唐诗文本自动分类的算法研究[1]

匡海波　陈小荷

（南京师范大学文学院，江苏南京　210097）

E-mail: mchypocn@hotmail.com

摘　要：本文旨在研究如何基于题材和感情，试对唐诗文本进行自动分类，即对现行的通用文本分类算法，做基于唐诗文本的比较实验和参数微调。本文的目的在于初步试探唐诗自动分类算法，发现算法的“关节”和“焦点”，为以后大规模唐诗文本的自动分类研究打下基础。

关键词：唐诗文本　自动分类　分类算法　参数微调

Research on Algorithms of Tang- Poetry-Text Automatic Classification

Kuang Haibo　Chen Xiaohe

(Department of Literature, Nanjing Normal University, Nanjing, 210097)

E-mail: mchypocn@hotmail.com

Abstract: This paper aims to study how to achieve the Tang- Poetry-Text Automatic Classification based on subjects and emotions. That is to make comparisons and parameter tunings of the existing common text classification algorithms, which are used to focus on Tang-Poetry-text. The purpose of this paper is also to preliminarily test the automatic classification algorithms of Tang-Poetry-Text and find the keys of the algorithms. The result is the basis for large-scale study of Tang- Poetry-Text Automatic Classification in the future..

Keywords : Tang- Poetry-Text, Automatic classification, classification algorithms, parameter tuning

1　引言

“弱水三千，我只取一瓢饮”——如何在海量文本中有序、实用地掌握我们所关注的信息，始终是中文信息处理的一个核心技术。基于这个考虑，文本自动分类技术一直是信息处理的热点课题，这项技术对于信息过滤、信息检索、信息挖掘都有极高价值。

另一方面，唐诗作为我国优秀的文学遗产之一，内涵丰富、样式繁多、流传广泛，影响深远。基于唐诗的计算机辅助研究，将为我们分析唐诗、理解唐诗提供一定的支持。因此，针对唐诗的自动分类技术既有理论价值，又不乏实用领域。

根据“程序=数据结构+算法”公式，面向唐诗的文本自动分类程序可以大致分解成数据结构部分和算法部分。唐诗文本数据结构可以描述为“唐诗文本模型”，即基于唐诗文本的特征性存储办法，且此办法来源于对唐诗的预处理结果，可期为针对分类算法的特定性存储结构。唐诗文本分类算法则是自动分类技术的核心。

因此，本文将着重于分类算法研究，在唐诗文本自动分类实验的基础上，对算法特质及结果做一些比较分析，并应用分类算法于一定数量的真实唐诗文本。

---

[1] [作者简介]匡海波，南京师范大学文学院本科生；陈小荷，南京师范大学文学院教授，博导，主要研究方向为计算语言学。

2 文本自动分类技术（Text Automatic Classification ）

文本自动分类系统接收待处理文本后，将根据分类体系和分类算法自动确定文本关联的类别。其本身是一个映射过程（赵敏涯，2009），它将系统接收的新文本映射到某一类别集中。本文不处理兼类现象，实际并非如此，因为一首唐诗可能兼有两种或两种以上类别的内涵与性质。

实际上，映射总是基于两个方面，一是基于数据结构，即文本模型部分，二是基于算法，即分类算法部分。因此我们还可以这样描述文本分类技术，如图2所示。

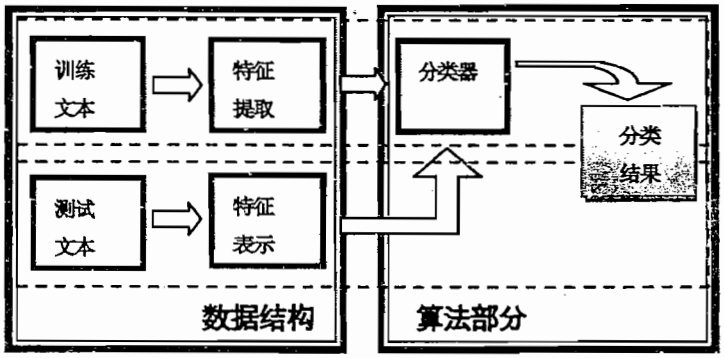

图 1

另一方面，自动聚类和自动归类是文本自动分类应用比较广泛的两种模式（余坎，2008）。其中，自动聚类技术是无指导的分类办法，即通过比较文本自身的特征，来确定相似类别。此算法属于动态规划，虽无明确的分类标准，但重在观察对象的相似度。自动归类技术更类似于有指导的分类办法，实际上是心理学“原型说”的应用。合适地找到原型，依赖统计；恰当地描述原型，依赖规则。该技术的核心是唐诗知识库和分类推理机的构建。本系统将采取比较快捷的自动归类策略，并以此建立分类系统。

