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

基于用户行为的微博网络信息扩散模型

刘红丽　黄雅丽　罗春海　胡海波

Modeling information diffusion on microblog networks based on users' behaviors

Liu Hong-Li　Huang Ya-Li　Luo Chun-Hai　Hu Hai-Bo

引用信息 Citation: Acta Physica Sinica, 65, 158901 (2016)　DOI: 10.7498/aps.65.158901
在线阅读 View online: http://dx.doi.org/10.7498/aps.65.158901
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2016/V65/I15

基于用户行为的微博网络信息扩散模型*

刘红丽 黄雅丽 罗春海 胡海波†

(华东理工大学管理科学与工程系, 上海 200237)

(2016 年 3 月 13 日收到; 2016 年 5 月 3 日收到修改稿)

利用新浪微博数据对用户行为进行分析, 在此基础上构建了基于用户行为的微博网络信息扩散模型 SIRUB, 同时计算了模型中各用户阅读微博和转发微博的概率. 在微博网络中的实验表明, 只有同时考虑阅读和转发概率时模型才能较准确地预测用户的转发行为. SIRUB 模型对用户转发行为预测的 F-score 最高为 0.228, 高于经典 SIR 模型和 SICR 模型, 此外该模型对微博扩散范围的预测其误差的均值和标准差也均小于 SIR 模型和 SICR 模型.

关键词: 微博网络, 用户行为, 信息扩散

PACS: 89.65.-s, 87.23.Ge **DOI:** 10.7498/aps.65.158901

1 引 言

近年来, 微博平台成为人们获取和传播信息的重要途径, 因而对微博网络中信息扩散的研究具有重大的社会经济意义 [1]. 微博网络中某些微博转发量可达到几十万甚至上百万, 信息受众更是数以千万计, 而另一些微博从一开始就被人们忽略, 造成这种扩散差异的原因是多方面的, 就此学者们对影响微博网络信息扩散的因素和扩散机制进行了大量的研究.

影响信息扩散的因素可以分为用户特征、社交关系特征和微博文本特征三个方面. 在用户特征影响信息扩散的显著性方面学者们存在不同观点, 有些学者认为转发关系中上游用户的粉丝数、朋友数会影响下游用户的转发行为 [2,3], 而另有研究则表明粉丝数对下游用户的转发行为影响不显著 [4,5]. 在社交关系特征上学者们则一致认为社交关系强度和信息扩散范围呈正相关关系, 如用户间主题兴趣相似度越高 [6,7]、交互强度越大 [5,8], 信息越容易在两个用户之间扩散. 微博文本特征也会影响信息扩散, 如微博和用户兴趣相似度正向影响用户转发行为 [5,9], 不同的信息不仅在用户和用户之间扩散的概率不同, 不同信息重复暴露对其被采用的边际贡献率也不同 [10], 学者们对不同的微博文本特征对信息扩散的影响显著性也存在不同的观点.

根据社交网络上信息扩散的机制和过程, 研究者们提出了各种理论模型 [11−14], 如基于传染病传播的模型、线性阈值模型、独立级联模型等, 这些模型为信息扩散研究提供了理论基础. Xiong 等 [15] 认为在微博环境下, 节点转发某条微博后, 微博将保留在他的主页上, 易感节点和传播节点接触要么转变成传播状态 (I), 要么转变成接触状态 (C), 而且只有接触状态的节点才能转变成免疫状态 (R), 模型达到稳态后网络中将包含 I 和 R 状态的节点; Prakash 等 [16] 提出了 Susceptible-Infected$_1$-Infected$_2$-Susceptible(SI_1I_2S) 模型, 研究了网络中两种竞争信息的扩散的结果, 研究发现处于优势的信息最终会“赢者通吃” (winner-takes-all), 将处于劣势的信息排挤出网络; Liu 等 [17] 将用户兴趣和信息内容结合起来提出了一种基于信息亲和机制的 Susceptible-Known-Informed-Refractory (SKIR) 扩散模型, 研究表明信息亲和阈值影响了信息的最终扩散范围. 目前绝大多数的信息扩散理

* 国家自然科学基金 (批准号:61473119, 61104139) 和中央高校基本科研业务费专项资金 (批准号:WN1524301) 资助的课题

† 通信作者. E-mail: hbhu@ecust.edu.cn

论模型是建立在传染病传播模型基础上的[11−13].

虽然在理论模型研究上取得了很大的进展, 但模型的参数设置往往缺少真实数据的支撑, 因而实验结果的可靠性受到了质疑, 如Goel等用结构化可传染性(structural virality)指标刻画了介于广播式扩散和传染病式扩散之间的扩散情况, 传染病模型仿真结果显示, 模型无法再现真实情况下结构化可传染性的多样性[18], 因此目前学者们在研究微博网络信息扩散时更偏向于数据驱动的模型. 如Liu等[7]根据用户主题兴趣和间接影响力, 提高了预测Twitter用户转发行为的准确度; Goyal等[19]在常规阈值模型基础上, 建立了静态模型、连续时间和离散时间模型, 并根据用户活动的先后顺序计算了基于影响力的转发概率. 有些学者根据用户的历史数据, 利用各种计算方法来预测未来的信息扩散[5,9,20], 也有学者研究了用户的在线阅读行为, 并基于用户粉丝阅读行为的分析计算了用户的影响力[21]. 此外, 分支过程也已广泛应用到数据驱动的信息扩散模型中[22−24].

信息在微博网络中扩散, 网络节点传播信息的前提是节点接收到了信息, 早期的信息扩散理论模型大多假设信息在一个封闭、同质的人群中扩散, 传播者和他的邻居会无差异地接触, 信息会被他的邻居无差异地传播. 但是在微博环境下, 信息量大、信息更新速度快, 用户的粉丝之间差异巨大, 无法保证用户所发的每条微博会被其每个粉丝阅读. 因此同转发行为一样, 用户的阅读行为也会影响微博网络中信息的扩散, 但是目前对用户阅读行为尤其是同时考虑阅读行为和转发行为对信息扩散的影响的研究仍不够深入. 本研究利用新浪微博数据, 首先分析用户的阅读行为和转发行为, 在此基础上构建基于用户阅读概率和转发概率的微博网络信息扩散预测模型, 并与其他模型的预测效果进行比较.

2 用户行为分析

2.1 数据描述

本研究利用新浪微博提供的API接口(http://open.weibo.com/), 从一个粉丝数和微博数较多的用户开始, 先将该用户加入爬取队列, 根据研究需要爬取该用户最新发布的100条微博, 对其中的每条微博, 再爬取该微博的原创微博(如果是转发微博)和转发微博以及原创微博和转发微博的用户信息, 并将这些用户加入爬取队列. 一个用户处理完后, 再提取爬取队列中的下一个用户进行相同处理, 重复上述操作.

从2014年10月15日开始到10月20日共收集了21992位用户的信息和这些用户发布的2076564条微博的信息, 之后收集了这些用户在2014年10月15日至2015年2月1日期间发表的9534792条微博, 最后收集这些用户的转发关系, 排除陌生人(即非本用户粉丝)转发, 共得到589626条关注关系. 本研究收集用户2014年10月15日前最新发表的100条微博和在2014年10月15日至2015年2月1日期间发表的微博, 用户在这些微博间若有转发关系, 则收集他们之间的关注关系, 由此可得到一个转发关系网络.

