新浪微博网信息传播分析与预测

曹玖新[1)] 吴江林[1)] 石 伟[1)] 刘 波[1)] 郑 啸[1),2)] 罗军舟[1)]

[1)](东南大学计算机科学与工程学院 南京 210096)

[2)](安徽工业大学计算机学院 安徽 马鞍山 243002)

摘 要 文中以新浪微博为研究对象,以分析新浪微博的信息转发与传播特征为研究目的,并对传播行为进行预测.在获取大量新浪微博在线数据的基础上,对各种可能影响用户转发行为的因素进行统计、分析,挖掘各种影响因素特征并进行建模.提出基于用户属性、社交关系和微博内容三类综合特征,使用机器学习的分类方法,对给定微博的用户转发行为进行预测.基于微博网关注关系拓扑,利用概率级联模型对给定微博的转发路径进行预测,为预测微博的影响范围提供依据.文中通过实验分析了新浪微博符合复杂网络特征、社交类特征对转发行为有重要影响,并验证了传播预测的有效性.

关键词 微博;转发;信息传播;预测;社交网络;社会计算

中图法分类号 TP393 **DOI号** 10.3724/SP.J.1016.2014.00779

Sina Microblog Information Diffusion Analysis and Prediction

CAO Jiu-Xin[1)] WU Jiang-Lin[1)] SHI Wei[1)] LIU Bo[1)] ZHENG Xiao[1),2)] LUO Jun-Zhou[1)]

[1)](*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)

[2)](*School of Computer Science, Anhui University of Technology, Maanshan, Anhui* 243002)

Abstract In this paper, research is conducted on Sina microblog for the purpose of analyzing information forwarding and propagation characteristics, as well as predicting propagation behavior. Based on a large number of online data from Sina microblog, a variety of possible factors that affect users' retweeting behavior have been analyzed and various features have been mined and modeled. Three comprehensive features, based on user attributes, social relations and microblog contents, are used to predict users' retweet behavior by machine learning classification algorithms. The microblog topology graph on following relation is constructed, and the cascade probability model is used to predict the propagation paths of a tweet, then a tweet's influence can be predicted. Experiment indicates that Sina microblog meets complex network characteristics, and social characteristics have a greater influence on forwarding behavior. Furthermore, it verifies the validity of propagation prediction.

Keywords microblog; retweet; information diffusion; prediction; social network; social computing

收稿日期:2013-06-20;**最终修改稿收到日期**:2014-01-23.**本课题得到国家"九七三"重点基础研究发展规划项目基金**(2010CB328104)、**国家自然科学基金**(61272531,61202449,61272054,61370207,61370208,61300024,61320106007)、**国家"八六三"高技术研究发展计划项目基金**(2013AA013503)、**高等学校博士点学科专项科研基金**(20110092130002)、**江苏省网络与信息安全重点实验室项目**(BM2003201)、**计算机网络和信息集成教育部重点实验室(东南大学)项目**(93K-9)**资助**.**曹玖新**,男,1967年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为服务计算、网络安全与社会计算.E-mail:jx.cao@seu.edu.cn.**吴江林**,男,1988年生,硕士研究生,主要研究方向为社会计算.**石 伟**,男,1987年生,硕士研究生,研究方向为社会计算.**刘 波**,女,1975年生,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为普适计算、社会计算.**郑 啸**,男,1975年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为服务计算、无线局域网.**罗军舟**,男,1960年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为下一代网络体系结构、协议工程、网络安全、网格与云计算、无线局域网.

1 引 言

随着互联网技术的普及,更多的人享受到互联网带来的便利和乐趣.尤其是近几年在线社交网络[1]的迅速发展,人们越来越多地参与到互联网上丰富的社交活动中.以国内最大的微博网站——新浪微博为例,截至目前新浪微博的注册用户已超过5亿,每天有超过1亿条微博内容产生.事实上,以微博为代表的社交网络已经成为当前最重要的媒体之一.因此,研究用户的在线行为以及信息的传播规律将有助于网络公司对每个用户的偏好进行更为准确地把握,并将其可能感兴趣的话题信息、其他用户或者用户社群推荐给用户;政府部门则可以通过预测消息传播范围和用户观点态度,准确判断舆论的热点问题,以便及时采取科学的控制和引导.

2 相关工作

在线社交网络的信息传播行为已经成为当前的研究热点.Boyd 等人[2-3]以 Twitter[2]为研究对象,对 Twitter 上人们 Retweet 的方式(类似国内微博网站上的"转发")、Retweet 的动机以及被 Retweet 信息的内容主题倾向进行了分析.Suh 等人[4]选取了 URL、标签、关注人数等因素,使用主成份分析方法(PCA)分析了影响用户转发的主要因素,最后结合所选因素应用广义线性模型分析影响因素与转发行为之间的关系.但这些研究仅仅是对转发行为的统计分析,缺少对用户行为的预测.文献[5]采用了基于概率的协同过滤模型,称为 Matchbox[6],作者选取了用户名、关注人数、微博包含的单词个数等特征对用户的转发行为进行预测,虽然这些特征在一定程度上反映了用户行为特点,但并不是最主要的影响因子.文献[7-8]中作者选取了22个影响因素,并采用因子图模型进行预测分析,对于用户转发行为预测取得了较高的精度,但由于对所选特征的量化处理过于简单,信息传播路径预测的精度较低.

Liben-Nowell 等人[9]较为全面地讨论了信息在真实社会网络中传播的特征及与之相关的一系列问题,并且指出:精确的预测信息传播路径是非常困难的;使用简单的预测模型往往与真实情况相距甚远,比如 Email 的实际传播模式与小世界传播模型不同,小世界模型人们之间的距离都比较短,但实际上电子邮件传播有的要经过数百个中间节点.在 Kossinets 等人[10]的工作中,作者首先对社会网络数据进行网络聚类,之后生成特征结构传播树并得到异步响应时间,最终提出并描述了一个概率模型.也有不少研究采用 SIR、SIS 等经典传播模型[11]对信息传播扩散进行分析研究.

研究者对不同类型的多种在线社交网络进行了研究,包括 Flickr[12]、Blogs[13]、Digg[14]以及 YouTube[15]等,这些研究揭示了信息在社交网络上传播所表现出来的规律.Fan 等人[16]对新浪微博的拓扑及信息扩散进行了研究,发现新浪微博的拓扑结构具有小世界及无标度特性,度分布服从幂律分布,热门事件的扩散拓扑呈现星形或两级结构.Webberley 等人[17]对信息在 Twitter 上传播的深度、产生影响的广度以及转发延迟进行了研究,作为对用户行为模式和转发规律的初步研究对本文的研究工作具有很好的借鉴意义.Galuba 等人[18]对 Twitter 上 URL 信息在用户之间传播规律进行了研究,提出了一个预测 URL 转发路径预测模型.Yang 等人[19]发现在 Twitter 上信息内容对相关用户的提及率是影响该信息传播速度、规模及范围的重要因素.

