户交互两个表征, 分析影响力形成的内因和外因两个动力学成因, 并基于多元线性回归模型, 提出一种度量用户社会影响力的方法. 其次, 在传统传染病 SIR (susceptible-infected-recovered) 模型基础上, 结合信息扩散与传染病蔓延相似的传播机理, 综合考虑信息传播的多源并发性和双向性, 引入影响力因子, 利用平均场理论改进得到一种基于用户影响力的信息传播模型. 实验表明, 该模型能有效地解释在线社会网络中信息传播的动力学原因, 感知社会网络中信息传播演化态势.

关键词: 信息传播, 传播动力学, 社交影响力, 平均场理论

PACS: 05.10.–a, 89.75.Fb **DOI:** 10.7498/aps.66.030501

1 引 言

面向社会性网络服务的在线社交网络是 Web2.0 体系下的典型应用, 当前关于在线社会网络的研究包括网络结构 [1]、群体互动 [2]、信息传播 [3,4] 三个核心要素. 在线社会网络的蓬勃发展和线上用户的急剧增长, 使其迅速成为人们信息传播、商品营销、购物推荐、观点表达、产生社会影响力的理想平台, 研究社会网络中的信息传播动力学成为热门课题之一 [5,6]. 鉴于信息收发方式多样化、传播效率核裂化、交互方式便捷化的特点, 有效控制舆情信息的准确性、真实性和及时性对规范线上网络信息起着至关重要的作用, 也为相关部门加强社会舆情监控和制定信息管控策略提供重要参考.

在线社会网络的信息传播问题研究, 主要涉及研究信息传播机理的解释模型和信息扩散规律的预测模型. 解释模型中, 学者们大多从网络结构或自定义的信息传播机理着手研究. 例如, Borge-Holthoefer 和 Moreno [7] 选择不同核数的节点作为模型的出发点, 探索核数在信息传播的作用. 王超等 [8] 结合遏制机制和遗忘机制提出新的信息传播模型. Lu 等 [9] 研究了竞争性和合作性实体对信息传播的影响. 预测模型归结起来可分为基于图的线性阈值模型 [10] 和独立级联模型 [11], 基于非图的传染病模型 [12] 和博弈论模型 [13].

传染病模型是信息传播领域较为成熟的模型 [14]. 传染病模型 [15] 认为, 当信息已知者对某个消息未知者的传播率大于某一临界值时, 信息已知者会将信息传播给该消息未知个体, 这个过程

* 国家重点基础研究发展计划 (批准号: 2013CB329606)、国家自然科学基金 (批准号: 61272400)、重庆市青年人才项目 (批准号: cstc2013kjrc-qnrc40004)、教育部–中国移动研究基金 (批准号: MCM20130351)、重庆市研究生研究与创新项目 (批准号: CYS14146)、重庆市教委科学计划项目 (批准号: KJ1500425) 和重庆邮电大学文峰基金 (批准号: WF201403) 资助的课题.

† 通信作者. E-mail: xiaoyp@cqupt.edu.cn

会持续到整个网络信息已知者处于某一稳定的状态. SIR (susceptible-infected-recovered) 模型在发展过程中出现了很多变种, 例如, 类似于SI模型的级联模型[16]、考虑到重复感染的SIS模型[17]以及异构网络中的SIRS模型[18,19]. Xiong等[20]提出一种基于转发机制的SCIR模型, 把网络中的人群划分为4类, 即易感人群S (susceptible)、读了信息但没有传播的人群C (contacted)、接收信息并继续传播的人群I (infected)和失去传播信息兴趣的人群R (refractory), 该模型认为I和R是信息传播最终稳定的状态. Li等[21]提出改进的SIQRS模型研究无标度网络中的传播动力学模型, 在模型中加入隔离人群Q(quarantined individuals), 并认为这些人在信息传播中起关键作用. Xiong等[22]引入潜伏者的角色, 提出了SILR(susceptible-infected-latent-refractory)模型. 由此可见, 学者们集中考虑SIR模型状态的划分问题, 少有量化导致状态改变的感染率和免疫率的研究.

另一方面, 影响力最大化问题[23,24]也是信息传播领域的关键问题之一, 其目的是通过提取异质网络中的多个特征, 与影响用户转发、评论行为的因素结合, 发现最大程度上影响网络功能与信息传播的节点集合[25,26]. 已有研究中, 学者们大都是基于网络拓扑结构[23,27,28]探讨信息的传播方式, 而在线社会网络用户数量巨大, 用户之间形成的关系非常复杂, 在这样的环境下对社交影响力的定性分析受到很多因素的干扰和影响[29]. 尽管有不少学者试图客观准确地理清影响力和其他因素之间的关系, 使用了包括随机化方法[30,31]、马尔可夫链[32]、块模型[33]在内的多种技术手段, 但最终难以很好地解决该问题. 这种局面不仅与社交影响力的复杂构成和信息传播的不确定性有关, 也与影响力自身的定义密切相关.

综上可知, 现有研究中普遍忽视了信息传播中用户相互影响的问题, 本文从个体记忆维度和用户交互维度两个角度量化群体间的影响力, 并认为影响力因素是传染病模型中状态转化的动力学成因, 利用平均场理论[34]对在线社会网络传播模式进行分析研究.

本文的创新点可以总结为如下两点.

1) 在影响强度计算上, 与当前研究工作中主要考虑网络结构不同, 本文综合考虑内部因素即个体记忆维度及外部因素即用户交互维度, 提出一种基于多元线性回归模型的用户影响力计算和衡量方法. 结合用户自身属性和个体行为习惯两个维度分析个体记忆原理, 利用图论中的最短路径法来度量社会网络中用户间信息交互经过某条边的流的总数来研究用户交互原理.

2) 在信息扩散建模上, 借鉴SIR模型机理, 本文引入用户影响力因子作为传染病模型中状态改变的参数, 运用平均场理论建立微分方程组, 并在此基础上给出新的信息传播动力学模型和验证方法, 有效避免了在模型中人为设定参数带来的随机性, 揭示信息传播中多因素耦合的本质规律.

本文的组织结构如下: 第一部分阐述模型建立的背景及其基础性工作; 第二部分依据在线社会网络信息关系网、信息传播属性、影响力形成的内因和外因给出属性的定义及问题的科学性描述; 第三部分对信息模型进行详细描述, 并给出相应的学习算法和平均场方程; 第四部分以真实的信息数据为背景, 运用研究思路对模型进行仿真实验, 并与真实的自然规律进行对比验证; 第五部分对本文所做的相关工作进行总结.

2 传播网络与问题定义

2.1 相关知识

首先, 设$G=\{V,E\}$为信息传播网络, 其中$V=\{v_1,v_2,\cdots,v_n\}$是社会网络中单个信息互动用户集合, $|V|=N$, 即用户总数, $E\subseteq V\times V$为用户间的朋友关系, 若存在边$e_{i,j}=\langle v_i,v_j\rangle$, 则用户$v_j$是用户$v_i$的粉丝, 表示信息可沿边$e_{i,j}$由用户$v_i$传向用户$v_j$.

然后, 设$A=\{(a,v_i,t)\}$为不同时间段的用户互动数据, 其中$\{(a,v_i,t)\}$表示用户v_i在t时间的动作a, A是用户群体T_k时间段的互动行为.

接着, 设个人记忆原理Inner和用户交互原理Outer两种度量用户影响力的特征量, 形式化表示群体事件扩散中用户行为动力学的内因和外因.