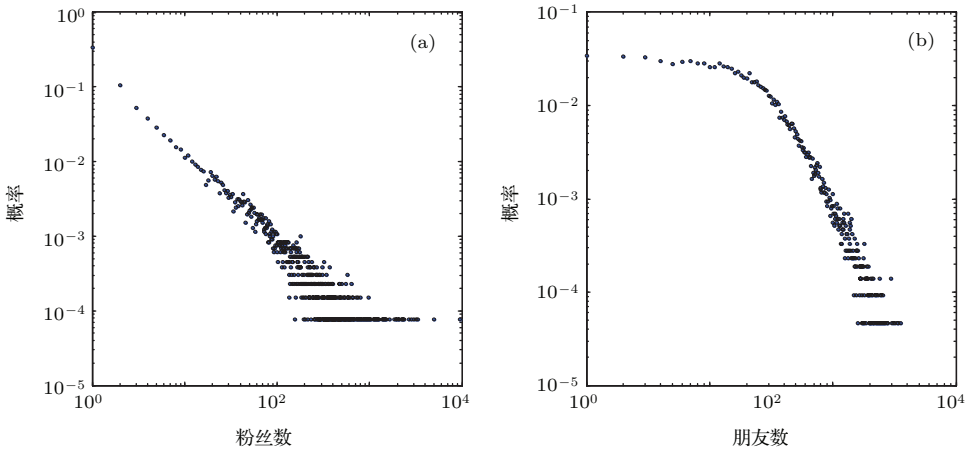

图1 有转发关系的用户的粉丝数(a)和朋友数(b)分布

Fig. 1. The distributions of the numbers of fans (a) and friends (b) of users among which there exists reposting relationship.

图1给出了该网络的出度和入度分布, 出度代表用户的粉丝数, 入度代表用户的朋友(用户关注的人)数. 图1(a)表明一段时间内与用户有转发关系的粉丝的数量近似于幂律分布. 虽然已有研究认为用户的粉丝数、朋友数均服从幂律分布, 但图1(b)表明只有当朋友数大于某一特定值(20)时才表现出幂律行为, 微博网络中超过一半的用户(53.9%)和少于20位的朋友有转发关系.

2.2 用户阅读行为

微博网络上用户的朋友发表的微博会按时间顺序显示在用户页面上, 用户登陆后按顺序翻阅微博. 阅读过程中用户如果觉得某条微博有趣、值得跟粉丝分享, 就会转发该微博. 微博发表的时间越长越被排在用户页面的后面, 用户一次登录一般不会翻阅完所有的微博, 以至于有些微博会被用户忽略, 这些微博即使用户感兴趣也不会被转发, 由此可见用户转发某条微博的前提是他必须阅读到该微博. 虽然用户的阅读行为不会被记录下来, 但是如果知道一位用户登陆微博的时间和用户登录后的信息阅读量, 就可以根据一条微博的发表时间来判断该用户会不会阅读到这条微博.

我们定义转发延迟为微博被创建和被转发之间的时间间隔, 根据用户在一天内各时间段登录微博的频率和用户的转发延迟时间, 可推测用户阅读到某条微博的概率. 为此, 本小节首先分析用户所发表的微博在一天内各时间段的分布, 之后推测用户登录微博的频率分布, 并分析用户转发延迟时间分布.

2.2.1 用户登录微博的行为

假设用户每次登录微博发文数量和登陆时间是不相关的, 即用户发文数量的时间分布只和各时间段用户登录微博的频率有关, 那么该时间分布可用于推测用户登录微博的概率在一天内的分布. 对微博发表时间进行分析, 可得原创微博与转发微博在一天内各时间段的分布, 如图2所示. 可见晚上11点至次日早上6点用户发文数量剧减, 早上6点以后开始增加, 早上8点到晚上8点发文数量分布较为均匀, 晚上9至10点处于一天的最高峰, 这说明用户一天内各时间段登录微博的频率是不同的.

2.2.2 转发延迟分析

用户阅读和转发一条微博的行为几乎是同时发生的, 所以可用转发微博的时间点表示阅读时间点. 转发延迟时间分布如图3所示, 该分布近似于幂律分布, 说明大部分转发延迟较小. 延迟小于8.77小时的占75%, 小于35.38小时的占90%, 说明微博消息具有很强的时效性, 发表时间越长的消息越少人去关注.

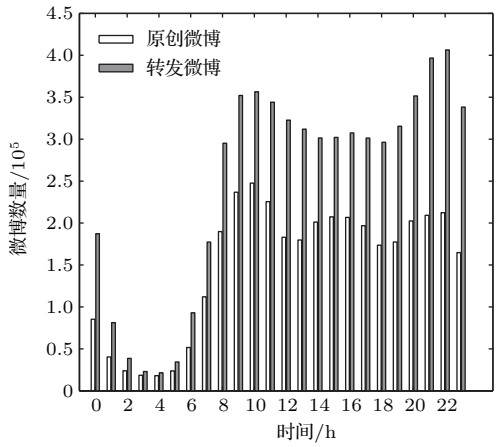

图2 原创微博与转发微博在一天内各时间段的分布

Fig. 2. The distributions of original and reposted microblog in a day.

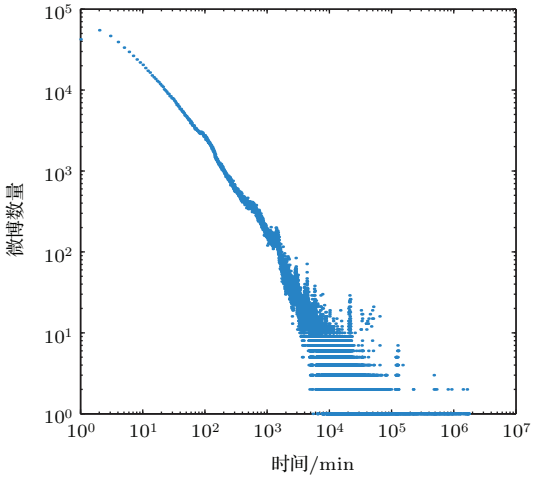

图3 微博转发延迟分布

Fig. 3. The delay distribution of microblog reposting.

2.3 用户转发行为的影响因素

本节对影响用户转发行为的因素展开分析, 其中包含上游用户的粉丝数、上游用户的微博平均转发数、用户交互强度、微博主题和用户主题兴趣相似度四个因素. 用户交互强度是指历史上用户相互转发对方微博的次数. 本研究对收集的微博进行清理、分词后, 用Twitter-Latent Dirichlet Allocation (Twitter-LDA)主题分析模型[25]对微博文本进行主题分析, 得到每个用户主题兴趣分布DT矩阵、

每条微博所属主题和每个主题的词汇分布. DT 为 $D \times T$ 矩阵, D 表示用户数量, T 表示主题数量, $DT(i,j)$ 以概率的形式表示用户 i 对主题 j 的感兴趣程度, 其值越大表明 i 对 j 越感兴趣. 从而我们可以得到微博所属主题在用户主题兴趣分布 DT 矩阵中的值, 并用以衡量微博主题和用户主题兴趣相似度.

