北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.039

唐诗题材自动分类研究

胡韧奋[1,†]　诸雨辰[2]

1. 北京师范大学中文信息处理研究所，北京 100875; 2. 北京师范大学文学院，北京 100875;
† 通信作者, E-mail: bnuhurenfen@126.com

摘要　将文本分类技术引入唐诗研究，首先将唐诗按照题材分为爱情婚姻、边塞战争、交游送别、羁旅思乡、山水田园、咏史怀古和其他 7 类，并据此提出唐诗题材自动分类模型。所选 500 首诗歌样本以《唐诗三百首》为基础并有所补充。采用向量空间模型(VSM)将唐诗文本转换为向量，通过卡方检验进行词语特征选择，最后基于朴素贝叶斯和支持向量机算法构造文本分类器，取得较好的题材分类效果。此外，还验证了作者关于题目、体制、作者等变量对题材分类产生影响的假设，为相关诗歌本体研究提供了科学依据。

关键词　唐诗；题材；文本分类；卡方检验；朴素贝叶斯；支持向量机
中图分类号　H087

Automatic Classification of Tang Poetry Themes

HU Renfen[1,†], ZHU Yuchen[2]

1. Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875; 2. School of Chinese Language and Literature, Beijing Normal University, Beijing 100875; † Corresponding author, E-mail: bnuhurenfen@126.com

Abstract　The authors propose a text classification model for Tang poetry. Firstly seven categories are defined for poetry themes: love and marriage, frontier war, friendship and farewell, journey and homesick, landscape and countryside, history and nostalgia, others. 500 Tang poems are selected as research samples, and they are represented in vectors with Vector Space Model (VSM). To reduce the vector dimensions, feature selection is made by Chi-square test. Two classifiers are built based on Naive Bayes and Support Vector Machine algorithms. The models perform well in classification experiment. Besides, the authors verify the positive effect of poetry titles, authors and types to poetry themes by text classification models, which could offer scientific reference to the related research of Tang poetry.

Key words　Tang poetry; themes; text classification; Chi-square test; Naive Bayes; support vector machine

唐诗是中国文学的瑰宝，具有极高的艺术价值。作为国人重要的文化记忆，其检索和分类工作一直是历代学者关心的学术命题。

古代文献学者主要通过编辑总集、类书、选本等方式整理唐诗文献，并在整理中体现其分类意识。当代一些学者将计算机技术引入唐诗研究，如北京大学计算语言学研究所胡俊峰等[1−2]开发了“唐宋诗计算机辅助研究系统”，以唐诗中的词汇为主要研究单位，实现语词检索、词频统计、意象索引等重要功能，为唐诗的检索和利用提供了极大便利。匡海波等[3]从题材角度出发，将文本分类技术引入唐诗研究，取得较好的实验效果，以边塞诗为例，其“最准确率”达到 60%，“次准确率”达到 26.7%。但是，其研究对象限定于 240 首五言绝句，研究结果无法反映唐诗总体风貌。此外，该研究没有将作者、诗体等因素纳入考虑对象，而这些对于准确定位诗歌的题材非常有意义。

在前人工作的基础上，本文开展唐诗题材分类

863 计划(2012AA011104)资助
收稿日期: 2014−07−27; 修回日期: 2014−10−13; 网络出版时间: 2014-12-01 10:20

技术研究, 将诗歌文本、诗歌题目、诗歌作者、诗歌体制均纳入分类特征, 所选样本也包含五言诗、七言诗、古体诗、近体诗等唐诗的基本体制, 以期全面而有效地实现唐诗文本的自动分类。我们选取500 首唐诗为实验对象, 将其按照题材分为爱情婚姻、边塞战争、羁旅思乡、交游送别、山水田园、咏史怀古和其他 7 类, 采用 VSM 模型将唐诗文本转换为向量, 通过卡方检验进行词语特征选择, 最后基于朴素贝叶斯和支持向量机算法进行文本分类。实验表明, 本文的分类模型取得了较好的分类效果, 分类模型中的相关技术对于辅助开展唐诗本体研究具有一定意义。

