分类___TP391___ 密级_____

西北師範大學

硕士学位论文

中文文本分类特征选择方法研究

陈建华

导师姓名职称：王治和 教授

专 业 名 称：计算机应用技术

研 究 方 向：数据库技术及应用(数据挖掘)

论文答辩日期：2012 年 5 月 学位授予日期：2012 年 6 月

答辩委员会主席：

评 阅 人：

二〇一二年五月

硕士学位论文

M.D Thesis

中文文本分类特征选择方法研究

Research of Feature Selection Method for Chinese Text Classification

陈 建 华

Chen Jian-hua

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包括其他人已经发表或撰写过的研究成果，也不包含为获得西北师范大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：陈建华 日期：2012.5.20

关于论文使用授权的说明

本人完全了解西北师范大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：陈建华 导师签名：王治和 日期：2012.5.20

摘　　要

随着科技的发展和网络的普及，人们可获得的数据量越来越多，这些数据多数是以文本形式存在的。而这些文本数据大多是比较繁杂的，这就导致了数据量大但信息却比较匮乏的状况。文本挖掘技术为解决这一问题提供了一个有效的途径。而文本分类技术是文本挖掘技术的一个重要分支，是有效处理和组织错综复杂的文本数据的关键技术，能够有效的帮助人们组织和分流信息。文本分类的两个重要的研究方向是:特征选择与文本分类算法。

特征选择是指从高维的文本特征空间中选择出最能代表文本内容的特征，好的特征选择方法一方面能够降低文本特征空间的维数，以利于提高文本分类的效率，另一方面好的特征选择方法通过去除对文本分类无效的特征也有利于提高文本分类的分类精度。而好的文本分类方法则能够直接有效地提高文本分类的效果。

目前在文本分类领域较常用到的特征选择算法中,仅仅考虑了特征与类别之间的关联性,而对特征与特征之间的关联性没有予以足够的重视。针对这种情况，本文提出一种基于类别区分度和关联性分析的综合特征选择算法。首先利用类别区分度提取出具有较强类别区分能力的特征词来降低特征空间的稀疏性，再通过特征的关联性分析衡量特征与类别的相关性以及特征之间的冗余度，最终选择出具有类别代表性且相互之间不存在冗余的特征词。经实验验证，该算法能有效地改善分类器的性能。

关键字：文本分类；特征选择；类别区分度；C-关联；F-关联；相关独立度

Abstract

With the development of technology and network's penetration , more and more data is available to people and most of these data is in the form of text. These unstructured form of data leads to a status with large volume of data but with relatively rare information. Text mining technology has provided an effective way to solve this problem. Text classification techno1ogy is a branch of text mining technology, which means it is one key technology of managementing and organizing complex text data effectively. Text mining can help people organize and stream information effectively. Two important research directions of text classification are: feature selection method and text classification algorithm.

Feature selection refers to select the feature terms which can best represent the characteristics of text from high-dimensional feature term space. Good feature selection method on one hand can reduce the dimension of the text feature space, resulting in the improvement of text classification efficiently, on the other hand good feature selection method can improve the accuracy of text classification through removing invalid feature terms. Good text classification method is able to improve text classification result directly.

Current feature selection algorithms frequently used in text categorization merely take the correlation between feature and class into account but pay less attention to correlation between the features. In view of this situation, this paper proposes a syntaxic feature selection algorithm, which based on category discriminating power and correlation analysis. The algorithm firstly uses discrimination power to extract the features that reveal larger differences among categories to reduce the sparsity of feature spaces, and then employs correlation analysis of features to measure relativity between features and categories and redundancy among features, so can acquire the feature subset which are more representative and have no redundancy each other. Experiments demonstrate that the proposed algorithm can improve the performance of the classifier effectively.

Keywords: text categorization; feature selection; category discriminating power; C-correlation; F-correlation; relevant independency

目　　录

第一章　绪论

1.1 研究背景和意义

随着信息技术的发展和 Internet 的迅速普及，网络信息资源呈现出了海量的特点。这些信息资源在给人们带来丰富知识和极大便利的同时，也暴露出了一些亟待解决的问题。其中，最主要的问题表现在这样的信息资源的增长速度远远超出了人们能够处理它们的能力。信息的极大丰富并没有提高人们吸收知识的能力，面对如此浩瀚的信息，人们更难快速得收集到自己所需要信息并从中获取知识。正如奈斯比特在《大趋势》一书中准确形容了人们目前所处的困境，即“信息是丰富的，而知识是贫乏的”。

文本挖掘技术是信息处理领域中的一个重要研究方向，主要用于基于文本信息的知识发现，即从非结构化的文本中发现潜在的有用信息。文本分类技术[1,2,3]则是文本挖掘的一个重要分支，其主要任务是根据未分类文本的信息判别其所属的类别，自动把文本分到预先设定的某个类别中，以帮助人们快速有效的找到所需要的信息。

由于文本数据具有非结构化的特点，所以在对文本数据进行文本分类之前需要对文本数据进行预处理，把非结构化的文本数据转变为结构化的形式，一般以特征向量空间模型[4,5,6]表示，以方便文本分类算法的处理。而用向量空间形式表示文本数据时，向量空间维数一般高达几万维甚至几十万维。即使经过进一步的处理，如去除停用词等仍然会有大量的高维的特征向量保留下来。理论上高维的特征向量应该有利于文本分类效率和精度的提高，而实际上并非如此，很多情况下，高维的特征向量在增加文本分类算法学习时间的情况下并没有提高文本分类的效率和文本分类的精度，反而在降低文本分类效率的情况下产生与之小的多的特征子集一样的分类精度。这样的情况也就导致了对文本特征进行特征选择成为文本分类的必要前提和基础。

特征选择[7,8,9]通过删除对文本分类没有多大贡献的特征词条，从而选择出对文本或类别具有较好代表性的特征词条。这样，一方面能够降低文本向量空间的维数，另一方面也能够提高特征词条对文本或类别的代表性，提高文本分类的准确性。除此之外，文本向量空间维数的降低也有利于分类算法的学习，使得更多的分类算法能够应用到文本分类技术中，为选择更好的文本分类算法提供条件。

随着网络上文本数据信息的爆炸式增长，特征选择方法越来越受到人们的关注，表现出了巨大的应用价值，因此对特征选择方法进行研究具有重要意义。

1.2 研究历史与现状

特征选择问题自 20 世纪 70 年代以来已得到了非常系统的发展。Siedlecki和Sklansky[10]讨论了如何评价一个特征选择算法，按照研究历史将特征选择算法分为过去的、现在的和将来的三类，相关文献发表于 1988 年，当时许多新的算法还未出现。Doak[11]，Jain等[12]研究了搜索的起点、方向和策略等问题，还研究了评价特征子集，即评价准则的问题。1997 年新加坡国立大学的M.Dash和H.Liu[13]对此以前的特征选择方法进行了总结，依据评判准则和特征选择策略将特征选择分为 15 种，其中特征选择搜索策略分为：启发式搜索策略、完全搜索策略和随机搜索策略；特征选择评判准则分为：距离度量、信息度量、相关度量、一致性度量和分类错误率度量。文中还提出一个特征选择框架，指出一个特征选择算法是由“特征子集生成”、“特征子集评价”、“停止条件”和“结果验证”四个部分组成的，其中“特征子集生成”决定了主要的搜索策略，它和“特征子集评价”中评价准则的确立是特征选择算法中起决定作用的两个问题。

根据样本中是否含有类别信息，特征选择可分为有监督的特征选择[13,14]和无监督的特征选择[15,16,17,18,19]。无监督的特征选择是指在数据集中，通过数据集中特征之间的关系进行特征选择的方式；有监督的特征选择是指在给定类别的前提下，利用特征之间和特征与类别之间的关系对特征集进行选择的过程。

特征选择按照和后续分类算法的结合方式可分为嵌入式、过滤式和封装式。

(1) 嵌入式特征选择

在嵌入式结构中，特征选择算法作为组成部分嵌入到分类算法里。如某些逻辑公式分类算法是通过向公式表达式中加减特征实现的[20]。类似的加减特征操作也构成一些复杂的逻辑概念推导的核心，只是通过不同特征组合形成更复杂的规则描述。最典型的即决策树算法，如Quinlan的ID3[21]和C4.5[22]以及Breiman的CART算法[23]等，算法在每一结点选择分类能力最强的特征，然后基于选中的特征进行子空间分割，继续此过程，直到满足终止条件，可见决策树生成的过程也就是特征选择的过程。

(2) 过滤式特征选择

过滤式特征选择的评估标准直接由数据集求得，独立于分类算法，如图 1.1 所示。最简单的过滤式特征选择于 20 世纪 60 年代早期提出[24]，该算法在特征间相互独立的假设下，研究各特征对于分类的可分性，按照某种搜索策略，选出符合要求的特征子集。

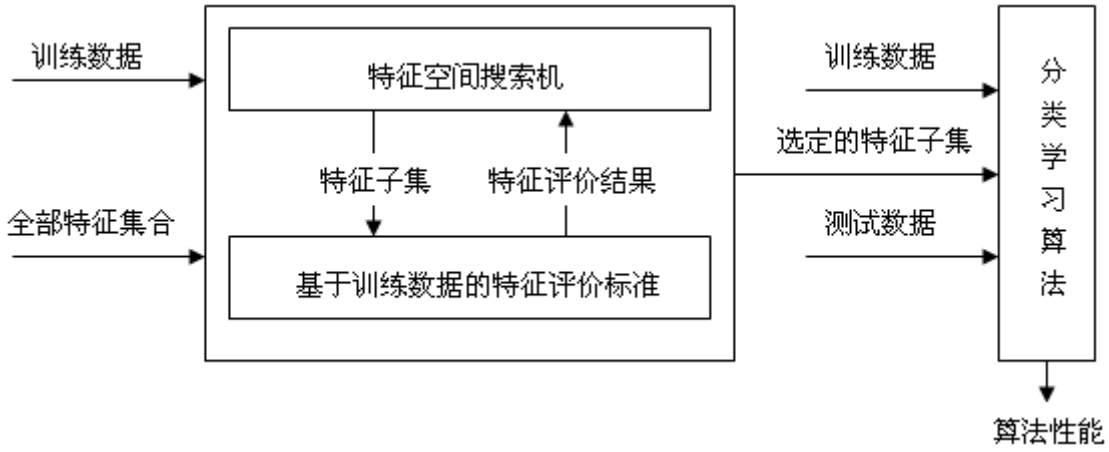

图 1.1 过滤式特征选择流程图

(3) 封装式特征选择算法

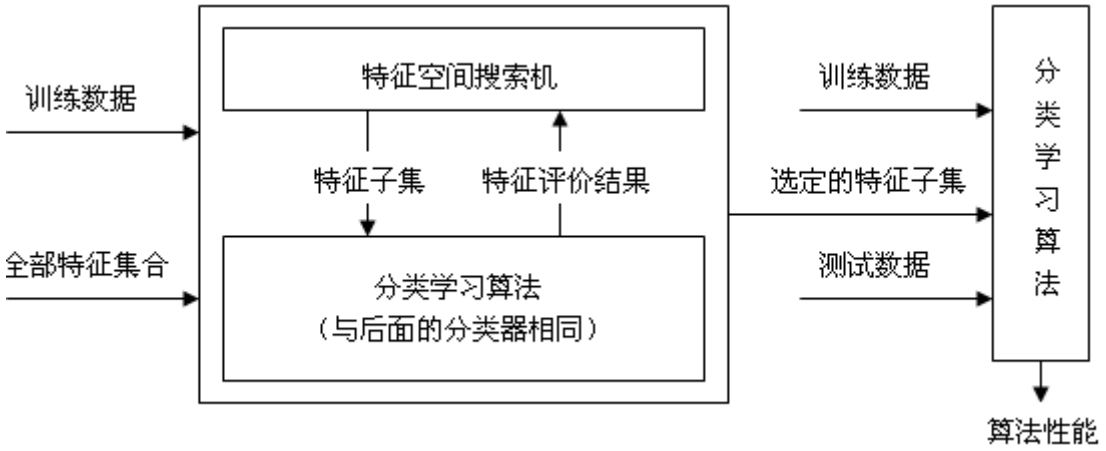

图 1.2 封装式特征选择流程图

封装式特征选择算法最早由John等在 1994 年提出[25]，如图 1.2 所示。与分类算法无关的过滤式特征评价会和后续的分类算法产生较大的偏差，而所选特征子集的分类识别封装式特征选择算法中用以评估特征的分类算法是没有限制的。

John等选用决策树[25]，Aha等将最近邻法IB1和特征选择算法相结合对云图进行分类研究[26]，Provan,Inza等则利用贝叶斯网络性能指导贪心地前向搜索算法[27,28]。

由于采用分类算法的性能作为特征评估标准，封装式特征选择算法比过滤式特征选择算法准确率高，但算法效率较低。因此，一些研究者努力寻找使评价过程加速的方法。Caruana等提出一种加速决策树的方法[29]，即在特征选择过程中大量减少决策树分支的数目。Moore等通过减少评估特征阶段的分类器的训练样本来提高特征选择的速度[30]。封装式方法的另一个缺点是过适应问题，但该问题主要发生在训练数据规模较小的情况[31]。

1.3 研究内容及组织结构

1.3.1 本文研究内容

在考虑特征冗余的特征选择方法中普遍认为，即使某个特征与类别具有很强的相关性，但如果它与已选的特征相关时，该特征的重要性降低，或者直接不会被选入特征子集，因为相比于已有的特征子集，它根本无法提供额外的与类别有关的信息。本文也采用这种思路对冗余因素进行处理。

本文采用“相关分析+冗余分析”特征选择的研究思路，根据特征的文档、类内、类间等分布信息考察特征的类别相关和分类贡献，根据分布信息的相似性进行特征相关冗余因素的度量和处理，在保证计算效率的同时优化特征选择，获得更好的特征子集用于文本分类。本文在文本特征选择方面所做的工作主要有：

（1）优先选择具有较高分类贡献的特征。

为了改善分类器的性能，较好的方法是通过增强分类器对类别的判别能力。本文首先提出特征的类别区分度，用来表示和度量特征词的类别区分能力，以选择出对类别具有较强区分能力的特征参与文本分类过程。

（2）对特征进行关联性分析，提出相关独立度来消除冗余。

特征词在文档内、类内和类间的分布信息常常会很相似，这些分布信息相似的特征之间有很强的相关性，因为分布相似的特征，它们对其所体现的某些类别具有很好的分类能力，但对其它类别的分类能力较弱，或者存在分类贡献的“雷同”现象。从分类系统的总体分类能力上考虑，最好选择具有各种不同的类别区分能力、不同分布信息特点的特征建立特征子集用于文本分类。因此本文提出相关独立度来表示在已知某一类别的情况下两个特征词之间的相互独立程度，用以判别并消除冗余的特征，建立最佳的特征子集。

（3）在判别冗余特征时避免候选特征之间的两两比较，以降低特征选择算法的时间复杂度。

如果直接对 N个特征逐对分析，算法的时间复杂度则为$O(N^2)$。在本文算法中，若判断出某一特征与已选特征之间有很小的相关独立度，即它和已选特征之间有很高的相关冗余度，也就是说明已选特征能够完全代表它所带有的类别信息，所以它不再参与和后续其它特征的比较而直接删除，这时算法的时间复杂度为$\mathrm{O}(N\log N)$。这样就避免了所有特征的逐对比较，降低了算法的时间复杂度。

1.3.2 论文组织结构

本文的组织如下：

第一章 绪论。介绍本研究课题的研究背景及研究意义、特征选择方法的研究历史和现状以及本文主要研究内容及论文的组织结构。

第二章 文本分类技术。对文本分类的相关技术进行详细的介绍。主要包括文本分类的流程、文本分类的预处理技术、文本的表示模型、常用的文本分类方法以及文本分类的实验评估方法。