本节从数据集中随机抽取20989个转发微博和21054个忽略微博(即用户阅读了微博, 但是未转发该微博. 为确保用户阅读了微博, 抽取的忽略微博的阅读延迟时间小于15 min)分别作为实验组和参照组, 首先检验同一影响因素在实验组和参照组中是否有显著差异, 之后比较同一影响因素在实验组和参照组中的累积概率分布.

实验组和对照组的四个转发影响因素的均值和标准差如表1所列, 各影响因素原始数值已经经过以自然常数为底的对数处理. 用户的交互强度原始值为非负数, 其值越大交互强度就越大. 微博主题和用户主题兴趣相似度原始值在0与1之间, 取对数后为非正数, 其值越大相似度也就越大. 以实验组和对照组的影响因素的均值相等为零假设 H_0, 统计分析表明在0.01显著水平下拒绝原假设, 认为上游用户的粉丝数、上游用户的微博平均转发数、用户交互强度、微博主题和用户主题兴趣相似度四个因素在实验组和参照组中有显著差异.

表1 各影响因素的统计信息

Table 1. The statistics of influencing factors.

影响因素	均值	标准差
上游用户粉丝数(实验组/参照组)	14.32/13.92	2.58/3.40
上游用户微博平均转发数(实验组/参照组)	5.24/4.38	3.92/5.60
用户交互强度(实验组/参照组)	1.69/0.92	1.48/1.19
微博主题和用户主题兴趣相似度(实验组/参照组)	−3.50/ − 4.31	1.41/1.66

图4是四个影响因素的累积概率分布图, 该分布差异越大说明影响因素越能区分转发行为和未转发行为. 可见, 虽然假设检验表明四个影响因素在实验组和对照组中有显著差异, 但粉丝数和微博平均转发数不能很好地区分转发行为和未转发行为, 尤其在粉丝数和微博平均转发数达到一定数量后, 累积概率分布差异很小, 仅依靠粉丝数或微博平均转发数不能很好地区分转发和未转发行为, 而用户交互强度、微博主题和用户主题兴趣相似度则可以很好的区分这两种行为.

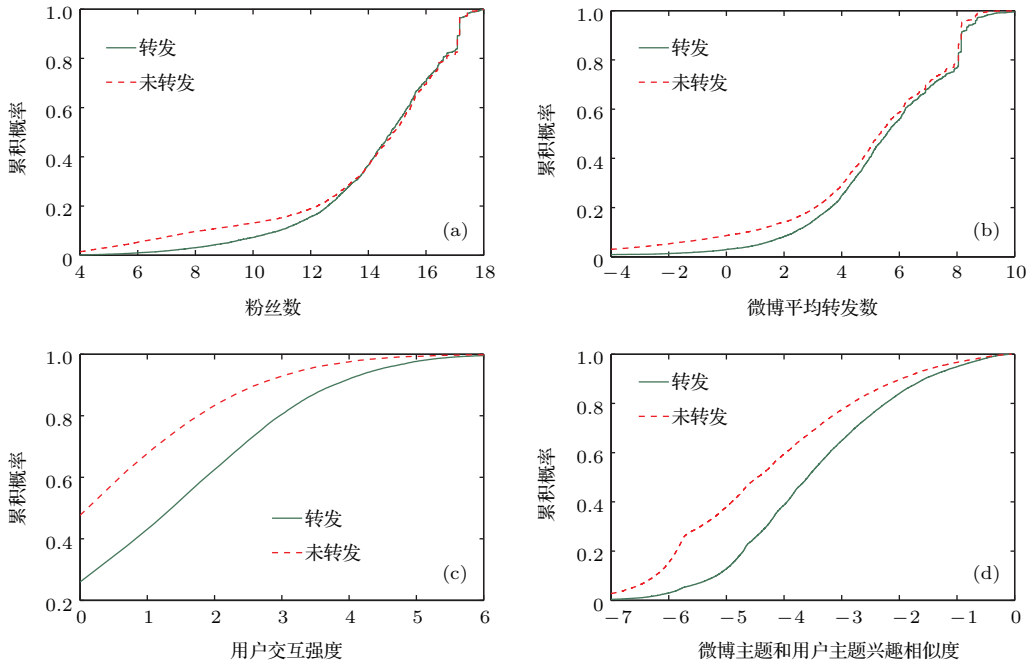

图4 各影响因素累积概率分布 (a) 粉丝数; (b) 微博平均转发数; (c) 用户交互强度; (d) 微博主题和用户主题兴趣相似度

Fig. 4. The cumulative probability distributions of various influencing factors: (a) the number of fans; (b) the average reposted number of microblog; (c) the intensity of users' interaction; (d) the similarity between microblog topics and users' topic interests.

3 微博网络信息扩散预测模型

3.1 模型的假设和信息扩散机制

基于微博用户行为的分析, 本研究假设: 用户只阅读和转发其关注的用户发表的微博; 用户只有在阅读完某条微博后才决定是否转发该微博; 用户每次登录微博后拥有一个固定不变的阅读量, 阅读量用阅读延迟时间表示, 假设用户 u 的阅读量为 t_{u_intvl}, 则该用户登陆后会无差别地阅读完阅读延迟小于 t_{u_intvl} 的微博.

显然, 用户是否转发微博是由用户阅读到这条微博的概率和用户对微博的转发概率决定的, 设 $p_1(\cdot)$ 为用户阅读到微博的概率, $p_2(\cdot)$ 为用户转发阅读到的微博的概率. 对一条微博而言, 不同用户的活跃程度、登录时间以及登陆后的阅读量不同, 他们阅读到该微博的概率也就不同, 此外, 不同用户对同一微博内容的感兴趣程度也不同, 因此 $p_1(\cdot)$、$p_2(\cdot)$ 会因人而异. 本研究基于用户行为提出可预测微博扩散效果的 Susceptible-Infected-Recovered based on Users' Behaviors(SIRUB) 模型, 模型将微博网络中的节点分为易感节点 (S)、转发节点 (I) 和免疫节点 (R). 易感节点指未阅读到微博的节点, 转发节点指转发了微博的节点, 免疫节点指阅读到微博但是没有转发的节点. 模型的扩散机制如下:

1) 当易感节点 (S) 所关注的节点变为转发节点 (I) 后, 他以概率 $p_1(\cdot)$ 阅读微博;

2) 阅读到微博的易感节点 (S) 以概率 $p_2(\cdot)$ 变为转发节点 (I), 以概率 $1-p_2(\cdot)$ 变为免疫节点 (R);

3) 转发节点不会改变状态, 一直处于已转发状态, 等待他的粉丝节点阅读微博.

3.2 阅读概率计算

假设用户 u 在一天中各时间段登录微博的概率密度为 $f_u(t)$, 则 u 在两个时间点 t_1, t_2 之间登录微博的概率为 $p'_u(t_1,t_2)=\int_{t_1}^{t_2} f_u(t)\mathrm{d}t$, 设 $\Delta t=t_2-t_1$, 当 Δt 足够小时有 $p'_u(t_1,t_2)=\Delta t\cdot f_u(t)$. 根据文献 [21], 本文利用公式 $p'_u(t_1,t_2)=1-(1-\rho_u(t_1,t_2))^{\alpha n_u}$ 来推测用户在时间区间 t_2-t_1 内登录微博的概率, 其中 $1/\alpha$ 表示用户每次访问微博平均发文数量, n_u 表示用户 u 一天平均发布微博的数量, αn_u 则表示用户 u 平均每天访问微博的次数, $\rho_u(t_1,t_2)$ 表示 u 在 t_2-t_1 时间区间内发布的微博数量与 u 发布的微博总数的比值.

