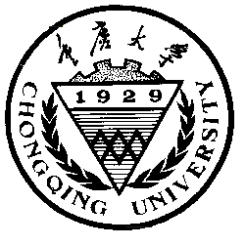
中文文本分类特征提取方法的研究与实现



重庆大学硕士学位论文

学生姓名：林少波

指导教师：杨　丹　教　授

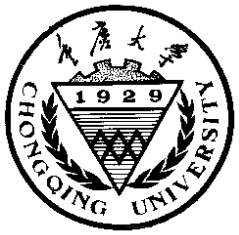
专　　业：计算机软件与理论

学科门类：工　学

重庆大学软件学院

二O一一年十一月

Research and Implementation of Feature Selection in Chinese Text Classification



A Thesis Submitted to Chongqing University

in Partial Fulfillment of the Requirement for the

Degree of Master of Science

By

Lin Shaobo

Supervised by Prof. Yang Dan

Major: Computer Software and Theory

College of Software Engineering of Chongqing University,

Chongqing, China

November 2011

摘　要

随着计算机网络技术的迅猛发展，文本信息数量呈现指数级的增长。文本分类作为一种有效的文本信息组织管理技术能很好地组织管理海量、异构的信息。在文本分类基础上通过信息检索、过滤等技术可以帮助人们从海量信息中快速，准确地查找相关知识信息，提高生活工作效率，因此对文本分类技术的研究具有较大的研究意义和实用价值。

本文首先对文本分类关键技术进行研究分析，在此基础上对特征提取方法进行了重点研究，提出了新特征提取方法，并利用新特征提取方法设计开发了一个中文文本分类系统，实验结果分析表明文中提出的特征提取方法取得了良好的实验效果。本文主要的研究工作如下：

① 分析了对文本分类过程及关键技术，研究了文本特征特征提取方法。通过对基于过滤模型的几种常用特征提取方法分析比较后，发现文本特征提取过程中负相关特征与弱相关特征对特征提取质量好坏易产生较大的干扰。为了避免这种干扰，本文提出一个基于类别正相关和类别强相关的特征提取方法 SP (Strong Correlation and Positive Correlation，正相关与强相关)，S P 方法通过优先选择正相关特征和强相关特征，有效地减少了负相关特征和弱相关特征的干扰，从而保证高质量文本特征的提取。

② 设计与实现了一个中文文本分类系统，把文本分类的特征提取方法 SP 应用到中文文本分类系统。文中对中文文本分类系统进行了总体设计和功能模块设计，分析研究汉语语法分析工具包 ICTCLAS 与全文检索工具包 Lucene，并将二者结合作为中文文本分类系统搭建解决方案，最终实现了中文文本分类系统。

③ 在中文文本分类系统上对特征提取方法进行大量的实验。把本文提出的新特征提取方法与常用的 DF、CHI、CC 等特征提取方法进行对比实验，利用多项常用的分类效果评价指标对实验结果进行综合性评价分析。实验结果表明 SP 方法通过提取高质量的特征词，构造低维的特征向量，能够有效地降低特征空间维度，在中文文本分类中表现出良好的特征提取效果，反映了类别间的差异度。

关键词：文本分类，特征降维，特征提取，类别正相关性，类别强相关度

ABSTRACT

With the development of society, especially the rapid development of network technology, various types of information get an exponential growth. Text classification can manage huge and heterogeneous data effectively. Information retrieval and filtering, which based on text classification, helps people get the required information in the huge data and helps people work more effectively. Text classification techniques have become popular and significant research topic.

This thesis does the detailed study and analysis on key techniques of text classification firstly, then focuses on the study of feature selection and proposes a new feature selection method. Finally, we design and realize the TC system by new method.

① Do analysis on the process and key techniques of TC, and do study on text feature selection methods. We find that negative feature and poor correlation feature effect the quality of selected feature by comparing several common methods which based filter model. Feature selection, this paper proposes a new approach of feature selection for TC, which is based on the strong class correlation and positive class correlation, named SP. SP can eliminate the effect of negative feature and poor correlation feature effectively by selecting positive and strong features, and then get high quality features.

② SP has been applied in designing and realizing the Chinese text classification system (CTCS), we do the overall design of CTCS and detailed design of modules of CTCS. This paper study on Chinese grammar analysis tool package ICTCLAS and Full-text search package Lucene, and then combines ICTCLAS and Lucene to be a solution of realizing CTCS, finally realize the CTCS.

③ We do many comparison experiments on new feature selection method SP and common method, such as DF, CHI.etc. This paper evaluates the result of classification by several classification performance evaluations. The result of experiments indicates the new feature selection method SP can select quality features, construct low-dimensional feature vector and reduce the dimensionality of feature space. SP has a good performance on feature selection in Chinese text classification, reflecting the degree of difference among classes.

Keywords: Text Classification, Feature Dimensionality Reduction, Feature Selection, Class Positive Correlation, Class Strong Correlation

目 录

1 绪 论

本章首先阐述了文本分类的研究背景、研究意义，以及当前国内外的研究状况和研究热点，然后介绍了本文的研究目的、研究内容和论文的组织结构。

1.1 研究背景及意义

随着社会的发展，我们处于一个信息爆炸的时代，各类的文本信息数量呈指数级的增长。期刊出版方面，全世界每年出版的期刊有数十万种，并且出版量每年以上万级别递增。在图书出版方面以中国为例，2004 年全国共出版图书 208294 种，总印数 64.13 亿册，并且出版量每年以数万种递增。图书馆藏书方面，2008 年中国最大图书馆北京图书馆藏书达到 2000 多万册，平均每天接受的新书达到上万种。特别是在网络方面，随着网络技术的发展，Internet 已经发展成为一个巨大的开放式、分布式的全球化信息空间，网络上的信息内容分门别类，各式各样，覆盖面极其广，包括人类社会活动中的生活，时事，经济，学习，工作各个发面的信息。当前 Internet 上的信息量呈亿级指数增长，惊人的信息增长速度使得人们对于海量信息的管理力不从心，海量、异构的信息中隐藏着对人类非常有帮助的知识信息，如何更好地组织管理海量、异构的信息，并且从中快速，准确的查找自己所需的相关知识信息，帮助人们提高生活工作效率，已经成为当今热点并且具有重大意义的研究课题。

文本自动分类是在给定分类体系的情况下，根据文本的内容自动将其分到一个或多个预定义类别。文本分类初期依靠人工分类方法实现，该方法不仅耗费大量人类，并且经常导致分类结果不一致，效率低下，人工分类方法已经不适应当今时代的需求。因此便产生可大量丰富的文本信息无法得到利用，而且海量的文本信息包括了结构化、非结构化、半结构化的文本信息，只有通过对文本信息组织管理，才能快速准确获得有效的相关信息，因此文本自动分类技术应运而生。文本自动分类是大规模文本信息处理的有效方法之一， 文本自动分类技术满足了人们对于信息的查准率、查全率等全面需求。

文本自动分类是一种有效的文本信息组织管理技术，以文本自动分类为基础可以实现许多与文本相关的具体应用，如信息检索，信息过滤、数字图书馆、文本数据库，因此有着广泛的应用前景和商业价值。

① 信息过滤

网络技术的迅猛发展使得网上在线文本资源数量十分庞大，并且这些文本资源具有高相似性等特征，因而对人们获取需要的相关领域的信息造成麻烦。信息

过滤技术通过对网上文本信息的筛选过滤获取自己需要的信息，屏蔽自己不感兴趣的信息，解决了获取有效信息的难题。基于以上描述我们可知信息过滤是二类的分类问题，将用户感兴趣的文本信息筛选出来主动推送给用户，将户不感兴趣的信息过滤掉，达到屏蔽无用信息的目的。

② 信息检索

信息检索将把大量的文本信息根据主题进行层次化分类，有效地对大量文本信息进行组织管理，简化了信息检索过程。如果按照类别，运用文本自动分类技术对信息检索结果进行分类区分，能够提高检索的查准率。目前很多 Web 搜索引擎站点都使用了 WEB 文档层次化分类组织。

