基于行为分析的微博传播模型研究

郑志蕴 郭 芳 王振飞 李 钝

(郑州大学信息工程学院 郑州 450001)

摘 要 随着微博的迅速兴起和其影响力的不断提高,提取微博信息传播特征和构建传播模型已成为了研究热点。针对用户转发行为,首先分析了信息传播机制;然后从影响用户转发行为的发布用户、接收用户、用户亲密度和信息时效性4个方面提取出8个特征因素进行建模;在借鉴传染病动力学SIR模型的基础上,引入用户行为分析和接触节点,提出基于用户行为分析的SCIR模型,并给出动力学方程;最后利用新浪微博真实转发数据验证模型的合理性。实验结果表明,考虑用户转发行为的8个影响因素,结合行为分析结果,能够较好地拟合信息传播过程。

关键词 微博,传播,SCIR模型,行为分析

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.007

Study on Microblog Propagation Model Based on Analysis of User Behavior

ZHENG Zhi-yun GUO Fang WANG Zhen-fei LI Dun

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract With the rapid rise of twitter and its influence continuing to improve, extracting the microblog information dissemination characteristics and building the propagation model have become a hot research topic. Forward for user behavior, firstly the information transmission mechanism was analyzed. Then accroding to eight factors extracted from publishing user, receiving user, user intimacy and information timeliness four aspects which affect the user behavior, the model was established. After that, the SCIR model was presented based on user behavior analysis and its dynamic equation was given. Finally the rationality of the model was validated by real forwarding data. Results show that forward considering user behavior influence factor, and combining the behavior analysis, can well fit information dissemination process.

Keywords Microblog, Propagation, SCIR model, Analysis of behavior

1 引言

近年来,随着社交网络的快速发展和普及,微博(Micro blog)作为一种全新的在线社交应用,其以接入便捷、内容简单、原创性、时效性和随意性等特点迅速聚集了大量用户,在全球范围内掀起一股热潮,并开始发挥越来越大的社会影响力。截止2014年第三季度,美国著名的微博客网站Twitter中活跃用户数量已达到2.84亿;到2015年3月,中国的新浪微博的用户总数量已超过6亿,月活跃用户增至1.67亿,是国内访问量最大的网站之一。2014年上半年的"马航事件"和2014年下半年的"冰桶挑战"凸显了新浪微博作为社交媒体所具有的快速的传播速度、深远的传播范围和积极的社会影响力。在2015年春晚直播期间,新浪微博推出"让红包飞"活动,互动人数高达3447万,话题量飙升至6133.4万条,春晚官方微博的粉丝增长率更是达到63.83%。

微博具有信息发布和用户交流两大功能,信息呈裂变式传播,其传播速度、广度和影响力远远超出传统媒体。分析微博网络中的用户行为、揭示微博传播规律以及构建信息传播模型具有非常重要的理论意义和应用价值。苑卫国等通过分析新浪微博的网络拓扑特征,发现微博网络具有小世界、无标度特性,属于BA无标度网络[1]。在行为分析方面,张旸等基于用户转发行为分别提取11个用户特征和11个文本特征,运用信息增益算法对各特征进行权重分析,构建了一种基于特征加权的预测模型[2];齐超等针对以前传播效果研究未考虑用户个体差异的问题,通过微博用户自身、用户关系和微博内容3方面提取9个相关特征,并结合逻辑回归(LR)方法构建基于行为分析的转发规模和传播深度预测模型[3];易兰丽以人类动力学理论为基础,综合运用复杂网络、概率论、管理学、统计学等多个学科的理论知识和方法,对用户的微博信息发布行为、转发行为和评论行为进行统计分析和理论建模[4];周东浩等结合网络结构特点、节点内容属性和历史传播数据等信息,提出基于随机游走模型的传播能力排序算法DiffRank来检测网络中可能的信息传播[5]。在构建传播模型方面,张彦超等考虑节点度和传播机理,综合复杂网络和传染病

到稿日期:2015-09-19 返修日期:2015-12-13 本文受郑州大学新媒体公共传播学科招标课题阶段性成果(XMTGGCBJSZ05),河南省科技攻关项目(142102310531)资助。

郑志蕴(1962—),女,博士,教授,CCF会员,主要研究方向为分布式计算、智能信息处理;郭 芳(1992—),女,硕士生,主要研究方向为智能信息处理和社交网络;王振飞(1973—),男,博士,副教授,CCF会员,主要研究方向为社交网络、信息安全,E-mail:iezfwang@zzu.edu.cn(通信作者);李 钝(1975—),女,博士,讲师,主要研究方向为社交网络、大数据。

