TSL：基于连接强度的 Facebook 消息流行度预测模型

王晓萌[1]，方滨兴[1,2]，张宏莉[1]，王星[1]
（1. 哈尔滨工业大学计算机网络与信息安全技术研究中心，黑龙江 哈尔滨 150001；
2. 广州大学网络空间先进技术研究院，广东 广州 510006）

摘 要：在线社交网络的迅速发展使信息呈现爆炸式增长，然而不同消息的流行度存在较大差异，对其准确预测一直是领域内的研究难点。流行度预测的任务是根据消息传播早期过程中涌现的特征预测其未来的传播趋势，现有基于传播网络特征与拟合函数的预测模型难以解决预测准确率低的问题，因此借助社会学中的弱连接理论，引入连接强度的概念，并融合消息传播早期的流行度构建多元线性回归方程，提出了一种针对 Facebook 知名主页的消息流行度的预测模型 TSL。通过在 Facebook 真实数据集（含 154 万次转发）上与其他具有代表性的基准模型进行比较，实验表明 TSL 模型可以对消息的最终转发流行度进行有效预测，预测性能优于同类方法。

关键词：在线社交网络；弱连接；流行度；信息传播
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019207

TSL: predicting popularity of Facebook content based on tie strength

WANG Xiaomeng[1], FANG Binxing[1,2], ZHANG Hongli [1], WANG Xing [1]
1. Research Center of Computer Network and Information Security Technology, Harbin Institute of Technology, Harbin 150001, China
2. Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China

Abstract: The rapid development of online social networks leads to an explosion of information, however, there are great differences in the popularity of different messages, and accurate prediction is always a great difficulty is the current study. Popularity prediction of online content aims to predict the popularity in the future based on its early diffusion status. Existing models for popularity prediction were mostly based on discovering network features or fitting the equation into a varying time function that the accuracy of current popularity prediction model was not high enough. Therefore, with the help of the weak ties theory in sociology, the concept of tie strength was introduced and a multilinear regression equation was constructed combined with the early popularity. A TSL model to predict the popularity of Facebook's well-known pages was proposed. The main contribution of this article was to solve the problem and few or no work based on sociology. A high linear correlation between the proportion of faithful fans was existed in Facebook homepage with frequent shares in the early and the future popularity. Compared with other baseline models, an experimental study of Facebook (including 1.54 million shares) illustrates the effectiveness of the proposed TSL model, and the performance is better than the existing similar methods.

Key words: online social networks, weak ties, popularity, information diffusion

1 引言

随着互联网技术的不断发展，尤其是 Web2.0 技术出现之后，在线社交网络应用逐渐涌现并迅猛发展，使人类使用互联网的方式产生了深刻变革。借助在线社交网络发布和接收信息的简便性，社交网络用户社区化、意见领袖化加速了新内容的创造与传播，各类话题和观点可以随时发布并爆炸式传

收稿日期：2019−06−21；**修回日期**：2019−08−30
基金项目：国家重点研发计划基金资助项目（No.2017YFB0803305, No.2016QY03D0501）
Foundation Item: The National Key R&D Program of China (No.2017YFB0803305, No.2016QY03D0501)

播扩散。根据欧盟社会计算报告，有别于以发布消息为主的 Twitter、微博和以共享空间为主的 YouTube 视频网站，Facebook 是一种典型的提供在线交友服务的社交网络[1]。以 Facebook 为代表的在线社交网络逐渐成为当代社会信息传播的重要集散地，其信息活跃性达到了前所未有的程度。很多名人或者组织都已在 Facebook 上开设公共主页，通过频繁地发布实时的动态消息来吸引公众注意力。例如，新闻主页抢发最新头条、电影主页网罗粉丝等。因此，针对 Facebook 的热门主页开展消息的流行度预测研究，如在线内容的转发量、视频的点击数量和在线新闻的评论数量等，对媒体广告投放效果评估、话题传播趋势预测[2]、电影票房收益评估[3]和选举预测[4]等问题都有着重要应用价值。

流行度预测问题本质上源于少数样本获得多数关注的统计分布上的不均匀现象，如财富分布、国家人口分布、交友网站的朋友数分布等。已有研究表明，大部分的网络内容只有很少的人关注，而少数内容却获得了大量的用户关注。针对这种非均匀分布现象的研究最早可以上溯至意大利经济学者维尔弗雷多·帕累托（Vilfredo Pareto）在研究 19 世纪英国人的财富和收益模式时，提出的著名的“二八定律”，即 20%的人口占据了 80%的社会财富。在互联网时代，Albert-laszlo 等[5]在《自然》杂志上发表的文章中指出，电影演员合作网、万维网、美国西部电力网等复杂网络的度分布符合幂律指数满足 $2<\gamma<3$ 的幂率分布。在线社交网络时代，Kwak 等[6]发现 YouTube 上 10%的最受欢迎的用户发布内容吸引了近 80%的用户关注，然而剩余的 90%内容至多获得了 20%的用户关注。在线社会网络中，流行度预测的主要任务是预测用户生成内容的流行度，该问题的一般定义为根据对用户生成内容发布后初期传播过程的观测，预测该内容在未来某个时间点的流行度值。

