基于属性融合的微博用户分类模型*

尹杰 张绍武 林鸿飞 魏现辉 刘晓霞
（大连理工大学信息检索研究室, 大连, 116024）

摘要：微博作为一种新媒体，每天活跃着数以亿计的用户，它为我们提供大量实时的信息。如何对具有相同兴趣爱好的用户进行分类逐渐成为研究的热点。本文提出一种多属性融合的用户分类模型，抽取用户的个人微博、社会化标签、用户简介、个人认证信息四个属性进行研究。首先将属性映射到语义空间，计算相同属性间的相似度；然后，利用本文提出的属性融合模型，对不同的属性进行融合。本文利用新浪微博名人堂中的用户信息进行分类，经实验表明多属性融合的用户分类模型更准确。

关键词：用户分类，属性融合，微博，新浪

A user classification model based on multi-attribute fusion

Jie Yin, Shaowu Zhang, Hongfei Lin, Xianhui Wei, Xiaoxia Liu
（Information Retrieval Laboratory, Dalian University of Technology, Dalian 116024）

Abstract: As one of the New Social Media, Weibo provides large amounts of information timely. Correspondingly, hundreds of millions of users may active on it every day. Obviously, more and more researches focus on how to classify the users with the same interests. In this study, we propose a user classification model based on multi-attribute fusion. The main attribute we focus on contains tweets, tags, brief and verify. First of all, we map attributes to the semantic space and calculate the similarity between the same attributes. Then, using the fusion model we proposed apply to attributes for classifying. In this paper, the database extracted from Hall of Celebrity in the Sina Weibo. In summary, our model based on multi-attribute fusion performs better than any other method in accuracy.

Key words:user classification, multi-attribute fusion, weibo, Sina, microblog

1 引言

普通的 web 网站，以组织内容为主体，即内容导向型；而社交网站的组织形式以用户为主体，当一个新用户加入，随之带来该用户的个人信息、所发布的内容以及与其他用户的链接关系[1]。因此用户量的迅猛增长，使得社交网络面临信息超载的问题。Pennacchiotti 等人认为寻找具有相同兴趣爱好的用户可以解决上述问题[2]，即用户分类可以作为信息过滤的一种手段。例如用户 Alice 对于音乐感兴趣，那么军事、房产等信息对她而言就是无价值信息，推荐给 Alice 关注的用户就应该避免出现与其不同类的用户，通过这样的方式过滤用户不期望关注的信息。

在微博平台下，影响用户分类的因素主要有三个方面：（1）用户历史行为；（2）用户个人信息；（3）网络结构特征。Rao 等人[3]对微博中用户的个人信息进行分析，通过性别、年龄、用户归属地、用户政治倾向四个方面对用户进行分类，文章利用用户的微博内容以及用户行为抽取特征词项。实验分析不同属性对于分类的作用及影响，并最终提出一种叠加的 SVM 分类方法。随后，Pennacchiotti 等人[4]提出一种机器学习的方法，该方法对于用户分类

*国家自然科学基金资助项目(编号：60973068、61277370)、教育部博士点基金(编号：2009004111002)和辽宁省自然科学基金（201202031）。作者简介：尹杰，女，硕士生，研究方向为情感分析与观点挖掘；张绍武，男，博士，副教授，硕士生导师，研究方向为搜索引擎、文本挖掘；林鸿飞，男，博士，教授，博士生导师，研究方向为搜索引擎、文本挖掘、情感计算和自然语言理解，hflin@dlut.edu.cn；魏现辉，男，硕士生，研究方向情感分析与观点挖掘；刘晓霞，女，硕士生，研究方向为社会计算。

问题的研究依据四个方面，包括用户的个人信息（性别、兴趣描述等）、用户的历史行为（评论、转发等）、用户发布微博的语言学特征以及用户的网络结构。相较于 Rao[3]等人的方法，Pennacchiotti 等人充分利用了用户携带的信息量以及平台的网络特征，综合衡量用户的价值。除此之外，Malouf 和 Mullen[5]提出一种基于图的用户分类方法，对用户的政治倾向进行分类，该方法主要利用微博内容，从消息中抽取情感词，利用情感倾向性将用户分为左派、右派和中立。Malouf 和 Mullen 首次将用户之间的关系网络引入分类中。与以往研究不同，Chen 等人[6]从信息推动的角度对用户的个人信息更新和分类问题进行了研究。他们认为用户的属性中只有一部分属性是决定分类效果优劣的关键，将其定为主属性包括时间属性、内容属性、关系属性。其余属性对于用户的兴趣分类影响较小，例如性别等常规因素。