最后, 设$D(v_i,t)$为用户v_i在时刻t的状态. 网络中的用户划分为3类, 每类个体集合都处于同一种状态, 即$D(v_i,t)=\{S,I,R\}\in\kappa$, 其中$\kappa$表示单个信息事件的传播行为. 每个用户有三种可能的状态, 分别为易感状态S (susceptible), 即消息未知者, 有可能被感染; 感染状态I (infected), 即消息已

知者, 具有传染性; 免疫状态R (recovered), 即消息免疫者, 对消息失去了兴趣.

2.2 特征提取与定义

在社会网络中, 信息传播是个体、团体之间的信息传递和交流. 一个人发布的消息会被其好友看到, 并以一定的概率分享、传播. 若好友对该信息内容不感兴趣, 则成为消息免疫者且不会传播. 本文针对热点话题传播, 挖掘影响用户参与话题讨论和转发等行为的内部、外部动力驱动因素, 具体从个体记忆和用户互动两个维度出发, 提取影响信息传播表征.

1) 个体记忆维度

在个体方面, 我们认为一个喜欢参与社会事件讨论的人可能在今后的生活中依然喜欢参与, 这种社交活动中的活跃分子对群体事件具有记忆效应, 这是由用户自身因素决定的. 在社会网络中, 用户对一个群体事件的关注程度反映了信息的受欢迎程度, 关注的人越多, 证明人们对该事件越感兴趣, 信息扩散的机会就越大. 也就是说, 研究用户个人兴趣在信息传播领域具有不可或缺的作用. 为了度量影响力的内部成因, 结合信息传播网络的静态拓扑结构, 本文提取用户内部属性列于表1. 以下对所提取的内部属性做具体阐述.

定义1 用户度数$Deg(v_i)$

用户度数(degree)定义为与用户v_i相关联的边的数目. 社会网络是有向图, 若存在边$v_i \to v_j$, 则用户v_j是用户v_i的关注者, 关注者总和记作$d^+(v_i)$; 若存在边$v_i \leftarrow v_k$, 则用户v_k是用户v_i的粉丝, 粉丝总和记作$d^-(v_i)$. 显然

$$Deg(v_i) = d^+(v_i) + d^-(v_i). \tag{1}$$

定义2 用户介数$C_{\rm b}(v_i)$

用户介数(between)定义为网络最短路径中经过该节点(或边)的概率之和, 描述了节点在网络中的影响力与中心性程度. 假设节点p和节点q之间的最短路径数为δ_{pq}条, 这两个节点之间经过节点v_i的最短路径数为$\delta_{pq}(v_i)$. 在此基础上, 用户v_i的介数可定义为

$$C_{\rm b}(v_i) = \sum_{p\in V}\sum_{q\neq p\in V}\frac{\delta_{pq}(v_i)}{\delta_{pq}}. \tag{2}$$

定义3 用户紧密度$C_{\rm c}(v_i)$

用户紧密度(closeness)定义为用户v_i与网络中其他用户的平均距离的长短, 考察用户v_i传播信息时不依靠其他用户的程度. 若社会网络中有N个用户, 求用户v_i到其他所有用户的最短距离, 记作$d(v_i, v_j)$, 则用户紧密度可定义为

$$C_{\rm c}(v_i) = \frac{N-1}{\sum\limits_{j=1}^{N} d(v_i, v_j)}. \tag{3}$$

表1 内部属性符号及描述

Table 1. Symbols and descriptions of internal attribute.

序号	符号	描述
1	$Deg(v_i)$	用户v_i的度数
2	$C_{\rm b}(v_i)$	用户v_i的介数
3	$C_{\rm c}(v_i)$	用户v_i的紧密度

为了便于描述, 本文统一用ψ_{ij}来表示用户v_i影响力内在驱动因素, 其中$j=1,2,3$分别代表上述3个静态属性.

2) 用户交互维度

现实世界中人们通过共同地域、相同活动、亲属关系等形式构建网络, 在线社会网络与现实世界不同, 主要通过信息发布、共享、扩散等信息交流的形式产生联系. 因此, 在线社会网络中信息是人们产生联系的载体, 用户线上的交互行为在信息传播影响力中扮演了重要角色. 为了对影响力形成的外部动态驱动因素进行定量分析, 结合促进信息传播的用户行为记录, 本文提取用户间交互的属性列于表2.

定义4 内容相似性$S(v_i)$

内容相似性(similarity)定义为用户v_i的个人兴趣与话题标签的相似程度. 从用户自定义的标签和热点话题中分别提取关键字, 用Jaccard系数进行归一化计算. Jaccard系数越大, 表明信息内容和用户个人兴趣有较大的相关性, 反之, 相关性较小. 令A为热点话题内容, B为用户历史行为数据的高频词汇, 则内容相似性为

$$S(v_i) = \frac{|A\cap B|}{|A\cup B|}. \tag{4}$$

定义5 意见领袖$L(v_i)$

意见领袖(leader)定义为对他人施加影响的活跃分子, 在信息传播中起到重要的中介或过滤作用. 用PageRank算法[35]计算得到的PR值作为

判定用户 v_i 是否为意见领袖的阈值, 用 φ 作为可调参数, 意见领袖定义为

$$L(v_i)=\begin{cases}1, & PR(v_i)>\varphi,\\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

定义 6 活跃用户 $A(v_i)$

$A(v_i)$ 代表目标用户 v_i 是否为活跃用户 (active user), 1 代表该用户是活跃用户, 0 代表该用户不是活跃用户. 相比非活跃用户, 我们认为活跃用户对信息传播所起的作用较大, 活跃用户定义为

$$A(v_i)=\begin{cases}1, & \text{Active}(v_i)>\tau,\\ 0, & \text{otherwise,}\end{cases} \tag{6}$$

其中, $\text{Active}(v_i)$ 代表用户 v_i 的活跃指数, τ 为可调参数.

$$\begin{aligned}&\text{Active}(v_i)\\ &=\rho\times Num[orig(v_i)]+Num[retw(v_i)],\end{aligned} \tag{7}$$

其中, $\rho\in[0,1]$ 为弱化系数, $N[orig(v_i)]$ 和 $N[retw(v_i)]$ 分别是用户 v_i 在话题发起前一个月每日发表微博和转发微博的数量.

定义 7 信息传播带动力 $I(v_i)$

$I(v_i)$ 指的是用户 v_i 发布信息后, 该信息由于该用户粉丝的浏览、评论、转发等历史行为在社会网络中不断扩散, 设 η 是弱化系数, $\overline{Num}[read(u_j)]$, $\overline{Num}[comt(u_j)]$, $\overline{Num}[ret(u_j)]$ 分别为用户 v_i 在所研究话题发起前一个月每条微博的平均阅读数、平均评论数、平均转发数. 综合不同的用户行为, 量化该用户的信息传播带动力为

$$\begin{aligned}I(v_i)=&\eta\times\overline{Num}[read(v_i)]+\overline{Num}[comt(v_i)]\\ &+\overline{Num}[ret(v_i)].\end{aligned} \tag{8}$$

表 2 外部属性符号及描述

Table 2. Symbols and descriptions of external attribute.

序号	符号	描述
1	$S(v_i)$	用户 v_i 标签与热点事件内容相似性
2	$L(v_i)$	用户 v_i 是否为意见领袖
3	$A(v_i)$	用户 v_i 是否为活跃用户
4	$I(v_i)$	信息传播带动力

为了便于描述, 本文统一用符号 χ_{ij} 来表示用户 v_i 影响力外部驱动因素, 其中 $j=1,2,3,4$ 代表上述 4 个动态属性.