1 分类体系和数据来源

1.1 唐诗题材分类

很早古人就有诗歌题材分类意识, 但他们对诗歌题材的分类往往是混杂的, 如南朝萧统[4]所编《文选》中的诗歌分类就杂合了题材(述德、咏史、游仙、招隐、行旅、军戎等)、体制(乐府、杂歌等)、功能(劝励、献诗、挽歌、咏怀等)、情感(哀伤等)甚至场合(公宴等)等诸多因素, 直到明代杨廉的《类编唐诗七言绝句》还有将题材(吊古、送别、征戍等)和功能(记行、写怀等)合一的倾向[5], 这种分法是基于诗歌创作的实际情况而论的, 虽然符合古代学术实用性、功能性的基本特点, 但把题材和功能(表达方式)在单一向度上混合起来不甚合理。此外, 古人分类还常有互相重合的问题, 如杨廉将闺情和宫词分为两类, 虽然二者书写场合有别, 但由于其书写对象和情感的相似性, 在现代学术视野下完全可以将其合并为爱情婚姻一类。

本文将文本分类技术引入唐诗研究，我们选择流传较为广泛的《唐诗三百首》作为基本的研究样本, 以其作为唐诗总体风貌的代表。为避免古人分类中的混杂和重合问题, 我们排除了功能的分类(如言志抒怀等), 而尝试将唐诗题材整合为以下 6 类: 爱情婚姻、边塞战争、羁旅思乡、交游送别、山水田园和咏史怀古。此外, 《唐诗三百首》中还有少量描述岁时节气、时事政治或具体状物的诗篇, 但由于其数量与其他题材大类不成比例, 暂将其统一归入“其他”一类。“其他”类的诗歌虽然难以自动归类, 但作为影响题材识别的重要因素, 仍然具有较高的分析价值。

1.2 唐诗样本数据

按照上述题材类别, 对《唐诗三百首》进行初步人工标注后, 各类别分布情况如下: 爱情婚姻 64 首; 边塞战争 24 首; 羁旅思乡 51 首; 交游送别 58 首; 山水田园 55 首; 咏史怀古 34 首以及其他 18 首。

考虑到边塞战争和咏史怀古诗数量较其他类略少, 为使题材分布尽可能均衡, 避免数据偏斜情况①对分类器性能造成影响, 我们又补充了近 200 首诗歌, 使样本整体数量为 500 首。最终各类总计为: 爱情婚姻 98 首; 边塞战争 55 首; 羁旅思乡 73 首; 交游送别 93 首; 山水田园 99 首; 咏史怀古 64 首以及其他 18 首, 这样既提高了一些类别的样本量, 同时兼顾了总体样本的比例分布情况。

2 唐诗文本分类模型

如图 1 所示, 本文分类模型主要分为 4 个阶段: 1) 对唐诗文本进行自动分词、去除标点等预处理操作; 2) 以 TF/IDF 值为权重将文本表示为向量; 3) 通过卡方检验进行特征选择; 4)采用朴素贝叶斯和支持向量机算法进行文本分类。

2.1 预处理

分词是中文文本分类预处理中的重要环节。与现代汉语文本相比, 唐诗中的单字词数量较多, 全切分算法似乎是一种可行的分词策略。但是, 有学者经语料观察指出: 除了双声叠韵词和专有名词外, 各类并列、偏正结构的多字词, 如“宝剑、北雁、悲伤、安排”等, 已经被大量使用, 而“白云”、“秋风”等一般意义上被视为词组, 但由于其在唐宋诗中具有特定的隐喻意也具有了词的性质[1]。因此, 本文采用面向现代汉语的 NLPIR/ICTCLAS2014 分词系统②来处理唐诗文本, 并在分词完成后进行去

