第 44 卷 第 11 期 计 算 机 工 程 2018 年 11 月
Vol. 44 No. 11 Computer Engineering November 2018

·开发研究与工程应用· 文章编号：1000-3428(2018)11-0282-07 文献标志码：A 中图分类号：TP393

基于转发行为影响因素的 SCIR 信息传播模型

张 永，华姗姗，张 航

（兰州理工大学 计算机与通信学院，兰州 730050）

摘 要：在线社交网络上信息传播模型的构建和信息传播规律的研究，对于抑制恶意信息传播，加强网络舆论监管力度具有重要意义。通过对热点网络事件的分析，在传统 SIR 模型基础上加入一类新节点，建立信息传播 SCIR 模型。以用户交互关系和转发行为影响因素为分析基点，对转发行为相关的 5 个特征量做出具体分析与算法描述。通过流行病动力学演化方程实现信息传播过程的理论分析，采用不同网络中 SIR 模型和 SCIR 模型进行差异性分析。实验结果表明，相对 SIR 模型，SCIR 模型具有更好的稳定性与信息传播覆盖性。

关键词：信息传播；转发行为；SIR 模型；SCIR 模型；行为影响力

中文引用格式：张 永，华姗姗，张 航. 基于转发行为影响因素的 SCIR 信息传播模型[J]. 计算机工程，2018，44(11)：282-288，299.

英文引用格式：ZHANG Yong，HUA Shanshan，ZHANG Hang. SCIR information propagation model based on influencing factors of forwarding behavior[J]. Computer Engineering，2018，44(11)：282-288，299.

SCIR Information Propagation Model Based on Influencing Factors of Forwarding Behavior

ZHANG Yong，HUA Shanshan，ZHANG Hang

（Department of Computer and Communication，Lanzhou University of Technology，Lanzhou 730050，China）

【Abstract】The research on the construction of information propagation model and the law of information transmission on the online social network is of great significance to restrain the spread of malicious information and strengthen the supervision of public opinion on the internet. Based on the analysis of hotspot events，a kind of new node named carrier node is added on the basis of the traditional Susceptible Infected Removed（SIR） model，and a new model named Susceptible Contacted Infected Removed（SCIR） model is established to describe the information diffusion better in online social network. As the analysis of the relationship between user interaction and forwarding behavior，the five characteristics of forwarding behavior are analyzed and described. By the theoretical analysis of the evolution equation of the epidemic dynamics，the difference between the SIR model and the SCIR model is given. Experimental results show that compared with the SIR model，the SCIR model has better stability and information coverage.

【Key words】information propagation；forwarding behavior；Susceptible Infected Removed（SIR） model；Susceptible Contacted Infected Removed（SCIR） model；behavior influence

DOI：10.19678/j.issn.1000-3428.0048694

0 概述

目前在 OSNs（Online Social Networks）上的用户不再是被动接受信息的媒体受众，而是通过建立单向或双向好友关系交流、分享信息资源，成为信息的制作者、分享者和传播者，积极参与到网络活动中[1]。Facebook、Twitter 作为国外主流的 OSNs 应用，其热门话题的传播速度明显优于报纸、杂志等传播途径。而国内流行的舆论扩散窗口新浪微博和即时通信工具，也是目前应用范围较广的在线社交网络服务平台。较之即时通信工具的私密性，新浪微博作为140 字内的短文本观点交流和转发他人言论的自由社交应用，已成为人们获取及时新闻和高舆论热点话题的重要工具之一。

由于流行病扩散与信息传播的相似性，用于研究疾病传播的动力学传染病模型也广泛应用于信息传播领域。为了使疾病传播模型更适用于在线社交网络上的信息传播研究，传染病模型的演化和改进也成为研究热点。文献[2]提出改进的 SI（Suscepticble Infected）模型。文献[3]提出考虑概率递减机制

作者简介：张 永（1963—），男，教授，主研方向为智能信息处理、数据挖掘；华姗姗、张 航，硕士研究生。

收稿日期：2017-09-15 **修回日期：**2017-10-26 **E-mail：**418060779@qq.com

(Decreasing Probability Susceptible Infected, DPSI)模型。文献[4]提出引入潜伏节点 E 的 SEIR 模型。文献[5]提出考虑直接免疫策略的改进 SIR(Susceptible Infected Removed)模型。文献[6]将网络流通量代入 SIR 模型中,发现均匀的负载分布有利于信息传播,大度节点对信息传播有双重作用,而通信量拥挤会阻碍信息传播。文献[7]提出的 SIHR(Susceptible Infected Hibernator Removed)谣言传播模型,考虑隐退节点 H,并引入遗忘机制和记住机制。

目前,有学者从谣言抑制、流行病阈值、用户行为影响因素分析方面展开社交网络上信息传播规律的研究[8-14]。文献[8]提出使用意见领袖方法从扩散源抑制不良言论的扩散。文献[9]采用局部策略方法从扩散过程实现有效抑制。文献[10]把疾病传播和信息传播分别放到接触层和通信层独立考虑并得出:疾病爆发会导致信息传播扩大化;疾病信息扩散可有效提高疾病爆发阈值,抑制疾病传播;信息传播阈值不变,但是提高了流行病阈值。文献[11]将自我意识、自我保护和免疫策略与 SIR 模型结合起来综合分析流行病阈值和免疫阈值的变化。文献[12]把意识分为私人意识和公众意识,从具体行为意识层面研究疾病的动态传播。文献[13]以 SIR 为基底模型研究记忆力对疾病传播阈值的影响。

