基于信息老化特征的微博传播模型研究

杨子龙[1] 黄曙光[1] 王 珍[1] 李永成[2] 肖 佳[3]
(电子工程学院网络系 合肥 230037)[1] (北方电子设备研究所 北京 100083)[2]
(北京邮电大学网络技术研究院 北京 100876)[3]

摘 要 随着微博的迅速兴起,提取信息传播特征和构建传播模型已成为研究热点。针对用户转发行为,首先分析信息转发结构,提取信息老化特征;然后结合转发时效性,基于平均转发概率的递减规律提出SIR的改进模型;最后利用真实转发数据验证了模型的合理性。结果表明,考虑信息时效性和老化特征,能够较好地拟合信息传播过程。进一步,将利用该模型分析不同节点传播影响力,发现其分布服从无标度特征。

关键词 SIR模型,转发长度,转发特征,信息老化,新浪微博
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.018

Study on Micro Blog Reposting Model Based on Characteristics of Information Obsolescence

YANG Zi-long[1] HUANG Shu-guang[1] WANG Zhen[1] LI Yong-cheng[2] XIAO Jia[3]
(Department of Network, Electronic Engineering Institute, Hefei 230037, China)[1]
(The Institute of North Electronic Equipment, Beijing 100083, China)[2]
(Institute of Networking Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)[3]

Abstract With the rapid development of the micro blog, extracting the characteristics of the message propagation and constructing the propagation model have already been a hot topic. Focused on users' reposting behavior, first we analyzed the structure of the message reposting and extracted the characteristics of information obsolescence. Then we proposed an improved SIR model based on the law of diminishing reposting probability combining the time-effectiveness of the message. At last, we made use of real reposting data to prove the validity of our model. The results show that taking the time-effectiveness and information obsolescence of message into account can fit the progress of the message propagation well. Furthermore, we took advantage of this model to analyze the scale-free property of the vertex influence distribution.

Keywords SIR model, Reposting length, Characteristics of reposting, Information obsolescence, Sina micro blog

1 引言

近年来,微博(Micro Blog)作为一种全新的在线社交应用,得到了快速发展。国内外一系列相关网站迅速崛起,例如美国的Twitter用户数量已经突破2亿,并且依然持续快速增长,是互联网上访问量最大的十个网站之一;国内的新浪微博用户数量已超过5亿,占据全国微博用户总量的57%,以及全国微博活动总量的87%,是中国访问量最大的网站之一。用户量的快速增加,使其逐渐成为发布、共享、传播信息的日常社交方式之一。微博兼具权威媒体发布平台与用户交流平台两大属性,信息可以在微博上裂变式传播,速度和广度远远高于传统媒体及Web应用,因此提取、分析微博中信息传播特征,构建相应的传播模型有着十分重要的理论价值和现实意义。

在微博信息传播过程中,可以选择不同的用户行为作为传播载体:Grabowski、郑蕾等[1,2]研究了微博中的关注行为;苑卫国等[3]研究的是双向关注行为;与关注的低成本不同,"转发行为"[4-10]更有利于实现信息的持续传播。关于提取转发行为中的信息传播特征,现有研究已取得了一定进展:Shaozhi Ye等[4]采集微博用户的转发信息,提取一般性新闻、爆炸性新闻的传播特征,并构建了不同的传播模型;Kwark等[5]和Akshay等[6]根据Twitter的关注、转发等关系构建用户关系网络,分析了网络拓扑特征;陈慧娟等[7]分析了影响信息传播的特征和因素;Zi Yang等[8]分析了影响转发行为的因素,如用户属性、信息内容、时间等;Bongwon等[9]研究了影响微博转发行为的用户属性,包括关注数、粉丝数、账号年龄等;李英乐等[10]在分析影响用户转发行为因素的基础上,提出了用户影响力、用户活跃度、兴趣相似度、微博内容重要性和用户亲密程度5项特征。在对传播特征分析中,除了上述文献所关注的用户属性、信息内容等特征,信息的转发长度也是传播特征的重要组成,于晶等[11]利用真实新浪微博数据,提取信息传播的网络结构及演化特征,分析了网络的循环