2.3 问题形式化

本文构建的模型旨在分析信息传播中的用户行为动力学成因, 挖掘用户个人和用户间的因素在信息传播中所起的作用, 如图 1 所示, 通过量化影响力强度进一步分析用户影响力对信息接收程度的影响, 并利用平均场理论挖掘信息传播态势. 其中, 对于影响力的度量, 考虑到线性回归模型能综合信息的多个特性, 本文设计了一种基于线性回归模型的影响力强度计算方法.

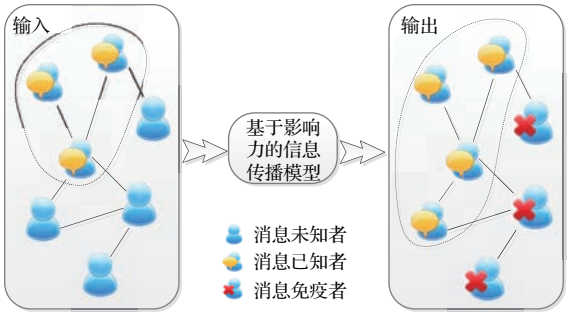

图 1 (网刊彩色) 问题概述

Fig. 1. (color online) Problem overview.

在 t 时间段内, 给定某个特定话题全网用户群体关系网 $G(V,E)$ 以及全网用户的历史行为数据 $A=\{(a,v_i,t)\}$, 解决如下问题.

1) 如何度量用户影响力? 本文模拟用户行为动力学, 结合影响力形成的内部静态要素 $f_{\text{internal}}(v_i)$ 和外部动态要素 $f_{\text{external}}(v_i)$, 用多元线性回归方法定量阐述用户影响力 $Inf(v_i)$.

2) 如何感知信息传播态势? 本文把影响力要素 $Inf(v_i)$ 与传染病模型结合, 为模型中感染率 λ 和恢复率 μ 提供理论依据. 本文通过二项分布计算得到的感染率和免疫率分别为 $\theta(t)$, $1-\theta(t)$. 接着, 依托平均场理论建立动力学微分方程组, 计算不同时间步的用户集合 $\{S_t\}$, $\{I_t\}$, $\{R_t\}$, 模拟不同阶段信息扩散的大致情况.

3 模 型

3.1 模型框架

为了解决上述问题, 构建模型系统框架如图 2 所示. 首先, 根据社会网络中用户个人属性、个人行为习惯和信息交互记录量化用户影响力的内因和外因, 即训练出个人记忆维度和用户交互维度. 接着, 计算处于同一状态下的用户群体相对于

另一状态用户群体的影响力均值作为感染率λ和恢复率μ. 最后, 基于平均场理论, 把λ和μ运用到传染病模型中, 作为微分方程状态转移的参数依据, 模拟信息传播趋势, 感知群体状态演化.

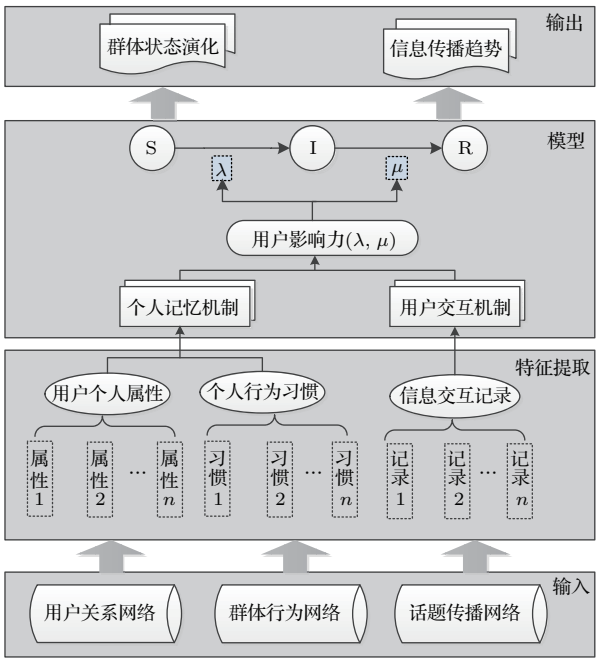

图 2 模型框架

Fig. 2. Model framework.

3.2 模型细化

根据信息传播网络拓扑关系和用户信息互动行为提取两种用户影响力要素, 即静态内部要素和动态外部要素. 形象表示如图3, 左边为个人所处网络静态环境, 右边为基于扩散信息动作的网络.

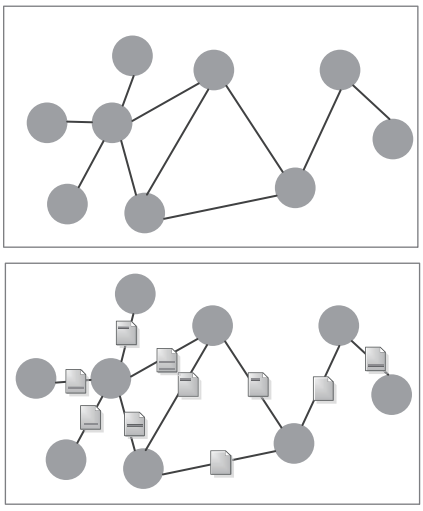

图 3 社交影响力内部和外部要素

Fig. 3. Internal and external factors of social influence.

我们认为信息的传播力不仅与用户的自身网络结构属性有关, 如用户度数、用户介数、用户紧密性等, 还与其外部行为属性有关, 如信息内容相似性、意见领袖、用户是否为活跃用户、信息传播带动力. 综合内因和外因, 用户v_i的影响力函数为

$$Inf(v_i) = \beta_0 + \beta_1 \times f_{\text{internal}}(v_i) + \beta_2 \times f_{\text{external}}(v_i), \tag{9}$$

这里的参数β_0, β_1, β_2是偏回归系数, 由多元线性回归模型训练拟合得出, 其中, β_1, β_2是测试个体内因和外因的权值系数, 反映网络结构和信息交互情况在影响力构成中的比重; $f_{\text{internal}}(v_i)$为用户的内部影响力,

$$f_{\text{internal}}(v_i) = \sum_{j=1}^{j} \frac{\psi_{ij}}{\max\limits_{v\in V}(\psi(v))} \quad (j = 1, 2, 3), \tag{10}$$

其中, ψ_{ij}表示用户v_i的静态结构属性, 可取度数、紧密度、介数等网络结构静态属性, $\max\limits_{v\in V}(\psi(v))$为归一化因子; $f_{\text{external}}(v_i)$为用户的外部影响力,

$$f_{\text{external}}(v_i) = \sum_{j=1}^{j} \chi_{ij} \times \left(\frac{1}{2}\right)^{\frac{t_i - t_i'}{\omega}} \quad (j = 1, 2, 3, 4). \tag{11}$$

由于信息话题影响力具有随着时间推移而逐渐降低的事实, 因此, 本文引入半衰期函数$(1/2)^{\frac{t_i - t_i'}{\omega}}$表示信息从发布到慢慢消亡的生命周期, 其中, t_i表示当前时间, t_i'表示用户v_i上次行为时间, ω为正则化因子, 本文中$\omega = 1000$; χ_{ij}表示用户v_i的动态行为属性, 可取内容相似性、意见领袖、活跃度、信息传播带动力等用户交互属性.

为了验证影响力对信息扩散的作用, 本文采用改进的SIR模型模拟信息传播的过程. SIR模型中的用户群体有三种状态: 易感染状态S (susceptible), 感染状态I (infected)、免疫状态R (recovered). 不同用户间的状态转移不仅依赖于用户自身的状态, 还与其邻居用户的状态相关.

