计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(1):28−34 [doi: 10.15888/j.cnki.csa.006147]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于情绪和兴趣的用户访问行为预测①

秦　锋[1], 陈　增[1], 郑　啸[1], 童　琨[2]

[1](安徽工业大学 计算机科学与技术学院, 马鞍山 243032)
[2](安徽祥云科技有限公司, 马鞍山 243032)

**摘　要**: 微博用户行为预测旨在研究用户的行为习惯, 本文主要从用户属性、用户兴趣和用户情绪三个方面, 对影响微博用户行为的因素进行研究分析, 提取影响用户行为的特征, 训练预测模型. 实验中还将情感和兴趣特征在预测模型中的作用进行了对比, 结果显示预测模型在转发行为预测的平均准确率能够达到 82.56%, 在评论行为预测的平均准确率能够达到 84.59%, 在点赞行为预测的平均准确率能够达到 79.35%, 表明了用户兴趣和情感特征对于微博用户行为预测结果提升中的有效性.

**关键词**: 用户行为; 微博; 情感分析; 兴趣; 预测

引用格式: 秦锋,陈增,郑啸,童琨.基于情绪和兴趣的用户访问行为预测.计算机系统应用,2018,27(1):28–34. http://www.c-s-a.org.cn/1003-3254/6147.html

## User Behavior Prediction Based on Emotion and Interest

QIN Feng[1], CHEN Zeng[1], ZHENG Xiao[1], TONG Kun[2]

[1](College of Computer Science and Technology, Anhui University of Technology, Maanshan 243032, China)
[2](Anhui Xiangyun Technology Co. Ltd., Maanshan 243032, China)

**Abstract**: Micro-blog user behavior prediction aims to study user behavior habits. This paper mainly studies the factors that affect the behaviors of users of microblogging from three aspects: the user attribute, user interest, and user's emotion. We extract the characteristics of the user behaviors, training and forecasting the model. The experimental results show that the average accuracy of forwarding behavior can reach 82.56% in the prediction, the average prediction accuracy of behavior in the comments reaching 84.59%, the prediction average accuracy of likes behavior rate reaching 79.35%, which indicates the effectiveness of user interest and emotion characteristics in the promotion of microblogging user behavior prediction.

**Key words**: user behavior; micro-blog; sentiment analysis; interest; prediction

## 1　引言

随着网络的快速发展, 为了对用户的社交网络访问行为进行预测, 将用户关注的内容呈现给用户, 实现个性化推荐, 并且对网络用户行为实时监管, 是当下网络发展过程中遇到的难题之一. 用户访问行为预测研究不仅仅实现对用户的个性化推荐, 展现了其商业价值, 同时也为网络信息传播、舆情监控、网络异常行为监控和热点提取等问题的研究提供帮助, 展现了其科研价值. 根据 CNNIC[1]发布的第 38 次中国互联网络发展状况统计报告, 直到 2016 年 6 月份, 我国微博用户规模为 2.42 亿.

## 2　相关工作

现在微博已成为在我国最广泛使用的社交网络, 分

① 基金项目: 国家自然科学基金(61402008); 安徽省高校自然科学研究重大项目(KJ2014ZD05); 安徽省高校优秀青年人才支持计划; 安徽省科技重大专项 (16030901060)
收稿时间: 2017-03-30; 修改时间: 2017-04-20; 采用时间: 2017-05-02; csa 在线出版时间: 2017-11-14

析研究微博用户的行为习惯对于了解社交网络信息的传递与扩散有着重要的参考价值. 目前国内外对于微博用户的行为研究主要根据用户的浏览和转发的历史行为或者用户关注对象特征等用户静态属性进行预测, 而忽略了用户本身的情绪和兴趣的影响. 在心理学研究中发现情绪对于用户行为有着直接的影响, 目前已经有部分学者将心理学模型于用到文本情感分析研究中, 本文主要从用户发布微博的文本信息进行研究分析, 将用户浏览微博时的情感和兴趣引入到预测模型中, 与用户的属性特征结合, 以此达到提升预测模型的效果.