上述研究均基于 Twitter，正如 Guo 等人[7]所言，中国的微博平台相对于 Twitter 存在一些特殊性。Wu 和 Wang[8]针对新浪微博做了相关研究，他们基于微博的内容对用户进行分类，参考 Google Directory 对 Web 网站的分类，文章将用户类别定义为 8 类，用户相似度计算采用互信息作为依据，最终利用朴素贝叶斯模型进行分类。随后，Zhou 等人[9]利用微博内容和用户行为对新浪微博加 V 用户的职业进行分类。抽取的用户行为主要有“@某人”和“#某话题”两类。实验将用户职业分为四类，包括体育、娱乐、IT 以及房产，最终抽取 21974 个特征，并使用 SVM 对用户进行分类。

本文通过抽取用户的微博历史记录、用户的行为以及个人信息对用户进行分类。首先将抽取的信息以用户微博、社会化标签、个人简介以及认证信息四种属性进行表示；其次，研究每一种属性对于用户分类的贡献；最后，综合考虑多种属性对于分类的影响，并提出一种基于属性融合的用户分类模型。同时，针对微博的短文本特性，本文通过引入同义词林的方式进行解决。

文章的组织结构如下：第二部分对研究方法中的相关工作进行说明，第三部分详细介绍本文的主要研究方法，第四部分是实验结果及分析，第五部分进行总结。

2 相关工作

2.1 用户信息

本文将微博平台下用户的信息分为两种类别，第一，显性信息，例如，用户的性别、年龄以及用户标签等，这些信息可以直接获取；第二，隐性信息，例如，用户参与的话题、@某人之间的互动、转发或评论行为等，这些信息不能直接获取，需要通过一些自然语言处理的方法在用户发布的消息记录上进行挖掘。

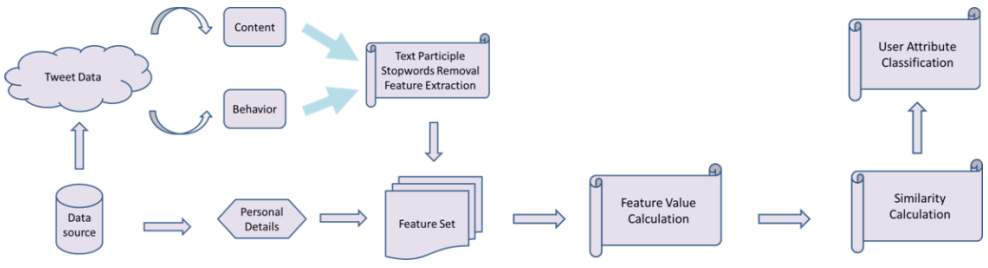

图 1　用户分类框架

Fig.1　The frame of user classification

2.1.1 用户隐性信息

本文对于用户建模时充分考虑微博信息的价值，从用户行为消息和常规微博消息两个方面抽取信息。首先将用户微博中含有@某人信息的微博单独提取，形成$content_1$，再将用户

微博中含有#话题类型的微博单独提取，形成 $content_2$；将剩余的微博作为 $content_3$。$content_1$和$content_2$属于用户行为产生的内容即图 1 中的 Behavior，$content_3$是常规的用户微博消息，即图 1 中的 Content，本文对三个方面的信息单独处理，分别抽取特征词，每个部分单独计算特征词权重，并最终形成微博部分特征词的向量。具体步骤如图 1 中对于 Tweet Data 的处理流程。

2.1.2 用户显性信息

针对分类问题，本文选用的显性信息包括三个方面，一是用户的个人简介；二是用户的标签信息；三是用户的认证信息。同时，对于三个方面信息的处理流程如图 1 中 Personal Details 所示。

（1）一般而言，用户通过简短的句子完成个人简介的描述，内容涉及用户的职业、兴趣、生活理念等，本文利用文本挖掘的方法有针对的提取用户简介中的特征词，形成简介的特征向量。

（2）标签作为新型的标注方式，同样被微博平台引入。用户对于标签的选择有两方面来源，一是平台自动提供，用户选择符合兴趣爱好的标签进行匹配；二是用户自由编写。微博平台提供的这种标签张贴形式带来新的问题，相同语义的内容，用户通过不同的词语进行标注，本文提出同义词林扩充的方法解决该问题。

