# 汉语文本按语体分类的研究*

王慧玲[@]　宋柔[#]　戴伟长[@]

@北京工业大学 北京 100022

E-mail:huiling_wang@263.net

#北京语言文化大学 北京 100069

E-mail:songrou@blcu.edu.cn

**摘要**：有导文本分类是指在给定的分类体系下，通过对训练语料的学习对每个类建立特征向量以表示该类，然后对每一个新的文本自动确定其类别的过程。本文介绍的文本分类的目的不是按专业领域对文本分类，而是按语体对文本分类。 这种分类研究对于统计语言模型的训练具有重要意义。本文以汉字的字频向量作为语体类别的表示方法，通过计算类别向量与文本向量的相似度来确定文本的类别。已应用该方法对几种不同语体的小说分类，取得了较好的分类结果。

**关键词**：语体；文本分类；向量空间模型

# The Research of Text Categorization on Style

Wang Huiling

Beijing Polytechnic University. Beijing 100022

E-mail:huiling_wang@263.net

Song Rou

Beijing Language and Culture University. Beijing 100069

E-mail:songrou@blcu.edu.cn

Dai Weichang

Beijing Polytechnic University. Beijing 100022

**ABSTRACT**: The text categorization with supervised learning is a procedure to retrieve the eigenvector for each category from the training corpus, and then assign a category to a text. This paper introduces a method that classifies texts on styles instead of domains. This study is significant to statistical language model. The method uses Chinese character frequency vector to represent categories and texts, and computes the comparability between categories and texts to predict the category. The method has been implemented on five corpus of classic novels, Jin Yong's novels, Gu Long's novels, translated novels, Chinese contemporary novels, and satisfactory

*本文得到国家自然科学基金（69882001）、国家 863 计划（863-306-ZD03-04-02）的资助

categorization effectiveness has been achieved.

**Keywords:** style ; text categorization ; Vector Space Model(VSM)

# 1. 背景

文本分类是指在给定的分类体系下，根据文本的内容自动确定文本类别的过程。该技术随着 Internet 的飞速发展，显得越来越重要。

目前流行的是从信息检索的需要出发，按领域来分类。在过去的几年中，Internet，尤其是 WWW 得到了飞速的发展. WWW 包含了多种类别和形式的信息，以其内容丰富吸引了大量用户。然而，由于 WWW 是一个开放性的全球分布式网络，并且网上的资源没有统一的组织和管理，造成了用户信息搜索的困难。因此，一些信息服务网站提供信息检索服务功能。但是，目前的检索系统普遍是基于关键字匹配的检索，检索结果包含了很多与用户需求无关的信息。若能在检索前先对信息分类，并要求用户除了提供检索关键字外还说明检索分类范围，则能使检索到的信息更好地满足用户的需求。然而，面对 Internet 这个信息海洋，完全采用人工分类组织的方法只是杯水车薪. 这就需要利用计算机对文本按领域进行自动分类标记，为文本信息的检索提供系统化的解决方案。

还有一种文本分类需求是按语体分类。本文中，语体指的是语言的风格。最大的语体类别是文言文和白话文。白话文中，政论、新闻、科技论文、法律文件、小说、散文、诗歌等也是不同的语体。此外，白话文中因夹杂文言成分的不同，语体也不同。例如金庸的武侠小说同翻译小说就很不一样。

对文本按语体分类，应用目标之一是统计语言模型训练语料的整理。语言信息智能输入输出系统的工作原理，核心是 n 阶马尔可夫链的统计语言模型。统计语言模型的最大问题是统计数据稀疏。为了缓解这一问题，通常的做法是加大训练语料的规模，从几百万字、几千万字，到几亿字、十几亿字。这种方法确实有一定效果，但也有弊病，主要是训练语料质地不均。不同专业领域的文本，所用术语很不相同；不同语体的文本，词语频率和词语之间邻接关系很不相同。过去研究工作者注重于专业领域不同所带来的影响，其实语体的影响非常大。以拼音汉字转换为例，如果训练语料的主体是报刊新闻，应用对象是小说时效果就比较差。为解决这个问题，有人建议用“均衡语料”，即把不同语体的文本按一定比例混合，用作训练语料。这样做的结果，训练语料是一个大杂烩，每一次的具体应用对象只是某一种特定语体的，训练语料的性质同应用对象还是不吻合，遇到该特定语体中的小概率词语邻接现象，还是会出现错误。比较合理的解决办法是对不同的语体采用不同的统计语言模型，不同的模型用不同语体的训练语料生成。应用时，根据特定应用对象的语体性质，自动挑选性质比较接近的统计语言模型。要达到这样的目的，必须先有能力对大规模语料库中的文本按语体分类。

