基于 LDA 模型和 SVM 方法的微博用户性别判别

孙世杰，李珠峰，濮建忠

（解放军外国语学院，河南洛阳 471000）

摘要：本文提出了一种微博用户性别判别的方法。通过对于不同性别用户的微博进行聚类，得到相关的区别性特征，并据此训练出对于用户性别进行分类的分类器。在构造过程中使用 LDA 模型充分挖掘不同性别用户之间的区别性特征，并利用 SVM 模型对所发现的特征进行学习。实验结果表明本方法能够较好地对不同性别的用户进行区分，得到相关的性别信息。

关键词：微博；用户信息识别；性别判别

Microblog users' gender discrimination based on the LDA model and the SVM method

SUN Shijie，LI Zhufeng，PU Jianzhong

（PLA University of Foreign Languages，Luoyang，Henan，471000，P. R. China）

Abstract：This paper presents a method aimed at discriminating the gender of micro-blog users. Relevant distinctive features can be obtained by classifying and clustering the gender of different users, and based on this a gender classifier comes into being. During the process of construction, the LDA model can be used to fully exploit the distinctive features between users of different gender, and the SVM model can be utilized for studying the characteristics which have been found. Experimental results show that this method can make a better distinction between users of different gender and gain relevant information on gender.

Key words：microblog; users' information identification; gender discrimination

1 引言

近年来，随着互联网的普及和相关技术的蓬勃发展，互联网的形式也已经发生了改变，由之前静态的、以网页信息为主的平台转化为由大量个体用户为主角的动态平台。在这一改变的过程中，微博以其较低的书写门槛和便捷的客户端组件形式成为了重要的推手。同时由于参与的网民众多，在微博平台上产生了大量的数据。这些数据使得通过相关的统计模型对语言特征与社会分类（social variables）之间的关联进行研究，成为可能。不论这些研究的目的是在于了解和发现这些形式上的差别，还是希望通过学习产生一个对隐含特征（latent features）进行判别的模型，这些研究都有一个暗含的假设，认为语言的选择与特定种群的人之间存在着相对稳定的关联。许多学者借助微博开展相关方面的研究，尤其是对用户性别和年龄与其语言特征关系的研究[1-4]。而对于通过这些特征对相关文本进行准确的判别，对

* **作者简介：**孙世杰（1988-），男，博士在读，主要研究方向：语言信息处理；李珠峰（1983-），男，博士在读，主要研究方向：语言信息处理；濮建忠（1968-），男，教授，主要研究方向：语料库语言学，语言信息处理。

于市场营销、个性化服务甚至法律调查等方面都有着十分重要的实用价值。

在本研究中，我们希望构建一个能够较好识别微博用户性别的分类器，并在研究的过程中对不同性别的人群的语言特征进行分类和总结。在研究中，我们将性别的识别问题看做是一个二分的问题，同时利用基于文本的特征对其进行分类研究。在第二部分将对算法进行说明。第三部分将对所得的试验结果进行分析和说明。第五部分将对我们研究的结果进行总结，并展望下一步的工作。

2 性别分类算法

根据微博的内容对微博用户进行性别的判别，本质上是一个分类的问题，而进行分类的依据就是根据不同性别用户之间的语言差异。因此对于特征的选取是构建分类器的关键，由于微博自身语言及其环境的特殊性，同时区别性特征并不仅仅由单个的词语反映，还会因为关注领域的不同等特征，反映出用户的性别。因此我们利用训练集通过聚类得到不同性别用户的区别性特征，并依此构建出分类器。本研究的算法流程如下（图 1）：

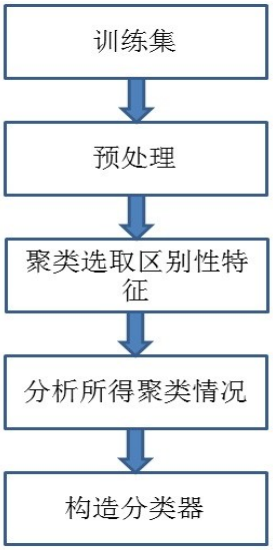

图 1 算法步骤

2.1 特征选取

2.1.1 研究现状

在微博中有很多的可以用于区分各类不同微博的特性，这些特性包括了时间印记（timestamps），用户界面的设置，头像，图标等，但是在本研究中我们所关注的主要是微博文字本身所具有的语言特征。

在我们已知的范围内，没有学者针对中文微博的用户性别判别开展研究，但之前一些学者对于 Twitter 中的性别判别进行的研究有一定的借鉴意义[5-9]，其中所包含的常见的区别性特征主要有如下几类（表 1）：