到稿日期:2014-02-10 返修日期:2014-05-20

杨子龙(1988—),男,硕士生,主要研究方向为复杂网络、社会网络分析,E-mail:zilongyang1988@gmail.com;**黄曙光**(1960—),男,教授,主要研究方向为复杂网络、社会网络分析、信息安全;**王 珍**(1981—),女,博士,工程师,主要研究方向为复杂网络、社会网络分析;**李永成**(1986—),男,博士,工程师,主要研究方向为复杂网络、社会网络分析;**肖 佳**(1983—),男,博士,主要研究方向为网络管理、社交网络。

结构、信息传播的路径长度以及信息传播网络的异质特征，但是并未对该特征的形成机制进行深入研究。

现有的信息传播模型一般将个体分为 3 个状态：易感状态(S)，感染状态(I)，恢复状态(R)。传染病模型主要有 SI、SIS、SIR、SIRS 模型。针对在线社会网络的信息传播模型，现有研究已取得了一定进展：张彦超等[12]结合复杂网络和传染病动力学理论，考虑了节点度和传播机理的影响，构造了基于在线社交网络的信息传播模型；熊熙等[13]研究社交网络中舆论观点扩散的形式与特征，提出了一种基于在线社交网络的观点传播模型；Fei Xiong 等[14]基于转发机制提出了一种改进的 SIR 模型，用以描述在线微博中的信息传播过程。与传染病的传播方式不同，信息在传播过程中具有生存周期，存在信息老化问题：李慧[15]借助文献老化的研究成果，论述了网络信息老化规律；龚思婷等[16]对网络信息在整个生命周期内信息价值的衍变过程进行了研究；文献[6,17,18]将信息老化规律的研究延展到微博中，针对转发行为的统计信息表明，随着信息转发长度的增加，用户对于信息的转发意愿不断降低。因此，本文根据信息的时效性，采用平均转发概率，提取转发行为中的信息老化特征，构建基于传播概率递减的 SIR 模型，并通过真实数据的转发长度拟合模型参数，验证了模型的合理性。

本文首先利用真实数据分析了微博中信息转发特征；然后构建基于传播概率递减的 SIR 模型；最后利用真实转发网络，基于改进 SIR 模型对信息传播进行模拟，并分析不同用户的传播影响力。

2 信息转发特征

用户的转发行为是微博中重要的信息传播机制。由于转发功能的存在，用户发布的信息得以快速扩散，这不仅影响与发布者直接相关的用户，还会影响与其间接相关的用户。Kwak 等[5]研究热门话题和一般话题发现，热门话题发布信息的最终转发次数都能够达到 1000 左右，信息的第二层、第三层以及第四层的传播加速了信息传播速度，而这些与用户粉丝数、关注数具有较弱相关性。因此，本文采用转发网络作为信息传播载体，对于某条信息，用户通过连续转发将信息逐层传播，从而在整个用户群体中形成巨大影响力。例如某一用户发布了一条信息，那么他所有的粉丝都会以广播的形式接收到，而其中部分用户会认同该信息，对该信息进行转发，重复上述过程，使得信息的影响力不断增加，传播范围不断扩大。

本节随机选取了某一时刻热门微博共计 60 条，平均每条信息的转发次数为 985.8，其中一条微博信息的转发结构如图 1 所示。

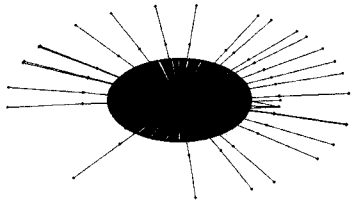


图 1 一条微博信息的转发结构图

在分析信息转发特征之前，首先定义信息转发长度用于表征转发用户与初始发布者的距离：如果信息直接转自初始发布者，则转发长度为 1；而对转发长度为 1 的信息再进行转发，则转发长度为 2，以此类推，每经过一次转发，信息转发长度就增加 1。在信息传播过程中，随着信息转发长度的增加，往往存在着老化特征。如果假定用户转发概率恒定，则通过持续转发可以得到较高的信息转发长度，但实际上，转发路径的长度通常较短：Kwak 等[5]分析 Twitter 的转发路径表明，97.6%的转发路径都小于 6，且最长距离也不会超过 11。沈珂轶等[17]认为这主要是因为随着信息传播层数的增加，人们对信息的兴趣是递减的。Eytan Baks 等[18]同样在研究中发现信息的最大转发长度只能达到 9，并且随着转发长度的增加，出现频率急剧下降。