第三章 文本特征选择方法。概述特征选择方法的意义，介绍常用文本特征选择方法的特点及评价。

第四章 基于类别区分度和关联性分析的综合特征选择。提出一种基于类别区分度和关联性分析的综合特征选择算法，通过类别区分度、特征的关联性分析等衡量特征词的类别代表性以及特征之间的冗余度。

第五章 实验设计与分析。介绍本文实验方法并对实验结果进行分析。

第六章 总结与展望。主要对论文所做得工作进行简单总结，同时提出了进一步研究的方向和需要开展的工作。

第二章 文本分类技术

2.1 文本分类定义

文本分类指预先定义文本的主题类别，然后按照待分类文本的内容将待分类文本划分到一个或若干个预定义的主题类别的过程。文本分类技术通过限定搜索范围来提高人们搜索信息的效率和准确度。简单来说，文本分类的任务为:在给定的分类体系下，根据每类样本的数据信息，建立相应的类别判定公式和类别判定规则，总结出分类规律。这样，当需要为待分类文本确定其类别时只需要根据已经总结的类别判定公式和类别判定规则就能够把待分类文本划分到相应的类别中去。

从数学的角度来看，文本分类实际上就是将未标注类别的文本映射到预定义类别的过程。文本分类的形式化定义为[1]：对于待分类的文本用文本集合D表示，$D=\{d_1, d_2, \cdots, d_n\}$，其中$d_n$表示待分类文本，$n$表示待分类文本的个数。分类系统预定义的类别用C表示，$C=\{C_1, C_2, \cdots, C_m\}$，其中$C_m$表示类别，$m$表示类别的种类。客观上，在待分类文本集合与预定义类别之间存在一个目标概念G，G表示为:

$$G: D \to C \tag{2.1}$$

G 把一个待分类的文本映射到一个预定义的类别中。对 D 中的文档 d 来说，G(d)是已知的。这样通过对文本分类训练文本集的有指导的学习就可以得到一个近似于 G 的分类模型 M，M 表示为:

$$M: D \to C \tag{2.2}$$

有了分类模型M之后，对于待分类文本d_n，其分类结果就可以用$M(d_n)$来表示。文本分类系统的搭建或者文本分类的学习目的就是要找到一个和G最相似的分类模型M。用公式表示为:

$$Min(\sum_{i=1}^{|D|} f(G(d_i) - M(d_i))) \tag{2.3}$$

其中|D|表示待分类文本集的大小，f为判断 G 和 M 是否相似的评估函数。

需要说明的是，待分类文本到预定义主题类别的映射可以是一对一的映射，即待分类文本只属于一个类别;待分类文本到预定义主题类别的映射也可以是一对二的映射，例如对垃圾邮件的判定就是一个一对二的分类映射;待分类文本到预定义主题类别的映射还可以是一对多的映射，即多类映射，通常情况下一般将多类映射问题转化为一对二映射问题进行研究。

2.2 文本分类任务的特点

文本分类就是将大量文本划分为一个或一组类别，使得各个类别代表不同的概念主题。这实际上是一个模式分类任务，所以很多模式分类的算法可以应用文本分类中。但是，文本分类是和文档的语义紧密相关，所以与普通的模式分类任务相比有许多独特之处[32]。

(1)高维特征空间

在文档特征提取的时候，有大量的候选特征。如果使用词语作为文档特征，即使一个 10000 篇左右的训练文档，一般也会产生上万的候选特征。如果使用这些特征来构造文档向量，那么向量空间的维数非常高。

(2)特征语义相关

考虑一种避免“高维灾难”的解决办法是，假设特征之间是相互独立的，即一个特征出现与否与其他的特征并无关系。但是，一般地，文本分类中很多特征包含一些相互依赖的关系，例如：“中共”、“中央”两个词共同出现的概率较大，存在相互依赖关系。

(3)特征分布稀疏

用特征词来表示文档的时候，往往特征维数非常高，而文档所出现的特征词只占总特征词的小部分。特别是对于一篇比较短的文档来说，特征空间中，仅仅出现少量的特征词，因此，多数特征词的出现频率都为零，导致了文档向量中大多数的特征的值都是 0，特征的分布非常稀疏。

(4)特征存在多义和同义现象

文本分类中一般使用词、短语等作为表征文档语义的文档特征。但是，这些特征往往无法清晰地表达一种含义。一个特征可能有多种含义，即多义现象，如：“教授”这个特征既可以表示一种职称的含义，也可以表示一种传授知识的含义。同时，许多相同的含义可以用不同的特征来描述，即同义现象，例如：“计算机”和“电脑”这两个特征都表示相同的含义。

(5)基本线性可分

文本分类中，大部分类别之间是基本线性可分的。所以一些复杂的、在其他模式分类任务中应用很成功的方法，在文本分类中未必会取得很好的效果。

2.3 文本分类流程

对于中文文本分类，其处理过程主要包括训练阶段和测试阶段。

在训练阶段，首先对中文训练集中的每一个文档进行分词，也就是将每一个文档分割成该文档所包含的单词。在经过分词后所形成的特征词集合中，有很多词对于文本分类的作用不大甚至还会影响分类效果，例如常用的一些虚词等，对于这些词我们就需要去除，这就是去停用词阶段。虽然去除了停用词，但对文档进行向量表示时，特征空间的维数很大，大大影响了分类的时间效率和空间效率。因此就需要在向量表示前对特征空间进行降维处理，从原有的特征空间中，挑选出对文本分类最重要或者影响最大的一些特征词。在降维处理后，为了区分不同特征词对分类的不同作用，还需要对特征词进行权重调整，然后进行文档的向量表示。最后用训练集中的文档向量进行分类器的构造。具体流程如图 2.1 所示。

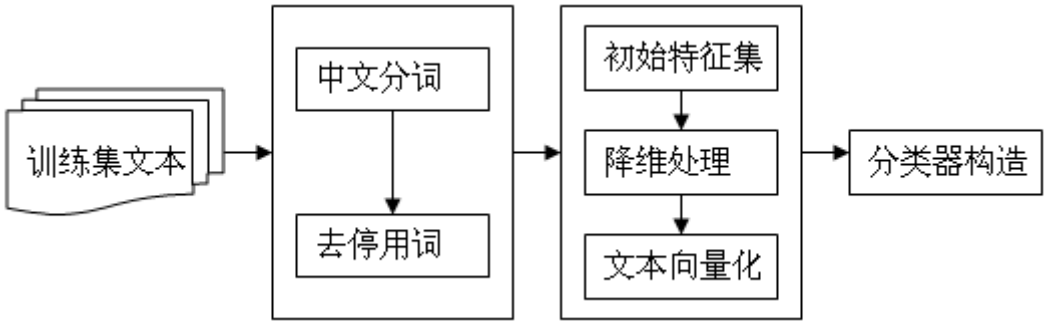

图 2.1 文本分类的训练过程

在测试阶段，我们同样需要对测试集中的文档进行中文分词和去停用词。然后用训练时生成的特征子集对测试集文档进行向量表示，最后用已构造的分类器进行分类。具体流程如图 2.2 所示。

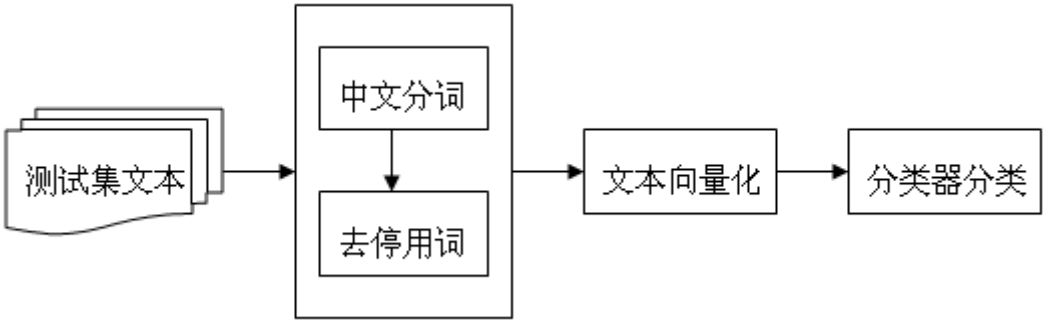

图 2.2 文本分类的测试过程

2.4 文本预处理技术

文本预处理是进行文本分类操作的第一步，预处理结果的好坏直接影响到文本分类的结果。不同语料库的存储格式各有不同，特别是有些语料是直接从互联网上获得的，内容复杂，格式不规范并且编码格式多样。因此，必须经过一定的处理，去除语料库中的噪声信息，对文本内容进行规范化处理，使得文本符合分类模型的输入要求。

文本预处理的目的是从文本语料库中规范地提取出主要内容，剔除与文本分类不相关的信息。这些操作主要包括以下几个方面:

(1)文本标记的处理

一般情况下，文本中除了表示文本内容的信息外，一般还会包括一些与内容无关的标记，如控制文本显示外观的标记、标点符号、图像、声音、动画等其它媒体信息，甚至有可能是乱码。这些标记与文本内容无关，对文本的分类没有帮助。由于中文文本分类处理对象是纯文本信息，所以说这些标记都是中文文本分类中的噪声数据，在对文本进行分类处理之前需要对待分类文本进行预处理，去除这些对分类没有贡献的标记。

(2)中文分词处理

由于中文文本的词与词之间不像英文的单词与单词之间具有一个形式化的分界符，所以在对中文文本进行分类处理之前需要对待分类的文本进行分词处理。中文文本的分词处理技术就是将连续的汉字序列按照一定的规则重新切分为词或词组的过程。由于汉语句子具有复杂性和多样性的特点，使得中文分词处理成为中文文本分类的一大难题。目前，常用的分词算法主要有以下三类[33,34]：

①基于词典匹配的分词方法

基于词典匹配的分词方法又称为机械性分词方法，其主要思想是:将待分词的中文字符串按照某种策略与词典中的词条进行匹配，如果在词典中找到了某个字符串，则表示识别出了一个词，匹配成功。按照对待分词字符串扫描方向的不同，基于词典匹配的分词方法可分为正向匹配、逆向匹配以及双向匹配；按照不同长度优先匹配情况，基于词典匹配的分词方法又可分为最大匹配和最小匹配。基于词典匹配的分词方法分词算法简单，分词效率也较高，但是它具有完全依靠分词词典的缺点。由于汉语语法复杂，导致分词词典具有不完备、规则不一致的问题，这也使得完全依靠分词词典的基于词典匹配的分词方法无法胜任大规模文本分词处理任务。目前常用的基于词典匹配的分词方法主要有:正向最大匹配、逆向最大匹配、最少切分、全切分等。

②基于统计的分词方法

从形式上看，词是稳定的字的组合，在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词[35]。也就是说字与字相邻出现的概率可以作为判断相邻字能否组成一个词的依据。基于统计的分词方法就是对待分词字符串中的各个字的组合频率进行统计，把各个字的组合频率作为判断其能否构成一个词的依据。当字与字之间的组合频率高于某个阈值时，便认为此字组可以构成一个词，否则则认为此字组不能组合成一个词。从基于统计的分词方法的分词依据我们可以看出该分词方法并不需要分词词典，完全依靠对待分词字符串序列进行字组频率统计来进行分词处理，所以该分词方法又称为无词典分词方法。基于统计的分词方法具有经常识别出共项频率高但并不是有意义词的缺点，由于汉语语法的复杂性，基于统计的分词方法的分词精准度较低，且时空开销比较大。在实际应用中，通常的做法是将基于词典匹配的分词方法和基于统计的分词方法相结合，即在利用一部基本的分词词典进行串匹配操作的同时使用基于统计的分词方法来识别出一些新的词，将串匹配分词与串频统计的方法相结合，在发挥串匹配分词切分速度快、效率高优点的同时，又发挥了串频统计分词方法识别新词、自动消除歧义的优点。

③基于理解的分词方法

基于理解的分词方法主要是通过让计算机模拟人类对自然语句的理解，来达到分词的目的。基于理解的分词方法的基本思想是:在对待分词字符串序列进行分词的同时，对该字符串序列进行句法、语义的分析，通过对句法和语义进行分析来处理歧义现象。基于理解的分词方法通常由三个部分组成:一是分词子系统，二是句法语义分析子系统，三是总控系统。其一般的处理流程是:首先总控系统控制协调整个分词操作，句法语义分析子系统对待分词字符串序列进行句法和语义上的分析，在此基础上分词子系统对待分词字符串序列进行分词以及歧义处理操作。从上面的处理流程可以看出，基于理解的分词方法模拟了人类对自然语言的处理，它需要使用大量的语言知识和信息。由于汉语语言知识的复杂性，目前还难以将各个汉语语言知识和信息组织成计算机可以直接理解的形式，因此目前基于理解的分词方法还处于试验阶段，仍存在大量值得研究的地方。

(3)去除停用词

一般情况下我们通过名词、动词和形容词等实词来体现文本的内容，而虚词以及在文本中经常出现但并不表示文本内容的词称为停用词。由于这些停用词并

不表示文本的实际意义，所以它们对文本分类没有任何贡献，相反它们反而会增加分类算法处理文本的时间和空间复杂度。所以为了降低存储空间，提高文本分类算法的分类效率和分类精度，我们需要对文本进行去除停用词的处理。通常情况下对文本进行去除停用词的处理是通过构造停用词表来实现的，即将分词所得到的文本初始特征词集中的每个词与停用词表中的词进行匹配，如果该词在停用词表中出现，则表示该词为停用词，应该去除；否则保留该特征。由于对文本去除停用词的处理依赖于停用词表，所以停用词表的完备性和科学性对去除停用词的处理结果有较大的影响。

2.5 文本表示模型

文本是由大量字符构成的字符串，是一种非结构化的数据，在进行文本分类的时候无法直接作为计算机分类算法的输入数据。所以，将文本表示成计算机所能够直接处理的数据形式是进行文本分类的必要操作，即对文本数据进行形式化表示。目前，常用的文本形式化表示方法主要有:布尔模型、向量空间模型和概率模型等。

2.5.1 布尔模型

布尔模型[36]是基于特征项的严格匹配模型。其基本思想是:通过建立对应于文本特征项的二值特征变量集合，把文本用这些特征变量来表示，如果文本中包含相应的特征项，则该特征变量取值为“True”，否则特征变量取值为“False”。在具体应用中，通常用特征变量数值“1”表示文本中包含该特征项，用特征变量数值“0”表示文本中不包含该特征项。相应的，在布尔模型中，用户的查询也表示为布尔表达式，检索时，根据用户提交的检索条件是否与文本中的逻辑关系一致将检索文本分为相关文本集和不相关文本集。

布尔模型具有结构简单，检索速度快的优点。但是由于布尔模型使用的是基于二元判定标准的匹配策略，对于文本检索，只有相关和不相关两种状态，缺乏对文本相关性排序的概念，从而限制了检索功能。其次，在实际应用中，将用户的查询转换为布尔表达式并不是一件容易的事情。

2.5.2 向量空间模型

向量空间模型[37]是由G.Salton教授等人在20世纪60年代提出的文本表示模型，它具有文本形式化表示效果好、应用广泛的特点。VSM模型最早应用于信息

检索领域，后来随着文本分类技术的发展，VSM模型又在文本分类领域得到了广泛的应用。VSM模型用向量的形式表示文本，是信息检索领域经典的文本表示模型。

向量空间模型的基本思想是:以特征向量的形式表示文本，两个文本之间的相似度通过文本特征向量之间的相关度来计算。向量空间模型中，文本D表示为由特征词条和特征词条的权重所组成的向量，形式为 $((t_1, w_1)(t_2, w_2)\cdots(t_n, w_n))$，其中 t_i 表示特征词条，w_i 则表示特征词条 t_i 的权重。而两个文本的相似度就通过两个文本向量之间的相关度来度量。目前，常用的计算文本特征向量相关度的方法主要有计算夹角余弦、向量内积和欧几里德距离等，常用的计算公式有：