表 1 先前研究中影响性别分类的特征

特征	示例	先前研究结果
代词（Pronouns）	you, u, I	女性特征
情感短语（Emotion Terms）	sad, glad, sick	女性特征
家庭短语（Family Terms）	mom, mommy, sister	女性特征
计算机媒介使用语言（CMC words）	lol, omg	女性特征
连接词（Conjunctions）	and, but, or	女性特征
附着语素（Clitics）	she'll, he's	女性特征
冠词（Articles）	a, an, the	男性特征
数字（Numbers）	1, 2-1, 500	男性特征
量词（Quanlifiers）	piece of, block of	男性特征
科技词语（Technology words）	solar energy	男性特征
前置词（Prepositions）	in, into, on	混合特征
脏话（Swear words）	damn, fuck, hell	混合特征
赞同词（Assent）	okay, yes, yess	混合特征
反对词（Negation）	no, nope	混合特征
表情符号（Emoticons）	:), :0	混合特征
怀疑的标记（Hesitation marker）	ur, er	混合特征

这些分类和特征的选取，大多是依靠社会语言学家们早期在人们交流过程中，以及在网络时代的早期基于相关的博客进行统计总结所得到的结果。囿于数据量的限制以及语言随时代的快速变化，其中的一些特征已经发生了变化。而一些针对微博开展的研究多数是基于英文的，其中 Burger（2011）的研究虽然考虑到了中文的情况，但是对其的处理只是简单地通过字一级进行研究，通过其结果也可以清晰地看到这一处理对作者最终的结论并未产生太大的影响。

2.1.2 针对中文微博的特征选取

针对中文自身是非粘着语，在处理中首先进行了分词，同时为了便于对其中的特征进行总结，对分词之后的文本进行了标注。由于处理的文本集是微博，其中的未登录词、网络用语、不规范的用词数量较多，因此在进行分词和标注过程中，我们通过人工总结出了微博中常用的特殊词语以及相关的符号集对 ICTCLAS 进行了改进，以使得最终所得到的特征集能够真正体现出不同性别使用者的语言特征，其中包含 65 个词语和 11 个符号。所扩充的词语和符号集部分如下表所示（表 2）：

表 2 扩充词语及符号集

词语	符号集
给力、神马、围观、纠结、淡定、浮云、犀利、鸭梨、你懂的……	“:)”、“;)”、“:-)”、“:D”、“;]”、“:]”、“:P”、“;P”、“:(”、“:-(”、“:\”

在进行切词和词性标注之后，与通常采用对相关的特征人工进行总结然后再将其在试验数据上进行验证不同，我们将这些标注的词语作为特征集，将其进行聚类，通过聚类的结果得到在中文微博中的男女用户的语言上存在着不同的特点，同时在聚类过程中，通过不同类中不同性别的用户的分布不同得到分类器的机器学习训练集。

在本试验中我们采用 LDA 模型作为聚类方法，同时将一个用户的所有微博聚合成一个大的用户文档，从而刻画出用户层面的背景。LDA 模型作为一个完全生成模型，对文本产生的过程进行了模拟，其本质上是一个三层贝叶斯模型，通过对文本的特征项进行映射，得到了在新的空间对文档简短的描述，保留了本质的统计信息，从而达到高效地处理大规模的文档集的目的。这一模型中包含了词、主题和文档三层结构，将每一个文档都表示为几个主题的混合，而每个主题转化为固定词表上的一个多项式分布，这些主题被文本集中的所有文档所共享；在文档层上，每一个文档集都由这些主题进行混合，利用狄利赫雷分布进行抽样，最终得到一个主题的比例用以表示文档。这种每个文档可以表现为多个主题混合的特性，使得在本研究中每个用户能够分在不同的多个类中，更好地多角度地挖掘出不同性别之间的差异性特征。

LDA 模型是建立在狄利赫雷分布的基础上的，是一个完全的生成概率模型，这个模型中利用随机混合的浅层主题组成文本，而将这些浅层主题看作是将所有词汇进行概率分布的结果。LDA 假设语料 D 中每一篇文本有如下的生成过程（Hofmann 1999）：

（1）选择$N\sim Poisson(\xi)$

（2）选择$\theta\sim Dir(\alpha)$

（3）对于每一个词w_n：

①选择一个topic $z_n\sim Multinomial(\theta)$

②从概率分布$p(w_n|z_n\beta)$中选择一个词w_n，p为在topic z_n下的一个多项式概率分布

算法在使用狄利赫雷分布时，一般假设狄利赫雷分布对称，这样就使得（1）式中的α同时满足（2）中的$\alpha_0=\sum_{k=1}^{k}\alpha_k$，且$\alpha_1=\cdots=\alpha_k$。在给定了α，β之后，topic z和文本w的联合概率就为：