动力学理论，建立信息传播SIR模型来分析节点行为规律[6]。在SIR模型的基础上，陈乾国等提出了一种无标度网络中带人工免疫的SIRS类传染病模型，并运用平均场理论分析该模型的动力学行为，对比了随机免疫和目标免疫两种人工免疫策略[7]；杨子龙等通过提取微博信息老化特征，结合转发时效性，基于平均转发概率的递减规律提出SIR的改进模型[8]；顾亦然等根据真实在线社交网络中的谣言传播特点和带有潜伏期的传染病模型，提出基于社交网络的谣言传播SEIR模型，给出重要熟人免疫策略以抑制谣言的传播[9]；王辉等在CSR模型基础上，从心理学上考虑个人接受阈值对接受概率的影响，改进了CSR模型的传播规则和动力学方程[10]。但是这些研究并没有同时考虑到微博用户群体的行为习惯差异性和用户间相似度两个因素对用户的转发行为的综合影响，不能实现差异性传播和精准预测。

本文在分析微博信息传播特征的基础上，从发布用户、接收用户、用户亲密度和信息时效性4个方面提取出8个特征对用户转发行为进行分析建模；然后在经典SIR模型的基础上引入接触节点，构建基于行为分析的微博传播SCIR模型；最后利用新浪微博真实数据对SCIR传播模型进行模拟，仿真结果显示该模型能够很好地刻画微博信息的传播过程。

2 信息传播机制分析

微博网络中，信息传播扩散主要依赖于用户的转发行为。一个用户发布微博后，粉丝会以一定概率看到，如果粉丝对该微博内容感兴趣，就会以一定概率转发，不感兴趣则不会转发微博。在此机制下，微博信息会沿粉丝关注关系进行传播。本文将微博网络中的注册用户定义为节点，用户之间的关注关系定义为连接节点的有向边，信息在微博网络中的传播过程如图1所示。

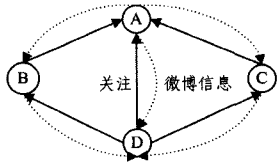


图1 微博信息传播网络图

图1中实线表示用户间的关注关系，虚线表示微博信息的传播路径，是关注关系的逆方向。用户B,C,D关注用户A，是A的粉丝，同时用户D也关注用户B和用户C，也是B和C的粉丝。当用户A发表微博后，其粉丝B,C,D同时看到微博信息，根据个人习惯、自身兴趣、微博内容及用户A对各自的不同影响力，决定是否转发这条微博。如果用户B或用户C转发该微博，则用户D会再次看到该微博信息。

研究表明，微博网络中用户的转发行为不仅与当前状态有关，用户自身行为习惯、发布用户影响力、用户之间相似度与亲密度和信息时效性等也是影响用户转发行为的关键因素。CHA等证明微博发布用户的影响力是影响用户转发行为的一个重要因素[11]。Crandall等[12]和Singla等[13]的研究表明节点间的相似度和节点间的影响力呈正相关关系，是影响微博信息传播的重要因素。同时，微博网络中的转发行为具有明显的时效性，毛佳昕等分析了新浪微博2009年9月到2012年11月的数据，统计发现50%的微博的转发延迟小于55min，90%的微博的转发延迟小于1153min(约19.2h)[14]。这表明微博网络中用户的转发行为受多种因素影响，对用户行为进行分析和建模是研究微博传播模型的重要基础。

3 构建基于行为分析的SCIR模型

传播模型对信息真实传播过程的刻画是否准确主要依赖于对网络特征提取和建模的准确度。本文利用从发布用户、接收用户、用户亲密度和信息时效性4个方面提取到的8个特征来优化SIR模型中对于用户转发概率的计算方法，同时引入一类新节点——接触节点，从而得到改进后的基于行为分析的SCIR模型。

3.1 特征提取

微博发布用户自身影响力是信息传播的主要因素，影响力作为用户的一个重要特征，是微博关系的基础。用户影响力越大，所受的关注度越高，微博被转发的概率就越大，对网络的影响作用也就越大。本文采用经典PageRank算法来计算微博发布用户的影响力，如式(1)所示。

$$R(u)=(1-d)+d\cdot\sum_{v\in Fans(u)}\frac{R(v)}{C(v)} \tag{1}$$

其中，u和v表示两个不同的用户；$R(u)$和$R(v)$分别表示u和v的影响力PR值；$Fans(u)$表示所有关注u的用户，即用户u的粉丝集合；$C(v)$表示用户v的关注数量；d是[0,1]之间的阻尼系数，表示用户u关注用户v的概率，$1-d$表示用户u关注一个随机用户的概率，引入阻尼系数保证算法具有收敛性，在实际应用中一般取值为0.85。这是个递归公式，一个用户的影响力取决于其粉丝的影响力，PR值是众多用户形成的一个分布概率，所有用户的PR值之和等于1。