③ 数字图书馆

在网络技术迅猛发展的过程中，数字图书馆应运而生，对图书进行数字化、快速有效的管理。在对图书馆中图书进行分类时，图书管理员无法准确掌握各个图书类别信息，又由于人工分类容易产生分类结果不一致性，所以文本自动分类技术便被应用于数字图书馆，进行图书分类管理，能够使得图书被客观准确地分类。

④ 文本数据库

伴随着文本信息量的迅速增长，组织、存储、查询文本信息已经无法满足文本数据库的管理。文本分类数据库管理需要多层次的服务支持，如数据挖掘等。而文本自动分类是文本数据挖掘的基石，是文本信息本身组织管理的有效手段之一，也是数据挖掘研究的重要支撑技术。

因此，对文本分类的研究具有极其重要的理论意义和广阔的应用前景，能够创造巨大社会效益以及商业价值。

1.2 国内外研究现状

Sebastiani 在文献[1]中在对文本分类发展历程的总结中指出国外的文本分类技术研究经历了四个发展阶段：第一阶段，在 1958 年至 1964 年之间主要进行文本分类技术的可行性研究；第二阶段，在 1965 年至 1974 年之间主要进行文本分类的实验研究；第三阶段，在 1975 年至 1989 年之间文本分类技术进入文本分类实际应用阶段；第四阶段，1990 年至今文本分类技术进入基于因特网的自动文本分类研究阶段。

文本自动分类的研究始于 20 世纪 50 年代末，H.P.Luhn 率先结合词频统计的思想进行文本分类，在这一领域进行了开创性的研究[2]。Maroon 于 1961 年发表了有关文本自动分类的第一篇论文[3]，对文本分类领域产生了深远的影响，至此以后许多学者如 Sparck，Salton 等都在文本分类领域进行了一系列卓有成效的研究工作

[4]。空间向量模型(Vector Space Model，VSM)于 1970 年由 Salon 等人提出[5]，该模型在良好的统计学方法基础上简明地对文本数据相关特性的抽象描述，成为文本分类领域的一种经典模型，至今仍被广泛应用。20 世纪 80 年代末基于知识工程技术构建的文本分类系统是非常流行的一种文本分类方法[6]。基于知识工程技术的文本分类方法是指根据领域专家获得的知识人工指定分类的规则。这种文本分类方法的缺点在于如何将获取知识转化为分类规则。需要知识工程师和领域专家对领域知识进行良好地沟通理解，否则知识工程师对知识的理解容易出现偏差。并且基于知识工程技术的文本分类技术适应性差，由于是面向专门领域定制，故无法适用于其他领域，需重新构建文本分类系统。在 20 世界 90 年代初期由于应用的需要，以及计算机硬件性能的提升，文本分类成为信息系统学科的一个主要研究分支，在这期间基于机器学习的文本分类方法[7]成为了研究的热点，基于机器学习的文本分类方法排除了人为对分类过程产生影响的因素。该方法注重了分类器的模型自动挖掘和生成及动态优化能力，在分类效果和灵活性上较之前的基于知识工程的文本分类方法有很大的提高，因此基于机器学习的文本分类方法成功替代了基于知识工程的文本分类方法成为文本分类领域研究和应用的经典范例。现今将语义分析方法与机器学习方法相结合进行文本分类处理成为了文本分类领域的研究热点之一。

现今，国外的文本分类技术在文本组织、电子会议、信息检索、邮件过滤等方面得到了广泛的应用[8]，文本分类技术已经进入了进入文本分类实际应用阶段阶段。其中较为成功的应用系统有：麻省理工学院（MIT）为白宫开发的邮件分类系统，Oracle 和 IBM 公司开发的电子邮件自动分类系统，卡内基集团为路透社开发的 Construe 系统[9]，美国 Carnegie Melton 大学开发的在线文本分类系统，美国 Just Research 公司开发的多类别文本分类系统。

中国对文本分类的研究工作大致是从 20 世纪 80 年代开始的，中国的文本工作大致上经历了可行性探讨、辅助分类系统和自动分类系统这三个阶段。南京林业大学侯汉清先生是国内对自动分类进行探讨的第一人，1981 年，侯汉清先生对自动分类进行探讨并从计算机分类检索、计算机管理分类表，计算机自动分类等几个方面对国外的研究现状进行了介绍[10]。自此掀开了国内对文本分类领域的研究热潮，中国科学院，以及以清华大学，复旦大学为首的高校在文本分类领域做了大量研究工作，并且取得了成果，开发了一系列基于知识工程技术和词典方法的文本分类系统[11]。大体上来说，中文文本分类技术还处于试验研究阶段，正在逐渐朝应用方向发展靠拢，学者们在充分认识到中文与英语的语言本质差异性，不能生搬硬套国外的研究成果。因而学者们在借鉴国外文本分类的研究策略之上，充分考虑了中文文本本身的语言特点，成功构建了中文文本分类体系。随着计算

机硬件技术，机器学习以及自然语言处理技术的快速发展，国内学者们已经研发了一系列的自动分类系统。例如，中科院研发的中文文本智多星分类系统、东北大学图书馆研发的图书分类专家系统、山西大学刘正瑛等人研发的金融自动分类系统、清华大学吴军研发的自动分类系统。文本分类技术现今急需在合理应用的实践过程逐步改善算法进而提高文本分类性能

虽然国内外研究者在文本分类领域理论研究取得一系列成果，并在实际应用中有不错的表现，但是文本分类技术在实际应用中仍然存在一系列亟待解决的问题，例如对海量文本集合如何进行快速并且准确分类的问题，如何让分类器解决语言兼容性的问题，如何在噪声环境中去除噪声达到优化分类性能目的等一系列问题。另外在特征空间降维，文本表示以及利用语义信息对文本内容进行分析进而增强分类算法性能等方面，需要做大量的研究工作。

1.3 本文主要研究内容

本文的主要研究内容包括以下三个方面：

① 本文通过对文本分类相关技术的研究分析，在此过程中重点研究分析了特征提取关键技术。通过对现有特征提取技术进行对比分析，结合类别正负相关性和类别强弱相关度的思想，提出了一种基于类别相关的新特征提取方法。

② 开发基于开源工具包 Lucene[12]与 ICTCLAS[13]的中文文本分类系统，支撑后续实验环节的顺利进行。首先研究分析系统需求，接着以需求为出发点，通过研究 Lucene 和 ICTCLAS 寻求满足系统需求的契合点，在此基础上制定需求关键功能解决方案，然后对系统进行模块设计，最后实现系统。

③ 在中文文本分类系统上对本文提出的新特征提取方法与常用的特征提取方法进行多组对比实验。利用多项常用的分类效果评价指标对多组对比实验结果进行综合性评价分析，验证本文提出方法的可行性和有效性。

1.4 本文的组织结构

第一章首先对文本分类的相关研究背景及意义进行介绍，接着对国内外文本分类技术的研究现状进行分析，最后明确指出本文主要研究内容及组织结构。

第二章着重对文本分类关键技术进行介绍和分析。首先明确了文本分类的概念，接着简要地介绍了文本分类过程。最后结合文本分类过程详细介绍了文本预处理、文本表示模型、文本特征降维方法、文本特征加权、文本分类算法和分类器性能评价等一系列与文本分类相关的技术。

第三章通过对现有特征提取技术进行对比分析，结合类别正负相关性和类别强弱相关度的思想，提出了一种基于类别相关的新特征提取方法。

第四章通过中文文本分类系统的总体设计，模块设计，具体实现三个章节详细介绍了中文文本分类系统从设计到实现的过程。

第五章在中文文本分类系统上对本文提出的新特征提取方法与常用的特征提取方法进行对比实验。结合分类效果评价指标评价分析相应实验结果，得出结论。

第六章总结了本文所做的相关研究工作，并展望了未来的研究方向。

2 文本分类的相关技术

2.1 文本分类概述

文本分类是一个有监督的学习过程，文本自动分类是在给定分类体系的情况下，根据文本的内容自动将其分到一个或多个预定义类别。文本分类的过程实际上是一个映射的过程，即根据预定义的类别中的文本数据信息总结归纳出分类的规律性，根据这个规律性将未标明类别的文档映射到预定义的类别中，该映射可以是一对一映射，也可以是一对多映射[14]。文本分类是模式分类和自然语言处理的交叉学科，因此模式分类的算法能够应用到文本分类中，而且它与文本语言关系密切，在这点上与普通模式分类有所区别。

