# 基于用户行为特征的微博转发预测研究

刘 玮[1),2),3)] 贺 敏[1),2),3)] 王丽宏[3)] 刘 悦[1)] 沈华伟[1)] 程学旗[1)]

[1)](中国科学院计算技术研究所 北京 100190)

[2)](中国科学院大学 北京 100049)

[3)](国家计算机网络应急技术处理协调中心 北京 100029)

**摘 要** 微博转发预测对微博话题检测和微博影响力评估具有重要意义,引起了学界和产业界的广泛关注.现有方法主要集中在微博属性及微博传播网络特征的研究,没有充分考虑转发行为的动态性和用户历史行为的规律性.文中从微博能见度和用户行为特征角度研究微博转发预测问题,(1)提出了基于用户活跃期和时间窗的转发行为、忽略行为、未接收行为识别方法,为模型训练和效果分析提供了更为准确的数据基础;(2)提出了基于时间衰减的用户兴趣计算模型,有效度量用户兴趣及其变化特性对用户转发行为的影响程度;(3)提出了用户转发率、交互频率等用户行为特征,有效度量了用户历史行为模式和用户影响力传递效应的差异性对用户转发行为的影响,最后融合上游用户特征、微博特征、转发用户兴趣和历史行为特征,提出了基于分类模型的转发行为预测方法.在真实数据上的实验结果表明,该方法能够有效提升预测准确性,并且能够在较小规模的训练集上取得好的预测效果.

**关键词** 转发预测;微博能见度;时间衰减;交互频率;历史行为;社交网络;社会媒体

**中图法分类号** TP18 **DOI 号** 10.11897/SP.J.1016.2016.01992

# Research on Microblog Retweeting Prediction Based on User Behavior Features

LIU Wei[1),2),3)] HE Min[1),2),3)] WANG Li-Hong[3)] LIU Yue[1)] SHEN Hua-Wei[1)] CHENG Xue-Qi[1)]

[1)](*Institute of Computing Technology, Chinese Academy of Sciencess, Beijing* 100190)

[2)](*Graduate University of Chinese Academy of Sciences, Beijing* 100049)

[3)](*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029)

**Abstract** Retweeting prediction is of great importance to the event detection and influence evaluation, which has attracted wide attention from both academic and industrial fields. Existing prediction methods are mostly concentrated in the study of the microblogs properties and diffusion network characteristics. They have not fully considered the dynamics of retweeting behavior and regularity of user's historical behavior. This paper investigated the microblog retweeting prediction problem from the view of microblog visibility and user behavior features, and (1) proposed a method to recognize retweeting behavior, ignoring behavior and un-received behavior based on user's activity and dynamic time window, which provided more accurate dataset for model training and effectiveness analysis; (2) presented user interest model based on dynamic user's interest and time attenuation, which is proved to be an effective measurement of user's interest and its change characteristics; (3) proposed several user behavior features of the user's retweeting rate and interaction frequency, which can effectively measure the impact of user's historical behavior patterns and user's influence

收稿日期:2015-09-15;在线出版日期:2016-02-16.本课题得到国家自然科学基金(61170230)、国家科技支撑计划项目(2012BAH46B01)、国家"八六三"高技术研究发展计划项目基金(SS2014AA012303)资助.**刘 玮**,女,1984 年生,博士研究生,高级工程师,中国计算机学会(CCF)会员,主要研究方向为社交网络数据挖掘、网络信息安全、信息过滤.E-mail:liuwei@isc.org.cn.**贺 敏**,女,1982 年生,博士研究生,高级工程师,主要研究方向为社交网络数据挖掘、话题发现.**王丽宏**,女,1967 年生,博士,教授级高级工程师,博士生导师,主要研究领域为网络信息安全、社交网络数据挖掘、舆情处理.**刘 悦**,女,1971 年生,博士,副研究员,主要研究方向为信息检索、社会计算.**沈华伟**,男,1982 年生,博士,副研究员,主要研究方向为社会计算、复杂网络、信息检索.**程学旗**,男,1971 年生,博士,研究员,博士生导师,计算机学会会员,主要研究领域为网络科学、网络信息安全、互联网数据挖掘.

transfer effect. Finally, this paper proposed a classification model based on retweeting behavior prediction method which is blend with upstream user's characteristics, microblog's characteristics, forwarding user's interest and user's historical behavior characteristics. Experimental results on real data show that the proposed method can improve prediction accuracy effectively, and achieve good results in the smaller size of the training set.

**Keywords**　retweeting prediction; microblog visibility; time attenuation; interaction frequency; historical behavior; social networking; social media

## 1　引　言

我国微博应用是人们表达观点和传递信息的重要社会媒体，截至 2015 年 6 月，我国微博用户规模为 2.04 亿[1]，用户之间结成复杂的关注关系，信息沿着用户间的关注关系进行传播，形成传播扩散网络. 微博 140 字的内容限制、“//@、#”符号标记语言以及单向认证的“弱关系”等机制，使得微博在内容生成方式、用户参与的广泛性和即时性、信息扩散模式和速度等方面均不同于新闻、论坛、博客等传播网络媒体[2]，微博能够更真实、全面、快捷地表达现实世界的社会关系和社会生活[3-5]. 微博转发是消息在微博网络中得到持续传播的重要方式，微博转发预测能够估计消息是否能获得转发及其转发规模，及早发现可能引发大规模爆发的微博，对微博话题检测、微博影响力评估、舆情监控及网络营销都具有重要价值.

基于微博特征的微博转发预测主要是针对某些消息具有更高的转发性这一现象，通过用户静态属性或消息本身特征来进行转发行为预测. 基于高转发率微博属性特征的预测方法没有充分考虑待预测用户的个体差异性和个人兴趣对转发决策的影响. 事实上，用户在阅读到一条微博时，会根据兴趣对微博价值和新颖性进行判断，然后决定是否进行转发. 用户兴趣可以从用户历史所发微博中分析获得，但通过用户历史微博来获取用户兴趣并进行微博转发预测具有时间局限性和内容局限性，时间局限性是指用户兴趣会随时间推移而变化，内容局限性是指微博只是用户进行网络交互的一种方式，无法完整反映用户所有兴趣. 所以，用户兴趣的计算要考虑兴趣的变化性和获取方式的多样性. 并且仅通过用户兴趣与微博的相似程度判断用户是否会转发某条微博是不准确的. 针对时间局限性，本文通过将用户所发微博的时间因素引入用户兴趣模型，从而更加真实地反映用户当前兴趣特征. 针对内容局限性，本文一方面融合标签和历史发布微博综合计算用户兴趣，另一方面，通过将用户转发行为特征和与关注用户的交互特征引入用户行为预测模型，从而更加全面的刻画影响用户转发行为的各类因素.

基于微博网络结构特征的微博转发预测主要是通过构建用户关注网络、用户转发网络或消息转发树，以关注、转发关系为连边，用户为节点，将用户转发和不转发作为两种节点状态，基于信息传播模型建立微博转发模型，预测用户转发行为，进而预测信息传播速度和规模. 该类方法需要建立完整的转发网络、前一时刻节点状态以及前后两个时刻的邻居节点状态，这需要获得完整的转发关系和历史转发日志数据，但是在实际转发预测问题中，网络节点规模巨大且存在大量不活跃节点，大部分情况下只能获取到部分用户转发数据和局部日志数据，建立完整的转发网络和节点状态是非常困难的，并且计算复杂度非常高. 本文主要针对基于微博和用户特征的转发预测方法开展研究和实验验证.

以新浪微博为代表的在线社交网络在中国广泛使用，掌握新浪微博用户的转发行为模式对分析国内社交网络信息发布和传播过程具有重要意义. 新浪微博与 Twitter 等国外社交网络存在诸多差异性，针对 Twitter 网络提出的消息转发预测方法难以很好的满足新浪微博网络中的转发预测需求. 樊鹏翼等人[6]通过对新浪微博开展网络测量，发现新浪微博网络具有小世界特性并且聚集系数高于 Twitter，说明其具有更紧密的网络结构，更利于消息的传播. 其次，由于新浪微博的关注上限限制，新浪微博网络的出入度不具有相关性. 第三，新浪微博中约有一半的用户对整个网络的信息传播贡献为零，且相比于原创微博，用户更倾向于对微博进行转发. 所以，新浪微博网络与 Twitter 网络在用户行为、网络结构和用户交互性方面存在差异性，需要研究针对新浪微博的转发行为预测方法.

微博转发预测的一个关键问题是如何准确判断用户转发微博和不转发微博，并将用户转发行为特

征引入到转发预测模型中.基于分类模型的转发预测方法在进行模型训练时,大多以显式的用户转发行为和不转发行为为正负样本,并未深入考虑其不转发样本是否出现在用户可见范围内,是用户真正判断为不转发的结果.毛佳昕等人[7]在研究用户社会影响力时通过构建用户阅读行为模型,考察了样本是否可见的因素对于用户转发行为的影响,但其行为模型是基于24小时的发帖行为建立的,以用户每天的发帖行为不变的假设为前提,并且没有充分考虑用户关注对象和微博内容的不同对用户转发行为的影响.

微博是否会被转发与待预测用户的个体行为具有紧密相关性,本文从微博对用户的能见度、用户兴趣和历史行为角度开展研究,主要贡献包括以下4个方面:(1)提出基于用户活跃期和动态时间窗的转发行为、忽略行为、未接收行为识别方法,构建了准确的模型训练数据集;(2)提出基于时间衰减的用户兴趣计算模型,有效度量了用户兴趣及其变化特性对用户转发行为的影响程度;(3)提出用户活跃度、转发率、交互频率等用户行为特征,有效度量了用户历史行为模式、用户影响力传递效应的差异性对用户转发行为的影响;(4)最后提出基于分类模型的转发行为预测方法,在真实数据上开展实验,验证了本文所提方法的有效性.