用户接收微博信息后对其进行转发的原因有3类：有转发微博的习惯、对微博内容感兴趣以及受到发布用户或是邻居节点的影响。有些用户习惯看完微博后转发，有些只是浏览却很少转发。不同的用户有不同的兴趣爱好，彼此关注点差别很大，例如有些用户喜欢美食和旅游，有些用户则对八卦或是体育乐此不疲。个人喜好的差异也会极大地影响用户在微博网络中的行为。同时，由于微博是一个社交网络，网络中互相关注的用户可能是现实中熟悉的人。当周围人都关注一个话题时，无兴趣的用户会受到影响开始关注此话题。基于这些影响因素，本文选择转发活跃度、兴趣度和邻居节点感染率3个特征来刻画以上因素对用户网络行为的影响。

T时间段内，用户u转发微博数占其所有发布微博总数的比例称为转发活跃度$A(u)$，如式(2)所示。

$$A(u)=\frac{\sum_{t\in T}f_t}{\sum_{t\in T}r_t} \tag{2}$$

其中，f_t表示用户u在第t天转发的微博数，r_t表示用户在第t天发布的微博总数。

根据用户u的历史兴趣空间，判断其对新微博w的感兴趣程度，用$M(u,w)$表示兴趣度。本文采用基于潜在语义的分析方法对用户兴趣空间及兴趣度进行建模，步骤如下：

Step 1 数据采样。收集用户u最近一段时间内所发布的历史微博信息，构成文档集合$D_u=\{d_1,d_2,\cdots,d_i,\cdots,d_n\}$，其中$d_i$表示用户$u$的第$i$条微博。

Step 2 兴趣提取。采用LDA(Latent Dirichlet Allocation)话题模型，自动从文档集D_u中挖掘出每篇文档对应的话题组合，构成用户u词语级的兴趣分布向量W_u，向量长度

代表兴趣个数，各个分量表示用户对该话题的兴趣度。

Step 3　停用词剔除。通过与CSDN(China Software Developer Network)提供的停用词列表进行比对，剔除掉 W_u 中类似于"的"、"啊"、"你"之类的使用频率高但对兴趣空间的建立不起作用的停用词，构成用户 u 的最终兴趣空间 I_u。

Step 4　微博主题分析。对微博内容 w 采用 LDA 话题模型进行主题分析，剔除停用词，构成 w 的主题空间 I_w。

Step 5　兴趣度计算。通过用户 u 的近期兴趣空间 I_u 和微博 w 的主题空间 I_w，采用 Jaccard 相似系数计算这两个离散分布向量之间的距离，即用户 u 对于微博 w 的兴趣度 $M(u,w)$，如式(3)所示。

$$M(u,w)=\frac{I_u \bigcap I_w}{I_u \bigcup I_w} \tag{3}$$

用户 u 的关注用户中已对微博 w 进行转发的数量与 u 关注用户总数的比值表示邻居节点感染率 $N(u,w)$，如式(4)所示。

$$N(u,w)=\frac{F(w)}{C(u)} \tag{4}$$

在微博网络中用户之间建立的是一种单向连接，属于弱连接关系，可能粉丝对其关注者比较了解，而被关注者对其粉丝却一无所知。用户之间的相似度和关注关系体现了用户的亲密度，网络结构越相似，兴趣空间越相近，交互越频繁，说明用户之间的关系就越亲密，彼此的喜好和情绪越容易相互感染，转发行为也就更容易发生。本文选取网络结构相似度、兴趣相似度和交互频度 3 个特征来衡量两个用户间的亲密度。

用户 u 和用户 v 的网络拓扑结构的相似程度表示网络结构相似度 $T(u,v)$，计算方法如式(5)所示。

$$T(u,v)=\frac{\sum_{k\in C(u)\cap C(v)}\frac{1}{\log|C(k)|}}{|C(u)\bigcup C(v)|} \tag{5}$$

其中，$C(u)$ 和 $C(v)$ 分别表示用户 u 和用户 v 所关注的用户集合，$C(k)$ 表示 u 和 v 共同关注的用户集合。

构建用户 u 和用户 v 的历史兴趣空间 I_u 和 I_v，以两个用户兴趣空间的相似度作为其兴趣相似度 $M(u,v)$。

$$M(u,v)=\frac{I_u \bigcap I_v}{I_u \bigcup I_v} \tag{6}$$

一段时间内用户 u 和 v 之间交流的频率用交互频度 $S(u,v)$ 表示。用户的交互行为包括转发、提及和评论 3 种，如式(7)所示。

$$S(u,v)=\frac{1}{6}\left(\frac{r_{uv}}{r_u}+\frac{c_{uv}}{c_u}+\frac{m_{uv}}{m_u}+\frac{r_{vu}}{r_v}+\frac{c_{vu}}{c_v}+\frac{m_{vu}}{m_v}\right) \tag{7}$$