（3）用户认证作为微博平台特有的方式，对于权威用户是一种官方的认可。本文主要关注用户被认证时的官方说法，即用户的认证信息，该内容主要涉及用户的职业，用户的贡献。

2.2 微博用户特性

2.2.1 用户关系

假设U_s、U_t为图$G(U,E)$中无直接连接的两个节点，即$U_s, U_t \in U$且$U_sU_t \notin E, U_tU_s \notin E$，$U_s$由于某种需求希望主动建立与$U_t$的联系，那么，$U_s$可以通过关注的方式形成一条指向$U_t$的连接，即$U_sU_t \in E$；同时对于$U_t$而言，入度增加 1 表示新增一个粉丝，即$U_s$是$U_t$的粉丝。

一般情况下，用户之间的关系抽象为网络中的边[10]，这些直接的连接可以是朋友、相同的兴趣爱好、甚至是为了获取某一方面的信息而形成的联系[11]。简而言之，微博中用户 A 创建与其他用户之前的连接，当用户 A 发布短文本信息时，就会出现信息的传播。微博不同于其他社交网络，用户 A 关注用户 B，他不需要得到 B 的允许，这就使得网络中出现一些特殊的节点，他们的入度很大[12]，即受到广泛的关注，本文将此类节点称为名人。该类特殊群体一般使用微博时间较长，用户的行为和发布的消息具有典型性，例如一条信息经由 *ID* 高的节点发布或转发，将会出现信息大量的转发，即 *ID* 高的节点更利于消息的传播[6]，因而本文将此类用户定为研究对象。

2.2.2 用户类别

本文从文本处理的角度为用户建模，将用户 U_i携带的所有信息以文本形式表示，不同属性以段落区分，则用户 U_i的第k个属性信息 U_i^k可以理解为整个文本的第k个独立段落。本文抽取用户众多属性中的 4 个属性进行处理和分析，即用户 U_i作为独立文本拥有 4 个段落，同时，本文将用户定义为多元组，形式如下：

$$U_i\langle U_i^1 \quad U_i^2 \quad U_i^3 \quad U_i^4\rangle = \langle\, tweet \quad tag \quad brief \quad verify \,\rangle \tag{1}$$

其中，*tweet* 表示个人微博、*tag* 表示社会化标签、*brief* 表示用户简介、*verify* 表示个人认证信息。个人微博作为一个用户携带信息的主体部分，承载着该用户在日常生活以及专业领域的所有信息，该部分内容由图 1 中的 Content 和 Behavior 组成，是分类的重要依据；社会化

标签是用户自主选择、贴给自己的标签，更有针对性的体现个人的喜好以及从事行业等；用户简介从主观的角度展现用户，与其余反映用户行为的方式相比，更直观、准确；认证信息作为是否为名人的重要指标，是借由新浪已有的认证体系，方便快速的定位用户的类别。对于用户归属的类别，本文选取 6 类，表示如下：

$$C\langle\sum_{i=1}^{6} C_i\rangle = \langle 娱乐 \quad 财经 \quad 汽车 \quad 房产 \quad IT \quad 军事\rangle \tag{2}$$

3 基于属性融合的用户分类

3.1 短文本处理

用户发布的微博是典型的短文本，考虑到短文本具有文本长度短、描述概念信号弱的特点, 因此，单纯的基于词项进行特征选择容易使文本表达主题分散，核心词易被赋予较低权重。上述问题的关键在于核心词的词频较低，语义不统一。

对于短文本的分类，学者们已经做了很多的尝试。Banerjee 等人[13]引入 Wikipedia 作为外部资源，查询 Wikipedia 中与已有特征词共现最多的词作为扩展特征词加入选取的特征词中，从而扩充特征词，提高了短文本聚类的准确性。Schonhofen[14]提出类似的方法，也是通过引入 Wikipedia 作为外部资源，进而扩充主题词，提高聚类的准确性。Phan 等人[15]利用 LDA 主题模型进行短文本、稀疏文本的分类，从而提高分类效果。本文对于该问题的解决参考前人的工作，引入同义词林进行核心词的扩充，从而有效的解决上述问题。

在梅家驹等人所撰写的《同义词词林》[16]的基础上，本文构造了一个支持微博的同义词词库，该词库主要收录微博中类别区分度高、专业性强的同义词。

3.2 属性计算

构建用户分类模型首先要针对用户进行建模；其次，对于抽取的属性进行特征词提取；最后，根据权重计算公式获得特征词的权重，从而构建属性特征向量。

（1）特征提取

对于类别信息为多类的分类问题，特征提取的目的不仅是寻找某用户具有代表特性的特征词，还需要考虑提取的特征词是否具有类别区分度。本文经过实验最终选择 χ^2 统计量（CHI）为特征选择方法，具体实验结果后续展示。

$$\mathrm{CHI}(t_k, C_i) = \frac{N\times[P(t_k,C_i)\times P(\overline{t_k},\overline{C_i}) - P(\overline{t_k},C_i)\times P(t_k,\overline{C_i})]}{P(t_k)\times P(C_i)\times P(\overline{t_k})\times P(\overline{C_i})} \approx \frac{N\times(A\times D - C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{3}$$

