社交网络突发事件传播速率模型研究

黄贤英，杨林枫*，刘小洋，何道兵，刘广峰，阳安志

(重庆理工大学计算机科学与工程学院 重庆 巴南区 400054)

【摘要】针对传统信息传播速率模型中对各因素描述不准确且仿真结果误差偏高的问题，该文提出了一种PPS信息传播速率计算模型。该模型选取Digg社交平台的数据集进行分析，拟合大量数据改进了传统模型中固有增长率及用户承载力的计算方法，并根据新闻投票量对新闻划分范围得到PPS模型，最后针对不同投票范围的新闻信息进行了仿真分析。仿真结果表明，在此平台上的新闻均经历增长期而到达稳定期，而通过传播速率分析图得出新闻在进入头版后增长速率最快。结合传统模型及算法进行准确率分析得到，该文提出的PPS模型在准确率上有了较大提升，证明提出的模型在分析社交平台的信息传播速率合理、有效。

关 键 词 固有增长率；传播速率；社交平台；用户承载力

中图分类号 TP393 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.03.024

Research on the Emergency Events Propagation Rate Model Based on Social Network

HUANG Xian-ying, YANG Lin-feng*, LIU Xiao-yang, HE Dao-bing, LIU Guang-feng, and YANG An-zhi

(College of Computer Science and Engineering, Chongqing University of Technology Banan Chongqing 400054)

Abstract According to the inaccurate description of various factors in the traditional information propagation rate model and the high error of simulation results, an information propagation rate model considering propagation speed (PPS) is proposed. First, the model selects the Digg social platform and analyzes the data set; Secondly, a large amount of data is used to improve the calculation method of intrinsic growth rate and user carrying capacity of the traditional model, then the PPS model is obtained according to the different news voting quantity, and finally the simulation analysis is carried out for different coverage stories. The simulation results show that the news on this platform has experienced the growth period and reached the stable period, and the speed of the news is the fastest after the front page is shown. Based on the traditional model and the algorithm, the proposed PPS model has a great improvement in accuracy, which proves that the model is reasonable and effective in analyzing the information transmission rate of the social platform.

Key words intrinsic growth rate; propagation rate; social platform; user carrying capacity

如今，社交网络成为信息传播和交流的重要平台，但谣言或欺诈信息泛滥的现象屡见不鲜。这类信息具有诱惑性，在传播过程中更容易引起人们的关注从而引发病毒式传播，为社会带来恶劣影响。因此，针对在线社交网络平台上突发信息传播的特征研究具有重要的现实意义。

近年来，国内外学者针对信息传播做了大量研究。对社交平台相关指标进行预测主要有3种方法：机器学习方法、改进现有的算法以及数学建模。

用机器学习算法可以进行信息传播分析或预测，优势是数据处理准确且算法成熟，但缺点是适用性低且误差较高。通过Twitter-LDA主题模型可以对社交网络的内容进行主题分析[1]，利用隐马尔可夫模型[2]可以预先识别社交网络中流行的虚假信息，主题分析及虚假消息识别都可以对研究内容精准把控。在对研究内容精准把控的基础上，基于用户属性、社交关系和社交平台内容3类综合特征，使用机器学习的分类方法[3]、引入多任务学习方法，以逻辑回归预测模型[4]可作为基准算法对用户转发行为进行预测。

通过改进现有的算法进行信息传播的研究，有很好的结合性且误差相对较低。如结合SIS和SIR两种经典传播模型[5-6]，可将信息传播看作一个易受感染的流行病过程，可以增加病毒传播模型的使用面。

收稿日期：2018 – 03 – 19；修回日期：2018 – 11 – 22

基金项目：重庆市教育委员会人文社会科学研究项目(17SKG144)；教育部人文社科青年基金(16YJC860010)；国家社会科学基金西部项目(17XXW004)

作者简介：黄贤英(1967 –)，女，教授，主要从事计算机应用方面的研究.