3 文本预处理阶段

一个完美的唐诗自动分类系统，可以将真实唐诗文本进行分类。但现在我们仍需要为唐诗文本做一些准备工作，即预处理。预处理过程中一系列问题的解决，将有助于我们提高分类的可行性和准确率。

3.1 唐诗文本的分词办法

唐诗文本如何分词，是预处理的重要策略之一。观察基于现代汉语分词办法的分词结果，唐诗文本与其说是词构成，不如说是字构成，全切分似也是一种策略选择。但有学者指出，词作为特征项要优于字和词组。同时，全切分策略很有可能将部分关键词（包括一些“未登录词”）肢解，这些“未登录词”恰恰是我们最容易把握唐诗的“关键词”。例如匈奴、长城、都护这些词语都显示了唐诗的边塞风格；而江南、鹊桥则充满婉约气息。因此，我们将采取现代汉语分词办法，只对结果进行一定的干预。

唐诗是一种“类似规则”的语言。韵律和规则的相对关系对于分词是有一定作用的，一个加进韵律规则、诗歌特征的分词系统必然会提高分词准确率。系统着眼于算法问题，因此只是基于韵律对部分结果进行干预，不对上述策略进行深究。

3.2 唐诗文本的预处理策略

文本预处理的后期工作主要是指一些调整策略，例如文本的降噪与增益过程。所谓文本降噪与增益，

是将一些严重干扰分类，区别特征淡化的词语从文本表征词典中除去，同时将一些特别明显的表证词进行一定的增益，以提高文本分类的效率。

如何降噪与增益，有待于实验微调。本系统的降噪增益过程，将有依赖于统计和规则的交互结合。在预处理阶段，系统将尽可能多地接受文本表示，这也和系统所依赖的语料库偏小，而相关知识不充裕有关。

4 唐诗本体的分类理论

人工确定分类标准是文本归类技术的首要问题。这将为系统设计一个个类别接口，使得符合各自标准的唐诗文本依毂而入。分类标准在一定程度上决定了系统的对象和效能。

唐·顾陶《唐诗类选》，宋·赵孟奎《分门纂类唐歌诗》，明·周叙《唐诗类编》等，都是古人对于唐诗分类研究的“开山之作”。当代学者张浩逊则从深层角度考虑了唐诗分类的特点和依据，更多地从作品义的角度分析了唐诗分类。总体看来，唐诗分类大致可以有这么几种角度：

（一） 按体裁分类

按照体裁分类，大体有绝句、律诗、排律三类，且均有五言、七言之分。不可否认，如此分类对于唐诗的整理和研究是有效的，但缺乏语义角度的考查。本文以此出发，着眼于“五言绝句”的分类研究。

（二） 按作者分类

此分类方案，易于搜索，便于查询，是编纂唐诗的传统方法之一。但对于唐诗的理解和深层研究，则只能做到引子的作用。另外，前人对此研究颇多，本系统就不做涉及。

（三） 按题材（主题）分类

明·敖英编纂的《类编唐诗七言绝句》采取了如此办法，此书专收七绝，分为吊古、送别、寄赠、怀思等类别。近代出版的唐诗分类词典如若按题材分，也大都采取这种杂糅的态势。即在题材中涵括主题，感情等作品义特征。

出于加工容量和加工精度的考虑，并融合一定的感情因素和题材内容，本系统着眼于“五言绝句”，将分类类别设为六大类，即边塞军旅、羁旅乡思、恋情爱情、山水田园、咏史怀古、咏物抒怀等。需要指出的是，唐诗本体知识表明，体裁和作者在一定程度上对自动分类有辅助作用。本分类系统将不考虑这些因素，将来如有可能，可以依此调节分类程序。