其中，N为用户的数目，A表示词项t_k与类别 C_i同时出现次数，B表示词项 t_k出现且类别C_i不出现次数，C表示词项 t_k不出现且类别C_i出现次数，D表示词项 t_k不出现且类别C_i不出现的次数。

（2）权重计算

对于类别区分度高的特征词，本文利用$TF\cdot IUF$进行特征词的权重计算，该公式为$TF\cdot IDF$的变形，即将用户信息表示成文本，该公式进行了平滑处理。

$$W\left(U_i^j, t_k\right) = \frac{tf_{U_i^j,t_k}\times iuf_{U_i^j,t_k}}{\sqrt{\sum_{k=1}^{n}\left[tf_{U_i^j,t_k}\times iuf_{U_i^j,t_k}\right]^2}} = \frac{\left(log\left(f_{U_i^j,t_k}\right)+1.0\right)\times log\left(N/n_{t_k}\right)}{\sqrt{\sum_{k=1}^{n}\left[\left(log\left(f_{U_i^j,t_k}\right)+1.0\right)\times log\left(N/n_{t_k}\right)\right]^2}} \tag{4}$$

其中，$W\left(U_i^j, t_k\right)$表示U_i的第 j 个属性中的特征词t_k的权重。$tf_{U_i^j,t_k}$表示用户U_i的第 j 个

属性中词项t_k的词项频率，$iuf_{U_i^j,t_k}$表示用户集合上词项t_k的重要性，即越多的用户信息中含有词项 t_k，说明该词项越重要。平滑方面，词项频率$tf_{U_i^j,t_k}$加 1 是为了保证频率为 1 的词项具有非零权值。

（3）属性空间向量表示

针对用户建模的最后步骤是将不同属性映射到语义空间。用户集合 U 的第 j 个属性，即U^j中所有词项构成词项集合$Term_{U^j}$，对于$\forall\, t_k \in U_i^j, U_i^j \in U^j$，用户$U_i$的第 j 个属性U_i^j向量定义为：

$$attr\left(U_i^j, t_k\right) = \begin{cases} W\left(U_i^j, t_k\right) & t_k \in Term_{U^j} \\ 0 & t_k \notin Term_{U^j} \end{cases} \quad (5)$$

其中，$W\left(U_i^j, t_k\right)$为公式（4）中定义的词项t_k的权重。

3.3 属性融合

本文对于用户分类的讨论基于用户的四个属性，首先分析每一种属性对于用户分类的影响；其次，利用属性之间相互融合，分析融合属性对于用户分类的影响。本文采用的分类方法为朴素贝叶斯。

（1）线性属性融合

本文采用的线性融合方法比较简单，对于$\forall\, t_k \in U_i^j, U_i^j \in U^j$，方法公式如下：

$$Merge(U_i, t_k) = \sum_{j=1}^{4} \lambda_j \times attr\left(U_i^j, t_k\right) \quad (6)$$

其中，λ_j分别代表微博、标签、简介、认证四个属性的相关性系数。为了最大化特征融合的效果，本文对于参数的设置采取平均的方式，即融合过程中，每种属性的重要性被看为相同的，此时，λ_j取值为所选属性个数的平均值。

（2）分类方法

本文采用朴素贝叶斯进行用户属性的分类，贝叶斯（Bayes）是一种基于概率的分类算法，以贝叶斯定理为理论基础，以条件概率的形式展现两种实体之间的关系，进而计算类别属性。

4 实验结果与分析

4.1 语料来源及实验流程

表 1 类别中的用户信息

Tab.1user information of each category

用户类别	用户数目
娱乐	140
财经	159
汽车	120
房产	120
IT	120
军事	100

本文数据集来自于新浪微博名人堂，选取 6 个类别进行爬取，即得到的数据集中用户的类别信息已标注。每个类别中用户的选择来自“分类人气榜 TOP20”。爬取过程中，由于每个大类中所含小类个数的不同，每个大类在最终选定的用户上呈现不同的数目，统计量如表 1 所示。用户的基本信息包括个人微博、社会化标签、用户简介、个人认证信息。为了更准确的展现数据集的信息，本文详细的统计数据集信息如表 2 所示。