通信作者：杨林枫，E-mai: 376985081@qq.com

国内外学者在信息传播数学建模上也做了很多贡献。在模型构建过程中考虑了互动的4个属性[7]：熟悉度、主动性、相似性和可信度。其中基于相似性得到的扩散预测(information-dependent embedding based diffusion prediction, IEDP)模型[8]和DL方程[9]，可在时间和空间特征上观测到扩散过程中的用户映射；在此基础上结合其他属性对国外流行的新闻聚集社交平台进行了分析预测，如水动力学模型[10]和线性扩散模型[11]。信息传播预测[12]可以帮助管理者利用社交网络引导舆情，扼制谣言的爆炸性传播。

然而，大部分研究虽然与传播速率相关，但直接针对传播速率预测的研究却很少。本文将基于线性常微分方程，提出一种社交网络突发事件传播速率模型，该模型可针对不同时刻预测新闻信息的传播速率，研究结果可用于广告的精准投放及新闻信息的即时推送。

1　数据选取及分析

1.1　数据选取

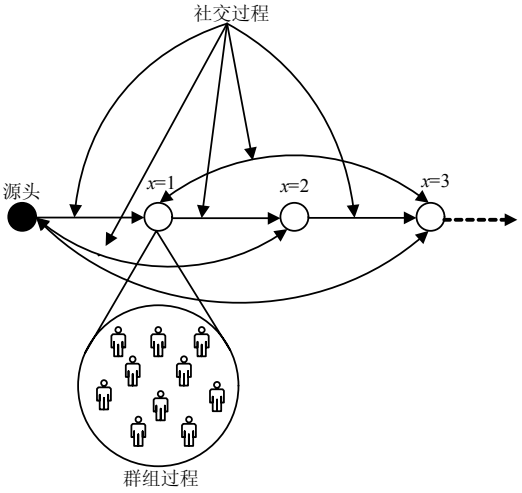


图1　Digg平台传播结构示意图

本文选取Digg社交网络平台进行分析。Digg是国外最流行的新闻聚集网站之一，平台用户可以将他们在专业新闻网站或博客中看到的新闻报道链接提交给Digg，其他用户可以对新闻进行投票和评论。新闻在Digg平台上的传播过程如图1所示。其中第一个把消息带到Digg网站的人被称为发起者或源头用户。当某新闻信息由源头发出后，有两种信息传播方式：1) 被源头用户的跟随者看到并进行投票，这类用户在投票后，其跟随者也能看到他投票的新闻并跟着投票，依此类推，从而形成一条信息传播链；2) 一旦新闻被提升到头版，无论直接或者间接的朋友，都能够查看新闻并进行投票。每个用户的投票数都是随机的，但大多数用户的投票数都集中在1～10票。因此，Digg平台数据的刷票率较低，真实性较高，可用于研究随机投票对信息传播过程的影响。

本文数据来自网络提供的Digg开源数据集，该数据集包含1 251条新闻信息，数据集中新闻信息均为Digg平台的突发新闻信息。总共收到了1 048 575次投票，涉及89 643名用户。

1.2　数据分析

为了观察一条突发新闻信息在发布后，受影响用户数量的变化情况，首先在数据集1 251个新闻信息中随机选取两个故事，对受影响的用户数量随时间变化进行绘图分析，如图2所示。