(1)夹角余弦

由于余弦计算得到的值正好是一个介于0到 1 的数，因此文本的相似度也可以用对应的向量之间的夹角余弦来表示。对于文本d_1和d_2，相似度可以表示为公式 2.4：

$$Sim(d_1,d_2)=\cos\theta=\frac{\sum_{i=1}^{n}w_{1i}w_{2i}}{\sqrt{(\sum_{i=1}^{n}w_{1i})^2(\sum_{i=1}^{n}w_{2i})^2}}\left(1\le i\le n\right) \tag{2.4}$$

(2)向量内积

该方法直接计算内积，计算复杂度低，但误差大。对于文本d_1和d_2，相似度可以表示为公式 2.5：

$$Sim(d_1,d_2)=d_1\cdot d_2=\sum_{i=1}^{n}w_{1i}w_{2i}\qquad(1\le i\le n) \tag{2.5}$$

(3)欧几里德距离

欧几里德距离简称欧式距离，对于文本d_1和d_2，相似度如公式 2.6 所示：

$$Sim(d_1,d_2)=D_E(d_1,d_2)=\sqrt{\sum_{i=1}^{n}(w_{1i}-w_{2i})^2}\qquad(1\le i\le n) \tag{2.6}$$

其中，n是特征总数，即特征空间的维度，w_{1i}表示文本d_1中特征t_i的权重，w_{2i}表示文本d_2中特征t_i的权重。

VSM模型的优点是:通过把文本相似度的计算转化为两个向量相关度的计算，通过计算两个向量的夹角或内积来度量两个文本的相似度，降低了文本相似度计算的复杂度。VSM 模型对特征项的权重以及相似度的计算并没有作严格的规定，

可以根据实际情况选择不同的权重评估函数和相似度计算方法，这使得向量空间模型的应用非常广泛。

VSM 模型的缺点是：认为文本的内容与特征项的位置以及顺序等信息没有关系，这使得向量空间模型损失了大量的有关文本结构和语义等的重要信息。

2.5.3 概率模型

概率模型是基于概率排队原则的文本表示模型。概率排队原则的基本思想是:当文本按概率降序的原则进行排列时可以获得最好的检索性能。对于用户给定的查询，概率模型计算所有文档的概率，然后依照文档概率的大小对文本进行降序排列。概率模型是利用词条与词条以及词条与文档之间的概念相关性来进行信息检索的文本表示模型，它克服了 VSM 模型和布尔模型忽略词条相关性的缺点。

概率模型中，用特征向量$d_i=(w_{i1}, w_{i2}, \cdots, w_{in})$表示文本D，特征向量$q_i=(w_{q1}, w_{q2}, \cdots, w_{qm})$表示用户查询串Q，其中向量$d_i$和$q_i$的权重计算都采用二值计算方法，即$w_{ij}\in\{0,1\}$, $w_{qj}\in\{0,1\}$, 1 表示特征项出现，0 表示特征项不出现。文本D与用户查询串Q的概念相关性计算公式如下所示:

$$p(d,q)=\sum \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \tag{2.7}$$

其中$p_i=r_i/r$, $q_i=(f_i-r_i)/(f-r)$。f为训练文档集中的文档总数，r为文档集中与用户查询相关的文档数，f_i表示训练文档集中包含特征项t_i的文档数，r_i表示r个相关文档中包含特征项t_i的文档数。

概率模型按照相关概率的降序排列规则对文本进行处理，综合考虑了文本集的整体情况。但是概率模型只采用了二值形式，对向量权重的计算较为简单。

2.6 文本分类方法

文本分类就是将待分类的文本按照其主题划分到预先定义好的一个或若干个类别中。文本分类方法是文本分类技术研究的重点，也是设计文本分类器的理论基础。根据分类规则的不同以及类别判断方法的不同，文本分类方法大体上可以分为以下两类:

(1)基于规则的分类分法[2]

基于规则的分类方法的基本原理是:由推理机根据领域专家系统知识推导分类。由此可以看出，该分类方法的基础和前提是领域专家知识，需要在领域专家编制大量推理规则的前提下才能进行有效的分类操作。基于规则的分类方法具有

分类正确率高、分类体系合理的优点。但是由于其分类的效果严重依赖推理规则，开发、管理、扩展费用高且领域性强，不易移植，所以在实际分类系统的设计中很少使用这种分类方法。

(2)基于统计的分类方法[38]

基于统计的分类方法的基本原理是:将文本看作是互不相关的特征词集合，忽略文本的语义结构信息，通过对训练样本的统计和计算，得到可以代表文本和类别的关键特征词条。分类器以所选择的特征形式化待分类的文本，然后分类器根据训练阶段所得到的文本与类别的关系来判定待分类文本的类别。基于统计的分类方法实现简单，同时由于其分类知识是通过对训练语料库分析得到的，分类依据可信度高，所以该分类方法还具有分类准确度高的优点。但是由于基于统计的分类方法实质上是非确定性的定量推理过程，所以存在对小类别文本忽视的缺点。

下面将详细介绍几种常用的文本分类方法。

2.6.1 Navie Bayes方法[39]

Navie Bayes 是一种以贝叶斯定理为理论基础的统计学的分类方法，是一种在已知先验概率和条件概率的情况下求后验概率的模式识别方法。Navie Bayes 分类方法是一种简单有效的分类方法。

Navie Bayes分类方法的基本思想是:在已知先验概率和条件概率的情况下，计算待分类文本属于各个类别的后验概率，然后将待分类文本分到后验概率最大的类别中。其中文本属于某个类别的概率为文本中各个特征词属于该类别概率的综合表达式。Navie Bayes的一个前提假设是：文本的特征词之间是相互独立的，即文本的一个特征词对分类的影响独立于其它特征词对分类的影响。Navie Bayes分类方法中，$d_i=(w_1, w_2, \cdots, w_i)$表示任一待分类的文本，$w_i$表示待分类文本中的特征词条，$c=\{c_1, c_2, \cdots, c_k\}$为预定义的文档类别，Navie Bayes的分类方法有如下定义:

(1)特征词条条件独立性假设:

$$p(d_i \mid c_j)=\prod_{k=1}^{r} p(w_{ik} \mid c_j) \tag{2.8}$$

(2)文本d_i属于类别c_j的概率：

$$p(c_j \mid d_i)=\frac{p(d_i \mid c_j)p(c_j)}{p(d_i)} \tag{2.9}$$

各项的计算方法如下：

$$p(c_j)=\frac{c_j\text{类文档数}}{\text{总文档数}}=\frac{N(c_j)}{\sum_k N(c_k)}\approx\frac{1+N(c_j)}{|c|+\sum_{k=1}N(c_k)} \quad (2.10)$$

$$p(w_i|c_j)=\frac{w_i\text{在类别}c_j\text{文档中出现的次数}}{\text{在}c_j\text{类别所有文档中出现的词的次数}}\approx\frac{1+N(\mathrm{c_j})}{\text{不同词的个数}+\sum_k N_{ki}} \quad (2.11)$$

由于Navie Bayes分类方法是在特征独立性假设的前提下进行文本分类操作的，该假设会影响Navie Bayes的分类结果。

2.6.2 KNN方法[40]

KNN分类算法是一种传统的基于统计的分类方法。KNN分类算法的基本思想为:在训练样本集中找到与待分类文本最近的K个文本，看这K个近邻文本中多数属于哪一类，就把待分类文本分到哪一类。KNN是一种简单高效的文本分类方法，其分类基本步骤如下:

(1)根据特征项集合扫描训练文本向量;

(2)对待分类文本进行向量表示;

(3)在训练文本集合中找到与待分类文本最近的K个近邻，近邻的判别标准一般采用文档向量余弦相似度方法来计算。K值的确定目前还没有好的方法，一般是先设定一个初始值，然后根据实验具体情况再对K值进行调整。

(4)依次计算待分类文本的K个近邻文本相对于各个类的权重，计算方法如下:

$$W(C_j)=\sum Sim(d,d_i)y(d_i,C_j) \quad (2.12)$$

其中，d表示待分类文本，d_i表示d的K近邻文本，$y(d_i, C_j)$为文档类别判定函数，如果文档d_i属于类别C_j则$y(d_i, C_j)$取值为1，否则为0。$Sim(d, d_i)$为待分类文本与K近邻文本的相似度，一般采用文档向量余弦相似度计算方法：

$$\cos(d,d_j)=\frac{d\cdot d_j}{|d||d_j|} \quad (2.13)$$

(5)根据各个类的权重计算结果，将待分类文本划分到权重最大的类别中。

KNN分类方法具有分类方法简单、易于实现、分类出错率低的优点。但是由于需要较大的空间来存储训练样本集，而且对于每个待分类的文本，都要计算其与训练样本集中各个文本的相似度，分类开销较大，因此该分类方法并不适用于大规模的数据集，相反在小规模数据集上该分类方法能够取得较好的分类效果。

2.6.3 类中心向量方法[41]

类中心向量方法是一种基于向量空间模型的简单分类方法。类中心向量分类方法的基本思想是:在分类器的训练阶段，利用训练样本集获得每个类所对应的中心向量。然后在分类阶段，将待分类的文本也用向量表示，计算待分类文本向量与训练阶段所获得的每个类的中心向量的相似度，然后将待分类文本划分到相似度最大的类别中去。如果希望待分类文本可以属于多个类别，则可以按照相似度降序的方法对类别进行排序，然后设定一个相似度阈值，将待分类文本划分到相似度大于等于相似度闭值的类别中去。目前，常用的类中心向量分类方法主要有:Rocchio 方法、Windrow-Hoff 方法、EG 方法等。

类中心向量分类方法具有分类方法简单易行、分类速度快的优点。

2.6.4 支持向量机方法

支持向量机最早是由Vapnik在 1995 年提出，并在 20 世纪 90 年代中后期得以发展和完善。支持向量机主要是根据统计学理论解决二分类模式识别问题[42]。Joachims最早将SVM方法应用于文本分类[43]。在文本分类问题中，SVM将分类问题转化为一系列二分问题。

SVM的目的是找到一个可以将训练样本集中的文本分为两类的超平面，以满足类别边界沿垂直于该超平面方向的距离最大，保证最小的分类错误率。用*(xi,yi)*表示线性可分样本集，其中$x_i \in R^d$，y_i为类别标识，$y_i \in \{-1,+1\}$，R^d为d维欧式空间。用方程g(x)=wx+b表示n维空间线性判别函数，wx+b=0 表示分类平面，经过对判别函数进行归一化处理，使得两类中的所有样本满足条件|g(x)|≥1，则两个类别之间的间隔为 2/‖w‖，这样为了满足类别边界沿垂直于超平面方向距离最大的条件，只需要使‖w‖取值最小即可。为了使分类超平面能够对所有样本正确分类，只需满足以下条件:

$$y_i[(wx_i)+b]-1 \geq 0, i=1,...,n \tag{2.14}$$

则‖w‖取值最小且满足上述条件的分类面就是所要求的最优分类面 H。可以将求解最优分类面的问题看作约束优化问题进行求解，使用 Lagrange 乘数法求解的最小值。

上面主要是支持向量机方法在两类分类问题中应用，对于多类分类问题，支持向量机的实现方法主要有：通过对一系列的两类分类器的组合实现多类分类问

题；通过合并多个分类面的参数到一个最优化问题，然后求解该最优化问题来实现多类分类。

由于支持向量机是针对有限样本情况的分类方法，它能够在有限样本情况下得到全局最优解。同时它对稀疏数据不敏感，能够更好的捕捉数据的内容特征，分类准确率高。其缺点是：难以根据实际问题选择合适的函数，参数调节比较困难，分类比较耗时。

2.6.5 决策树方法

决策树方法是一种多级分类方法，它通过分级的形式把复杂的多类别分类问题转化为若干个简单的分类问题。它采用自上而下的递推方法，通过对实例的推导学习得到分类规则。

决策树分类方法的的基本思想是:在训练阶段从根节点开始对训练语料库中的样本进行测试，根据测试结果将训练语料库中的样本划分为若干个样本子集，每个样本子集构成决策树的一个子节点，递归这一过程，直到各个子节点中的训练样本子集都属于同一类或满足终止条件。这样，就得到了一颗决策树，该决策树由一个根节点，若干个内部节点和若干个终止节点组成。每个终止节点代表一个类别。然后当一个待分类样本出现时，就利用训练阶段所得到的决策树对该待分类文本进行分类操作，把它划分到决策树的某一叶子节点，即某一类别中。

决策树分类方法具有抗噪声能力强、分类精度高的优点。其缺点是在处理大规模样本数据集时分类效率不高。目前，典型的决策树方法主要有 CART 方法、C4.5 方法以及 ID3 方法。

2.6.6 神经网络方法[44,45,46]

神经网络分类方法通过模拟人脑神经网络的基本组织特性来完成文本的分类操作。神经网络分类系统通常为三层组织结构，即输入层、输出层和至少一个隐层。其中输入层神经元的个数代表样本的特征数，输出层神经元的个数代表样本类别数。神经网络实际上是由多个输入、输出连接组成的，其中每个输入、输出连接都有一定的权重。神经网络分类方法的基本思想为:训练阶段，利用训练样本集对神经网络分类系统中的每一个输入、输出连接的权重进行调整，以期得到具有最佳分类效果的神经网络分类器；分类阶段，当一个新的待分类文本到来时，利用训练阶段所得到的分类器将待分类文本从神经网络输入层传输到一个合适的神经网络输出层，完成待分类文本的类别判定工作。

目前，常用的神经网络模型主要有：多层感知机、自适应映射网络等。神经网络分类方法具有自适应性强、鲁棒性以及容错性高的优点。但是由于神经网络分类方法采用“黑盒”策略，缺乏解释能力，且其分类效果在很大程度上依赖于训练样本集，分类训练过程慢，所以并不适用于大规模训练语料库的训练学习[47]。

2.7 实验评估方法

分类评价指标是指在实验过程中使用的一些用来评价分类器分类准确度的量化指标。对于文本分类系统，文本分类评价指标的选择是需要考虑的一个关键点。目前已经有很多种分类评估方法被提出，其中部分评价指标都是从某个角度来对分类器的效果进行评测，即衡量分类器在某个方面的性能。对于文本分类系统的评估测试，国际上有通用的评价指标，其中最常用的包括查全率、查准率、宏平均、微平均等[48]，下面将详细介绍。

2.7.1 查全率与查准率

查全率是指在一个类别中，分类器正确判断为该类的文本数与属于该类的文本总数的百分比。查全率考察的是分类器的完备性，查全率越高说明分类器在该类别上可能漏掉的文本越少。查全率计算公式如下所示:

$$R_j = \frac{TP_j}{TP_j + FN_j} \tag{2.15}$$

其中TP_j指被分类器正确分到类别c_j中的文档数，FN_j表示实际上属于类别c_j但分类器并没有将其正确分到类别c_j中的文档数。

查准率是指在一个类别中分类器正确分类的文本数与分类器在该类别上实际分类的文本总数的百分比。查准率考察的是分类器的正确性，查准率越高则表明分类器在该类别上出错的概率越小。查准率计算公式如下所示:

$$P_j = \frac{TP_j}{TP_j + FP_j} \tag{2.16}$$

其中TP_j指被分类器正确分到类别c_j中的文档数，FP_j指实际上不属于类别c_j但却被分类器错误的分到类别c_j中的文档数。

2.7.2 宏平均与微平均

查全率、查准率是从类别的角度来衡量分类器的分类效果，所以查全率和查准率指标只能代表局部意义。而实际上当评估一个分类方法的分类效果时还需要

综合各个类别的分类效果以综合评估一个分类器的分类效果。综合评估分类器分类效果的方法主要有两种，即宏平均和微平均。

宏平均的计算方法是:首先计算各个类别的查全率和查准率，然后取各个类别查全率和查准率的算术平均。从宏平均计算方法可以看出宏平均强调每个类别对整体结果的影响，宏平均的计算公式如下:

$$MacAvg\_\mathrm{Re}call=\frac{\sum_{j=1}^{|c|}\mathrm{Re}call_j}{|C|} \tag{2.17}$$

$$MacAvg\_\mathrm{Pr}ecision=\frac{\sum_{j=1}^{|c|}\mathrm{Pr}ecision_j}{|C|} \tag{2.18}$$