针对 Facebook 的消息流行度预测具有较高研究价值，但是也存在很多难点。首先，数据规模庞大。惠普实验室的 Golder 等[7]发现 Facebook 的好友数（度值）的中值为 144，均值为 179.53。截止到 2015 年 9 月，Facebook 拥有约 10 亿用户，包括社科、名人、政府机构等知名主页以及众多普通用户主页，其用户规模与繁荣程度已经在某种程度上可以理解为人类社会在网络社会的映射；其次，信息传播与演化受多种因素控制，如关系结构、网络群体和信息内容等，其中很多因素由于测量手段以及隐私保护等因素的限制而难以获取；最后，多种干扰因素导致传播过程具有随机性，不同的信息间也存在着竞争与抢占关系。

虽然实时预测一个消息的流行度演化很难，但是基于信息发布后一段时间内的传播表现来预测最终流行度是可行的。其中最著名的就是 Szabo 等[8]于 2008 年提出的 SH（Szabo and Huberman）模型，他们发现文章评分网站 Digg 上的新闻讨论帖、视频分享网站 YouTube 上的视频的早期转发量和最终转发量在进行取对数后存在一定程度的线性相关，并首先提出了基于线性回归（linear regression）的流行度预测模型。Facebook 作为典型的在线交友网络，其消息传播速度介于 Digg 和 YouTube 之间，传播机制也更接近人类社会网络，因此有必要针对 Facebook 的消息流行度预测问题展开研究。现在方法很少从社会学角度研究信息传播的机理，并且对转发过程中潜在用户的特征挖掘不充分。然而已有研究表明，社交网络中的弱连接可以增益信息传播，它们对信息传播的深度和广度起到非常重要的作用[9]，通过对一些 Facebook 知名主页的消息流行度演化趋势进行分析，本文发现那些在传播早期就能聚集较多弱连接用户参与转发的消息，其最终流行度都很高，且消息的最终转发流行度与早期传播过程中的忠实粉丝所占比例在双对数坐标系中存在线性相关。

基于以上发现，本文提出了一种面向在线交友网络的流行度预测模型。为了更好地量化信息传播过程中的弱连接用户的参与程度，本文提出了连接强度的概念，并根据消息传播早期的流行度与连接强度构建多元线性回归方程，然后通过用户活跃度对预测方程进行修正，最终得到基于连接强度的预测模型（TSL, tie strength based linear）。本文将提出的模型与一些代表性的基准方法如 SH、DSH（depth based SH）和 RPP（reinforced Poisson process）进行比较，实验验证了所提模型对 Facebook 主页消息的最终流行度预测效果较好。本文贡献介绍如下。

1) 将“弱连接理论”引入流行度预测问题，并发现连接强度这一新的流行度预测特征。

2) 提出了基于早期流行度和连接强度的线性回归模型 TSL。

3) 引入多个基准模型，针对 Facebook 主页消息的流行度预测效果进行对比分析。

2　相关工作

近年来，流行度预测问题受到了越来越多研究者的关注，并涌现出了许多模型与方法，主要可以分为基于群体状态的方法、基于回归/分类的方法和基于时间序列的方法。

基于群体状态的方法是将社交网络中的节点分成几种状态，通过模拟群体状态转移过程，建立信息传播模型来分析流行度演化趋势，主要包括传染病模型、级联传播模型等。在传染病模型中，系统中的个体一般被分为几类，每一类个体都处于同一种状态。基本状态包括：易感状态 S（susceptible），即健康的状态，但有可能被感染；感染状态 I（infected），即染病的状态，具有传染性；移除状态 R（recovered），即感染后被治愈并获得了免疫力或感染后死亡的状态。Abdulah 等[10]利用传染病模型对 Twitter 消息的传播进行了研究，他们认为在社交网络中处于感染状态（I 类）的节点发布相关消息，则其粉丝成为新的易感者，总的人数不断增大。Matsubara 等[11]发现博客流行度服从幂率分布，且用户关注呈现周期性变化，在传统 SI 模型基础上提出了一种动态感染率的流行度预测模型。Li 等[12]考虑网络底层拓扑特征对传播的影响，针对人人网的外源性视频流行度预测问题，提出了一种基于网络级联流行度预测方法。

基于回归/分类的方法通过发现信息传播过程中的关键影响因素，并探寻这些因素与消息流行度之间的关系，从而将流行度预测转化为分类或回归问题进行求解。这类方法关注的重点在于提取对于分类或回归有效的特征，能对未来流行度的数值给出一个具体的预测，例如 Szabo 等[8]发现早期的某个特定时间的流行度与传播晚期的流行度都取对数之后有强线性关系，并率先用回归方法预测最终流行度。Chang 等[13]发现视频网站的电视剧单集流行度与历史发布过的剧集的流行度存在相关性，其收视群体中的随机观看者随着时间推移越来越少，并基于以上发现提出了一种改进的回归模型。Bao 等[14]发现早期传播网络的密度和消息转发深度与最终流行度存在线性相关，并基于这 2 个特征提出了一种改进的 SH 模型。Kim 等[15]发现博文早期浏览量与最终浏览量有关，提出了一种基于指数函数的回归模型。Cheng 等[16]从时间角度分析了在线社交网络的热点话题传播规律，提出了一种自回归移动平均模型预测回帖数量。朱海龙等[17]提出了一种基于传播加速度的微博流行度预测方法，该方法首先提出传播加速度概念，并结合早期流行度建立多元回归模型对微博转发数量进行预测。