表 2 数据集统计

Tab.2statistics of dataset

内容	数量	平均值
用户	759	1.00
微博	1806316	2379.86（Max=33458）
标签	5255	6.92
简介	642	0.85
认证	759	1.00
关注链	380336	501.10（Max=2000）
粉丝链	649006342	855080.82（Max=14707929）

实验流程如下：第一，对数据集进行初步处理，去除无用微博，例如只包含“转发微博”的微博信息，并将用户的四部分信息进行分离；数据集的分词工作选用中科院的分词系统 ICTCLAS。第二，为了使得本文的实验重点集中在特征融合对于分类的影响上，在语料预处理阶段，首先，确定停用词对于分类的影响；其次，通过引入同义词林的方式解决 3.1 中短文本概念词稀疏的问题；最后，确定不同的特征提取方法对于分类的影响。后续所有实验均在三种影响因素最优下进行。随后，通过实验分析词性标注、情感本体对于分类的影响。第三，属性对于用户分类的影响分两个方面展示，单一属性以及属性融合对于分类的影响，并作全面的分析。

本文采用五倍交叉验证方法来进行实验，即将 80%的用户信息作为训练集，另外 20%的用户信息作为测试集。用户文件使用训练集的数据进行文件表示，实验目标是训练一个分类模型，从而正确地将用户分类到标注的集合中。

4.2 评价方法

本文采用如下四种常见的方法对分类结果进行评价：

（1）Precision：准确率评价分类正确的用户占所有用户的概率，该数值越大，说明分类效果越好，本文用 *P* 代表准确率。

（2）Recall：召回率评价分类正确的用户占已标注的该类用户的概率，该数值越大，说明分类效果越好。本文用 *R* 代表召回率。

（3）F-Measure：F 值度量是综合召回率和准确率效果的评价方法，并定义为召回率和准确率的调和平均数，该数值越大，说明分类效果越好，本文用 *F* 代表 F 值度量[17]。

$$F = \frac{1}{\frac{1}{2}\times\left(\frac{1}{R}+\frac{1}{P}\right)} = \frac{2RP}{(R+P)} \tag{7}$$

（4）ROC Area：ROC 曲线由 TP(True positive rate)、FP(Flase positive rate)构成，以 TP 作为 Y 轴，FP 作为 X 轴。通常，分类器的分类效果越好它所对应的 TP 值就越高、FP 越低，映射到 ROC 空间，即 ROC 曲线下面覆盖的面积越大分类效果越好[18]。

4.3 实验结果及分析

4.3.1 语料预处理影响

（1）停用词对于分类的影响：

停用词 Stopwords 在常规文本处理时，通常被认为是噪音[19]；但考虑到微博短文本的特性，为了使研究结果更严密、可信，本文通过实验明确停用词对于分类效果的影响。

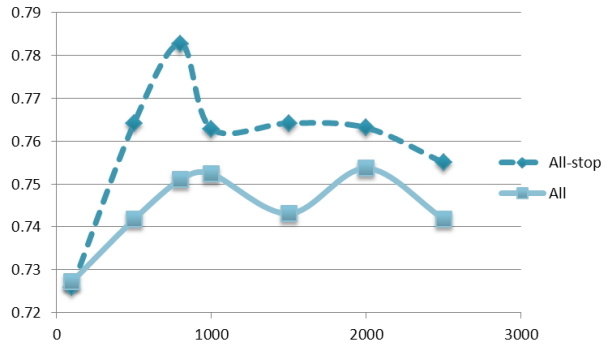

图 2 停用词对于分类影响

Fig.2The influence of stopwords on classification

其中，X 轴为选取的用户特征词项的个数，Y 轴为分类的 *F* 值；实线 All 为去除停用此前，相应虚线 All-stop 为去除停用词后。从图 2 可以看出，短文本处理和常规文本处理一样，停用词均为用户分类中的噪音，需要进行去停的处理。因此，下文实验均在去除停用词的数据集上进行。

（2）同义词林对于用户分类的影响：

引入同义词林前后，本文实验在整个数据集上进行。首先在相同的条件下进行实验，即 Cilin(tf· idf)与 All(tf· idf)；为了比拟引入同义词林之后的结果，图 3 展示了引入同义词林前的最好结果 All(Chi-square)。