以上文献虽都有行为因素分析,但多数源于人行为意识的主观因素,而在实际社交网络中,信息转发是信息传播的重要途径。因此,本文通过对影响用户转发行为相关因素的分析定义转发影响力公式,并以此分析用户间的信息传播效率。

1　舆论事件分析

2016 年 11 月 20 日 0 点—2016 年 11 月 23 日 16 点京昆高速多车相撞(舆论总数为 22 206)和日本福岛地震(讨论量为 52 849)成为实时热门搜索新闻事件,利用新浪微舆情工具(http://wyq.sina.com/login.shtml)进行分析,结果分别如图 1(a)和图 1(b)所示。从图 1 可以看出,京昆高速多车相撞事件呈现蒲公英式传播模型[15],以源点为中心的大范围传播后出现碎片化小范围扩散。即在首轮传播热度过后趋于平淡,很少转发或者只出现少量小伞状二次传播。而日本福岛这一实时事件呈现双子星传播模式[15],整个传播过程中仅存在 2 个影响力比较接近的中心节点。林丹出轨这一娱乐事件一周讨论量高达1 064 494。同样使用新浪微舆情分析工具进行事件全网传播分析,结果表明,其传播方式完整地呈现出波纹式传播模型[15],如图 1(c)所示。这种传播方式主要集中在以源点博文为核心的传播圈内,表现出从中心向外围扩散的特点。首次传播热度过后,将很少或不出现二次传播热潮。

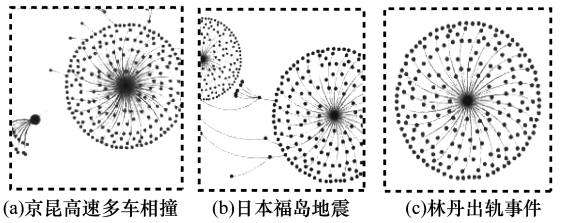


图 1　热点舆论事件分析

综上分析,热点事件在网络上的传播并不是无规律可循,一般会存在 1 个或 2 个高潮时期,但是舆论热潮过后事件就会趋于平静。事件传播通常是以大 V 用户为传播源头,他们的粉丝为信息散播分支,实现树型扩散式传播。由于分支越来越多,事件传播规模会随之扩大,但事件影响力却变小。随着时间的推移,讨论数或传播量会降低至零,表明事件全网传播进入尾声,该事件不再具有传播价值。为了更好地描述以上事件传播过程,本文通过 SCIR 模型动态模拟事件传播过程,并将该模型的传播效果与 SIR 模型进行比较,从而更加深刻具体地描绘社交网络上信息传播的复杂性。

2　转发行为影响因素分析

2.1　用户关系和行为分析

如果从关注机制来考虑社交网络中的用户间关系,那么用户 u 与 v 之间的交互关系可分为陌生人、关注、被关注、双向关注 4 种。其中,以 B 为源传播者,上述 4 种交互关系如图 2 中的(4)、(3)、(2)、(1)所示。同样以 B 为中心的用户行为传播信息树图如图 3 所示,以 B 为信息传播源点呈树型扩散,信息覆盖范围越来越广。

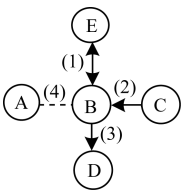


图 2　以 B 为中心的用户关系

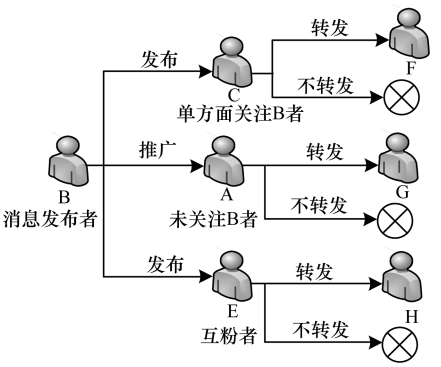


图 3　以 B 为源点的用户行为树

2.2 特征提取

2.2.1 消息转发力度

微博中消息传播主要是靠转发行为实现,因而转发力度可以作为影响转发因素的重要参数。本文将潜在转发人群的概率定义为转发力度。从图 2 和图 3 综合分析可得,用户 B 发布信息的极大化传播跟粉丝的转发力度密不可分。因此,本文根据用户之间关注行为的单向性、双向性以及特定时间内兴趣所趋路转粉联系将文中的粉丝用户群分为纯粉、互粉和路转粉 3 类,但各类粉丝对于消息传播的影响程度是有区别的。如图 2 所示,纯粉指只存在对用户 B 的单向关注行为的粉丝群;互粉指用户 B 关注该用户并且此用户也关注用户 B;路转粉指本来未关注用户 B,但因为兴趣所趋,成为用户 B 的粉丝,但用户 B 并未关注该用户。则转发力度公式如下:

$$Weight_{\text{ret}}=\frac{p_1\cdot F_{\text{Pure}}+p_2\cdot F_{\text{Mutual}}+p_3\cdot F_{\text{passerby}\rightarrow\text{fan}}}{Fans} \tag{1}$$

$$Fans=F_{\text{Pure}}+F_{\text{Mutual}}+F_{\text{passerby}\rightarrow\text{fan}} \tag{2}$$