阅读延迟主要是由微博发表时间和用户登陆时间间隔引起的, 这种时间间隔越大, 微博被阅读的可能性也就越小, 超过一定时间间隔后微博就不会再被阅读到. 用二值函数 $h_u(t_u-t_{w_v})$ 表示用户 u 在时间 t_u 登陆微博后是否会阅读到所关注用户 v 于时间 t_{w_v} 发表的微博 w_v, 如果 u 阅读到了微博 w_v, 则该函数为 1 否则为 0, 用 t_{last} 表示 u 某次登陆后阅读最后一条微博的时间点, 则

$$h_u(t_u-t_{w_v})=\begin{cases}1 & t_u-t_{w_v}<t_u-t_{\mathrm{last}},\\ 0 & t_u-t_{w_v}>t_u-t_{\mathrm{last}}.\end{cases}\tag{1}$$

模型假设每位用户都有一个不变的阅读量, 用户登录微博后能否阅读到具体的某条微博是由用户登陆时间和该微博的发布时间决定的, 因此 (1) 式可写为

$$h_u(t_u-t_{w_v})=\begin{cases}1 & t_u-t_{w_v}<t_{u\_\mathrm{intvl}},\\ 0 & t_u-t_{w_v}>t_{u\_\mathrm{intvl}}.\end{cases}\tag{2}$$

2.2.1 节分析了各用户在一天内各时间段使用微博的概率, 至此可以得出用户 u 在 t_u 时刻登录并且阅读到微博 w_v 的概率

$$\begin{aligned}p_1(u,w_v,t_u)&=\Delta t\cdot f_u(t_u)\cdot h_u(t_u-t_{w_v})\\ &\approx(1-(1-\rho_u(t_u,t_u+\Delta t))^{\alpha n_u})\\ &\quad\cdot h_u(t_u-t_{w_v}).\end{aligned}\tag{3}$$

3.3 转发概率计算

为了预测用户的转发行为, 需要选取影响转发行为的因素. 本研究选取包括微博发布者特征、微博文本特征和社交关系特征三方面的 16 个影响因素来预测用户未知的转发行为, 如表 2 所列. 其中微博平均转发数是指上游用户平均每条微博被转发的次数, 微博平均转发率是指上游用户平均每条微博被其粉丝转发的比例. 用户主题兴趣相似度的计算需要用到用户主题兴趣分布 DT 矩阵, 根据文献 [6] 对用户 u 和 v 主题兴趣差异的定义: $\mathrm{dist}(u,v)=\sqrt{2D_{\mathrm{JS}}(u,v)}$, 可以测量两位用户间主题兴趣的差异, 其中 $D_{\mathrm{JS}}(u,v)$ 表示两位用户主题分布的 Jensen-Shannon 散度, 其他因素可从微博本身特征中直接获取. 因素 6, 10—14 只有是和否两种情况, 统一采用 1 表示是, 0 表示否, 因素 1—5, 7, 8, 16 原始数值都经过了底为自然常数的对数处理.

表 2 影响用户转发行为的因素

Table 2. The factors that influence users' reposting behavior.

特征类别	序号	特征名称
上游用户特征	1	粉丝数
	2	微博数
	3	微博平均转发数
	4	微博平均转发率
	5	注册时间
	6	是否认证
社交关系特征	7	用户交互强度
	8	用户主题兴趣相似度
微博文本特征	9	微博长度
	10	是否包含 URL
	11	是否提及他人
	12	是否包含 Hashtag
	13	是否非转发微博
	14	是否包含图片
	15	发文时间 (取值范围为 0—23)
	16	微博主题和用户主题兴趣相似度

对用户 u 转发行为的预测属于二分类问题, 将微博历史数据 M_u 作为训练集, 通过对用户 u 的历史行为的分析预测该用户对未知微博的转发概率. 本研究采用逻辑回归模型计算用户的转发概率:

$$\begin{aligned} p_2(u, w_v) &= p(y_u = 1|\boldsymbol{X}_{w_v}) \\ &= 1/[1 + \mathrm{e}^{-(b+\boldsymbol{B}_u \cdot \boldsymbol{X}_{w_v})}], \end{aligned} \tag{4}$$

其中 y_u 表示用户 u 对微博的转发决策, $y_u = 1$ 表示用户转发该微博, 否则为 0, $\boldsymbol{X}_{w_v}$ 表示微博 w_v 的特征集合, 包含了表 2 中提到的 16 个特征, 它为模型自变量, $\boldsymbol{B}_u$ 为自变量系数, b 为常系数. 在微博集合 M_u 下, 用 y_1, y_2, $\cdots$, $y_{|M_u|}$ 表示转发决策观察值, 对逻辑回归模型进行最大似然估计, 得到似然函数

$$L(\boldsymbol{B}_u, b) = \prod_{i=1}^{|M_u|} p_i^{y_i} (1 - p_i)^{(1-y_i)}, \tag{5}$$

p_i 表示 (4) 式的转发概率, 当 $\nabla_{\boldsymbol{B}_u} \ln L(\boldsymbol{B}_u, b) = \partial \ln L(\boldsymbol{B}_u, b)/\partial \boldsymbol{B}_u = \boldsymbol{0}$, $\partial \ln L(\boldsymbol{B}_u, b)/\partial b = 0$ 时可得 $\boldsymbol{B}_u$, b, 对于要预测的微博只要知道它的特征集合就可以用 $\boldsymbol{B}_u$ 和 b 计算用户 u 的转发概率.

了解各影响因素的系数的分布有助于深入理解微博网络中的信息扩散. 本研究对收集的 21992 位用户进行逻辑回归分析, 除去没有关注关系的用户共得到 20978 位用户, 因此得到一个 20978×16 的系数矩阵, 每个因素的系数分布如图 5 所示. 我们发现, 第一, 影响因素系数分布大部分属于正态分布, 只有少部分向左或向右偏, 如注册时间长度、是否非转发微博等, 说明大部分影响因素只对部分用户的转发行为有显著影响, 并且这种影响既有正向又有负向, 只有少部分影响因素偏负向或正向; 第二, 影响因素系数大小不同, 有的偏大, 有的则偏小, 如是否非转发微博、平均转发率, 说明各影响因素对用户转发行为的影响程度存在差异; 第三, 用户交互强度的系数分布在横坐标为 0.085 时有一个异常峰值, 本文在确认数据处理无误的情况下尚无法解释该异常峰值出现的原因.