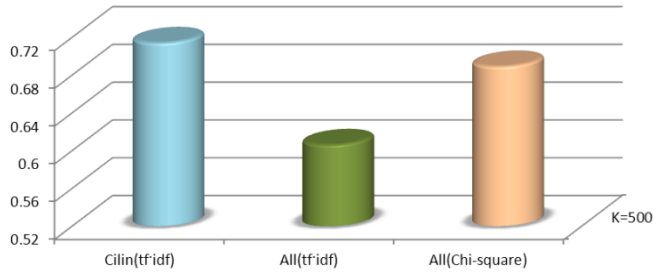

图 3 同义词林对于分类影响

Fig.3 The influence of Tongyici Cilin on classification

其中，K=500 表示三组实验均取 500 个词项为特征。该 K 值的选择依据引入同义词林后最好的组别 All(Chi-square)具有的特征项个数为基准。从图 3 可知，引入同义词林 Cilin 的分类效果明显好于其他组别，尤其是同样以 tf·idf 为权重计算公式的组别 All(tf·idf)；即使与未引入同义词林的最好组别 All(Chi-square)相比，引入同义词林也使得分类效果更加明显。实验表明，引入同义词林解决了短文本“文本长度短、描述概念信号弱的特点，文本表达主题分散”的问题。

（3）特征提取方法对于用户分类的影响

特征提取作为文本处理的基本方法，其价值不言而喻。对用户进行建模时，特征词的选择尤为重要，直接影响分类的结果。

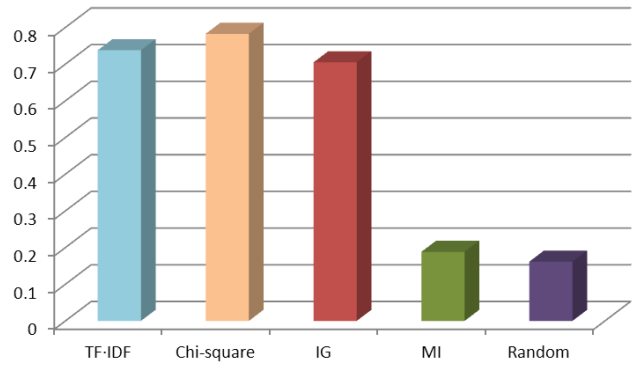

图 4 特征提取对于分类影响

Fig.4 The influence of feature extraction on classification

结果差异源于不同的特征提取方法侧重点不同。由图 4 可知，Random 方法作为 baseline，效果最差，从侧面说明进行特征提取对于分类是必要的；MI 的效果明显不好原因在于只考虑文档的 df，而对于每个词项在用户文档中的 tf 没考虑；Chi-square 的效果最好，在于它兼顾了词项的 tf 和 df。本文下述实现均选择 Chi-square 作为特征提取方法。

4.3.2 基于词性标注的用户分类

词性作为重要的文本特征，是语义网络的基础，本文期望通过实验研究词性对于用户分类的影响。由于介词、连词、语气词多为停用词，已经由实验证明为分类过程中的噪音，因此，本文主要选取名词、形容词、动词、副词作为影响用户分类的词性。

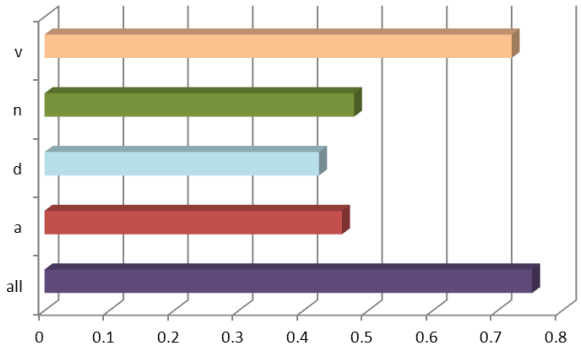

图 5 词性标注对于分类影响

Fig.5 The influence of POS on classification

从图 5 可知，动词和名词对用户分类的影响最大，而动词分类的影响高于名词的原因是中科院的系统 ICTCLAS 对于一些既是名词又是动词的词项标注为动词。例如，销售、建筑、赛车、编程均是兼有名词和动词词性的词项，而 ICTCLAS 均将其标注为动词，从而出现了动词的影响高于名词。

4.3.3 基于情感本体的用户分类

用户通过微博分享个人的观点、发布信息，使得网络中储藏着大量的具有情感倾向性信息的内容。本文从这些情感倾向性的消息入手，通过实验分析情感对于分类的影响。

表 3 情感对于分类影响

Tab.3 The influence of POS on classification

指标/特征数	Precision	Recall	F-Measure	ROC Area
50*6	0.619	0.607	**0.579**	0.762
100*6	0.626	0.607	0.579	0.761
200*6	0.617	0.588	0.557	0.748
500*6	0.608	0.547	0.516	0.722
1000*6	0.598	0.499	0.455	0.691
1500*6	0.594	0.476	**0.414**	0.676