基于时间序列的方法是假设消息的转发过程在时间维度上具有延续性，利用观测所得的历史不同时间点上的数值序列进行建模并预测未来变化趋势。Crane 等[18]通过分析 Youtube 网站的 500 万段视频的传播过程，发现绝大部分（约 90%）视频的传播过程可以用泊松过程进行精确刻画，剩余视频的传播过程在经历流行度的峰值之后其单位时间内增加的流行度服从幂律分布。Yang 等[19]研究了用户生成内容流行度随时间的变化模式。该研究通过对 5.8 亿条推文和 1.7 亿篇博客文章流行度随时间消涨模式的聚类分析，挖掘出 6 类形态各异的流行度时序模式。Lerman 等[20]在 Digg 网的消息投票模型中考虑了消息的兴趣度和可见度，并利用所得模型进行消息最终获得票数的预测。Gao 等[21]提出了一种基于动态泊松过程的改进方法，该方法建模了信息传播过程中新颖性随着时间的衰减过程以及优先连接机制。

虽然上述方法已在流行度预测问题上取得了一些进展，但是针对 Facebook 这种超大规模在线社交网络的预测效果仍然有待提高。造成这种现象的主要原因是 Facebook 用户群体庞大，消息的转发迅速，传播机理更为复杂。基于群体状态的方法从微观角度利用数学模型推演信息传播的过程，但模型中的节点属性与状态转移概率过于理想化，仅适用于在网络拓扑已知条件下进行粗粒度的传播范围估计。基于时间序列的方法的本质是利用拟合函数刻画实时流行度演化趋势，这类方法针对短期预测有较好的效果，但是随着预测时间的增加，误差积累导致预测精度逐渐降低。基于回归/分类的方法旨在建立信息传播早期流行度与未来流行度的映射关系，需要对流行度演化数据进行特征提取，适用于长期预测。本文针对 Facebook 的信息传播机制进行深度分析，提出了一种基于回归分析的流行度预测模型，该模型首先根据社会学中的“弱连接理论”以连接强度的形式作为关键特征引入回归方程，同时结合早期流行度对消息最终流行度进行预测，实验表明该方法可以有效地提升预测性能。

3　问题定义

本文的研究对象为 Facebook 主页的用户生成

消息（user generated content）。用户可对这些消息进行评论、点赞以及转发操作。相比于评论数和点赞数，消息的转发数量可以更显著地反映信息的传播能力，因此本文将采用消息的转发数来刻画 Facebook 信息传播的流行度。

对于任意 Facebook 开放主页上用户发布的消息，人们可以确定其发布时间以及截止观测时的所有转发者 ID。对于给定消息 m，定义其发布时间为 T_0，预测时间为 T_{predict}，参考时间为 $T_{\text{reference}}$。流行度预测示意如图 1 所示，其中参考时间为预测任务采集早期信息传播情况所需要的时间长度，这段时间的信息传播特征被用于模型训练。预测时间是从消息发布时间 T_0 开始直至预测任务所设定的目标时间，消息的转发流行度随着目标时间的增长而不断增加，当时间超过消息生命周期后流行度近似保持不变，一般可以认为 $T_0<T_{\text{reference}}<T_{\text{predict}}$。进一步地，本文将消息 m 接收到第 i 次用户转发的时间用 t_i 表示，截至 $T_{\text{reference}}$ 时刻的转发过程可以记为 $\{t_k^m\}$，其中 $k\in(0,\ n_m)$，n_m 为全部训练时间段 $[0,\ T_{\text{reference}}]$ 内消息 m 获得的转发数，将 B_m 记为消息 m 在参考时间 $T_{\text{reference}}$ 的实际转发数，则 B'_m 为消息 m 在 T_{predict} 时刻的分享数预测值。

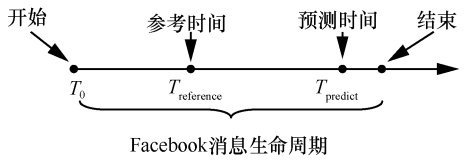


图 1 流行度预测示意

综上所述，流行度预测问题定义如下：已知消息 m 从发布时间 T_0 到参考时间 $T_{\text{reference}}$ 的转发数的累积过程 $\{t_k^m\}$，估计消息 m 从发布时间 T_0 到预测时间 T_{predict} 所取得的转发数 B'_m。

4 流行度预测

4.1 数据集

本文通过模拟用户以及页面解析的方式爬取了部分 Facebook 主页数据进行实验分析，随机选取了一些 Facebook 中排名前 100 的最热门主页作为数据抓取对象，包含名人主页、新闻主页以及娱乐主页等，基于这些主页抓取了 2016 年 1 月 1 日—12 月 31 日这些主页的所有历史发布信息共 3 775 条，并将这些消息送入爬取列表，采集转发过这些消息的用户 ID，总计得到消息的 154 万次转发。

Facebook 数据采集详细情况如表 1 所示，本文将已抓取主页分为 2 类，A 兴趣类（国家地理 Geographic、福克斯新闻 Fox News 等）；B 娱乐类（哈利波特 Harry Potter、电影明星威尔史密斯 Will Smith 等）。

表 1 Facebook 主页采集信息

主页	类别	消息总数/条	转发数/次	点赞数/次
Geographic	A	256	177 130	469 498
History	A	474	76 022	168 570
NBA	A	714	330 728	849 734
Call of Duty	B	294	38 357	122 610
Grey's Anatomy	B	360	64 571	288 365
Fox News	A	562	215 803	570 201
The Simpsons	B	395	157 205	390 057
Will Smith	B	151	144 793	451 108
Barack Obama	A	248	211 807	498 824
Harry Potter	B	321	120 824	429 731