这里要特别说明的是, 该模型的建立基于以下3个假设.

1) 由于信息传播具有爆发性、时长短的特点, 我们认为在研究时间段内用户群体粉丝的增长和减少互相持平, 故不考虑出生率和死亡率等种群因素. 参与信息传播的用户总数始终保持一个常数N, 所以$S + I + R = 1$.

2) 信息传播为接触性传播, 一个新用户与信息已知者接触就必然具有一定的传染率.

3) 假定用户被传播信息后, 经过一段时间(8 h)就会对该消息失去兴趣, 从而变成免疫者. 经过对以上数据的统计, 本文设定消息已知者变为消息免疫者的时间为8 h.

当个体处于感染状态时, 以λ的概率感染处于易感染状态的邻居个体, 以μ的概率恢复为免疫状态. 通过不同状态间的变化, 研究信息传播机制, 动力学方程如下:

$$\begin{cases} \mathrm{d}S/\mathrm{d}t = -\lambda S(t)I(t), \\ \mathrm{d}I/\mathrm{d}t = \lambda S(t)I(t) - \mu I(t), \\ \mathrm{d}R/\mathrm{d}t = \mu I(t), \\ S + R + I = 1. \end{cases} \tag{12}$$

个体感染具有单向性, 用户接受信息的顺序为未感染状态、感染状态、免疫状态. 因此, 假设一个处于某个状态的用户v_i有m个邻居, 则其中k个邻居状态发生改变的概率满足二项分布,

$$P(X = k) = \binom{m}{k} Inf(v_i)^k (1 - Inf(v_i))^{m-k} \quad (k = 0, 1, 2, \cdots, m), \tag{13}$$

则任一用户在时刻t改变状态的概率为

$$\theta(t) = \sum_{k=0}^{m} \frac{k}{m} \binom{m}{k} Inf(v_i)^k (1 - Inf(v_i))^{m-k} \quad (k = 0, 1, 2, \cdots, m). \tag{14}$$

结合平均场方程式(12)得

$$\begin{cases} \mathrm{d}S/\mathrm{d}t = -\overline{\theta(t)}S(t)I(t), \\ \mathrm{d}I/\mathrm{d}t = \overline{\theta(t)}S(t)I(t) - (1 - \overline{\theta(t)})I(t), \\ \mathrm{d}R/\mathrm{d}t = (1 - \overline{\theta(t)})I(t), \\ S + I + R = 1. \end{cases} \tag{15}$$

针对在线社会网络中信息传播模式的特点, 结合传染病动力学原理, 提出在线社会网络中新的信息传播扩散模型. 模型考虑不同关键用户对信息传播机理的影响力度, 挖掘在信息传播扩散过程中各因素的地位, 建立不同用户节点的演化方程组, 模拟信息传播扩散的过程, 分析不同类型的用户在网络中的结构特征以及影响信息传播的主要因素.

3.3 模型算法

传统的SIR模型的状态转变参数缺少理论依据, 人为设定的参数具有较大的随机性, 从而产生一定的误差, 导致预测值与当前实际值差距较大的现象. 针对以上局限性, 进行了一些改进, 并以此为基础结合在线社会网络行为属性, 利用平均场理论的基本思想和方法, 建立基于用户影响力的SIR传播模型. 通过模型学习算法, 获取残差值, 并分

表3 模型算法

Table 3. Model algorithm.

Algorithm 1 Model algorithm

Input

network: $G = (V, E)$ action history in different times: $A = \{(a, v_i, t)\}$

Output

factor influence $Inf(v_i)$; infection rate $\theta(t)$, immunization rate $1 - \theta(t)$; $\{S_t\}, \{I_t\}, \{R_t\}$

for each user u_i **do**

  initialize S_0, I_0, R_0

  for each topic t_j of u_i **do**

    compute dynamic influence factor function $\{f_{\text{internal}}(v_i)\}$, $\{f_{\text{external}}(v_i)\}$ from Eqs.(1)—(8);

    perform multiple linear regression to calculate influence factors $Inf(v_i)$ from Eqs.(9)—(11);

    obtain residual, confidence interval, parameter β_0, β_1, β_2

    //perform binomial distribution

    calculate $P(X = k)$ and max term k calculate state changed probability $\theta(t)$ from Eq.(14)

  end for

  execute mean field theory from Eq. (15)

  result: get all influence factor $Inf(v_i)$, infection rate$\theta(t)$, immunization rate $1 - \theta(t)$,

  user set $\{S_t\}, \{I_t\}, \{R_t\}$ at different time

end for

析网络静态因素和交互动态因素对用户影响力的权重地位. 总体来说, 模型学习算法分为两步:1) 调节学习参数, 并根据平均场理论获取不同时间段的信息扩散情况; 2) 通过多元线性回归算法, 挖掘影响力动力学成因. 表3列出了本文所提出模型的算法实施过程.

模型学习算法首先是初始化不同状态的用户群, 利用模型进行首次估计, 得到初始状态用户的影响因子; 然后利用多元线性回归模型不断调节参数, 重复获取不同状态节点的影响权重, 将其作为用户影响力的量化值, 每个节点状态改变的概率作为模型的传染率和免疫率. 本文根据每个信息的初始用户群体数量, 结合信息的生命周期特征, 预测信息的变化趋势. 由于每次建立模型求解用到的原始数据都为话题数据, 保留了真实性. 不断调节学习参数, 对模型进行修正, 找到与实际最相符的状态转移概率, 减小误差, 提高精确度. 其中, 多元线性回归算法对影响力的度量与信息传播SIR模型密切相关, 时间复杂度为$T_{\text{influence}} = O(N)$, 影响因子选择阶段的时间复杂度$T_{\text{factor}} = O(N)$, 信息传播模型执行的时间复杂度$T_{\text{model}} = 1$. 因此, 算法的时间复杂度$T = T_{\text{influence}} + T_{\text{factor}} \sim O(N)$.

4 实验验证与结果讨论

4.1 数据描述

在人们的社交活动中, 很多时候信息是以话题的形式产生和传播的. 研究工作发现, 不同话题具有不同的影响力, 即便是相同的话题在不同的人群中也不尽相同. 所以, 使用不同话题作为信息传播的基本研究对象, 能够从多角度对用户影响力进行细致刻画. 本文基于腾讯微博真实数据进行了实证验证, 选取的热点话题包括“冯小刚电影《私人定制》”(话题 A)、“爸爸去哪儿”(话题 B)、“熊猫女孩儿急需救助”(话题 C), 针对这3 个不同主题的群体事件对模型进行实验验证. 表4列出3个话题的统计数据.

表4 话题数据统计

Table 4. Statistics of topics.

数据集	时间区间	用户数	粉丝数	网络节点数	网络边数	用户行为数
话题 A	2013.12.19—2014.01.04	3504	898126	901038	1399436	26417807
话题 B	2014.05.14—2014.09.04	7022	879780	884839	1075051	40680234
话题 C	2014.02.25—2014.09.07	9626	534459	543824	582529	21449323

本文利用真实信息传播趋势和模型估计值对信息传播模型进行校验, 真实走势与模型估计值越接近, 模型越优; 利用残差检验方法对影响力度量进行校验, 残差序列越平稳, 置信区间落在$stats \in [-1, 1]$越多, 则模型拟合越佳; 利用线性回归模型验证社交影响力形成的内因和外因对信息传播的驱动力大小; 在模型的计算感染率和免疫率基础上, 预估信息传播的最大感染峰值, 为舆情传播提供可控可管思路.