3.4 微博扩散范围

令 $S(t)$, $I(t)$ 和 $R(t)$ 分别表示易感节点、转发节点和免疫节点在 t 时刻的数量, F_v 表示用户 v 的粉丝集合, 则微博 w_v 在 t 时刻的扩散过程为

$$\begin{cases} S(t+\Delta t) - S(t) \\ = -\displaystyle\sum_{v \in I(t)} \sum_{u \in (F_v \cap S(t))} (1-(1-\rho_u(t, t+\Delta t))^{\alpha n_u}) \\ \quad \times h_u(t - t_{w_v}), \\ I(t+\Delta t) - I(t) \\ = \displaystyle\sum_{v \in I(t)} \sum_{u \in (F_v \cap S(t))} (1 - (1 - \rho_u(t, t+\Delta t))^{\alpha n_u}) \\ \quad \times h_u(t - t_{w_v}) \cdot \dfrac{1}{1 + \mathrm{e}^{-(b+\boldsymbol{B}_u \cdot \boldsymbol{X}_{w_v})}}. \end{cases} \tag{6}$$

因为各用户的 $1/\alpha$ 和阅读量不能从用户的历史数据中计算出来, 所以本研究根据用户同质相聚的现象为每个用户的各度粉丝设置相同的 $1/\alpha$ 和阅读量, 并在后续实验中为每个用户的粉丝寻找最优的 $1/\alpha$ 和阅读量. 直接求解 (6) 式较困难, 用离散求解的方法可得在稳态时微博 w_v 被转发的数量的期望 I_{w_v}, 依此可测量微博的扩散效果:

$$\begin{aligned} I_{w_v} = & \sum_{u \in F_v} (1 - (1 - \rho_u(t_{w_v}, t_{w_v} + 1))^{\alpha n_u})\, h_u(1) \\ & \times \frac{1}{1 + \mathrm{e}^{-(b+\boldsymbol{B}_u \cdot \boldsymbol{X}_{w_v})}} + \sum_{u \in F_v} \Bigg[\sum_{\tau=2}^{t_{u\_\mathrm{intvl}}} (1 - (1 \\ & - \rho_u(t_{w_v}, t_{w_v} + \tau))^{\alpha n_u}) \cdot h_u(\tau) \\ & \times \frac{1}{1 + \mathrm{e}^{-(b+\boldsymbol{B}_u \cdot \boldsymbol{X}_{w_v})}} + I_{w_u} \Bigg], \end{aligned} \tag{7}$$

其中 τ 是离散计算的时步, I_{w_u} 是微博从节点 u 开始的扩散范围. 实验在用户转发关系网络中进行, 计算时步 τ 取 1/4 h.

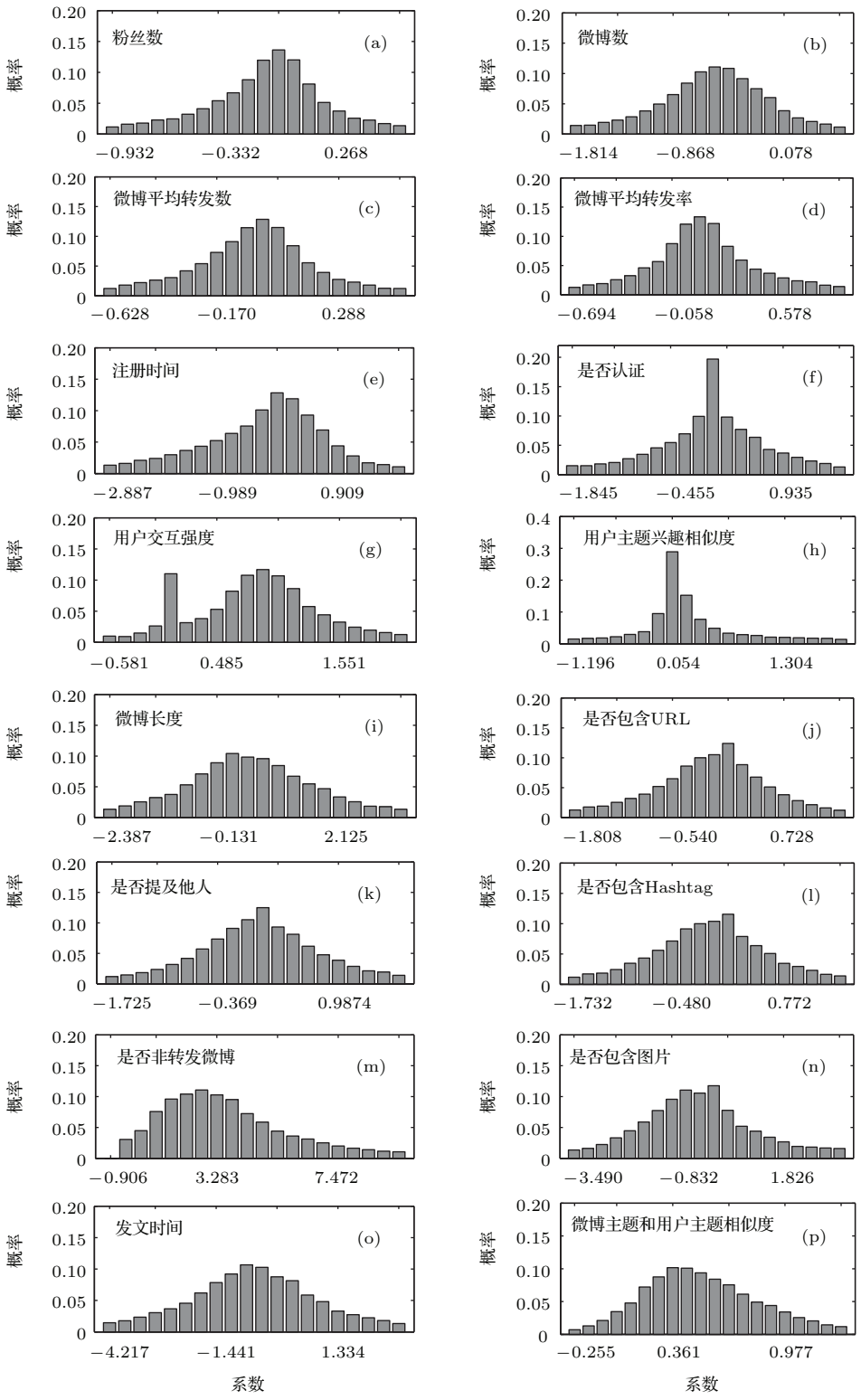

图 5 各影响因素系数概率分布图 (a) 粉丝数; (b) 微博数; (c) 微博平均转发数; (d) 微博平均转发率; (e) 注册时间; (f) 是否认证; (g) 用户交互强度; (h) 用户主题兴趣相似度; (i) 微博长度; (j) 是否包含 URL; (k) 是否提及他人; (l) 是否包含 Hashtag; (m) 是否非转发微博; (n) 是否包含图片; (o) 发文时间; (p) 微博主题和用户主题兴趣相似度

Fig. 5. The probability distributions of coefficients of various influencing factors: (a) The number of fans; (b) the number of microblog; (c) the average reposted number of microblog; (d) the average reposted rate of microblog; (e) registration time; (f) authentication; (g) the intensity of users' interaction; (h) the topic interest similarity of users; (i) microblog length; (j) including URL; (k) mentioning other users; (l) including Hashtag; (m) not reposted microblog; (n) including pictures; (o) posting time; (p) the similarity between microblog topics and users' topic interests.