4.2 Facebook 消息的生命周期

社交网络的信息传播存在特定的生命周期，所以预测任务的首要问题是选取合适的时间粒度与时间窗口。一方面，本文需要在消息生命周期未知的条件下，提前设置预测时间 T_{predict} 的取值范围，而且预测任务的目的是估计消息最终的转发量，所以基于完整性的考虑，预测时间应涵盖绝大部分转发过程，这样才能得到较为真实传播情况。另一方面，消息发布早期往往会获得更多的关注与传播，参考时间 $T_{\text{reference}}$ 设置越大则转发量累积越多，更容易估计最终转发流行度，而 $T_{\text{reference}}$ 设置越小则预测难度越大。基于实时性的考虑，参考时间则应尽可能地缩短以提升预测模型的响应速度。例如文章评分网站 Digg 上的推送新闻的生命周期较短，往往只需要一天时间就可以达到 80%的最终总评论量[8]，而视频分享网站 YouTube 上的内容生命周期较长，平均 7 天内的用户转发量只占最终转发量的 50%。本文首先分析了 Facebook 消息流行度的时间特征。Facebook 消息的生命周期如图 2 所示，其中纵坐标表示一条消息在每小时内所获得的平均转发量，横坐标表示距离消息发布时刻的时间长度。从图 2 可以看出，消息发布后在前几小时内流行度较高，但会在前 24 h 内迅速衰减，在 150 h 之后每小时增量衰减为 0，因此本文将预测时间 T_{predict}

设置为 7 天。此外，用户的转发行为在消息发布后的前 12 h 最为集中，因此基于实时性的考虑将 $T_{\text{reference}}$ 设置为 3 h。

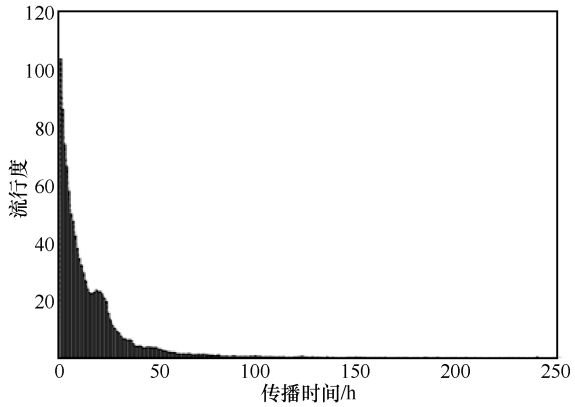


图 2　Facebook 消息的生命周期

4.3　用户活跃度（user activity）

很多在线社交网站的用户活动都具有周期性规律，Facebook 用户的转发行为也具有周期性。以一天为例，本文统计了数据集中所有主页每小时的消息平均转发量，Facebook 用户活跃度的变化趋势如图 3 所示，其中横坐标表示一天中的第几小时，纵坐标表示该小时一条消息所获得的平均转发量。从图 3 可以看出，用户的活跃度在不同时间段内存在显著差异，每小时用户转发数越多，说明用户在该时段活跃度越高。在凌晨 4 时至中午 12 时这段时间的用户活跃度最低，而 18 时至 22 时为转发最频繁时间段，符合用户的使用习惯与作息规律。此外，这种周期性差异可能会导致一个在冷门时间段发布的热点消息并没有在当下时刻引起足够多的关注，但是会在热门时间段内得到更多的转发，因此有必要在信息传播早期将所有消息的传播能力进行统一比较。本文引入了相对活跃度的概念，相对活跃度是一个一维向量，表示一天中第 i 小时的用户相对活跃强度。其计算过程为，首先求解数据集中所有消息平均每小时转发量 M，然后计算第 i 小时总转发量 $S[i]$（$1 \leqslant i \leqslant 24$），则第 i 小时的相对活跃度为

$$S'[i] = \frac{S[i]}{M} \tag{1}$$

式(1)从比例上反映出 Facebook 平台上每天任意小时内的用户活跃度，本文将在后面部分引入这个公式对预测模型进行修正。

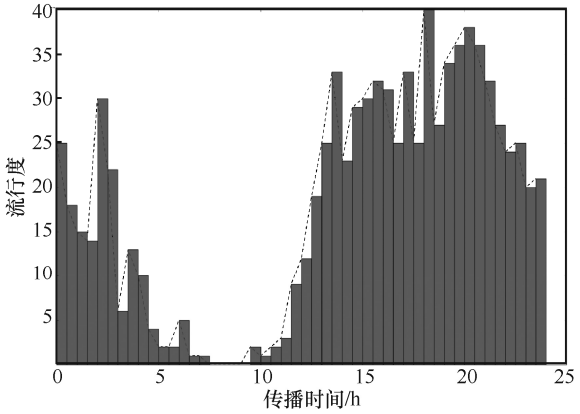


图 3　Facebook 用户每日活跃度

4.4　Facebook 中的弱连接现象

Facebook 用户可以浏览其他用户的页面墙并关注成为其粉丝，或者接受其他用户的好友邀请，这种好友关系在宏观层面上构成了一种较为稳定的拓扑结构。然而相比于静态的好友关系拓扑，根据主页发布信息的转发情况构建的交互图可以更好地反映出信息的传播能力。Ferrara 等[22]发现 Facebook 中的弱关系边对传播有很明显的增益效果。本文发现 Facebook 的信息转发规律符合社会学中的弱连接理论，通过将主页与历史上所有转发过该主页消息的用户构成一个交互图，将其中极少参与转发的用户称为弱连接节点，将那些经常参与转发的忠实粉丝称为强连接节点，并基于节点交互关系提出了连接强度的概念，量化了用户对于主页消息的转发的频繁程度。连接强度系数 f 表示用户 j 相对于主页 k 的转发频率，具体计算式为

$$f = \frac{c_{jk}}{\sum_{j=1}^{n_k} c_{jk}} \tag{2}$$

其中，c_{jk} 是用户 j 转发主页 k 所有发布消息的总频度，n_k 为历史上参与主页 k 上消息转发的所有用户数，则 f 为用户 j 在主页 k 的连接强度系数。