文本分类过程总的分为训练过程阶段和分类过程阶段。根据文本分类过程中涉及到的相关技术环节可将文本分类过程细分为：文本预处理、文本表示、文本特征降维、文本特征加权、分类方法的选择和分类性能评价五个部分。文本分类流程如图 2.1：

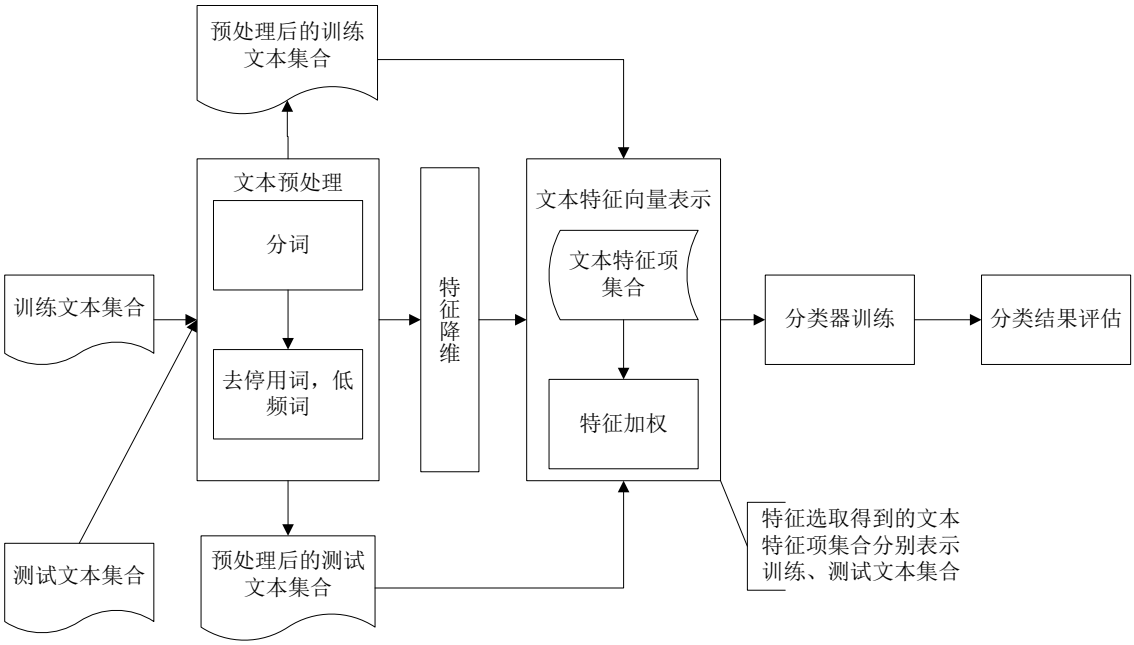


图 2.1 文本自动分类过程

Fig.2.1 The Progress of Text Classification

2.2 文本预处理

文本预处理将不规范的文本信息转换成能够让计算机处理的文字信息，将非结构化的文本转换成能够让计算机处理的结构化文本。文本预处理过程的工作主要

是统一规范文档格式，以方便于后续工作。文本预处理起到过滤噪声数据的效果，其结果将直接影响文本分类的性能，是文本分类中关键技术之一。

本文研究对象是中文文本集合，中文文本预处理过程主要包括中文分词和去停用词。中文分词处理是将一段中文的字序列切分成词序列的过程。中文分词与英文分词不同，英文文本中以单词为基本单位，英文文本中将空格作为自然分界符号，词与词之间以空格隔开。中文文本中以字为基本单位，将标点符号与换行符号作为自然分界符号，句子与句子之间以标点符号隔开，段落与段落之间以换行符隔开，但是词语之间没有一个形式化的分解符。因此对中文分词可以看成是将一段连续的中文字序列按照一定规则重新组合成词序列的过程。

中文文本中通过名词、动词等实词来具体表示一篇文章的内容信息，对文本分类意义重大。同时中文文本中也包括了频繁出现的“我”，“的”，“了”，副词，语气词等虚词，此类词对文本分类毫无贡献，不能反映文本主题，这类词的存在只会对文本分类产生干扰，应该予以过滤。这类词在自然语言处理领域被称为停用词。在文本分类处理过程中，我们一般通过对文本分词，过滤停用词以及虚词对文本集合进行初步降维，只保留对文本分类有贡献的词语。

2.2.1 中文分词

中文分词是中文信息处理的基础，已经被广泛应用于中文信息领域的信息检索、机器翻译、自动摘要、汉字智能输入、中文校对、语音合成、汉字简繁体转换等技术中[15-17]。

① 中文分词的难点

中文是一种十分复杂的语言，人理解尚且有困难，更不用说让计算机理解中文。中文分词面临两大难题：歧义识别、未登录词识别。

1）歧义识别

中文分词过程中遇到的歧义字段有三种类型：组合型歧义、交叉型歧义和真歧义。

组合型歧义是指存在字串 A、B、C，词表 W，在字段 AB 中存在 $AB \in W$，$A \in W$，$B \in W$。例如：组合型歧义字段“将来”可切分成“我/将来/要/当/科学家”和“我/将/去/成都/工作”。虽然在组合型歧义字段在中文中所占数量较少，但是由于中文语言本身的复杂性，使得对组合型歧义的处理相对比较困难。

交叉型歧义是指存在字串 A、B、C，词表 W，在字段 ABC 中 $AB \in W$ 并且 $BC \in W$。例如：交叉型歧义字段“为人民”可切分成为“为人/民”和“为/人民”。虽然交叉型歧义字段在中文中所占的数量比较大，但其处理方法相对简单，易于实现。

真歧义是指如果某一歧义字段，不结合其它信息，人也无法判断出其正确的切分方式，那么这样的歧义字段就叫做真歧义字段。例如：真歧义字段“羽毛球拍

卖完了”可切分成为“羽毛/球拍/卖/完/了”和“羽毛球/拍卖/完/了”。像这种由中文本身的二义性引起的真歧义，如果不知道上下文信息，无法做出正确的切分。

2）未登录词识别

未登录词指的是未包含在词表中但必须切分出来的新词。在现今社会中，随着 Internet 的高速度发展，语言在不断的发展和变化，随之而来的就是大量新词的出现，所以任何词表无法包含所有的词语。

未登录词主要包括人名、地名、机构名、产品名、简称等。以人名为例，未登录的人名很容易与常规词形成交叉型歧义。例如对 “李紫没喝一瓶酒”这句话进行切分，可能切分出“李紫没/喝/一瓶/酒/”和“李紫/没/喝/一瓶/酒/”。未登录词的识别对中文分词有着重要的作用，对未登录词的识别直接影响到最后词频的统计，如果不能很好识别未登录，则最后词频的统计结果会产生较大的偏差。目前未登录识别的准确率已经作为评价分词系统一项重要指标。

② 中文分词算法简介

随着专家学者们在中文分词领域中的研究的开展和深入，众多中文分词算法被提出来，总的来说中文分词算法可以分为三类：基于词典的分词算法、基于理解的分词算法和基于统计的分词算法。这三类方法也是现今中文分词领域研究的三个主要方向。

1）基于词典的分词算法

基于词典的算法又被称为机械匹配算法，主要思想是基于字符串匹配的机械分词，即根据一定的扫描顺序和匹配原则将待分词的中文字串与一个由人工维护的庞大分词词典中的词语进行匹配，若在词典中找到相应的字符串，则成功切分一个词。基于词典的算法由扫描方向、字符串匹配原则和分词词典三大部分构成。按照文本扫描方向的不同可对字符串进行正向匹配、反向匹配和双向匹配。按照字符串匹配原则的不同可对字符串进行最大匹配、最小匹配、逐词匹配和最佳匹配。分词词典是指人工事先建立好的分词词典和分词规则，并且需要人工经常维护更新，根据应用领域的不同，各个领域的应用会维护自己的专业词典。在实际应用中我们通常将不同的匹配方法结合使用，由于汉语单字成词的语言特点，所以基于最小匹配原则的分词方法一般很少使用。故常用的基于词典的算法有正向最大匹配方法和反向最大匹配方法。

a. 正向最大匹配方法

正向最大匹配法是一种最基本的机械匹配的分词方法，以“长词优先”为原则。正向最大匹配法的分词过程可以描述为：读入待切分语句，去除标点符号，将语句分成若干待切分段，若词典中最长词的单字个数为 N，则对待切分语句段以首字为开始从左到右选取长度为 N 的匹配字段与字典中的词进行匹配，若匹配成功，