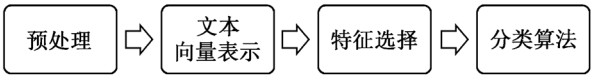

图 1 唐诗文本分类流程

Fig. 1 Flow chart of classification model of Tang poety

① 数据集关于类别的分布往往是偏斜(skewed)或称不均衡的, 在数据偏斜的情况下, 样本无法准确反映整个空间的数据分布, 分类器容易被大类淹没而忽略小类。

② http://ictclas.nlpir.org/

除标点符号等操作。

2.2 文本向量表示

向量空间模型(VSM, Vector Space Model)由Salton 等[6]提出, 20 世纪 70 年代以来, 该模型被广泛地应用于文本分类、检索和相似度计算等任务。在 VSM 模型中, 一篇文本可视为由词语构成的特征项集合, 每一个特征都有一定的权重, 由此可以构成一个多维向量: $\boldsymbol{d}=\{t_1, w_1; t_2, w_2;\ldots; t_n, w_n\}$。其中, $\boldsymbol{d}$ 表示文本, t_i 表示文本中的一个特征(即词语), w_i 为 t_i 对应的权重。

本文采用相对词频信息表示特征权重, 可通过 TF-IDF(Term Frequency-Inverse Document Frequency)公式[7−8]计算:

$$w(t_i,d)=\mathrm{tf}(t_i,d)\times\mathrm{idf}(t_i,d)=\mathrm{tf}(t_i,d)\times\log\left(\frac{N}{n_t}\right), \tag{1}$$

其中, $\mathrm{tf}(t_i, d)$ 为词语 t_i 在唐诗 d 中出现频率, $\mathrm{idf}(t_i, d)$为词语 t_i 的逆文本频率指数, 公式为 $\log(N/n_t)$, N 为唐诗总数, n_t 为包含词语 t_i 的唐诗数量。

2.3 文本特征选择

将文本表示为向量后, 通过特征选择的方法降低向量空间维数, 以避免过拟合问题并提升分类效果。对于唐诗文本分类而言, 特征选择即选出对于分类贡献较大的特征(词语), 比如“将军”、“怀古”、“妾”等, 停用对于分类几乎没有贡献的特征(词语), 如“有”、“无”等。

文本分类中, 常用的特征选择方法有互信息(MI)、信息增益(IG)、文档频率(DF)和卡方检验(CHI)等。Yang 等[9]对这 5 种方法进行分析和比较, 得出卡方检验(CHI)和信息增益(IG)方法效果最佳的结论。本文采用卡方检验进行特征选择, 令 t_i 表示特征(词语), c_i 表示类别, 先计算 4 个观察值: 1) 包含词语 t_i 且属于类别 c_i 的唐诗数量, 设为 A; 2) 包含词语 t_i 但不属于类别 c_i 的唐诗数量, 设为 B; 3) 不包含词语 t_i 但属于类别 c_i 的唐诗数量, 设为 C; 4) 不包含词语 t_i 且不属于类别 c_i 的唐诗数量, 设为 D。

特征 t_i 和类别 c_i 之间的关联度可通过卡方值体现, 计算公式为

$$\chi^2(t_i,c_i)=\frac{N(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)}。 \tag{2}$$

式(2)还可进一步化简, N 为唐诗总数, $A+C$ 为类别 c_i 中所有唐诗数量, $B+D$ 为其他类别中所有唐诗数量, 对同一类别唐诗中的所有词来说, 这几个值都是恒定不变的。由于我们只关心词语和某一类别开方值的大小排序, 故而可省去这 3 项数值, 将公式简写为

$$\chi^2(t_i,c_i)=\frac{(AD-BC)^2}{(A+B)(C+D)}。 \tag{3}$$

通过卡方检验, 我们得到 500 首唐诗中 4441 个词条分别与 6 个类别之间的卡方值, 对每个词条取 6 个卡方值中最大的一项, 进行降序排列, 便得到一份词语对分类的贡献排名。接下来, 可根据卡方值排名从高向低选择一定数量的词条作为特征, 其具体数目需要在实验中不断调整, 以达到最优效果。

值得一提的是, 卡方值反映了各词语与不同主题之间的关系, 检验结果不仅可应用于文本分类任务中的特征提取, 也可辅助开展唐诗本体研究。例如, 爱情婚姻诗中出现的高频意象有帘、罗、粉、珠等, 它们都有细腻、柔软的特点, 将女性的柔婉美展现出来; 又如山、塘、帘等, 它们具有阻隔的空间内涵, 借此把相思之情具象化。可见, 卡方检验提供的词表对于文学研究者细致而准确地进行意象层面的研究也有较大意义。