通过反复实验可以发现，在传播早期强连接用户比例较小且弱连接比例较大的消息，其最终流行度都很高。将各个主页 top 1%最频繁参与转发的用户作为强连接节点，并在参考时间 $T_{\text{reference}}$ 设置为 3 h 的条件下，将主页发布的消息中强连接节点所占比例与该消息最终流行度构成一组点对，图 4 描绘了数据集中所有消息发布后前 3 h 内强连接用户所占比例与发布 7 天后的最终流行度在双对数坐标系中构成的散点图，其中横纵坐标均以自然对数为底。

从图 4 中可以明显地观测出，这些点对在双对数坐标系中呈现较为明显的线性相关。根据弱连接理论，本文可以这样认为，如果在传播早期转发该消息的人中有较多忠实粉丝，那么传播过程会更局限于较为封闭的社区从而导致最终流行度较小；如果一个消息在传播早期可以吸引很多具有弱连接关系的陌生人进行转发，更容易扩散至多个圈子被更多的人关注并转发，从而获得较大的最终流行度。

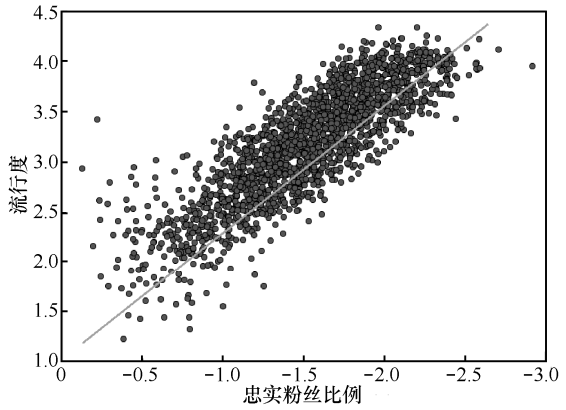


图 4 传播早期强连接用户参与比例与最终流行度的关系

通过将连接强度作为一个预测最终流行度的重要指标，在 SH 模型基础上添加连接强度这一特征，构建了一个二元线性模型

$$\ln B'_m = \alpha_1 \ln B_m + \alpha_2 \ln f + \alpha_3 \tag{3}$$

$$B'_m = \exp[\alpha_1 \ln B_m + \alpha_2 \ln f + \alpha_3] \tag{4}$$

将每个主页历史发布消息总条数的 75%作为训练集，剩余 25%作为测试集，并采用最小二乘法估计进行训练，得到参数 α_1、α_2 和 α_3。

进一步地，考虑到信息发布时间会对用户活跃度产生一定影响，从而导致早期流行度的观测值与真实传播能力不相符，因此本文引入相对流行度 B_m^*对其进行修正

$$B_m^* = \frac{B_m}{S'[i]} \tag{5}$$

将式(5)代入预测模型中的早期流行度 B_m 项，得到最终的 TSL 预测模型，其计算式为

$$B'_m = \exp[\alpha_1 \ln B_m^* + \alpha_2 \ln f + \alpha_3] \tag{6}$$

5 性能测试与分析

5.1 实验环境及数据

本节通过实验验证 TSL 模型的性能，数据集为表 1 所示的从 2016 年 1 月 1 日—12 月 31 日的部分热门 Facebook 主页数据，包含这些主页历史发布消息 3 775 条以及 154 万次转发 ID。

实验环境为 Intel 酷睿 i5-6500@ 3.2 GHz 四核，8.00 GB 内存，操作系统为 Microsoft Windows 7，编程语言为 Python。

5.2 对比模型

为了比较并验证本文提出的基于弱连接理论的流行度预测模型，通过将本文模型与其他 3 种主流模型进行对比来说明本文提出模型的有效性，参与比较的基准模型介绍如下。

1)SH 模型

SH 模型[8]是 Szabo 和 Huberman 研究在线视频与图片分享流行度时提出的经典模型，该模型基于早期流行度与最终流行度值存在对数坐标系下的线性关系。其计算式为

$$\ln N(T_{\text{predict}}) = p\ln\varphi + q\ln N(T_{\text{reference}}) + \sigma \tag{7}$$

其中，$N(T_{\text{predict}})$为最终流行度，φ 为通过最大似然估计得到的模型参数，σ 为修正项。这种线性回归方法可以用来做长期预测，但是由于特征选取比较简单，预测精度较低。

2) DSH 模型

DSH 模型是 Bao 等[14]提出的一种改进的线性回归模型，该模型首先测定了微博最终流行度和连边密度（link density）之间的关系。他们发现微博的最终流行度和连边密度之间存在着很强的负相关性，这表明低连接度和高传播深度的群体会更加促进微博流行度的提升。基于以上发现，研究者改进了 SH 模型。改进后的模型为

$$\ln\hat{p}_k(t_r) = \beta_1 \ln p_k(t_i) + \beta_2 \ln\rho_k(t_i) + \beta_3 \tag{8}$$