### 2.1 用户行为预测

随着微博用户规模的不断扩大, 微博在人们的日常生活中的地位也更加重要, 国内外对于微博网络中用户行为也有了更多的研究. 张旸等人[2]根据分析 Twitter 中用户转发行为的特点, 根据选取特征的重要性排名, 提出了基于特征加权预测模型, 使用机器学习的方法验证了模型的有效性. 清华大学的 Tan 等人[3]通过构建社交网络结构, 分析用户属性和用户行为历史, 提出 NTT-FGM 模型以便更好地预测用户行为. 曹玖新等人[4]以新浪微博为研究对象, 对各种影响用户转发微博的因素统计分析, 并且根据分析的特征进行建模研究. 最终选取用户特征、社交特征和微博特征构建转发预测模型, 通过机器学习的方法验证模型的效果. Xu Zhiheng 等[5]从个人用户的转发行为的视角对 Twitter 的社会特征、内容特征、Twitter 特征和作者特征构建预测模型, 实验中使用 C4.5 决策树、支持向量机、逻辑回归三种分类算法, 并提出了“leave-one-feature-out”的方法确定了影响用户转发行为的特征是密切相关的. 刘玮等[6]将影响用户转发行为的因素分为三类: 用户行为因素、微博因素、用户兴趣因素. 通过分析各方面的特征建立预测模型 UBF-RPM 模型, 实验表明效果提升 3.59%. 李志清等[7]分析了影响用户转发行为的各类因素, 通过将 LDA 概率主题模型挖掘微博的隐含主题特征, 与微博特征和用户特征结合建立微博转发预测模型, 实验结果表明融合特征对转发行为预测的有效性.

### 2.2 文本情感与兴趣

微博短文本情感分析是通过对微博文本内容的分析, 挖掘用户在微博文本中表达的情感色彩, 同时这是微博短文本情感分析的工作核心. 如今国内外在微博短文本情感方面的研究非常多, Pak 等人[8]从语言学的角度对抓取的 Twitter 微博进行分析, 构建语料库, 建立情感分类器, 并且在 NB、SVM 和 CRF 实现. Sriram 等[9]考虑到微博文本的特有特征, 如作者信息、发布时间等, 通过实验说明在文本分类任务时加入这些特征后, 分类性能得到了提高. 国内外对于微博用户兴趣的研究同样取得了很大的进展. Shen 等[10]假设用户的兴趣分布可以用各种实体表示, 利用主题算法对知识库进行实体训练以及上下文语义关联, 构建用户兴趣模型并完成实体链接任务. 邱云飞等人[11]结合微博短文本数据集, 给出微博短文本重构概念, 对微博的原始特征进行扩充, 让聚类效果有所提升, 而且根据重构特征建立用户兴趣模型. 王岩等[12]根据微博数据存在大量链接的特点, 抽取 HTML 元素组成文档链, 根据共现阈值构造主题抽取模型, 并且实现话题的情感分析. 陈文涛等人[13]通过对 TwitterLDA、UserLDA 以及 AuthorLDA 的对比实验, 分析了三种 LDA 模型优势所在, 同时详细介绍了通过主题模型来构建用户兴趣模型的方法和技术.

## 3 社交网络用户行为

### 3.1 社交网络用户行为特点

网络用户行为的一个子类——社交网站用户行为, 不但拥有其父类的特征, 自身同时具有独特的个性. 我们把社交网络中的行为特点总结归纳成下面的 4 点.

1) 交互性. 当用户浏览社交网络平台的时候, 会通过信息的发布、转发、评论等行为与好友进行交互, 在信息转发等传递过程中, 用户的信息交互促进朋友关系的发展, 也会吸引更多新的用户加入.

2) 消息快速扩散性. 社交网络中, 用户之间构建了庞大的复杂的用户关系网络, 用户发布或者分享的消息能够快速的在用户间传播, 随着社交平台的多样化, 信息的传播速度也大大提升.

3) 保密性. 多数的社交平台使用中, 不要求用户实名认证, 对用户的信息最大程度上给予保护.

4) 不确定性. 现在随着各式各样的社交网络平台的出现, 使得原本繁杂的网络环境更加的复杂, 社交用户的群体也有着很大的区别, 这些都让用户在社交网络中的行为变得更加复杂多变, 难以预测.

### 3.2 微博用户行为对比

在微博平台中, 对于所有用户均可见的行为有转发、评论和点赞 3 种, 还有一种收藏行为除了用户本

身之外的其他均不可见, 所以对于微博用户行为的研究中不考虑收藏行为.