以上针对在线社交网络中信息传播路径和影响范围的分析预测大多得到的是信息在特定网络上传播所呈现的统计规律,例如文献[9]的研究结果说明 Email 网络的传播树呈现传播广度小、传播深度大的特性,但是并没有针对特定信息预测其传播可能经过的路径.同时部分研究将网络中的节点视作具有相同或相近的行为模式[18],这显然是不合理的,因为网络中的每个个体具有各自独立的行为特征.另外,目前大多数研究工作是针对国外的社交网络,如 Twitter、Facebook、Flickr 等,针对国内在线社交网络的研究还比较少.随着我国经济的发展,互联网普及率大大提高,我国的网民数量比例已超过国内人口总数三分之一,仅新浪微博的注册人数就超过了5亿,地域性也是在线社交网络的一个主要特性,不同国家的社交网络在用户行为、信息传播上可能呈现出不同的特性,因此,针对国内在线社交网络的研究更为迫切.

针对以上存在的问题及不足,本文首先根据实际社交网络新浪微博在线数据,对各种可能影响用户转发行为的因素进行统计分析,获得各种因素对

用户转发行为的影响关系，然后重点研究微博类社交网络上用户的转发行为模式，给出用户转发预测模型，并以此建立信息传播路径级联概率模型，实现对用户行为及信息传播影响趋势的预测.

3　数据描述

我们基于新浪微博的开放接口[①]开发了爬虫程序来抓取需要的数据. 程序依照广度优先的策略，从一个特定的用户开始，爬取该用户最近发表的 100 条微博，对于其中的每条微博，再爬取该微博的转发微博以及转发该微博的用户，将这些用户添加至待爬取队列. 结束对一个用户的处理之后，再取出待爬取用户队列中的第一个用户，继续同样的处理，循环往复. 爬取程序不间断运行了一周时间，通过这种方式，最终从 1 935 391 个用户中获得了 10 785 921 条微博消息. 经过初步统计，我们发现其中 28.98%的微博是原创的，71.02%的微博是转发产生的. 由于我们的目标是发现转发模式，通过该方法抓取的转发样本要比其他方法得到的数据集比例更大.

3.1　数据完整性分析

我们按照广度优先的次序对转发网络进行遍历能够获得较为完整的子图. 同时在抓取过程中过滤掉活跃度过低的用户，这些用户的行为比较随机，历史行为也比较少，不具有代表性. 随后我们又抓取了所有用户之间的 137 284 538 条关注关系，得到了完整的关注拓扑.

通过对关注网络拓扑进行分析，发现入度分布近似满足幂律分布，如图 1 所示；出度分布满足 150 定律[②]，如图 2 所示，由于新浪微博对普通用户关注人数的限制为最多 2000 人，因此在横轴 2000 的位置附近出现了一个人数高峰.

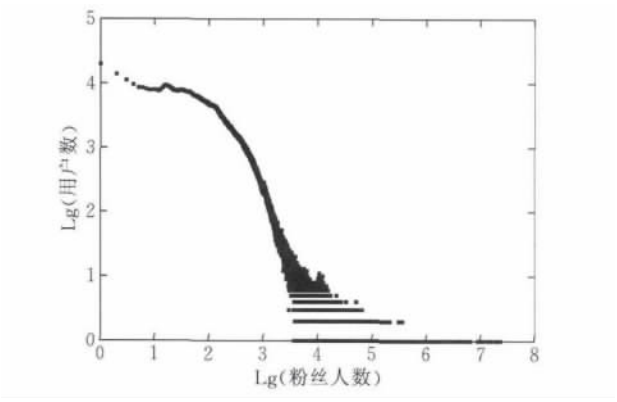

图 1　关注网络拓扑入度分布

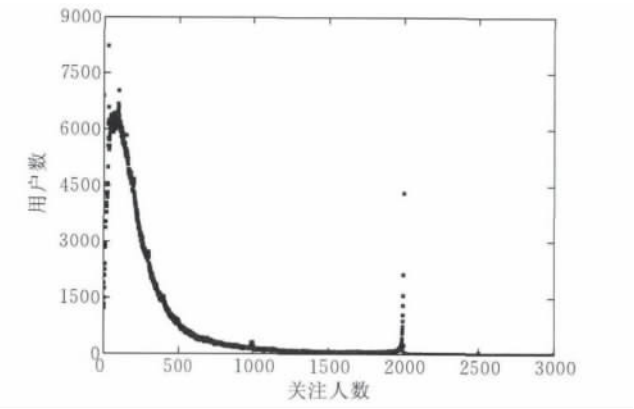

图 2　关注人数分布

我们对关注网络进行采样，并计算出样本网络的聚类系数为 0.168，直径为 7，平均距离为 3.068，可以看出新浪微博关注网络具有较高的聚类系数和较小的平均距离，符合小世界特征. 但是节点之间平均距离与我们的常识不太一致，不同于人人网、Facebook 等网络，微博关注网络是有向的，而平均距离却更小（Facebook 上用户之间的平均距离为 4.74[③]），这是一个有趣的现象. 说明微博上人与人之间的联系更为紧密，也就是说消息往往只需要经过很少的跳数就能传播到网络上的其他人.

我们对转发树的拓扑也进行了分析，列举出了几种常见模式，图 3 是 4 条微博的转发树拓扑，可以看出转发树的拓扑结构主要分为两类：星形结构

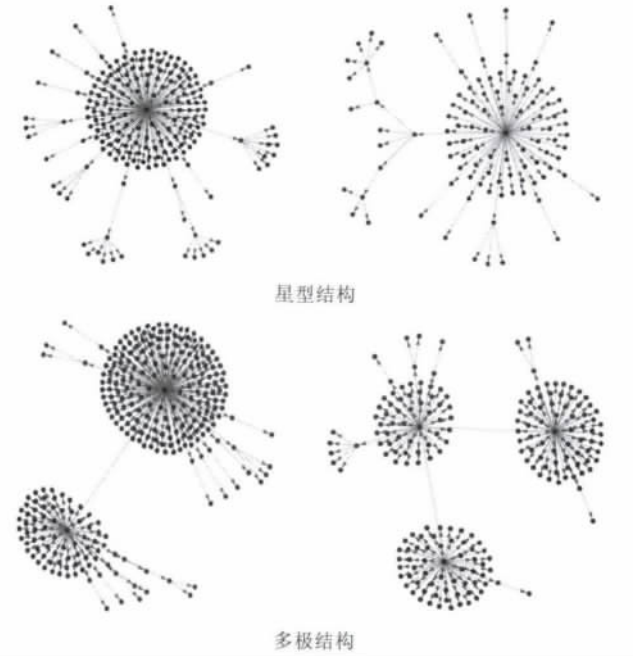

图 3　转发树拓扑

① http://open.weibo.com
② http://en.wikipedia.org/wiki/Dunbar's_number
③ https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook

和多极结构. 星形结构往往是以微博原创者为中心，其第一层转发者绝大部分都是他的粉丝，再向外则只有少量转发；多极结构的形成往往是因为有多个入度较大的节点参与转发，高入度节点的每次转发都会引起一次转发高峰，因此可以认为推动微博在网络上扩散的原因不仅仅在于微博内容，更在于是否有高入度的节点参与转发.

分析显示，不管何种传播模式都有一个共同的特点，即传播树在深度上都比较小，广度却比较大. 图 4 反映转发树最大深度分布近似符合幂律分布，也就是说大多数的传播路径都比较短. 这也在一定程度上解释了消息为什么能在微博网络上迅速传播.