其中，r_{uv} 为用户 u 转发 v 的微博次数，c_{uv} 为评论次数，m_{uv} 为提及次数，r_u 为该时间段内用户 u 转发微博的总次数。$S(u,v)$ 是利用转发、评论和提及这 3 种行为占用户总行为的比值来衡量交互频度，比仅利用行为次数来判断亲密度更准确，更能体现一方对于另一方的重要性。

微博网络带有明显的时效性。文献[15]对微博的转发路径长度的统计分析表明，信息在距离初始节点较短的距离内具有较强的影响力，在网络的第二层、第三层和第四层传播速度最快，随着传播距离的增加，影响力迅速衰减。本文通过转发长度来衡量信息的时效性，如式(8)所示。

$$O(u,v,w)=\frac{l_{uv}+1}{l_{uv}^2} \tag{8}$$

其中，$O(u,v,w)$ 表示初始用户 v 发布的微博 w 经过转发后到达用户 u 时的时效性，l_{uv} 表示 u 到 v 的最短信息转发长度。如图 2 所示，初始发布者 A 的微博经 B 和 C 转发后到达用户 D 的路径包括 AD，ABD，ACD，$ABCD$ 4 条，对应的信息转发长度分别是 2，3，3，4，则 $l_{DA}=\min\{2,3,3,4\}=2$。

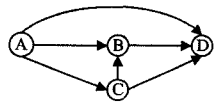


图 2　微博转发路径

3.2　转发概率计算

在已知微博发布用户 v 和目标微博内容 w 的条件下，对微博接收用户 u 采用式(9)计算微博的转发概率。

$$\begin{cases} P(u,v,w)=\dfrac{1}{1+e^{-\omega Y(u)}} \\ Y(u)=[R(v),A(u),M(u,w),N(u,w),T(u,v), \\ \qquad M(u,v),S(u,v),O(u,v,w)]^{-1} \\ \omega=[\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5,\lambda_6,\lambda_7,\lambda_8] \end{cases} \tag{9}$$

其中，$P(u,v,w)$ 表示用户 u 对用户 v 发布的微博 w 的转发概率；$Y(u)$ 表示影响用户 u 转发行为的特征向量，其元素是式(1)—式(8)的计算结果；ω 表示特征权重向量，分别对应特征向量 $Y(u)$ 中 8 个不同特征对用户转发行为的影响权重。权重的取值采用极大似然函数的方法计算。

3.3　基于行为分析的 SCIR 模型

在刻画社交网络中的信息传播过程时，经典 SIR 信息传播模型借鉴传染病动力学理论，将网络中的节点分为 S(未感染节点)、I(传播节点)和 R(免疫节点)3 类。未感染节点表示节点尚未接触到微博信息，但有被感染的概率；传播节点表示节点已看到微博信息并转发该微博；免疫节点表示节点已经转发过该微博并对其失去兴趣，不会再进行转发。对于特定的微博信息，网络中的节点在这 3 种类型间进行转换，一个未感染节点 u 收到该微博信息后，以一定概率进行转发，类型由 S 变为 I，网络中从 u 出发的边所连接的节点收到这条信息，并以一定的概率继续转发该微博，同时节点 u 的类型由 I 变为 R，之后与其他节点互不影响。

本文在 SIR 模型的基础上，通过用户行为分析来优化转发概率的计算方法，并引入一类新节点——接触节点，从而构建基于行为分析的 SCIR 微博传播模型。接触节点表示已经获知目标微博信息但还未对该微博进行转发时的用户。对于某条目标微博，网络中节点分为未感染节点 S(Susceptible)、接触节点 C(Contracted)、传播节点 I(Infected)和免疫节点 R(Removed) 4 种类型。基于微博信息传播模型，定义网络中节点的类型转换过程，如图 3 所示。

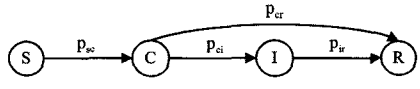


图 3　SCIR 模型节点类型转换图

如果一个未感染节点 S 与一个传播节点 I 接触，则 S 会以概率 1 转换为接触节点 C，其中未感染节点与传播节点的接触概率为 p_{sc}，p_{sc} 称为信息接触概率。接触节点 C 以概率 p_{ci} 转换为传播节点 I，p_{ci} 称为接触节点 C 对目标微博的转发概率。接触节点 C 以概率 p_{cr} 转换为免疫节点 R，p_{cr} 称为接触节点 C 对目标微博的直接免疫概率。传播节点 I 以概率 p_{ir} 转换为免疫节点 R，p_{ir} 称为传播节点 I 对目标微博的间接免