我们通过对某一认证用户一个星期内发布的微博的点赞数、评论数及转发数的对比, 我们发现三种行为之间的操作次数的变化趋势呈现出一致性, 所以认为 3 种行为操作之间具有正相关的关系, 如图 1 所示.

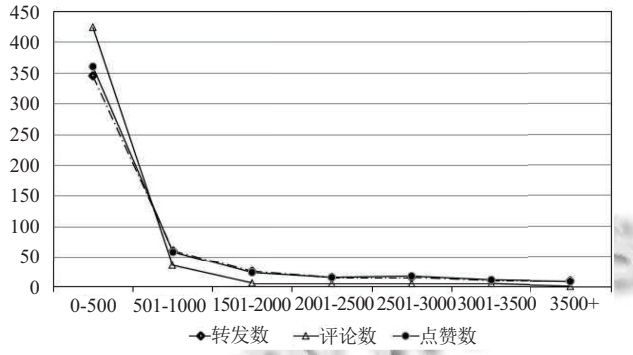



图 1 微博转发、评论、点赞数对比

经过对该认证用户的这一个星期内发布的微博的点赞数、评论数及转发数的对比分析, 我们通过 SPSS 工具对微博的三种操作次数进行相关性分析, 结果发现其显著性 p=0.2%<1%(显著性水平), 说明三种行为之间都具有强正相关关系, 其相关系数都超过了 99%, 这说明微博用户的转发、评论和点赞行为之间有直接的关系.

## 4 用户访问行为预测模型

### 4.1 用户特征

本文选取的属性特征有微博等级、粉丝数量、微博累计数量、认证类别、微博使用天数、是否是会员、会员等级、关注用户数量等. 这些用户主要特征能够很好地帮助我们识别用户的类型, 其中关注用户数量能够反映当前用户微博被转发的可能性大小.

### 4.2 文本情感模型

本文情感特征是通过短文本情感分析方法, 对文本进行识别分析, 最终得到文本的情感特征, 这里情感特征主要分为三类, 包括: 正向情感特征, 中性情感特征, 以及负向情感特征. 由于微博的特性, 本文采用微博短文本情绪分类方法, 主要选取的文本特征包括情感词典、否定词、表情及标点符号、词性标注特征等.

#### 4.2.1 情感特征

目前主流的情感分析算法, 很多都在使用情感词典, 但是目前主流的情感词典中 HowNet 和 NTUSD 词典只有情感极性分类, 没有情感强度的划分, 所以本文在情感词的识别任务中, 我们根据 HowNet、DLUT、NTUSD 情感词典, 构建 AHUT 情感词典. 格式如表 1 所示.

表 1 AHUT 情感词典示例

| 词语 | 极性强度值 | 词语 | 极性强度值 |
|---|---|---|---|
| 苟延残喘 | –1 | 深意 | 1 |
| 剥削 | –3 | 气壮如牛 | 3 |
| 早衰 | –5 | 瑞雪 | 5 |
| 脏乱 | –7 | 叹为观止 | 7 |
| 罪状 | –9 | 英姿飒爽 | 9 |
| 长长短短 | 0 | | |

注: 负号表示该词情感极性为负, 无符号表示该词情感极性为正, 数值表示词语情感强度的强弱, 其中0表示中性.

在文本情感计算中还引入程度副词, 表情符号及特殊标点符号 (如“? ”、“! ! ! ”、“……”等等), 使情感计算更加准确, 一般情况下仅有一个程度副词修饰该情感词. 程度级别副词词典由 HowNet 提供, 共包含 219 个词汇, 分为 6 个级别, “极其/最”, “很”, “较”, “稍”, “欠”, “超”. 还将否定词加入到特征中去, 本文中用到的否定词如表 2 所示.

表 2 否定词列表

| 否定词 |
|---|
| 并非、未、没、不然、不、否、不得、不能、不可、没有、莫、非、无、不够、绝非、不是、不行、不要 |

综合考虑情感模型的特征, 情感计算公式如下:

$$Sent_i = \sum_{j,h,k} s_{i,j} \times adv_{j,h} \times (-1)^k + \sum_n Sy_n \quad (1)$$

其中, $s_{i,j}$表示文本 $i$ 中的情感词, $adv_{j,h}$为修饰情感词 $s_j$ 的程度副词, 而 $k$ 表示情感词前后窗口中否定词出现的次数, $adv_{j,h}$表示修饰情感词的程度副词, $Sy_n$ 表示表情符号以及特殊符号的情感强度.