2.4 其他分类特征

本文在实验中选择诗歌题目、诗歌体制和作者作为重要的分类参考, 三者对诗歌题材均有较为重要的影响。

2.4.1 诗歌题目

诗歌题目对诗歌起着直接解释说明的作用, 如与边塞诗直接相关的题目常见《出塞》、《将军行》等, 与交游诗直接相关的题目常见《赠某某》、《送某某》; 与咏史诗直接相关的题目常见《咏史》、《某某怀古》。此外, 对于大部分乐府诗来说, 乐府旧题本身就有题材倾向性, 如《长干行》之于爱情诗、《战城南》之于边塞诗, 这些诗题对诗歌文本类别的判定无疑具有极高的参考性。

2.4.2 诗歌体制

关于诗歌体制对于题材的影响, 古往今来论诗者常有敏感细致的总结, 例如清代王士禛辨析五言诗和七言诗不同的表现手法, 说: “五言著议论不得, 用才气驰骋不得。七言则须波澜壮阔, 顿挫激昂, 大开大阖耳。”(《答刘大勤问》) [10], 袁枚[11]也提出“凡咏险峻山川, 不宜近体”, 刘熙载认为“五言质, 七言文”(《诗概》) [12]所以田家诗多作五言而

少作七言。因而，我们选择诗歌体制因素作为分类特征之一，希望通过实验来验证其与唐诗题材之间的关联。诗歌体制可分为 7 类：杂古、五古、七古、五律、七律、五绝、七绝。其中古诗不讲平仄的格律、长短不限、用韵不限；律诗八句、中间两联对仗、押平声韵；绝句四句、押平声韵。

2.4.3 作者

作者也会对诗歌的题材分布产生一定影响，因为每一位诗人往往有自己最为擅长的题材，如明代王世贞在《艺苑卮言》中评价李白和杜甫说："太白以气为主，以自然为宗，以俊逸高畅为贵；子美以意为主，以独造为宗，以奇拔沈雄为贵"，"五言律、七言歌行，子美神矣，七言律，圣矣。五七言绝者太白神矣，七言歌行，圣矣，五言次之。"[13]每个诗人都有驾驭不同题材、体制的天赋，使得他们写作某类诗作时尤其得心应手。因而，我们也将作者因素纳入分类特征集。

2.5 分类算法

本文采用朴素贝叶斯(Naive Bayes)和支持向量机(SVM, Support Vector Machine)两种算法构造唐诗文本分类器，并采用 SMO 算法优化 SVM 分类器的训练过程。接下来，分别对这些分类算法的原理进行简单介绍。

2.5.1 朴素贝叶斯算法

朴素贝叶斯算法以贝叶斯定理为基础，是文本分类领域应用最为广泛的算法之一[14]。分类过程中，贝叶斯算法通过估测文本属于每个类别的概率，并选择概率最大项作为分类结果进行输出，计算公式如下:

$$P(c_i|d)=\frac{P(d\,|\,c_i)P(c_i)}{P(d)}, \tag{4}$$

其中，d 表示一首唐诗文本，c_i 表示类别，$P(c_i|d)$为唐诗文本 d 属于类别 c_i 的概率。$P(c_i)$表示 c_i 类唐诗的出现概率，可通过 c_i 类唐诗数量除以唐诗总数得到，$P(d)$是该唐诗的出现概率，由于我们的数据中文本不存在重复，故该值为固定常数。$P(d|\ c_i)$为 c_i 类中唐诗 d 出现的概率，其求解为算法的关键部分。

朴素贝叶斯算法的重要假设是特征之间相互条件独立，对唐诗文本分类而言，即一首诗中各个词之间在概率上彼此独立。那么，令唐诗 $d=w_1, w_2, \ldots, w_i, \ldots, w_{n-1}, w_n$, $P(d|c_i)$即可按照如下公式展开:

$$P(d|c_i)=P(w_1\,|\,c_i)P(w_2\,|\,c_i)\cdots P(w_j\,|\,c_i)\cdots P(w_n\,|\,c_i) \tag{5}$$

其中，$P(w_j|c_i)$代表 c_i 类别唐诗中词语 w_j 的出现概率，可通过多项式模型[15]进行求解，其计算公式如下:

$$\mathrm{P}(w_j|c_i)=\frac{TF(w_j,c_i)+1}{\sum_{|V|}^{k=1}TF(w_k,c_i)+|V|}, \tag{6}$$