则该匹配字段作为一个词被切分出来，若匹配失败，则将该匹配字段的最后一个字去掉，剩下的字符串重新组成新的匹配字段再进行匹配，一直匹配失败则重复上述过程直到成功切分出一个词。对正向最大匹配过程循环执行，直到切分出语句中所有词。

b. 反向最大匹配法

反向最大匹配方法与正向最大匹配方法相似，不同之处在于匹配方向相反，反向最大匹配方法的分词过程可以描述为：读入待切分语句，去除标点符号，将语句分成若干待切分段，若词典中最长词的单字个数为 N，则对待切分语句段以段位字为开始从右到左选取长度为 N 的匹配字段与字典中的词进行匹配，若匹配成功，则该匹配字段作为一个词被切分出来，若匹配失败，则将该匹配字段的最前一个字去掉，剩下的字符串重新组成新的匹配字段再进行匹配，一直匹配失败则重复上述过程直到成功切分出一个词。对正向最大匹配过程循环执行，直到切分出语句中所有词。

基于词典的分词算法的思想原理简单易于实现，无需考虑相关的语义信息，关键在于维护一个包含足够大的词典，基于词典的分词算法的不足在于匹配长度较难选择，若匹配长度过短，语句的切分完整性和准确性无法保证，若匹配长度过长，将会增加算法的运算复杂度，降低了分词的效率。由于缺乏统一标准的词集支持，因此基于词典的分词方法存在着交叉型歧义和组合型歧义的问题。

2）基于理解的分词算法

基于理解的分词算法是通过句法、语义分析，让计算机获取关于句法、语义信息，从而通过人工智能方式模拟人对句子的理解。在分词的过程中，利用句法、语义信息来处理词语歧义的情况。基于理解的分词算法包含三个主要部分：分词子系统、句法语义子系统和总控部分。

通过总控部分的协调处理控制，分词子系统在进行歧义情况处理时可以得到从句法语义子系统输出的有关句法，语法以及语义信息，从而准确的进行分词。这个过程完成了计算机模拟人理解句子语义的过程。

基于理解的分词算法需要使用大量的句法、语义信息。由于中文知识的复杂性，难以将中文知识转化为计算机可理解的形式，因此目前基于理解的分词系统还处在实验阶段。

3）基于统计的分词算法

基于统计的分词算法以语言事实为依据，将词是稳定的字的组合作为前提，从概率论的角度出发，认为在上下文中，相邻字的联合概率越高，即字与字相共现的频度越高，则越可能组成一个词。

基于统计的分词算法通过建立数学统计模型，以字与字相邻共现的概率作为

组成词的可信度评价依据，对语料中相邻共现的各个字的组合的频度进行统计，计算它们的相关度。常见基于统计的分词算法有互信息统计算法，N-Gram 算法，隐马尔科夫算法。其中最常用基于统计的分词算法是互信息统计算法[18]。

互信息是信息论中作为衡量两个信号关联程度的一种尺度，后来引申为描述两个随机变量间关联程度的度量。我们利用互信息公式计算字符 a 和字符 b 的互信息。通过互信息来描述字符 a 和字符 b 之间的关联程度。互信息计算公式如公式 1 所示。

a. 若 MI(a,b)>0，则表示字符 a，b 之间具有可信的关联性，随着互信息值的增大，字符 a，b 的关联性也将增强。

b. 若 MI(a,b)=0，则表示字符 a，b 之间关联性具有不确定性。

c. 若 MI(a,b)<0，则表示字符 a，b 之间几乎无关联性，随着互信息值的减小，字符 a，b 的关联性也将减弱。

$$MI(a,b)=\log\frac{p(a,b)}{p(a)p(b)} \quad (1)$$

其中 p(a)表示字符 a 出现的概率，p(b)表示字符 b 出现的概率。p(a,b)表示字符 a 和字符 b 同时出现的联合概率。MI(a,b)表示字符 a 与字符 b 的互信息值。

基于统计的分词算法也叫无词典分词算法，无须维护庞大的分词字典是该算法的一个优点。基于统计的分词算法基于语言事实，通过对真实语料库中各个文本的字符串统计，能够客观的反应各个字符在语料库中的分布规律。算法简单明了，因而基于统计的分词方法具有较好的实用性。

基于词典的分词算法、基于理解的分词算法和基于统计的分词算法三种方法各有优缺点，并不存在哪个方法更优更有效的说法。现今成熟的分词系统，通过应用的领域情况，结合实际需求，通过综合各种分词算法，起到发挥各个算法长处，互相弥补的作用。通过上述手段更好的适应和满足实际生活中的应用需求。

2.2.2 停用词删除

停用词（StopWords）指的是虽然在文本集合中出现频度很高，但是对分类毫无贡献，存在只会增大特征空间维度，增加分类运算复杂度的无用词。如语气词、副词、连词、介词等虚词。在文本分类之前，需要引入停用词表来过滤掉停用词。

停用词的建立方式可以分为人工建立和基于概率统计的自动建立停用词表。人工建立停用词表是根据语言学专家的主观判断选择某些词集或是对特定的某一应用领域选择特定的词来构成停用词表；英文停用词表，比较著名的是 Van Rijsbergen[19]发表的停用词表以及 Brown Corpus 停用词表[20]。关于中文停用词的研究，虽然当前已有一些较好的停用词表，但其构造与选取语料相关，针对不同应用很难直接应用，目前可以查到的中文停用词表正在不断完善和扩充中。基于概

率统计的自动建立是基于词频信息构建停用词表，或者从初步的分词结果中得到部分停用词，然后在之后的分词过程中不断更新并根据切分结果进行验证。基于概率统计的自动建立主要通过采取熵、联合熵和基于 TF/IDF 词语 KL 分布的重采样技术自动获取停用词表。

2.2.3 词条选择

在向量空间模型中，空间向量由文本的特征项构成，未做特征选择之前文本集合中所有词条都可以作为表示文本的特征项。由于文本集合中的词汇量庞大，导致文本特征空间经常高达几万维，甚至更高。不能体现文本内容信息的虚词对文本分类无贡献，若作为文本特征用于文本分类将产生负面影响，因此应被视为是噪声数据。

词条选择，对文本特征空间进行压缩，能够减少存储空间，提高文本分类准确度，降低运算复杂度，提高程序效率。所以在预处理中进行词条选择作为特征空间初降维最有必要性。所以我们通过选择能表示文本内容信息的名词、动词，剔除一些对类别没有贡献或者贡献很小的虚词。因此对文本词条的过滤，可看作特征空间的初步降维，具有必要性。

我们通过对文本集合中的文本进行分词并加上词性标注，并且运用正则表达式(Regular Expression)进行匹配选择，将句子中能代表文本内容的实词如名词、动词的等选择出来。

2.3 文本表示模型

文本可以看出是一个数量庞大的字符串，并且由于文本结构的复杂多样性，在实际应用文本分类技术的过程中，需要从文本中提取出能代表文本内容的特征，并且将其转化为计算机能够处理的结构化数据形式，从而使得提取的文本特征能够直接应用于文本分类算法，使得计算机能够高效的处理文本信息。在提取文本特征的时候在减少数据处理量的同时要尽可能的保留文本中的语义信息。现有文本分类技术通常有两种文本表示模型[21]，即布尔模型和向量空间模型。