#### 4.2.2 情感模型构建

结合当前短文本情感分析的研究, 选取在短文本分析中重要的特征构建本文微博情感分析模型, 其中微博情感强度计算如算法 1 所示.

算法 1. 文本情感计算算法

输入: $T = \{T_1, T_2, \cdots, T_n\}$ //文本列表
输出: $S = \{S_1, S_2, \cdots, S_n\}$ //文本情感向量

FOR each $T_i \in \{T_1, T_2, \cdots, T_n\}$ DO
//文本预处理, 分词, 去除 URL, @用户名以及停用词等
$T_i$←preprocessing ($T_i$)

$T_i$←{$w_1, w_2, \cdots, w_m$} //微博词向量
IF ($T_i$ contain ($S_{i,j}$))
    $Sent_i$ += $S_{i,j}$ *$adv_{i,j}$*$(-1)^k$;
IF ($T_i$ contain ($Sy_n$))
    $Sent_i$ += $Sy_n$;
END FOR
IF ($Sent_i$==0) //根据情感词强度计算
//情感强度, 然后根据对文本标记
    $T_i$_Lable = 0;
ELSE IF ($Sent_i$>0)
    $T_i$_Lable = 1;
ELSE
    $T_i$_Lable = –1;
END IF

### 4.3 用户兴趣模型

为了将用户兴趣更好的分类展示, 我们对新浪微博主页的热门类别与各种门户网站上的类别目录进行分析研究, 最后确定将微博用户兴趣映射到 10 个较为常见的分类中, 分别是: 新闻、娱乐、体育、财经、科技、时尚、汽车、旅游、教育、文化. 并且由此构建用户兴趣向量$I_j = \{I_{j,1}, I_{j,2}, \cdots, I_{j,10}\}$. 如某用户对美食和娱乐的内容兴趣度较高, 则其对应的兴趣向量为 $I$ = {1,0,1,0,0,0,0,0,0,0}. 本文用户兴趣模型从用户标签特征和文本隐含主题特征两个方面提取用户兴趣.

#### 4.3.1 用户标签特征

用户个人标签是指描写职业、个性或者兴趣等的短语或者有关自我介绍的词组等, 这些标签在很大程度上反映了用户的兴趣领域, 但是也有一些不利之处, 如微博中很多用户是没有设置自己的标签或者标签是随便填写, 不能真实的体现用户的兴趣等.

#### 4.3.2 基于改进 TF-IDF 兴趣关键词提取

TF-IDF (Term Frequency-Inverse Document Frequency) 是文本分类研究中的常用技术, 是用来统计文档中每个词汇对于该文档的影响力大小的工具[14]. TF-IDF 的主要思路是: 如果在某个文档中一个词语出现的次数较多, 而在其余文档中出现次数较少, 则这个词语就能够很好的把该文档与其他的区别开来. TF-IDF 值等于: $TF_{i,j} \times IDF_i$, $TF_{i,j}$表示词频, 即 $w_i$ 在文档 $j$ 中出现的频率, $IDF_i$ 为 $w_i$ 在训练语料上的逆文档频率值.

在选择特征方面, TF-IDF 方法和信息增益 (Information Gain) 方法忽略了特征词类间分布情况; 而卡方检验 (Chi-square test) 方法和互信息 (Mutual Information) 方法有低频词倾向, 夸大了低频词的作用. 文档分布方差反映的是不同类别文本间特征词分布差异, 词概率分布方差则可以修正文档分布方差的低频词缺陷. 根据这两类方差的特点, 将其与 TF-IDF 计算融合到一起中, 能够在一定程度上提升主题关键词的提取效果. 例如“手机”既有可能在“科技”类别的新闻中出现, 又可能出现在“时尚”类别新闻中等等. 为了保证“类别”专有特征的选择效果并且保证主题关键词的提取准确率, 我们用词的类间概率分布方差和文档分布方差乘积的对数来更新 TF-IDF 的特征权重.