其中，$\hat{p}_k(t_r)$ 为 t_r 时刻的流行度，即晚期流行度；$p_k(t_i)$ 为 t_i 时刻的流行度，即早期流行度；$\rho_k(t_i)$ 为 t_i 时刻的连边密度；β_1、β_2、β_3 为数据集中训练出的参数。

3) RPP 模型

RPP 模型是一种基于动态泊松过程的时间序列模型[21,23]，通过结合时间松弛方程、线性增强方程和事件映射过程，可以模拟新颖性随时间衰减的过程。该模型针对短期预测效果较好，如时效性较强的微博、新闻等。

4) TSL 模型

本文提出的基于弱连接理论的线性回归模型，如式(6)所示。

5.3　评价指标

均方根误差（RMSE, root mean square error）是在有限测量次数下，测量值与真实值差的平方均值的平方根，在评价拟合效果方面被广泛使用，也能够体现出样本的离散程度。RMSE 越小表示测试数据与真实值偏差程度越小，其具体计算式为

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=0}^{n}(X_{\mathrm{obs,i}}-X_{\mathrm{model},i})^2}{n}} \tag{9}$$

其中，$X_{\mathrm{obs},i}$ 为 n 个测试样本真实数据的第 i 个结果，$X_{\mathrm{model},i}$ 为模型输出数据的第 i 个结果。

平均绝对百分误差（MAPE, mean absolute percentage error）是一种预测模型的常用评价方法，它通常使用百分比的形式展现。MAPE 越小，说明模型输出与真实值偏差越小。具体计算式为

$$\mathrm{MAPE}=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_t-F_t}{A_t}\right| \tag{10}$$

其中，A_t 为样本的真实值，F_t 为模型输出值。

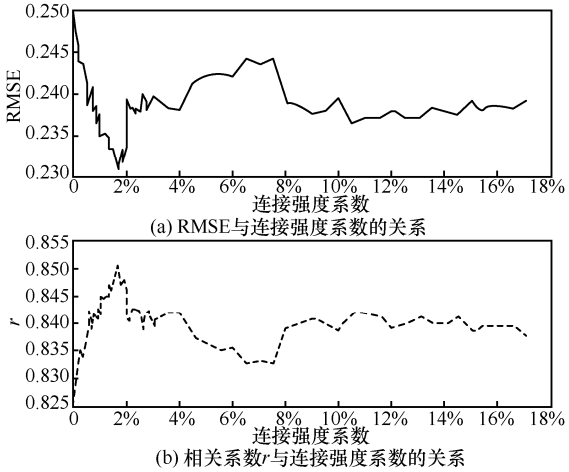
皮尔逊相关系数（Pearson correlation coefficient）用于评价线性相关变量（X 与 Y 之间相互关系）之间相关关系密切程度的统计指标。皮尔逊相关系数的取值范围为−1~1，当绝对值为 1 时，称 X 与 Y 完全相关；当绝对值为 0 时，称 X 与 Y 不相关；当绝对值大于 0.8 时，X 与 Y 高度相关；当绝对值小于 0.3 时，X 与 Y 低度相关。皮尔逊相关系数 r 的计算式为

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{11}$$

5.4　实验结果

本节设置了 3 组实验，首先分析模型中连接强度系数 f 的取值对预测性能的影响，然后在各个主页数据集上测试本文 TSL 模型与 SH 模型的预测性能，最后对所选主页进行分类，用多个模型进行对比分析。

首先，为了取得最优预测效果，需要预先设定连接强度系数 f 的值，在这个过程中有 2 个问题：当 f 取值过小时，训练数据也减少，从而导致预测模型失真，因为个别消息在发布前几小时可能并没有忠实粉丝进行转发，其次强连接用户为 0 会导致不可进行取对数操作，因此本文假设主页自身就是一个忠实转发者，这样任意发布的消息至少存在一个忠实粉丝进行转发，从而不会造成点对的缺失；当 f 取值过大，即强连接节点门槛较低时，模型逐渐退化为 SH 模型。Fox News 主页经平滑后的随 f 取值变化的预测性能趋势如图 5 所示，其中参考时间 $T_{\mathrm{reference}}$ 取值为 3 h，横坐标为连接强度系数 f。从图 5 可以看出，传播初期随着 f 逐渐增大，RMSE 逐渐减小，相关系数 r 逐渐增大。当 f=1.8%时，2 个指标同时达到极值点。



(a) RMSE与连接强度系数的关系

(b) 相关系数r与连接强度系数的关系

图 5　连接强度系数 f 的取值对 Fox News 主页预测性能的影响

其次，为了验证模型的正确性，本文选取同样基于线性回归方法预测最终流行度的 SH 模型进行比较。采用 SH 模型的 Fox News 主页预测散点如图 6(a)所示。其中横坐标为信息发布后参考时间 $T_{\mathrm{reference}}$ 取值为 3 h 的早期流行度，纵坐标为消息发布 7 天后的最终流行度，横纵坐标均以自然对数为底。将数据集中 75%的点对作为训练集，数据集中另外 25%的点对作为测试集。采用基于连接强度的 TSL 模型预测效果如图 6(b)所示。从图 6(b)可以明显看出，采用该模型训练后离散点有减少趋势，更多的点都汇聚在直线上，拟合效果更优，这说明融合连接强度的二元线性回归模型可适用于流行度预测。