4.2 实验结果与讨论

通过观察上述社会网络中的3个话题在整个生命周期中的演化趋势, 发现或多或少存在信息传播的爆发期, 结合用户影响力, 选定最活跃的时间段进行本文信息传播模型的验证.

本文把每2 h参与话题的人作为一个时间段分片, 设定用户转发或评论话题的有效时间为24 h, 如果超过24 h 没有重新参与话题, 则认为该用户对该话题失去兴趣. 通过观察, 从话题发布算起, 话题A的活跃区间为94—472 h, 话题B的活跃区间为16—2730 h, 话题 C 的活跃区间为18—3598 h. 图4是3个话题不同时间段的实际参与信息传播情况与模型估计值对比.

在图4中, 左边一列图4(a), (c), (e)依次为模型估计的话题 A、话题 B、话题C的信息未知者、信息已知者、信息免疫者的比例. 网络总数为最终传播信息的人群及其所有会接受到该信息的粉丝. 初始免疫者为网络中内因和外因值均为0 的用户群, 即先剔除网络中尤其不活跃的用户; 初始信息已知者为两个时间片内知道该信息的用户; 网络中剩下的用户群为信息未知者. 从实验图中可看出, 在信

息传播过程中, 信息未知者密度不断较少, 信息免疫者密度持续增大, 信息已知者先增大后减小, 总有一个时间段信息已知者会达到峰值, 称为信息爆发期; 随着时间的推移, 三种状态会之间趋于稳定, 即信息处于消亡期, 验证了信息已知者会逐渐演变为信息免疫者, 网络中最终用户群的状态为信息未知者和消息免疫者两种人群. 右边一列图 4 (b), (d), (f) 为活跃时期真实的消息已知者和模型估计值对比, 从图中大致观察到本文所建模型的精确度与真实信息传播趋势较为接近.

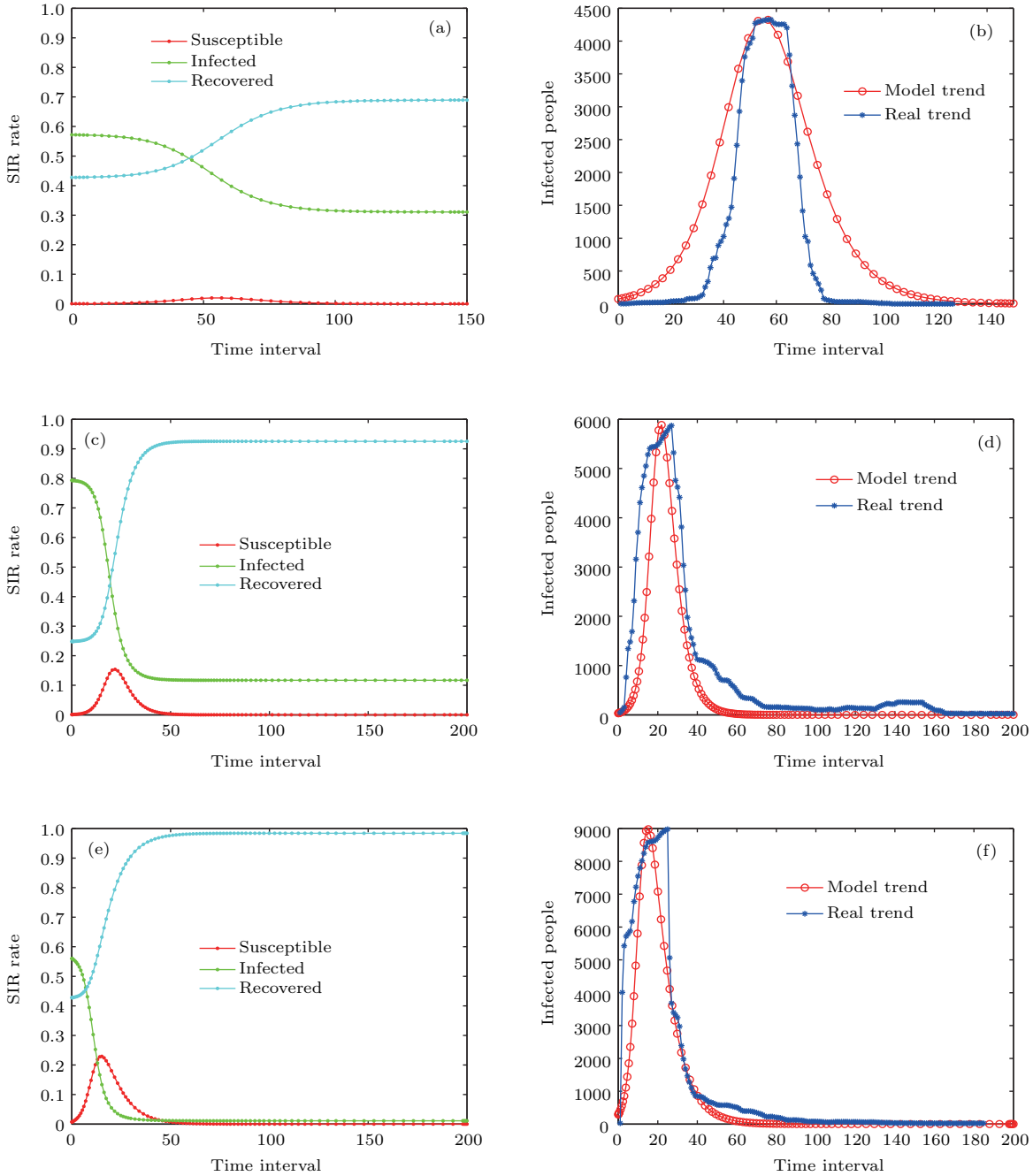

图 4 (网刊彩色) 真实情况和模型结果对比 (a) 话题 A 的 SIR 估计比例; (b) 话题 A 的真实感染人群和模型估计值, 总人数 $N = 216792$; (c) 话题 B 的 SIR 估计比例; (d) 话题 B 的真实感染人群和模型估计值, 总人数 $N = 38188$; (e) 话题 C 的 SIR 估计比例; (f) 话题 C 的真实感染人群和模型估计值, 总人数 $N = 39197$

Fig. 4. (color online) Contrast between real situation and model results contrast: (a) SIR estimated proportion of topic A; (b) contrast between real infected people and model estimation of topic A, $N = 216792$; (c) SIR estimated proportion of topic B; (d) contrast between real infected people and model estimation of topic B, $N = 38188$; (e) SIR estimated proportion of topic C; (f) contrast between real infected people and model estimation of topic C, $N = 39197$.

本文选定初始信息已知者作为研究对象, 运用多元线性回归模型验证影响力内因和外因的拟合程度. 通过残差衡量标准, 获得如图5所示的信息残差序列图, 横坐标为不同用户的编号, 纵坐标为残差值.