$p(\theta,,z,w|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta)$（3）

这里的θ和z都为隐含变量，对其求边缘概率可得一篇文本的概率是：

$p(w|\alpha,\beta)=\int p(\theta|\alpha)\left[\prod_{n=1}^{N}\sum_{z_n}p(z|\theta)p(w|z,\beta)\right]d\theta$（4）

在抽取样本值时我们采用了 Gibbs 抽样的方法，其完整概率模型如下：

$$w_i|z_i,\phi_{z_i}\sim Discrete(\phi^{(z_i)})$$

$$\phi \sim Dirchlet(\beta)\text{（6）}$$

$$z_i|\theta^{(di)} \sim Discrete(\theta^{(di)})$$

$$\theta \sim \text{Dirichlet}\ (\alpha)\text{（7）}$$

在 LDA 模型对文本集主题建模的 Gibbs 算法的过程如下：

（1）z_i初始化为 1 到 T 的某个随机整数，i 取 1…N，N 是文本集中词汇记号的个数，这是马尔可夫链的初始状态；

（2）i 从 1 循环到 N，对主题词汇进行分配，获取马尔可夫链的下一个状态；

（3）不断迭代第二步足够多次，当认为马尔可夫链接近目标分布，记录z_i的当前值，从而通过词汇对于主题的后验概率P(w|z)间接估算出Φ和β的值。

我们在试验中使用了 GibbsLDA[10]作为试验数据的建模平台。这是一个采用 Gibbs 抽样进行参数估计和推理的一个试验平台。在进行聚类后得到的 theta 文档中，反应出了文件中的每一行表示一个文本在主题集上的概率分布，也就是概率$\text{p}(\text{topic}_t|\text{document}_m)$，也就是话题与用户之间的关系，我们将此作为权重，按照下式计算其中男性对于各个话题的因子载荷（factor loading）。

$$\text{FL} = \sum_{i=1}^{i} p_i\text{（8）}$$

其中 FL 表示男性用户对于话题的因子载荷，i 等于男性用户的数量，p_i表示某个男性用户对于某一话题的因子载荷。

结果中的 tassign 中包括了训练集中每个文本中每一个词项的主题分配，每一行代表一个文本，包括了列表<$word_{ij}$>:<topic of $word_{ij}$>。利用词与所在类之间的关系，我们可以得到一个词语编号和它们所属类别编号的矩阵，这一矩阵可以作为学习分类的特征。

2.2 分类器的构建

在得到聚类的类别和词语与类别之间的关系之后，以及男性对于每个话题的因子载荷，我们利用 SVM 的方法来构建分类器。

SVM 是一个基于监督式学习方法一般化的线性分类器，它广泛应用于统计分类以及回归的分析中，能够在最小化经验误差的同时，将几何边缘最大化。在处理线性可分的数据，可以描绘出一条直线直接将各个元组分离开来，但是对于非线性不可分的数据，SVM 采取了一种非线性映射，将原训练数据映射到较高的维度，这就使得高维特征的空间可以采用线性算法来对样本中非线性特征进行线性分析。在新映射形成的高维空间中，SVM 利用支持向量和边缘函数，也就是输入的基本训练元组和支持向量的定义，并基于结构风险的最小化理论寻找到线性分离的最佳超平面，也就是将一类元组与其他相分离的决策边界。利用 SVM 进行的学习可以表示为凸优化问题，因此可以利用已知的有效算法来获取目标函数的全局最优解，而不同于其他分类的算法通常采用基于贪心的学习策略得到局部的最优解。

由于 SVM 模型从本质上来讲是为二值分类问题设计的，因此我们将聚类所得的类进行划归，在高维空间中，SVM 分类问题可以简化为一个线性分类问题。我们令男性=1，女性=-1，假设$\text{f(x)} = \text{wx} + \text{b}$可以划分，其中，w 和 x 均为多维向量。与传统 SVM 分类不同的是，

传统 SVM 分类中若$f(x) > 0$，则$sgn(f(x)) = 1$,否则$sgn(f(x)) = -1$;而对于我们的分类器而言，对任意文本向量x，若其计算的$FL > 0.5$，则$sgn(f(x)) = 1$,否则$sgn(f(x)) = -1$。事实上，在这里 f(x)和 FL 的计算式等价的，都是向量元素的加和，在此基础上按照传统的 SVM 分类方法将该分类器转化为一个优化求解:

$$\min \frac{1}{2}||w||^2 \ s.t., y^i(w^T x_i + b) \geq 1, i = 1, \ldots, n \quad (9)$$

这里由于篇幅限制不再赘述。

这样就将词与主题的关系也划归至男性的主题与女性主题的分类中来，以此来进行分类器的训练，在得到分类器后对试验集中的数据进行试验。

3 试验结果及分析

3.1 数据

在本研究中所使用的数据来自新浪微博。新浪微博是一个由新浪网推出的，提供微博服务的类 Twitter 平台，用户可以借由这一平台通过网页、WAP 页面、手机短信、彩信等发布消息（也就是微博）。作为国内使用人数最多的微博平台，