本文采集 60 次微博信连续息转发过程，分析最大转发长度(见图 2)可以发现：一方面，信息的最大转发长度均大于 1，说明在信息传播过程中存在转发行为；另一方面，最大转发长度平均值为 3.28，众数为 3，最大长度为 11，这说明转发机制并不能使传播无限制地进行，转发长度依然被限制在一定的范围内。

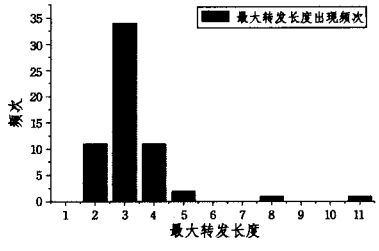


图 2 最大转发长度统计结果

为了进一步了解微博信息的转发特征，对 60 次微博信息转发过程中所有信息转发长度进行统计(见图 3)，分析表明：随着信息转发长度的增加，相应转发长度的信息在所有信息中所占的比例迅速下降，这说明随着传播距离的增大，微博的传播能力逐渐减弱。信息在距离初始节点较短的距离内具有较强的影响力，这就使得信息获得了相应的间接影响力，而随着传播距离的增加，这种影响力则迅速衰减，从而使得信息不能无节制地传播，使其影响力局限在一定的范围内。这实际反映出信息在传播过程中其价值是在不断老化衰减的，如图 3 所示，信息的老化规律基本符合幂律递减规律。

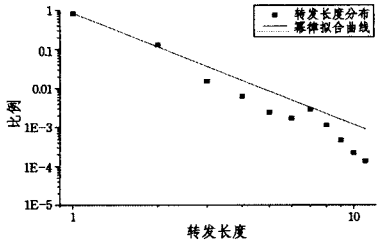


图 3 转发长度分布拟合对比

3 基于信息老化特征的改进 SIR 模型

在信息传播过程中，经典的信息转发模型(网络 SIR 模型)中状态为 S(易感状态)的个体可以通过被传染而以一定概率变为状态 I(感染状态)，而状态为 I 的个体则可能以一定概率恢复为 R(恢复状态)。其中，针对微博中一条特定信息的转发，感染则对应转发信息，尚未转发信息的用户对应 S，已经转发过信息的用户则不会对该信息进行重复转发，从而对应 R，而正在转发信息的用户对应 I。一个尚未转发过信息的用户 u，收到信息并以一定概率进行转发后，其状态变为 I，那么在有向网络中，沿着网络中从 u 出发的边与 u 相连的用

户就会收到这条信息，并以一定的概率对信息继续转发，同时，用户 u 转发过信息后状态变为 R，随后既不会影响其他用户也不会受到其他用户影响，过程如图 4 所示。

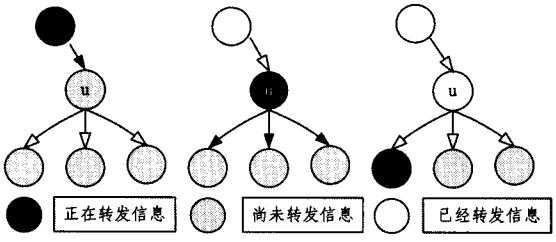


图 4　传播模型示意图

信息的转发行为具有明显的时效性，Kwak 等[5]研究发现：有一半的 Twitter 博文是在一个小时之内被转发的，75% 的在一天之内转发；沈珂轶等[17]的研究也表明，平均 90% 的微博在 25.1 分钟后就无人问津了，而平均 95% 的微博在发布 51.7 分钟后，就不再被转发。大多数的转发行为都发生在收到微博后的 75.4 分钟内。这都说明了信息转发存在明显的生命周期，信息由发布到被转发只能发生在一个较短的时间段内，那么在这个时间段内，设其粉丝用户接收并对信息进行转发的概率为 p，而由于正在转发的信息的用户不会再被其它用户转发也不会再次转发信息，因此其状态以 100% 的概率在下一个时间段变为 R。