其中,p_1、p_2、p_3 指各类粉丝的影响权重,为了便于权重衡量,本文规定纯粉、互粉和路转粉的权重分别为 2、1、0.5。

2.2.2 用户影响力

在信息传播过程中,信息传播个体的传播影响力决定着消息的传播广度和流行时长。一般认为粉丝数多的用户影响力值也相对较高。粉丝对影响力用户博文转发量越多,消息的传播面越广,用户影响力也会因此而提高。但是由于消息来源的广度与关注者数正相关。用户关注者数越多,接收信息的范围会随之扩大,那么该用户微博被转发的概率会增加,影响力也会提高。因此,文中用户影响力评估主要考虑 2 个方面:粉丝对博文原创者发布信息的转发传播和消息来源的接受广度。综合考虑关注者和粉丝活跃度的 PageRank[16] 用户影响力评估算法描述如下:

$$R(u)=(1-d)+d\Big(\sum_{v\in Fans(u)}\eta_{u,v}\cdot R(v)+\sum_{w\in Followers(u)}\zeta_{u,w}\cdot R(w)\Big) \tag{3}$$

其中,$R(u)$ 指的是用户 u 的影响力,d 为阻尼系数,一般情况下取 $d=0.85$,假定初始 R 值为 1。$Fans(u)$ 表示节点 u 的粉丝总集合,$Followers(u)$ 表示节点 u 的关注者总集合,$\eta_{u,v}$ 是节点 u 分配给粉丝节点 v 的 R 值的比例,$\zeta_{u,w}$ 是节点 u 分配给关注者节点 w 的 R 值的比例。

$$\eta_{u,v}=\frac{A_v}{\sum_{k=1}^{N}A_k},\zeta_{u,w}=\frac{A_w}{\sum_{m=1}^{M}A_m} \tag{4}$$

其中,A_k 为节点 u 的第 k 个粉丝节点的活跃度,A_m 为节点 u 的第 k 个关注者节点的活跃度,N 为节点 u 的粉丝总数,M 为节点 u 的关注者总数。

对于粉丝节点 v,其活跃度 A 表示为:

$$A=\frac{\sum_{t=1}^{T}(Rt_t+Thu_t+C_t)}{T} \tag{5}$$

对于关注节点 u,其活跃度 A' 可以表示为:

$$A'=\frac{\sum_{t=1}^{T}(F_t+F'_t+P_t+Rt_t+Thu_t+C_t)}{T} \tag{6}$$

其中,F_t、F_t'、P_t、Rt_t、Thu_t、C_t 分别指 T 天内用户关注朋友数目、用户粉丝数、发布微博数、转发微博数、点赞数和评论数。

2.2.3 个体信任水平的衡量

用户之间的信任水平使用二值函数来衡量,其中 0、1 分别表示个体间的非信任和信任状态:值为 1 表示 2 个个体处于不同状态,接触后状态发生改变,信息可以传播;值为 0 表示 2 个个体属于相同状态,状态不发生变化,消息不会传播。在信息传播过程中,携带者节点接触传播节点,那么信任值为 1,携带者会感染成为信息传播者,进行消息传播;而携带者接触携带者,信任值为 0,接触个体仍处于原态,不具备信息传播能力。式(7)中 $state_i$ 表示的是节点 i 的状态,i 指的是处于 S、C、I、R 任意一种状态。

$$tr=\begin{cases}0, state_i=state_j\\ 1, state_i\neq state_j\end{cases} \tag{7}$$

2.2.4 内容相似性

用户之间进行信息交换,趣味是否相投至关重要。用户对于获取信息的感兴趣程度,文中表述为兴趣相似值。用户之间的兴趣相似值将从两用户发布博文的内容相似度来具体衡量。即发布博文内容相似性越高,用户间的兴趣相似值越高,那么其中一个用户对另一用户发布博文转发的概率也会增大。

本文内容相似度算法只适用于长度不大于 140 字的博文,并且使用空间向量余弦算法[17] 来计算 2 个用户博文的相似比重,其算法描述如下:

1)对于微博中出现频率高但无意义的词、符号、表情、标点、特殊连接词等进行剔除,比如:吃、这、的、会、了、吧、哎呀;[、] 、《 》; 、 、 ;。、-、,、……;_、#、@ XXX、&、http://XXX 等。

2)将剔除掉 1)中所列无意义字符后的博文 U 分为若干关键词词组,则用户 K 的一篇博文可记为:$U_K=\{u_1,u_2,\cdots,u_n\}$,其中,$u_i$ 表示连续划分的第 i 个关键词组,$1\leqslant i\leqslant n$。

3)对 N 词博文中出现的 M 次 w 词组,计算其频次 F:

$$F=\frac{N_w}{N} \tag{8}$$

而 w 的权重指数 $index$ 为:

$$index=\ln\frac{N}{N_w} \tag{9}$$

针对关键词的权重计算公式如下:

$$W=F\times index \tag{10}$$

4)假设每个博文中词与词之间不相关,那么博文 C_u 和 C_v 的内容相似性计算公式如下:

$$Similarity(C_u,C_v)=\frac{(\sum_{i=1}^{n}F_i\cdot index_i)\times(\sum_{j=1}^{n}F_j\cdot index_j)}{\sqrt{(\sum_{i=1}^{n}(F_i\cdot index_i)^2)\times(\sum_{j=1}^{n}(F_j\cdot index_j)^2)}} \tag{11}$$