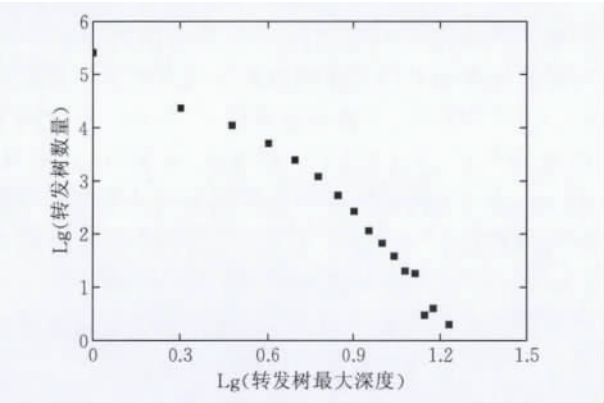

图 4 转发树最大深度分布

3.2 转发样本识别

转发样本比较容易识别. 以新浪微博为例，用户点击转发按钮后会弹出转发对话框，如图 5 所示. 用户可以添加一段评论，点击发布后，产生一条新的微博，该微博的格式形如：//@UserScreenName：Previous Hop Weibo Text.

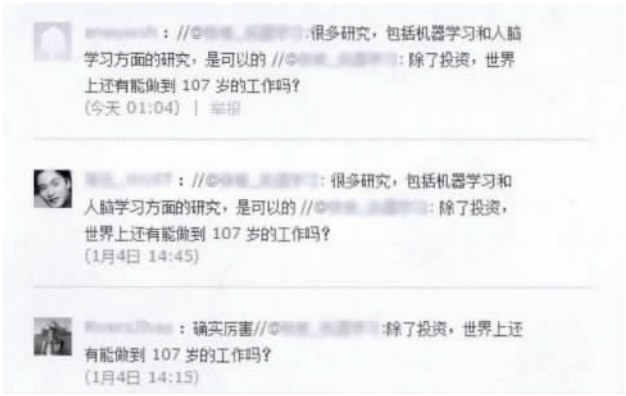

图 5 转发微博

因此如果微博中出现了“//@UserScreenName：”这样的模式，说明该微博是通过转发 UserScreenName 用户的微博产生的. 新浪微博相应的 API 也会返回转发微博的源微博标志(ORIMID，对于原创微博返回 0)，因此通过该方法能够识别转发样本.

3.3 忽略样本识别

为了研究微博如何在网络上传播，我们需要知道在何种情况下，人们愿意将微博分享给自己的关注者. 因此需要判定用户看到微博时的两种行为：转发行为和忽略行为. 我们将转发样本作为正例，将忽略样本作为负例.

定义 1. 若用户 u 转发了在 t 时刻发表的微博，则将他关注的用户在$[t-\Delta t, t+\Delta t]$时间区间内发表的且未被 u 转发的微博称为忽略样本.

相比于转发样本识别，忽略样本的识别较为困难，原因在于用户忽略动作行为无法显式地体现在数据集中. 用户没有转发微博可能并不是用户主观忽略微博行为，也可能是由于用户不在线而错过的消息. 为了解决这个问题，我们通过用户的转发动作来识别忽略行为，以提高样本的准确度.

算法描述如下：

算法 1. 微博忽略样本识别算法.

输入：用户 u_i 关注的用户发表的微博集合 P_i；
用户 u_i 转发微博集合 R_i

输出：用户 u_i 忽略的微博列表 I_i

1. 对于任意一条微博 $m \in R_i$ (t_m 为 m 的发表时间)：
2. 取出 P_i 中发表时间满足 $t \in [t_m-\Delta t, t_m+\Delta t]$ 的微博，记为 Q_m；
3. 对于任意一条微博 $w \in Q_m$：
4. 若 $w \notin R_i$，则将 w 加入 I_i；
5. 返回 I_i.

以新浪微博为例，当用户登录微博后，他所关注的用户最近发表的微博会按照时间的倒序展现在页面上. 用户通常从上至下顺序阅读，再通过点击“下一页”按钮阅读更早的微博. 当用户遇到感兴趣并认为值得转发的微博(假设该微博发表时间戳为 t_1)，他们会点击转发按钮，并最终在时间 t_2 生成新的微博，如图 6 所示.

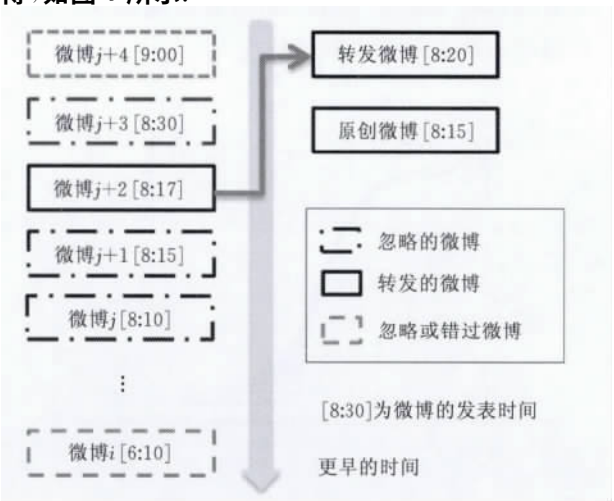

图 6 用户忽略行为识别

因此，我们认为其关注的用户在 t_1 前后一小段时间内$[t_1-\Delta t, t_1+\Delta t]$发表的微博已经被该用户阅读，如果没有转发则是该用户主观忽略的微博. Δt 越小，忽略动作的识别越准确. 在本文中，Δt 取 15 min.

4　影响用户转发行为因素分析

在本小节中，通过归一化处理，我们将分析对比可能促进或者制约用户转发行为的若干影响因素，并归纳其各自特征.

4.1　用户粉丝人数

在众多影响用户转发行为的特征中，微博原创者的影响力可能会对下游用户的行为产生影响. 而直接衡量一个用户的影响力比较困难，采用原创者的粉丝数（关注他的人数）可以在一定程度上体现一个用户的影响力.

从图 7 可以看出，当用户粉丝数处于一个比较小的量级时，随着用户粉丝数的增加，转发可能性反而降低. 这个与我们直观上的认识不太一致，分析原因可能是：结合之后的特征分析，我们发现用户之间的交互特征才是影响用户转发行为最主要的特征，而微博上绝大多数的用户都是“草根”，和名人交互的机会不多，所以平时转发的往往都是身边朋友的微博，而这些人的关注人数都是比较少的，因此导致粉丝数较少的用户微博被转发的可能性反而高. 我们可以观察用户微博的转发次数和粉丝人数的比值（在之后的分析中我们也考虑了这个因素），名人虽然粉丝数多，但是转发数与粉丝数的比值往往比不上普通用户，这也说明用户粉丝多少与微博被转发的可能性不一定成正比关系. 而粉丝人数与微博被转发次数是正相关的，我们随机抽取了 11 978 条原创微博，图 8 展示了微博发布者粉丝数与该微博被转发次数之间的关系，从图上可以看出粉丝人数和转发次数成正相关.