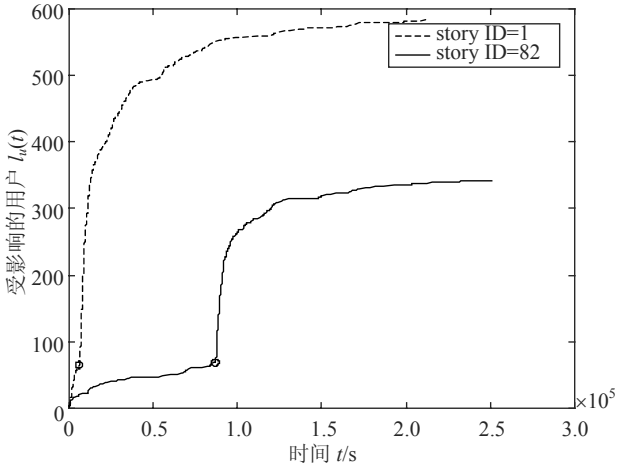


图2　新闻信息传播示意图

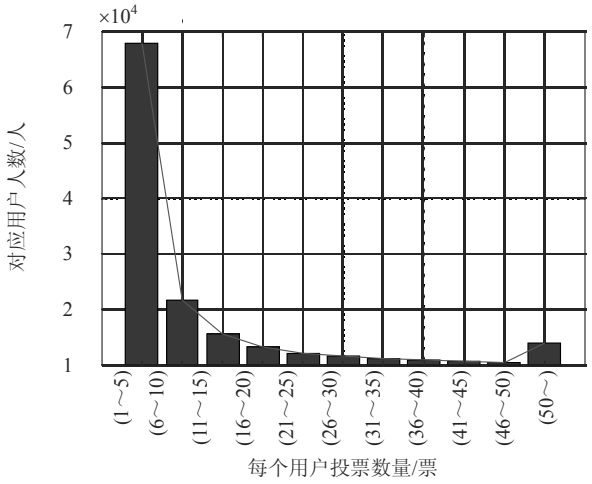


图3　用户投票数统计图

由图2可以看出story1在发出后得到用户的迅速响应并投票，受影响用户的数量急剧增长，最后达到一个稳定增长的阶段。仔细观察能在增长的阶段发现一个不太明显的拐点，这个点是该篇新闻被提到了头版的象征；而story82则是先平稳增长，然后迅速增长，最终达到平稳的一个过程，这个新闻中的拐点比较明显，说明该篇新闻进入头版之前等待

的时间比较久，在进入头版之后，大量用户为其投票，受影响用户数量开始增长迅速。

下面将对数据集中所涉及的89 643名用户所投的1 048 575票进行统计分析，以此来判断Digg平台用户投票的真实性和有效性，如图3所示。

可以看出，整个趋势呈递减趋势，且投票在1～5票的用户最多。有84%的用户对不同新闻的总投票数在1～15票之间。

2 传播速率PPS模型

针对上述分析，本文提出基于社交网络Digg平台的传播速率(PPS)模型，用于分析Digg平台上新闻信息的传播速率与受影响用户增长的关系。

模型中用 I_u 表示每条新闻中受影响的总用户数量；$I_u(t)$ 表示随时间变化的用户数量；$\mathrm{d}I_u(t)/\mathrm{d}t$ 表示在 t 时刻受影响用户的增长率，也就是传播速率。而与 $\mathrm{d}I_u(t)/\mathrm{d}t$ 相关的两个因子分别为受影响用户的固有增长率 gr 和用户的承载力 $K(I_u)$。受影响用户的固有增长率 gr 代表一条新闻发布之后，在不受外界因素的影响下，为其投票用户的增长率；$K(I_u)$ 表示新闻信息受不同程度外界因素影响的承载力。本文在训练模型时，随机选取数据集中的80%(1 000条新闻)作为训练集，20%(251条新闻)作为测试集。

关于受影响的用户的固有增长率，随着时间的迁移，受影响的用户的实时数量是逐渐递减的。根据数据集中1 000条新闻的时间t和投票数进行拟合得到固有增长率 gr 的变化为：

$$\mathrm{gr}(t)=1.5\mathrm{e}^{-1.5(t-1)}+\omega \tag{1}$$

式中，ω 为调控参数，根据总投票量的不同，调控参数会取到不同的值，分析大量数据得到二者之间的关系是总投票量越大，调控参数越小。

而关于用户的承载力 $K(I_u)$，这个参数会随着受影响用户数的变化而变化。在一条新闻信息进入头版之后，其承载力会变大；而当受影响的用户数量增大时，承载力也会变大。本文制定用户承载力的变化方式如下：最初承载力为 K_1，新闻进入头版后的承载力为 K_2，最后达到稳定状态后的承载力为 K_3。各个状态的承载力的转换与受影响用户 I_u 的变化有关，具体转换过程如图4所示。

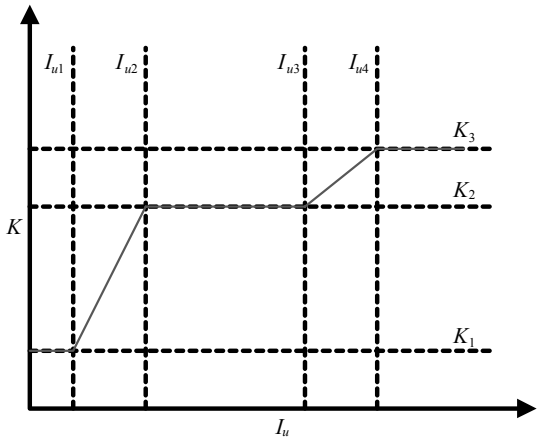


图4 承载力系数变化关系

由此得到 K 的函数拟合式为：

$$K(I_u)=\begin{cases}K_1 & I_u\leqslant I_{u1}\\ \dfrac{K_2-K_1}{I_{u2}-I_{u1}}(I_u-I_{u1})+K_1 & I_{u1}<I_u\leqslant I_{u2}\\ K_2 & I_{u2}<I_u\leqslant I_{u3}\\ \dfrac{K_3-K_2}{I_{u4}-I_{u3}}(I_u-I_{u3})+K_2 & I_{u3}<I_u\leqslant I_{u4}\\ K_3 & I_u>I_{u4}\end{cases} \tag{2}$$