文本按语体分类采用的基本方法仍然是在按领域分类方面采用的基本方法—向量空间模型方法，但文本按语体分类与按领域分类是有区别的， 对文本按领域分类实质上是在内容

上的分类，也就是在词汇、句子乃至整个篇章所传达的主要思想上的分类，而按语体分类实质上是在形式上的分类，也就是在文章所采用的体裁及其语言特色上的分类。这也就决定了文本按语体分类与按领域分类在解决方法上有一定的区别。

按领域分类，特别关注的是得以区分不同领域的词语。因此，特别高频的词，即通用于各种领域中的词，不在考虑之列，称作“停用词”。但这类词，在不同的语体中表现为不同的频率。如古典白话小说中“道”（意为“说”）的频率超过“的”的频率，大陆小说则反过来；又如政论文、古典白话小说、现代武侠小说中代词用得较少，翻译小说中代词却比比皆是；口语性白话文本中用“时候”、“这个”、“很”、“因为”、“所以”、“的”等，而文言性强的文本对应地用“时”、“该”、“甚”、“因”、“故”、“之”等。

按领域分类通常不关心词语之间的排列关系，但这类性质对于区别语体则是重要的。如翻译小说中的对话，“某某说”之类的文字常常出现在直接引语之后，或者两段直接引语中间，“说”后用逗号；章回小说很少有这种行文方式，一般是在直接引语前有“某某道”，“道”与引号之间是冒号。

按领域分类需要关心词的意义。比如“计算机”就是“电脑”，“国库券”是“证券”的一种。按语体分类则不关心这些问题。

## 2. 分类方法

长期以来，文本分类一直是自然语言处理的一个重要的应用领域。90 年代以来，基于机器学习的文本分类方法已成为文本分类的主流技术。基于机器学习的文本分类方法通常由训练和测试两个阶段组成，在训练阶段，从训练文本中学习分类知识，建立分类器；在测试阶段，则根据分类器将被测试文本分到最可能的类别中 。

本文讨论了单层和两层两种分类方法，它们都是采用向量空间模型的基本思想。

向量空间模型（VSM）是 Salton 等人于 60 年代末提出来的，它是目前多数已有的分类系统的基础。在向量空间模型中，文本和类别要被表示成特征向量，特征向量中的每个数字与文本中一特征项（可以是字或词等）出现的频率相联系，通过计算特征向量之间的距离，来判断文本之间的相似度。在文本分类中，使用该方法首先学习训练语料以选取用以分类的特征集，然后对每一个测试文本求出其相应于上述特征集的特征向量，依次计算该特征向量与各个类的特征向量的距离，选取距离最小的类别作为该测试文本所属的类别。

### 2.1 语体类别和文本的表示

一般我们把原始数据组成的空间叫测量空间，把分类识别赖以进行的空间叫做特征空间，通过变换，可把在维数较高的测量空间中表示的模式变为在维数较低的特征空间中表示的模式。在特征空间中的一个模式通常也叫做一个样本，它往往可以表示为一个向量，即特征空间中的一个点。

目前很多的文本分类系统均采用词作为表示文本的特征项，除词外，字也可以在某种程

度上表示文本，特别是在按语体对文本进行分类的方面。语体中最大的两类是文言文和白话文，它们之间存在着不少的差异，其中之一是词汇的差异。由于社会的变迁，文言文中的一些词，随着旧事物的消亡已变得没有用了，而白话文中的许多词是文言文中所没有的。而且，文言文以单音节词为主，而白话文中双音节词占绝对优势。例如，白话文中的“现在”的意思，在文言文中由“今”来表达，文言文中用“处”，而不用“地方”等等，这样就使得同一汉字在不同语体中有不同的字频，因此本文采用字频向量作为文本和语体类别的特征向量。