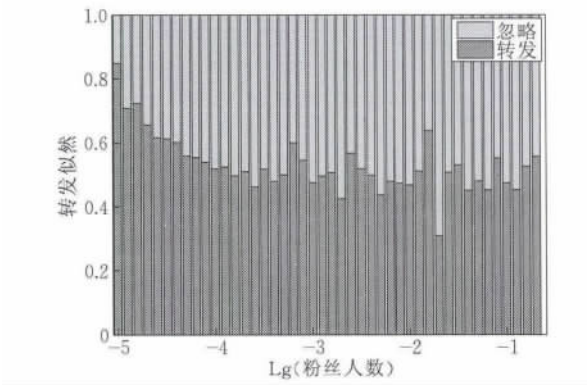

图 7　微博原创者粉丝人数与转发似然之间的关系①

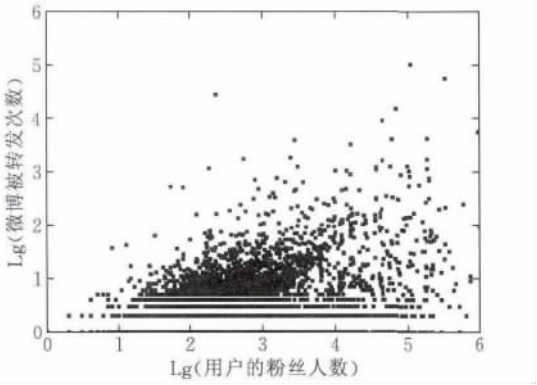

图 8　微博发布者粉丝人数与微博被转发次数之间的关系

我们使用CDF(Cumulative Distribution Function，累积概率分布函数)图描述一个特征对于用户不同行为的区分度，图 9 描述了用户粉丝数这个特征在被转发微博和被忽略微博上的累积分布函数，从图上可以看出用户粉丝数这个特征在转发和忽略上的分布较一致，因此仅仅使用该特征还不能很好地预测用户的转发行为. 粉丝人数在一定程度上表征了用户在微博网络上的影响力（粉丝越多，微博的受众也越多），然而微博上也有不少的僵尸用户，因此衡量一个用户的影响力，仅仅依赖粉丝数量是不全面的.

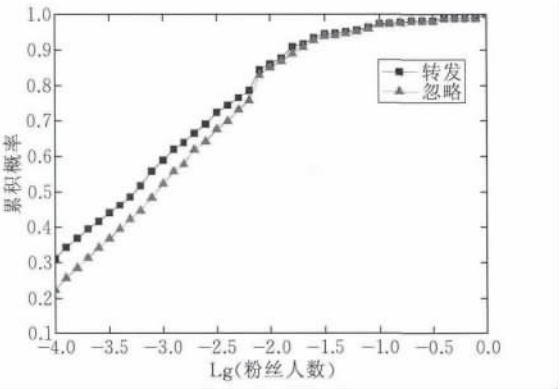

图 9　粉丝数量在转发和忽略上的分布

4.2　用户的 PageRank 值

与其他社交网络如人人网、Facebook 不同，微博的关注网络是有方向性的. 一个用户的影响力也可以通过他的粉丝质量来体现，即如果一个用户的粉丝影响力越大，那么说明该用户也具有较大的影响力.

基于以上的考虑，本文采用 PageRank 算法[20]计算用户在关注网络中的 PageRank 值，作为用户影响力的度量指标之一.

用户节点 PageRank 值计算公式如下：

① 我们对所有量化后的特征均进行了归一化处理.

$$pr_i=\frac{1-q}{N}+q\sum_{j\in Follower(i)}\frac{pr_j}{|Friend(j)|} \quad (1)$$

其中，pr_i 代表用户 i 的 PageRank 值，$Follower(i)$ 代表用户 i 的粉丝集合，$Friend(j)$ 代表用户 j 关注的用户集合，q 为阻尼系数，N 为用户总数.

通过新浪微博开放接口，我们获得了数据集中 193 万用户之间的 137 284 538 条关注关系. 使用 Graphchi 程序包[①]计算用户的 PageRank 值.

从图 10 可以看出，微博原创者的 PageRank 值与转发之间的关系呈现出的趋势与用户粉丝数特征相同，从 CDF 图(图 11)也可以看出 PageRank 也不能很好地预测用户的转发行为.

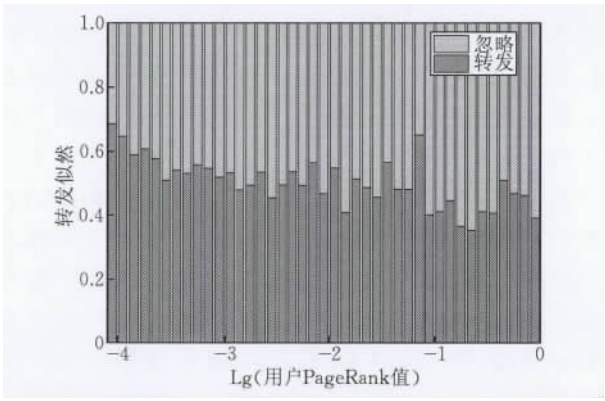

图 10 微博原创者 PageRank 值与转发似然之间的关系

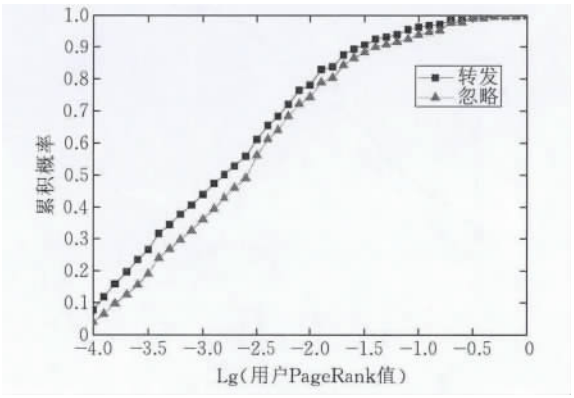

图 11 PageRank 值在转发和忽略上的分布

4.3 用户的转发活跃度

我们按如下方式定义单个用户的转发活跃度 f_{rr}：

$$f_{rr}=\frac{n_{\text{repost}}}{n_{\text{post}}} \quad (2)$$

其中，n_{repost} 代表用户最近发布的微博中转发微博的数量，n_{post} 代表用户最近发布微博的数量. 这个值越大说明用户在微博上转发的频率越高，高转发频率的用户对于微博的扩散起了比较大的推动作用.

从图 12 和图 13 可以看出用户转发活跃度与转发行为之间存在着较强的关系，整体上随着用户转发活跃度的上升，微博被转发的概率也随之上升.

4.4 与上游用户的交互强度

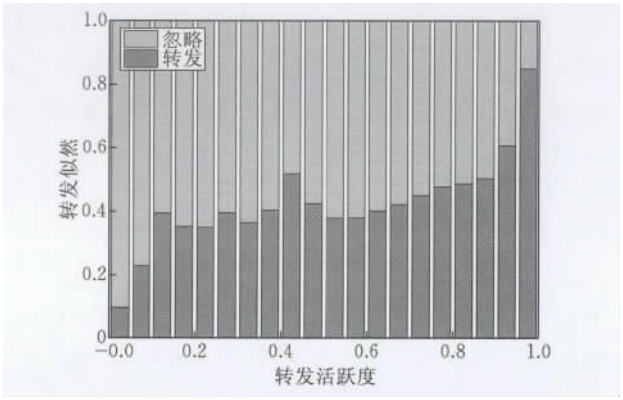

图 12 用户转发活跃度与转发似然之间的关系

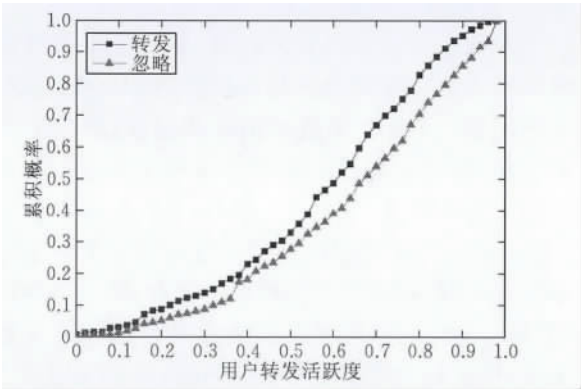

图 13 用户转发活跃度在转发和忽略上的分布

用户之间的历史交互频率可能会影响用户的转发行为，因此本文分析被预测用户和上游传递者之间的交互强度与转发似然之间的关系. 我们按如下方式定义用户 u 与上游传递者 v 之间的交互强度 f_{uv}：

$$f_{uv}=\frac{n_{uv}}{n_u} \quad (3)$$

其中，n_{uv} 表示 v 的微博出现在 u 的转发微博中的次数，n_u 表示 u 转发微博的总次数. 该值越大说明用户与上游用户之间的交互强度越大.