5 唐诗文本的数据结构表示

5.1 文本特征表示

文本自动分类算法依赖文本所表示的类别特征，即文本的数据结构。因此需要用准确而恰当的方法将文本表示成分类函数可以接受的文本特征形式。

最经典的文本特征形式表示是上世纪60年代Salton等人提出的向量空间模型（VSM，Vector Space Model）。该模型将文本组织成一组词条向量，且向量以权重为分量。即$D=\{t_1,w_1;t_2,w_2\ldots\ldots t_i,w_i\ldots\ldots t_n,w_n|1\le i\le n\}$，其中$t_i$表示文本的特征项序列，$w_i$为第i项特征的权重。特征项权重的表示，系统采取相对词频的统计办法，该统计量可以运用TF-IDF公式计算，公式如下所示：

$$w(t,d)=\frac{tf(t,d)\times\log(\frac{N}{n_t}+L)}{\sqrt{\sum_{t\in d}\left[tf(t,d)\times\log(\frac{N}{n_t}+L)\right]^2}}$$

其中，w（t,d）为词t 在文本d 中的权重，tf（t,d）为词t 在文本d 中的词频，N 为训练文本的总数，n_t为训练文本集中出现t 的文本数。

5.2 文本结构组织

系统约定，训练语料收入分类特征较明显的唐诗文本，以组织文本模型，训练分类器。系统以“小型化”的原则收入《全唐诗》中合计240 首960 句“五言绝句”唐诗文本，并且尽量做到作者、年代的相对平衡。

训练完毕的抽象向量模型可以如表1 所示（表为“边塞军旅”类别向量模型）。不难看出，系统采取的向量模型既有典型的代表词，也有一些应列入停用词表的非典型词。如何对文本向量进行修饰完善，可以在实验中进行调试。

同理可以构建其余五类别的向量模型。对比观察已完成的各类别文本结构组织模型，不难发现特征词有交叉成分、或相关度较低成分，这些都可能影响具体的分类效果。设想“边塞军旅”和“羁旅乡思”区分度应该相对较低；而中古汉语的一些基础词重复出现也不偶然，其余特征本文就不一一赘述。

word	w(t, d)
不	0.111229
马	0.161909
塞	0.136622
里	0.077368
出	0.089438
边	0.103772
征	0.130764
行	0.083324
胡	0.104911
人	0.052970
入	0.078707

表 1

6 文本分类算法

分类算法是系统的核心部分。以下具体介绍系统采取的几种主流算法。

6.1 简单向量求矩法

顾名思义，简单向量求矩法就是通过对比新文本与模型中心文本的向量距离，判定文本距离最低，即文本内容最相似的类别。模型中心文本的获取可以采取算术平均的方式。其公式具体表述为：

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

其中，d_i为文本的特征向量，d_j为第j 类的中心向量，M 为特征向量的维数，W_k为向量的第K 维。

分类系统需要关注各大类的封闭测试结果。系统抽样选取了15 组样本结果，其中重点描述如下三种类别的测试结果（表3）：

表 3

	边塞军旅	恋情爱情	咏物抒怀
最接近次数	9	2	5
次接近次数	4	3	4
第三接近次数	2	6	4

观察表3得知，系统对“边塞军旅”类的最准确判断率达60%，而次准确判断率达26.7%，两项结果之和为86.7%。[2]系统约定“接近原则”，即如果分类文本的接近百分比排在第一或第二位就算判断基本正确。因此系统测评的结果可以统计如下表。

	边塞军旅	羁旅乡思	恋情爱情	山水田园	咏史怀古	咏物抒怀
准确率	86.7%	86.7%	33.3%	86.7%	73.3%	60%

6.2 贝叶斯算法（Bayes）

贝叶斯公式作为概率论的经典公式，在分类算法上也有其应用模型，即贝叶斯分类器。基于贝叶斯公式的算法，其基本思路是计算文本属于各类别的概率，将文本分到概率最大的类别中。算法主要步骤如下：

1）重新计算文本数据结构模型中的w值，即特征词的类别概率向量（$w_1,w_2,w_3,\ldots\ldots w_n$），计算公式如下所示：

$$w_k = P(W_k|C_j) = \rho * \frac{1 + \sum_{i=1}^{|D|} N(W_k, d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N(W_s, d_i)}$$

其中，$P(W_k|C_j)$为词K在C_j中出现的比重，D为该类的训练文本数。N（W_k，d_i）为词K在C_i中的词频，|V|为总词数，$\sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N(W_s, d_i)$为该类所有词的词频和。ρ参数的作用是调整公式的结果更利于统计。