本文第1节引言部分介绍问题背景和研究现状;第2节介绍相关工作;第3节是问题描述和研究对象;第4节介绍忽略样本识别方法;第5节介绍用户转发特征分析和计算方法;第6节介绍实验结果和分析;第7节是总结和下一步工作.

## 2 相关工作

针对微博转发预测的代表性研究包括,Suh等人[8]针对某些消息具有更高的转发性这一现象,基于Twitter数据研究了多种微博转发的影响因素,建立的特征空间包括URL、标签、关注人数、粉丝人数等,通过主成分分析和广义线性模型的分析方法,研究各影响因素与微博转发之间的关系,结果表明,是否含有URL和标签以及粉丝关注数特征对转发影响较大.张旸等人[9]通过分析多种用户特征和文本特征在转发微博和不转发微博中的区分度,提出了特征加权预测模型,将转发预测问题转化为二类分类问题,并且采用信息增益法对各特征的重要性进行了分析,结果表明"用户粉丝数"和"用户被提及数"重要性较高.Yang等人[10]根据转发关系建立微博转发树,通过截取Twitter消息中的RT@username提取微博转发关系建立微博转发树,然后基于因子图模型建立转发预测模型,以用户为节点,转发关系为连边,将用户转发和不转发作为两种节点状态,该模型将节点属性、前一时刻节点状态以及前后两个时刻的邻居节点状态作为训练数据获得模型参数,然后进行节点状态预测.但是在实际转发预测问题中,建立完整的转发树是很困难的,且计算复杂度高.曹玖新等人[11]将微博转发的影响因素分为用户特征、社交特征和微博特征,通过分类模型研究这些特征对微博转发的影响,并利用单跳转发的预测来预测转发路径形成的概率.Wu等人[12]对Twitter用户行为数据进行了分析,研究发现用户同配性概率、二步传播行为比重以及推文的寿命根据用户类型不同而表现出巨大差异.Tan和Wang等人[13-14]提出NTT-FGM模型来预测用户行为,定义并计算行为偏好、朋友影响和自相关性及其对用户行为的影响大小,进而通过求解条件概率问题来实现用户转发行为预测.上述研究主要针对用户静态属性或消息本身特征分析消息是否会被转发,没有充分考虑转发行为的动态性和用户历史行为规律对用户转发行为的影响.

## 3 问题描述和研究对象

### 3.1 问题描述与相关概念

(1)关注网络

用有向无权图 $G=(V,E)$ 表示用户关注网络,节点 $u_i \in V$ 表示网络中的第 $i$ 个用户,边 $e_{ij} \in E$ 表示用户 $i$ 和 $j$ 存在一条关注关系,即用户 $i$ 关注了用户 $j$.

(2)转发网络

用有向有权图 $\hat{G}=(\hat{V},\hat{E},W)$ 表示用户转发网络,节点 $u_i \in V$ 表示网络中的第 $i$ 个用户,边 $\hat{e}_{ij} \in \hat{E}$ 表示用户 $i$ 和 $j$ 存在转发关系,即用户 $i$ 转发了用户 $j$ 的微博,$w_{ij} \in R$ 表示用户 $i$ 和 $j$ 连边的权重,本文定义为用户 $i$ 转发用户 $j$ 的微博数量.

(3)上游用户、转发用户

如果一条微博的转发路径用 $(u_0,\cdots,u_k,u_{k+1},\cdots,u_n)$ 表示,则 $u_k$ 在该条信息的转发路径上是 $u_{k+1}$ 的上游用户,$u_{k+1}$ 是转发用户.

转发预测是研究以社交网络为基本网络结构、以由用户发起的、经关注关系传递而形成的消息传播过程,这里假设信息仅通过关注网络传递.转发预

测针对信息传播的上游用户属性、微博内容属性、转发用户与上游用户的交互性属性等，研究用户间影响、用户行为模式的动态性、微博内容与用户兴趣相似性等因素对用户转发行为的影响.

若用户 $u_i$ 发布了一条微博信息 $m$，$y=f(u_i,u_j,m)$ 表示 $u_i$ 的粉丝 $u_j$ 是否会转发 $m$，当 $y=+1$ 表示转发，当 $y=-1$ 表示不转发. 转发预测问题的形式化描述为：在已知 $u_i$、$u_j$、$m$ 的条件下，寻找目标函数 $y=f(u_i,u_j,m)$，$y\in\{+1,-1\}$，将 $u_i$、$u_j$、$m$ 映射到两个类别中，使得分类准确性最高.

### 3.2　研究对象

本文选取国内用户量最大的社交网络平台——新浪微博作为研究对象，考虑到新浪微博含有大量粉丝数少、活跃度低的账号，容易造成关注网络和转发网络过于庞大和稀疏，为了提高网络连通度，我们选取新浪微博中粉丝数最多的 70 万账号，采用“滚雪球”策略，采集了这些账号所发的微博消息、Profile 信息、关注关系等信息以及被这些账号转发过消息的账号和对应信息，经过垃圾微博去除等预处理过程后，构建了包括 600 万个微博账号、约 12 亿条微博消息的微博网络. 选取其中 2012 年 4 月 1 日至 4 月 30 日的微博进行标注，提取该段时间的发帖用户和转发用户及其个人信息，关注关系，最终构建局部转发预测实验数据集，包括 140 922 个用户，3 847 724 条微博，15 228 294 条关注关系. 如图 1 所示，用户粉丝数与发微博数的关系表明，用户发微博越多，获得的关注数也越多，即粉丝数越多.

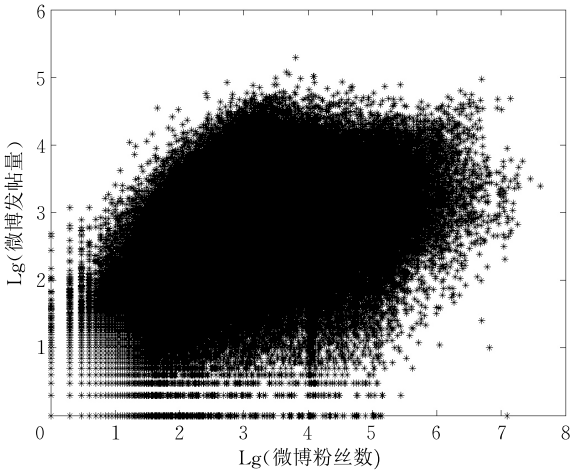


图 1　用户粉丝数与发微博数的关系

## 4　忽略样本识别方法

为了更准确的分析对用户转发微博起关键作用的影响因素，需要准确识别用户转发微博和不转发微博，对数据集中的用户微博对进行类别标注，将用户转发的微博标注为正样本，将用户不转发的微博标注为负样本. 正样本可以通过用户显式地转发标记识别，负样本的识别则由于没有显式标记而比较困难. 本节主要介绍基于动态时间窗的忽略行为识别方法.

转发微博能够代表用户感兴趣的微博，而不转发微博代表了用户不感兴趣的微博. 但微博空间用户数庞大，每天产生大量微博信息，如果用户关注人数较多，大量的信息将被淹没和冲逝，并不是每条微博都能在用户使用微博的时间区间内被用户看到，经过用户判断后再被决定转发还是忽略. 还没被用户看到就被冲逝的微博，我们称之为不可见微博. 要准确分析用户的转发行为，就必须将忽略微博与不可见微博区分开来，将用户真实的忽略微博作为负样本.

要准确区分哪些微博是用户看到但未转发的，哪些微博是用户没有看到的，需要掌握用户使用微博的时间区间，但是该信息无法通过网页爬取获取，我们只能通过爬取到的用户发帖时间推断用户访问微博的时间区间. 用户发帖行为包括原创和转发，我们验证用户转发微博时间分布和原创微博时间分布是否具有一致性. 我们统计了用户原创发帖数和所有发帖数(原创数加转帖数)在 1 天 24 小时内的发帖时间分布.

从图 2 可以看出，用户 24 小时发帖行为符合作息规律，每天 11 点、15 点、22 点用户发帖数量达到高峰. 上午 8 点开始，发帖数量逐渐上升，到 11 点午休前达到一次高峰. 下午 13 点开始，发帖数量逐渐上升，到 15 点达到第二次高峰. 晚上从 20 点开始，发帖数量逐渐上升，到 22 点达到第三次高峰. 用户原创发帖数量和全部发帖数量(原创数加转帖数)的

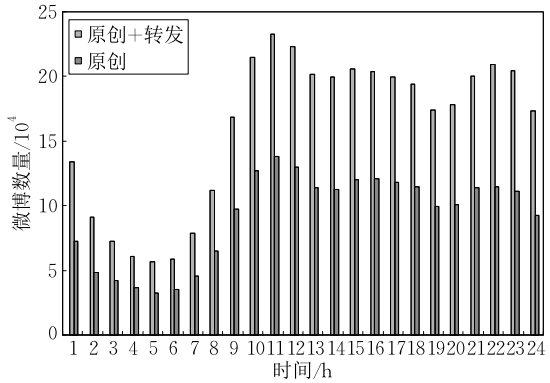


图 2　一天内发布和转发微博的时间分布

分布具有一致性，转发行为与原创发帖时间是一致的，所以我们使用用户总的发帖时间来代表用户使用微博的时间，进而计算用户访问微博的时间区间.