① 布尔模型

布尔模型(Boolean Model)是一种基于集合理论和布尔代数理论的分类模型，布尔模型可以看作是向量模型的一种特例，根据特征项在文档中出现与否，特征项值只能取 1 或 0。若某一特征在文本当中出现，则该特征在当前文本当中的值取 1，否则，该特征在当前文本中的值取 0。布尔模型不能很好体现文本特征的重要程度，通常情况下布尔模型的效果不如其他文本表示模型。但某些情况下，使用布尔模型表示文本进行分类所得到的效果并不比以其他文本表示模型差。基于布尔模型的常用文本分类方法包括关联规则方法、决策树方法和 Boosting 方法。

② 向量空间模型

向量空间模型（Vector Space Model,VSM）由 Gerard Salton 和 McGill 于 1969 年提出，向量空间模型具有简单、高效的特点，是信息检索领域最经典的文本表示模型。在实际应用中，向量空间模型成功地运用于著名的 SMART 系统中。现在多数文本分类方法都采用向量空间模型。

向量空间模型利用文本特征向量表示文本，在该模型中，用文本集合中的词条作为表示文本的特征项,所有特征项构成表示文本的特征向量，文本空间被视为一组词条向量所张成的向量空间。假设一个文档集合 D 包括 n 个文档，文档集合表示为$D=\{d_1,d_2,d_3,.....,d_n\}$，文本集合中所有词条表示为$T=(t_1,t_2,t_3,...,t_k)$，$t_n$ 表示为文本集合中的特征词条，每个文档特征项对应一个权值，每个文档中特征项分布的差异性，使得特征项的权值不同，则每个文本都可以由特征向量$W_{d_n}=\{w_{t_1},w_{t_2},w_{t_3},..w_{t_k}$ 唯一表示，W_{d_n} 表示文本 d_n 的特征向量，w_{t_k} 表示特征词 t_k 在文本 d_n 中对应的权值。向量空间模型的构造过程如图 2.2 所示：

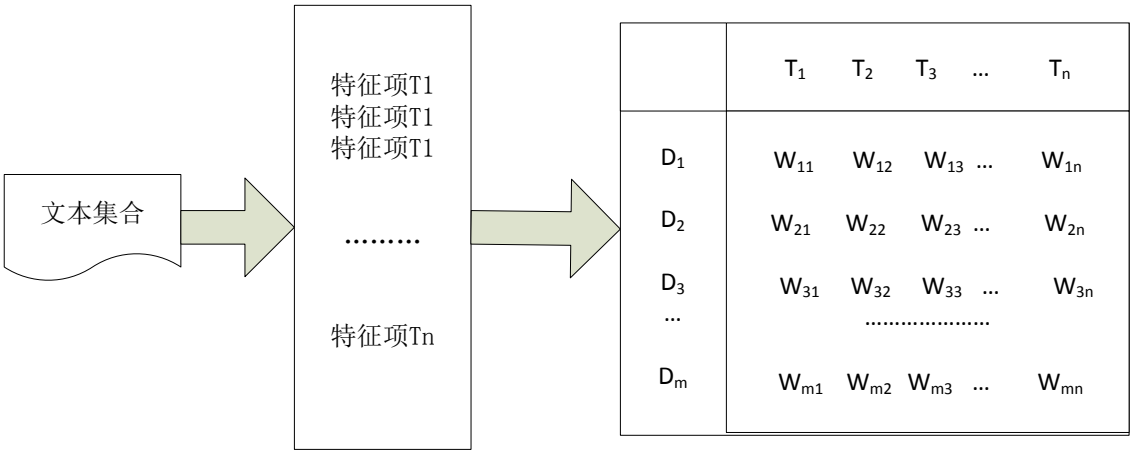


图 2.2 向量空间模型构造过程

Fig.2.2 The Process of Creating Vector Space Model

向量空间模型在知识表示上的优越性，使其成为一种被广泛使用的文本表示模型，它将文本集合中的文档转化为空间向量，在向量空间中，每个文本以空间中的点的形式存在，简化了文本表示形式，方便计算各个文本间的关系，如相似度等。由于布尔模型不能很好体现文本特征的重要程度，本文亦将向量空间模型作为文本表示模型，向量空间模型的产生提高了自然语言文本的可操作性和运算能力，从而为文本处理领域提供了良好理论基础。

向量空间模型是不考虑特征词出现的位置、次序以及特征词在文本中的上下文关系的词袋（Bag of Words）文本表示模型。在向量空间模型中文本被视为一系

列相互独立的特征词的集合，仅将文本中单词出现的频率作为文本分类的唯一依靠的信息。向量空间在简化文本表示操作的同时丢失了特征词上下文信息、文本结构信息和语义信息，而这些信息对自然语言处理往往有重要的价值。

向量空间模型是建立在文本所有的特征是相互正交的这一假设的基础之上，并没有充分考虑文本特征项之间的关联关系。由于自然语言本身具有丰富且复杂的特点，故向量空间模型基于的假设有悖与自然语言存在的客观事实。目前已经有很多改进的文本表示模型提出，但是效果并不好，由于自然语言语义的复杂性，如何寻求适合的数学模型来反映自然语言语义是众多研究者一直努力的方向。

2.4 文本特征降维

文本特征降维是文本分类过程中最核心的环节。文本特征降维技术一般可以按选取集合的不同分为:特征析取和特征提取。在上面已经提到，我们使用向量空间模型来表示文本，在实际应用中，文本集合中包含的词条高达数万，甚至更高，这就造成了特征空间的“高维性”。若使用如此高维的向量表示一个文本，绝大多数特征项都不会出现在当前表示的文本中，这样就造成了特征空间的“高稀疏性”。特征空间的“高维性”和“高稀疏性”导致了很多分类算法无法应用，即使能够应用，也严重影响了分类的准确度和时间。故文本降维技术应运而生，通过对特征空间的降维，不仅提高了分类速度，而且过滤了噪声数据，提高精度的同时还有助于解决过拟合问题[22]（基于样本文本集合中的训练集合的分类器，对样本集合中的测试集合的分类效果好，但是测试集合换成其他文本集合分类效果差）。

2.4.1 文本特征析取

特征析取也称为特征重构，特征析取是依据一定的原则将原始特征空间映射到一个新的低维特征空间，在低维特征空间中各维度特征项相互独立，集中体现了原始空间中对分类有帮助的特征信息。因此提取出来的特征集合 T′是对原始特征集合 T 采用合并、转化、归纳等手段重构获得，T′不是 T 的子集。文本特征析取中常用的方法有主成分分析[23]（Principal Component Analysis ， PCA ），潜在语义索引[24]（Latent Semantic Indexing，LSI）等。

2.4.2 文本特征提取

特征提取是根据一定的特征提取度量标准从测试集合的初始特征集合中提取出相关的原初始特征子集达到降低特征向量空间维度的目的。在特征提取的过程中不相关以及冗余的特征将会被删除。特征提取作为学习算法数据预处理方法，可以很好地提高学习算法的准确性，减少学习算法耗费的时间。可以得知若学习算法使用的特征充斥着不相关性，冗余性，以及干扰性，那么学习算法的结果必定很差。在实际应用中，特征提取如何得到一个最优的特征子集是一个 NP 难题。

特征提取可按照特征提取策略、特征降维模型加以区分。

① 按特征提取策略区分

按特征提取策略，可以将特征提取方法分为全局特征提取与局部特征提取。面对不同的数据集我们应该选择适合的特征提取策略。

所谓局部特征提取是依据给定的特征评价函数对每个类别中的存在的所有特征进行计算，分别统计每个类别特征空间中各个局部特征项的值，将各个特征项按其特征值排序，并根据设置的阈值选择出合适规模的类别特征子集，然后将各个类别的特征集合组合成一个新的特征空间，该特征空间是各个类别特征子集的集合。局部特征提取是依据给定的特征选择度量，对每个特定的类别分别选择一组最优特征，主要考虑这个类别内部的共性和特征的局域特性。

全局特征提取依据给定的特征评价函数对每个类别中的存在的所有特征进行计算，然后结合每个特征对应各个类别的值，通过全局值计算方法对每个特征进行唯一全局值的计算，将各个特征项按其全局特征值排序，并根据设置的阈值选择出合适规模的全局特征子集。全局特征提取是对分类涉及到的所有类别选择一个共同的最优特征集合，主要考虑类别之间的差异性。