疫概率。免疫状态为网络的吸收态，凡是进入免疫状态的节点，其状态在微博传播过程中不再发生变化。

假定目标微博的传播过程中网络中用户数量和彼此之间的关注关系不会发生变化，即节点总数和网络结构不变。设网络中总用户数量为 N，t 时刻网络中 4 类节点的数量分别是未感染节点 $S(t)$、接触节点 $C(t)$、传播节点 $I(t)$以及免疫节点 $R(t)$，则 $S(t)+C(t)+I(t)+R(t)=N$。由于一条目标微博通常是由一个用户发出，因此 $I(0)=1$，$C(0)=R(0)=0$。根据图 3 中的节点类型转换过程，可将 SCIR 模型用微分方程组描述如下。

$$\begin{cases}\dfrac{\mathrm{d}S(t)}{\mathrm{d}t}=-p_{sc}I(t)K(t)\\ \dfrac{\mathrm{d}C(t)}{\mathrm{d}t}=p_{sc}I(t)K(t)-p_{ci}C(t)-p_{cr}C(t)\\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t}=p_{ci}C(t)-p_{ir}I(t)\\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t}=p_{cr}C(t)+p_{ir}I(t)\end{cases}\tag{10}$$

其中，第一个式子表示网络中未感染节点数量的变化率；第二个式子表示接触节点数量的变化率；第三个式子表示传播节点数量的变化率；第四个式子表示免疫节点数量的变化率。

(1)信息接触概率 p_{sc} 由网络拓扑结构和用户群体确定。

(2)$t=1$ 时，$K(t)$ 表示网络中节点的平均出度；$t>1$ 时，$K(t)$ 表示网络平均额外出度[16]。

(3)转发概率 p_{ci}：对发布用户 v、接收用户 u 和目标微博 w，根据式(9)计算出用户 u 的转发概率 $P(u,v,w)$，则令 $p_{ci}=P(u,v,w)$，实现信息的差异化传播。

(4)本文假设用户转发过目标微博后就会对其失去兴趣不再转发，所以间接免疫概率 $p_{ir}=1$。

4 实验与分析

根据 SCIR 模型，以微博转发网络为信息传播载体，模拟信息在网络中的传播过程，验证 SCIR 模型的效果和性能。

4.1 实验数据

基于新浪微博开放的应用程序接口(Application Programming Interference，API)，采用广度优先策略，抓取新浪微博用户消息内容。

(1)数据采集：首先将选定时间段内的微博信息作为原始数据，然后对某一条用户 i 转发自用户 j 的微博，分别提取 i 和 j 的 id，检查现有网络中是否存在表示用户 i 和 j 的节点以及表示 i 和 j 间转发关系的边，若不存在，则将其添加到网络中。本文采集 2015 年 1 月 1 日－2015 年 6 月 30 日时间段内热门微博的用户转发信息，得到数据原始信息。

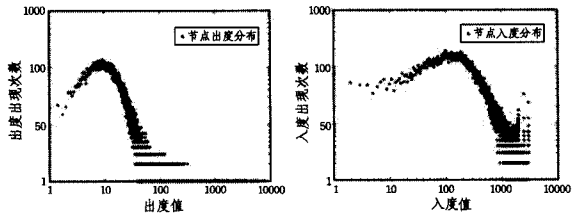
(2)数据预处理：该过程是用户行为分析和传播预测必不可少的环节。原始数据中存在部分活跃度、参与度极低的用户以及一些可能的“僵尸粉”、“水军”等，需要对数据进行清洗，删除无用数据，去掉重复信息，过滤干扰因素，以便提高对模型性能分析和评估的准确率。

首先基于 SVM 分类器对原始数据进行垃圾用户检测[17]，然后去除粉丝数低于阈值 10、日均发布信息数量低于阈值 0.1 的用户，最终得到了一个由 63641 个节点和 2775981 条边构成的转发网络，网络模型参数如表 1 所列。

表 1 网络模型参数

参数	结果	参数	结果
节点数	63641	边数	2775981
好友关系总数	1391718	度相关系数	0.033495
出度平均值	1099.75	出度最大值	2813741
入度平均值	397.56	入度最大值	90242
聚类系数	0.20319	平均路径	非连通图

网络的出、入度分布如图 4 所示，近似服从幂率分布。



(a)节点出度分布图　(b)节点入度分布图

图 4 微博有向转发网络出入度分布

(3)用户行为分析：将用户 u 在该段时间内所发布的信息汇总为文档 D_u，所有用户文档组成文档集 D，对微博转发网络进行特征提取，用于计算接收用户对于目标微博的转发概率。