$\mathrm{TF}(w_j,c_i)$ 为词语 w_j 在 c_i 类唐诗中的出现频次，$\sum_{k=1}^{|V|}\mathrm{TF}(w_k,c_i)$ 为类 c_i 唐诗中的特征总数，$|V|$为特征词表中的词语总数。

2.5.2 支持向量机算法(SVM 算法)

SVM 算法由 Cortes 等[16]1995 年首先提出，其分类思路为从训练样本中寻找能够确定一个最优超平面的支持向量，该超平面试图将空间中的点最大间隔地分成两类[17]。SVM 是文本分类领域的经典算法，具有很好的分类效果，但是，其训练过程中计算开销较大。为解决这一问题，研究者们相继提出 $\mathrm{SVM}^{\mathrm{Light}}$、SMO 等改进算法来提升训练速度。SMO(Sequential minimal optimization, 序列最小优化算法)由 Platt[18] 1998 年提出，目前在 SVM 训练中得到广泛应用，本文也采用该算法来优化 SVM 分类器的训练过程。

SVM 的分类规则函数可以表示为

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}c_i\alpha_iK(di,\mathrm{d})+b) \tag{7}$$

其中，c_i 为文本类别，α_i 为拉格朗日乘子，d_i 为训练文本，d 为待分类文本。对于 SVM 分类器而言，核函数 K 的选择和惩罚因子 C 的设定十分重要，经反复试验，我们选用唐诗分类效果最好的多项式核函数(Polynomial Kernel)，惩罚因子 C 设定为 1。拉格朗日乘子 α_i 可通过优化下面的目标函数求解:

$$\mathrm{Maximise}L_D=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_ic_i\alpha_jc_jK(d_i,d_j), \tag{8}$$

$$\text{其中}\sum_{i=1}^{1}\alpha_ic_i=0,\alpha_i\geqslant 0\ \text{。}$$

SMO 算法的改进体现在对目标函数的优化上，它将 SVM 的二次规划问题分解成一系列子问题解决，每一步选择两个拉格朗日乘子进行优化，这样避免了复杂的数值求解过程，也降低了计算机存储开销。

3 实验和数据分析

本文以 500 首唐诗为测试样本，基于朴素贝叶斯和 SVM 算法分别构造两种分类器。为保证模型评价的客观和准确，避免因数据分布不均造成的过

拟合现象，我们采用十折交叉验证(10-fold cross validation)的方法，即将数据集分成 10 份，轮流将其中 9 份作为训练数据，一份作为测试数据，进行 10 次试验，分别记录准确率(*P*)、召回率(*R*)和 *F* 值作为评价指标，最终取 10 次测试的平均值作为实验结果。

实验流程为：以唐诗文本为基础，先后加入诗题、体制和作者三项特征，考察分类效果。需要说明的是，我们按照卡方检验值从高到低选择特征，经反复实验调优确定为卡方检验值高于 200 的 3123 个特征。实验结果如表 1 所示。

从表 1 可以看出，基于两种算法构造的分类器中，唐诗题目、体制和作者三项特征的加入使得准确率、召回率和 *F* 值均得到提升，说明三项特征对于唐诗文本分类有不同程度的影响，这也直接验证了作者个性与诗体特征和诗歌题材之间有着重要的关联。

为了更好地对比两种算法的分类效果，我们以综合准确率和召回率的 *F* 值为指标，绘制两种算法的分类效果折线图。如图 2 所示，在不同特征集的训练和测试上，基于 SVM 算法构造的分类器表现均优于朴素贝叶斯算法。

表 1　唐诗文本分类实验结果

Table 1　Experiment result of Tang poetry classification

分类特征	Naive Bayes			SVM		
	P/%	*R*/%	*F*/%	*P*/%	*R*/%	*F*/%
诗文	62.9	63.2	63.0	65.5	65.4	65.4
诗文+题目	65.3	65.8	65.5	71.7	72.2	71.9
诗文+题目+体制	65.6	66.2	65.9	72.3	72.8	72.5
诗文+题目+体制+作者	66.8	67.4	67.1	74.7	75.2	74.9