对 $K(I_u)$ 值的影响主要是进入头版时受影响的用户的数量，以及进入平稳状态时受影响的用户数量。经过对1 000条新闻信息的投票数进行分类，并统计各种状态下的受影响的用户边界值来得到 $K(I_u)$ 的各个指标取值情况。上式分段函数中，每个阶段 I_u 的取值情况如表1所示。

表1中 I_{u1} 为新闻没有进入头版时期的取值，此时新闻受影响的用户呈现自增长状态，存在较多不定因素，根据1 000条新的总投票数统计得到 I_{u1} 及其他因子的取值。

表1 用户承载力参数范围

受影响用户数量(票数)	I_{u1}	I_{u2}	I_{u3}	I_{u4}	K_1	K_2	K_3
$0<I_u\leqslant 1\,000$	40～90	25% I_u	55% I_u	I_u	40～90	60% I_u	I_u
$1\,000<I_u\leqslant 2\,000$	90～100	25% I_u	50% I_u	I_u	90～100	62% I_u	I_u
$2\,000<I_u\leqslant 3\,000$	100～110	25% I_u	48% I_u	I_u	100～110	64% I_u	I_u
$3\,000<I_u\leqslant 4\,000$	100～110	25% I_u	46% I_u	I_u	100～110	62% I_u	I_u
$4\,000<I_u\leqslant 5\,000$	110～150	25% I_u	46% I_u	I_u	110～150	62% I_u	I_u
$5\,000<I_u$	150～	25% I_u	$\leqslant$ 46% I_u	I_u	150-	$\leqslant$ 62% I_u	I_u

对受影响用户在 t 时刻的值 $I_u(t)$、受影响用户的固有增长率 gr、用户的承载力 $K(I_u)$ 进行拟合，最后定义本文PPS模型的传播速率为：

$$V_{\text{model}} = \text{gr} \times I_u(t) \times \left(1 - \frac{I_u(t)}{K(I_u)}\right) \tag{3}$$

式中，V_{model} 在理论意义上的计算方法为：

$$V_{\text{model}'} = \frac{\mathrm{d}I_u(t)}{\mathrm{d}t} \tag{4}$$

根据式(3)可预测一条突发新闻信息受影响用户增长的速率。预测结束后，需要与真实数据的传播速率进行比对。真实数据中传播速率为：

$$V_{\text{real}} = \frac{\mathrm{d}I_u(t)}{\mathrm{d}t} = \frac{I_u(t+\Delta t) - I_u(t-\Delta t)}{2\Delta t} \tag{5}$$

在计算真实数据的传播速率时，由于时间在快速增长期过于密集，故对整条数据进行一次筛选，筛选规则是取到第一个数据点后，在时间间隔大于500 s时再取第2个点，依此类推。

相比真实传播速率 V_{real}，模型传播速率 V_{model} 的准确率 δ 为：

$$\delta = 1 - \frac{\left|V_{\text{model}} - V_{\text{real}}\right|}{V_{\text{real}}} \tag{6}$$

$$\overline{\delta} = \frac{\delta_1 + \delta_2 + \cdots + \delta_n}{n} \tag{7}$$

3　仿真分析

3.1　模型分析

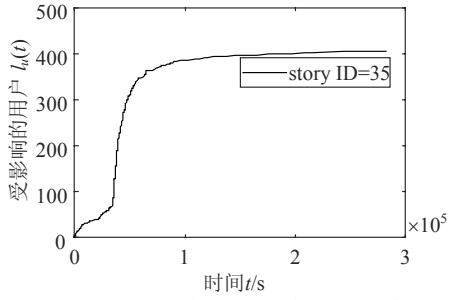
根据训练集中1 000条新闻训练得到PPS模型。下面在测试集中选取实验数据，首先在251个新闻信息中，对票数范围进行从小到大排序并统计范围内的新闻篇数，如表2所示。