利用平均场论考虑网络传播的平均情况。一条特定信息通常是由某一个用户发出，那么 $I(0)=1, R(0)=0$。在时刻 1，转发的期望值为网络平均出度乘以转发概率，即：

$$I(1)=p\cdot K_{out} \tag{1}$$

其中，K_{out} 表示网络中节点的平均出度。同时有 $R(1)=1$。在时刻 2 以后，感染节点的平均出度不再为网络中节点平均出度，而是网络平均额外出度[19]，用 EK_{out} 表示，那么当 $k>1$ 时，有：

$$\begin{aligned} I(k)&=I(k-1)\cdot p\cdot EK_{out}\\ &=(K_{out}/EK_{out})\cdot(p\cdot EK_{out})^k \end{aligned} \tag{2}$$

$$R(k)=\frac{K_{out}\cdot(p\cdot EK_{out}-(p\cdot EK_{out})^{k+1})}{EK_{out}\cdot(1-p\cdot EK_{out})}+1 \tag{3}$$

根据实际信息转发长度特征，若 $I(k)$ 随着 k 的增加而递减，则需要 $(p\cdot EK_{out})$ 小于 1，但这样会出现两个问题：1）实际 $I(k)$ 更加符合幂律递减，而固定概率的传播只能按照指数递减，两者有较大差异；2）即使按照指数进行拟合，p 会是很小的值，根据公式(3)，每个节点最终只能影响 0.18 个人，会导致传播无法进行，这显然与实际情况相违背。

因此假设传播概率随时间变化，根据上一节描述的信息转发特征，即 $I(k)\sim k^{\lambda}$，设 $I(1)=p(1)\cdot k_{out}=a$，其中 $a>0$ 为常数，那么：

$$I(k)=I(k-1)\cdot p(k-1)\cdot EK_{out}=a\cdot k^{\lambda} \tag{4}$$

当 $k>1$ 时，可以得到传播概率：

$$p(k)=(k/(k+1))^{\lambda}EK_{out}^{-1} \tag{5}$$

那么对于 $R(k)$，当 $k>1$ 时：

$$\begin{aligned} R(k)&=a\sum_{x=1}^{ste_{\max}}x^{\lambda}+1\\ &\approx\frac{((ste_{\max}+1)^{\lambda+1}-1)\cdot p(1)\cdot K_{out}}{\lambda+1}+1 \end{aligned} \tag{6}$$

其中，$ste_{\max}$ 表示最大传播步长。

4　实验分析

本节根据上节提出的改进 SIR 模型，以微博转发网络为信息传播载体，模拟信息传播过程，并进一步利用该模型分析节点影响力的分布。

4.1　微博转发网络

为了研究微博中的信息传播，本文首先构建微博转发网络。研究用户限定为新浪名人堂用户，这些用户是经过新浪人工审核通过的具有行业影响力的认证用户，这一方面确保了用户的真实性和用户的质量，另一方面也有效地减少了工作量。

数据采集过程如下：首先将限定时间段内的微博消息作为原始数据，然后对于某一条用户 i 转发自用户 j 的信息，微博信息中会包含用户 i 和 j 的 id，检查现有网络中是否有表示用户 i、j 的节点以及其之间是否有边，若不存在这样的节点或边，则将其添加到网络中。按照这样的方式，本文分析 2012.09.23－2012.10.23 这一个月内的名人堂用户转发信息，最终得到了一个由 92933 个节点、1083584 条边构成的转发网络，网络参数如表 1 所列。

表 1　网络参数统计

参数	结果	参数	结果
节点数	92933	平均最短路径长度	2.99
边数	1083584	强连通分支规模	57532
出/入度平均值	11.7	弱连通分支规模	90242
入度最大值	860	聚类系数	0.0105
出度最大值	8173	有效直径	2.94