设 $w_i$ 是文本集中的一个词, 词 $w_i$ 的类间概率分布方差为:

$$\mathrm{var}(w_i) = \sqrt{\frac{\sum_j p\left(w_i, c_j\right) - \overline{p(w_i)}}{c}} \tag{2}$$

$c$ 为类别总数, $p\left(w_i, c_j\right) = \dfrac{N\left(w_i, c_j\right)}{\sum_i N\left(w_i, c_j\right)}$是词 $w_i$ 在类别 $c_j$ 中的出现的概率, 同理, 定义词 $w_i$ 的类间文档分布方差为:

$$\mathrm{var}_D(w_i) = \sqrt{\frac{\sum_j p_D\left(w_i, c_j\right) - \overline{p_D(w_i)}}{c}} \tag{3}$$

其中$p_D\left(w_i, c_j\right) = \dfrac{D\left(w_i, c_j\right)}{D\left(c_j\right)}$是特征词 $w_i$ 的文档概率.

文档 $j$ 中词 $w_i$ 的 TF-IDF 特征修正权重是:

$$weight(w_i, j) = TF_{i,j} * IDF_i * \log(\mathrm{var}(w_i) * \mathrm{var}_D(w_i)) \tag{4}$$

#### 4.3.3 LDA 主题特征

现实的微博网络环境中, 微博的文本内容在很大程度上影响用户是否浏览、转发该信息, 每个用户都有自己独特的兴趣爱好, 关注科技方面但是不懂体育的用户在浏览微博时, 对“大数据”为主题的微博的兴趣度要比“NBA 比赛”为主题的微博的兴趣度高很多. 因此, 微博文本的内容隐含主题特征对于微博用户行为的影响非常大. 本文通过使用 LDA 模型对用户一定时间段内的微博文本提取特定主题数的主题词语分布, 实现了文本内容到主题向量的映射. 本文的 LDA 主题模型使用的是 LDA 开源工具 JGibbLDA, LDA 模型中的主要参数 alpha 默认为 50/K(K 是主题数目), beta 默认值取 0.1.

#### 4.3.4 用户兴趣模型构建

为了更加准确地提取用户的兴趣, 我们构建用户兴趣模型, 将用户的标签兴趣 $Q$ 与文本实时兴趣 $P$ 根据公式计算, 得出最终用户兴趣 $I$. 标签兴趣 $Q$ 是根据标签词语和用户兴趣类别关键词的相似度计算得出,

实时兴趣 $P$ 是根据 TF-IDF 提取的关键词与 LDA 模型输出的主题分布进行相似度计算, 如算法 2 所示.

算法 2. 用户短期兴趣提取算法

输入:

$T=\{T_1, T_2, \ldots, T_n\}$ //用户微博集合

$F=\{F_1, F_2, \ldots, F_m\}$ //用户特征向量

输出:

$P_u(t_0,t)=\{i_1,i_2,\ldots,i_{10}\}$ //用户兴趣向量

```
SHORT_INTEREST PROCEDURE
FOR i = 0, 1, 2, …, n DO
//通过文档主题生成模型获取关键词分布
    K = {K1, K2, …, Kl}←Ti;
//TF-IDF 算法处理过程
    Words = {W1, W2, …, Wh}←Ti;
    FOR j = 0, 1, 2, …, h
        FOR k = 0, 1, 2, …, l
        If (Similar(Kk, Wj) < α)
        remove Kk from K;
        END FOR
    END FOR
Short ← K;
Similar(Kj, Wi)←JaccardSimilarty(Kj, Wi);
END FOR
```

根据用户微博的发布时间, 将其短期兴趣分为 $k$ 个时段的实时兴趣 (本文中时间间隔取一周), 根据兴趣衰减函数, 得到用户在$(t_0,t)$时间内的用户兴趣, 其公式如下:

$$I_u(t)=\alpha Q_u(t_0,t)+(1-\alpha)\times\sum_{i=1}^{k}\left(1-e^{-i}\right)\times P_u(t_0+(i-1)\times\Delta t, t_0+i\times\Delta t) \quad (5)$$

如图 2 所示, 介绍了用户兴趣提取实现过程.