表2　不同票数范围的新闻数统计表

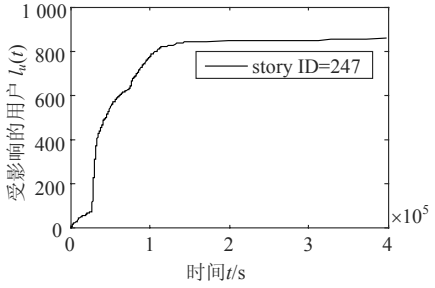
票数范围	0～1 000	1 001～2 000	2 001～3 000	3 001～4 000	4 001～
新闻篇数	188	33	19	7	4

由上表可得投票数在1 000票以下的新闻数量最多，超过4 000票的新闻篇数不到2%。在仿真实验中，根据表2每个范围新闻数量的比例值，从0～1 000票范围内随机选择两篇新闻进行仿真分析，余下4个票数范围各随机选择1篇新闻进行分析，仿真分析结果如下。

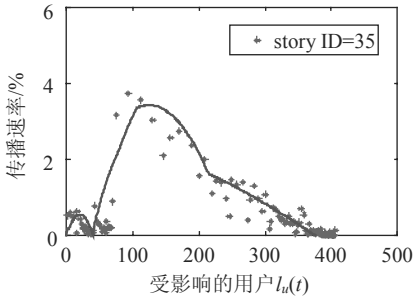
如图5所示，图5a和5b为所选的两个新闻信息的传播示意图。可以看出，其中一条新闻在发布之后缓慢上升，进入头版后开始激增，最后达到平稳状态。图5c和5d为对应新闻信息的传播速率随受影响用户增加的变化趋势，点代表真实数据的计算结果，曲线代表模型的仿真结果。可得新闻进入头版前的速率变化，进入头版之后的速率急速上升，到很少用户关注该新闻最后退出头版，致其速率缓慢下降到达最小值的过程。模型预测曲线与真实数据点的趋势图比对基本一致。



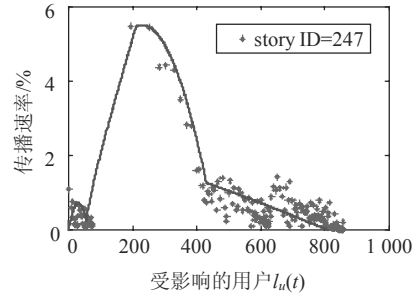
a. ID为35的新闻信息随时间变化的传播趋势



b. ID为247的新闻信息随时间变化的传播趋势



c. ID为35的新闻信息传播速率随受影响用户增加的变化趋势

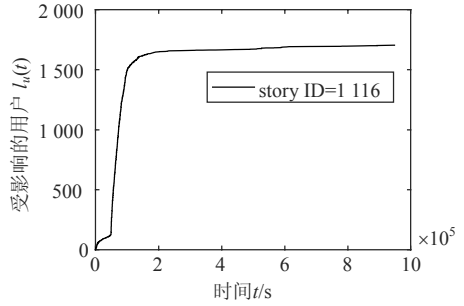


d. ID为247的新闻信息传播速率随受影响用户增加的变化趋势

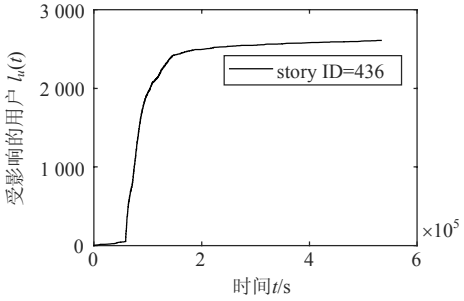
图5　票数在0～1 000的传播分析图

图6a和图6b分别为票数在1 001～2 000和2 001～3 000中的随机一条新闻，其展现的是随时间

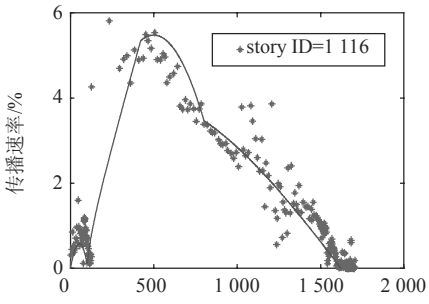
增加受影响的用户变化趋势图，可以看出同样存在进入头版前的增长期，激增期以及最终的稳定期。图6c和6d是受影响的用户与对应的传播速率的指数图，图中的点代表真实数据计算的传播速率，而曲线代表PPS模型拟合，两者有明显重合区域，但是图6d的准确率较图6c有明显下降，原因是数据集票数多的新闻数量少，导致训练出的模型在拟合票数高的新闻时，准确率会偏低。