用户和上游传递者之间的交互强度与转发似然之间的关系如图 14 所示.

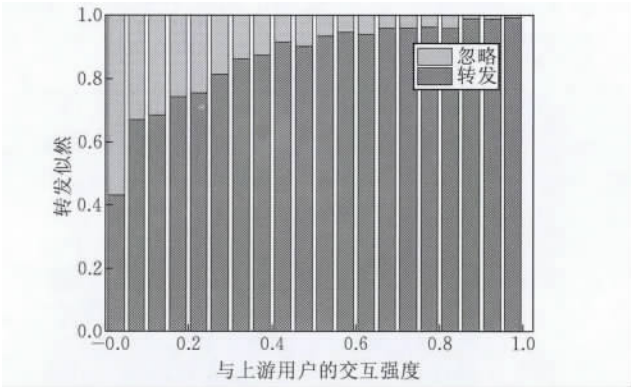

图 14 上游传递者之间的交互强度与转发似然之间的关系

① http://code.google.com/p/graphchi

从图 14 可以看出，如果用户与上游用户之间的交互强度越大，那么该用户转发上游用户微博的可能性也越大，与我们的直观认识一致，而且从图 15 可以看出用户交互强度在转发和忽略上的区分度较大，因此该特征可以很好地用于用户转发行为的预测.

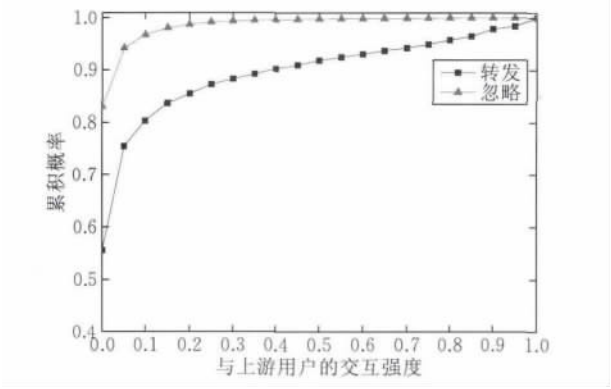

图 15　用户之间的交互强度在转发和忽略上的分布

4.5　微博内容与用户历史兴趣相似度

该特征考虑微博内容与用户兴趣偏好的相近程度，用户的兴趣偏好可以通过分析用户的历史转发记录得出. 我们直观地认为微博内容越符合用户的兴趣爱好，它被该用户转发的可能性也就越大.

由于用户的兴趣偏好具有时效性，用户最近一段时间转发的微博最能代表用户近期的兴趣偏好，因此本文仅将用户最近转发的信息集合作为历史记录. 既保证了预测精度，又降低了计算代价. 首先对待预测微博和待预测用户的文本进行分词，并将这两段语料表示成向量空间模型(VSM)[21]，向量中的每个元素为对应词的 TF-IDF 值①，将两者的余弦值作为当前信息与该用户兴趣偏好的相关性度量. 余弦值越大，说明文本之间的夹角就越小，两段语料也就相似.

信息 C_{content} 和用户历史转发记录集合 C_{history} 的向量空间计算过程如下：

(1) 采用 ICTCLAS 汉语分词系统②对 $C=C_{\text{content}}\cup C_{\text{history}}$ 进行分词，得到词汇字典 $D=\{w_1,w_2,\cdots,w_n\}$，其中 w_i 为 C 出现过的词语(不包括“是”、“的”等停止词)，n 为出现过的单词个数.

(2) 对 D 中的每个词语计算其 TF-IDF(Term Frequency-Inverse Document Frequency)值，作为单词的权重，记为 t_i

$$t_i=TF_i\times IDF_i=\frac{n_i}{\sum_1^k n_k}\times\log\frac{|C|}{|\{c:w_i\in c\}|} \tag{4}$$

其中 n_i 代表词语 w_i 在文本中的词频，k 表示文本中出现过的不同单词总数. $\log\frac{|C|}{|\{c:w_i\in c\}|}$ 为词 w_i 的逆文档频率(IDF)，是词语重要性的度量，这里由微博总数除以包含该词语的微博数目，再将商取对数得到.

(3) 生成向量空间表示

$$V_{\text{content}}=(t_1,t_2,\cdots,t_n),$$
$$V_{\text{history}}=(t'_1,t'_2,\cdots,t'_n) \tag{5}$$

用当前信息和用户历史转发记录的向量空间模型的余弦值作为内容相关性的度量值：

$$F_{\text{content}}=\cos\langle V_{\text{content}},V_{\text{history}}\rangle=\frac{V_{\text{content}}\cdot V_{\text{history}}}{|V_{\text{content}}|\times|V_{\text{history}}|} \tag{6}$$

在给定相似度数值度量方法之后，似然分布如图 16 和图 17 所示. 从图中可以看出如果微博内容与用户兴趣相似度越高(相似度大于 0.3)，那么用户越有可能转发该微博；如果相似度很小甚至为 0，用户仍然有一定的概率转发该微博. 因此，我们可以认为微博内容与用户兴趣偏好的相似度能够用于预测他们的转发行为.

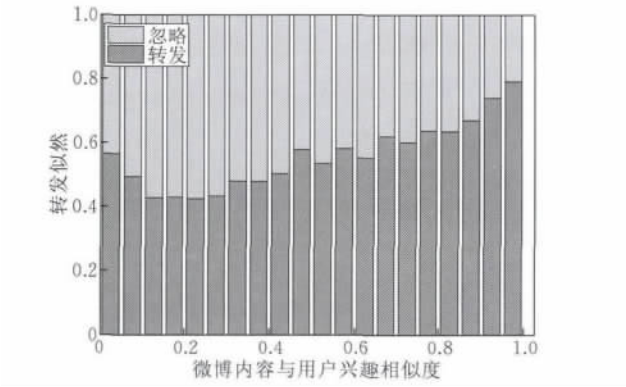

图 16　内容相似度与转发似然之间的关系

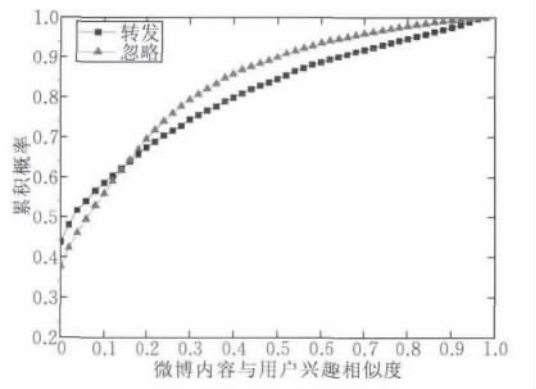

图 17　内容相似度在转发与忽略上的分布

① http://zh.wikipedia.org/wiki/TF-IDF
② http://www.ictclas.org

4.6 与转发路径上游用户间的兴趣相似度

考虑到用户的行为也受到具有相同兴趣的传递者的影响，我们分析了用户与相邻传递者、微博原创者之间的兴趣相似度对转发的影响.