由于忽略微博难以显式地从数据集中提取，我们假设关注用户在用户使用微博的时间区间内发表的微博，用户 $u$ 才可以看到，进而根据个人偏好选择转发行为和忽略行为. 我们将忽略样本定义为：若用户 $u$ 在时刻 $t$ 发表了微博，表明用户 $u$ 在时刻 $t$ 在访问微博，则将他关注用户在 $[t-\Delta t_u, t+\Delta t_u]$ 时间区间内发表的且未被用户 $u$ 转发的微博称为忽略微博，即用户真正的不转发微博，时间窗口 $\Delta t_u$ 的大小决定了微博对用户可见概率的大小.

接下来我们计算用户使用微博的时间窗口. 用户 $u$ 将接收到的微博数量取决于其关注的用户发帖数量，关注的用户越多，接收到的微博数量也越多. 用户 $u$ 使用微博的时间是有限的，接收到的微博越多，每条微博能够被用户 $u$ 看到的停留时间就越短，距离用户发帖时间越远的微博，被用户看到的概率就越小，所以，用户 $u$ 关注的用户越多，关注用户所发微博刷新的速度越快，每条微博对用户 $u$ 的可见时长越短，也就是时间窗口 $\Delta t_u$ 越小.

为了给用户设定合适的时间窗口，我们统计了微博转发时延分布情况，如图 3 所示，20% 的用户转发行为发生在微博发出的 30 分钟，只有 20.2% 的转发行为发生在 24 小时以上. 我们据此设定时间窗口分为 15 分钟、30 分钟、10 小时、24 小时 4 个级别.

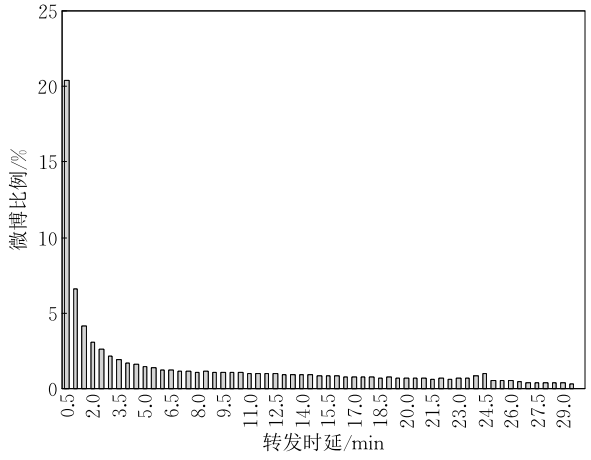


图 3 转发时延分布

我们根据关注用户数量划分用户的时间窗口，计算过程如下：首先统计用户 $u$ 的关注用户数量，然后基于分段函数计算微博停留的时间窗口：

$$\Delta t_u = 15\times 60\times \Theta(N_{\text{follow}}-100)+ 30\times 60\times \Theta(N_{\text{follow}}-50)\times \Theta(100-N_{\text{follow}})+10\times 60\times 60\times \Theta(N_{\text{follow}}-10)\times \Theta(50-N_{\text{follow}})+ 24\times 60\times 60\times \Theta(10-N_{\text{follow}}) \quad (1)$$

$N_{\text{follow}}$ 表示用户的关注好友数量，$\Theta(x)$ 定义如下：

$$\Theta(x)=\begin{cases}1, & x\geq 0\\ 0, & x<0\end{cases} \quad (2)$$

例如用户关注了 80 个好友，则其时间窗口 $\Delta t_u$ 为 30 分钟. 确定时间窗口后，进行忽略行为的识别，识别过程如图 4 所示，用户 $u$ 的关注用户所发微博按发表时间倒序排列，发表时间如图 4 所示，第 $j+2$、$j+4$ 微博被用户 $u$ 转发，转发时间分别是 5:52 和 6:23，依据式(1)时间窗口 $\Delta t_u$ 为 30 分钟，则用户访问微博的时间区间为 [5:22, 6:22]，[5:53, 6:53]，落入时间区间但未被转发的微博为 $j+1$、$j+3$、$j+5$，该 3 条微博为用户 $u$ 的忽略微博即不转发微博，未落入时间区间的微博为第 $j$、$j+6$ 条，该 2 条微博为不可见微博，无法判断用户 $u$ 是否对微博内容是否感兴趣.

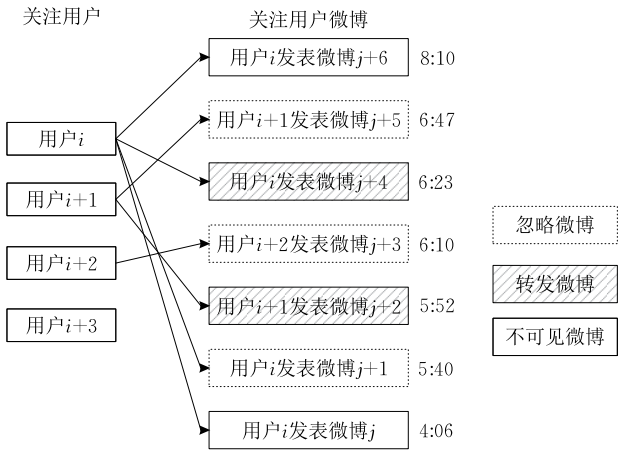


图 4 忽略微博样本识别

## 5 转发特征分析

### 5.1 用户行为特征

转发用户的行为特征是影响其是否转发微博的重要因素，用户一段时间内的发帖数量能够代表用户在社交网络中的活跃程度，相对于不活跃的用户，活跃用户在上线时间内，更倾向于显式的表现自己是否对某条微博感兴趣. 基于这一现象，我们计算了每个用户在一个月内的发帖数量来分析用户的活跃程度. 图 5 结果表明，社交网络中的活跃用户数量符合幂率分布，经常发微博的用户占比较小，大部分用户的发帖行为较少，用户使用微博的频率和时间受多种因素影响，具有一定随机性.

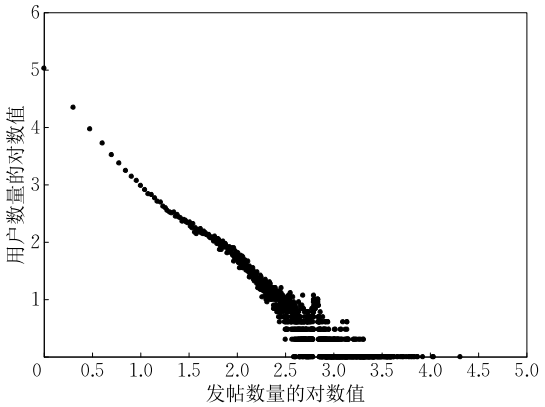


图 5　用户发帖数量分布

为了进一步说明用户发帖的稀疏性，我们统计了每月用户平均发帖数，结果如图 6 所示，91.7%的用户每月发帖数为 30 条，即平均 1 天发布 1 条微博，真实微博网络中的用户发帖非常稀疏，所以基于统计的用户行为建模方法在遇到用户发帖数据量较少的情况下准确性会受到影响. 通过发帖数量作为特征来预测用户的转发行为会由于大多用户发帖行为具有随机性和稀疏性而导致效果不佳. 所以我们提出将用户转帖数量与发帖数量的比值作为表征用户发帖行为的特征，将用户行为特征从发帖数量这一绝对值转变为发帖类型比例这一相对值.

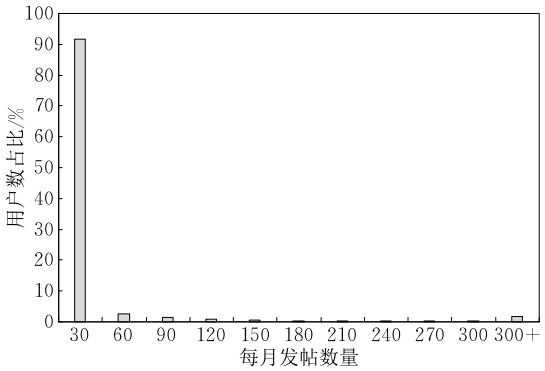


图 6　用户每月发帖数量分布

用户发帖数量中转帖的比例能够表明用户在参与微博网络时更倾向于采用原创发帖行为还是转帖行为. 我们定义转发概率是被转发微博数量和全部微博的比值，转发率是用户发帖数中转帖数的比例. 如图 7 所示，转发概率与转发率的关系显示出用户转发比例不同对应的转发概率不同，总体显示为转发率越高的用户转发微博的概率越大，转发率越低的用户转发微博的概率越小.

当转发率低于 0.5 时，转发概率平均为 40%，也就是说当一个用户的发帖中 50%以下是转贴时，在接收到一个新消息时，该用户只有 40%的概率会进行转发. 在实验部分的结果也表明，转发率特征能够提升转发预测准确性.