网络的出入度分布如图 5 所示，近似服从幂律分布。

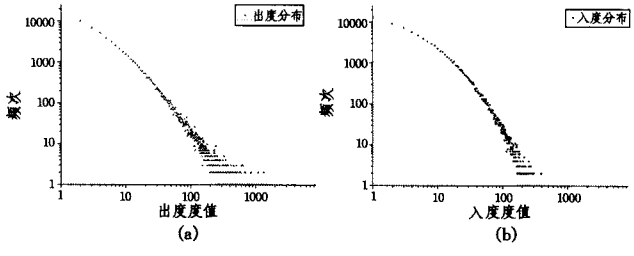


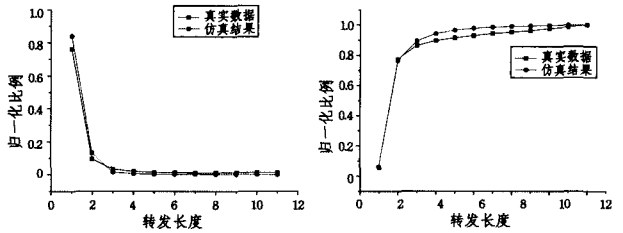
图 5　微博有向转发网络出入度分布

4.2　传播仿真

对于真实转发数据，前文已知其转发长度分布近似服从幂律分布，进一步通过 origin 软件，本文对分布曲线进行幂律函数 $y=a\cdot x^{b}$ 拟合，得到参数 $a=0.838, b=-2.85$，该分布函数实际对应于式(4)中 $I(k)$ 服从的幂律函数 $I(k)=a\cdot k^{\lambda}$，因此利用得到的拟合参数可以确定式(4)－式(6)的指数参数 λ，使模型特征与实际数据相吻合。根据建立的 SIR 模型，对网络中的信息传播进行模拟，实验次数为 20000，每次随机选取初始节点通过转发进行信息传播，观察不同状态的节点比例变化情况。微博的全体用户数量庞大，当某一个用户发出一条信息，其最终影响的用户数量相对于整个网络而言十分微小，例如在采集热门微博转发数据中，平均每条信息会被转发 1000 次，与微博用户数量相比并不显著。同样，在我们的仿真实验中，绝大部分信息影响的节点不超过节点总数的 1%，因此网络中未受到信息影响的节点即状态为 S 的节点的变化幅度微小，本文主要关注状态为 I 和 R 的节点的变化情况。

仿真结果如图 6 所示。由于限定在名人堂用户的微博转发网络仅仅是现实网络的一个极其简化的版本，因此仿真结果与真实结果在绝对数量上有着较大的差异。本文主要关注的是趋势的变化而不是绝对数量的变化，因此为了合理比较，

本文对所有结果都进行了归一化处理，对于状态为 I 的节点，利用感染节点数总和进行归一化；而对于状态为 R 的节点，则利用最终人数进行归一化，从而观察真实数据与仿真结果的变化趋势。真实数据即为前面采集的微博信息转发数据，本文统计了不同转发长度的信息占所有信息的比例。



(a) 状态为 I 的节点比例随转发长度的变化　(b) 状态为 R 的节点比例随转发长度的变化

图 6　改进的 SIR 模型模拟结果与真实数据对比

通过图 6 可以发现，仿真算法可以较好地对真实数据进行拟合，这说明仿真算法在一定程度上反映了真实信息转发规律。在整个传播过程中，已经转发过信息的用户以及正在转发信息的用户在初始时刻迅速增加，而随着转发长度的增加，其增长速度迅速下降。

4.3　不同节点传播范围分布

实际上不同根节点作为初始节点的最终传播范围即代表了不同节点的传播影响力，文献[3,20-22]等通过不同角度研究发现不同节点间影响力差异巨大。本文同时模拟了不同节点作为初始传播节点的转发情况(见图 7)，结果表明：相较于度分布，传播范围分布表现出较强的无标度特征，即不同个体影响力的差异明显。