文献[25]指出在均衡数据中全局特征选择优于局部特征选择，因为单独利用局部特征提取的缺点是局部特征提取不能有效地选取到能代表所有类别的特征集合，忽视了全体训练样本和类别的整体性。文献[26]指出在偏斜数据集中使用加权局部特征选择优于全局特征，因为偏斜数据中各个类别间文档数量上差距很多，基于全局选择的特征集合不能代表少数类，如果使用无加权特征选择从各个类别中选择相同数目的特征也必将影响特征选择效果。基于以上分析，由于我们实验中采用的是均衡数据集，故使用全局特征选择策略。

② 按特征降维模型区分

按特征降维模型，可将特征提取分为基于 Filter 模型、Wrapper 模型和混合式的特征提取方法。基于 Filter 模型的特征提取方法与基于 Wrapper 模型的特征提取方法的区别在于所采取的评价标准是否和具体的分类器有关。

1）基于 Filter 模型的特征提取方法

如图 2.3 所示基于 Filter 模型的特征提取方法的计算依赖于样本数据集合本身，采取一种和分类器无关的评价标准，该评价标准称为特征评价函数，通过特征选择方法对所有特征进行过滤，过滤掉冗余、无关的特征，保留与类别相关的特征，通常认为相关度较大的特征子集将有利于学习算法的准确率，故选择和目标函数相关度大的特征子集，使用特征选择方法选择出来的特征集合表示样本从而进行分类器的训练。

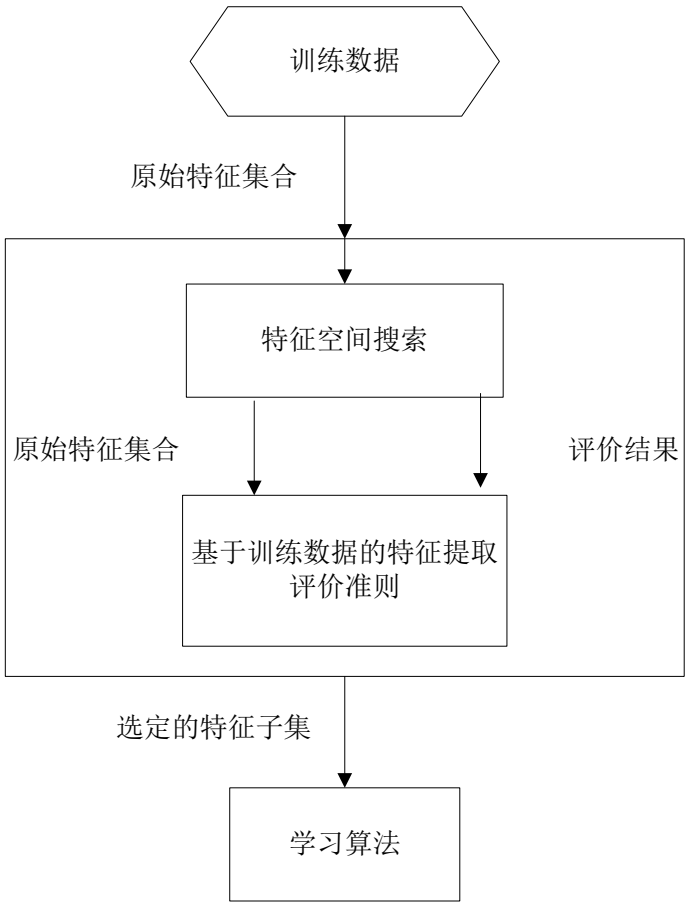


图 2.3 基于 Filter 模型的特征提取

Fig.2.3 The Filter Model of Feature Selection

根据特征评价函数度量特征优劣的不同，特征提取方法又可以分为：基于距离度量、相关性度量、信息量度量和一致性度量的特征提取方法。

基于距离度量的特征提取方法目的是要找到一个特征子集能够令各个类之间的距离最大，使得类别的可分性最大，分类的错误率最低。距离度量主要采用欧式距离，所有评价标准中距离度量所依据的理论基础最完善。

基于相关性度量的特征提取方法用统计学中的相关性刻画了两个随机变量之间的相关程度。若特征子集T_1和类别 C 之间的相关系数大于T_2和 C 之间的相关系数，则根据相关性准则特征子集T_1好于T_2。可以使用线性相关系数(correlation coefficient) 来衡量向量之间线性相关度。

基于信息量度量的特征提取方法的理论依据是信息论[27]。信息量度量关键在于评测类别从特征中获取的信息来源，如果类别 C 从特征子集T_1中获取的信息比特征子集T_2多，则可以认为特征子集T_1好于特征子集T_2。目前常用的信息度量有熵度量和互信息两大类。熵度量通过比较特征子集T_1与特征子集T_2的出现能够消除多少文本属于类别 C 的不确定为依据，特征子集T_1较特征子集T_2消除的不确定

性多，则特征子集T_1好于特征子集T_2，在接下来将要介绍的常用特征提取方法中信息增益 IG 正式利用了这一准则。互信息方法通过计算特征子集 T 和类别 C 之间的互信息量，作为对特征子集好坏的一个评价。从数学角度看来，若特征子集 T 包含类别 C 的统计互相信息量越大，基于特征子集 T 的分类器性能越好。由于互信息计算量交大，当实际应用中特征子集维数规模增多时，通常采用互信息量的一个近似值来代替。常用特征提取方法中互信息（MI）正是利用这一准则。

基于一致性度量的特征提取方法致力于找出能够与完整特征集分类效果一致的最小特征子集，同样一致性度量可以用不一致率来描述。不一致率即一个数据集合中“不一致”的样本占数据集样本总数的比例[28]。如果得到的不一致率为 0，则认为一致率为 100%。若用同一个特征向量对两个样本进行特征描述，对应两个样本，特征向量值一致，而所属两个样本所属的类别不一样，则称这两个样本是“不一致”的样本。例如样本 1 与样本 2 属于不同的分类，但在特征子集T_1和T_2上的取值完全一样，那么特征子集{T_1，T_2}不应该选作最终的特征集。由于不一致率低的特征子集包含更多能区别不同样本的特征项，故不一致率低的特征子集要好于不一致率高的特征子集。一致性度量只适用于离散型特征情况，对于连续型特征不适用。一致性度量有计算复杂性低、满足单调性的优点。

基于 Filter 模型的特征提取具有通用性强，算法复杂度低，高效率的特点，适用于大规模数据集的优点。文献[29]指出选择可最优化分类准确率的特征子集和寻找和类别相关的特征子集是不同的任务，故基于 Filter 模型的特征提取所得到的特征子集并非最优。

2）基于 Wrapper 模型的特征提取方法

如图 2.4 所示，基于 Wrapper 模型的特征提取是基于分类错误度度量的特征提取方法，该方法认为与学习算法无关的基于 Filter 模型的特征评价提取出来的特征会使得分类算法产生较大的偏差，而学习算法基于所选特征子集的性能好坏是更好的特征评价标准。不同学习算法偏好不同的特征子集，既然经过特征提取出来的特征子集将应用于之后将使用的学习算法上，那么应用特征子集后学习算法的性能好坏就是最好的特征评价标准。因此在基于 Wrapper 模型的特征提取中将使用当前特征子集后学习算法的性能好坏作为特征提取的评价标准。基于 Wrapper 模型的特征提取对于特定的分类器来说可找到最优的特征子集，算法准确率高。但是基于 Wrapper 模型的特征提取通用性差，对不同的学习算法需要重新进行特征提取。每次特征提取都需要进行学习算法计算，计算量大，对小规模测试数据集容易出现过拟合现象。