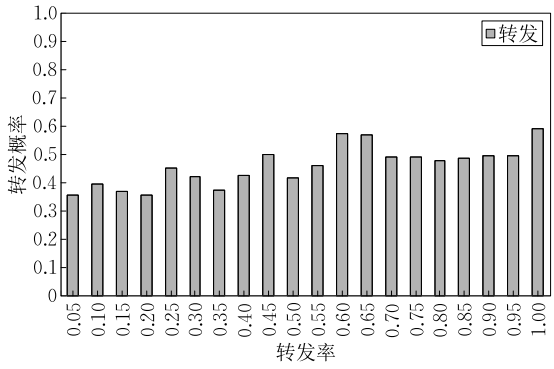


图 7　转发概率与转发率的关系

## 5.2　微博特征

微博内容是否能获得转发，能否取得较高的流行度很大程度上取决于微博内容所承载的信息量大小，以及微博内容情感因子大小[15]. 鉴于微博属流式短文本数据，信息产生和消亡速度快，且内容长度有限，我们采用实体词和情感词作为微博消息内容信息量大小的计算载体. 用实体词个数表示微博信息量大小的量化特征，包含的实体词越多，微博所含信息量越丰富，与用户关心内容或社会热点具有相关性的概率越大. 用情感词个数表示微博情感因子的量化特征，包含的情感词越多，表示微博含有更加强烈的情感因素，越容易吸引用户转发.

针对实体词提取，本文利用 ICTCLAS 分词软件对微博内容进行词性标注，提取类别为机构名、人名、地名、时间词的词，进而计算实体个数. 我们统计了含有不同实体词个数的微博所占比例，以及含有不同实体词个数的微博被转发的比例，如图 8 所示，柱状图表示不同实体词个数的微博所占比例，曲线图表示不同实体词个数的微博被转发的比例. 从柱

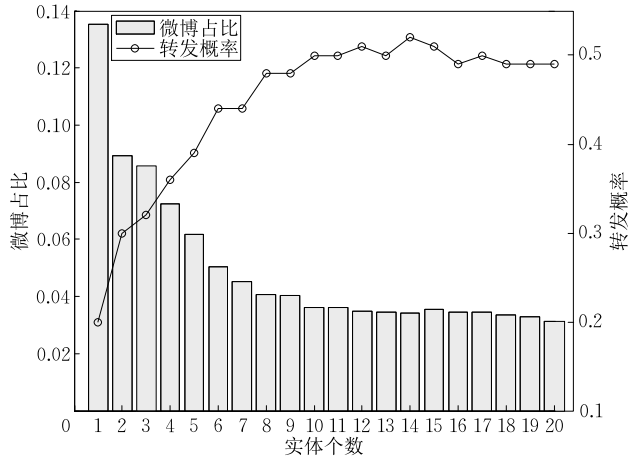


图 8　转发概率与实体个数分布的关系

状图我们可以看出微博占比随着实体词个数的增加而降低,大部分微博含有 5 个以下实体词,这与微博长度较短、实体词个数较少的现实情况是相符的.从曲线图我们可以看出,随着实体词个数的增加,转发概率逐渐增加,这是因为微博所含实体词个数越多,表达的内容越丰富,越容易引起用户的转发意愿,所以转发概率越大.

针对情感词提取,本文利用情感词词典、微博表情符号等特征,对微博进行了情感词提取,情感词词典采用知网发布的"情感分析用词语集(beta 版)"里中文情感分析用词语集中的中文正负情感词①,其中包含 836 个中文正面情感词语,1254 个中文负面情感词语,表 1 分别列举了 30 个中文正负面情感词.

**表 1 中文正负面情感词示例**

| 序号 | 中文负面情感词语 | 序号 | 中文正面情感词语 |
|---|---|---|---|
| 1 | 哀愁 | 1 | 爱不释手 |
| 2 | 哀怜 | 2 | 爱戴 |
| 3 | 哀悯 | 3 | 爱抚 |
| 4 | 哀伤 | 4 | 爱好 |
| 5 | 哀痛 | 5 | 爱护 |
| 6 | 哀怨 | 6 | 爱慕 |
| 7 | 哀恸 | 7 | 拜服 |
| 8 | 哀矜 | 8 | 表扬 |
| 9 | 懊悔 | 9 | 表彰 |
| 10 | 懊恼 | 10 | 称道 |
| 11 | 懊丧 | 11 | 称快 |
| 12 | 抱不平 | 12 | 称赏 |
| 13 | 抱憾 | 13 | 称颂 |
| 14 | 抱恨 | 14 | 称叹 |
| 15 | 抱怨 | 15 | 称羡 |
| 16 | 暴跳 | 16 | 称谢 |
| 17 | 悲哀 | 17 | 称心 |
| 18 | 悲悯 | 18 | 称许 |
| 19 | 悲戚 | 19 | 称誉 |
| 20 | 悲凄 | 20 | 称愿 |
| 21 | 悲伤 | 21 | 称赞 |
| 22 | 悲痛 | 22 | 崇拜 |
| 23 | 悲痛欲绝 | 23 | 崇敬 |
| 24 | 悲怆 | 24 | 崇尚 |
| 25 | 悲恸 | 25 | 崇仰 |
| 26 | 鄙视 | 26 | 答礼 |
| 27 | 鄙夷 | 27 | 答谢 |
| 28 | 鞭挞 | 28 | 电贺 |
| 29 | 贬斥 | 29 | 感恩 |
| 30 | 变心 | 30 | 感激 |

另外,在微博系统中含有丰富的表情符号,自由、便捷的消息发布特点也让用户能够更加方便地使用表情符号来表达自己的情感,所以可以将微博表情符号也作为提取微博情感的一个来源.我们抓取了新浪微博平台上的 300 个表情符号,通过人工标记区分出正面情感符号 218 个和负面情感符号 82 个,一个表情符号代表一个情感词语.表 2 分别列举了 20 个正负面表情符号.

**表 2 正负面情感表情符号表**

| 序号 | 负面情感表情符号 | 序号 | 正面情感表情符号 |
|---|---|---|---|
| 1 | [吐] | 1 | [狂笑] |
| 2 | [怒骂] | 2 | [哈哈] |
| 3 | [闭嘴] | 3 | [笑哈哈] |
| 4 | [抓狂] | 4 | [好激动] |
| 5 | [鄙视] | 5 | [好喜欢] |
| 6 | [伤心] | 6 | [乐乐] |
| 7 | [失望] | 7 | [太开心] |
| 8 | [悲催] | 8 | [偷乐] |
| 9 | [呜呜] | 9 | [偷笑] |
| 10 | [阴险] | 10 | [阳光] |
| 11 | [阴天] | 11 | [赞] |
| 12 | [生病] | 12 | [bm 可爱] |
| 13 | [咆哮] | 13 | [给力] |
| 14 | [泪流满面] | 14 | [可爱] |
| 15 | [泪] | 15 | [耶] |
| 16 | [讥笑] | 16 | [嘻嘻] |
| 17 | [惊恐] | 17 | [威武] |
| 18 | [din 抓狂] | 18 | [爱你] |
| 19 | [bm 生气] | 19 | [花心] |
| 20 | [bm 血泪] | 20 | [ppb 鼓掌] |

通过微博内容情感词提取和情感表情符号提取,我们统计了含有不同情感词个数的微博所占比例,以及含有不同情感词个数的微博被转发的比例.如图 9 所示,柱状图表示不同情感词个数的微博所占比例,曲线图表示不同情感词个数的微博被转发的比例.从柱状图我们可以看出微博比例随着情感词个数的增加而降低,80%的微博只含有一个情感词,这是由于微博长度限制,用户在一条微博中倾向于表达一种情感.从曲线图我们可以看出,随着情感词个数的增加,转发概率逐渐增加,这是因为情感词

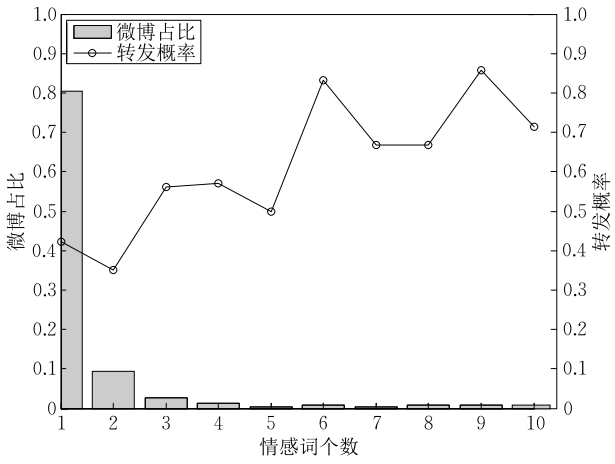


图 9 转发概率与情感词个数分布的关系

① 知网"情感分析用词语集(beta 版)". http://www.keenage.com/html/c_index.html

较多的微博更容易得到转发，其往往携带了更多衍生信息，情感表达方式也较为激烈，易于使得消息得到传递、发展和演化. 因此，针对微博内容，赋予消息内容一定情感词，有助于提高其传播概率.

### 5.3　用户兴趣特征

用户往往希望能看到并转发自己感兴趣的内容[16]，有效提取用户转发兴趣是提高转发预测准确性的关键问题之一. 在社交网络中，通过用户转发微博行为获取用户兴趣是一个具有挑战性的问题，第一，用户兴趣具有一定时间延续性，使得能够通过分析用户历史微博来提取用户兴趣，但是用户兴趣同时也具有时间衰减性，用户兴趣会随着多种因素变化而变化，所以越早的微博对提取用户兴趣的作用越小，在微博转发预测研究中，要重点考虑用户兴趣的动态变化特性. 第二，用户兴趣不仅反映在转发微博中，也体现在标签等用户属性中，而前者是具体化的内容，后者通常是概括的类别性描述，如何融合两者所含用户兴趣信息并综合计算用户兴趣是一个重要问题.

针对上述问题，本文提出基于线性加权的兴趣计算模型，融合用户标签信息和用户历史发布微博计算用户兴趣特征，然后计算用户兴趣与待预测微博之间的相似度即用户兴趣特征，用于进行用户转发预测.