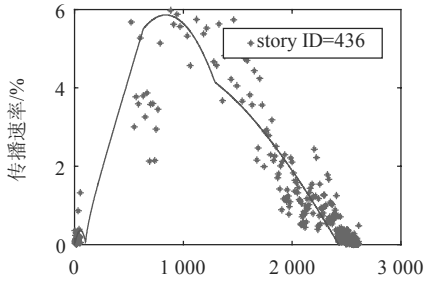
a. ID为1 116的新闻信息随时间变化的传播趋势



b. ID为436的新闻信息随时间变化的传播趋势



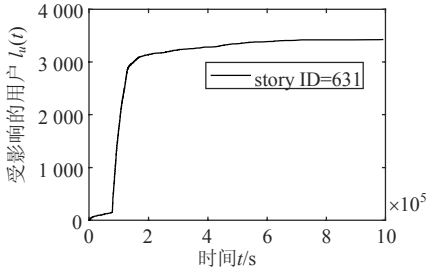
c. ID为1 116的新闻信息传播速率随受影响用户增加的变化趋势



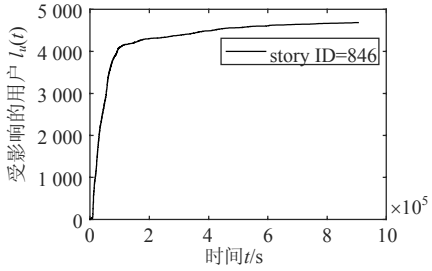
d. ID为436的新闻信息传播速率随受影响用户增加的变化趋势

图6 票数在1 001～3 000的传播分析图

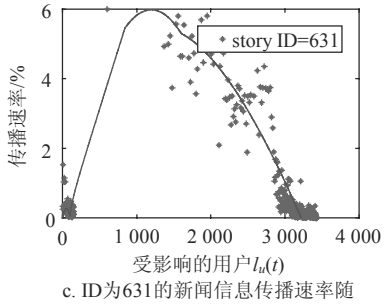
图7为两个票数较高的新闻信息，图7a中的新闻信息在经历了很短时间的缓慢增长后就进入了头版达到了激增期，最后到达了平稳期。图7c和7d两图中部的点分散度过大，是因为新闻过度热门，投票的用户实时变化散度大，而图中右边的点很稠密，是因为激增期时间短，且投票量多，可近似看作重尾分布。从这两个新闻明显看出真实数据与模型曲线较之前的数据存在较大误差，原因是数据集票数多的新闻数量少，训练出的模型在拟合此类新闻时准确率偏低。