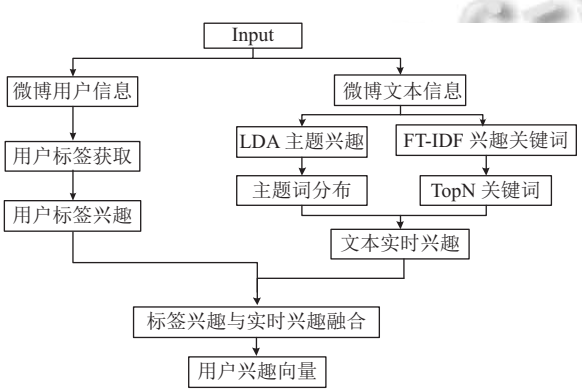



图 2 用户兴趣提取过程

### 4.4 微博用户访问行为预测模型

将情感特征提取模型和用户兴趣模型, 获取的微博情感, 用户兴趣以及用户特征融合, 建立微博用户的行为预测模型, 预测模型特征如表 3 所示.

表 3 预测模型特征

| 特征属性 | 编号 | 特征描述 |
|---|---|---|
| 用户特征 | 1 | 关注数 |
| | 2 | 粉丝数 |
| | 3 | 微博数 |
| | 4 | 认证类别 |
| | 5 | 活跃天数 |
| | 6 | 是否是会员 |
| | 7 | 会员等级 |
| | 8 | 微博等级 |
| | 9 | 历史被转发数 |
| 情感特征 | 10 | 正向情感 |
| | 11 | 中立 |
| | 12 | 负向情感 |
| 主题特征 | 13-22 | 主题向量 |

根据选取的特征, 构建模型输入向量, 根据分类器输出行为预测结果向量$\left(u_i, text_{i,j}, x_{i,j}, y_{i,j}, z_{i,j}\right)$, 当$x_{i,j}=-1$时表示不会进行转发操作, 当$x_{i,j}=+1$时, 表示进行转发操作; 当$y_{i,j}=+1$时, 表示评论, 当$y_{i,j}=-1$时, 表示不评论; 当$z_{i,j}=-1$时表示不点赞, 当$z_{i,j}=+1$时表示点赞. 如图 3 所示.

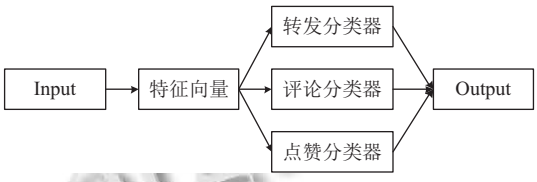



图 3 预测模型

## 5 实验结果分析

### 5.1 数据集与评价指标

数据集由北京理工大学网络搜索挖掘与安全实验室张华平博士分享的五百万微博语料提取所得, 提取规则为: 用户微博数超过 2000, 并且相邻两篇微博发布时间的间隔要小于 24 小时. 一共选取 200 位用户大概 450 000 条微博文本. 每条数据记录的属性是: uid: 用户编号, weiboId: 微博编号, created_at: 发表时间, favorited: 是否收藏, reposts_count: 转发次数, comments_count: 评价次数, attitudes_count: 点赞次数, text: 微博内容.

为了评估预测分类效果, 我们采取常见的评价标准, 准确率 $P$(Precision)、查全率 $R$(Recall) 和 $F$ 值 ($F$-

*measure*), 作为我们的评价标准, 点赞行为实验结果以表 4 的形式表示.

表 4 实验结果统计表

| | 预测点赞 | 预测未点赞 |
|---|---|---|
| 实际点赞 | *a* | *b* |
| 实际未点赞 | *c* | *d* |

那么, *P*、*R* 和 *F-measure* 的具体计算公式如下:

$$P=\frac{a}{a+c},\ R=\frac{a}{a+b},$$
$$F-measure=\frac{2\times P\times R}{P+R} \tag{6}$$

同理我们可以计算得出评论行为和转发行为的分类预测的准确率 *P*(Precision)、查全率 *R*(Recall) 和 *F* 值 (*F-measure*).

### 5.2 微博情感与兴趣实验结果与分析

在微博文本情感特征提取模型中, 我们采用的是目前短文本分类常用的分类器, 包括朴素贝叶斯 (NB)、K-近邻 (KNN)、支持向量机 (SVM)、TF-IDF 文本分类算法四种文本分析主流算法. 情感特征提取实验中采用 5 折交叉验证实验, 其平均性能如表 5 所示.