4.2 实验方法和参数学习

将 6 个月的数据划分为相互独立的 3 组分别进行训练和测试。第 1 组选取第 1 个月的数据作为训练数据，用于建立 SCIR 模型。将第 4 个月的数据进行测试，用以分析模型的性能。以此类推，每一组的训练数据和测试数据时间间隔为 3 个月，最后取各组结果的平均值。SCIR 模型中需要学习的参数是式(9)中的特征权重向量 ω，选择命中率作为目标函数，使得命中率最高的特征权重向量为 ω 的最优取值。

4.3 传播仿真

基于 SCIR 模型，对网络中的信息传播进行模拟，实验次数为 2000，每次随机选择名人堂中的一个用户作为初始节点，通过转发进行信息传播，观察网络中不同类型的节点数量的变化情况。图 5 为采用 SCIR 模型模拟目标微博在转发网络中传播时，网络中未感染节点、接触节点、传播节点和免疫节点的密度随时间的演化情况。$S(t)/N$ 在信息传播初期呈迅速递减的趋势，因为网络中一旦某个节点转换为传播节点，所有粉丝节点以概率 $p_{sc}=1$ 转换为接触节点，表明微博网络中信息的裂变式传播模式。$C(t)/N$ 在话题传播初期呈迅速增加的趋势，并在短时间内达到峰值，之后由于网络中未感染节点 S 数量的减少以及 C 类型节点向 I 类型、R 类型的转变，其密度随时间递减，并最终在网络中消失；$I(t)/N$ 在信息传播初期呈增加趋势，达到峰值后逐渐递减，最终在网络中消失；$R(t)/N$ 在传播初期随时间递增，之后逐渐稳定，最终趋近于 1，即免疫状态是网络中的吸收状态。

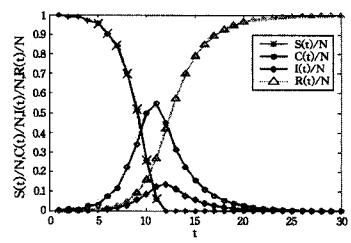
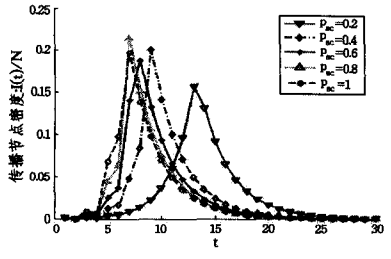


图 5 不同类型节点密度随时间变化情况

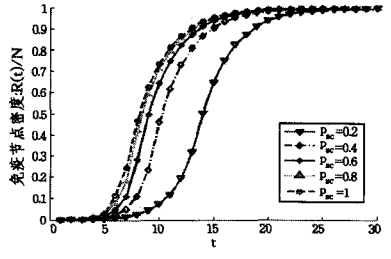
图 6 表示在初始条件保持不变的情况下，外部影响概率 p_{sc} 取不同值时，传播节点密度 $I(t)$ 及免疫节点密度 $R(t)$ 随时间的变化关系曲线。

(1)在网络达到稳态之前，$I(t)$随 p_{sc}取值的增大而增加，因为 p_{sc}表示未感染节点通过各种途径与感染节点的接触概率，p_{sc}的取值越大，未感染节点的转发行为受外部环境的影响就越大，图 6 的仿真结果与实际舆情话题的传播规律相符。

(2)免疫节点的最终密度随着 p_{sc}取值的增大而增加，因为 p_{sc}取值的增大，会使网络中更多的未感染节点转变为传播节点，这些传播节点最终会因为失去对话题的兴趣而成为免疫节点。



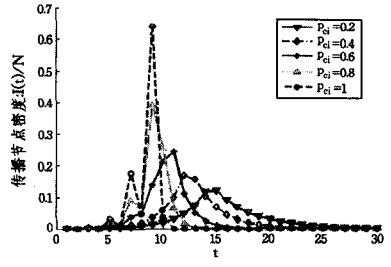
(a)传播节点密度 $I(t)/N$ 随时间变化的情况



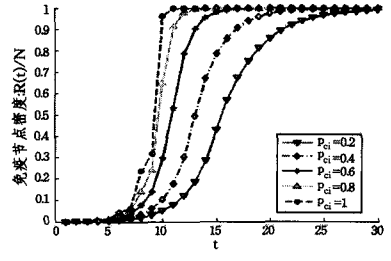
(b)免疫节点密度 $R(t)/N$ 随时间变化的情况

图 6　不同 p_{sc}值时传播节点和免疫节点密度随时间变化图

图 7 分别给出了在其他初始条件不变的情况下，转发概率 p_{ci}取不同值时，传播节点密度 $I(t)$及免疫节点密度 $R(t)$随时间的变化曲线。



(a)传播节点密度 $I(t)/N$ 随时间变化的情况



(b)免疫节点密度 $R(t)/N$ 随时间变化的情况

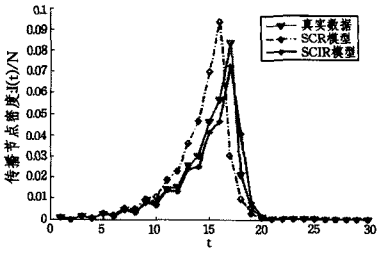
图 7　不同 p_{ci}值时传播节点和免疫节点密度随时间的变化图

(1)在网络达到稳态之前，$I(t)$随 p_{ci}取值的增大而增大，而 $R(t)$随 p_{ci}取值的增大而减小。因为转发概率 p_{ci}的增大表明处于接触状态的节点转发话题的可能性增加，传播的次级联效应由此增强，转发节点密度 $I(t)$也随之增大。