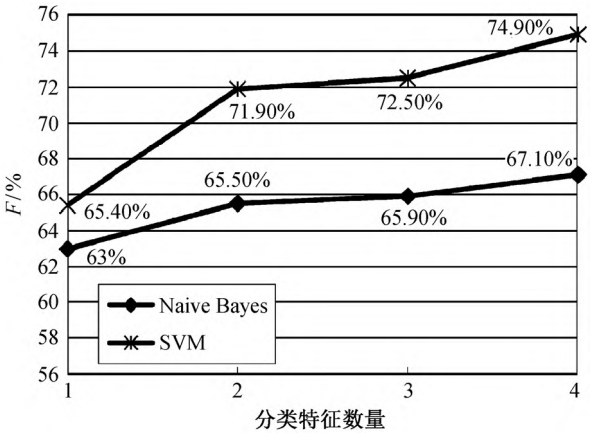

图 2　分类器效果折线图

Fig. 2　Line chart of classification result

文献[3]中，以五言绝句为对象进行测试，最佳实验结果为 66.7%。与其相比，本文研究面向各个类别唐诗，通过卡方检验进行特征选择，并考虑体制和作者对于题材的影响，在题材分类工作上取得较大提升。不过，仍有一批唐诗未能实现准确分类，为考察分类错误原因，探索进一步提升系统性能的可能，我们从 SVM 分类器第 4 次实验结果中提取各类别唐诗的分类 *F* 值，绘制折线如图 3 所示。

由图 3 可知，边塞战争诗、山水田园诗和咏史怀古诗分类效果最好，均超过 80%，爱情婚姻和交游送别诗也取得较高 *F* 值，羁旅思乡诗分类效果略低，而“其他”类别诗则难以进行准确分类。其主要原因在于，分类模型以文本中的词语为主要特征，每个类别唐诗的分类效果均依赖于诗中词语与该类别之间的关联度。例如，边塞战争诗中词语特征非常明显，卡方检验值排名最高 20 个词中有 10 个词与边塞战争类别关联最大，分别为“塞”、“军”、“将

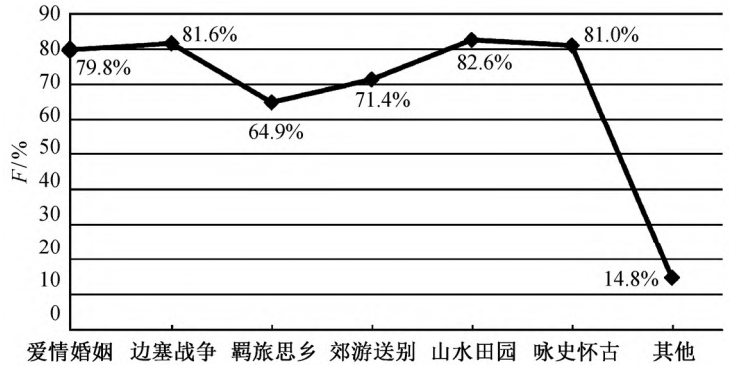

图 3　唐诗各类别分类 *F* 值折线图

Fig. 3　Line chart of classification *F*-measure of Tang poetry

军”、“战”、“胡”、“将”、“笛”、“马”、“出塞”和“弓”, 故而, 边塞战争诗虽然在样本集中不占数量优势, 也能取得很好的分类效果。如果该类别诗中词语特征性不够强, 或者特征性强的词语占比例较低, 则会影响分类效果, 如羁旅思乡诗多以景物描写为主, 且诗中卡方检验值高的词语数量较少, 排名前 20 的词中仅有一个词——“乡”, 故而分类 F 值略低。此外, 以时事政治、状物寄托和岁时节气为主题的唐诗被归入“其他”, 由于诗中词语不具有统一特征, 且样本数量过少(仅占总数 3.6%), 因而无法为分类器准确识别。