2）按下面的公式计算新文本di属于类Ci的概率：

$$P(C_j|d_i;\hat{\theta}) = \frac{P(C_j|\hat{\theta})\prod_{k=1}^{n} P(W_k|C_j;\hat{\theta})^{N(W_k,d_i)}}{\sum_{r=1}^{|C|} P(C_r|\hat{\theta})\prod_{k=1}^{n} P(W_k|C_r;\hat{\theta})^{N(W_k,d_i)}}$$

其中，$P(C_j|\hat{\theta})$为第J类文档训练文档数/总训练文档数，|C|为类别和，$N(W_k,d_i)$为W_k在di中的词频。

3）比较新文本属于各类的概率，将文本分到概率最大的那个类别中。

贝叶斯算法的统计结果参照简单向量求距法，区分较为明显。通过对训练语料的计算测试，其结果仍可以统计如下表，其中参与结果计算的文本仍采取抽样原则取出。

	边塞军旅	羁旅乡思	恋情爱情	山水田园	咏史怀古	咏物抒怀
准确率	100.0%	100.0%	93.3%	100.0%	100.0%	100.0%

7 分类算法的参数微调

算法微调将起催化剂的作用。从上述结果来看，分类系统需要一定的微调，尤其是数据模型的初期设立和算法变量的值选择。

7.1 停用词表的设计

有学者曾经指出，文本自动分类的困难之一是特征空间的高维性和文档表示向量的稀疏性(牛延莉，张化，2008)。对于小规模的唐诗文本分类系统来说，这种情况同样存在。本系统首先采取一种基于规则

[2]最准确率=最接近数/所有判断数，次准确率=次接近数/所有判断数

的降噪办法，即停用一部分公有的非典型词。实际操作时，初步只抽取了5个非典型词。

通过统计手段，可以得出如下5个停用词（表5）。实际上，如下五个停用词并非都为虚词（停用词表往往倾向于选择一些对于表达意义关联不大的虚词，助词等），本文不深究其“规则性”。

停用词	不	人	一	月	里

表 4

通过对简单向量求距法的实验，结果表明，“咏物抒怀”类的分类准确率从60%提高到了66.6%，“恋情爱情”类从33.3%调整到了40%。贝叶斯算法实验结果也亦有所增加。可以预见，如果有效地增加一些合适的停用词，系统分类效果还会有所提高。

7.2 参数的信息增益

针对技术性的概率公式，同样需要进行一定的参数微调，以提高其效率。系统着眼于两部分进行测试，初步发现，基于小规模唐诗语料而言，单纯的概率参数微调，效果并不明显。

1. 相对词频再增益

观察系统发现，唐诗文本模型的概率矩阵稀疏度不亚于现代汉语，并且高频词数甚至远远低于现代汉语。基于解决相对词频数据稀疏的考虑，系统采取的办法是参数平滑（parameter smoothing）。目前，系统的参数平滑实验，并没有取得特别突出的效果。或许，唐诗文本的后期数据修饰，更依赖于规则。

2. 贝叶斯算法ρ 值

贝叶斯算法中的类别概率向量公式应用到一个额外的技术参数ρ，ρ 值的意义是使结果更利于统计。从系统观察得出，ρ 值的设定取决于语料规模，本系统合适的ρ 值大约为1000-3000，其中以2000为佳，其比较如下所示：

ρ =2000	ρ =1000
Δ [3]=87774772.5978	Δ =589.046
Δ =121802189.0389	Δ =817.400
Δ = 4648311945.7900	Δ =2000.443
Δ =5013.9401	Δ =0.013
Δ =290.6796	Δ =0.012
Δ =2476.8019	Δ =0.025

不难看出，如果Δ 值太小，将不利于系统的概率运算。因此需要调节ρ 值来得出合适的Δ 值。在系统硬件的制约和影响下，兼顾语料规模和概率稀疏，多次实验调节ρ 值是比较可信的做法。

8 算法比较分析

系统的分类结果出现了一定程度的类别杂糅。实际上，系统的分类难度比现代汉语分类更加复杂（例如有学者针对新闻分类的类别为政治、经济、文化、新闻、教育等，比之于本系统的分类办法，特征更为明显），因此一定程度的杂糅也是情理之中。