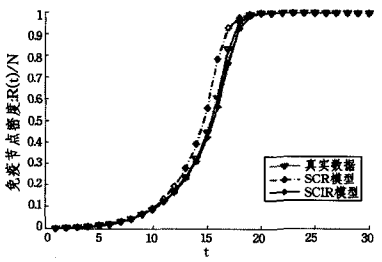
(2)p_{ci}取值的改变不会影响网络中各类节点的最终密度，即 $I(t)$将最终趋近于 0，$R(t)$最终趋近于 1。

(3)网络达到稳定状态的时间随着 p_{ci}的增加而延长，由于 p_{ci}的增加使得网络中传播节点数目随之增加，从而导致需要更长的演化时间才可以使传播过程达到稳定状态。

本实验构建的微博传播网络只是现实网络的简化版本，为了进行合理比较，本文对结果进行归一化处理。图 8 分别给出 SCIR 模型关于转发节点密度和免疫节点密度的仿真结果与真实数据的对比。由图 8 可知，仿真算法可以对真实数据进行拟合，说明仿真算法能够反映真实信息转发规律。



(a)传播节点



(b)免疫节点

图 8　SCIR 模型模拟结果与真实数据对比

结束语　信息转发是微博中信息传播的重要方式。本文从实际数据出发，提取了微博信息转发过程中发布用户、接收用户、用户亲密度和信息时效性 4 个方面的 8 个特征进行用户行为分析和建模，利用分析结果优化了 SIR 模型中对于用户转发概率的计算方法，并引入了一种新的节点类型——接触节点，来刻画已经接触到目标微博但尚未做出行动的这类用户状态，提出了基于行为分析的微博传播 SCIR 模型。为了验证模型仿真效果，本文收集了新浪名人堂 2015 年 1 月 1 日－2015 年 6 月 30 日期间的热门微博的用户信息和转发数据，构建有向转发网络，采用 SCIR 模型对其中的信息传播进行模拟。仿真结果呈现出与实际数据相符合的传播效果，证明提出的 SCIR 模型在一定程度上体现了微博网络中的信息转发规律，具有一定参考价值。

参 考 文 献

[1] Yuan Wei-guo, Liu Yun, Cheng Jun-jun, et al. Empirical analysis of microblog centrality and spread influence based on Bi-directional connection[J]. Acta Physica Sinica, 2013, 62(3): 502-511 (in Chinese)
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)：502-511

[2] Zhang Yang, Lu Rong, Yang Qing. Predicting Retweeting in Microblogs[J]. Journal of Chinese Information Processing, 2012, 26(4): 109-114(in Chinese)
张旸，路荣，杨青. 微博客中转发行为的预测研究[J]. 中文信息学报，2012，26(4)：109-114

(下转第 70 页)

[12] Yao Y Y, Zhao Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356-3373

[13] Chen H, Yang J A, Zhuang Z Q. The core of attributes and minimal attributes reduction in variable precision rough set[J]. Chinese Journal of Computers, 2012, 35(5): 1011-1017

[14] Jia X Y, Liao W H, Tang Z M, et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2013, 219(10): 151-167

[15] Jia X Y, Tang Z M, Liao W H, et al. On an optimization representation of decision-theoretic rough set model[J]. International Journal of Approximate Reasoning, 2014, 55(1): 156-166

[16] Wang G Y, Ma X A, Yu H. Monotonic uncertainty measures for attribute reduction inprobabilistic rough set model[J]. International Journal of Approximate Reasoning, 2015, 59: 41-67

[17] Chan C C. A rough set approach to attribute generalization in data mining[J]. Journal of Information Sciences, 1998, 107: 169-176

[18] Li T R. A rough sets based characteristic relation approach for dynamic attribute generalization in data mining[J]. Knowledge-Based Systems, 2007, 20: 485-494

[19] Chen H M, Li T R, Qiao S J, et al. A rough set based dynamic maintenance approach for approximations in coarsening and refining attribute values[J]. International Journal of Intelligent Systems, 2010, 25(10): 1005-1026