在对实验数据进行分析总结的基础上, 我们认为, 本文文本分类模型还有待在以下几个方面进行改进: 1) 样本数量仍然较少, 亟须通过增加样本缓解数据稀疏问题, 提升分类器效能; 2) 本文特征选择和分类器构造完全基于统计方法, 而唐诗是规则性较强的文本, 可在今后的分类工作中进行一定的规则干预; 3) 词语特征由分词得到, 其中, 很多词语表达相同或相近含义, 如“明月”和“月”、“出塞”和“塞”等, 特征选择和权重确定需要考虑同义和近义因素。

4 总结

文本分类模型中的相关技术对于唐诗研究有着重要意义, 它可直接服务于唐诗资源库的建设以及分类检索技术的实现, 提高研究者材料搜集、整理的效率。同时, 也可以为传统的基于经验性的诗歌意象研究、诗歌主题研究提供科学准确的统计数据依据, 间接服务于唐诗本体研究。

本文将文本分类技术引入唐诗题材研究, 采用 VSM 模型将唐诗文本转换为向量, 通过卡方检验进行词语特征选择, 最后基于朴素贝叶斯和 SVM 算法构造文本分类器, 取得了较好的题材分类效果。分类除了调用一般的词语特征外, 还考虑到诗歌体制和作者因素, 并通过实验证实了二者与唐诗题材之间的关联性。然而, 现有系统仍存在一些不足, 比如诗歌题材分类中“其他”项所包含题材的处理问题, 我们计划通过增加样本数量、考虑结合人工规则、调整特征选择及权重等方法来改进并提升分类性能。

此外, 中唐以后发生诗歌转型, 导致诗人关注题材发生了变化, 这突出地反映在“唐宋诗之争”的讨论中。我们对宋诗的初步观察和分析后, 已经发现宋人一方面积极关注时政民生, 展现忧国忧民的一面, 一方面又不断吟咏自然风物, 展现风流自赏的一面, 时事政治与岁时节气题材的诗歌大量涌现, 以宋诗作为对象, 研究这两类题材的分类依据, 显然比唐诗更为有效。进一步也说明时代特征之于诗歌文本分类的意义。这些有待于未来的研究中做更深入的讨论, 以期科学地探索计算机辅助古典诗歌研究的思路和方法, 从而将古诗的规律和韵味更好地呈现出来。

参考文献

[1] 胡俊峰, 俞士汶. 唐宋诗之计算机辅助深层研究. 北京大学学报: 自然科学版, 2001, 37(5): 727–733

[2] 胡俊峰, 俞士汶. 唐宋诗中词汇语义相似度的统计分析及应用. 中文信息学报, 2002, 16(4): 39–44

[3] 匡海波, 陈小荷. 唐诗文本自动分类的算法研究 // 第五届全国青年计算语言学研讨会论文集. 武汉, 2010: 399–405

[4] 萧统, 编. 文选. 李善, 注. 上海: 上海古籍出版社, 1986: 12–22

[5] 孙琴安. 唐诗选本六百种提要. 西安: 陕西人民教育出版社, 1980: 110

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

[7] Jones K S. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 1972, 28(1): 11–21

[8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988, 24(5): 513–523

[9] Yang Y, Pedersen J. A comparative study on feature selection in text categorization // ICML. Nashville, 1997: 412–420

[10] 王士禛, 等. 诗问四种. 周维德, 笺注. 济南: 齐鲁书社, 1985: 78

[11] 袁枚. 随园诗话. 顾学颉, 校点. 北京: 人民文学出版社, 1982: 455

[12] 郭绍虞. 清诗话续编. 上海: 上海古籍出版社, 1999: 2434

[13] 丁福保. 历代诗话续编. 北京: 中华书局, 2006: 1005–1006

[14] 李静梅, 孙丽华, 张巧荣, 等. 一种文本处理中的朴素贝叶斯分类器. 哈尔滨工程大学学报, 2003, 24(1): 71–74

[15] McCallum A, Nigam K. A comparison of event models for naive bayes text classification // AAAI workshop on learning for text categorization. Madison, 1998: 41–48

[16] Cortes C, Vapnik V. Support-vector networks. Machine learning, 1995, 20(3): 273–297

[17] 何建兵, 何清, 史忠植. 基于 SMO 的多层次文本分类法研究. 计算机工程与应用, 2006, 13: 152–154, 167

[18] Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines // Scholkopf B, Burges C, Smola A. Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1998: 185–208