微平均的计算方法是:首先计算所有类别中正确分类的文档总数和错误分类的文档总数，然后再求所有类别的查全率和查准率。从微平均的计算方法可以看出微平均强调大类别对整体结果的影响，微平均的计算公式如下：

$$MicAvg\_\mathrm{Re}call=\frac{\sum_{j=1}^{|c|}TP_j}{\sum_{j=1}^{|c|}(TP_j+FN_j)} \tag{2.19}$$

$$MicAvg\_\mathrm{Pr}ecision=\frac{\sum_{j=1}^{|c|}TP_j}{\sum_{j=1}^{|c|}(TP_j+FP_j)} \tag{2.20}$$

2.7.3 F_β测量值

由于查全率反映的是分类器的完备性而查准率反映的是分类器的准确性，一般情况下查全率会随着查准率的升高而降低，两者相互矛盾。所以为了科学的度量分类器的分类效果，十分有必要综合考虑这两个指标。目前常用的衡量分类器整体效果的方法主要有：Break-even point以及F_β值方法，Break-even point方法将在下一小节中介绍。

F_β值是由Van Rijsbergen在1979年首先提出的，F_β值将查全率和查准率结合为一个指标，以综合考虑分类器的分类效果，两者的相对重要性用参数β表示，F_β值的计算公式如下：

$$F_{\beta}=\frac{(1+\beta)\times P_j\times R_j}{\beta^2\times P_j+R_j} \tag{2.21}$$

其中P_j为查准率，R_j为查全率，β是调节参数(0≤β≤+δ)，用于对查全率和查准率的重要性相关度进行调节。从公式可以看出，当β=0时，F_{β}的值即为查准率，当β=+δ时，F_{β}的值就会与查全率一致。通常情况下取β=1平衡查全率和查准率的重要性，当β=1时，F_{β}即为在实际应用中比较广泛使用的F_1，F_1的计算公式如下：

$$F_1=\frac{2\times P_j\times R_j}{P_j+R_j} \tag{2.22}$$

F1值的宏平均和微平均的计算公式如下:

$$MacAvg\_F_1=\frac{\sum_{j=1}^{|C|}F_{1j}}{|C|} \tag{2.23}$$

$$MicAvg\_F_1=\frac{\sum_{j=1}^{|C|}(R_j*P_j)}{\sum_{j=1}^{|C|}(R_j+P_j)} \tag{2.24}$$

2.7.4 BEP(Break-even point)

分类器的查全率和查准率之间是相互矛盾的，当查全率提高时，相应的查准率就会降低；而当查全率降低时，相应的查准率就会提高。为了平衡查全率和查准率这两个指标，将查全率和查准率调整到一个相等的值，这个值即为分类器的break-even点[49]。BEP的计算比较复杂，通常情况下采用查全率和查准率的算术平均值来代替BEP的值，BEP的计算公式如下：

$$BEP_j=\frac{R_j+P_j}{2} \tag{2.25}$$

BEP的宏平均、微平均计算公式如下所示：

$$MacAvg\_BEP=\frac{\sum_{j=1}^{|C|}BEP_j}{|C|} \tag{2.26}$$

$$MicAvg\_BEP = \frac{\sum_{j=1}^{|c|}(R_j + P_j)}{2} \tag{2.27}$$

2.8 本章小结

本章主要介绍了中文文本分类的几个关键技术，包括文本预处理、文本分类的表示模型以及文本分类的评价指标等。着重讨论了几个经典的文本分类方法，并对各种方法的优缺点进行了分析。

第三章 文本特征选择方法

3.1 文本特征选择概述

作为数据预处理的关键步骤，特征选择是数据挖掘、机器学习和模式识别中的一个重要研究问题。对特征选择的研究开始于 20 世纪 60 年代，早期用于信号处理问题和统计学研究，特点是涉及特征较少且依据特征间独立性假设。20 世纪 90 年代起机器学习问题开始涌现，面对大规模数据的处理，已有的特征选择方法在准确性和效率等方面受到很大挑战，人们开始致力于研究适应大规模数据处理的特征选择算法。目前，多样化和综合性的特征选择研究蓬勃发展，多种新的评价标准和搜索算法被研究并用于特征选择，而且出现了多种技术相结合的研究方法。

3.1.1 特征选择的意义

特征选择是指从一组原始特征集合中选择具有代表性的特征子集，使其保留原有数据的大部分信息，即所选择的特征子集几乎可以像原来的全部特征一样用来正确区分数据集中的每个数据对象。

具有d个特征的特征集合含有2^d个特征子集，穷举所有子集是不现实的也是不可能实现的。通常的做法是根据某种准则获取一个较小的特征子集，使得分类任务的效果与原始特征集下的效果近似甚至更好。这是因为，特征选择通过删除与分类任务无关和冗余的特征，使得原来复杂的数据模型被简化，得到的简化模型常常是对数据集更精确的描述而且易于理解。

对高维数据进行特征选择的作用和意义可总结如下[9]：

(1) 方便可视化和数据理解。高维数据通常难以进行全面的可视化处理，人类自身也难以理解高维数据隐含的某些关系。另外，在高维数据上进行数据挖掘的结果通常也很难被理解和解释。

(2) 降低数据度量和存储的条件。

(3) 降低学习器训练和使用的时间。数据的维度是影响学习器训练时间的一个重要因素。数据维度越高，学习器训练的代价越大。另外，如果数据维度过高，学习器训练完成后将不能及时完成分类预测等任务，影响其应用。

(4) 战胜维度灾难从而提高分类预测的精度。特征选择算法采用各种策略有针对性地选取最有效特征，平衡特征个数与样本个数之间的关系，最终达到改善学习器性能的目的。

3.1.2 特征选择的分类

本质上，多数特征选择方法可以看作是一个搜索问题[50]，搜索的目的是在所有可能的特征子集中获得评估值最大的特征子集作为最优解。已证明，最优特征子集的搜索是一个NP问题，因此通常使用启发式方法寻找近似最优解。

根据特征子集的获取过程是否和最终的归纳学习算法相关，特征选择方法分为 Wrapper 模型方法和 Filter 模型方法。前者直接使用具体归纳学习算法并把它的性能作为评价标准来选择特征。后者独立于具体的归纳学习算法，特征选择时只考虑数据集内在的本质。一般来说，Wrapper 模型考虑归纳学习算法的影响，通常可以对预选的归纳学习算法有较好的性能。但是，Wrapper 模型方法获得的特征并不一定对更多或其它的学习算法有较好的适用性，而且，其计算复杂度取决于学习算法，代价通常非常高昂，所以，当特征数量很大时，Filter 模型是理想的选择。因此在文本分类领域，通常选择 Filter 模型。Filter 型特征选择算法的一般步骤为：首先，对每个特征计算重要度值，然后对这些值进行排序，根据预先设定的阈值剔除低值特征，或选择预定个数的特征，得到的特征子集将作为分类算法的输入。

从特征评价的方式上，特征选择方法可以分为特征子集评价的方法和单一特征评价的方法。前者从特征子集整体考虑，评价和比较多个特征子集的类别区分能力，选择区分能力最强的特征子集。后者对特征的类别区分能力进行单独评价，然后选择一系列相对区别能力较强的特征组成特征子集。一般来说，因为考虑了特征子集的整体特征，使用特征子集评价的特征选择方法通常比单一特征评价的选择方法表现更优[51,52]。

根据评价函数是否依赖类别信息，特征选择方法可以分为无监督的特征选择和有监督的特征选择。前者不依赖分类系统的类别信息，主要包括文档频率、Term 强度等。后者依赖类别信息，主要包括信息熵、互信息、χ^2 统计量等。文本分类中，考虑类别信息进行特征选择所得的分类效果往往好于不考虑类别信息的情况。

有监督的特征选择根据其选择的范围可以分为局部的特征选择和全局的特征选择。前者在每一个类中都选择一个特征子集；后者在整个的数据集选择一个特征子集，即所有类别的文档都使用同一个特征子集，适用于所有分类算法。

3.2 文本特征选择的特点

文本特征选择就是从总体的词项集合中，选择那些具有较强的类别区分能力的词项作为特征项。

文本特征选择的功能可概括为两个方面：

(1) 避免过拟合，提高分类准确度。

如果经过某种学习之后的分类模型对于训练文档可以有很高的自动分类精度，但对训练集外的文档适应性差很多，特征选择可以使得分类模型对训练集和待分类文档有基本一致的适应性。

(2) 通过降低特征空间维度，降低计算的时间和空间复杂度，提高分类精度。

在文本分类系统中，经过分词、过滤停用词等处理后，使用向量空间模型表示的文本数据存在高维特性，即文档空间涉及的总词项数很大，即文档向量维数很大，上万甚至更高的数量级。但是，每个文档可能涉及的词项只是其中的一小部分，而且，词项集合中还存在大量的对分类没有贡献的噪音词，如冷僻词、常用高频词等，噪音词参与分类算法将对分类精度有很大影响。因此，通过特征选择降低特征空间维度，不但能够大大降低计算的时间和空间复杂度，而且能够提高分类精度。

数据空间存在的高维特性，如文本领域、基因分析领域的数据空间的高维性，常常使得很多特征选择技术难以适用。Wrapper 模型方法计算复杂度高，对于高维数据几乎无法实现。完全特征子空间搜索方法也存在同样的问题，它的时间复杂度至少是$O(N^2)$,对于成千上万甚至更高维的文本数据特征空间来说，这种最优化的搜索策略几乎是行不通的。

因此，在文本分类领域常用的特征选择方法，如文档频率、互信息等可以归结为基于 Filter 模型的单一特征选择，此类特征选择方法研究的重点是用来衡量特征词重要性的评价函数。其过程是：引入特征之间的条件独立性假设，通过构造一个评价函数，单独对每个特征进行评价，根据评价值从大到小进行排序，然后选择评估值较大的特征构造最佳的特征子集。

这种方法显然不能保证找到最优特征子集，甚至也不能保证找到一个较好的次优特征子集，但由于其计算量非常小，从而使得高维特征空间通过特征选择进行降维成为可能。业已证明，即使在各个原始特征相互独立的假设条件下，各个最优特征构成的子集未必是最优特征子集，但是文本分类的大量研究实践也表明，很多对单个特征考察优劣的特征选择方法仍有一定的有效性。

3.3 常用的特征选择方法

3.3.1 文档频率

一个特征词条t_i的文档频率是指在训练语料库中出现特征词条t_i的文档数[53]。文档频率特征选择方法的基本思想是：首先设定最小和最大文档频率阈值，然后计算每个特征词条的文档频率，如果该特征词条的文档频率大于最大文本频率阈值或小于最小文档频率阈值，则删除该特征词条，否则保留。文档频率特征选择方法是基于如下假设：即如果特征词条的文档频率过小，则表示该特征词条是低频词，没有代表性；相反如果特征词条文档频率过大，则表示该特征词条没有区分度，这样的特征词条对分类都没有多大的贡献，所以将它们删除并不会影响分类效果。

特征词条文档频率用 DF 表示，计算方法为：

$$DF(t_i, c_j) = \frac{\text{类别}c_j\text{中包含特征词条}t_i\text{的文档数}}{\text{类别}c_j\text{的总文档数}} \tag{3.1}$$

文档频方法是一种简单高效的特征选择方法，具有相对于训练文本集规模的线性计算复杂度，能够应用于大规模训练语料库的统计。但是文档频特征选择方法具有如下缺点:首先文档频特征选择方法在对特征词条进行选择操作时认为文文档频率过小的特征词条是低频词，认为它们不含有或含有很少的类别信息，所以将它们删除并不会影响分类器的分类效果。而实际上，这一假设是不全面的，存在文档频率低却能很好反映类别信息的特征词条，文档频率特征选择方法将该类特征词条过滤掉，影响了分类器的分类效果；其次文档频率特征选择方法只考虑了特征词条是否在文档中出现，忽略了特征词条在文档中出现的次数这一重要信息。

3.3.2 信息增益

信息增益是一种基于熵的评估方法，在机器学习领域具有较为广泛的应用。信息增益表示某特征词条在文本中出现前后的信息熵之差。信息增益特征选择方法的基本思想是：计算每个特征词条的信息增益，然后按照信息增益值的大小对特种词条进行降序排列，然后通过选择预定义的特征词条个数的特征或通过删除信息增益值小于预定义信息增益阈值的特征来实现特征选择操作。

用 IG 表示特征词条的信息增益，其计算公式如下:

$$IG(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t) + P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) \tag{3.2}$$

其中$P(c_i)$表示c_i类文档在语料中出现的概率，$P(t)$表示语料中包含词条t的文档的概率，$P(c_i/t)$表示文档包含词条t时属于c_i类的条件概率，$P(\bar{t})$表示语料中不包含词条t的文档的概率，$P(c_i \mid \bar{t})$表示文档不包含词条t时属于c_i的条件概率，m表示类别数。

信息增益特征选择方法的不足在于它考虑了特征词条不出现的情况，虽然特征词条不出现也可能对文本分类有贡献，但是实验表明，这种贡献往往小于考虑特征词条不出现情况所带来的干扰[1]。

3.2.3 互信息

互信息是信息熵的引申概念，它根据特征词条t_i与类别c_j之间的相关程度来度量特征词条与类别的相关度。特征词条t_i与类别c_j的互信息计算公式如下:

$$MI(t_i, c_j) = \log \frac{P(t_i, c_j)}{P(t_i) \times P(c_j)} \tag{3.3}$$

其中$P(t_i, c_j)$表示包含特征词条t_i且属于类别c_j的文本概率，$P(t_i)$表示训练语料库中包括特征词条t_i的文本概率，$P(c_j)$表示训练语料库中c_j类文本出现的概率。

当特征词条与类别相互独立时，即满足$P(t_i, c_j)=P(t_i) \times P(c_j)$时，$MI(t_i, c_j)$的值为 0；当特征词条$t_i$很少在类别$c_j$中出现时，$MI(t_i, c_j)$的值为负数；当特征词条$t_i$集中出现在类别$c_j$中时，$MI(t_i, c_j)$的值很大。也就是说$MI(t_i, c_j)$的值越大，该特征词条被选择的可能性也越大。

在实际应用中，通常用训练语料库中各类文本出现的概念来对互信息进行近似计算。计算公式如下:

$$MI(t_i, c_j) = \log \frac{X \times N}{(X+Y) \times (X+Z)} \tag{3.4}$$

其中X表示包含t_i且属于类别c_j的文档频数，Y表示不包含t_i但属于类别c_j的文档频数，Z表示包含t_i但不属于c_j的文档频数，N表示语料库中的文本总数。

互信息的缺点在于：它没有考虑特征词条出现的频率，在很大程度上受到特征词条的边缘分布的影响[54]。根据互信息的计算公式可知，在相同的条件概率下，稀有词的互信息值更大。这一特点导致了互信息特征选择方法在进行特征选择操作时倾向于选择稀有特征词的缺点。因此在对出现频率差别较大的特征词条进行特征选择时，互信息特征选择方法的效果并不是很好。