其中，表 3 中最左一列代表每个实验组别选取特征的个数，其余列代表不同评价指标下的分类效果，每一行展示同一组别的不同评价指标效果。为了充分说明情感对于分类的影响，本文选择的特征数范围较大$[50*6,1500*6]$。由 F-Measure 一项可知，抽取情感词作为特征，并不能很好的区分不同的类别。实验结果说明微博平台下的用户无论处于哪种类别，都会发表大量关乎心情的微博，使得情感词对于分类的效果不明显。

4.3.4 基于属性融合的用户分类

(1) 单一属性对于用户分类的影响

表 4 单一属性对于分类影响

Tab.4 The influence of single attribute on classification

组别	属性	特征个数	Precision	Recall	F-Measure	ROC Area
1	tweet	200*6	0.800	0.768	0.767	0.857
2	tag	70*6	0.718	0.585	0.599	0.747
3	brief	50*6	0.695	0.555	0.548	0.721
4	verify	100*6	0.826	0.823	**0.822**	0.894

由表 4 可知，从单一属性考虑，认证信息 verify 分类效果最好，主要原因是认证信息经过新浪微博的官方过滤，相当于官方给用户的标签，因此对于用户的定位更加准确，并隐含有类别信息。相比之下，用户的简介 brief 分类效果较差，主要原因是用户对于简介的用途理解各异，有些用户在个人简介部分描述自己的兴趣爱好，很可能这些信息与他所属的类别无关，相当于在简介这个属性中引入了噪音，从而降低了分类的准确率。更深入探究，除去 verify 这样的专业信息，从普遍意义上考虑，用户微博 tweet 蕴含大量的可挖掘信息，即通过用户的日常行为可以很好的预测用户所属类别。标签 tag 属性分类效果相对较差的原因在于标签稀疏性问题，用户为自己贴上的标签带有随意性，同一个语义可能使用不同的词语表示，因而分类效果较差。

(2) 多属性融合对于用户分类的影响

为了更充分的利用用户的信息，除了单一属性的实验，本文对多属性融合的组别进行多种形式的组合，从两个属性融合到四个属性融合共进行 11 组实验，具体信息如下。

表 5 多属性对于分类影响

Tab.5 The influence of multiple attribute on classification

组别	属性	特征个数	Precision	Recall	F-Measure	ROC Area
1	tweet+tag+brief+verify	(200+70+50+100)*6	0.874	0.872	**0.871**	0.922
2	tweet+tag+verify	(200+70+100)*6	0.870	0.869	**0.868**	0.921
3	brief+tag+verify	(50+70+100)*6	0.869	0.868	0.867	0.920
4	tweet+brief+verify	(200+50+100)*6	0.852	0.850	0.850	0.909
5	tweet+tag+brief	(200+70+50)*6	0.772	0.767	0.765	0.858
6	tag+verify	(70+100)*6	0.866	0.865	**0.864**	0.919
7	brief+verify	(50+100)*6	0.848	0.846	0.846	0.907
8	tweet+verify	(200+100)*6	0.838	0.836	0.833	0.901
9	tweet+tag	(200+70)*6	0.792	0.788	0.787	0.872
10	tweet+brief	(200+50)*6	0.732	0.721	0.716	0.830
11	tag+brief	(70+50)*6	0.723	0.674	0.674	0.801

本文多属性融合实验部分，充分考虑每一种属性可能的组合方式，并根据 3.3.3 中的融合方法进行计算，最终使用 3.4.2 中的评价指标进行分析。从表 3.5 可知，四类属性同时作用时分类效果最好，即本文提出的属性融合公式；但是综合各种属性组合方式发现，并不是选取的属性越多分类效果越好，而是体现为某一类属性对于分类的效果影响较大的现象。本文 verify 对于分类的影响较大，含有 verify 的组别分类效果明显好于其他组别，多属性融合的 11 组实验结果中，7 组含有 verify 的组别在同类量级的融合中效果更好。同时，tweet 作为文本的主体部分，含有 tweet 的组别对于分类影响次之，从最后三组实验结果可见一斑。

纵观单属性和多属性的 15 组实验结果，多属性融合的实验结果 F 值 0.871 明显好于单属性最好结果 F 值 0.822，即实验结果说明属性融合的用户分类方法可以更好的体现用户所属的类别。