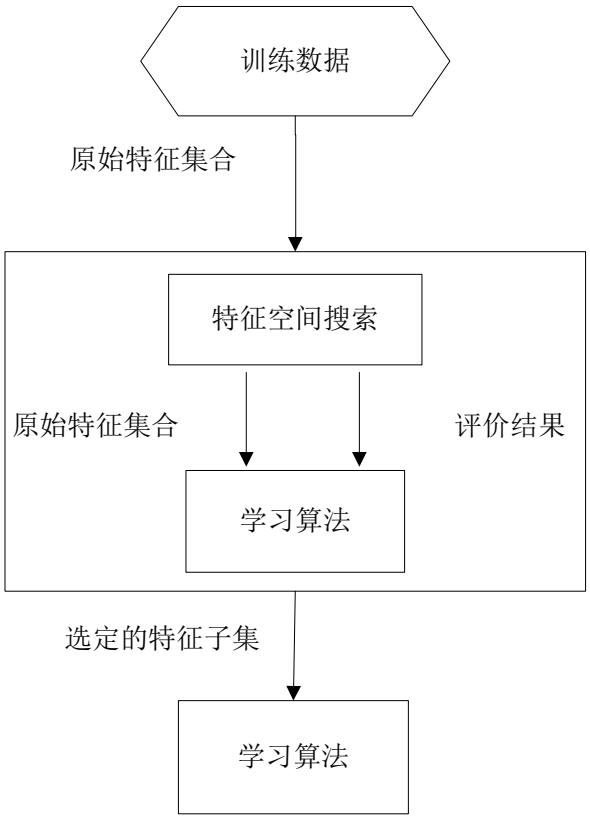


图 2.4 基于 Wrapper 模型的特征提取

Fig.2.4 The Wrapper Model of Feature Selection

3）基于 Filter 和 Wrapper 模型混合式的特征提取方法

基于 Filter 和 Wrapper 模型混合式的特征提取方法认为该方法中结合 Filter 和 Wrapper 模型是两种互补的模式，该特征提取方法的原理是：先基于 Filter 模型的特征提取方法快速过滤删除无关或冗余特征，得到经过初降维的中间特征子集，然后利用基于 Wrapper 模型的特征提取方法进一步细化。现已有一种综合 Filter 和 Wrapper 优点的组合式特征选择算法。如 Huang 提出的一种两阶段组合式特征选择算法，Das 提出的组合式算。

2.5 文本特征加权

文本特征加权是指对构成文本向量空间的特征集中的每一个特征根据某个的标准赋予相应权重的过程。特征词权重刻画该词特征词在描述此文档内容时所起作用的重要程度。特征词权重计算唯一的准则就是要最大限度的区分不同文档。最常见的文本特征加权方法是 TF-IDF[30]。

传统的特征权算法 TF-IDF 主要考虑特征项的频率信息以及反文档频率信息。如公式 2.1 所示，TF-IDF 的计算公式为：

$$tfidf(t_i,d_j)=tf(t_i,d_j)\log(\frac{N}{N(t_i)}) \quad (2.1)$$

在很多情况下还需要将向量归一化，如公式 2.2 所示，TF-IDF 的归一化计算公式为:

$$W_{ij}=\frac{tfidf(t_i,d_j)}{\sqrt{\sum_{i=1}^{n} fidf(t_i,d_j)^2}} \quad (2.2)$$

如果 $n(t_i,d_j)>0$，$tf(t_i,d_j)=1+\log n(t_i,d_j)$， 否则 $tf(t_i,d_j)=0$。

2.6 文本分类算法

文本分类算法是文本自动分类的核心部分，在研究和发展的过程中，涌现了很多经典的文本分类算法。目前经典的文本分类算法大部分都是基于机器学习方法，分类算法大致可以分为三类：1、基于统计的方法，如 K 近邻（KNN）[31]，朴素贝叶斯(Naive Bayes)[32]，支持向量机（SVM）[33]；2、基于规则的方法，如决策树（Decision Tree），关联规则；3、基于连接的方法，如人工神经网络（ANN）。接下来我们将主要介绍基于统计的方法。

2.6.1 K 近邻

K 近邻分类算法是著名的模式识别系统学方法之一，是一个理论上成熟的分类算法。K 近邻分类算法是基于实例的文本分类方法，具有简单、稳定、有效的特点，因而在文本分类领域广泛地使用。

K 近邻分类算法的原理是通过找到与待分类文本相似度最接近的 k 个训练集文本，然后统计 k 个训练集文本的类别归属情况，若 k 个训练集文本中，属于类别 C 的文本占多数，则待分类文本属于类别 C。文本相似度的计算通常采用欧式距离或余弦相似度公式。在 K 近邻分类算法中 k 值由人工指定，k 值是基于经验的参数。由于在实际应用中，待分类文本需要计算它与所有文本间的相似度，然后将它与所有文本间的相似度进行排序，取出前 k 个与它最相似的文本，所以 k 的大小不会影响到算法的计算复杂度和时间。为了更好地说明问题，我们将列出 K 近邻分类算法的具体步骤。

算法步骤如下：

1）计算特征选择环节获得的文本特征词集合各个特征词对应每个经过预处理的训练集文本的权重，然后将经过预处理的训练集文本用特征向量表示。

2）计算特征选择环节获得的文本特征词集合各个特征词对应经过预处理的待分类文本的权重，然后将经过预处理的待分类文本用特征向量表示。

3）通过相似度计算得出与待分类文本最相似的前 k 个训练集文本。

4）统计 k 个训练集文本的类别归属情况。

5）将待分类文本分到 k 个训练集文本归属类别占多数的类别中。

K 近邻算法是一种惰性学习算法，在训练阶段 K 近邻算法的工作只是将训练集文本表示成特征向量形式。在进行分类时，需要经过计算相似度和排序两个环节，在分类时，计算时间复杂度高，若训练集中文本总数为 n，则 K 近邻算法的分类时间复杂度为 O（n），可见 K 近邻算法的分类时间复杂度与训练集中文本总数成正比。在进行大规模的文本分类计算时，该算法运行期间开销大，无法满足实时性要求。因此 K 近邻算法适用于小规模训练集合的文本分类。

2.6.2 朴素贝叶斯

朴素贝叶斯分类算法是一种基于概率统计的文本方法算法，是一种简单有效的文本分类算法。朴素贝叶斯算法以贝叶斯定理为理论基础，该算法成立的前提是假设表示文本各个属性（即文本的特征向量中的各分量）之间相互独立，不相关。通过已知的先验概率和条件概率来计算一个未知类别的文本属于各个类别的概率，则未知类别文本属于概率最高的类别。该算法可应用到大规模文本集合中，具有方法简单、分类准确率高、速度快的优点。但是由于朴素贝叶斯分类算法基于的假设太过严格，使得该分类算法的应用范围受到限制。在现实应用中，若不能满足独立性假设，则该算法的分类准确率会有一定程度的下降。

假设训练集中存在 k 个类别，类别集合表示为$C=(c_1,c_2,c_3,....,c_k)$，文本特征词集合表示为$T=(t_1,t_2,t_3,...,t_k)$，各个文本特征对给定类别的影响相互独立。

如公式 2.3 所示，类别c_i的先验概率为：

$$P(c_i)=\frac{N_i}{N}，\ \text{i=1,2,3,…..,k} \tag{2.3}$$

其中，N_i表示训练集中属于c_i类文本数目，N 表示训练集中文本总数目。

如公式 2.4 所示，未知类别文本 d 属于类别c_i的条件概率$P(d\mid c_i)$为：

$$P(d\mid c_i)=P((t_1,t_2,...,t_k)\mid c_i)=\prod_{i=1}^{k}P(t_k\mid c_i) \tag{2.4}$$

其中t_k为表示文本的特征集合中第 k 个特征词，$P(t_k\mid c_i)$表示特征词t_k在属于类别c_i的文档中出现的概率。

根据贝叶斯定理，如公式 2.5 所示，类别c_i的后验概率$P(c_i\mid d_i)$为：

$$P(c_i\mid d)=\frac{P(d\mid c_i)P(c_i)}{P(d)} \tag{2.5}$$

$$P(d)=\sum_{i=1}^{k}P(c_i)P(d\mid c_i) \tag{2.6}$$

如公式 2.6 所示，其中 P(d)对于各个类别来说是常数，代表文本 d 中的特征词集合在文本集合中出现的概率，故可以舍去，公式 2.5 简化后如下：

$$P(c_i\mid d)=P(d\mid c_i)P(c_i) \tag{2.7}$$

结合公式 2.4 和公式 2.7 可得：

$$P(c_i \mid d) = P(c_i)\prod_{i=1}^{k} P(t_k \mid c_i) \quad (2.8)$$

利用(2.8)式计算出每一个类别对文本 d 的后验概率值后，选择概率值最大的类别作为文本 d 的类别。

2.6.3 支持向量机

支持向量机分类算法（Support Vector Machine，SVM）[33-34]是由V.Vapnik 于1995 年提出的一种基于统计学习理论的机器学习方法。SVM 分类算法具有良好的性能，广泛地应用于文本分类、图像处理等领域，SVM 分类算法已经成为文本分类领域研究的热点之一。