4 实验与结果分析

本节用收集的数据验证SIRUB模型对信息扩散预测的有效性. 前2/3的数据用于计算模型参数 n_u, ρ_u, 逻辑回归模型系数, 并为用户各度粉丝确定最佳的阅读量 t_{u_intvl} 和每次登陆发文数量 $1/\alpha$, 从而确定 $p'_u(t_1,t_2)$ 和 $h_u(\tau)$, 后1/3的数据用于检验模型的有效性.

经典Susceptible-Infected-Recovered (SIR)模型是其他传染病模型的基础, 而Susceptible-Infected-Contacted-Recovered (SICR)模型是文献[15]依据微博转发机制提出的信息扩散模型, 较符合微博网络环境, 因此本研究将SIRUB模型与经典SIR和SICR模型的预测结果进行比较. 对于经典SIR模型, 易感节点(S)和转发节点(I)接触后以概率 β 变为转发节点, 同时转发节点以概率 γ 变为免疫节点(R), 不失一般性本文假设 $\gamma=1$, 并利用训练数据集以0.001的增量在0至1之间寻找最优的 β. 而对于SICR模型, 易感节点(S)和转发节点(I)接触后以概率 β 变为转发节点, 未变为转发节点(I)的则变为接触节点(C), 接触节点(C)或者以概率 δ 自发的变为免疫节点(R), 或者以概率 β 再次被它的处于转发状态的邻居节点感染. 本研究将 δ 设定为0.4, 并同样利用训练数据集以0.001的增量在0至1之间为SICR寻找最优的 β.

为了更好地比较不同模型的预测效果, 本节以微博平均转发次数最多的50个用户和他们所发的微博作为研究对象, 分析模型对用户转发行为的预测效果以及模型对微博扩散效果的预测准确度.

4.1 对转发行为的预测

筛选的50位用户在测试集中共有2306条微博, 我们以这些微博在微博网络中的22249个转发行为为预测样本, 实验时, 以这50位用户为信息源, 每条微博从信息源开始, 预测微博在网络中扩散时哪些用户转发了该微博. 对于一个样本微博 w_v, 可以用SIRUB模型计算用户 u 从用户 v 阅读并转发微博 w_v 的概率 $p_1(u,w_v)\cdot p_2(u,w_v)$(下文称为预测概率), 模型的预测概率越大表示用户越有可能转发微博. 为了衡量SIRUB模型的预测效果, 本文给定不同的阈值 θ, 考察模型在某一阈值下预测概率大于该阈值的用户的准确率、召回率和 F-score值. 准确率是指预测概率大于该阈值的用户中真实转发的用户所占比例, 召回率是指预测概率大于阈值并且真实转发了微博的用户占所有真实转发用户的比例, 分别用 P 和 R 表示准确率和召回率, 则 $F\text{-score}=2PR/(P+R)$, F-score折中考察了准确率和召回率. 我们先分析SIRUB模型的阅读概率和转发概率对转发行为的预测效果, 再将预测效果与经典SIR和SICR模型比较.

4.1.1 阅读概率和转发概率对SIRUB模型的影响

考察SIRUB模型在只考虑阅读概率和转发概率或者同时考虑两者时对用户转发行为的预测效果, 结果如图6所示. 图6(a)表明三条曲线均随阈值的增加而增大, 在阈值为0.99时分别达到最大值0.029, 0.018和0.321, 同时注意到, 在同时考虑阅读概率和转发概率时预测准确率远高于只考虑其一的准确率, 表明同时考虑两种概率可以提高识别用户转发行为的准确率. 其原因是阅读概率 $p_1(\cdot)$ 越大表明用户登录微博的概率越高, 同时登录微博后的信息阅读量越大, 而转发概率 $p_2(\cdot)$ 越大表明用户对微博越感兴趣, 因此 $p_1(\cdot)\cdot p_2(\cdot)$ 越大的用户越有可能转发微博. 对于一条微博而言, 用户登录概率高或者用户对该微博感兴趣, 都不能很准确地判断用户的转发行为, 只有在用户登录概率高同时对微博很感兴趣时, 才能较为准确地判断用户的转发行为.

图6(b)表明只考虑阅读概率或转发概率时的召回率比同时考虑两者的高, 原因是同时考虑二者时用于预测用户转发行为的概率比只考虑其一时的低. 图6(c)表明从综合指标 F-score来看, 同时考虑阅读概率和转发概率时的预测效果远比只考虑其一时要好.

4.1.2 模型预测效果比较

对于经典SIR和SICR模型, 本文通过训练寻找最优的转发概率, 用此概率预测用户的转发行为. 表3给出了经典SIR、SICR和SIRUB模型的最优预测结果. 可见SIRUB模型的准确率为0.201, 优于SIR和SICR模型, 同样SIRUB模型的 F-score为0.228, 也优于SIR和SICR模型. SIR和SICR模型的召回率都比SIRUB模型高, 这是因为二者对转发行为的预测是一个以转发概率 β 为参数的0-1分布, β 即为转发用户的微博的粉丝数量占用户所有粉丝数量的比例, 因此可以认为 β 也是SIR、SICR模型预测的召回率 R, 而准确率 P

是相对不变的. 由 F-score=$2PR/(P+R)$, 可得 F-score=$2P/(P/R+1)$, F-score 是 R 的单调增函数, 因此当转发概率在最大值附近时, 召回率 R 和 F-score 取得最大值.

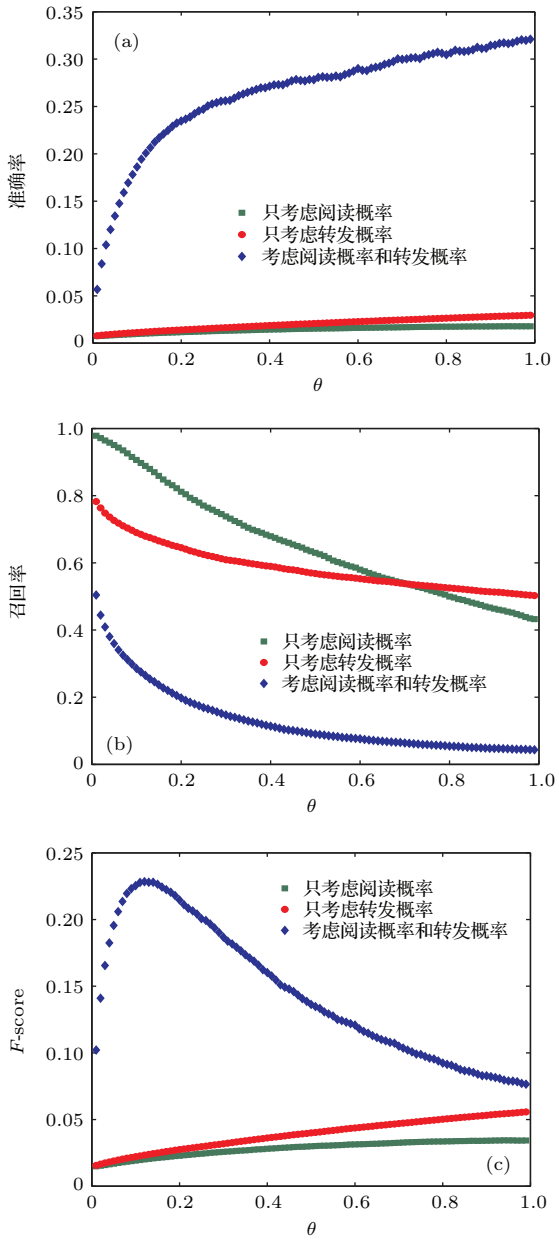

图 6 (网刊彩色) 不同阈值下 SIRUB 对转发行为的预测效果 (a) 准确率; (b) 召回率; (c) F-score

Fig. 6. (color online) The prediction results of SIRUB model for reposting behavior under different thresholds: (a) Precision; (b) recall; (c) F-score.