3.2.4 χ^2统计

χ^2统计量方法通过计算特征词条t_i与类别c_j的相关程度来进行特征选择操作，并假设特征词条t_i与类别c_j满足具有一阶自由度的χ^2分布[55]。特征词条与类别的相关度与χ^2的值成正比，χ^2值越大，表示该特征词条所携带的类别信息量也越多，则被选择的几率也就越大。

χ^2统计量的计算公式如下所示：

$$CHI(t_i,c_j)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{3.5}$$

其中A表示包含特征词条t_i且属于类别c_j的文本频率，B表示包含特征词条t_i但不属于类别c_j的文本频率，C表示不包含特征词条t_i但属于类别c_j的文本频率，D表示不包含特征词条t_i也不属于类别c_j的文本频率。N=A+B+C+D为总的文本数。

对于多类分类问题，特征词条t_i的计算方法主要有以下两种：

(1)计算特征词条t_i相对于每个类的χ^2统计量值，然后取最大的χ^2统计量值作为该特征词条的最终χ^2值。计算公式如下:

$$CHI_{max}(t_i)=\max_{j=1}^{m}CHI(t_i,c_j) \tag{3.6}$$

其中m为类别数。

(2)计算特征词条t_i相对于每个类的χ^2统计量值，然后取这些值的平均值作为该特征词条的最终χ^2值。计算公式如下：

$$CHI_{avg}(t_i)=\sum_{j=1}^{m}P(c_j)CHI(t_i,c_j) \tag{3.7}$$

3.4 特征选择方法比较

上述各种特征选择方法有各自的优点也有各自的缺点，特征选择方法的效果与分类文档集以及分类算法的选择有一定的关系，因此即使是相同的特征选择方

法由于采用的分类文档集不同或采用的分类算法不同也可能得到不同的分类结果。一些研究者在对多种特征选择方法和分类方法作了大量的对比实验后，得到如下结论[3]：

(1) χ^2 统计量特征选择方法的特征选择效果最好，获得的分类精度最高；

(2)其次是信息增益特征选择方法，其分类精度接近 χ^2 统计量方法；

(3)互信息特征选择方法与 χ^2 统计量以及信息增益特征选择方法相比，在分类精度上有所差异；

(4)文档频特征选择方法是一种高效且成本较低的特征选择方法，通常情况下其特征选择效果优于信息增益。

由于分类文本集的不同，或者所选择的分类算法的不同，上述结论并不一定是绝对的，在某些情况下可能会有差异，但是此结论对于如何验证本文所提特征选择方法的有效性和可行性具有很好的指导意义。

3.5 本章小结

本章对特征选择方法进行了概述，对文本特征选择的主要特点进行了简要的说明,主要介绍了几种常用的特征选择方法。最后对这几种特征选择方法进行了对比分析，为提出本文的特征选择方法奠定了理论基础。

第四章 基于类别区分度和关联性分析的综合特征选择

目前在文本分类领域较常用到的特征选择算法中,仅仅考虑了特征与类别之间的关联性,而对特征与特征之间的关联性没有予以足够的重视。本文提出一种基于类别区分度和关联性分析的综合特征选择算法, 首先利用类别区分度提取出具有较强类别区分能力的特征词来降低特征空间的稀疏性,再通过特征的关联性分析衡量特征与类别的相关性以及特征之间的冗余度,最终选择出具有类别代表性且相互之间不存在冗余的特征词。

本文中的特征选择算法以信息论量度为基本工具,综合考虑了计算代价问题。算法在保留类别相关特征的同时识别并摒弃了冗余特征,能够得到一个更优的特征子集。

4.1 特征的冗余和相关性

4.1.1 特征的冗余

文本数据中,特征和特征之间往往不是独立的,而是存在着某种相关性[56,57]。列举两种可能存在的现象:

(1)特征词在语义方面存在的同义或近义的现象非常普遍,如“计算机”、“电脑”等均是计算机类的类别特征词。

(2)“篮球”、“姚明”等特征词虽然不是同义词,但在区分体育与其它类别时可能有较强的“雷同作用”。

特征的冗余性和特征之间的相关性密切相关,如果两个特征之间的相关性很强,我们认为这两个特征是存在相互冗余的。所以,当只考虑特征和类别之间相关性时,获得的特征子集会存在大量的冗余特征。

4.1.2 特征的相关性

在理论上,较多的特征应该能提供比较强的识别能力,但是当面对实际的数据学习过程时,对于有限的训练数据,过多的特征不仅大大减慢学习程序的速度,同时也会导致分类器对训练数据过度适应的问题,特别是那些与类别不相关的特征和冗余特征,更会误导分类学习算法。因此,我们需要选择出适合文本分类的好的特征子集。那究竟什么样的特征子集才会被称之为好的呢?通常,一个特征子集被定义为好的,则表明特征与类别是强相关的,而互相是不关联的。

为了对相关特征和冗余特征进行分析，下面我们介绍几个特征相关性的定义:

John、Kohavi和Pfleger将特征分成三个互不相交的类别：强相关、弱相关和不相关的特征[58,59]。

设F是所有特征的集合，F_i是其中的一个特征，$S_i=F-\{F_i\}$，C是给定的类别，则相关性的三种类别在下面正式给出：

定义 4.1 强相关性：如果特征F_i满足公式(4.1)，则称特征F_i是与类别C强相关的。

$$P(C|F_i,S_i)\neq P(C|S_i) \tag{4.1}$$

定义 4.2 弱相关性：如果特征F_i满足公式(4.2)，则称特征F_i是与类别C弱相关的。

$$P(C|F_i,S_i)=P(C|S_i),\ \text{且存在}S_i'\subset S_i,\ \text{使得}P(C|F_i,S_i')\neq P(C|S_i') \tag{4.2}$$

定义 4.3 不相关性:如果特征F_i满足公式(4.3)，则称特征F_i是与类别C不相关的。

$$\forall S_i'\subseteq S_i,\ P(C|F_i,S_i')=P(C|S_i') \tag{4.3}$$

一个特征的强相关性表明该特征对一个最优的特征子集来讲总是必需的，在不影响最初的类别分布的情况下该特征不能被除去；一个特征的弱相关性表明该特征对一个最优的特征子集并不总是必需的，但是在某种条件下可能加入到一个最优的特征子集中去；不相关性表明该特征在最优特征子集中总是不必要的。

因此，一个最优的特征子集应该包括所有强相关特征和部分弱相关特征，而不应该包含不相关的特征。也就是说，在特征子集的选择过程中，仅仅去掉不相关特征是远远不够的，因为在弱相关特征中还存在一部分特征为冗余特征。所谓的冗余特征是指该特征与已选特征子集中的某一特征具有很强的相关性。冗余特征的存在仍然会影响分类的性能和效率，如 Naive Bayes 分类器对冗余特征很敏感。所以在特征子集的选择过程中，我们不单单要剔除不相关特征，还要同时除去冗余特征，剔除冗余特征可以在尽量减少信息损失的前提下对特征维数进行更有效的缩减。

为了能够在剔除类别不相关特征的同时，仍然能够除去冗余特征，必须要给出一种算法来判别特征与类别之间，以及与其它的特征之间究竟是强相关的，弱相关的，或是冗余的，以更好地进行特征的筛选。本文首先利用类别区分度提取

出有较强类别区分能力的特征词，再通过特征的关联性分析，衡量候选特征词与类别之间的相关性，即先保证已选特征与类别的关联度最强，然后计算其它特征与已选特征之间的关联度，当某个特征与已选特征有很高的相关度，即使这个特征与类别具有很强的关联性，本文算法也不会将其选入特征子集，因为相比于已有的特征子集，它根本无法提供额外的与类别有关的信息。同时，由于特征之间的相关性或冗余性与类别信息之间是存在联系的，本文算法在计算特征之间的相关度时也考虑了类别信息对两个特征相关冗余的影响。

4.2 类别区分度

现有的一种特征选择算法DPM(Discriminating Power Measure)[60]，是通过计算每个特征在某一类别和剩余其它类别中的文档频，比较了特征对一个类别和对其它类别的贡献，提取出具有强类别区分能力的特征词。在研究此特征选择算法的基础上对该算法加以改进，提出了类别区分度的计算公式，该公式同时考虑了每个特征的类别频次在计算特征类别区分能力方面的重要性。经实验验证，通过类别区分度筛选得到的特征子集能够获得较好的分类效果。

4.2.1 DPM 特征选择算法

DPM 算法的目标是提取出具有最强类别区分能力的特征，算法步骤如下：

步骤 1 假设共有 m 个类别，对每一个类别 $j(1\leqslant j\leqslant m)$，考虑每个特征在类别 j 里的文档频以及在剩余其它类别里的文档频：

特征f_i在第j个类别中的文档频计算公式和在除了第j个类别之外其它类别中的文档频计算公式分别为(4.4)和(4.5)：

$$\mathrm{DF}_j^{\mathrm{in}}=\frac{n(\mathrm{doc}(f_i)_{\mathrm{all}},\mathrm{cat}_j)}{n(\mathrm{doc}_{\mathrm{all}},\mathrm{cat}_j)} \tag{4.4}$$

$$\mathrm{DF}_j^{\mathrm{out}}=\frac{n(\mathrm{doc}(f_i)_{\mathrm{all}},\text{collection-cat}_j)}{n(\mathrm{doc}_{\mathrm{all}},\text{collection-cat}_j)} \tag{4.5}$$

式中f_i表示第i个特征,cat_j表示第j个类别,collection-cat_j表示文档集中除类别j之外其它类别中的文档，$\mathrm{doc}(f_i)_{\mathrm{all}}$表示包含$f_i$的所有文档，$\mathrm{doc}_{\mathrm{all}}$表示文档集中所有的文档，$n(\mathrm{doc}(f_i)_{\mathrm{all}},\mathrm{cat}_j)$ 表示在类别cat_j中包含f_i的所有文档数，$n(\mathrm{doc}_{\mathrm{all}},\mathrm{cat}_j)$ 表示类别cat_j中的所有文档数。同理，$n(\mathrm{doc}(f_i)_{\mathrm{all}},\text{collection-cat}_j)$ 表示除了类别cat_j剩余其它类别中所有包含f_i的

文档数，$n(\text{doc}_{all},\text{collection-cat}_j)$ 表示除了类别cat_j剩余其它类别中所有的文档数。

步骤 2 计算f_i在类别cat_j内和类别cat_j外文档频的绝对差：

$$d_j = |\,\text{DF}_j^{\text{in}} - \text{DF}_j^{\text{out}}\,| \tag{4.6}$$

步骤 3 对每一个类别$j(1\leqslant j\leqslant m)$，计算关于$f_i$的$d_j$值，将所有类别的$d_j$值相加，作为特征$f_i$的DPM值：

$$\text{DPM} = \sum_j d_j \tag{4.7}$$

步骤 4 对特征集里的每一个特征计算相应的 DPM 值，如果某个特征的 DPM 值达到了一个事先给定的阈值，则留下该特征，否则删除。

实验证明，DPM算法提高了分类器的精度，并且性能优于词频、互信息、Minimum aggregation、Entropy等特征选择方法[60]。但该算法只考虑特征发生的文档频而没有考虑单词发生的词频。如果某一特征只在某一类别的少量文档中频繁出现，则通过DPM计算公式计算出来的DPM值很低，在特征选择时这种特征就会被排除掉，但这种在少量文档中频繁出现的特征很有可能对分类的贡献很大。并且假设特征f_1和f_2在所有类别中的文档频相同，那么该算法认为这两个特征的贡献是相同的，而忽略了它们在文档中出现的次数，这显然比较片面。因此，在改进后的类别区分度公式中既考虑了特征的文档频又考虑了特征的类别频次。

4.2.2 类别区分度

定义 4.4 类别频次。特征f_i的类别频次是指特征f_i在某一类别中出现的次数与该类别特征词总数之比，记为TF。

特征f_i在第j个类别中类别频次计算公式为：

$$\text{TF}_j^{\text{in}} = \frac{n(f_i,\text{cat}_j)}{n(\text{f}_{\text{all}},\text{cat}_j)} \tag{4.8}$$

特征f_i在除第j个类别之外剩余其它类别中类别频次计算公式为：

$$\text{TF}_j^{\text{out}} = \frac{n(f_i,\text{collection-cat}_j)}{n(\text{f}_{\text{all}},\text{collection-cat}_j)} \tag{4.9}$$

式中$n(f_i,\text{cat}_j)$表示类别cat_j中f_i出现的次数，$n(f_{all},\text{cat}_j)$表示类别cat_j中所有特征词个数，$n(f_i,\text{collection-cat}_j)$表示除类别$\text{cat}_j$剩余其它类别中$f_i$出现的次数，$n(f_{all},\text{collection-cat}_j)$表示除类别$\text{cat}_j$剩余其它类别中所有的特征词个数。

定义 4.5 类别区分度。表示特征f_i的类别区分能力，用DP(Discriminating power)表示：

$$\mathrm{DP}(f_i)=\sum_{j=1}^{m}\left(\frac{\mathrm{DF}_j^{\mathrm{in}}\times\mathrm{TF}_j^{\mathrm{in}}-\mathrm{DF}_j^{\mathrm{out}}\times\mathrm{TF}_j^{\mathrm{out}}}{\sum_i \mathrm{DF}_j^{\mathrm{in}}\times\mathrm{TF}_j^{\mathrm{in}}}\right)^2 \tag{4.10}$$

其中m为类别的个数，$\mathrm{DF}_j^{\mathrm{in}}$和$\mathrm{DF}_j^{\mathrm{out}}$的计算同公式(4.4)和(4.5)。从分析可知，DP(f_i)综合考虑了特征词的文档频和类别频次在评价特征类别区分能力方面的重要性，DP(f_i)越大则表明特征f_i的分类能力越强，该特征越重要。本文算法首先通过计算特征的类别区分度DP(f_i)选择出有较强类别区分能力的特征，同时去除噪声数据来降低特征空间的稀疏性以及后续冗余处理的时间复杂度。接下来通过实验来验证改进后的类别区分度相对于DPM算法的优越性。

4.2.3 实验结果及分析

(1) 实验设计

实验语料库选用TanCorpV1.0 中文文本分类语料库[61]，该语料库由谭松波博士收集整理。实验选取其中的六个类别，依次是财经、教育、科技、人才、体育、艺术，共 2100 篇文档，每个类别中文档数均为 350 篇，文档集在去除停用词后共 38897 个不同的特征词。

为了评估特征选择算法的有效性，实验在数据集上使用 Weka 软件中 Naive Bayes 算法的十折交叉验证法对六个类的文档进行测试，关于 Weka 软件的介绍见第 5 章。实验在分别运行 DPM 算法、文档频方法以及类别区分度 DP 后比较对应性能的变化。

(2) 实验结果比较

表 4.1 最好情况下准确率和召回率的比较

类别	准确率%			召回率%		
	DP	DPM	DF	DP	DPM	DF
财经	91.2	90.7	89.4	88.3	89.4	86.9
教育	83.5	82.6	76.9	80.9	77.4	77.1
科技	95.9	95.3	95.2	93.4	92.6	90.3
人才	86.6	85.0	83.7	92.3	90.9	89.4
体育	98.0	98.2	97.1	98.0	96.0	94.3
艺术	90.2	87.1	85.4	92.3	92.3	88.6
平均	90.9	89.8	87.9	90.9	89.8	87.8

用 Naive Bayes 作为分类器，TF-IDF 公式计算特征权重，在最好的情况下 DPM 算法、文档频方法以及 DP 算法在每个类别中的分类准确率和召回率如表 4.1