表 5 常用分类器性能对比

| 分类器 | 平均准确率 (%) | 平均召回率 (%) | 平均 F 值 (%) |
|---|---|---|---|
| NB | 77.25 | 77.86 | 77.55 |
| KNN | 83.56 | 84.25 | 83.90 |
| SVM | 85.04 | 84.92 | 84.98 |
| TF-IDF | 82.78 | 83.74 | 83.26 |

通过对比实验, 我们可以看出在情感特征提取中 SVM 分类算法表现的效果最好.

用户兴趣模型的分类结果如图 4 所示.

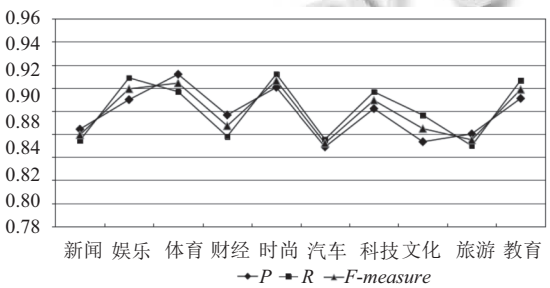



图 4 用户兴趣分类结果

### 5.3 微博用户行为预测结果与分析

#### 5.3.1 特征选取对比实验

为了验证情感特征和兴趣特征的重要性, 我们在无情感特征 (选取用户特征和兴趣特征)、无主题特征 (选取用户特征和情感特征) 和全特征 (选取全部三种特征) 三种情况下的准确率、查全率和 F-measure. 实验中是以 LibSVM 为分类器. 实验结果对比如图 5 所示.

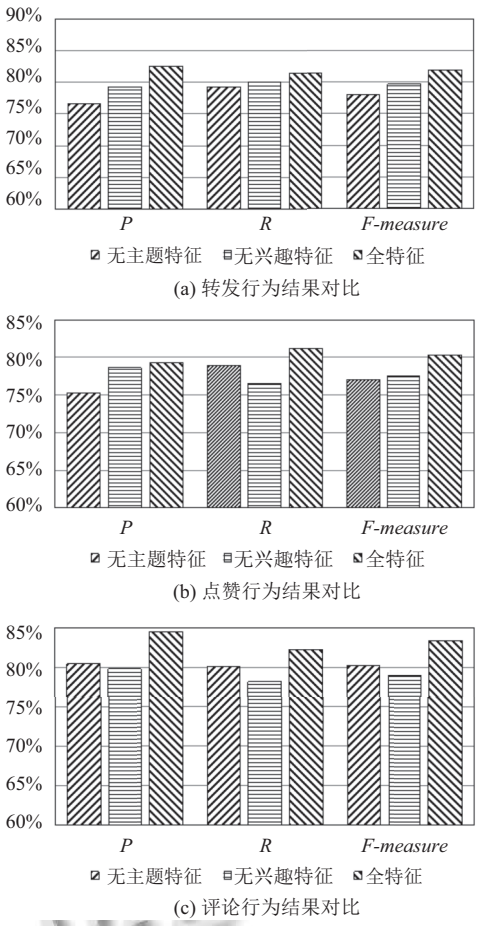



图 5 预测模型特征选取对比结果

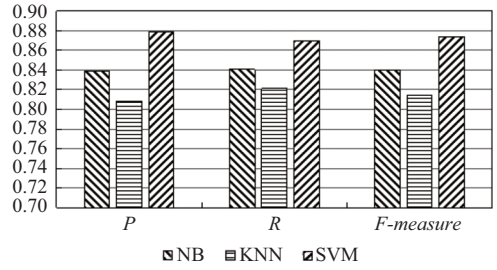
通过统计图我们可以很清楚地观察到, 在特征选取时只考虑用户情感或者兴趣, 无论是用户的点赞行为、转发行为还是评论行为预测的准确率和召回率都比全特征时的高, 因此情绪特征和兴趣特征对用户行为的预测是有效的.