国标 GB-2312 中的一二级字库共有汉字 6763 个，如果将这些汉字都作为文本和类别的特征，将因计算量极大而导致无法计算，而且这些汉字中的相当一部分在语料库中出现的频率很低，即使不考虑它们，对计算结果的影响也不大。因此，我们首先要解决的问题是特征的选择，要对所有出现的汉字进行筛选，得到最能反映分类本质的特征汉字，这就是特征选择的过程。假设要识别的文本和类别有 t 种特征观察量，将其抽象为 t 维空间中的向量，形式如下：$X=(x_0,x_1,\cdots,x_{t-1})$，其中每一维为一个特征，表示某个汉字在此文本或类别中出现的频率。

## 2.2 相似性的度量

为了将样本进行分类，就需要研究样本和类别之间的关系，与哪一类别最相似就将样本归为哪一类。我们采用目前应用较广的用在空间中定义的某种距离来度量相似性的方法，与哪一类别的距离最近就将样本归为哪一类。距离有多种定义方法，本文采用下面两种方法：

假设有两个向量：$X(x_1,x_2,\cdots,x_t)$，$Y(y_1,y_2,\cdots,y_t)$

(1) 绝对值距离　　$d_1(X,Y)=\sum_{k=1}^{t}|x_k-y_k|$

$d_1(X,Y)$的值愈小，表明 $X$和$Y$ 之间的距离愈小，值愈大，表明 $X$和$Y$ 之间的距离愈大。

(2) 夹角余弦　　$$d_2(X,Y)=\frac{\sum_{k=1}^{t}x_k y_k}{\sqrt{\left(\sum_{k=1}^{t}x_k^{\,2}\right)\left(\sum_{k=1}^{t}y_k^{\,2}\right)}}$$

$d_2(X,Y)$的值愈大，表明 $X$和$Y$ 之间的距离愈小，值愈小，表明 $X$和$Y$ 之间的距离愈大。

## 2.3 特征选取

特征选取的过程分 2 步：

(1) 特征汉字的粗选取：从每个类别的样本中挑选出出现频率在一个给定阈值以上的汉

字，这些汉字的并集构成粗选取的结果；

(2) 特征汉字的细选取：从粗选取的结果汉字集中，挑选出在不同样本中出现频率的差别超过某个阈值的汉字，以排除因平均分布在各类语体中而不具有分类意义的汉字，这一步就是要通过对训练语料的学习，确定频率差别的阈值，进而选定特征汉字。

特征选取的算法如下：设有 k 类语体，$T_0, T_1, \cdots, T_{k-1}$ 为它们各自的训练语料，$T = \bigcup_{0 \le i < k} T_i$

(1) 对 GB-2312 汉字集中的每个汉字，统计它们在各类语体中出现的频率 $freq(x, T_i)$（$x \in GB$，$GB$ 指 GB-2312 汉字集，下同）及每个汉字的平均字频 $freq(x, T)$。

(2) 设定字频阈值 $\delta$，进行粗选取： $S_1$ 为结果汉字集

$$S_1 = \bigcup_{0 \le i < k} \{x \in GB \mid freq(x, T_i) > \delta\}$$ 根据实践经验将 $\delta$ 设为千分之一

(3) 进行细选取：由 $\sigma$ 决定的结果汉字集为 $S_2^\sigma$，用下式表示汉字 $x$ 在不同语体中出现频率的差别： $dif(x) = Max[\underset{0 \le i < k}{Max}[freq(x, T_i)] / freq(x, T), freq(x, T) / \underset{0 \le i < k}{Min}[freq(x, T_i)]]$