表 2 给出了所有主页的拟合结果的详细参数。从表 2 可以看出，各主页 RMSE 指标均在 0.35 以下，说明误差较小，而相关系数 r 达到 0.8 以上，为高相关。此外还发现连接强度 f 的最优解因主页的异同而波动较大，而转发数较多的 Harry Potter、The Simpsons 主页分别为 0.1%和 0.4%，转发数最多的 NBA 主页的值却接近 10%，由此可见，连接强度与转发用户的数量并没有直接关联。另一个发现是，娱乐类主页的连接强度普遍小于 1%，如 Harry

Potter、The Simpsons 等名人或电影的公共主页。而 NBA、History、Fox News 等属性鲜明的兴趣类主页连接强度都较大，这意味着这些主页中有比例更高的忠实粉丝进行规律性的转发，连接强度特征将更适合预测转发流行度。

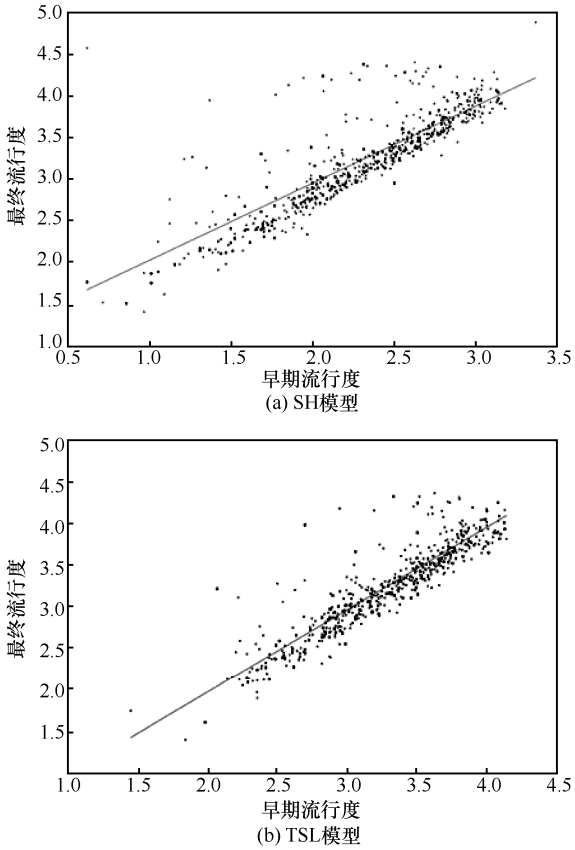


图 6 Fox News 主页的预测效果

表 2 Facebook 主页预测结果

主页	RMSE	r	f	转发用户数
Geographic	0.278 325 94	0.854 637 37	1.9%	177 130
History	0.327 377 78	0.800 401 19	2.4%	76 022
NBA	0.286 945 86	0.877 307 71	9.5%	330 728
Call of Duty	0.191 666 00	0.937 414 42	0.4%	38 357
Grey's Anatomy	0.261 488 96	0.858 801 76	0.7%	64 571
Fox News	0.308 733 30	0.851 034 12	1.8%	215 803
The Simpsons	0.246 806 58	0.903 169 68	0.4%	157 205
Will Smith	0.217 223 48	0.902 040 75	0.3%	144 793
Harry Potter	0.195 716 76	0.889 507 83	0.1%	120 824
Barack Obama	0.288 655 17	0.766 377 99	1.7%	211 807

接下来，本文将 TSL 模型与其他 3 种较为主流的流行度预测模型在表 1 所示的 A（兴趣类）、B（娱乐类）2 类数据集中进行预测效果对比，如图 7 所示。其中，参考时间 $T_{\text{reference}}$ 设置为 3 h，通过调整预测时间 T_{predict} 来观察各模型的长期预测效果。对于 RPP 模型，本文将初始参数□设置为 10，图 7(b)给出了这几种模型在娱乐类主页数据集上的 MAPE 测度评价。从图 7(b)可以看出，RPP 模型在中短期的预测误差要优于其他模型，但随着预测时间 T_{predict} 的增长和转发量的积累，TSL 模型的长期流行度预测效果逐渐显现优势。在兴趣类主页数据集上，如图 7(a)所示，当 $T_{\text{predict}} \geqslant 4.5$ 天时，TSL 模型的长期流行度预测优势更为明显，表明 TSL 模型对于长期预测性能更优。这可能是由于兴趣类主页的关注群体较为固定，忠实粉丝群体转发活动较为规律，在这种场景下连接强度对最终流行度有更强的指示作用。