从图 18 和图 19 可以看出，用户间兴趣相似度与转发似然之间的关系和微博内容相似度与转发似然之间的关系较为相似，用户之间的兴趣相似度越高，转发概率也就越大. 特别的，当用户之间的兴趣相似度比较小时，仍然有一定的概率转发. 这也比较符合人们的猎奇心态，当看到比较新奇的内容时，往往倾向于转发，分享给自己的粉丝.

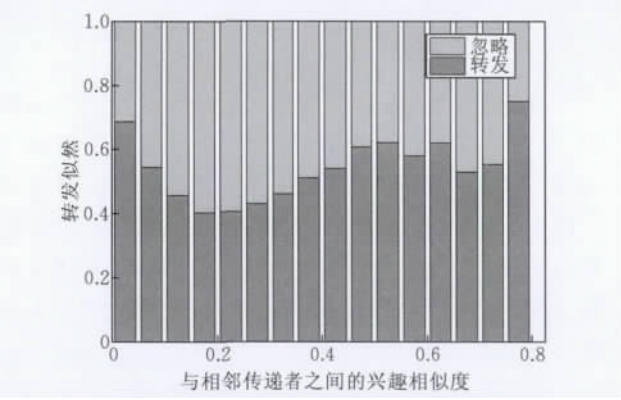

图 18 和相邻传递者之间的兴趣相似度与转发似然之间的关系

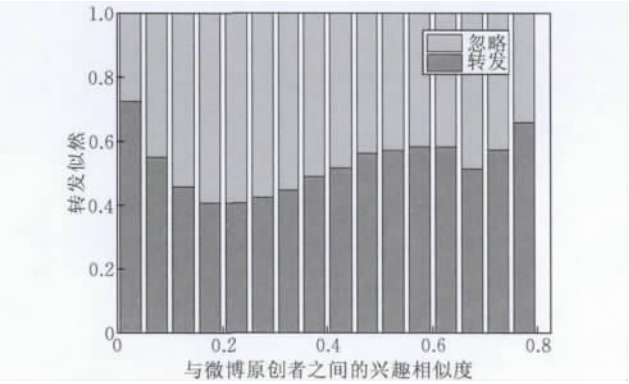

图 19 和微博原创者之间的相似度与转发似然之间的关系

除了以上这些特征，我们还考虑了其他一些特征，例如原创微博的发布时间，用户每条微博的平均转发率，原创以及上游用户是否是认证用户，原创微博的文本长度等.

我们将所有的特征分为 3 类：

(1) 用户特征. 该类特征只与用户个体相关，可独立计算. 例如用户粉丝数、用户 PageRank 值、是否是认证用户等.

(2) 社交特征. 该类特征主要体现两个用户之间的社交特性，例如交互强度、用户之间的兴趣相似度等.

(3) 微博特征. 该类特征从微博文本中提取，包括内容相似度、发表时间等.

表 1 给出了所有的特征以及其说明，其中 1～7 为用户特征、8～11 为社交特征、12～15 为微博特征.

表 1 选择的特征及说明

序号	特征名称
1	上游用户粉丝数
2	上游用户 PageRank 值
3	待预测用户转发活跃度
4	上游用户每条微博平均被转发次数
5	上游用户每条微博平均转发率(转发率:转发次数/粉丝数目)
6	上游用户是否是认证用户
7	微博原创者是否是认证用户
8	与上游用户之间的交互强度
9	与上游用户之间的兴趣相似度
10	与微博原创者之间的兴趣相似度
11	微博内容与被预测用户之间的兴趣相似度
12	微博内容的长度
13	原创微博的发表时间段(范围是 0～23)
14	上跳微博的发表时间段(范围是 0～23)
15	微博中所有词语的 TF-IDF 和

5 用户转发行为预测

在本节中，我们将通过监督学习框架刻画本问题，并使用多种算法对用户的转发行为进行预测.

5.1 问题描述

对于用户转发行为预测的问题描述如下：给定微博关注网络 $G\langle U,E\rangle$，G 是有向网络，U 是网络中所有用户的集合，E 是关注网络中关注关系的集合. 用户 u_{origin} 在时刻 t 发布或者转发一条微博 m，该条微博经过的传播路径记为 $P\langle u_{\text{origin}},u_1,\cdots,u_n\rangle$，$n$ 为传播路径的长度，若用户 u 关注了传播用户 u_n，预测用户 u 是否会转发该微博或者转发该微博的概率 p.

5.2 实验设置

(1) 数据集. 对于转发行为预测问题，我们从爬取的数据中共提取出 439 607 个转发样本，203 156 个忽略样本. 在预测过程中，采用十折交叉验证，将全部数据划分为训练集和测试集. 对于传播路径预测问题，我们从数据集中共提取了 12 284 条转发路径作为测试集.

(2) 预测方法. 我们选择逻辑回归、朴素贝叶斯以及贝叶斯网络等方法对该问题进行求解.

在贝叶斯方法中，需要估算每个因子的类条件概率密度. 图 20 和图 21 是与上游用户话题相似度

特征在转发和忽略行为中的类条件概率密度分布，从图上我们无法判断它们的类条件概率密度函数形式，因此采用 Parzon Window 进行非参数估计.

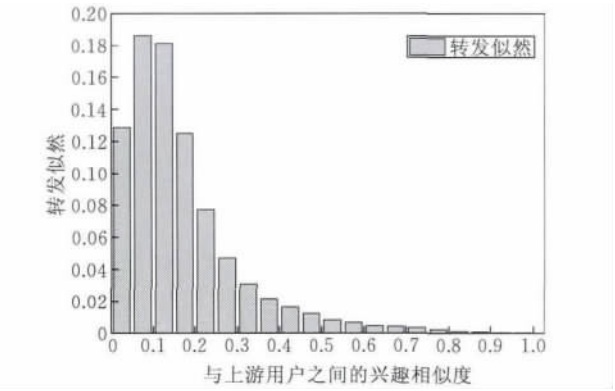

图 20　与上游用户相似度在转发类中的概率密度分布

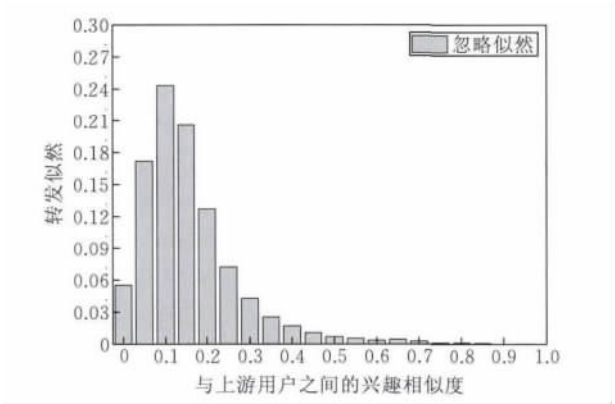

图 21　与上游用户相似度在忽略类中的概率密度分布

非参数方法的优势在于能够处理任意形式的密度函数，不必假设密度函数的参数形式，但是需要的样本数量要远多于参数方法. 由于我们拥有足够多的样本，因此采用非参数方法能够很好地利用已知样本对总体分布密度函数进行估计. 具体计算方法如下：

$$p_f(x|C)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\varphi\left(\frac{x-x_i}{h}\right) \tag{7}$$

其中 $p(x|C)$ 是所要估计的因子 f 在转发或忽略类 $C\in\{repost,ignore\}$ 中的条件概率密度，n 为相应类中的样本个数，h 是窗口宽度，φ 是窗口函数，我们采用标准正态分布作为窗口函数.