新浪微博在进行推广的过程中为了提高其知名度，邀请了明星和名人加入，并对他们进行了实名的认证，进行认证之后的用户其用户名会有加上特殊的标记——一个字母“V”。同时随着新浪微博的发展和用户人数的增加，相关认证也已经扩展到了一些具有一定社会地位的普通人群中。因为考虑到在普通的用户中，会存在着用户对于资料中的信息空缺不填以及错误填写的情况，会影响到对于性别的识别和对相关特征的抽取，因此在构建微博的数据库时，我们选取了已经进行了验证的用户作为研究的对象。

本试验的数据集中包含了男女认证用户各 500 人，每名用户抽取 400 条微博（其中用户所有公开微博少于 400 条的，抽取其所有微博），其中男女各 400 人作为训练集，其余部分作为试验集进行分类器的验证。

3.2 实验结果及分析

在对训练集和试验集进行分词后，我们对训练集中的用户利用 GibbsLDA 进行聚类，聚类数为 30，迭代次数为 3000 次，其他参数采用默认参数。在进行计算之后得到的 30 类中有 18 个为男性主题，12 个为女性主题。这些主题的分布如下表所示：

表 3 主题分布

因子载荷	主题数量
0.3-0.2	3
0.4-0.3	4
0.4-0.5	5
0.5-0.6	5
0.6-0.7	6
0.7-0.8	4
0.8-0.9	3

SVM 算法采用使用台湾大学的 libsvm 算法库，并使用默认参数，通过训练得到分类器。

利用此分类器，我们对男女各 100 人进行了分类，所得到的正确率为 83%，得到了较好的效果。我们对分类错误的用户进行了分析。得到主要错误的原因有以下四点：

（1）在错误分类的 34 个用户中有 9 个，经过对其所发微博词语中包含的词语与聚类所得各类中的词语进行比照，发现与这些用户用词所相似的类中，男性用户的因子载荷都在 0.4 与 0.6 之间，也就是男女区分度不高的类中，

（2）另外还有分类错误的 10 个是语言整体风格较为偏异性的用户，例如蔡康永、郭敬明等人的就被错误地判断为女性。

（3）在预处理过程中，由于所使用的分词工具对于网络表达方式的识别能力还比较低，因此对一部分词语的分词错误导致最终的聚类和分类结果都受到了一些影响。

（4）在聚类和分类过程中，对于参数的设置没有进行进一步的调整，可能不能很好地满足本研究的需要。

4 结论和展望

通过本文的研究，我们可以发现在中文微博中，男女用户在语言的特征上存在着比较明显的差异，同时同性别的用户之间的用语上存在着较大程度的重复，本研究所构造的分类器也取得了较好的效果。

下一步为了适应互联网时代的大数据需求，我们会进一步改进方法，使得分类器能够进行不断自学习和自适应，同时提高其准确性，与其它的相关研究进行比较。并会对不同性别的用户的语言特征进行总结和归纳，得到一些具有更好适应性的特征。

参考文献

[1] Schler, J., Koppel, M., et al. Effects of age and gender on blogging[C]. //, Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006:27-29.

[2] Argamon, S., Koppel, M., et al. Mining the blogosphere: Age, gender, and the varieties of selfexpression[J]. First Monday, 2007, 12(9).

[3] Mukherjee, A. & Liu, B. Improving gender classification of blog authors.[C]. //, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010:207–217.

[4] Rao, D., Yarowsky, D., et al. Classifying latent user attributes in Twitter[C]. //, Proceedings of the 2nd international Workshop on Search and Mining User-Generated Contents, 2010:37–44.

[5] Rayson, P., Leech, G., et al. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus[J]. International Journal of Corpus Linguistics, 1997, 2(1):133-152.

[6] Koppel, M., Argamon, S., et al. Automatically categorizing written texts by author gender[J]. Literary and Linguistic Computing, 2002, 17(4):401-412.

[7] Argamon, S., Koppel, M., et al. Gender, genre, and writing style in formal written texts [J]. Text, 2003,

23(3):321-346.

[8] Nowson, S., Oberlander, J., et al. Weblogs, genres and individual differences[C].//Proceedings of the 27th Annual Conference of the Cognitive Science Society, 2005:1666-1671.

[9] Burger, J., D., Henderson, J., et al. Discriminating gender on Twitter[C]. //In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011:1301-1309.

[10] Phan, X. & Nguyen, C. GibbsLDA++, A C/C++ implementation of latent dirichlet allocation (LDA) using Gibbs Sampling for parameter estimation and inference [EB/OL]. 2009 [2012-5-12]. http://gibbslda.sourceforge.net/