5 总结与展望

微博作为一种维系社会关系、获取信息的手段，受到越来越多人的关注。名人作为微博网络中的一个特殊群体，对于信息的发布、传播等具有重要的影响。随着网络中用户量的快速增长，如何寻找具有相同兴趣爱好的用户成为一个研究热点。本文以新浪微博作为载体，抽取用户众多属性中的个人微博、社会化标签、用户简介、认证信息四个属性进行研究。通过剖析不同的属性对于用户分类的影响，并最终提出基于属性融合的用户分类方法。实验表明，本文提出的方法更好的适用于用户分类。同时，本文从词性角度和情感角度对用户进行分类，展现并分析各组实验结果。

进一步工作有三点：第一，本文采用的属性融合方法是最简单的线性融合，下一步可以考虑采用逻辑回归等方法进行属性的融合；第二，对于属性的选择，本文抽取四种，下一步工作中可以抽取更多属性进行研究；第三，每类属性中参数的选择，本文选择最简单的平均形式，下一步可以优化参数，提高对分类影响大的属性的权重。

参 考 文 献

[1] Pan S J, Ni X, Sun J T, et al. Cross-Domain Sentiment ClassificationVia Spectral Feature Alignment[C]. In Proceedings of the 19th International Conference on World Wide Web. Raleigh, NC, USA, 2010: 751-760.

[2]Pennacchiotti M, Silvestri F, Vahabi H, et al. Making Your Interests Follow You on Twitter[C]. In Proceeding of the International Conference on Information and Knowledge Management (CIKM). New York, NY, USA, 2012: 165-174.

[3]Rao D, Yarowsky D, Shreevats A, et al. Classifying latent user attributes in twitter[C]. In Proceedings of the 2nd international workshop on Search and mining user-generated contents. New York, NY, USA, 2010: 37-44.

[4]Pennacchiotti M, Popescu A M. A machine learning approach to twitter user classification[C].In Proceedings of Fifth International AAAI Conference on Weblogs and Social Media (ICWSM). Barcelona, Spain, 2011:281-288.

[5]Malouf R, Mullen T. Graph-based user classification for informal online political discourse[C]. In Proceedings of the 1st Workshop on Information Credibility on the Web (WICOW). Miyazaki, Japan, 2007.

[6]Chen C, Liao G P, Shi X H, et al. The Model of User Interest Update and User Classification in Personal Information Push Service [J]. Procedia Environmental Sciences, 2011, 10: 262-268.

[7]Guo Z, Li Z, Tu H. Sina Microblog: An Information - driven Online Social Network[C]. In Proceedings of the International Conference Cyberworlds. Banff, Canada, 2011: 160-167.

[8]Wu X, Wang J. Micro-blog in China: identify influential users and automatically classify posts on Sina micro-blog[J]. Journal of Ambient Intelligence and Humanized Computing, 2012: 37-42.

[9]Zhou M, Xu Y, Zhao X. Study of Feature Extract on Microblog User Occupation Classification[C]. Information Science and Engineering (ISISE), 2012 International Symposium on.IEEE. Shanghai, China, 2012: 20-23.

[10] Moreno, J.L. Emotions mapped by new geography[J]. New York Times, 1933.

[11] M. Cha, H. Haddadi, F. Benevenuto, et.al. Measuring User Influence in Twitter: The Million Follower Fallacy [C].In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.Washington, DC, USA, 2010: 10-17.

[12]Kwee A T, Lim E P, Achananuparp P, et al. Follow Link Seeking Strategy -- A Pattern Based Approach [C]. In Proceedings of the 6th SNA-KDD Workshop on Social Network Mining and Analysis. Beijing, China, 2012.

[13]Banerjee S, Ramanathan K, Gupta A. Clustering short texts using Wikipedia[C]. In Proceedings of the 30th annual international ACM SIGIR conference. New York, NY, USA, 2007: 787-788.

[14]P. Schonhofen. Identifying Document Topics Using the Wikipedia Category Network[C].In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC, USA, 2006: 456-462.

[15]Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]. In Proceeding of the 17th international conference on World Wide Web. New York, NY, USA, 2008: 91-100.

[16]百度, 百度百科 [EB/OL]. http://baike.baidu.com/view/1355899.htm.

[17]Chai J Y, Sun J. Hybrid FEMfor deformation of soft tissues in surgery simulation[C]. Medical Imaging and Augmented Reality. Washington, DC, USA, 2001: 298-303.

[18]Xiang G, Fan B, Wang L, et al. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus[C]. In Proceeding of the International Conference on Information and Knowledge Management (CIKM). New York, NY, USA, 2012: 1980-1984.

[19] Hong L, Davison B D. A classification-based approach to question answering In discussion boards[C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, 2009: 171-178.