所示。具体看来，不同的特征选择算法在不同的数据集上各有优势，但总体来说，改进后的 DP 算法在分类性能上优于 DPM 算法和文档频方法。

比较在不同的特征数目下分类器的性能。实验结果如表 4.2 所示。

在特征维数变化的情况下，考察分类器的性能变化情况，可以很好的反映一个分类器对数据样本变化的敏感程度。从表 4.2 可以看出，随着特征维数从小到大变化，分类器的F_1测试值也随着特征个数的增加而升高，并将到达一个相对稳定的水平。同时改进后的DP算法的性能并没有因维数的变化而出现较大的波动，因此该算法具有比较稳定的性能。从表 4.2 还可以看出，改进后DP的F_1测试值在特征个数为 2500 的时候达到最高，为 0.908，以后基本维持在 0.904 左右。DPM和DF的F_1测试值分别在特征维数为 2000 和 3500 的时候达到最高，为 0.897 和 0.878，以后分别基本维持在 0.890 和 0.874 左右。对比可知，改进后的DP具有较好的特征选择效果，优于其它两种特征选择算法。这是由于改进的DP在选取特征时，由单词的文档频率和类别频次两个因素同时决定特征的类别区分能力，克服了只有一种因素决定的片面性，取得了较好的效果，因此对DPM算法的改进是有效的。

表 4.2 特征维数变化下的 F1 测试值比较

特征维数	DPM	DF	DP
50	0.828	0.539	0.827
100	0.830	0.712	0.841
300	0.842	0.754	0.869
500	0.871	0.796	0.886
800	0.886	0.835	0.900
1000	0.887	0.859	0.898
1200	0.889	0.862	0.903
1500	0.889	0.869	0.901
1800	0.895	0.872	0.903
2000	**0.897**	0.873	0.903
2500	0.894	0.877	**0.908**
3000	0.890	0.877	0.907
3500	0.893	**0.878**	0.904
4000	0.887	0.876	0.900
4500	0.885	0.874	0.906
5000	0.889	0.873	0.905

4.3 特征的关联性分析

4.3.1 特征的关联

首先区别特征之间的两种类型的关联[62]：

定义 4.6 C-关联：任何的特征f_i和类别C之间的关联叫做C-关联。

定义 4.7 F-关联：特征f_i与f_j(i≠j)之间的关联叫做F-关联。

如果某个特征与类别C有很强的C-关联，但与其它特征有较弱的F-关联，那么这个特征就是需要的。因此特征选择必须解决两个问题：a.如何衡量一个特征与类别之间C-关联的强弱，即如何选择出有很强C-关联的特征词。b.若特征f_i和f_j(i≠j)与类别C均有较强的C-关联，且f_i和f_j之间存在F-关联，如何确定其是否为冗余特征。

4.3.2 特征 C-关联的度量

(1)关联性的度量方法

变量关联性的度量方法大体上可分为两类:一类是线性关联，包括：Pearson积矩相关、线性关联系数、最小衰减误差平方和最大信息压缩等。另一类建立在信息理论上，如熵等。

①线性关联

现假设有两个变量 X 和 Y,则两个变量之间的线性关联可以通过下式来衡量：

$$r_{XY}=\frac{\sum_i (x_i-\bar{x_i})(y_i-\bar{y_i})}{\sqrt{\sum_i (x_i-\bar{x_i})^2}\sqrt{\sum_i (y_i-\bar{y_i})^2}} \tag{4.11}$$

这种度量方法称为 Pearson 积矩相关。其中：$\bar{x_i}$、$\bar{y_i}$ 是 X 和 Y 的均值。r 的取值范围为[-1，1]，如果变量 X 和 Y 之间彼此完全相关，则 r 的值为 1 或者-1；如果二者之间彼此完全独立的话，r 的值为 0。

线性关联简单易度量，但是由于变量之间的关联是极其复杂的，如果总是假设两个变量之间的关联是线性的也是不太合理的，为了克服这个缺陷，我们从另外一个途径来寻找关联性度量的方法。

②基于信息论的关联性度量

熵是通讯与信息理论中一个非常重要的概念，它衡量的是一个随机变量取值的不确定性程度。而就数据集合而言，熵可以作为数据集合的不纯度或者说不规则程度的量度，所谓的不规则程度指的是集合中数据元素之间依赖关系的强弱。设X是n维随机变量，取值概率$P_i=P(X=x_i)$(i=1,2,…,n)，则X的熵定义为：

$$H(X)=\sum_{i=0}^{n} p_i \log_2 \frac{1}{p_i} \tag{4.12}$$

在随机试验之前，我们只了解各取值的概率分布，而做完随机试验后，我们就确切地知道了取值，不确定性完全消失了。这样，通过随机试验我们获得了信息，且该信息的数量恰好等于随机变量的熵，在这个意义上，熵可以作为信息的度量，因此本文中特征 C-关联的度量采用基于信息论的方法。

(2)特征 C-关联的度量方法

由公式(4.12)可知，信息熵度量了时间的不确定性，即信源提供的平均信息量的大小。在事件 Y发生的前提下，事件 X仍存在的不确定性定义为：

$$H(X|Y)=\sum_{j} P(Y_j)\sum_{i} P(X_i|Y_j)\cdot \log_2 \frac{1}{P(X_i|Y_j)} \tag{4.13}$$

获得信息 I(X|Y)定义为：

$$I(X|Y)=H(X)-H(X|Y) \tag{4.14}$$

获得信息 I(X|Y)对于变量 X和 Y是对称的，即：I(X|Y)= I(Y|X)。假设有属性 X、Y与 Z，若有 I(X|Y)> I(Z|Y)，则说明 Y与 X的相关性大于 Y与 Z的相关性。但式(4.14)在处理多值属性时会有偏差，且在计算特征的关联性时，必须把关联值规格化来确保特征之间有可比性，因此采用变换后的

$$S(X,Y)=2[\frac{I(X|Y)}{H(X)+H(Y)}] \tag{4.15}$$

来衡量特征的C-关联[62]。S(X,Y)的取值范围在[0,1]之间，当S(X,Y)=1 时，表明X与Y的关联性大到可以互相替代；S(X,Y)=0 时，表示X与Y是相互独立的。

4.3.3 相关独立度

提出相关独立度(relevant independency)来表示在已知某一类别的情况下两个特征词之间的相互独立程度，也就是两个特征词关于某一已知类别的独立程度。通过计算相关独立度来衡量特征之间的 F-关联，即度量其它特征与已选特征之间的冗余度大小，以最大程度的消除冗余，获得最优的特征子集合。

定义 4.8 条件互信息：在联合集XYZ中，在给定z_k的条件下，x_i与y_j之间的互信息定义为条件互信息，即

$$I(x_i;y_j|z_k)=\log \frac{p(x_i|y_j z_k)}{p(x_i|z_k)} \tag{4.16}$$

可以发现条件互信息提供了一个很好的消除冗余特征的方法，因为它能很好的表达待选择特征与已选特征之间的关系及特征自身的类别区分能力。但条件互信息是非对称的，即若有特征X_i、X_j以及类别Y，$I(X_i;Y|X_j)\neq I(X_j;Y|X_i)$。考虑到对称的公式将更加方便于特征之间关联程度的计算与比较，因此提出一个新的度量两个特征之间关联程度的公式：

$$D(X_i,X_j)=I(X_i;Y|X_j)+I(X_j;Y|X_i) \tag{4.17}$$

式(4.17)有如下性质：

(1) $D(X_i,X_j)$是非负的，即$D(X_i,X_j)\geq 0$。

证明：由于$I(X_i;Y|X_j)=H(X_i|X_j)-H(X_i|Y,X_j)$，同时

$H(X_i|X_j)\geq H(X_i|Y,X_j)$

则$I(X_i;Y|X_j)\geq 0$。同理，$I(X_j;Y|X_i)\geq 0$。即$D(X_i,X_j)\geq 0$。

(2) 如果$X_i=X_j$，则$D(X_i,X_j)=0$。

证明：由于$I(X_i;Y|X_j)+I(X_j;Y|X_i)=2I(Y;X_i,X_j)-I(Y;X_i)-I(Y;X_j)$

如果$X_i=X_j$，则有$I(Y;X_i,X_j)=I(Y;X_i)=I(Y;X_j)$。

即$I(X_i;Y|X_j)+I(X_j;Y|X_i)=0$。

(3) $D(X_i,X_j)$是对称的，即$D(X_i,X_j)=D(X_j,X_i)$。此性质可以从式(4.17)明显得到。

为了更好的比较特征间的关联度，将$D(X_i,X_j)$规格化，使它的值在[0,1]之间，即：

$$RI(X_i,X_j)=\frac{I(X_i;Y|X_j)+I(X_j;Y|X_i)}{2H(Y)} \tag{4.18}$$

证明：由于$I(X_i;Y|X_j)\leq H(Y|X_j)\leq H(Y)$，$I(X_j;Y|X_i)\leq H(Y|X_i)\leq H(Y)$，所以$I(X_i;Y|X_j)+I(X_j;Y|X_i)\leq H(Y|X_j)+H(Y|X_i)\leq 2H(Y)$；因此$RI(X_i,X_j)$的取值范围在[0,1]之间。

式(4.18)衡量了在现有类别Y的情况下，特征X_i和特征X_j之间的独立程度，即为相关独立度。本文算法通过相关独立度来衡量其它待选特征与已选特征之间的F-关联。当$RI(X_i,X_j)=1$时，表示特征X_i和特征X_j关于类别Y是独立的，存在最弱的F-关联；当$RI(X_i,X_j)=0$时，表示特征X_i和特征X_j关于类别Y提供等同的信息量，存在最强的F-关联。

4.4 综合特征选择算法

本文算法首先通过计算每个特征词的类别区分度，得到一个由较强类别区分能力的特征组成的候选特征集合。然后计算候选集合中所有特征的 C-关联，依据大小对其降序排序，将得到的特征序列重新放入候选集合中。选取序列集中第一个特征，即序列集中具有最大 C-关联的特征，将其从序列集中移出放入结果集，判断候选集合中剩余其它特征与该特征的相关独立度是否满足大于等于某阈值 α 的关系，若满足则留下，若某一特征不满足则认为这一特征是冗余的，将其删除。再将剩下序列集中 C-关联最大的特征移出加入到结果集，并且判断候选集合中剩余其它特征与该移出特征的相关独立度是否满足大于等于某阈值 α 的关系，根据判断结果确定冗余特征并删除，直到候选特征集为空。

算法主框架描述如下：

输入：特征集$F=\{f_i|i=1,2\cdots n\}$，类别C，阈值 δ 、 α 。

输出：特征集 result。

```
result=Φ, temp=Φ;
for each fi∈F do begin
if(DP(fi)>= δ) temp←fi
else delete fi
end for            //获得候选特征集合 temp
for each fi∈temp do begin
calculate S(fi,C)
end for            //计算候选特征集合中每个特征的 C-关联
sort(temp)         //将候选特征集合中的特征按 C-关联值降序排序
do{
select first fi∈temp
result←fi,temp←temp\{ fi }
for each fj∈temp do begin
calculate RI(fi,fj)  //计算已选特征与其它候选特征之间的相关独立度
if(RI(fi,fj)<α) delete fj
end for
}while(temp≠Φ)//计算候选特征与已选特征间的 F-关联，消除冗余特征
```

4.5 算法时间复杂度分析

从算法的描述中可以看出，算法的时间复杂度主要由算法中消除冗余部分的多重循环决定。如果直接对 N 个特征逐对分析，算法的时间复杂度则为 $O(N^2)$。在本文算法中，若判断出某一特征与已选特征之间有很小的相关独立度，即它和已选特征之间有很高的相关冗余度，也就是说明已选特征能够完全代表它所带有的类别信息，所以它不再参与和后续其它特征的比较而直接删除，在理想的情况下，集合 temp 中的所有特征均被删除，而最坏的情况下没有任何特征被删除，因此平均来说，假定每次循环删除其中一半特征，则算法的时间复杂度为 $O(N\log N)$。这样就避免了所有特征的逐对比较，降低了算法的时间复杂度。

4.6 本章小结

本章首先分析了在文本特征选择过程中考虑特征的类别区分能力的同时剔除冗余特征词的重要性，接着提出一种新的综合特征选择算法，对该特征选择方法进行了详细的介绍，对算法的框架进行了描述，并分析了该算法的时间复杂度。

第五章 实验设计与分析

5.1 实验环境的构造

为了对前面提出的综合特征选择方法进行验证和评估，本文构造了一个集分词、特征选择、分类于一体的实验环境。其中分词、特征选择和分类 3 个模块之间相互独立但它们之间的接口是统一的。也就是说各个模块可以很方便地调用其它模块，某一模块所作的修改对其它模块是透明的。今后的研究工作中对任何一个模块进行改进时不会引起其它模块的变动。具体的构造方法如下：

5.1.1 实验环境的系统结构

本文的实验环境系统结构如图 5.1 所示：

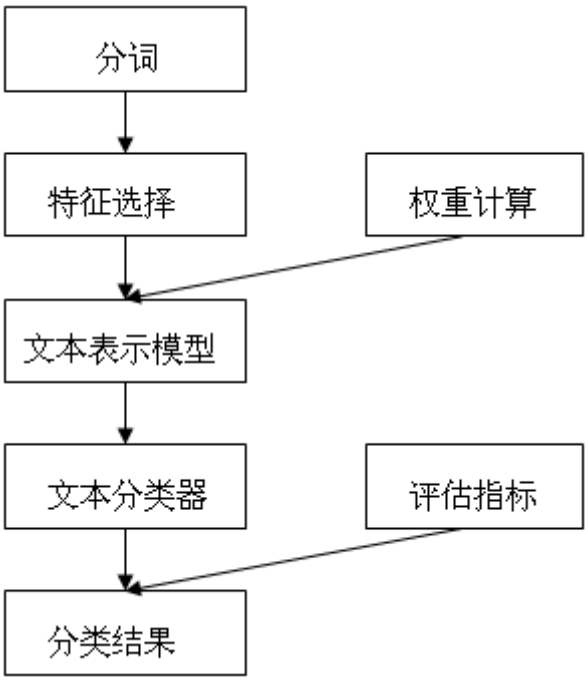

图 5.1 实验环境系统结构

本文的中文文本分类采用的处理流程是这样的：文本集中的文本经过分词预处理后，文本就表示成了一个个特征词组成的集合。通过特征选择，最能代表某一类特征的特征词被选择出来，再经过权重计算，文本表示成了能被分类器识别的模型即文本表示模型，本文采用的文本表示模型是向量空间模型。文本表示模型经过分类器的处理和评估指标的衡量就得到了分类结果，通过对分类结果的分

析，可以判断我们提出的方法是否有效可行。在这个处理流程中，分词、特征选择和分类是三个比较重要的模块，本文分别运用三个开源工具构造了三个相应的系统，下面分别介绍一下。

5.1.2 分词系统

中文词法分析是中文信息处理的基础与关键。分词系统能否达到实用性要求主要取决于两个因素：分词精度与分析速度，这两者相互制约，难以平衡。大多数系统往往陷入“快而不准，准而不快”的窘境。中国科学院计算技术研究所在多年研究工作积累的基础上，研制出了汉语词法分析系统 ICTCLAS(Institute of Computing Technology,Chinese Lexical Analysis System)。中国科学院汉语词法分析系统 ICTCLAS 主要功能包括中文分词、词性标注、命名实体识别、新词识等，同时支持用户词典。中国科学院计算技术研究所先后精心打造五年，内核升级 6 次，目前已经升级到了 ICTCLAS3.0。

ICTCLAS3.0 研制出了完美 PDAT 大规模知识库管理技术，在高速度与高精度之间取得了重大突破，该技术可以管理百万级别的词典知识库，单机每秒可以查询 100 万词条，而内存消耗不到知识库大小的 1.5 倍。基于该技术，ICTCLAS3.0 分词速度单机 996KB/s，分词精度 98.45%，API 不超过 200KB，各种词典数据压缩后不到 3M，是当前世界上最好的汉语词法分析器。