在一个近 2.2 万个节点、59 万条边的微博网络中预测一条微博从某一条边被转发是一项非常困难的任务, 且预测难度随着网络规模的增大而增加. 文献 [9] 提出的因子图模型预测的 F-score 为 0.325, 文献 [20] 利用随机场方法在 1000 个节点的网络中可以获得 F-score 为 0.662 的预测效果, 但该方法的预测效果随着网络规模的增加而下降. 文献 [20] 没有进一步增大网络, 本文的网络规模是文献 [20] 的 22 倍, 文献 [9] 也没有提供实验网络的规模, 所以本文暂不能对各方法进行评价.

表 3 经典 SIR, SICR 和 SIRUB 模型对转发行为的预测结果

Table 3. The prediction results of classic SIR, SICR and SIRUB models for reposting behavior.

	准确率	召回率	F-score
SIR	0.020	0.947	0.039
SICR	0.019	0.985	0.037
SIRUB	0.201	0.265	0.228

4.2 对微博扩散效果的预测

SIRUB 模型可以预测一条微博在微博网络中的扩散效果, 我们以筛选的 50 位用户的测试集微博在网络中的转发次数为预测对象, 对于用户 v 可得其所发微博的平均转发次数

$$\bar{I}_v = \sum_{w_v \in W_v} I_{w_v}/|W_v|,$$

其中 W_v 表示用户 v 所发微博的集合. 图 7 按实际平均转发量从小到大给出了微博转发次数最多的 50 位用户的微博预测结果和真实值, 可见 SIRUB 模型的预测结果和真实转发量较为接近, 两者走势相似, 曲线波动较小, 说明该模型可以较准确地预测用户在一段时间内所发微博的平均转发数. SIR 和 SICR 模型的预测曲线走势相似, 不随真实值而改变, 预测结果和真实情况差异较大.

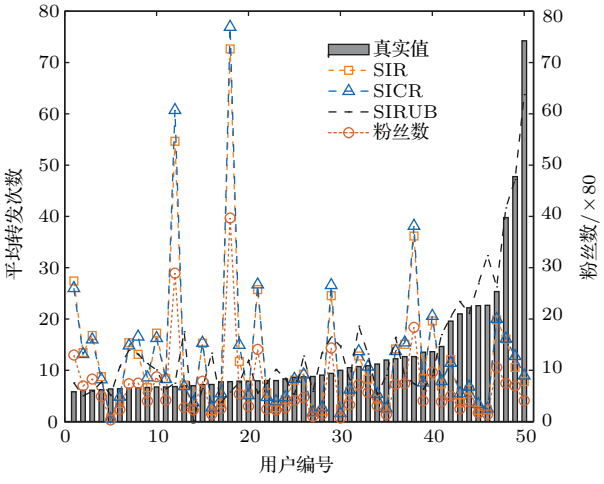

图 7 (网刊彩色) 用户真实平均转发数和模型预测的平均转发数

Fig. 7. (color online) Users' real average reposting numbers and the predicted numbers by models.

图7同时给出了各用户的粉丝数量. 可见SIR和SICR模型预测值与用户的粉丝数走势相同, 而SIRUB模型则与粉丝数没有明显的联系. 原因是SIR和SICR模型假设用户粉丝阅读到了微博同时以相同的转发概率转发微博, 粉丝的数量成了影响用户微博扩散范围的惟一因素. 但正如前文所述, 用户的粉丝之间差异巨大, 粉丝的活跃程度、信息阅读量不同, 不能保证用户所发的每条微博都会无差异地被他的粉丝阅读到, 同时各种影响因素会导致不同粉丝有不同的转发倾向, SIRUB模型同时考虑了用户的阅读行为和转发倾向, 所以较准确地预测了用户微博的转发范围. 虽然SICR模型根据微博环境对SIR模型扩散机制做了修改, 但它没有考虑用户的行为, 在预测效果上并不比SIR 模型好.

进一步地, 图8给出了三种模型预测结果绝对误差的均值和标准差. SIRUB模型的误差均值和标准差总体上均小于SIR和SICR模型, 表明该预测准确度比SIR和SICR模型好, 预测结果也较稳定.

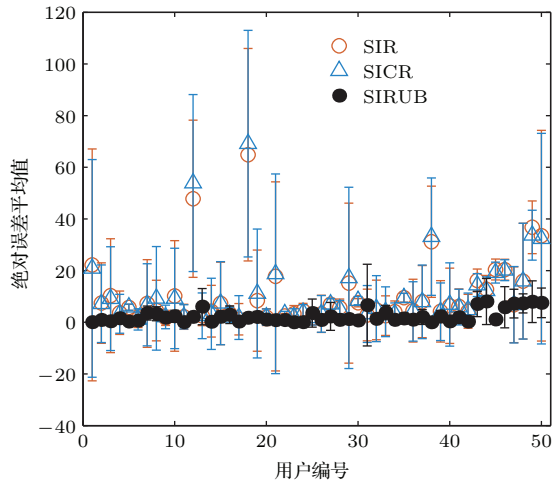

图8 (网刊彩色) SIR, SICR和SIRUB模型预测的平均转发数绝对误差

Fig. 8. (color online) The absolute errors of the predicted average reposting numbers for SIR, SICR and SIRUB models.

5 总结与展望

社交网络对用户行为的记录为研究网络上的信息扩散提供了前所未有的机会, 本研究在分析用户阅读行为和转发行为的基础上, 构建了基于用户行为的SIRUB信息扩散模型. 从对用户转发行为的预测结果看, 只有同时考虑阅读概率和转发概率时SIRUB模型才能较为准确地预测用户的转发行为, SIRUB模型的最优预测结果F-score高于经典SIR和SICR模型. 从对扩散范围的预测结果看, SIRUB 模型的预测效果比较接近真实值, 误差相对较小, 预测结果较稳定. 而SIR和SICR模型预测曲线随用户粉丝数的变化而波动, 与真实值相差较远. 无论是对用户转发行为的预测还是对微博扩散范围的预测, SIRUB模型的预测效果均优于经典SIR和SICR 模型.

本研究仍存在需要改进的地方. 由于历史数据没有记录各用户每次登录的信息阅读量和发文数量, 通过有效途径获取或者推测该类信息是本研究急需改进的地方. 用于计算用户转发概率的各影响因素之间可能存在一定的关联, 例如注册时间越长用户所发的微博数有可能会越多, 这对SIRUB模型的预测准确度有一定的影响. 此外验证其他诸如贝叶斯网络、朴素贝叶斯方法等对用户转发概率计算的有效性也是本研究的下一步研究内容.