相比较于现代汉语文本，唐诗文本数据稀疏和维度偏低的特点都是显见的，这给分类带来了不小的难度。以杜甫的《蜀相》为例，文本维度为 45，而向量模型的概率权重皆为 1。这样的文本要进行分类，难度自然是可以预期的。

上文的封闭测试结果表明，贝叶斯的效果要优于简单向量求距法，这在现代汉语中也是比较普遍的（赵敏涯，2009）。贝叶斯算法封闭测试效果极佳，与训练语料规模不大有关联，同时贝叶斯的过度拟

[3]Δ 为中间运算值，Δ 值越大，则概率越大。

合问题也使然。如何使语料扩大后保持准确率，有待进一步的研究。

比较突出的是，“边塞军旅”的效果较好，而“爱情恋情”的效果较不佳。仔细分析前一类的文本模型，可以看出文本向量维数并不见得很高，但频率普遍不低；而后一类则特征相对缺乏，难以提取合适的模型特征。如何更加有效地提取特征质量，是算法发展的一个关键。

至于KNN算法，支持向量机(SVM)算法，神经网络算法等，本系统未能一一测试。就目前系统的语料规模而言，数据稀疏问题比较明显。以KNN算法为例，K值的取量一直是算法优劣的关键，目前的K值一般初始值定为几百到几千，这样的K值对于小规模唐诗文本分类测试而言，是需要进行调整的。

9 小型开放测试与总结

通过上述封闭测试和参数微调，现对一小部分唐诗文本进行开放测试。系统随机选取了《全唐诗》中另外24首“五言绝句”唐诗文本，其结果可以如下所示：

算法名称	简单向量求矩法	贝叶斯算法
最准确率[4]	50%	66.7%
次准确率	25%	16.7%

结果表明，特征越明显的唐诗文本，分类效果越好。而某些寓意深刻的唐诗，理解起来就比较棘手。因此，如何提高系统的效率，重点在于挖掘分类特征，尤其是某些类别的不明显特征。

为了更好的观察分类算法，系统进行了小规模的“反向测试”，即对分类结果中概率最小的类别，作否定判断，其中准确率普遍高于80%。这对于分类系统的功能完善有一定的参考价值，可以帮助我们排除一些远端的类别特征。

总体来看，唐诗文本自动分类研究仍应采取经验主义的原则，因此分类有赖于概率算法的选择。这样的分类过程不仅有助于我们研究唐诗本体的特征，也有助于我们继续中文信息处理的方法改善。但就唐诗而言，我们还不应忽略唐诗的“理性成分”，规则干预在某种程度上将成为唐诗计算机辅助研究的重要环节。相比较现代汉语，唐诗的“规则性”要丰富和广泛得多。

我们坚信，唐诗文本分类研究的“东风”将为唐诗的深层研究打下基础，唐诗文本的计算机辅助研究终将会“直挂云帆济沧海”。

参 考 文 献

[1]Charles Day. The Computation of Poetry. Computing in Science & Engineering.

[2]余坎. 中文文本自动分类技术的研究. 理工高教研究. 2008年8月第4期.

[3]赵金仿, 赵艳, 缪建明. 网页信息抽取及其自动文本分类的实现. 计算机技术与发展. 2008 年10 月第10期.

[4]蒲筱哥. 自动文本分类方法研究述评. 情报科学. 2008 年第26卷第3期.

[5]董乐红, 耿国华, 周明全. 一个中文文本自动分类器的设计. 计算机应用与软件. Vol125 No. 4.

[6]马建斌, 李谨, 滕桂法, 王芳, 赵洋. KNN和SVM算法在中文文本自动分类技术上的比较研究. 河北农业大学学报. 2008年第3期.

[7]易勇, 郑艳, 何中市, 李良炎. 基于机器学习的古典诗词作者的判别研究. 心智与计算, Vol. 1, 2007.

[8]徐平，徐建中. 基于量子自组织网络的Web 文本自动分类方法. 情报科学. 2009 年1 月第1期.

[9]赵敏涯. 文本自动分类算法的比较与研究. 电脑知识与技术. 2009.

[10]牛延莉，张化. 文本自动分类研究进展. 软件导刊. 第7卷第4期. 2008年.

[11]张浩逊. 谈谈唐诗的分类研究. 吴中学刊. 1997年第4期.

[12]吴企明. 《唐诗分类研究》前言. 无锡教育学院学报. 1999年12月第13卷第4期.

[4] 最准确率，次准确率算法类比上文。