由于用户兴趣标签词通常为概念性描述，而微博内容则为对象的具体性描述，所以无法直接对两者进行计算. 本文采用基于知网的词汇相似度计算方法[17]，通过标签词和微博关键词两两之间的词语距离，计算两者之间的相似度，如果两者所包含词语两两之间相似度之和大于预设阈值，则表示用户所发微博与用户标签内容相匹配，用户所发微博内容体现了标签所代表的兴趣，标签内容对用户兴趣的贡献由其匹配命中的用户发布微博内容替代，从而在计算用户兴趣时实现了对标签特征和用户所发微博特征的有效融合.

我们提出支持兴趣衰减的用户兴趣模型，如式(3)所示：

$$I_u(t)=\alpha \boldsymbol{Q}_u(t_0,t)+(1-\alpha)\sum_{i=1}^{k}(1-e^{-i})\cdot \boldsymbol{P}_u[t_0+(i-1)\times\Delta T,t_0+i\times\Delta T] \quad (3)$$

其中，$\boldsymbol{Q}_u(t_0,t)$代表由用户标签匹配命中的用户发布微博所提取的兴趣向量，表示用户相对长期、稳定的关注内容或兴趣. $\Delta T$ 表示兴趣函数的衰减窗口大小，将用户历史发布的微博按时间划分为 $k$ 份，$k$ 表示时间间隔序列的长度，观测时间周期固定时，$k$ 越大，时间间隔越小，兴趣计算的精度越高. $\boldsymbol{P}_u[t_0+(i-1)\times\Delta T,t_0+i\times\Delta T]$表示由第 $i$ 个窗口内的用户发布微博所提取的兴趣向量，即通过对用户历史发布微博内容分析所提取的兴趣向量. $(1-e^{-i})$表示兴趣随时间的衰减因子，用户越早发布的微博对用户近期兴趣的表征作用越小. $\alpha$ 表示长期兴趣和近期兴趣的权重因子，本文在实验中设定为 0.5.

通过式(3)计算获得用户兴趣向量后，然后计算用户兴趣与待预测微博之间的相似度. 由于微博内容非常短，特征词的出现能够更好地描述用户兴趣特征，所以，本文采用 Jaccard 距离计算微博关键词与用户兴趣向量的相似度. 整体用户兴趣特征计算过程如算法 1 所示.

**算法 1.**　User interest feature calculation algorithm.

输入：
$T=\{T_1,T_2,\cdots,T_m\}$//user tag keyword set
$S=\{S_1,S_2,\cdots,S_m\}$ //user history microblog set
$S_{pre}=\{w_1,w_2,\cdots,w_n\}$ //feature vector of microblog content to be predicted
输出：$sim(T,S,S_{pre})$ //interest similarity
PROCEDURE
FOR EACH $S_i\in\{S_1,S_2,\cdots,S_m\}$ DO
　//word_segmentation, stopword_filtering, wordfrequency_counting
　$S_i\leftarrow preprocessing(S_i)$
　//content feature vector of each microblog
　$vector(S_i)\leftarrow\{w_1,w_2,\cdots,w_n\}$
　$time(S_i)\leftarrow$posting time of $S_i$
END FOR
//calculate the user interest vector from the user's historical microblog
FOR $i=1,\cdots,k$ DO
　FOR EACH $S_j\in\{S_1,S_2,\cdots,S_m\}$ DO
　　IF $time(S_j)\in[t_0+(i-1)\times\Delta T,t_0+i\times\Delta T]$ THEN
　　$P(t_0,t)\leftarrow P(t_0,t)+(1-e^{-i})vector(S_j)$
　　END IF
　END FOR
　$i\leftarrow i+1$
END FOR
//calculate the interest vector from the user's historical microblog which is matched by the user's tag
FOR EACH $T_i\in\{T_1,T_2,\cdots,T_m\}$ DO
　FOR EACH $S_j\in\{S_1,S_2,\cdots,S_m\}$ DO
　　IF $sim(T_i,S_j)>\omega$ THEN
　　$Q(t_0,t)\leftarrow Q(t_0,t)+vector(S_j)$

END IF
END FOR
END FOR
//calculate user interest vector
$I_u(t) \leftarrow \alpha Q(t_0,t)+(1-\alpha)P(t_0,t)$
//calculate the similarity between the characteristics of the microblog content and the user's interest characteristics
$sim(T,S,S_{pre}) \leftarrow JaccardSimilarity(I_u(t),S_{pre})$

我们统计了不同用户兴趣特征值对应转发概率的分布,如图 10 所示,柱状图表示不同用户兴趣特征值对应的微博所占比例,曲线图表示不同用户兴趣特征值对应的微博被转发的比例.从柱状图我们可以看出微博占比随着微博与用户兴趣相似度的增大而降低,微博与用户兴趣相似度为 0.05 的微博占 25%,大部分微博与用户兴趣相似度较低.从曲线图我们可以看出,随着微博与用户兴趣相似度的增大,转发概率逐渐增加,微博与用户兴趣越相似,用户越可能转发该微博.在本文实验部分针对单项特征贡献的实验结果可以看出,待预测微博与用户兴趣的相似度即用户兴趣特征能够将转发预测准确性提升 1.54%,微博与用户兴趣的相似度能够很好的区分用户的转发与不转发行为.

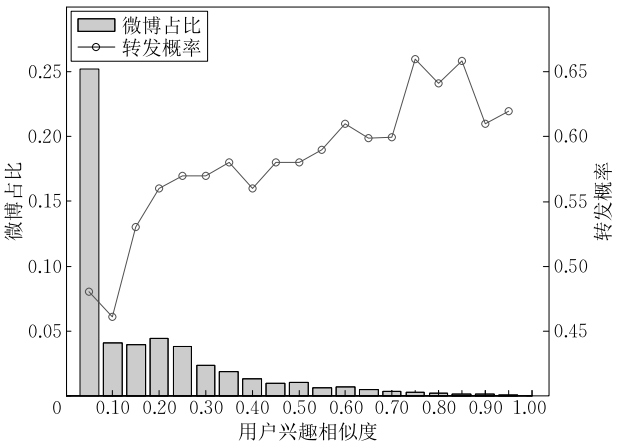


图 10 微博与用户兴趣相似度分布

### 5.4 用户交互性特征

社交网络建立了以用户影响力为信息传播动力的信息扩散机制,用户关注的人数众多,但上游用户对其产生的影响力具有较大差异性[18].研究表明,微博用户更倾向于被与自己紧密相关的"小圈子"影响[19-20],研究转发用户与上游用户的交互性特征能够获取用户的转发源倾向性,找出对用户影响力大的上游用户.用户 $u$ 与其关注用户 $v$ 的互动频率计算方式如下:

$$f_{uv}=\frac{n_{uv}}{\sum_{v\in V_u} n_{uv}} \tag{4}$$

其中,$V_u$ 表示用户 $u$ 关注的用户集合,$n_{uv}$ 表示用户 $u$ 转发的来自 $v$ 的微博数量,$\sum n_{uv}$ 表示用户 $u$ 从关注用户 $v$ 转发的所有微博.

我们统计了不同互动率特征值对应转发概率的分布,如图 11 所示,曲线图表示不同互动率特征值对应的微博被转发的比例,可以看出,互动率越大转发概率也越大.用户偏向于从其过去经常转发微博的关注用户转发微博,用户影响力倾向于通过交互频率最大的关注关系进行传递.用户与上游用户间的互动频率能够较好的区分转发与不转发行为.

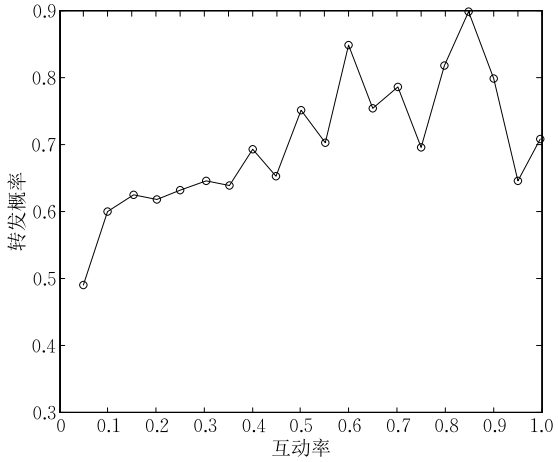


图 11 互动率分布

## 6 实验结果与分析

### 6.1 实验设置和评价方法

为了验证本文方法的有效性,我们利用新浪微博数据开展实验验证工作,在本文构建的新浪微博数据集基础上,进行用户转发微博和不转发微博标注,最终构建了包含 13 902 178 条记录的实验数据集,数据集中每条记录代表一个四元组($u_i$,$u_j$,$m$,$flag$),表示用户 $u_i$ 发布的微博 $m$ 是否被其粉丝 $u_j$ 转发,$flag=+1$ 表示转发,$flag=-1$ 表示不转发.其中,正样本 1 446 533 条,负样本 12 455 645 条.转发微博的识别通过以下过程实现:提取微博转发标志,若 rid 不等于空表明该条微博为转发微博,提取微博内容中的"//@UserScreenName:",获取转发上游用户.不转发微博即忽略样本的识别通过第 4 节介绍的忽略样本识别方法实现.

本文开展 3 组实验.首先,验证所提出的情感词

数量、实体词数量、用户转发率、交互频率、兴趣相似度特征对提升转发预测准确性的贡献，并通过与已有基于微博特征方法的对比，验证本文方法能够提升预测准确性. 其次，验证所提出忽略样本识别方法对提高转发预测准确性的有效性. 最后，通过验证本文方法在不同分类器和在不同规模训练集得到的预测准确性，验证方法的稳定性.