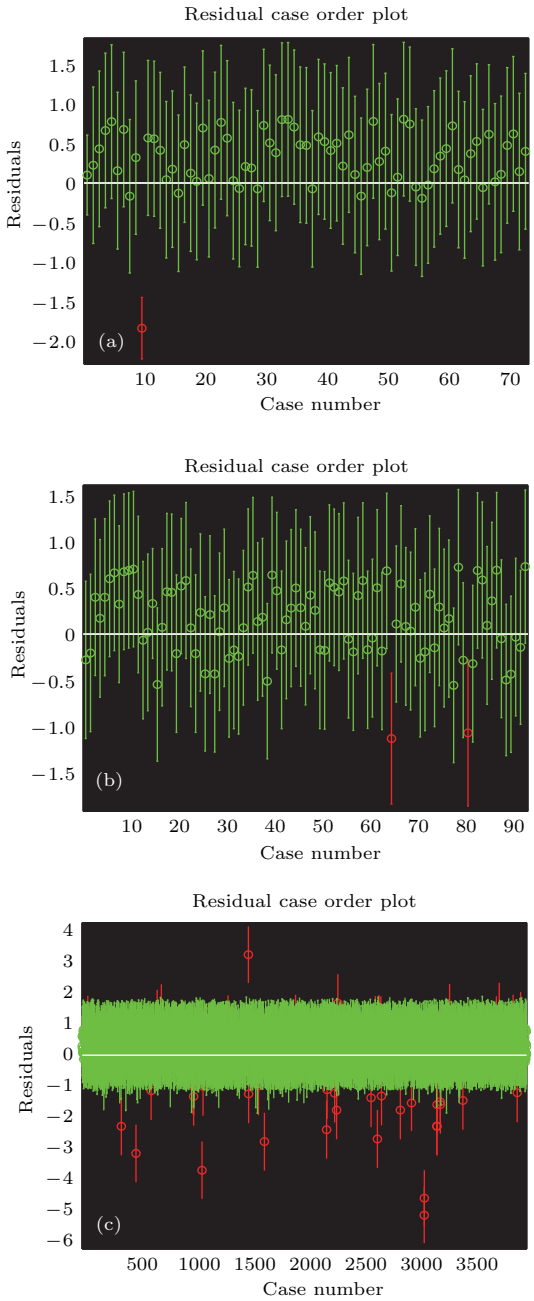

图5 (网刊彩色) 残差序列图 (a) 话题 A; (b) 话题 B; (c) 话题 C

Fig. 5. (color online) Residual plot: (a) Topic A; (b) topic B; (c) topic C.

通过图5可以观察到, 残差序列整体上较平稳, 基本上都落在[−1, 1]的区间, 说明观察值和估计值之差较小, 预估方程符合客观事实, 即内部静态因素和外部动态因素在群体影响力的度量上有一定的相关性.

为了衡量不同静态内部属性对用户内部影响力的影响, 本文分别选取度数、介数、紧密度两两组合进行实验, 得出不同指标与信息已知群体的关系如图6所示. 横坐标表示时间序列, 纵坐标表示消息已知者累积值.

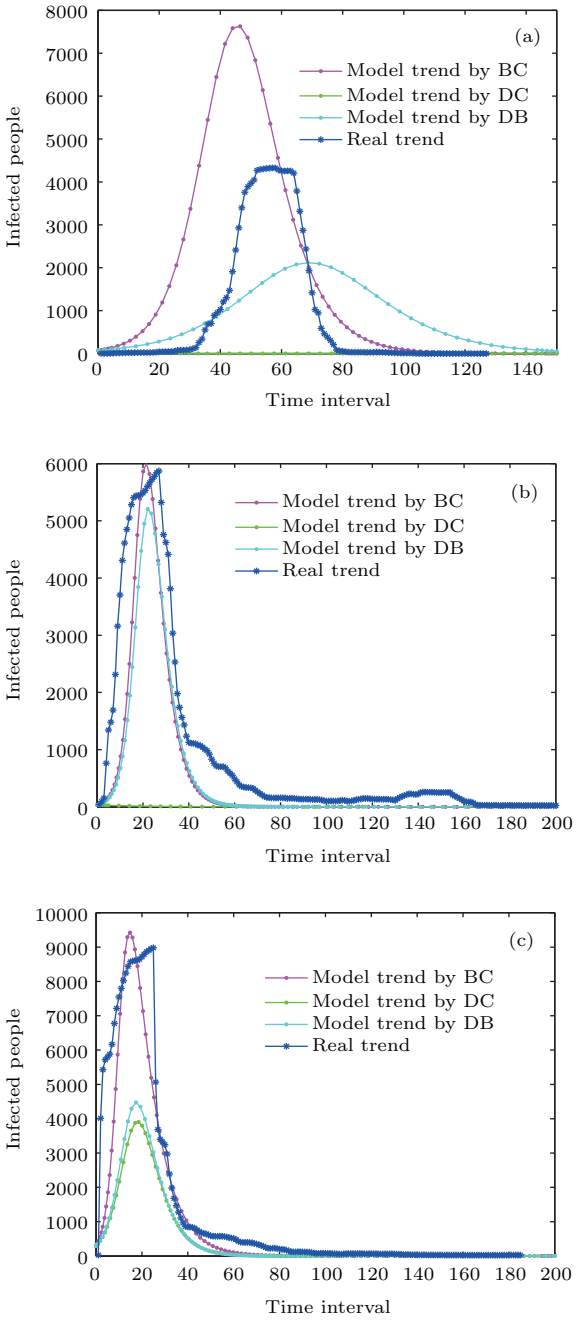

图6 (网刊彩色) 用户内部因素影响变化趋势 (a) 话题 A 内因影响量化图; (b) 话题 B 内因影响量化图; (c) 话题 C 内因影响量化图

Fig. 6. (color online) Trends of user internal influence: (a) Internal factors quantization of topic A; (b) internal factors quantization of topic B; (c) internal factors quantization of topic C.

如图6 (a) 所示, 当除去网络中的度数影响时, 信息扩散峰值普遍增加, 尤其是对话题A 的影响较为明显. 经计算对比, 在初始传播节点中, 话题A 的平均度数为1069.14, 话题B的平均度数为877.3, 话题C的平均度数为956.31. 由此可见, 信息在初始传播节点平均度数大的网络中传播较广, 这与现实是相符的, 往往朋友多的人比朋友少的人更容易成为意见领袖, 促进信息的传播. 如图6 (b) 所示, 当除去网络中介数的影响时, 信息扩散受到较大影响, 甚至难以扩散. 这是因为介数是用来评价节点的流量承载能力的指标, 常用来衡量网络中某一用户向所有用户之间传递信息的重要程度. 若除去重要的节点, 将影响其周围众多节点接收信息. 如图6 (c) 所示, 当除去网络紧密度时, 信息受众最大峰值将会下降. 这是由于网络紧密度反映给定节点到其他节点中可达程度, 若网络不紧密, 信息传播会受到限制.

为了进一步分析内因和外因对信息传播的影响, 本文采用时间分片的方法, 截取信息扩散活跃阶段进行研究, 尝试将社交影响力的内因和外因认同度量化. 如图7所示, 横坐标为活跃时间段 (每2 h为一个时间片), 纵坐标为信息扩散的驱动因子认同度.

通过以话题A、话题B、话题C为例的实验发现, 在线社会网络信息传播的驱动因子在不同时间段会动态变化, 但整体波动不大. 即便静态因素和动态因素均存在一定的差异, 但就总体而言, 在整个信息传播过程中, 外因的作用力比内因的作用力大. 在信息传播后期, 内因的量化值基本上为0, 即内因所起作用不大. 这种情况与现实社会是相符的, 人们在社交网络中不是独立的个体, 人与人之间的信息互动交流是促进其传播最直接的成因, 而不仅仅是依靠网络的静态结构. 也就是说, 针对信息传播和话题演化, 个体记忆并不占据优势地位, 相反却是用户交互在信息传播的生命周期起主导作用.

接下来, 在前面构建的信息传播模型的基础上, 分别固定平均场方程中的感染率和免疫率, 分析三类信息的传播广度. 如图8所示, 从上述三类信息中计算得到的信息传播影响力出发, 取 $\delta = 0.05$ 的步长, 在 (0, 1) 范围内分析信息传播过程中的最大覆盖范围.