其中,C_u、C_v 分别表示用户 u 和用户 v 的一篇博文,i 表示 C_u 中的特征词数,j 表示 C_v 中的特征词数。

2.2.5　有效转发率

有效转发率在信息传播过程指的是用户 u 的博文被用户 v 转发的有效概率,算法如下:

$$Fr_{v\to u}=\frac{q_1R_{\text{original}}+q_2R_{\text{forward}}}{R_u} \tag{12}$$

其中,q_1、q_2 分别指的是用户 v 转发用户 u 的原创微博和转发微博的权重,R_{original}是用户 u 的原创微博数,R_{forward}是用户 u 的转发微博数,R_u 是用户 u 总的博文数。

2.3　影响力计算

针对转发过程中已知源点用户 u 对于转发用户 v 在目标微博内容[18] w 下的影响力 $Influence(u,v,w)$ 计算公式如式(13)所示。

$$\begin{cases}Influence(u,v,w)=\dfrac{1}{1+\dfrac{1}{Y^{\vartheta}}}\\ Y=[Fr_{v\to u}^{Weight_{\text{ret}}},R(u),tr,Similarity(C_i,C_j)]\\ \vartheta=[\vartheta_1,\vartheta_2,\vartheta_3,\vartheta_4]\end{cases} \tag{13}$$

其中,$Influence(u,v,w)$表示用户 u 的博文内容 w 对于用户 v 在各特征向量影响下的具体转发行为影响力,Y 表示 2.2 节中各个特征量,特别指出 $Fr_{v\to u}^{Weight_{\text{ret}}}$表示转发力度为 $Weight_{\text{ret}}$时博文 w 的有效转发 $Fr_{v\to u}$的指数特征量,对应特征量值的计算式如式(1)、式(3)、式(7)、式(11)和式(12)所示。ϑ 是特征量权重,值采用最大似然估计算法来计算。由式(13)得出,$Influence(u,v,w)$收敛,且取值区间为[0.5,1]。

3　SCIR 模型

在现实生活中,当个体处于感染状态时,要通过接种、免疫、隔离等保护措施避免其他个体不被感染是比较困难的。但是在 SI 演化过程中引入过渡型节点,对处于过渡状态的个体采取接种免疫等有效保护措施可有效减少个体转变为感染者的概率,甚至降低疾病的有效传播速率以及爆发规模,从而实现有效抑制疾病传播的目的[18]。与文献[16]中将 C 节点定义为接触状态相比,本文将无意识接收信息但并未进行信息传播的 C 类型节点定义为携带者节点,使其更贴近现实生活。本文使用文献[16]中 SCIR 模型状态转换过程分析,考虑式(13)中转发行为影响力对信息传播的影响,以及建模过程不同网络中模型的对比分析、关键参数分析和网络信息覆盖率讨论,从而构建出文中提出的信息传播 SCIR 模型。

SCIR 模型中信息传播存在如下规则:个体未从任何渠道获取热点事件传播源,那么在时间片 $1/\delta$ 后个体将以概率 δ 自动蜕变为持久免疫节点,不再参与信息传播;一个源点传播个体发布一条博文后,其粉丝对用户发布的博文以概率 α 进行转发或者评论以传递给更多的人;当未知者浏览传播信息时,他已了解信息但是并未产生传播意识,那么未知者将以概率 ε 转变成观望者;观望者可能因为对传播信息感兴趣以概率 p 对其进行转发或只是出于单纯的浏览,一段时间后直接以遗忘速率 μ 将浏览过的信息淡忘蜕变为免疫态;而原有博文被广泛传播或者热潮过后可能失去自身吸引力,以概率 β 被人们淡忘或者不再具有传播价值,那么它将不再以任何形式被传播。

SCIR 模型中仍设人口总数不变,并且认为转发影响力 $Influence(u,v,w)$ 等于直接传播速率 p,用 $S(t)$、$C(t)$、$I(t)$、$R(t)$分别表示 t 时段易感染者、携带者、感染者、康复者的人口总数,假定人口总数为 $N(t)$,那么 $S(t)+C(t)+I(t)+R(t)=N(t)$。对应态人口密度分别记作 $s(t)$、$c(t)$、$i(t)$、$r(t)$,则各态人口密度变化公式表示如下:

$$\begin{cases} s \xrightarrow{\delta} r \\ c \xrightarrow{\mu} r \\ i \xrightarrow{\beta} r \\ \qquad\qquad \Rightarrow \\ s + c \xrightarrow{\varepsilon} c + c \\ c + i \xrightarrow{p} i + i \\ s + i \xrightarrow{\alpha} i + i \end{cases}$$

$$\begin{cases} \dfrac{\mathrm{d}s(t)}{\mathrm{d}t} = - <k>\alpha s(t)i(t) - <k>\varepsilon s(t)c(t) - \delta s(t) \\ \dfrac{\mathrm{d}c(t)}{\mathrm{d}t} = <k>\varepsilon s(t)c(t) - <k>pc(t)i(t) - \mu c(t) \\ \dfrac{\mathrm{d}i(t)}{\mathrm{d}t} = <k>\alpha s(t)i(t) + <k>pc(t)i(t) - \beta i(t) \\ \dfrac{\mathrm{d}r(t)}{\mathrm{d}t} = \delta s(t) + \mu c(t) + \beta i(t) \\ s(t) + c(t) + i(t) + r(t) \equiv 1 \\ \delta + \varepsilon + \alpha = 1 \\ p + \mu = 1 \end{cases} \tag{14}$$