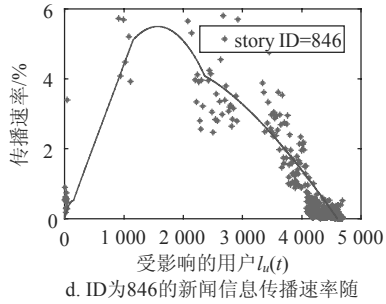
a. ID为631的新闻信息随时间变化的传播趋势



b. ID为846的新闻信息随时间变化的传播趋势



c. ID为631的新闻信息传播速率随受影响用户增加的变化趋势



d. ID为846的新闻信息传播速率随受影响用户增加的变化趋势

图7 票数在大于3 000的传播分析图

3.2 准确率分析

针对实验的6组数据，分别计算本文模型与真实

数据的准确率，及其他模型、算法与真实数据的准确率，并将计算得到的准确率进行对比，结果如图8所示。

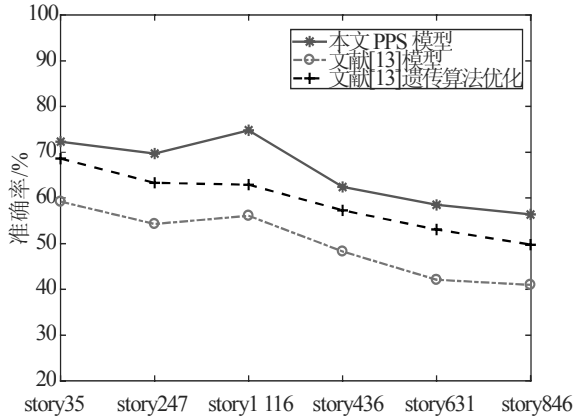


图8　各模型算法间准确率对比图

由上图可以看出，本文基于数据分析，对固有增长率 gr 以及用户承载力 $K(I_u)$ 的计算方法进行改进后，得到的模型结果的准确率较传统方法有明显提高。

对随机抽取的6组实验数据进行仿真分析后，本文对数据集中1 251条新闻传播速率进行准确率分析。由于数据量过大，这里针对每条新闻的时间，设置新的时间间隔筛选规则为：取到第一个数据点后，在时间间隔大于2 000 s时再取第二个点，以此类推。对计算1 251条突发新闻，分别计算本文模型与真实数据的准确率及文献[13]模型与真实数据的准确率。对比结果如图9所示。

数据集中1 251条新闻预测后进行误差分析得到的传播速率平均准确率为67.28%。其中在1 140条新闻预测中，本文PPS模型优于传统模型，占数据集的91%，说明本文提出的模型能有效预测新闻传播速率。

本文PPS模型预测在构建过程中，其构建思想、构建步骤均适用于其他社交平台。但是由于社交平台都有极强的个性化属性，不同的社交平台中数据字段都不完全相同，因此本文构建的模型仅适用于Digg社交平台。

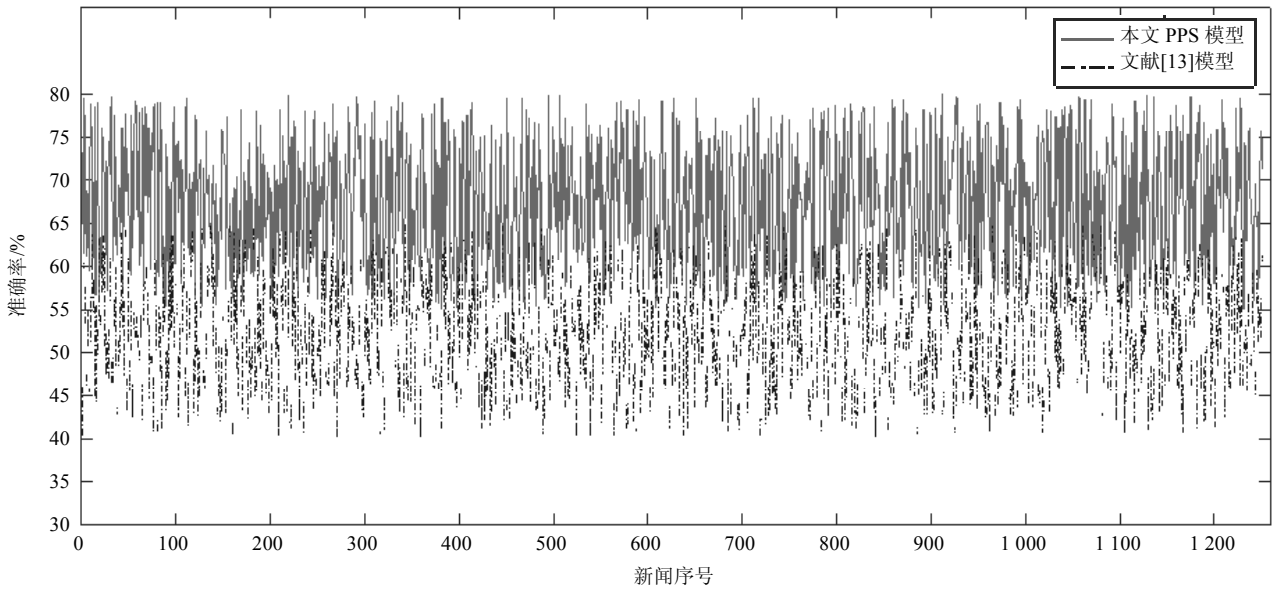


图9　数据集所有新闻两种模型准确率对比图

4　结 束 语

本文利用了Digg平台上的数据，对固有增长率及用户承载力两个指标提出了新的计算方法，并得到了不同投票范围的取值方式，构建出PPS信息传播速率模型。最后在不同票数范围内选取6个突发新闻事件信息进行仿真分析，根据仿真图可以看出模型预测与真实传播速率有较高的重合性。对计算得到的准确率分析后，发现本文模型较传统模型算法都有了比较大的提高，证明本文的模型有效。但由于社交平台的随机性，在分析一些特殊的新闻信息时，会出现误差较大的情况。下一步的研究工作将通过分析不同用户的自身意愿，来判断他们是否会在预计时间内进行投票，用以分析更加准确的传播速率。