则 $S_2^\sigma = \{x \in S_1 \mid dif(x) > \sigma\}$

$\sigma$ 是汉字出现频率差别阈值，其值通过学习获得，该过程如下：

设得到的 $S_2^\sigma$ 中的汉字个数为 $n^\sigma$，则得到的特征汉字序列为 $C^\sigma = (C_0^\sigma, \cdots, C_{n^\sigma - 1}^\sigma)$，各类语体的特征向量为：$V_i^\sigma = (freq(C_0^\sigma, T_i), freq(C_1^\sigma, T_i), \cdots, freq(C_{n^\sigma - 1}^\sigma, T_i))$ $(0 \le i < k)$

对每个类别的语料 $T_i (0 \le i < k)$ 选取 m 个训练样本，为 $t_{i,0}, t_{i,1}, \cdots, t_{i,m-1}$，对每个 $t_{i,j}$ $(0 \le i < k, 0 \le j < m)$ 求得其特征向量：$V_{ij}^\sigma = (freq(C_0^\sigma, t_{ij}), \cdots, freq(C_{n^\sigma - 1}^\sigma, t_{ij}))$

为确定 $\sigma$ 的值，我们建立一个 k 行 m 列矩阵 $M_{km}^\sigma$，其元素为

$$w_{\substack{i,j \\ 0 \le i < k \\ 0 \le j < m}}^\sigma = \begin{cases} 1 & i = \arg\min\limits_{0 < q < k} [d(V_{ij}^\sigma, V_q^\sigma)] \\ 0 & i \ne \arg\min\limits_{0 < q < k} [d(V_{ij}^\sigma, V_q^\sigma)] \end{cases}$$

即，如果 $t_{i,j}^\sigma$ 与 $T_i$ 距离最近，则 $w_{i,j}^\sigma$ 为 1，否则为 0。

进一步，我们令 $M^{\sigma}=\sum_{i=0}^{k-1}\sum_{j=0}^{m-1}w_{ij}^{\sigma}$，它表示：对于选定的某个$\sigma$，$k\times m$个训练样本中有多少个被正确分类。

设$\sigma$的取值范围为$[\sigma_0,\ \sigma_1]$，将该区间分成r等份，每一等份长度为$\varepsilon=\frac{\sigma_1-\sigma_0}{r}$，统计$\sigma=\sigma_0,\ \sigma_0+\varepsilon,\cdots,\ \sigma_0+l\varepsilon\ \ (l\in[0,r])$时各自的$M^{\sigma}$值，设其中$M^{\sigma}$最高时的$l$为$l_e$，即 $M^{\sigma_0+l_e\varepsilon}=\underset{1\le l<r}{Max}M^{\sigma_0+l\varepsilon}$，则细选取的阈值为$\sigma_e=\sigma_0+l_e\varepsilon$。

## 2.4 测试

对未知语体的文本$t_x$，求其特征向量

$$V_x^{\sigma_e}=(freq(c_0^{\sigma_e},t_x),freq(c_1^{\sigma_e},t_x),\cdots,freq(c_{n^{\sigma_e}-1}^{\sigma_e},t_x))$$

若$d(V_x^{\sigma_e},V_i^{\sigma_e})=\underset{0\le j<k}{Min}d(V_x^{\sigma_e},V_j^{\sigma_e})\ \ (0\le i<k)$则认为$t_x$属于第$i$类语体。

## 2.5 两层分类方法基本思想

两层分类方法是基于模式识别中的多级分类器方法和上述单层分类方法实现的。利用两层分类方法首先将一个多类别分类问题转化为若干个（这里设为2个）简单的分类问题来解决。它不是试图用一种算法，一个决策规则去把多个类别一次分开，而是采用分级的形式，使分类问题逐步得到解决。

## 2.6 两层分类方法实现过程

设有k类语体，$T_0,T_1,\cdots,T_{k-1}$为它们各自的训练语料，$T=\bigcup_{0\le i<k}T_i$

（1） 对GB-2312一二级字库中的6763个汉字，分别统计它们在各类语体中的出现频率$freq(x_i,T_i)\ (x\in GB)$及每个汉字的平均字频$freq(x,T)$；