其中，$<k>$ 为初始接入网络的节点单位时间度平均值，α、p 分别表示 S 态、C 态接触 I 态变为感染者的速率，α 定义为外部感染速率，p 定义为内在传播速率，$p = Influence(u, v, w)$，ε 表示用户无意识情况看到过传播信息，但是并未转发的速率，δ 表示用户微博中从未出现传播信息，或用户对微博中其他用户的信息并不感兴趣，在短时间后，直接蜕变为免疫者的概率，μ 表示用户在无意识情况下浏览过传播信息，但是还未转发该信息，已经忘记该内容的存在，因此指的是遗忘率，β 表示感染个体的康复速率。

初始状态，假设 $S(0) = N - 1$，$I(0) = 1$，$C(0) = 0$，$R(0) = 0$，即除了一个传播者节点，其余节点全部为易感染节点。特别地，$<k> = 1$，$\varepsilon = p = \mu = \delta = 0$，SCIR 模型将蜕变为经典的 SIR 模型。

4 实验分析

本文在 Windows 7 系统中用 Anaconda 2 和 PyCharm 5.0.1 搭建 python 运行环境，并在 python 2.7 版本中使用 networkx 工具包生成 BA 和 WS 网络。然后用生成的网络为底图构建对应网络下的 SCIR 模型，并对式(13)中的权重 ϑ 进行参数学习，以最优目标选择作为优值选取条件，辅之模型中的其他变量参数建模。最后在对应网络下进行实验结果对比分析和在新浪微博上进行验证。其中，文中所用的新浪数据来源获取网址为：http://www.nlpir.org/?action-viewnews-itemid-299。

BA 网络分布不均匀，用 hub 型节点作为影响信息传播的枢纽节点，与真实网络中的领袖型传播节点极为相似，但是与小世界网络相比其聚类系数偏低。而小世界网络虽然有着高聚类系数，但是多数节点均匀分布，大度节点极少。新浪微博具有幂律分布特性，属于异质网络，超级传播者对信息传播影响比较明显，但是网络信息覆盖率比前两者低。假设文中网络为有向无权网络，文中出现的图均无单位，网络参数设置和不同网络条件下的度分布图分别如表 1 和图 4 所示。

表 1 网络参数设置

网络名	节点数	边数	平均度	度相关系数	聚类系数
BA	1 000	3 984	3	-0.066 96	0.037 11
WS	1 000	5 000	5	-0.014 72	0.261 79
新浪微博	3 982	6 803	8	-0.133 04	0.083 81

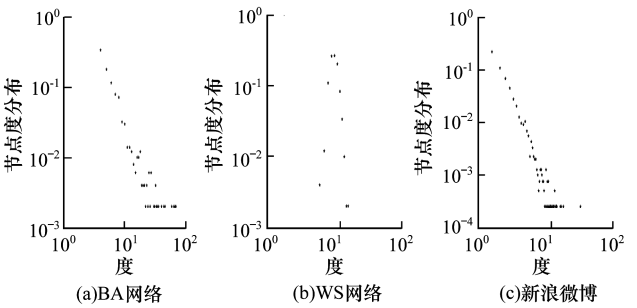


图 4 各网络的度分布

4.1 SIR 模型和 SCIR 模型的对比分析

实验选取一个节点为初始传播节点 I，其余节点全部为易感节点 S，设传播轮数为 15 次并且进行 200 次迭代实验。设初始传播速率、康复速率、节点转变速率均为 0.5，SIR 和 SCIR 模型各态节点在 BA 网络、WS 网络以及新浪微博中的密度变化分别如图 5～图 7 所示。其中，横轴 t 为传播轮次，纵轴为各节点密度分布情况。综合对比图 5～图 7 中各态节点密度变化情况，可以看出 SCIR 模型比 SIR 模型具有更好的网络适应性，新浪微博的稳定性低于 BA 网络和 WS 网络，网络信息覆盖率相对而言比较低，信息无法达到全网覆盖。