参考文献

[1] Xu X K, Hu H B, Zhang L, Wang C J 2015 *Computational Communication on Social Networks* (Beijing: Higher Education Press) p8 (in Chinese) [许小可, 胡海波, 张伦, 王成军 2015 社交网络上的计算传播学 (北京: 高等教育出版社) 第8 页]

[2] Suh B, Hong L, Pirolli P, Chi E H 2010 *IEEE Second International Conference on Social Computing* Minneapolis, MN, USA, August 20–22, 2010 p177

[3] Zhang Y, Lu R, Yang Q 2012 *J. Chin. Inf. Process.* **26** 109 (in Chinese) [张旸, 路荣, 杨青 2012 中文信息学报 **26** 109]

[4] Kwak H, Lee C, Park H, Moon S 2010 *Proceedings of the 19th International Conference on World Wide Web* Raleigh, NC, USA, April 26–30, 2010 p591

[5] Cao J X, Wu J L, Shi W, Liu B, Zheng X, Luo J Z 2014 *Chin. J. Comput.* **37** 779 (in Chinese) [曹玖新, 吴江林, 石伟, 刘波, 郑啸, 罗军舟 2014 计算机学报 **37** 779]

[6] Weng J, Lim E P, Jiang J, He Q 2010 *Proceedings of the Third ACM International Conference on Web Search and Data Mining* New York City, NY, USA, February 3–6, 2010 p261

[7] Liu L, Tang J, Han J, Jiang M, Yang S 2010 *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* Toronto, ON, Canada, October 26–30, 2010 p199

[8] He X, Cheng S, Chen W, Jiang F 2013 *International Conference on Information Society* Toronto, ON, Canada, June 24–26, 2013 p130

[9] Yang Z, Guo J, Cai K, Tang J, Li J, Zhang L, Su Z 2010 *Proceedings of the 19th ACM International Conference*

on Information and Knowledge Management Toronto, ON, Canada, October 26–30, 2010 p1633

[10] Romero M D, Meeder B, Kleinberg J 2011 *Proceedings of the 20th International Conference on World Wide Web* Hyderabad, India, March 28–April 1, 2011 p695

[11] 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, 胡元萍, 刘志宏, 马建峰 2014 物理学报 **63** 180501]

[12] Wang J L, Liu F A, Zhu Z F 2015 *Acta Phys. Sin.* **64** 050501 (in Chinese) [王金龙, 刘方爱, 朱振方 2015 物理学报 **64** 050501]

[13] Li W, Tang S, Fang W, Guo Q, Zhang X, Zheng Z 2015 *Phys. Rev. E* **92** 042810

[14] Wang X J, Song M, Guo S Z, Yang Z L 2015 *Acta Phys. Sin.* **64** 044502 (in Chinese) [王小娟, 宋梅, 郭世泽, 杨子龙 2015 物理学报 **64** 044502]

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

[16] Prakash B A, Beutel A, Rosenfeld R, Faloutsos C 2012 *Proceedings of the 21st International Conference on World Wide Web* Lyon, France, April 16–20, 2012 p1037

[17] Liu H, Xie Y, Hu H, Chen Z 2014 *Int. J. Mod. Phys. C* **25** 1440004

[18] Goel S, Anderson A, Hofman J, Watts D J 2016 *Manage. Sci.* **62** 180

[19] Goyal A, Bonchi F, Lakshmanan L V S 2010 *Proceedings of the Third ACM International Conference on Web Search and Data Mining* New York City, NY, USA, February 3–6, 2010 p241

[20] Peng H K, Zhu J, Piao D, Yan R, Zhang Y 2011 *IEEE 11th International Conference on Data Mining Workshops* Vancouver, BC, Canada, December 11, 2011 p336

[21] Mao J X, Liu Y Q, Zhang M, Ma S P 2014 *Chin. J. Comput.* **37** 791 (in Chinese) [毛佳昕, 刘奕群, 张敏, 马少平 2014 计算机学报 **37** 791]

[22] Iribarren J L, Moro E 2011 *Phys. Rev. E* **84** 046116

[23] Golub B, Jackson M O 2010 *Proc. Natl. Acad. Sci. USA* **107** 10833

[24] Iribarren J L, Moro E 2009 *Phys. Rev. Lett.* **103** 038702

[25] Zhao W X, Jiang J, Weng J, He J, Lim E P, Yan H, Li X 2011 *Proceedings of the 33rd European Conference on Information Retrieval Research* Dublin, Ireland, April 18–21, 2011 p338

Modeling information diffusion on microblog networks based on users' behaviors*

Liu Hong-Li Huang Ya-Li Luo Chun-Hai Hu Hai-Bo†

(*Department of Management Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*)

(Received 13 March 2016; revised manuscript received 3 May 2016)

Abstract

Online social networks, such as Facebook, Twitter and YouTube, play a vital role in information sharing and diffusion, and recently many dynamics models on social networks have been proposed to model information diffusion. However most models are theoretical, their parameters do not come from realistic data and their validity and reliability have not been evaluated empirically. In the paper we first analyze the users' behaviors of reading and reposting microblog in Sina Weibo, a Twitter-like website in China, and find that users' number of fans, the average reposted number of users' microblog, the intensity of users' interaction and the similarity between microblog topics and users' topic interests can significantly influence reposting behavior. Then we propose an information diffusion model Susceptible-Infected-Recovered based on Users' Behaviors (SIRUB) on microblog networks, compute the users' probability of reading microblog in the model according to the probability of their logging on microblog in a day, and obtain the reposting probability utilizing the logistic regression which considers 16 possible factors influencing users' reposting behavior. The 16 factors can be divided into three categories: the characteristics of microblog publishers, microblog text features and social relationship characteristics. We utilize the beginning 2/3 microblog data to obtain model parameters and logistic regression coefficients, and the remaining 1/3 data to examine the validity of the model. The experiments on Sina Weibo network show that the model can predict users' reposting behavior accurately only when it considers both reading and reposting probabilities. F-score which considers precision and recall is used to assess prediction effect of the model. The highest F-score for the prediction of SIRUB model on users' reposting behavior is 0.228 which is much larger than those of classical Susceptible-Infected-Recovered (SIR, F-score=0.039) and Susceptible-Infected-Contacted-Recovered (SICR, F-score=0.037) models. The prediction on the spreading scope of microblog for SIR and SICR models is related with users' number of fans while for SIRUB model not. For SIRUB model the mean and standard deviation of the errors of prediction on spreading scope are smaller than those of SIR and SICR models. These results indicate that users' behaviors of reading and reposting microblog should be appropriately taken in account when modeling information diffusion on microblog networks, and that, in general, the prediction performance of the data-driven SIRUB model proposed in the paper is better than those of SIR and SICR models regardless of the prediction of users' reposting behavior or diffusion scope of microblog.

Keywords: microblog network, user behavior, information diffusion

PACS: 89.65.-s, 87.23.Ge **DOI:** 10.7498/aps.65.158901

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61473119, 61104139), and the Fundamental Research Funds for the Central Universities, China (Grant No. WN1524301).

† Corresponding author. E-mail: hbhu@ecust.edu.cn