（2） 求各类之间的相关系数矩阵以对初始语体类别进行分类：

设特征向量：$X(x_1,x_2,\cdots,x_n)\ \ Y(y_1,y_2,\cdots,y_n)$　均值：$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$　$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$

协方差：$\mathrm{cov}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

标准差：$S_x=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$　$S_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$　相关系数：$r(X,Y)=\frac{\mathrm{cov}(X,Y)}{S_xS_y}$

$r_{ij}$ 为 $T_i$ 与 $T_j$ 两类间的相关系数，$r_{ij}$ 愈大，表明 $T_i$ 与 $T_j$ 愈相似。若 $r_{ml}=\underset{\substack{0\le i<k\\0\le j<k}}{Min}r_{ij}$（$0\le m,l<k$），则第一子类集合 $P_1$ ：$\{T_m\}$，第二子类集合 $P_2$ ：$\{T_l\}$。对其余语体类别 $i(i\neq m,l)$，若 $r_{im}>r_{il}$ ，则 $P_1\leftarrow P_1+\{T_i\}$，否则 $P_2\leftarrow P_2+\{T_i\}$。

（3） 对 GB-2312 一二级字库中的 6763 个汉字，分别统计它们在 2 个子类中的平均字频，利用单层分类方法的原理，通过对训练语料的学习，获得用于一二两个子类分类的特征汉字集 1；

（4） 对第一子类，分别统计其中每个语体类别的字频及该子类汉字的平均字频，利用单层分类方法的原理，通过对第一类训练语料的学习，获得用于第一子类分类的特征汉字集 2；

（5） 对第二子类，分别统计其中每个语体类别的字频及该子类汉字的平均字频，利用单层分类方法的原理，通过对第二类训练语料的学习，获得用于第二子类分类的特征汉字集 3；

（6） 测试：对每一被测试文本，统计其字频，利用特征汉字集 1 判断它所属的子类，再由该子类分类时的特征汉字集确定其最终所属的语体类别。

## 3．文本分类试验与结果分析

### 3.1 文本分类试验

3.1.1 所使用的训练语料是从黄金书屋网站上不同语体类别的网页上下载的文章。作为初次试验，我们选择了古典白话小说（如《红楼梦》）、古龙武侠小说、金庸武侠小说、外国翻译小说、现代大陆小说等 5 种语体类别。它们的训练语料字数如下：

| | | | |
|---|---|---|---|
| 古典小说类训练集字数： | 9311229 | 古龙小说类训练集字数： | 11164637 |
| 金庸小说类训练集字数： | 6424234 | 翻译小说类训练集字数： | 11451174 |
| 大陆小说类训练集字数： | 14876821 | 训练集总字数： | 53228095 |

3.1.2 单层分类方法得到的特征汉字集：

绝对值距离方法：曰 它 她 队 剑 德 氏 斯 杨 玉 军 韦 克 尔 国 宝 等62个汉字；

夹角余弦距离方法：曰 它 她 队 剑 德 氏 斯 杨 玉 军 韦 克 尔 国 等216个汉字；

3.1.3 单层分类方法测试结果：

封闭测试结果：

| | | |
|---|---|---|
| 文本总数：6442 | 绝对值方法判对数：5691 | 正确率：88.3% |
| 文本总数：6442 | 夹角余弦方法判对数：5526 | 正确率：85.8% |

开放测试结果：

| | | |
|---|---|---|
| 文本总数：2029 | 绝对值方法判对数：1805 | 正确率：88.9% |
| 文本总数：2029 | 夹角余弦方法判对数：1597 | 正确率：78.7% |