预测准确性的评价方法采用准确率、召回率和 $F$ 值.

准确率($Precision$)，考察转发预测模型的准确性，其数学公式为

$$Precision=\frac{\text{判断正确的记录数目}}{\text{判断为该类别的记录数目}} \tag{5}$$

召回率($Recall$)，考察转发预测模型的全面性，其数学公式为

$$Recall=\frac{\text{判断正确的记录数目}}{\text{应判断为该类别的记录数目}} \tag{6}$$

$F$ 值($F$-Measure)，是准确率和召回率的综合度量指标，其数学公式为

$$F\text{-Measure}=2\frac{precision\times recall}{precision+recall} \tag{7}$$

### 6.2　实验过程与结果分析

#### 6.2.1　UBF-RPM 转发预测过程

我们将本文提出的基于用户行为特征的微博转发预测方法称为 UBF-RPM(User Behavior Features based Retweeting Prediction Method)，基于本文提出的用户行为特征、微博特征、用户兴趣特征、用户交互性特征等转发特征，采用经典分类器，建立转发预测模型进行微博转发预测.

我们设计了 UBF-RPM 转发预测实现过程示意图来进一步说明所提转发预测方法的训练和预测过程. 如图 12 所示，主要包括转发用户兴趣特征、转发用户行为特征、微博内容特征、用户特征分析 4 个部分组成，基于本文提出的转发特征计算方法获取微博转发特征进行转发预测.

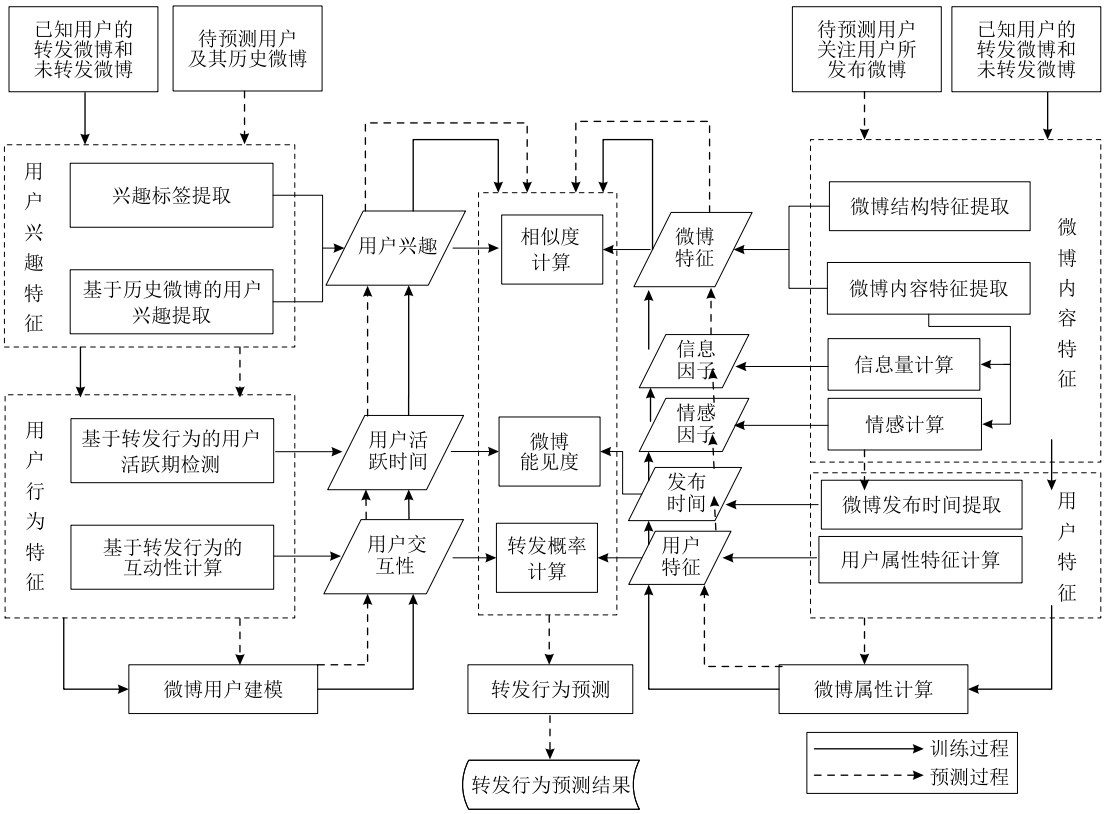


图 12　UBF-RPM 转发预测方法的实现过程

(1)用户兴趣特征计算过程包括基于标签提取和历史微博的用户兴趣分析、微博内容与用户兴趣相似度计算. 其中兴趣标签通过微博用户信息采集与解析，提取兴趣标签信息，建立长期兴趣关键词向量，计算与标签兴趣匹配的用户发布微博，然后汇聚用户发布的历史微博，融合标签和历史微博计算用户兴趣，最后计算用户兴趣与待预测微博内容的相似度，得到用户兴趣特征.

(2)转发用户行为特征计算过程包括用户活跃时间、用户转发率以及用户与上游用户的互动频率

等指标的计算.通过用户活跃时间及其关注用户的数量计算用户使用微博的时间窗口,确定微博是否会出现在用户使用微博的时间区间内,进而结合关注用户发布微博的时间构建训练样本集.通过用户转发率计算用户上网发帖更倾向于原创还是转帖.互动频率则是计算用户与其上游用户之间转发关系的紧密程度.

(3)微博内容特征计算包括微博内容和结构特征的提取和计算、微博信息量计算以及微博情感词计算.微博内容提取用于计算用户兴趣特征,已在用户兴趣特征计算部分介绍.结构特征包括微博是否包含链接、是否包含图片等影响微博转发的因素,在本文作为对比特征使用.微博信息量和微博情感词计算结果形成微博特征.

(4)上游用户特征包括用户发布微博的时间和用户属性特征.用户发布微博的时间用于与待预测用户使用微博的时间区间一起计算待预测微博能见度,构建训练样本集.用户属性主要考虑上游用户特征,具体包括用户是否加 V 和粉丝数,作为本文所提特征的对比特征使用.

在模型训练过程中,通过各类特征计算,然后建立基于分类器的转发预测模型,利用转发和不转发微博训练模型参数.在预测过程中,输入待预测用户关注用户所发布的微博和待预测用户历史微博,计算用户兴趣、活跃时间和用户交互性、微博特征等,基于分类模型进行分类,实现对转发行为的预测.

### 6.2.2 UBF-RPM 转发预测方法的预测结果

为了验证本文提出的用户转发特征的有效性,选择发帖用户、微博结构、用户发帖相关特征作为基准方法.相比于目前研究较多的预测微博是否会被转发的问题,本文的研究目标是局部转发预测问题,即预测特定用户是否会转发其关注用户的某条微博,所以我们选择对前者预测效果较好的特征作为本文方法的基础特征.具体包括发帖用户是否加 V、用户粉丝数、微博是否含有 url、微博是否含有视频、用户发帖数,称为基础特征,如表 3 中特征 1 至 5 所示.本文 UBF-RPM 预测方法中提出的特征包括情感词数量、实体词数量、用户转发率、交互频率、用户兴趣相似度特征,如表 3 中特征 6 至 10 所示.

我们将基础特征和 UBF-RPM 特征输入 C4.5 决策树分类器,采用 4 折交叉验证计算平均 F 值,对比转发预测效果.结果如表 4 所示,UBF-RPM 能够将预测准确性提升 3.59%,相比于用户和微博静态特征,本文所提出的微博内容特征、用户与上游用户的交互频率、微博与用户兴趣的相似度特征,综合考虑了微博内容信息量、情感倾向性、与用户兴趣相似性、用户间交互特性这些对用户转发行为起关键作用的因素,有效提升了转发行为的预测准确性.

**表 3 特征集合及说明**

| 序号 | 特征名称 | 特征描述 |
|---|---|---|
| 1 | Certified | 上游用户是否加 V |
| 2 | Follower number | 上游用户粉丝数 |
| 3 | Contains url | 微博是否含有 url |
| 4 | Contains video | 微博是否含有视频 |
| 5 | Tweet number | 待预测用户发帖数 |
| 6 | Emotional words number | 微博包含情感词数量 |
| 7 | Entity number | 微博包含实体词数量 |
| 8 | Retweeting rate | 待预测用户转发率 |
| 9 | Interaction rate | 待预测用户与上游用户的交互频率 |
| 10 | Interest similarity | 微博内容与待预测用户兴趣相似度 |

**表 4 基准方法与 UBF-RPM 预测准确性对比**

| 特征 | 准确率 | 召回率 | F 值 | 提升率/% |
|---|---|---|---|---|
| Basic_Features | 0.779 | 0.779 | 0.779 | — |
| UBF-RPM | 0.807 | 0.807 | 0.807 | 3.59 |

为了验证 UBF-RPM 各特征对提高预测准确性的贡献程度,我们分别对各个特征相对于基础特征的预测准确性进行了对比实验.结果如表 5 所示,交互频率特征的提升效果最高达到 1.80%,其次是用户兴趣相似度特征,提升效果达到 1.54%,实体词数量特征的提升效果为 0.77%,情感词数量特征的提升效果为 0.51%,用户转发率特征的提升效果为 0.39%,各特征一起使用能够获得最高的提升效果 3.59%.