汉语分词牵涉到汉语分词、未定义词识别、词性标注以及语言特例等多个因素，大多数系统缺乏统一的处理方法，往往采用松散耦合的模块组合方式，最终模型并不能准确有效地表达千差万别的语言现象，而 ICTCLAS 采用了层叠隐马尔可夫模型（Hierarchical Hidden Markov Model），将汉语词法分析的所有环节都统一到了一个完整的理论框架中，获得最好的总体效果，相关理论研究发表在顶级国际会议和杂志上，从理论上和实践上都证实了该模型的先进性。

ICTCLAS 全部采用 C/C++编写，支持 Linux、FreeBSD 及 Windows 系列操作系统，支持 C/C++/C#/Delphi/Java 等主流的开发语言；支持 Lucene。

本文的分词系统使用的是 ICTCLAS 的一个 Java 开源版本，该版本没有去除停用词的功能，本文的分词系统对 ICTCLAS 做出了改进，加入了去除停用词的功能。

5.1.3 特征选择系统

本文的特征选择系统是在 Lucene 的基础上做的，这里先介绍一下 Lucene。

Lucene 是 apache 软件基金会 jakarta 项目组的一个子项目，是一个开放源代码的全文检索引擎工具包，即它不是一个完整的全文检索引擎，而是一个全文检索引擎的架构，提供了完整的查询引擎和索引引擎以及部分文本分析引擎（英文与德文两种西方语言）。Lucene 的目的是为软件开发人员提供一个简单易用的工具包，以方便的在目标系统中实现全文检索的功能，或者是以此为基础建立起完整的全文检索引擎。

全文检索是指计算机索引程序通过扫描文章中的每一个词，对每一个词建立一个索引，指明该词在文章中出现的次数和位置，当用户查询时，检索程序就根据事先建立的索引进行查找，并将查找的结果反馈给用户的检索方式。这个过程类似于通过字典中的检索字表查字的过程。

全文检索的方法主要分为按字检索和按词检索两种。按字检索是指对于文章中的每一个字都建立索引，检索时将词分解为字的组合。对于各种不同的语言而言，字有不同的含义，比如英文中字与词实际上是合一的，而中文中字与词有很大分别。按词检索指对文章中的词，即语义单位建立索引，检索时按词检索，并且可以处理同义项等。英文等西方文字由于按照空白切分词，因此实现上与按字处理类似，添加同义处理也很容易。中文等东方文字则需要切分字词，以达到按词索引的目的，关于这方面的问题，是当前全文检索技术尤其是中文全文检索技术中的难点，在此不做详述。

全文检索系统是按照全文检索理论建立起来的用于提供全文检索服务的软件系统。一般来说，全文检索需要具备建立索引和提供查询的基本功能，此外现代的全文检索系统还需要具有方便的用户接口、面向 WWW 的开发接口、二次应用开发接口等等。功能上，全文检索系统核心具有建立索引、处理查询返回结果集、增加索引、优化索引结构等等功能，外围则由各种不同应用具有的功能组成。结构上，全文检索系统核心具有索引引擎、查询引擎、文本分析引擎、对外接口等等，加上各种外围应用系统等等共同构成了全文检索系统。

本文的特征选择系统就是利用 Lucene 的全文检索功能，对分词后的文本建立一个词与文档的倒排索引，特征选择方法就能很方便的统计特征词在文本中出现的频率和词出现在文本中的频率，从而可以用相应的特征选择方法将特征词对文本分类的贡献计算出来，通过比较特征词对分类的贡献大小就能选择出最具代表性的特征。特征选择出来后，经过权重计算，本文的特征选择系统将文本表示成能被分类系统处理的文本表示模型。

5.1.4 分类系统

本文的分类系统是基于 Weka，下面先介绍一下 Weka。

Weka 的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是一款免费的，非商业化的，基于 JAVA 环境下开源的机器学习以及数据挖掘软件。

Weka 自 1993 年由位于 New Zealand 的 the University of Waikato 进行开发，最初的软件基于 C 语言实现。1997 年，开发小组用 JAVA 语言重新编写了该软件，并且对相关的数据挖掘算法进行了大量的改进。2005 年 8 月，在第 11 届 ACM SIGKDD 国际会议上，the University of Waikato 的 Weka 小组荣获了数据挖掘和知识探索领域的最高服务奖，Weka 系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一。

Weka 作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互式界面上的可视化。而开发者则可使用 Java 语言，利用 Weka 的架构开发出更多的数据挖掘算法。

本文的分类系统接收从特征选择系统得到的文本表示模型，用户可以选择 Weka 提供的 Naive Bayes、KNN、支持向量机、决策树等多种分类方法对文本集进行分类。

5.2 实验设计

本文实验选用 2 个中文语料，语料 1：由谭松波等人提供的 TanCorpV1.0, 选取其中的六个类别，依次是财经、教育、科技、人才、体育、艺术，共 2100 篇文档，每个类别中文档数均为 350 篇，文档集在去除停用词后共 38897 个不同的特征词；语料 2：从复旦大学中文文本语料库中选取 10 个类的部分文档，其类别文档分布见表 5.1。

本文用 TF-IDF 公式计算特征权重,其定义如下：

$$W_{ik} = tf_{ik} \times idf_k \tag{5.1}$$

其中，tf_{ik}表示项T_k在文档D_i中的文档内频数，$idf_k=\log(N/n_k)$表示项T_k的反文档频率，n为T_k的文档频率。由于式(5.1)倾向于选择内容比较长的文档，因此，通常标准的TF-IDF公式是对公式(5.1)归一化的公式，如公式(5.2)所示：

$$W_{ik}=\frac{tf_{ik}\times\log(N/n_k)}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2\times[\log(N/n_k)]^2}} \tag{5.2}$$

表 5.1 语料 2 中文档分布情况

类别	文档数	类别	文档数
经济	519	环境	327
体育	359	艺术	294
计算机	628	太空	192
政治	405	历史	466
农业	550	军事	75

为了评估本文特征选择算法的有效性，实验在数据集上使用 Weka 软件中 Naive Bayes 算法的十折交叉验证法对选用文档进行测试。十折交叉验证法是将数据集分成十份，轮流将其中 9 份作为训练集，1 份作为测试集进行实验，每次实验都会得出相应的准确率，十次结果的准确率的平均值即为对算法精度的估计。

5.3 实验结果及分析

5.3.1 不同特征维数下的性能比较

在语料 1 上设定阈值 δ=0.9，在语料 2 上设定阈值 δ=1.5，反复试验阈值 α 以选择不同的特征维数，实验在分别运行本文方法、类别区分词(CDW)、MI、IG 和 χ^2 统计算法后比较对应性能的变化。

在不同特征数目下各种算法的 F1 测试值如表 5.2 和表 5.3 所示。在特征维数变化的情况下，考察分类器的性能变化情况，可以很好的反映一个分类器对数据样本变化的敏感程度。从表 5.2、5.3 可以看出：

(1) 不同的特征选择算法在不同的数据集上各有优势，并且每种方法获得的性能一般都随特征数目的增多而逐渐提高，并将到达一个相对稳定的水平。同时本文特征选择算法的性能并没有因维数的变化而出现较大的波动，因此该算法具有比较稳定的性能。

(2) 语料 1 中的数据集是均匀分布的，语料 2 中的数据集分布不均匀。当语料不同时，五种算法所表现的性能也不同，但实验结果表明数据集的不均匀分布并没有对上述五种算法的性能产生大的影响，并且本文算法在数据集分布不均匀的情况下表现出更好的性能。

(3) 具体看来，CDW 和 MI 算法的性能稍差，并且 CDW 算法在特征维数越高的情况下性能越好，而 MI 和本文算法在特征维数较低的情况下表现出优势，这也体现出本文算法能有效的提取出有类别代表性的特征词，有较好的降维效果。

(4) 从实验结果看出， IG、χ^2 统计和本文算法性能比较接近，但总体来说本文算法的性能优于上述四种特征选择算法，即本文提出的算法是有效的。

表 5.2 五种特征选择算法在语料 1 上的 F1 测试值(%)

特征数	CDW	MI	IG	χ^2 统计	本文方法
100	80.266	81.317	82.519	82.035	82.616
500	80.879	82.978	84.387	84.459	84.529
1000	83.296	83.015	85.612	86.927	86.170
2000	86.530	85.635	88.793	88.243	88.657
3000	87.616	86.464	**88.936**	89.052	**89.924**
4000	87.763	**86.975**	88.109	**89.876**	89.903
5000	**87.938**	86.616	87.666	89.762	89.628

表 5.3 五种特征选择算法在语料 2 上的 F1 测试值(%)

特征数	CDW	MI	IG	χ^2 统计	本文方法
100	69.172	72.381	75.061	74.508	75.331
300	72.573	76.543	79.898	79.863	80.627
500	79.839	80.925	80.256	81.294	82.519
1000	80.312	83.219	84.924	85.329	85.143
2000	84.619	85.243	86.726	86.753	87.235
4000	86.732	**86.889**	87.221	87.190	**87.961**
6000	**86.798**	86.877	87.138	**87.692**	87.679
8000	86.775	86.793	**87.276**	87.429	87.590

5.3.2 各个类的分类情况比较

为了便于对比分析，本文在语料 2 上使用 χ^2 统计、IG 和本文算法进行 Navie Bayes 分类，各个类别的查全率对比图、查准率对比图以及 F1 值对比图分别如图 5.2、图 5.3、图 5.4 所示。其中各个图的横坐标为类别坐标，纵坐标为分类结果度量指标。

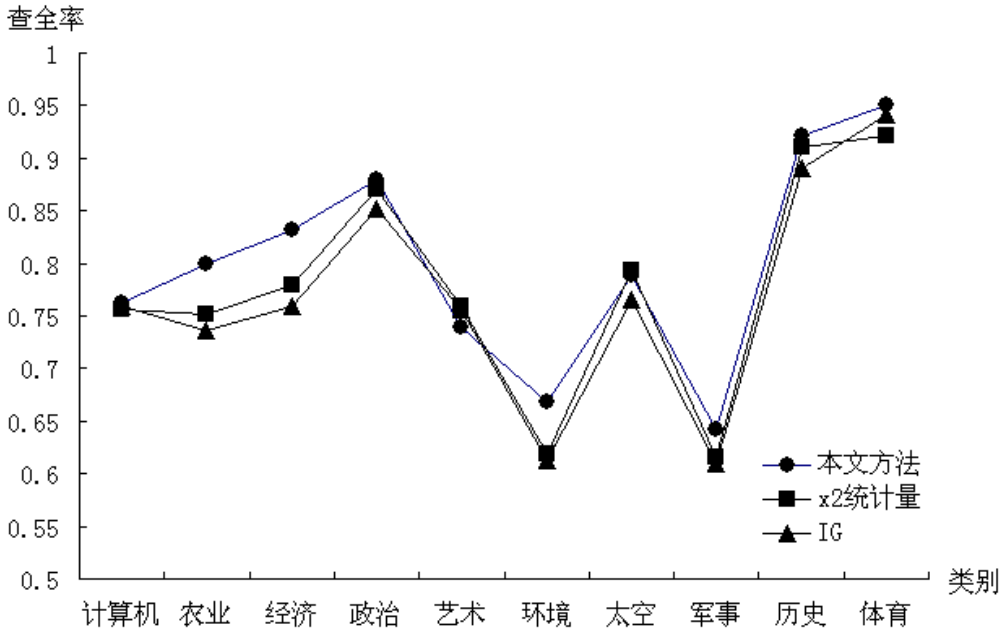

图 5.2 类别查全率对比图

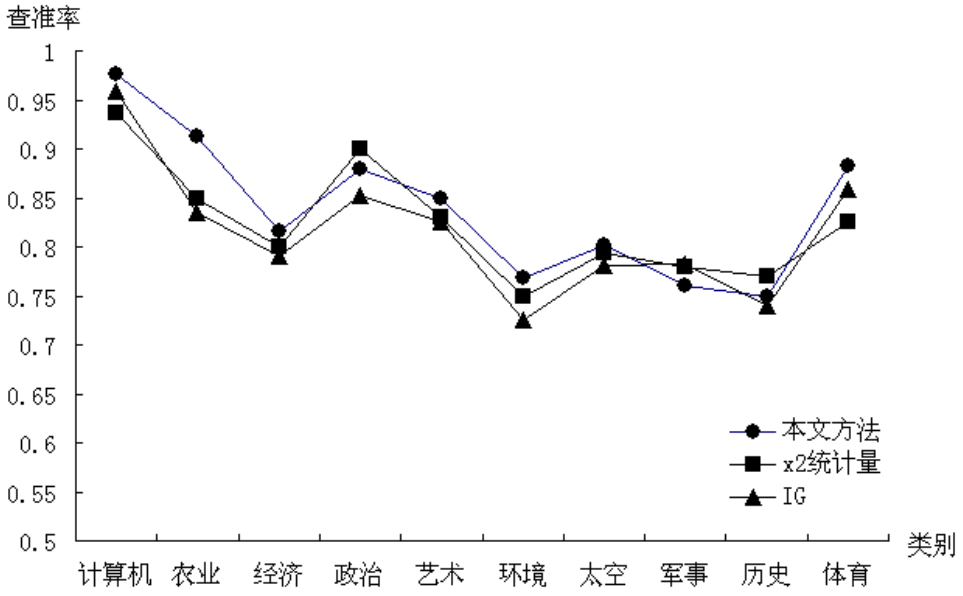

图 5.3 类别查准率对比图

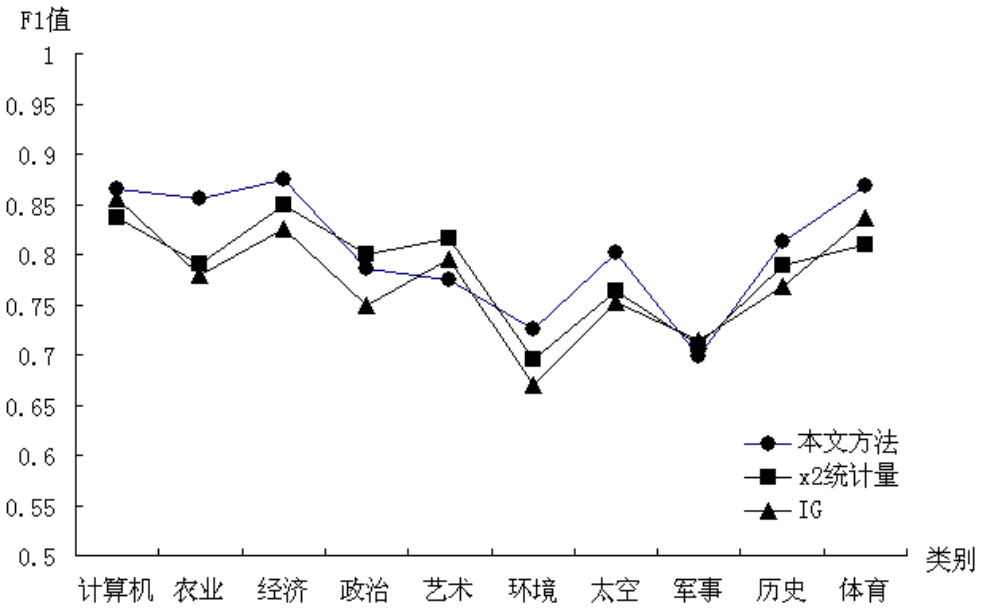

图 5.4 类别 F1 值对比图

从上述各类别查全率、查准率和 F1 值对比图分析可得：使用本文所提出的综合特征选择方法进行文本分类所得的绝大多数类别查全率、查准率和 F1 值相对于使用 χ^2 统计量特征选择方法以及信息增益特征选择方法进行文本分类所得的类别查全率、查准率和 F1 值都有所提高。这表明本文所提出的特征选择方法是有效可行的。