本文还得到重庆市高校网络舆情与思想动态研究咨政中心项目(KFJJ2017024)的资助，在此表示感谢。

参 考 文 献

[1] 罗春海, 刘红丽, 胡海波. 微博网络中用户主题兴趣相关性及主题信息扩散研究[J]. 电子科技大学学报, 2017, 46(2): 458-468.
LUO Chun-hai, LIU Hong-li, HU Hai-bo. Research on correlation of users' topic interests and topic information diffusion in microblog networks[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(2): 458-468.

[2] 谢柏林, 蒋盛益, 周咏梅, 等. 基于把关人行为的微博虚假信息及早检测方法[J]. 计算机学报, 2016, 39(4): 730-744.
XIE Bo-lin, JIANG Sheng-yi, ZHOU Yong-mei, et al. Misinformation detection based on gatekeepers' behaviors in microblog[J]. Chinese Journal of Computers, 2016, 39(4): 730-744.

[3] 曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779-790.
CAO Jiu-xin, WU Jiang-lin, SHI Wei, et al. Sina microblog information diffusion analysis and prediction[J]. Chinese Journal of Computers, 2014, 37(4): 779-790.

[4] 唐兴, 权义宁, 宋建锋, 等. 微博个性化转发行为预测新算法[J]. 西安电子科技大学学报, 2016, 43(4): 51-56, 62.
TANG Xing, QUAN Yi-ning, SONG Jian-feng, et al. Novel algorithm for predicting personalized retweet behavior[J]. Journal of Xidian University, 2016, 43(4): 51-56, 62.

[5] 王伟, 舒盼盼, 唐明, 等. 网络传播动力学模拟方法评述[J]. 电子科技大学学报, 2016, 45(2): 288-294.
WANG Wei, SHU Pan-pan, TANG Ming, et al. Simulation methods for spreading dynamics on networks: A recitation[J] Journal of University of Electronic Science and Technology of China, 2016, 45(2): 288-294.

[6] KANDHWAY K, KURI J. Using node centrality and optimal control to maximize information diffusion in social networks[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2016, 47(7): 1099-1110.

[7] WANG D, MUSAEV A, PU C. Information diffusion analysis of rumor dynamics over a social-interaction based model[C]/International Conference on Collaboration and Internet Computing. Pittsburgh, PA, USA: IEEE, 2017: 312-320.

[8] GAO S, PANG H, GALLINARI P, et al. A novel embedding method for information diffusion prediction in social network big data[J]. IEEE Transactions on Industrial Informatics, 2017(99): 2097-2105.

[9] WANG F, WANG H, XU K. Diffusive logistic model towards predicting information diffusion in online social networks[C]//International Conference on Distributed Computing Systems Workshops. Macau, China: IEEE, 2011: 133-139.

[10] HU Y, SONG J, CHEN M. Modeling for information diffusion in online social networks via hydrodynamics[J]. IEEE Access, 2017(99): 1.

[11] 彭川, 李元香, 莫海芳. 在线社会网络中的多源信息扩散问题研究[J]. 计算机应用研究, 2015, 32(10): 2947-2950.
PENG Chuan, LI Yuan-xiang, MO Hai-fang. Research on multiple sources information diffusion in online social networks[J]. Application Research of Computers, 2015, 32(10): 2947-2950.

[12] ZHANG X, CHEN X, CHEN Y, et al. Event detection and popularity prediction in microblogging[J]. Neurocomputing, 2015, 149(8): 1469-1480.

[13] DAVOUDI A, CHATTERJEE M. Prediction of information diffusion in social networks using dynamic carrying capacity[C]//2016 IEEE International Conference on Big Data. Washington, DC, USA: IEEE, 2016: 2466-2469.

编 辑 蒋 晓