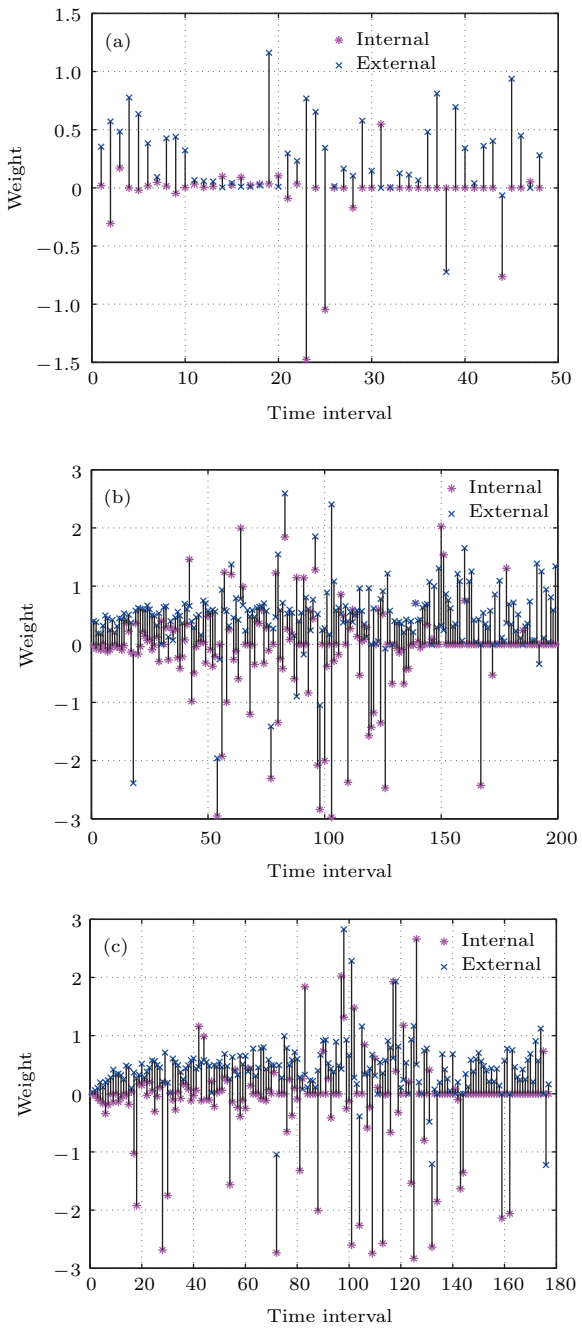

图7 (网刊彩色) 信息传播驱动因子的变化趋势 (a) 话题A驱动因子量化图; (b) 话题B驱动因子量化图; (c) 话题C驱动因子量化图

Fig. 7. (color online) Trends of information dissemination driving factors: (a) Driving factors quantization of topic A; (b) driving factors quantization of topic B; (c) driving factors quantization of topic C.

从图8的实验结果可以发现, 最大感染峰值MIP(max infected peak) 在三类信息中明显地表现为随着感染率的增大而上升, 随着免疫率的减小而下降. 最大感染峰值的变化表明用户单位时间内信

息的传播力度是有差异的. 显然, MIP 值越大, 表明单位时间内信息扩散的速度越快. MIP 作为衡量用户传播信息速度和信息覆盖程度的指标, 很大程度上受感染率和免疫率的直接影响. 如果感染率很大, 信息很快就会扩散, 一旦控制不当, 甚至能影响整个网络; 如果免疫率很小, 信息也会传播得很快, 并且覆盖整个网络. 特别地, 如果感染率控制在较小的范围, 免疫率控制在较大的范围, 信息就能较好地维持在初始感染水平, 知道信息的用户不会爆发式增长. 所以, 相关部门可以从信息感染率和免疫率两方面着手, 对网络舆情进行有效管控.

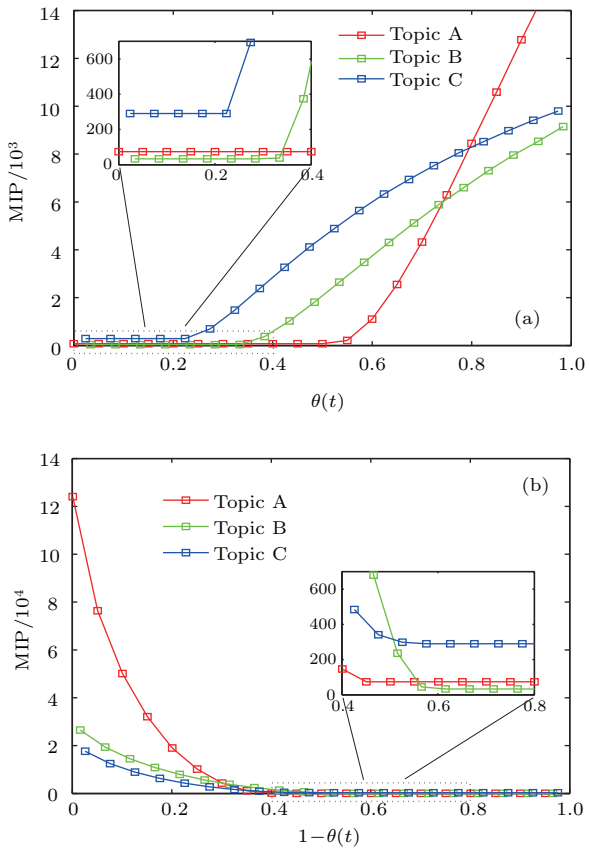

图 8 (网刊彩色) 感染人数峰值 (a) 感染峰值随 $\theta(t)$ 的变化情况: topic A, $1-\theta(t)=0.301$; topic B, $1-\theta(t)=0.266$; topic C, $1-\theta(t)=0.126$; (b) 感染峰值随 $1-\theta(t)$ 变化情况: topic A, $\theta(t)=0.699$; topic B, $\theta(t)=0.734$; topic C, $\theta(t)=0.874$

Fig. 8. (color online) Peak of infected people: (a) Peak change with $\theta(t)$, topic A, $1-\theta(t)=0.301$; topic B, $1-\theta(t)=0.266$; topic C, $1-\theta(t)=0.126$; (b) peak change with $1-\theta(t)$, topic A, $\theta(t)=0.699$; topic B, $\theta(t)=0.734$; topic C, $\theta(t)=0.874$.

5 结 论

通过分析人们相互之间的影响模式和信息传播方式, 既能够从社会学角度加深理解人们的社会行为, 为公共决策和舆情导向等提供理论依据, 同时还能促进政治、经济和文化活动等多个领域的交流和传播, 具有重要的社会意义和应用价值. 考虑到在线社交影响力及其传播过程受诸多因素制约, 而且不同因素之间往往相互影响, 本文对具有特定属性的用户深入探讨并对用户与信息之间的交互行为进行分析和度量. 在此基础上, 本文采用改进的 SIR 模型作为信息传播模型, 将用户群体划分为信息未知者、信息已知者、信息免疫者三种状态, 并把量化后的影响力作为状态转移参数的理论支撑. 研究表明, 本文优化的模型可以从更为宏观的层面上理解社交影响力的作用原理和信息传播机制, 解释信息传播的内部和外部动力学成因, 探知挖掘人类的行为特征和行为规律, 促进该领域知识的进步和发展, 为社会感知信息传播态势和制定信息传播管控策略提供理论依据.