5.4 本章小结

本章详细阐述了本文中文文本分类系统的框架以及各个模块的设计，介绍了实验数据来源，通过在相同的实验数据集上分别使用本文所提出的特征选择方法和几种经典的特征选择方法进行 Navie Bayes 文本分类，对所得到的类别查全率、查准率、F1 值以及不同维数下 F1 值的对比分析，证明了本文所提出的综合特征选择方法的有效性和可行性。

第六章 总结与展望

6.1 总结

本文在研究文本分类的基础上，重点研究了文本分类中的特征选择技术，在此过程中做了如下工作:

（1） 对文本分类技术中几种常用的特征选择方法进行研究，在此基础上发现目前常用的特征选择方法大多都是仅仅考虑了特征与类别之间的关联性,而对特征与特征之间的关联性没有予以足够的重视，因此通常得到的特征子集里会存在大量冗余特征，这些冗余特征无法提供额外的与类别有关的信息，并且会影响分类精度。

（2） 提出一种新的衡量特征词类别区分能力的方法，即类别区分度。类别区分度综合考虑了特征词的文档频和类别频次在评价特征类别区分能力方面的重要性，一个特征类别区分度的值越大则表明该特征的分类能力越强，此特征越重要。因此首先通过计算特征的类别区分度选择出有较强类别区分能力的特征，同时去除噪声数据来降低特征空间的稀疏性以及后续冗余处理的时间复杂度。

（3） 对特征进行关联性分析，提出相关独立度来消除冗余。相关独立度表示在已知某一类别的情况下两个特征词之间的相互独立程度，通过计算相关独立度来衡量特征之间的 F-关联，即度量其它特征与已选特征之间的冗余度大小，以最大程度的消除冗余，获得最优的特征子集合。

（4） 提出综合特征选择算法，该算法在保证较高分类精度的同时避免了候选特征之间的两两比较，降低了特征选择算法的时间复杂度。并在相同的实验数据集上分别采用本文所提特征选择方法和其它经典特征选择方法进行 Navie Bayes 文本分类,通过对采用不同特征选择方法进行文本分类所得的类别查全率、查准率、F1 值以及特征维数变化下的 F1 值的对比分析，证明了本文所提出的综合特征选择方法的有效性和可行性。

6.2 展望

本文的下一步工作主要有以下几个方面：

(1) 本文所提的特征选择方法涉及到类别区分度阈值 δ 和相关独立度阈值 α 两个参数，其中任何一个参数的设置对本文所提的特征选择方法都有一定程度

的影响，本文将进一步对这两个参数进行研究，寻找最优的参数值，进一步提高本文所提的综合特征选择方法的有效性。

(2) 本文实验选取的是复旦大学自然语言处理小组整理的语料集和谭松波博士等收集整理的语料库，不同的语料集对分类效果的影响是不同的，本文下一步应该在不同的语料集上对所提出的综合特征选择方法进行研究。

(3) 语言学上认为处于文章首段、尾段、段首、段尾的句子更能代表文章的中心思想，因此，在文本分类时，来自于首段、尾段、段首、段尾的特征词应该赋予更高的权重以提高文本分类的准确度，论文下一步将在该方面继续研究。

参考文献

[1] Yiming Yang,Jan O.Pedersen.A Comparative Study on Feature Selection in Text Categorization[C].Proceedings of ICML-97,14th International Conference on Machine Learning,1997:410-420.

[2] Fabrizio Sebastiani.Machine learing in automated text categorization[J].ACM Computing Surveys,2002,34(1):1-47.

[3] 荣光.中文文本分类方法研究[D].济南：山东师范大学，2009:21-23.

[4] 庞剑锋.基于向量空间模型的文本自动分类系统的研究与实现[J].计算机应用研究，2001:18(9):23-26.

[5] 丁琼.基于向量空间模型的文本自动分类系统的研究与实现[D].上海:同济大学,2007:24-26.

[6] 苏力华.基于向量空间模型的文本分类技术研究[D].西安:西安电子科技大学,2006:15-18.

[7] 周茜,赵明生.中文文本分类中的特征选择研究[J].中文信息学报.2004,18(3):17-23.

[8] Shan Songwei,Feng Shicong,Li Xiaoming.A comparative study on several typical feature selection methods for Chinese web page categorization[J]. Journal of the Computer Engineering and Application.2003,2003,39 (22):146-148.

[9] Guyon Isabolle,Elisseeff Andre.An introduction to variable and feature selection [J].The Journal of Machine Learning Research,2003,3:1157-1182.

[10] W.Siedlecki and J.Sklansky.On antomatic Feature selection[J].Int.J.Pattern Recognition Art,Intell,vol.2,no.2,1988:197-200.

[11] J.Doak,An evaluation of feature selection methods and their application to computer security[J].Technical report.Davis CA:University of California, Department of Computer Science,1992:10-22.

[12] A.K.Jain,R.P.W.Duin,J.Mao,Statistical pattern recognition:A review[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2000:4-37.

[13] M.Dash and H.Liu,Feature Selection for Classification[J].Intelligent Data Analysis,Elsevier,Vol.I,No.3,1993:123-129.

[14] S.M.Weiss and C.A.Kulikowski,Computer Systems That Learn[J].Morgan Kaufmann Publishers.San Mateo,California,1991:133-149.

[15] J.G.Dy and C.E.Brodley,Feature subset selection and order identification for unsupervised learning[C].In Proceedings of the Seventeenth International Conference on Machine Learning,2000:247-254.

[16] M.Dash,H.Liu and J.Yao.Dimensionality reduction of unsupervised data[C].In Proceedings of the Ninth IEEE International Conference on Tools with AI(ICTAI97),November,1997,Newport Beach,California,1997:532-539.

[17] M.Dash and H.Liu.Handling large unsupervised data via dimensionality reduction[C].In Proceedlings of 1999 SIGMOD Research Issues in Data Mining and Knowledge Discovery,Workshop,1999:221-232.

[18] L.Talavera.Feature selection as a preprocessing step for hierarchical clustering [C].In Proceedings of International Conference on Machine Learning,1999:145-157.

[19] P.Mitra,C.A.Murthy,S.K.Pal.Unsupervised feature selection using feature similarity[J].IEEE Trans Pattern Recognition and Machine Intelligence,2002, 3(24):301-312.

[20] A.L.Blum.Learning Boolean Functions in an Infinite Attribute Space[J]. Machine Learning,1992,9(4):356-373.

[21] J.R.Quinlan.Learning efficient classification procedures and their application to chess end games[J].Machine Learning:An artificial intelligence approach,San Francisco,CA:Morgan Kaufmann,1983:463-482.

[22] J.R.Quinlan.C4.5:programs for machine learning[J]. San Francisco:Morgan Kaufmann,1993:242-267.

[23] L.Breiman,Friedman J H,et al.Classification and Regression Trees[J].Wadsforth International Group,1984:221-244.

[24] TM.Cover.The best two independent measurements are not the two best[J].IEEE Transactions,Syst Man Cybern,1974,4(2):116-117.

[25] G.John,R.Kohavi,K.Pfleger.Irrelevant features and subset selection problem[C]. The Eleventh International Conference on Machine Learning,1994:121-129.

[26] D.W.Aha,R.L.Bankert.Feature selection for case-based classification of cloud types[J].In:working notes of the AAAI94 workshop on case-based reasoning,1994:106-112.

[27] G.M.Provan,M.Singh.Learning Bayesian networks using feature selection [C].In:Proc.5th Intern Workshop on AI and Statistics,1995:450-456.

[28] I.Inza,P.Larraaga,B.Sierra.Feature subset selection by Bayesian networks based on optimization[J].Artificial Intelligence,2001,123(1):157-184.

[29] R.A.Caruana,D.Freitag.Greedy attribute selection[C].The Eleventh intl.conf.on Machine Learning,1994:28-36.

[30] A.W.Moore,M.S.Lee.Efficient algorithms for minimizing cross validation error[C].The Eleventh Intl.Conf.on Machine Learning,1994:190-198.

[31] R.Kohavi,G.H.John.Wrappers for feature subset selection[J].Artificial Intelligence journal,special issue on relevance,1997,97(1):273-324.

[32] 李荣陆.文本分类及其相关技术研究[D].上海:复旦大学,2005.

[33] 史伟.中文自动分词关键技术研究与实现[D].四川：电子科技大学,2008.

[34] Kai Ying Liu,Jia Heng.Research of automatic Chinese word segmentation [C].Proceedings of 2002 International Conference on Machine Learning and Cybernetics,2002:805-809.

[35] 揭春雨,刘原,梁南元.论汉语自动分词方法[J].中文信息学报,1989,3(1):2-9.

[36] Kevin Chen-Chuan Chang,Hector Garcia-Molina,Andreas Paepcke.Boolean Query Mapping Across Heterogeneous Information Sources[J].IEEE Transactions on Knowledge and Data Engineering,1996,8(4):515-512.

[37] Ricardo Baeza-Yates,Berthier Ribeiro-Neto.Modern Information Retrieval [C].Chinese Machine Press,2003:143-145.

[38] C.Apte,F.Damerau,SM.Weiss.Text mining with decision rules and decision tree[C].In Proceedings of the Conference on Automated Learning and Discovery Workshop 6:Learning from Text and Web,1998:501-507.

[39] A.McCallum,K.Nigam.A comparison of event for naïve bayes text classification[A].Learning for text categorization: papers from the 1998 workshop[C]. AAAI Press,1998:41-48.

[40] Makato Jwayama.A comparison of category search strategies[C].ACM Conference on Research and Development on Information,1995:123-124.

[41] W.Cohen,Y.Singer.Context-sensitive learning methods for text categorization [C].In Proceeding of 19th Annual International ACM Conference on Research and Development in Information Retireval SIGIR,1996:207-308.

[42] V.Vapnik.Nature of Statistical Learning Theory[M].New York:Springer Press,2000.

[43] T.Joachims.Text categorization with support vector machines learning with many relevant feature[C].Proceedings of 10th European Conference on Machine Learning,1998:137-142.

[44] 高亚波.文本分类系统的设计与实现[D].北京:北京交通大学,2008.

[45] F.Sebastiani.A tutorial on automated text categorization[C].Proceedings of ASAI-99,1st Argentinian Symposium on Artificial Intelligence,1999:7-35.

[46] Wiener,Pedersen,Weigend.A neural network approach to topic spotting [C].Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval,1995:317-332.

[47] 肖兆武.信息检索系统中信息模型的建立方法研究[J].科技情报开发与经济,2005,15(8):74-76.

[48] Yiming Yang and Xin Liu,A re-examination of text categorization methods [C],Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval,1999:42-49.

[49] Fabrizio Sebastiani.Machine learning in automated text categorization[J].ACM Computing Surveys,2002,334(1):1-47.

[50] Yang Y,Zhang J,Kisiel B.A calability analysis of classifiers in text categorization[A]. Proceedings of the 26th ACM International conference on Research and Development in Information Retrieval[C].Toronto:ACM Press,20 03:96-103.

[51] Qu G,Hariri S,Yousif M.A new dependency and correlation analysis for features[J].IEEE Transaction on Knowledge and Data Engineering,2005, 17:1199-1207.

[52] Wang G,Lochovsky F.H,Yang Q.Feature selection with conditional mutual information maxmin in text categorization[A].Proceedings of the 13th ACM Conference on Information ang Knowledge Management[C].New York:ACM press,2004:342-349.

[53] 许高建,路遥,胡学刚,涂立静.一种改进的文本特征选择方法研究与设计[J].苏州大学学报,2008,8(11):31-32.

[54] 张俊丽.文本分类中的关键技术研究[D].湖北:华中师范大学,2008.

[55] 邓琦,苏一丹,曹波,闭剑婷.中文文本体裁分类中特征选择的研究[J].计算机工程,2008,34(23):89-91.

[56] Jensen R,Shen Q.Fuzzy-rough attribute reduction with application to web categorization[J].Fuzzy Sets and Systems,2005:469-485.

[57] Iosif E,Potamianos A.Unsupervised Semantic Similarity Computation Between Terms Using Web Documents[J].IEEE Transactions on Knowledge and Data Engineering,2010,22(11):1637-1647.

[58] I Guyon,A Elisseeff.An introduction to variable and feature selection. Journal of Machine Learning Research.2003(3):1157-1182.

[59] L Yu,H Liu.FCBF-Feature Selection for High-Dimensional Data.In Proceedings of the twentieth International Conference on Machine Learning,Washington DC,USA.2003:856-863.

[60] Chih-Ming Chen,Hahn-Ming Lee,Yu-Jung Chang.Two novel feature selection approaches for web page classification [J].Expert Systems with Application,2009, 36:260-272.

[61] 谭松波,王月粉.中文文本分类语料-TanCorpV1.0[EB/OL].http://www.searchforum.org.cn/tansongbo/corpus.htm.

[62] Lei Yu,Huan Liu.Efficient feature selection via analysis of relevance and redundancy[J]. Journal of Machine Learning Research,5(2004):1205-1224.

致　　谢

在论文完成之际，回首三年的求学历程，脑海中浮现的是老师们对我的悉心指导，在此向他们表达最诚挚的感谢。

首先谨向我尊敬的导师王治和教授致以诚挚的谢意和崇高的敬意,本论文在选题及研究过程中得到王老师的悉心指导，王老师丰富的实践经验、渊博的专业知识、务实忘我的工作作风、以及宽厚待人的处事态度使我受益匪浅。更让我明白了做事应有的踏实和执着。在今后的工作和生活中，王老师做人做事的态度必将让我终生受益。

本论文得以定稿，离不开三年来学院所有老师的帮助和指导，尤其要感谢蒋芸老师、杨勇老师、贾俊杰老师、王小牛老师、蔺想红老师、冯慧芳老师等的悉心指导和帮助。正是由于他们的精心授课和热心辅导才使我克服重重困难，完成学业。

感谢我的师兄王凌云、席元鸿，师姐杨天霞、王华、王晓霞、张国荣，他们是我学习上的第二导师，给了我极大的帮助；感谢三年来一起学习的许虎寅和樊东辉，他俩是这三年来给我最大帮助和支持的人；感谢师妹潘丽娜、党辉、任钊婷和师弟杨晏，感谢一起生活三年的李小艳、闫秋粉、焦卫丽、姜金娣、王明芳、万红娟等，有你们的陪伴使我的研究生生活才如此丰富多彩；感谢所有提到和没提到的同学、朋友们，谢谢你们在我三年的研究生学习和生活中给我的大力帮助。

衷心感谢在百忙之中抽出时间审阅本论文的专家教授，感谢答辩委员会的各位老师和专家们对我的论文提出的宝贵建议，为我今后的学习和研究开拓思路。

最后，感谢我的父母这么多年来的培养，他们在各个方面都给予了我支持和鼓励，他们是我最大的支柱和动力。

攻读硕士期间参与的项目和公开发表的论文

参与的项目

[1] 2009.11 参与第五届全国高等学校计算机课件大赛中网站的制作,获得一等奖

[2] 2010.11 参与第十届全国多媒体课件大赛中网站的制作,获得二等奖

公开发表的论文

[1] 陈建华,王治和,蒋芸,许虎寅,樊东辉. 一种改进的文本分类特征选择算法[J].微电子学与计算机, 2011,28(12):180-183.

[2] 陈建华,王治和,蒋芸.基于类别区分度和关联性分析的综合特征选择[J],计算机工程(已录用), 2011.

中文文本分类特征选择方法研究

作者：陈建华
学位授予单位：西北师范大学