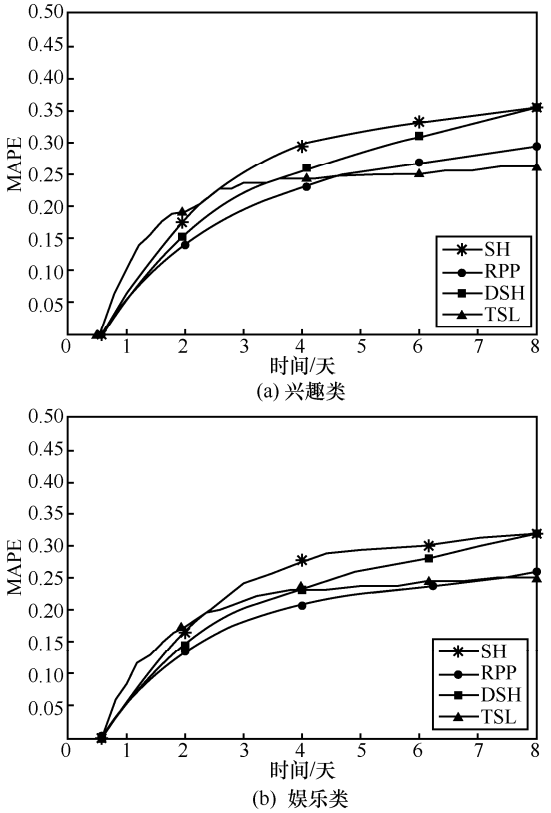


图 7 各模型的 MAPE 随时间变化趋势

6 结束语

本文研究了 Facebook 消息转发流行度的早期传播特征与最终流行度之间的关系，提出了一种 Facebook 流行度预测 TSL 模型。首先介绍了社会学中的弱连接理论，提出了连接强度系数，进而发现在信息传播早期具有强连接属性的忠实

粉丝比例与最终流行度在双对数坐标系中呈现线性相关。其次，通过融合早期流行度与连接强度系数提出了一种面向 Facebook 交友网络的流行度预测模型。最后，根据 Facebook 真实数据集的实验分析表明，所提模型可以对消息的最终转发流行度进行有效预测，相较于同类主流模型有较好的预测效果。

参考文献：

[1] HUIJBOOM N, VAN DEN BROEK T, FRISSEN V, et al. Key areas in the public sector impact of social computing[J]. European Communities, 2009, 3.

[2] GUPTA M, GAO J, ZHAI C X, et al. Predicting future popularity trend of events in microblogging platforms[J]. Proceedings of the American Society for Information Science and Technology, 2012, 49(1): 1-10.

[3] ZHAO Y, QIN B, LIU T, et al. Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog[J]. Multimedia Tools and Applications, 2016, 75(15): 8843-8860.

[4] BORAH P. Political Facebook use: campaign strategies used in 2008 and 2012 presidential elections[J]. Journal of Information Technology & Politics, 2016, 13(4): 326-338.

[5] ALBERT-LASZLO B, REKA A. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[6] 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.

[7] GOLDER S A, WILKINSON D M, HUBERMAN B A. Rhythms of social interaction: messaging within a massive online network[C]//Communities and Technologies 2007. Springer, 2007: 41-66.

[8] SZABO G, HUBERMAN B A. Predicting the popularity of online content[J]. Communications of the ACM, 2010, 53(8): 80-88.

[9] FRIEDKIN N E. Information flow through strong and weak ties in intraorganizational social networks[J]. Social Networks, 1982, 3(4): 273-285.

[10] ABDULLAH S, WU X D. An epidemic model for news spreading on twitter[C]//2011 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2011: 163-169.

[11] MATSUBARA Y, SAKURAI Y, PRAKASH B A, et al. Rise and fall patterns of information diffusion: model and implications[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012: 6-14.

[12] LI H, MA X, WANG F, et al. On popularity prediction of videos shared in online social networks[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013: 169-178.

[13] CHANG B, ZHU H, GE Y, et al. Predicting the popularity of online serials with autoregressive models[C]//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. ACM, 2014: 1339-1348.

[14] BAO P, SHEN H W, HUANG J, et al. Popularity prediction in microblogging network: a case study on sina weibo[C]//Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 177-178.

[15] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv Preprint, arXiv:1408.5882, 2014.

[16] CHENG J, ADAMIC L, DOW P A, et al. Can cascades be predicted?[C]//Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014: 925-936.

[17] 朱海龙, 云晓春, 韩志帅. 基于传播加速度的微博流行度预测方法[J]. 计算机研究与发展, 2018, 55(6): 1282-1293.
ZHU H L, YUN X C, HAN Z S. Weibo popularity prediction method based on propagation acceleration[J]. Journal of Computer Research and Development, 2018, 55(6): 1282-1293.

[18] CRANE R, SORNETTE D. Robust dynamic classes revealed by measuring the response function of a social system[J]. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649-15653.

[19] YANG J, LESKOVEC J. Patterns of temporal variation in online media[C]//Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 177-186.

[20] LERMAN K, HOGG T. Using a model of social dynamics to predict popularity of news[C]//Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 621-630.

[21] GAO S, MA J, CHEN Z. Modeling and predicting retweeting dynamics on microblogging platforms[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 107-116.

[22] FERRARA E, DE MEO P, FIUMARA G, et al. The role of strong and weak ties in Facebook: a community structure perspective[J]. arXiv Preprint, arXiv:1203.0535, 2012.

[23] BAO P, SHEN H W, JIN X, et al. Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes[C]// Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 9-10.

[作者简介]

王晓萌（1987− ），男，黑龙江哈尔滨人，哈尔滨工业大学博士生，主要研究方向为在线社交网络、信息传播预测、舆情安全等。

方滨兴（1960− ），男，江西万年人，中国工程院院士，哈尔滨工业大学教授、博士生导师，主要研究方向为计算机网络与信息安全理论与技术、并行计算等。

张宏莉（1973− ），女，吉林榆树人，博士，哈尔滨工业大学教授、博士生导师，主要研究方向为网络与信息安全、网络测量与建模、网络计算、并行处理等。

王星（1981− ），男，重庆人，博士，哈尔滨工业大学助理研究员，主要研究方向为网络与信息安全、网络舆情监控、知识迁移。