在贝叶斯网络中，我们使用 K2 方法学习贝叶斯网络结构.

6　实验结果与分析

6.1　评价指标

预测结果以混淆矩阵的形式表示(图 22). 为了评价预测模型的效果，我们选用信息检索的评价指标，包括查准率、查全率和 $F1$ 度量. 查准率指一类中被正确预测的微博占预测为该类微博的比例，例如，被转发的微博这一类中，准确率为 $a/(a+c)$，查全率为一类中被正确预测的微博占该类实际的全部微博的比例，例如还是被转发的微博这一类中，查全率为 $a/(a+b)$，$F1$ 度量是一个综合指标，可以用来同时描述查准率和查全率，计算方式如式(8)所示.

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{8}$$

		预测	
		被转发	未被转发
实际	被转发	a	b
	未被转发	c	d

图 22　混淆矩阵

我们还分析了各类特征对于预测的结果的影响程度，评价指标采用 ROC(Receiver Operating Characteristic，受试者工作特征曲线，又称感受曲线). ROC 以真正类率为纵坐标，以负正类率为横坐标. 传统的评价方法有一个共同特点，必须将预测结果分为两类，再进行统计. ROC 曲线的评价方法与传统的评价方法不同，没有这个限制，而是根据实际情况，允许有中间状态，可以把结果划分为多个有序分类. ROC 曲线越靠近左上角代表预测方法的效果越好，也可通过 ROC 曲线下方的面积(AUC)大小进行比较，AUC 越大，说明预测方法的效果越好.

6.2　转发行为预测结果

转发行为预测结果如表 2 所示. 从表 2 可以看出，使用贝叶斯网络方法进行预测的效果最好. 朴素贝叶斯是基于因子互相独立的假设，而实际上影响因子之间往往存在着一定的联系，图 23 展示了微博原创者 PageRank 值与用户粉丝数之间的关系.

表 2　用户转发行为预测结果

方法	查全率	查准率	$F1$ 度量
贝叶斯网络	0.762	0.761	0.761
朴素贝叶斯	0.721	0.716	0.715
逻辑回归	0.671	0.667	0.667

为了定量计算用户 PageRank 值和粉丝数之间的相关性，我们使用式(9)来计算二者的相关系数.

$$R(X,Y)=\left|\frac{Cov(X,Y)}{\sqrt{Var(X)\times Var(Y)}}\right| \tag{9}$$

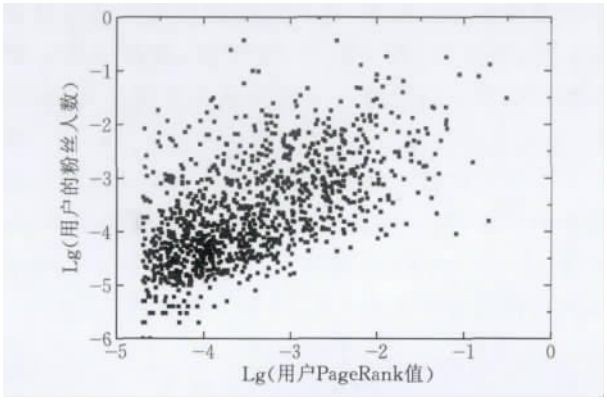

图 23 用户 PageRank 与规格化后的粉丝数之间的关系

其中，$Cov(X,Y)$是两个特征的协方差，定义如下：

$$Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{u}_x)(y_i-\bar{u}_y) \quad (10)$$

$R(X,Y)$是介于 0 和 1 之间的数，值越大则两个特征越相关. 若为 0，则说明 X,Y 不相关；为 1，则说明 X,Y 线性相关.

我们抽取了部分用户作为样本，提取出他们的 PageRank 值和规格化后的粉丝数，计算出两者的相关系数为 0.575，说明用户 PageRank 值与粉丝数这两个特征存在着较强的相关性.

如图 23 所示，一些特征之间存在着某些联系，由于这种特征之间的相关性，研究某种类型的特征而不是每一个具体的特征对于用户转发行为的影响更有意义. 因此，本文分别利用三类特征对转发行为进行预测，以比较不同类别特征在转发预测中的影响. 预测结果如表 3～表 5 所示.

表 3 使用用户类特征预测转发行为的预测结果

方法	查全率	查准率	$F1$ 度量
贝叶斯网络	0.647	0.757	0.698
朴素贝叶斯	0.603	0.795	0.686
逻辑回归	0.612	0.647	0.629

表 4 使用社交类特征预测转发行为的预测结果

方法	查全率	查准率	$F1$ 度量
贝叶斯网络	0.737	0.648	0.690
朴素贝叶斯	0.726	0.606	0.661
逻辑回归	0.726	0.507	0.579

表 5 使用微博类特征预测转发行为的预测结果

方法	查全率	查准率	$F1$ 度量
贝叶斯网络	0.581	0.622	0.619
朴素贝叶斯	0.579	0.521	0.549
逻辑回归	0.549	0.725	0.625

图 24 画出分别使用三类因子预测用户转发行为的 ROC 曲线. 从图中我们可以看出，对用户转发行为影响最大的是社交类特征，而微博本身的特征对于转发行为的影响在三个类中是最小的. 这充分说明了微博是一个社交平台，用户的社交需求远远高于内容需求.

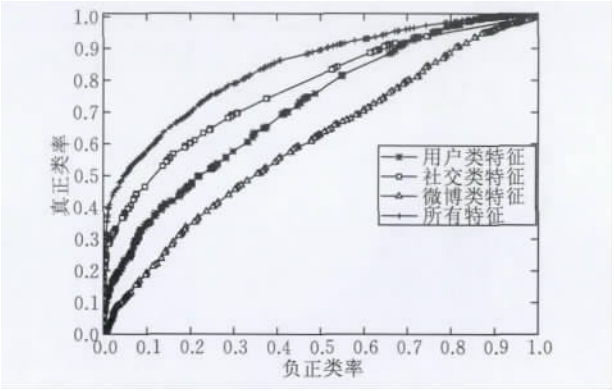

图 24 三类特征的 ROC 曲线

6.3 传播路径预测结果

由于数据抓取阶段我们按照转发网络进行广度优先遍历，因此数据集中保留了比较完整的转发路径，从中提取出了 12284 条传播路径，我们将长度为 n 的传播路径表示为$\langle origin_mid, mid_1, mid_2, \cdots, mid_n, ignore_uid\rangle$，其中 $origin\_mid$ 为原创微博的 ID，mid_i 为转发微博的 ID，$ignore\_uid$ 忽略该微博的用户 ID，也就是说微博经过上游用户的转发传播到该用户，但该用户并没有转发，因此采用该方法可以表示一条完整的传播路径.