参考文献

[1] Cai M, Du H F, Feldman M W 2014 *Acta Phys. Sin.* **63** 060504 (in Chinese) [蔡萌, 杜海峰, Feldman M W 2014 物理学报 **63** 060504]

[2] Zhou B, He Z, Jiang L L, Wang N X, Wang B H 2014 *Sci. Rep.* **4** 7577

[3] Huang J, Li C, Wang W Q, Shen H W, Li G, Cheng X Q 2014 *Sci. Rep.* **4** 5334

[4] Zheng M, Lü L, Zhao M 2013 *Phys. Rev. E* **88** 012818

[5] Chen D B, Huang S, Shang M S 2011 *Comput. Sci.* **38** 118 (in Chinese) [陈端兵, 黄晟, 尚明生 2011 计算机科学 **38** 118]

[6] Chen D B, Gao H 2012 *Chin. Phys. Lett.* **29** 048901

[7] Borge-Holthoefer J, Moreno Y 2012 *Phys. Rev. E* **85** 026116

[8] Wang C, Liu C Y, Hu Y P, Liu Z H, Ma J F 2014 *Acta Phys. Sin.* **63** 180501 (in Chinese) [王超, 刘骋远, 胡远萍, 刘志宏, 马建峰 2014 物理学报 **63** 180501]

[9] Lu W, Chen W, Lakshmanan L V S 2015 *Proceedings of the VLDB Endowment* Hawaii, USA, August 31–September 4, 2015 p60

[10] Elias Boutros K, Dilkina B, Song L 2014 *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* New York, USA, August 24–27, 2014 p1226

[11] Singer Y 2012 *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* Washington, USA, February 8–12, 2012 p733

[12] Liu J G, Lin J H, Guo Q, Zhou T 2016 *Sci. Rep.* **6** 21380

[13] Montanari A, Saberi A 2010 *Proc. Natl. Acad. Sci.* **107** 20196

[14] Yuan X P, Xue Y K, Liu M X 2013 *Chin. Phys. B* **22** 030207

[15] Pastor-Satorras R, Castellano C, van Mieghe P, Vespignani A 2015 *Rev. Mod. Phys.* **87** 925

[16] Kempe D, Jon K, Éva T 2003 *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Washington, USA, August 24–27, 2003 p137

[17] Liu Q M, Deng C S, Sun M C 2014 *Physica A* **410** 79

[18] Li C H, Tsai C C, Yang S Y 2014 *Commun. Nonlinear Sci.* **19** 1042

[19] Chen L, Sun J 2014 *Physica A* **410** 196

[20] Xiong F, Liu Y, Zhang Z J, Zhu J, Zhang Y 2012 *Phys. Lett. A* **376** 2103

[21] Li T, Wang Y, Guan Z H 2014 *Commun. Nonlinear Sci.* **19** 686

[22] Xiong F, Wang X M, Cheng J J 2016 *Chin. Phys. B* **25** 108904

[23] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nat. Phys.* **6** 888

[24] Hu Q C, Zhang Y, Xu X H, Xing C X, Chen C, Chen X H 2015 *Acta Phys. Sin.* **64** 190101 (in Chinese) [胡庆成, 张勇, 许信辉, 邢春晓, 陈池, 陈信欢 2015 物理学报 **64** 190101]

[25] Lü L Y, Chen D B, Ren X L, Zhang Q M, Zhang Y C, Zhou T 2016 *Phys. Rep.* **650** 1

[26] Wu Y, Yang Y, Jiang F, Jin S, Xu J 2014 *Physica A* **416** 467

[27] Liben-Nowell D, Kleinberg J 2008 *Proc. Natl. Acad. Sci.* **105** 4633

[28] Lü L Y, Zhou T, Zhang Q M, Stanley H E 2016 *Nat. Commun.* **7** 10168

[29] Myers S A, Zhu C, Leskovec J 2012 *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Beijing, China, August 12–16, 2012 p33

[30] Aral S, Walker D 2012 *Science* **337** 337

[31] La Fond T, Neville J 2010 *Proceedings of the 19th International Conference on World Wide Web* Raleigh, USA, April 26–30, 2010 p601

[32] Li P, Zhang J, Xu X K, Small M 2012 *Chin. Phys. Lett.* **29** 048903

[33] Newman, M E J 2012 *Nat. Phys.* **8** 25

[34] Barabási A L, Albert R, Jeong H 1999 *Physica A* **272** 173

[35] Pedroche F, Moreno F, González A, Valencia A 2013 *Math. Comput. Model* **57** 1891

An information diffusion dynamic model based on social influence and mean-field theory*

Xiao Yun-Peng† Li Song-Yang Liu Yan-Bing

(*Chongqing Engineering Laboratory of Internet and Information Security, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*)

(Received 7 June 2016; revised manuscript received 18 October 2016)

Abstract

With the development of online social networks, they rapidly become an ideal platform for information about social information diffusion, commodity marketing, shopping recommendation, opinion expression and social consensus. The social network information propagation has become a research hotspot correspondingly. Meanwhile, information diffusion contains complex dynamic genesis in online social networks. In view of the diversity of information transmission, the efficiency of propagation and the convenience of interaction, it is very important to regulate the accuracy, strengthen the public opinion monitoring and formulating the information control strategy.

The purpose of this study is to quantify the intensity of the influence, especially provides a theoretical basis for studying the state transition of different user groups in the evolution process. As existing epidemic model paid less attention to influence factors and previous research about influence calculation mainly focused on static network topology but ignored individual behavior characteristics, we propose an information diffusion dynamics model based on dynamic user behaviors and influence. Firstly, according to the multiple linear regression model, we put forward a method to analyze internal and external factors for influence formation from two aspects: personal memory and user interaction. Secondly, for a similar propagation mechanism of information diffusion and epidemics spreading, in this paper we present an improved SIR model based on mean-field theory by introducing influence factor.

The contribution of this paper can be summarized as follows. 1) For the influence quantification, different from the current research work that mainly focuses on network structure, we integrate the internal factors and external factors, and propose a user influence evaluation method based on the multiple linear regression model. The individual memory principle is analyzed by combining user attributes and individual behavior. User interaction is also studied by using the shortest path method in graph theory. 2) On modeling the information diffusion, by referring SIR model, we introduce the user influence factor as the parameter of the state change into the epidemic model. The mean-field theory is used to establish the differential equations. Subsequently, the novel information diffusion dynamics model and verification method are proposed. The method avoids the randomness of the artificial setting parameters within the model, and reveals the nature of multi-factors coupling in the information transmission.

Experimental results show that the optimized model can comprehend the principle and information diffusion mechanism of social influence from a more macroscopic level. The study can not only explain the internal and external dynamics genesis of information diffusion, but also explore the behavioral characteristics and behavior laws of human. In addition, we try to provide theoretical basis for situation awareness and control strategy of social information diffusion.

Keywords: information dissemination, diffusion dynamics, individual influence, mean-field theory

PACS: 05.10.–a, 89.75.Fb **DOI:** 10.7498/aps.66.030501

* Project supported by the National Basic Research Program of China (Grant No. 2013CB329606), the National Natural Science Foundation of China (Grant No. 61272400), the Chongqing Youth Innovative Talent Project, China (Grant No. cstc2013kjrc-qnrc40004), the Foundation of Ministry of Education of China and China Mobile (Grant No. MCM20130351), the Chongqing Graduate Research and Innovation Project, China (Grant No. CYS14146), the Science and Technology Research Program of the Chongqing Municipal Education Committee, China (Grant No. KJ1500425), the WenFeng Foundation of Chongqing University of Post and Telecommunications, China (Grant No. WF201403).

† Corresponding author. E-mail: xiaoyp@cqupt.edu.cn