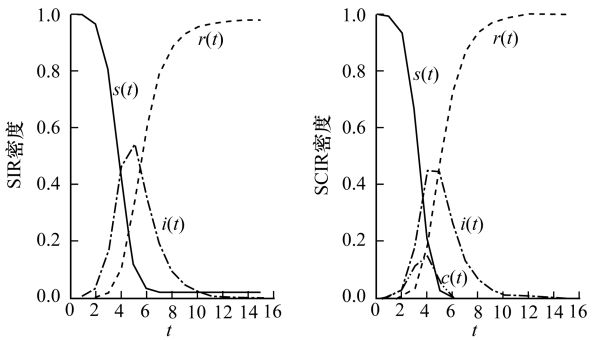


图 5　BA 网络中节点密度变化

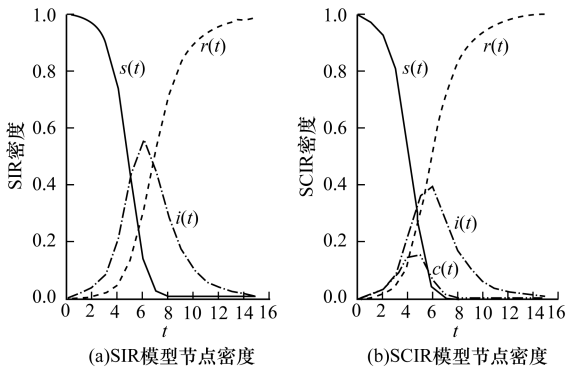


图 6　WS 网络中节点密度变化

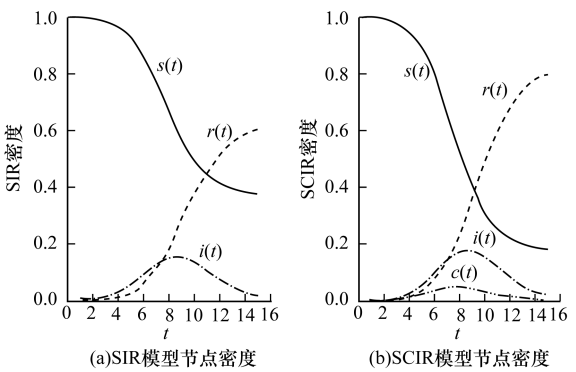


图 7　新浪微博中节点密度变化

在图 5 中,BA 网络中随着时间的推移,2 个模型$s(t)$节点都从 1 降为 0,而且在 t 为 2 到 6 时间段内急剧下降。这是由于几乎全是 S 节点的初始传播条件发生变化,I 节点和 C 节点数目不断增长所致。$c(t)$和 $i(t)$节点在增至峰值后也呈下降趋势,最终趋于 0。而 R 是唯一在一定时间后趋于 1 的节点。对比图 5(a)和图 5(b)可以看出,SCIR 模型中因为 C 节点的引入,$i(t)$的峰值明显下降,I 类节点处于爆发高潮时期的最大感染人口比出现近乎 0.1 的减幅。从图 5(b)可以看出,在时间为 4 时,S 节点和 C 节点的密度几乎同时抵达峰值,说明此时信息传播进入爆发高峰期。在 t 为 10 时,SCIR 模型中 $r(t)=1$,信息实现全网覆盖,但 SIR 模型中却仍然存在少数 S 节点,使得 $r(t)$在传播结束时刻仍无法增至 1。

从图 6 可以看出,在 WS 网络中,SIR 模型中 $i(t)$在 $t=6$ 时刻达到峰值 0.6,SCIR 模型到达峰值时间稍有延迟,且峰值降至 0.4,但是曲线波动趋于平缓。如图 6(b)所示,$c(t)$变化曲线与 $i(t)$变换曲线形状非常相似,且 $c(t)$的峰值出现时刻明显早于 $i(t)$。从图 5 得出的 SCIR 模型比 SIR 模型稳定性更好的结论在图 6 也得到验证。

新浪微博中 2 种模型的对比如图 7 所示。SCIR 模型中初态的 S 节点密度尾值较 SIR 模型中$s(t)$值从 0.4 下降至 0.2,终态的 R 节点密度值较 SIR 模型的 $r(t)$值从 0.6 上升至 0.8,但感染人口密度变化并不明显。上述情况表明信息传播结束时,新浪微博中仍存在部分感染节点 S,并未实现网络中所有节点都转变为免疫节点 R 的理想态,信息传播不能实现全网覆盖性传播。对比图 7(a)和图 7(b),图 7(b)的情况比图 7(a)有明显改进。

在不同网络中,SIR 模型与 SCIR 模型随着初始有效感染速率 λ 变化下的传播节点 I 的最大密度(最大感染比例)对比情况如图 8 所示。其中,横轴表示初始有效感染速率,纵轴表示最大感染比例。2 种模型在不同网络中最大感染比随着 λ 增大呈现增长趋势。但 SCIR 模型的整体最大感染比值相对 SIR 模型偏小,且增长趋势更加平缓。SIR 模型在 BA 和 WS 网络中,$\lambda<0.1$ 区间段类似线性增长,之后增长趋势相对平缓,且整体感染比高于 SCIR 模型。但在新浪微博中,2 个模型除了初始阶段的明显差异,后期增长趋势吻合度很高。

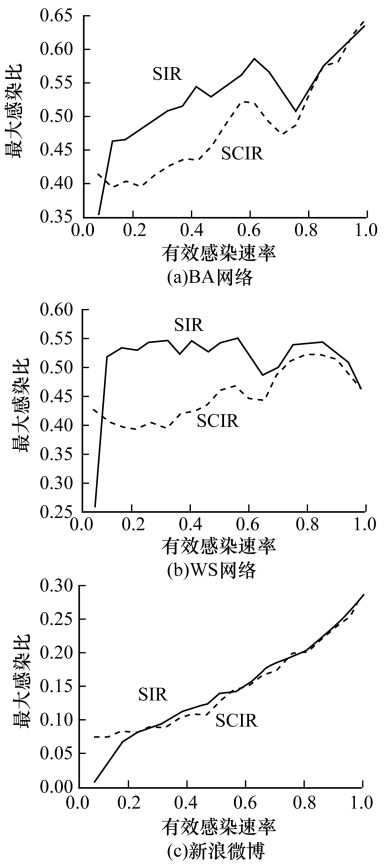


图 8　不同网络中最大感染比变化情况对比

4.2 新浪微博中康复速率 β 对 $r(t)$ 的影响

康复速率 β 为 0.2、0.4、0.6、0.8、1.0 时免疫节点密度 $r(t)$ 随 t 的变化趋势如图 9 所示，式(15)中的其他参数均给定值。其中，$\delta=0.2$，$\varepsilon=0.3$，$\mu=\alpha=p=0.5$。$r(t)$ 随着 β 值的增大而呈上升趋势，说明 β 值对于免疫节点密度的增大有正向促进作用，且 β 取值不同，曲线斜率变化也不同，说明 β 对于 $r(t)$ 的影响强度是不等的。