我们采用级联预测的方法对传播路径进行预测，将路径预测精度作为评价指标. 成功预测一条传播路径是指对于路径上的每个转发行为都成功预测，并且对最后用户的忽略行为也做出正确预测. 预测结果如表 6 所示.

表 6 路径预测结果

路径长度	被正确预测的路径数目	路径总数	预测精度
2	4630	10849	0.4268
3	417	1198	0.3481
4	70	192	0.3646
5	15	37	0.4054

本文与现有相关研究进行了对比，文献[22]将特征分成社交特征和微博特征两类，发现社交特征能更好地预测用户转发行为，与本文的结论一致，但由于选择的特征较少，且量化方式较简单，预测精度并不高. 本文通过 3 种方法对转发行为预测均取得了较高的查全率、查准率和 $F1$ 度量，并在此基础上采用级联方式对传播路径进行预测分析.

7 总结与展望

本文主要考察了与用户转发行为有关的若干因

素，并分析了这些因素对用户转发行为的影响，利用机器学习中的分类算法来预测一条微博是否会被转发，最后利用对单跳的转发预测来预测微博的传播路径. 实验表明 76.2%的微博能被正确预测，对于较短的转发路径预测精度能达到 40%左右.

然而，本文仍然存在一些需要改进的地方. 例如，我们考虑的特征还不够全面，随着新浪微博的发展以及互联网越来越开放，我们可以获取到更丰富的用户属性，这些都可以用来提高预测精度. 另外，新浪微博上有不少的“僵尸粉”、“水军”，这些用户的转发行为往往与正常用户不同，然而在本文中并没有识别并剔除这类用户，对预测精度也造成了一定的影响，如果能够从数据中识别出这类用户并删除他们的行为记录，对提高预测精度也会有所帮助.

参 考 文 献

[1] Boyd D, Ellison N B. Social network sites: Definition history and scholarship. Journal of Computer Mediated Communication, 2007, 13(1): 210-230

[2] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter//Proceedings of the Hawaii International Conference on System Sciences. Hawaii, USA, 2010: 1-10

[3] Kwak H, Lee C, Park H, Moon S B. What is Twitter, a social network or a news media?//Proceedings of the World Wide Web Conference. Raleigh NC, USA, 2010: 591-600

[4] Suh B, Hong L, Pirolli P, Chi E H. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network//Proceedings of the IEEE International Conference on Social Computing-SocialCom. Palo Alto, USA, 2010: 177-184

[5] Zaman T R, Herbrich R, van Gael J, Stern D. Predicting information spreading in Twitter//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2010, 104(45): 598-601

[6] Stern D, Herbrich R, Graepel T. Matchbox: Large scale online Bayesian recommendations//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 111-120

[7] Yang Zi, Guo Jingyi, Cai Keke, et al. Understanding retweeting behaviors in social networks//Proceedings of the 19th International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1633-1636

[8] Yang Zi. Predictive models in social network analysis[M. S. dissertation]. Tsinghua University, Beijing, 2011(in Chinese)
(杨子. 社会网络分析中的预测模型[硕士学位论文]. 清华大学, 北京, 2011)

[9] Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet chain-letter data. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(12): 4633-4638

[10] Kossinets G, Kleinberg J, Watts D. The structure of information pathways in a social communication network//Proceedings of the 14th ACM SIGKDD. New York, USA, 2008: 435-443

[11] Zhou Tao, Fu Zhong-Qian, Niu Yong-Wei, et al. Research on spreading dynamics on complex networks. Progress in Natural Science, 2005, 15(5): 513-518(in Chinese)
(周涛, 傅忠谦, 牛永伟等. 复杂网络上传播动力学研究综述. 自然科学进展, 2005, 15(5): 513-518)

[12] Cha M, Mislove A, Adams B, Gummadi K P. Characterizing social cascades in Flickr//Proceedings of the 1st Workshop on Online Social Networks. New York, USA, 2008: 13-18

[13] Leskovec J, McGlohon M, Faloutsos C, et al. Cascading behavior in large blog graphs//Proceedings of the 7th SIAM International Conference on Data Mining. Minnesota, USA, 2007: 101-121

[14] Wu F, Huberman B A, Adamic L A, Tyler J R. Information flow in social groups. Physica A, 2004, 337(1): 327-335

[15] Szabó G, Huberman B A. Predicting the popularity of online content. Communications of the ACM, 2008, 53(8): 80-88

[16] Fan Pengyi, Li Pei, Jiang Zhihong, et al. Measurement and analysis of topology and information propagation on Sina-Microblog//Proceedings of the Intelligence and Security Informatics. Beijing, China, 2011: 369-401

[17] Webberley W, Allen S, Whitaker R. Retweeting: A study of message-forwarding in Twitter//Proceedings of the 1st IEEE NSS Workshop on Mobile and Online Social Networks. Milan, Italy, 2011: 13-18

[18] Galuba W, Aberer K, Chakraborty D, et al. Outtweeting the Twitterers: Predicting information cascades in Microblogs //Proceedings of the 3rd Workshop on Online Social Networks. Berkeley, USA, 2010: 3-3

[19] Yang Jiang, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, USA, 2010: 355-358

[20] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web. Stanford University: Technical Report SIDL-WP-1999-0120, 1999

[21] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613-620

[22] Petrovi S, Osborne M, Osborne M. RT to Win! Predicting message propagation in Twitter//Proceedings of the 25th Conference on Artificial Intelligence. San Francisco, USA, 2011: 508-511

CAO Jiu-Xin, born in 1967, Ph.D., professor, Ph.D. supervisor. His research interests include service computing, network security and social computing.

WU Jiang-Lin, born in 1988, M.S. candidate. His research interest is social computing.

SHI Wei, born in 1987, M.S. candidate. His research interest is social computing.

LIU Bo, born in 1975, Ph.D., associate professor. Her research interests include pervasive computing and social computing.

ZHENG Xiao, born in 1975, Ph.D., professor. His research interests include service computing and wireless local area network.

LUO Jun-Zhou, born in 1960, Ph.D., professor, Ph.D. supervisor. His research interests include next-generation network architecture, protocol engineering, network security, grid and cloud computing, and wireless local area network.

Background

In order to better understand the regularity of information diffusion in social networks, we study the Sina microblog based on a large-scale statistical analysis. Retweeting is a key mechanism for information diffusion in online social networks, and it is the mechanism of retweeting that leads to fast and wide diffusion for information in microblog. We study the key factors which affect users' retweet behavior and give their formal descriptions. Three comprehensive features, based on user attributes, social relations and microblog contents, are used to predict users' retweet behavior by machine learning classification algorithms. The microblog topology graph on following relation is constructed, and the cascade probability model is used to predict the propagation paths of a tweet, then a tweet's influence can be predicted.

This work is supported by National Basic Research Program (973 Program) of China under Grant No. 2010CB328104, National Natural Science Foundation of China under Grant Nos. 61272531, 61202449, 61272054, 61370207, 61370208, 61300024, 61320106007, National High Technology Research and Development Program (863 Program) of China under Grant No. 2013AA013503, China Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 2011009213002, Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No. BM2003201 and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No. 93K-9.