3.1.4 两层分类方法初步分类结果：

第一子类：古典小说、金庸小说　　第二子类：翻译小说、古龙小说、大陆小说

3.1.5 两层分类方法中初步分类时使用的特征汉字集：

绝对值距离方法：像 众 很 宝 甚 便 官 兵 日 什 样 弟 与 之 王 见 等83个汉字；

夹角余弦距离方法:像 众 很 宝 甚 便 官 兵 日 什 样 们 现 她 爷 地等20个汉字；

3.1.6 两层分类方法中子类分类时使用的特征汉字集：

第一子类分类时使用的特征汉字集：

绝对值距离方法:曰 韦 她 姐 剑 掌 功 且 钱 氏 瞧 力 脸 玉 今 刀 等208个汉字；

夹角余弦距离方法：曰 韦 她 姐 剑 掌 功 且 钱 氏 瞧 力 脸 玉 今 等51个汉字；

第二子类分类时使用的特征汉字集：

绝对值距离方法：队 斯 德 剑 克 尔 军 玉 国 武 掌 部 忽 特 吗 利 等89个汉字；

夹角余弦距离方法:队 斯 德 剑 克 尔 军 玉 国 武 掌 部 忽 特 吗 利等89个汉字；

3.1.7 两层分类方法测试结果：

封闭测试结果：

| | | |
|---|---|---|
| 文本总数：6442 | 绝对值方法判对数：5628 | 正确率：87.4% |
| 文本总数：6442 | 夹角余弦方法判对数：5727 | 正确率：88.9% |

开放测试结果：

| | | |
|---|---|---|
| 文本总数：2029 | 绝对值方法判对数：1781 | 正确率：87.8% |
| 文本总数：2029 | 夹角余弦方法判对数：1787 | 正确率 ：88% |

## 3.2 结果分析

由上述结果可知，

(1) 单层的夹角余弦方法较差，其他三种搭配方法基本正确率相同；

(2) 按语体分类的确可以获得较高的区分度（开放测试正确率达到88%-89%）；

(3) 封闭测试和开放测试的结果差别并不大，说明分类器已经得到了比较充分的训练，所抽取的特征汉字具有普遍性和有效性。

分类错误的原因分析如下：

(1)名字引起的误判：在某篇文章中若某个名字多次出现，且这个名字中含有特征汉字，

较易引起错误。例如：大陆小说训练集中的《十面埋伏》，由于文中“罗维民”、“魏德华”是主要人物，两个名字多次出现，而“罗”、“德”是典型的外国人名用字，是“翻译小说”和“大陆小说”进行区别的比较重要的特征，从而导致多篇文章被判为“翻译小说”类；又例如，“宝”字，因在《红楼梦》中出现频率较高而在其他类别的语体中出现频率较低，所以被选为古典小说特征汉字，但它在其他古典小说文本中并不常出现。类似这样的汉字还有古龙小说中的“楚”，“留”，“香”等。这样的特征字选取导致错误分类。

(2)由于训练语料中文章的选取不是很严格而导致的误判：例如，大陆小说训练集中的《哈尔滨人》一文，由于是随笔，较少人物、事件，而使得大陆小说方面的特征削弱，导致误判；

(3)由于古典小说训练语料中的一些文章因时代等差别引起的语言差异而导致的误判：古典小说类的主要特征是“甚”多“很”少，“日”多“天”少，“无”多“没”少，“何”多“什么”少，起代词作用“之”多等。而《老残游记续》中不说“甚”，而说“很”，多用“没有”，而少用“无”等；《官场现行记》中既用“众人”，又用“大家”，多用“日”，而少用“天”，多用“什么”，而少用“何”，起代词作用的“之”少。因此，《老残游记续》和《官场现行记》的一些章节被误判。

## 4．今后的工作

目前只是采用汉字字频作为文本和类别的特征，有可能忽略了其他对文本分类来说有意义的信息，比如汉字之间出现频率的相对差别、汉字的接续关系、用词的不同、词间关系的不同等等。在今后的工作中，将同时考虑其他的具有分类意义的特征，以各个特征的权重作为系数定义一个新的判别函数，有望进一步提高正确率。

## 参考文献

[1] 陈小荷等译，语言研究中的统计方法，北京语言文化大学出版社，2000

[2] 张尧庭等，多元统计分析引论，科学出版社，1997

[3] 边肇祺等，模式识别，清华大学出版社，2000

[4] 黄萱菁等，独立于语种的文本分类方法，中文信息学报，2000，14（6）：1-7

[5] 吴赣等，WWW 页面的文档分类技术，计算语言学文集，清华大学出版社，1999

[6] 胡裕树等，现代汉语，上海教育出版社，1995