**表 5 单项特征对提升预测准确性的贡献**

| 特征 | 准确率 | 召回率 | F 值 | 提升率/% |
|---|---|---|---|---|
| Basic_Features | 0.779 | 0.779 | 0.779 | — |
| Basic_Features+Emotional words number | 0.781 | 0.785 | 0.783 | 0.51 |
| Basic_Features+Entity number | 0.786 | 0.785 | 0.785 | 0.77 |
| Basic_Features+Retweeting rate | 0.783 | 0.782 | 0.782 | 0.39 |
| Basic_Features+Interaction rate | 0.794 | 0.793 | 0.793 | 1.80 |
| Basic_Features+Interest similarity | 0.792 | 0.791 | 0.791 | 1.54 |
| UBF-RPM | 0.807 | 0.807 | 0.807 | 3.59 |

### 6.2.3 忽略样本识别方法对提高转发预测准确性的有效性

本文实现了毛佳昕等人[7]提出的用户阅读模型,并以此作为基线方法与所提出的忽略样本识别方法进行对比和效果分析.该论文的目标是通过建

立用户阅读模型计算用户对微博的转发概率，预测微博的被转发次数，进而估计用户影响力大小. 利用该方法进行转发预测的主要原理和实现过程如下：

(1) 首先对用户发帖行为离散化，统计每个用户在 24 小时时间区间里的发帖概率，通过拉普拉斯平滑处理，得到以下用户 $u$ 的发帖概率分布：

$$\rho_u(t)=\frac{\text{用户 } u \text{ 在 } t \text{ 小时发布的微博数量}+1}{\text{用户 } u \text{ 发布的微博总数}+24} \tag{8}$$

(2) 然后假设用户每次访问微博时按照此概率分布独立地选择 1 个小时访问微博，于是得到用户浏览微博的概率分布，其中 $\alpha n_u$ 表示用户平均每天的发帖数.

$$p_u(t)=1-(1-\rho_u(t))^{\alpha n_u} \tag{9}$$

(3) 考虑微博时效性影响，计算用户在 $t$ 时刻阅读一条 $t_s$ 时刻发布的微博的概率 $\lambda_u(t_s,t)$：

$$\lambda_u(t_s,t)=p_u(t)\cdot\frac{1}{t-t_s+1} \tag{10}$$

(4) 考虑用户转发偏好，但不考虑用户 $u$ 转发不同关注用户的概率，偏好的计算公式如下：

$$q_u=\frac{\text{用户 } u \text{ 历史转发的微博数量}}{\text{用户 } u \text{ 时间线上出现的微博数量}} \tag{11}$$

(5) 最后计算用户 $u$ 转发微博的概率：

$$\boldsymbol{P}_u(t_s,t)=\lambda_u(t_s,t)\cdot q_u \tag{12}$$

我们计算了数据集中每个用户 $u$ 与其关注用户所发微博之间的转发概率，通过转发概率大小来预测用户是否转发，方法如下：

$$f(v,u,m)=\begin{cases}+1, & \boldsymbol{P}_u(t_m,t)\geqslant\varepsilon\\ -1, & \boldsymbol{P}_u(t_m,t)<\varepsilon\end{cases} \tag{13}$$

当转发概率大于等于 $\varepsilon$ 时，预测结果为转发，小于 $\varepsilon$ 时预测结果为不转发. 实验结果表明当 $\varepsilon=0.007$ 时，分类准确性最高，与本文 UBF-RPM 方法的对比结果如表 6 所示.

**表 6　毛佳昕等人[7]方法与 UBF-RPM 方法的准确性对比**

| 方法 | 准确率 | 召回率 | $F$ 值 | 提升率/% |
|---|---|---|---|---|
| 毛佳昕等人方法 | 0.671 | 0.669 | 0.670 | — |
| UBF-RPM | 0.807 | 0.807 | 0.807 | 20.45 |

毛佳昕等人方法的预测准确性为 0.67，UBF-RPM 方法的预测准确性为 0.807，提升效果达到 20.45%. 这是因为毛佳昕等人方法提出的用户阅读模型是以每天 24 小时为窗口，但如前文统计数据所示，91.7%的用户平均每天只发布 1 条微博，真实微博网络中的用户发帖非常稀疏，每天访问微博的时间会受诸多因素影响，具有一定随机性，导致基于用户发帖量统计结果的用户阅读模型的准确性受到影响；该方法假设了用户平均每天发布微博数量不变，但是用户发帖会受到个人或者热点事件等因素的影响；模型没有考虑用户关注用户数量的不同而导致的消息流逝速率不同，很多消息在未来得及被用户阅读就被大量新产生的微博湮没；第四，模型没有考虑用户个人兴趣对转发微博的影响. 而 UBF-RPM 方法在进行模型训练时以用户真实发帖时间作为用户使用微博的时间标志，以其关注用户数为指标估计消息流逝的速度，进而计算用户浏览微博的时间区间，避免预测模型受到用户发帖时间随机性和发帖数量稀疏性的影响，使得分类模型更加准确. 其次，提出的基于用户标签和历史发布微博的兴趣融合计算方法，充分考虑了用户兴趣与微博内容匹配程度对用户转发行为的影响.

### 6.2.4　不同分类器和训练数据规模下的预测准确性

为了验证 UBF-RPM 特征效果的稳定性，我们对基础特征和 UBF-RPM 特征在 C4.5 决策树分类器和贝叶斯网络分类器下的预测准确性进行了对比实验，并且通过交叉验证方法验证了 UBF-RPM 在不同规模训练数据下预测效果的稳定性. 交叉验证是常用的精度测试方法，将数据集分成 $N$ 份，轮流将其中 $N-1$ 份做训练，剩余 1 份做测试，$N$ 次的结果均值作为算法准确性的估计.

图 13 是在 C4.5 决策树分类器下基础特征和 UBF-RPM 特征 2 折到 10 折交叉验证准确性的对比结果. UBF-RPM 的准确性一致地高于基础特征，在 2 折交叉验证实验下，UBF-RPM 相比基础特征的提升率为 3.49%，在 10 折交叉验证实验下，UBF-RPM 相比基础特征的提升率为 3.59%. UBF-RPM 方法 2 折交叉验证准确性达到 0.8，随着训练数据规模的增大，UBF-RPM 的预测准确性稳步提高，从 2 折到 10 折，准确性提高了 1.13%，而基础特征的准确性在 6 折实验时出现下降. 这说明 UBF-RPM 方法能

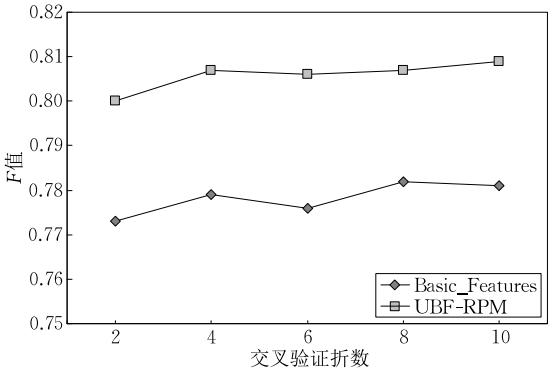


图 13　基于决策树分类器的预测准确性

够在训练集规模较小的情况下获得更高的准确性，并且准确性能够随训练数据规模的增大而逐步提升.

图 14 是在贝叶斯网络分类器下基础特征和 UBF-RPM 特征 2 折到 10 折交叉验证准确性的对比结果. UBF-RPM 的准确性同样一致高于基础特征，在 2 折交叉验证实验下，提升率为 2.65%，在 10 折交叉验证实验下，提升率为 1.68%. UBF-RPM 方法 2 折交叉验证准确性达到 0.776. 同样地，随着训练数据规模的增大，预测准确性稳步提高，从 2 折到 10 折，UBF-RPM 的准确性提高了 1.16%，而基础特征的准确性提高了 2.12%，表明 UBF-RPM 方法能够在训练集规模较小的情况下更快地达到最高预测准确性.

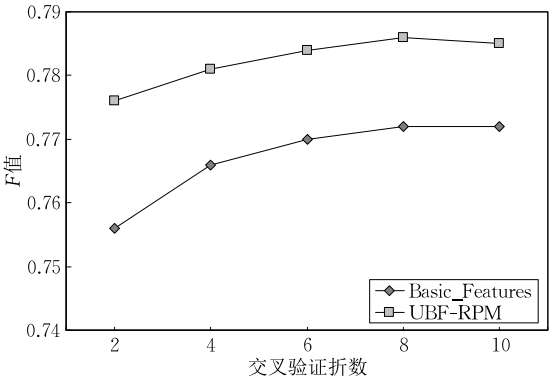


图 14 基于贝叶斯网络分类器的预测准确性

通过在不同分类器和训练数据规模下的预测准确性实验结果表明，本文基于上游用户、微博内容、转发用户提出的微博内容信息量、情感倾向性、用户兴趣相似性、用户交互特性等特征能够有效地预测用户对于特定微博的转发行为，在不同分类器下的预测准确性都高于使用用户和微博静态特征的分类方法，基于 C4.5 决策树分类器的预测准确性更高，所以本文之前的实验都采用 C4.5 决策树分类器. 其次，UBF-RPM 方法能够在较小规模的训练集上获得更高的准确性，并且随着训练数据规模的增大，预测准确性稳步提升. 这是因为 UBF-RPM 特征融合了微博内容、用户兴趣、用户交互性等特征，这些特征能够更准确的刻画用户转发行为特征，具有相对稳定性，所以在较小规模数据集上训练得到的预测模型能够获得更好和更稳定的预测效果.