[20] Liu D, Li T R, Ruan D, et al. An incremental approach for inducing knowledge from dynamic information systems[J]. Fundamenta Informaticae, 2009, 94(2): 245-260

[21] Luo C, Li T R, Zhang J B. Dynamic maintenance of approximations in set-valued ordered decision systems under the attribute generalization[J]. Information Sciences, 2014, 257(2): 210-228

[22] Zhang J B, Li T R, Chen H M. Composite rough sets for dynamic data mining[J]. Information Sciences, 2014, 257: 81-100

[23] Yao Y Y. Two semantic issues in a probabilistic rough set model [J]. Fundamenta Informaticae, 2011, 108(3): 249-265

(上接第 45 页)

[3] Qi Chao, Chen Hong-chang, Yu Yan. Micro-blog information diffusion effect based on behavior analysis[J]. Journal of Computer Applications, 2014, 34(8): 2404-2408(in Chinese)
齐超,陈鸿昶,于岩. 基于行为分析的微博信息传播效果[J]. 计算机应用,2014,34(8):2404-2408

[4] Yi Lan-li. Research on Statistical Characteristic Analysis and Modeling for Behavior in Microblog Community Based on Human Dynamics[D]. Beijing: Beijing University of Posts and Telecommunications, 2012(in Chinese)
易兰丽. 基于人类动力学的微博用户行为统计特征分析与建模研究[D]. 北京:北京邮电大学,2012

[5] Zhou Dong-hao, Han Wen-bao. DiffRank: A Novel Algorithm for Information Diffusion Detection in Social Networks[J]. Chinese Journal of Computers, 2014, 37(4): 884-893(in Chinese)
周东浩,韩文报. DiffRank:一种新型社会网络信息传播检测算法[J]. 计算机学报,2014,37(4):884-893

[6] Zhang Yan-chao, Liu Yun, Zhang Hai-feng, et al. The Research of Information Dissemination Model on Online Social Network [J]. Acta Physica Sinica, 2011, 60(5): 66-72(in Chinese)
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):66-72

[7] Chen Qian-guo, Zhang Zi-li. Dynamics Behavior and Immune Control Strategies of SIRS Model with Immunization on Scale-free Complex Networks[J]. Computer Science, 2013, 40(6): 211-214(in Chinese)
陈乾国,张自力. 无标度网络上带人工免疫的 SIRS 模型动力学行为及其免疫控制策略[J]. 计算机科学,2013,40(6):211-214

[8] Yang Zi-long, Huang Shu-guang, Wang Zhen, et al. Study on Micro Blog Reposting Model Based on Characteristics of Information Obsolescence[J]. Computer Science, 2014, 41(12): 82-85 (in Chinese)
杨子龙,黄曙光,王珍,等. 基于信息老化特征的微博传播模型[J]. 计算机科学,2014,41(12):82-85

[9] Gu Yi-ran, Xia Ling-ling. The Propagation and Inhibition of Rumors in Online Social Network[J]. Acta Physica Sinica, 2012, 61(23): 238701(in Chinese)
顾亦然,夏玲玲. 在线社交网络中谣言的传播与抑制[J]. 物理学报,2012,61(23):238701

[10] Wang Hui, Han Jiang-hong, Deng Lin, et al. Dynamics of Rumor Spreading in Mobile Social Networks[J]. Acta Physica Sinica, 2013, 62(11): 110505(in Chinese)
王辉,韩江洪,邓林,等. 基于移动社交网络的谣言传播动力学研究[J]. 物理学报,2013,62(11):110505

[11] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in twitter: the million follower fallacy [C]// Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media(ICWSM 2012). Menlo Park: AAAI Press, 2010: 10-17

[12] Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities[C]// Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, ACM, 2008: 160-168

[13] Singla P, Richardson M. Yes, there is a correlation: From social networks to personal behavior on the Web[C]// Proceeding of the 17th International World Wide Web Conference. Beijing, China, 2008: 665-664

[14] Mao Jia-xin, Liu Yi-qun, Zhang Min, et al. Social Influence Analysis for Micro-Blog User Based on User Behavior[J]. Chinese Journal of Computers, 2014, 37(4): 791-795(in Chinese)
毛佳昕,刘奕群,张敏,等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报,2014,37(4):791-795

[15] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media? [C]// Proceddings of the 19th international conference on World Wide Web. ACM, 2010: 591-600

[16] Newman M. Network: an introduction [M]. New York: Oxford University Press, 2009: 449

[17] Li He-yuan, Yu Xiao-ming, Liu Yue, et al. Research on Detecting Spammer in Micro-blogs[J]. Journal of Chinese Information Processing, 2014, 28(3): 62-67(in Chinese)
李赫元,俞晓明,刘悦,等. 中文微博客的垃圾用户检测[J]. 中文信息学报,2014,28(3):62-67