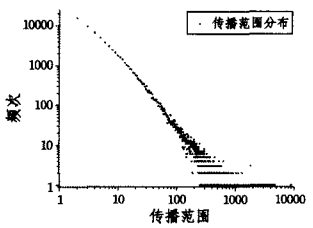


图 7　不同节点传播范围

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的信息老化特征，并将这种规律与 SIR 模型相结合，提出传播概率递减变化的改进网络 SIR 模型。为了验证模型仿真效果，本文收集了名人堂用户信息，构建有向转发网络作为载体，对其中的信息传播进行模拟。仿真结果显示出与实际数据类似的传播递减效应，这说明提出的模型在一定程度上体现了信息转发规律，具有一定参考意义。同时，以不同节点作为信息源，发现节点传播影响力分布服从幂律特征，不同节点具有显著差异。

接下来，我们准备从以下几个方面进行更为深入的研究：1)模型中考虑信息内容对于转发行为的影响。2)分析用户关系网络拓扑结构对于信息传播结构的影响。3)分析不同用户转发行为的差异及相应的影响。

参 考 文 献

[1] Grabowski A, Kosinski R A. Percolation in Real On-line Networks[J]. Acta Physica Polonica B, 2010, 41(5): 1135

[2] 郑蕾，李生红. 基于微博网络的信息传播模型[J]. 通信技术，2012，45(2)：39-41

[3] 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

[4] Ye Shao-zhi, Wu Fe-lix. Measuring Message Propagation and Social Influence on Twitter. Com[C]// Proceedings of the Second International Conference on Social Informatics. Springer Berlin Heidelberg, 2010: 216-231

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

[6] Java A, Song Xiao-dan, Finin T, et al. Why We Twitter: Understanding Microblogging Usage and Communities[C]// Proceedings of the 9th Webkdd and 1st Sna-kdd 2007 Workshop on Web Mining and Social Network Analysis. ACM, 2007: 56-65

[7] 陈慧娟，郑啸，陈欣. 微博网络信息传播研究综述[J]. 计算机应用研究，2014(2)：333-338

[8] Yang Zi, Guo Jing-yi, Cai Ke-ke, et al. Understanding Retweeting Behaviors in Social Networks[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, 2010: 1633-1636

[9] Suh Bong-won, Hong Li-chan, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[C]// 2010 IEEE Second International Conference on Social Computing (SocialCom). IEEE, 2010: 177-184

[10] 李英乐，于洪涛，刘力雄. 基于 SVM 的微博转发规模预测方法[J]. 计算机应用研究，2013，30(9)：2594-2597

[11] 于晶，刘臣，单伟. 在线社会网络中信息传播的结构研究[J]. 情报科学，2013，31(12)：136-140

[12] 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

[13] 熊熙，胡勇. 基于社交网络的观点动力学研究[J]. 物理学报，2012，61(15)：104-110

[14] Xiong Fei, Liu Yun, Zhang Zhen-jiang, et al. An Information Diffusion Model Based on Retweeting Mechanism for Online Social Media[J]. Physics Letters A, 2012, 376(30): 2103-2108

[15] 李慧. 从文献信息老化到网络信息老化的研究分析[J]. 情报科学，2010，28(3)：384-388，394

[16] 龚思婷，孙建军. 网络信息生命力评价——基于网络信息的增长与老化模型[J]. 情报杂志，2012，31(5)：75-79

[17] 沈珂轶. 社会网络的社团发现与动态特性研究[D]. 上海：上海交通大学，2011

[18] Bakshy E, Hofman J M, Mason W A, et al. Everyone's an Influencer: quantifying Influence on Twitter[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 65-74

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

[20] 刘志明，刘鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程，2011，29(6)：8-16

[21] 肖宇，许炜，商召玺. 微博用户区域影响力识别算法及分析[J]. 计算机科学，2012，39(9)：38-42

[22] Weng J, Lim E P, Jiang J, et al. Twitterrank: Finding Topic-sensitive Influential Twitterers [C]// Proceedings of the Third Acm International Conference on Web Search and Data Mining. ACM, 2010: 261-270