## 7 结束语

本文研究用户对于其关注用户的某条微博是否会转发这一局部转发预测问题. 围绕社交网络中信息传播过程，针对新浪微博网络，分别从信息发送者、传播内容、接收者的角度研究转发行为预测问题，提出了基于微博能见度和用户兴趣的转发行为预测方法. (1) 在研究用户活跃时间和微博能见度的基础上，提出了基于动态时间窗的转发行为、忽略行为、未接收行为识别方法，设计了数据集构建方法，为模型验证和效果分析提供了更为准确的数据基础和测试环境，并通过对比实验验证了该方法能够提高预测模型的准确性；(2) 研究并提出了基于时间衰减的用户兴趣模型，融合了用户长期兴趣和短期兴趣特征，有效度量了社交网络信息传播过程中，用户兴趣及其变化特性对用户转发行为的影响程度；(3) 研究并提出了转发率等用户行为特征，以及交互频率等用户交互特征，有效度量了用户历史行为模式、用户影响力传递效应的差异性对用户转发行为的影响；(4) 融合上游用户特征、微博特征、转发用户兴趣和历史行为特征，提出了基于分类器的用户转发行为预测模型，在真实数据下，通过准确率、召回率、准确性 $F$ 值指标对特征在不同分类模型和训练数据下的效果进行了验证和对比，并对结果进行了分析和说明. 结果表明，本文提出的特征有效提升了预测准确性，在 C4.5 决策树分类器下的预测准确性最高，并且能够在较小规模的训练集上达到最好的预测效果.

本文所提出的基于上游用户特征、微博特征、转发用户兴趣和历史行为特征的转发预测方法，为展现和认识社交网络中微观层面用户转发行为模式，提供了有效的途径和方法. 提出的基于时间区间的用户忽略行为识别、基于时间衰减的用户兴趣模型、用户交互频率计算等方法能够有效应用于社交网络话题检测、用户行为分析、影响力传播模型等相关研究. 下一步可以进一步研究用户对不同主题下微博的转发倾向性，分析不同主题下的微博内容和传播规律对用户转发行为的影响，进一步提高转发预测准确性.

## 参 考 文 献

[1] 36th Statistical Report on Internet Development in China. China Internet Network Information Center, Beijing, 2015 (in Chinese)
(第 36 次中国互联网络发展状况统计报告. 中国互联网信息中心(CNNIC), 北京, 2015)

[2] Liu Wei, Wang Li-Hong, Li Rui-Guang. Topic-oriented measurement of microblogging network. Journal on Communications, 2013, 34(11): 171-178(in Chinese)
(刘玮，王丽宏，李锐光. 面向话题的微博网络测量研究. 通信学报，2013，34(11)：171-178)

[3] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media//Proceedings of the 19th International Conference on World Wide Web. Ralcigh, USA, 2010: 591-600

[4] Chew C, Eysenbach G. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. PLoS One, 2010, 5(11): e14118

[5] Java A, Song X, Finin T, et al. Why we Twitter: Understanding microblogging usage and communities//Proceedings of the Joint 9th WebKDD and 1st SNA KDD Workshop'07. San Josc, USA, 2007: 56-65

[6] Fan Peng-Yi, Wang Hui, Jiang Zhi-Hong, et al. Measurement of microblogging network. Journal of Computer Research and Development, 2012, 49(4): 691-699(in Chinese)
(樊鹏翼，王晖，姜志宏等. 微博网络测量研究. 计算机研究与发展，2012，49(4)：691-699)

[7] Mao Jia-Xin, Liu Yi-Qun, Zhang Min, et al. Social influence analysis for micro-blog user based on user behavior. Chinese Journal of Computers, 2014, 37(4): 791-800(in Chinese)
(毛佳昕，刘奕群，张敏等. 基于用户行为的微博用户社会影响力分析. 计算机学报，2014，37(4)：791-800)

[8] Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network //Proceedings of the IEEE Second International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust. Minneapolis, USA, 2010: 177-184

[9] Zhang Yang, Lu Rong, Yang Qing. Predicting retweeting in microblogs. Journal of Chinese Information Processing, 2012, 26(4): 109-114(in Chinese)
(张旸，路荣，杨青. 微博客中转发行为的预测研究. 中文信息学报，2012，26(4)：109-114)

[10] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1633-1636

[11] Cao Jiu-Xin, Wu Jiang-Lin, Shi Wei, et al. Sina microblog information diffusion analysis and prediction. Chinese Journal of Computers, 2014, 37(4): 779-790(in Chinese)
(曹玖新，吴江林，石伟等. 新浪微博网信息传播分析与预测. 计算机学报，2014，37(4)：779-790)

[12] Wu Fang, Huberman B A. Novelty and collective attention. Proceedings of the National Academy of Sciences, 2007, 104(45): 17599-17601

[13] Tan C H, Tang J, Sun J, et al. Social action tracking via noise to tolerant time-varying factor graphs//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1049-1058

[14] Wang C, Tang J, Sun J, et al. Dynamic social influence analysis through time-dependent factor graphs//Proceedings of the International Conference on Advances in Social Networks Analysis and Mining. Kaohsiung, China, 2011: 239-246

[15] Hogg T, Szabó G. Diversity of user activity and content quality in online communities//Proceedings of the 3rd International Conference on Weblogs and Social Media. Menlo Park, USA, 2009: 58-65

[16] Weng J S, Lim E P, Jiang J, et al. TwitterRank: Finding topic-sensitive influential Twitters//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 261-270

[17] Liu Qun, Li Su-Jian. Word similarity computing based on how-net. International Journal of Computational Linguistics & Chinese Language Processing, 2002, 7(2): 59-76(in Chinese)
(刘群，李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学，2002，7(2)：59-76)

[18] Wu S, Hofman J M, Mason W A, et al. Who says what to whom on Twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 705-714

[19] Zhang J, Liu B, Tang J, et al. Social influence locality for modeling retweeting behaviors//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2761-2767

[20] Huang J M, Cheng X Q, Shen H W, et al. Exploring social influence via posterior effect of word-of-mouth recommendations //Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 573-582

**LIU Wei**, born in 1984, Ph. D. candidate, senior engineer. Her research interests include social network data mining, network information security, and information filtering.

**HE Min**, born in 1982, Ph.D. candidate, senior engineer. Her research interests include social network data mining, topic detection.

**WANG Li-Hong**, born in 1967, Ph. D., professor of engineering, Ph. D. supervisor. Her research interests include network information security, social network data mining, public opinion processing.

**LIU Yue**, born in 1971, Ph. D., associate professor. Her research interests include information retrieval and social

computing.

**SHEN Hua-Wei**, born in 1982, Ph.D., associate professor. His research interests include social computing, complex networks, information retrieval.

**CHENG Xue-Qi**, born in 1971, Ph.D., professor, Ph.D. supervisor. His research interests include network science, network information security, Web data mining.

**Background**

This work is supported by Grants from the National Natural Science Foundation of China (No. 61170230), the National Key Technology R&D Program (No. 2012BAH46B01), and the National High Technology Research and Development Program of China (No. SS2014AA012303).

Microblog retweeting prediction is of great significance to topic detection and influence evaluation, which has attracted wide attention from both academic and industrial fields. Current microblog retweeting prediction methods can be divided into two categories, which are based on the characteristics analysis of the high retweeting microblogs and the information diffusion model based on the microblog network. Prediction methods which are based on high retweeting microblogs features used the users and microblogs properties and statistical characteristics to predict retweeting behavior. They have not fully considered the dynamics of retweeting behavior and regularity of user's historical behavior. Additionally, there is no in-depth consideration of real users' reading behavior which can be introduced into the prediction model. Prediction methods which are based on information diffusion model needed to obtain complete forwarding relations and history retweeting log data. However, in real retweeting prediction tasks, there are a large number of nodes in network and most of them are inactive nodes. In most cases, only partial user data and log data can be obtained by crawling. It is very difficult to establish complete forwarding network and node state, let alone its high complexity.

The authors investigated the microblog retweeting prediction problem from the view of microblog visibility and user behavior features, and (1) proposed a method to recognize retweeting behavior, ignoring behavior and un-received behavior based on user's activity and dynamic time window, which provided more accurate dataset for model training and effectiveness analysis; (2) presented user interest model based on dynamic user's interest and time attenuation, which is proved to be an effective measurement of user's interest and its change characteristics; (3) proposed several user behavior features of the user's retweeting rate and interaction frequency, which can effectively measure the impact of user's historical behavior patterns and user's influence transfer effect. Finally, this paper proposed a classification model based on retweeting behavior prediction method which is blend with upstream user's characteristics, microblog's characteristics, forwarding user's interest and user's historical behavior characteristics. Experimental results on real data show that the proposed method can improve prediction accuracy effectively, and achieve good results in the smaller size of the training set.

The proposed retweeting prediction method which is based on upstream user's characteristics, microblog's characteristics, forwarding user's interest and user's historical behavior characteristics, shows an effective way to understand user's retweeting behavior patterns from the micro level of the social network. The time interval based user's ignoring behavior recognition method, time attenuation based user interest model and user's interaction frequency calculation method can be applied to user's behavior analysis, influence propagation mechanism analysis and other related research effectively. Their work sheds light on future works, in which they may study the user's retweeting tendency on different topics, and analyze the influence of microblogs content and diffusion model on user's retweeting behavior under different topics, which will get further improvement on the retweeting prediction accuracy.