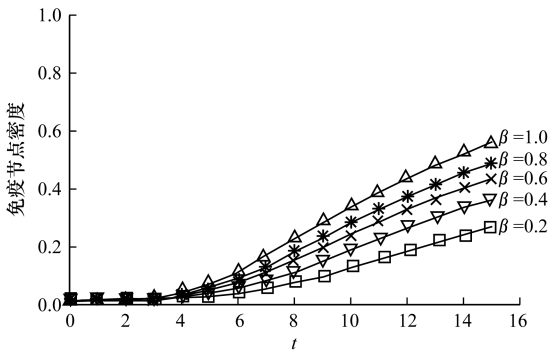
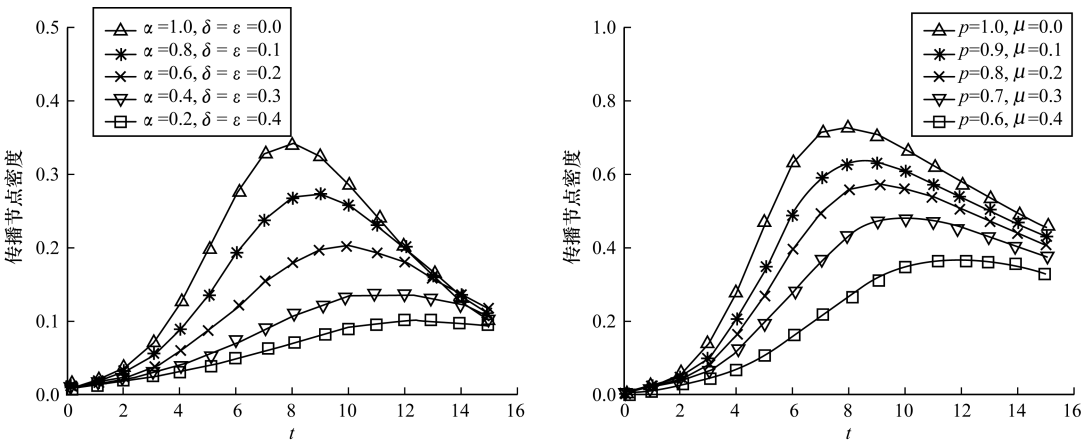


图 9 免疫节点密度变化趋势

4.3 新浪微博中外部感染速率 α 和内在传播速率 p

从图 10 可以看出，新浪微博中较之于外部感染速率，内在传播速率对传播节点密度 $i(t)$ 的影响较为显著。虽然 $i(t)$ 随着 α 的增大也出现了明显的增幅，但是 $i(t)$ 最大只增长到 0.3；而在 p 的影响下，$i(t)$ 最大增至 0.8，说明外部感染速率虽然对传播者密度有影响，但是不明显，而内在传播速率对传播者密度变化影响非常明显。从影响时长来说，内在传播速率 p 的影响时长也比外在感染速率长。因为在 t 接近传播终态时，α 影响下的终态 $i(t)$ 值在 0.1 附近，而 p 影响下的 $i(t)$ 值却在 0.4 附近，说明图 10(a)中 $i(t)$ 到达零点花费的时长要比图 10(b)中的少，后者 $i(t)$ 波动的时间更久，即内在传播速率 p 对 $i(t)$ 的影响周期大于外部感染速率 α。由于 α 对 $i(t)$ 的影响区间在[0.1,0.3]，纵轴最大值如果仍设为 1，参数曲线波动幅度仅占到差不多整幅图的 1/3，视图水平偏低，因此图 10(a)中纵轴最大值减半，设为 0.5。



(a)密度随 α 变化　　(b)密度随 p 变化

图 10 传播节点密度变化趋势

5 结束语

本文在传染病动力学 SIR 模型的基础上引入携带者节点 C 构成 SCIR 模型，将转发行为影响力取值作为感染节点感染速率的概率变量。实验结果表明：SCIR 模型中感染节点比例与 SIR 模型相比有明显下降；SCIR 模型中的免疫节点到达稳态时，信息覆盖率比 SIR 模型高；新浪微博中信息传播覆盖度相比其他 2 种网络偏低，信息传播受事件热度影响显著；初始有效感染速率对感染节点比例增长有正向促进作用，但最终会趋于平稳，且 SCIR 模型中感染节点增长比 SIR 模型更平缓；参数变化对节点比例提升有影响，会加快信息传播进程，但是不会改变传播阈值。

本文假设网络拓扑为无向图，所用网络均取自真实网络的一部分。文中假设的网络结构是理想化设想，但现实生活中社交网络的结构并非单一特定化而是具有复杂社会性，如何根据信息本身以及初始传播速率等因素研究复杂分层网络中的信息传播规律，将是下一步的研究方向。

参考文献

[1] 张 赛，徐 恪，李海涛. 微博类社交网络中信息传播的测量与分析[J]. 西安交通大学学报，2013，47(2)：124-130.

[2] 李 娟，周学广，陈 斌，等. 基于概率递减机制的信息传播模型构建与分析[J]. 海军工程大学学报，2016，28(1)：11-14，98.