#### 5.3.2 常见分类器对比实验结果分析

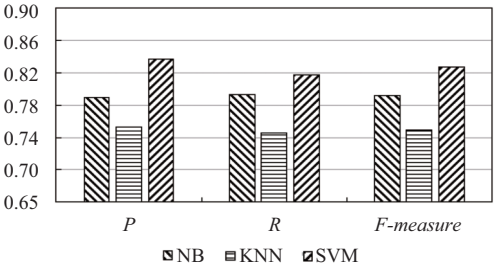
根据本文构建的微博用户预测模型, 我们分别使用朴素贝叶斯、K 近邻、支持向量机 3 种常用分类算法进行实验, 实验采取的是数据的 5 折交叉验证, 分别实现了转发、评论、点赞 3 种行为的预测分析, 如图 6 是 5 折交叉实验的平均结果.

通过实验结果对比, 我们发现行为预测模型在朴素贝叶斯和 K 近邻分类器上对用户行为预测的准确

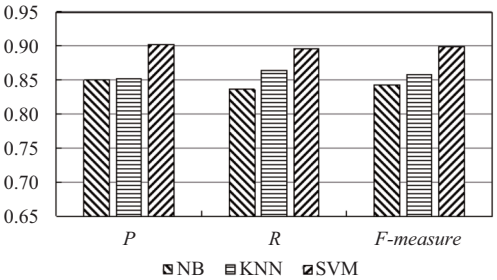
率、召回率都在 75% 以上, 在支持向量机分类算法上表现得很好, 最高的准确率接近 90%, 所以认为该微博用户行为预测模型是有效的, 但是相比较转发和点赞行为的预测结果, 评论行为的预测效果表现较差, 根据分析我们猜测评论用户对微博的关注重点与转发和点赞的用户有所差别, 比如当用户看到一些实用技巧分享的微博, 可能会进行转发或者点赞, 但是不一定会评论, 相对于评论行为, 用户可能会更加倾向于点赞和转发.




(a) 转发行为预测结果




(b) 评论行为预测果




(c) 点赞行为预测结果

图 6 行为预测结果对比

## 6 总结与下一步工作

本文主要对于用户的行为进行研究, 建立了基于用户情感和兴趣的行为预测模型, 通过使用常用的分类算法对微博用户的三种行为进行对比实验, 通过统计实验结果的准确率、召回率和 F 值, 证明了预测模型的可行性与有效性, 本文下一步工作就是将微博用户的一些历史行为、关注用户列表等因素融入到预测模型中, 也可以在微博文本分析中将图片、视频等融入到情感模型中, 提升用户情感分析的准确率.

## 参考文献

1 中国互联网络信息中心. 第 38 次《中国互联网络发展状况统计报告》. 北京: 中国互联网络信息中心, 2016.
2 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, 121.
3 Tan CH, Tang J, Sun JM, *et al*. Social action tracking via noise tolerant time-varying factor graphs. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA. 2010. 1049–1058.
4 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.
5 Xu ZH, Yang Q. Analyzing user retweet behavior on twitter. Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey. 2012. 46–50.
6 刘玮, 贺敏, 王丽宏, 等. 基于用户行为特征的微博转发预测研究. 计算机学报, 2016, 39(10): 1992–2006. [doi: 10.11897/SP.J.1016.2016.01992]
7 李志清. 基于 LDA 主题特征的微博转发预测. 情报杂志, 2015, 34(9): 158–162.
8 Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the 7th Conference on International Language Resources and Evaluation. Valleta, Malta. 2010.
9 Sriram B, Fuhry D, Demir E, *et al*. Short text classification in twitter to improve information filtering. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland. 2010. 841–842.
10 Shen W, Wang JY, Luo P, *et al*. Linking named entities in tweets with knowledge base via user interest modeling. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, IL, USA. 2013. 68–76.
11 邱云飞, 王琳颖, 邵良杉, 等. 基于微博短文本的用户兴趣建模方法. 计算机工程, 2014, 40(2): 275–279.
12 王岩. 基于共现链的微博情感分析技术的研究与实现[硕士学位论文]. 长沙: 国防科学技术大学, 2011.
13 陈文涛, 张小明, 李舟军. 构建微博用户兴趣模型的主题模型的分析. 计算机科学, 2013, 40(4): 127–130, 135.
14 王甜甜, 康宇. 方差和词向量用于文本降维的研究. 计算机系统应用, 2016, 25(11): 29–34. [doi: 10.15888/j.cnki.csa.005473]