[3] 李 娟，陈 斌. 基于 SI 模型的信息传播研究[J]. 信息系统工程，2015(8)：135-136，138.

[4] 王 超，杨旭颖，徐 珂，等. 基于 SEIR 的社交网络信息传播模型[J]. 电子学报，2014，42(11)：2325-2330.

(下转第 299 页)

Web pages[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2003: 601-606.

[3] ZHAI Y, LIU B. Web data extraction based on partial tree alignment [C]//Proceedings of the 14th International Conference on World Wide Web. New York, USA: ACM Press, 2005: 76-85.

[4] BUTTLER D, LIU L, PU C. A fully automated object extraction system for the World Wide Web [C]// Proceedings of the 21st International Conference on Distributed Computing Systems. Washington D. C., USA: IEEE Computer Society, 2001: 361-370.

[5] CHANG C, HSU C, LUI S, et al. Automatic information extraction from semi-structured Web pages by pattern discovery[J]. Decision Support Systems, 2003, 35(1): 129-147.

[6] LU Y, HE H, ZHAO H, et al. Annotating structured data of the deep Web[C]//Proceedings of the 23rd International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2007: 376-385.

[7] 扬少华,林海略,韩燕波. 针对模板生成网页的一种数据自动抽取方法[J]. 软件学报,2008,19(2):209-223.

[8] LIU W, YAN H, XIAO J. Automatically mining review records from forum Web sites[C]//Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery. Washington D. C., USA: IEEE Press, 2010: 2450-2455.

[9] 刘　伟,严华梁,肖建国,等. 一种 Web 评论自动抽取方法[J]. 软件学报,2010,21(12):3220-3236.

[10] ZHAO H, MENG W, WU Z, et al. Fully automatic wrapper generation for search engines[C]//Proceedings of the 14th International Conference on World Wide Web. New York, USA: ACM Press, 2005: 66-75.

[11] KAI S, LAUSEN G. ViPER: augmenting automatic information extraction with visual perceptions[C]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2005: 381-388.

[12] GUPTA S, KAISER G, NEISTADT D, et al. DOM-based content extraction of HTML documents[C]// Proceedings of the 12th international conference on World Wide Web. New York, USA: ACM Press, 2003: 207-214.

[13] 曹冬林,廖祥文,许洪波,等. 基于网页格式信息量的博客文章和评论抽取模型[J]. 软件学报,2009,20(5):1282-1291.

[14] MCCALLUM A, NIGAM K, RENNIE J, et al. Building domain-specific search engines with machine learning techniques [EB/OL]. [2017-08-01]. https://www.researchgate.net/publication/228738940_Building_Domain-Specific_Search_Engines_with_Machine_Learning_Techniques.

[15] BERGMARK D, LAGOZE C, SBITYAKOV A. Focused crawls, tunneling, and digital libraries[C]//Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries. Berlin, Germany: Springer, 2002: 91-106.

编辑　司森森

(上接第 288 页)

[5] 丁　鑫,刘其成,张　伟. 一种改进的微博网络信息传播预测模型[J]. 中国科学技术大学学报,2014,44(7):582-589,598.

[6] PU Cunlai, LI Siyuan, YANG Xiaxia, et al. Traffic-driven SIR epidemic spreading in networks[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 446: 129-137.

[7] ZHAO Laijun, WANG Jiaia, CHEN Yucheng, et al. SIHR rumor spreading model in social networks[J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(7): 2444-2453.

[8] 陈　波,唐相艳,于　泠,等. 基于胜任力模型的社交网络意见领袖识别方法[J]. 通信学报,2014,35(11): 12-22.

[9] MEHTA A, GUPTA R. An efficient local strategy to control infornation spreading in network[J]. Computer Science, 2015, 138(4): 273-275.

[10] WANG W, TANG M, YANG H, et al. Asymmetrically interacting spreading dynamics on complex layered networks[J]. Scientific Reports, 2014, 4(7502): 5097.

[11] GAO B, ZHAO D W. Competing spreading processes and immunization in multiplex networks[J]. Chaos Solitons & Fractals the Interdisciplinary Journal of Nonlinear Science & Nonequilibrium & Complex Phenomena, 2016, 93: 175-181.

[12] AGABA G O, KYRYCHKO Y N, BLYUSS K B. Mathematical model for the impact of awareness on the dynamics of infectious diseases [J]. Mathematical Biosciences, 2017, 286: 22.

[13] SAEEDIAN M, KHALIGHI M, AZIMI-TAFRESHI N, et al. Memory effects on epidemic evolution: the susceptible-infected-recovered epidemic model [J]. Physical Review E, 2017, 95(2-1): 022409.

[14] POZZANA I, SUN K, PERRA N. Epidemic spreading on activity-driven networks with attractiveness [J]. Physical Review E, 2017, 96(4-1): 042310.

[15] 易成岐. 社会网络的信息传播规律研究[D]. 哈尔滨: 哈尔滨工业大学,2013.

[16] 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.

[17] 文本相似度算法 [EB/OL]. [2017-09-01]. http://www.cnblogs.com/liangxiaxu/archive/2012/05/05/2484972.html.

[18] 郑志蕴,郭　芳,王振飞,等. 基于行为分析的微博传播模型研究[J]. 计算机科学,2016,43(12):41-45,70.

编辑　赵　辉