SVM 分类算法主要针对两类分类问题，建立在统计学习理论中结构风险最小化的基础之上。SVM 分类算法的思想是通过将输入空间中线性不可分的数据映射到高维空间，使得数据成为线性可分，通过定义适当的内积可实现非线性转换。在数据线性可分的情况下需要在高维空间寻找一个最优分类面，所谓最优分类面是指能够将两类样本最大限度地分开的一个超平面，并且两类数据都存在一个与最优分类面最近的向量，最优分类面使得两个向量间的距离最大。最优分类面作为两类的边界，保证最小的分类错误率。

对于大于两类的多类别文本分类问题，因为每个最优分类面只能区分两个类别，就需要对每个类构造一个最优分类面，利用构造出的多个超平面，可区分当前类别与其他类别。

SVM 分类算法能够较好的解决高维、非线性、小样本等问题，相对其他分类算法而言，分类的精度和准确率较高，不容易产生过拟合现象。SVM 分类算法的不足在于核函数的选取较为困难，相关参数不易调整，训练时间长。

2.7 分类器性能评价

2.7.1 评价方法

随着文本分类领域的研究和发展，出现了众多分类器，需要提出系统的方法来评价分类器的好坏。评价分类器准确性的方法有保持法（Holdout）、交叉验证（Cross validation）等。保持法是将给定的数据集切分成训练集和测试集两个相互独立的数据子集。训练集和测试集的划分比例一般为 2:1，训练集占数据集的 2/3，测试集占数据集的 1/3。利用训练集归纳建立分类模型，在测试集上进行分类模型的准确率评估，分类器的准确率即在测试机上的准确率评估。交叉验证方法是将数据集合切分成若干个大小相同的子集，在检验过程中这些子集既被作为测试集又被作为训练集使用。每个子集作为训练集的次数一样，每个子集作为测试集当

且仅有一次。以二折交叉验证方法为例，将数据集切分成大小相同独立的两个子集s_1和s_2，首先其中一个子集s_1作为训练集，另一个子集s_2作为测试集进行验证，然后s_2作为训练集，s_1作为测试集进行验证，实现了交叉验证。k-折交叉验证方法是将数据集合切分出 k 个大小相同独立的子集$S=\{s_1,s_2,s_3,...,s_k\}$，进行 k 次验证即 k 次训练和测试。在进行第 k 次迭代时，将s_k作为测试集，其他的子集作为训练集。在进行 k 次迭代后，准确率的估计值为 k 次迭代正确分类的样本数除以原始数据中的总样本数。

2.7.2 评价指标

文本分类器的性能主要从效率和预测准确率两方面评价。效率的评估指标即分类器训练的时间和预测的时间，预测准确率的评估指标较为多样。在文本分类研究中，主要还是关注文本分类正确率，本文着重讨论文本分类正确率的评价，即文本分类的效果评价。文本分类效果的评价方法主要由查全率（Precision，准确率），查准率（Recall，召回率）和 F1 值。查全率和查准率以及 F1 值都由四个分类结果数据元素 TP、FP、TN 和 FN 构成，分类结果数据元素如表 2.1 所示：

表 2.1 分类结果数据元素表

Table 2.1 Data Elements of Classification Result

类别c_i	Relevant	Not Relevant
Retrieved	TP_i	FP_i
Not Retrieved	FN_i	TN_i

表 2.1 中，Relevant 指的是原本属于类别c_i的文本，Not Relevant 指的是原本不属于类别c_i的文本，Retrieved 指检索后被分到类别c_i的文本，Not Retrieved 指检索后没被分到类别c_i的文本，则表中 TP 表示原本属于类别c_i并且被正确分到类别c_i的文本数，FP 表示原本不属于类别c_i但是被错误分到类别c_i的文本数，FN 表示原本属于类别c_i但是被错误分其他类别的文本数，TN 表示原本不属于类别c_i也没有被分到类别c_i的文本。

对于类别c_i，如公式 2.9 所示，分类查全率定义为：

$$\mathrm{Re}call_i=\frac{TP_i}{TP_i+FN_i} \tag{2.9}$$

对于类别c_i，如公式 2.10 所示，分类查准率定义为：

$$\mathrm{Pr}ecision_i=\frac{TP_i}{TP_i+FP_i} \tag{2.10}$$

已知测试集类别c_i中有$\mathrm{TP}_i+\mathrm{FN}_i$个文本，分类后有$\mathrm{TP}_i$个文本正确分到类别$c_i$

中，查全率代表被正确分到类别c_i的文本数占原类别c_i文本数的比例，考察的是分类结果的完整性。

已知测试集类别c_i中有$\mathrm{TP}_i+\mathrm{FN}_i$个文本，分类后有$\mathrm{TP}_i$个文本正确分到类别$c_i$中，其他类别中有$\mathrm{FP}_i$个文本被错误分到类别$c_i$中，查准率代表分类以后类别$c_i$中真正原本属于类别$c_i$的文本所占比例，考察的是分类结果的正确性。

由于查全率和查准率是两个互相矛盾的指标，查全率的提高需要以查准率的下降为代价，故提出 F-测量(F-Measure)指标，F-测量是将查全率和查准率综合考虑的指标。如公式 2.11 所示，F-测量(F-Measure)定义如下：

$$F_\beta=\frac{(\beta^2+1)PR}{\beta^2P+R} \tag{2.11}$$

其中β是权重调整因子，使得查全率和查准率体现出不同的权重。当β为 1 时，表示查全率与查准率权重相同。此时 F-测量被称为F_1，如公式 2.12 所示：

$$F_1=\frac{2PR}{P+R} \tag{2.12}$$

查全率、查准率和 F1 值都是表示局部意义，相对单个类别而言。通过宏平均（Macro-Averaging）和微平均（Micro-Averaging）可以综合所有类别上的查全率，查准率和 F1 值来全局评价分类器的分类效果。用|C|表示测试集文本的总数，宏平均和微平均定义如公式 2.13-2.17 所示：

宏平均：

$$MacroAvg\_\mathrm{Re}call=\frac{\sum_{i=1}^{|C|}\mathrm{Re}call_i}{|C|} \tag{2.13}$$

$$MacroAvg\_\mathrm{Pr}ecision=\frac{\sum_{i=1}^{|C|}\mathrm{Pr}ecision_i}{|C|} \tag{2.14}$$

$$MacroAvg\_\mathrm{F1}=\frac{\sum_{i=1}^{|C|}F1_i}{|C|} \tag{2.15}$$

微平均

$$MicroAvg\_\mathrm{Re}call=\frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}(TP_i+FN_i)} \tag{2.16}$$

$$MicroAvg\_\mathrm{Pr}ecision=\frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}(TP_i+\mathrm{FP}_i)} \tag{2.17}$$

$$MicroAvg\_F1 = \frac{\mathrm{Re}call \times \mathrm{Pr}ecision \times 2}{\mathrm{Re}call \times \mathrm{Pr}ecision} \quad (2.18)$$

微平均和宏平均的不同之处在于，宏平均是是求每个类别性能指标的平均值，微平均是求每个文本性能指标的平均值。在文本集合中各个类别中文本分布不均衡的情况下，一般说来，微平均受大类影响较大，而宏平均故相对受稀有类别影响较大，所以分类器性能评估指标的选择视实际应用的需求而定。

3 基于类别相关的新文本特征提取方法

3.1 文本特征提取的过程

基于第二章 2.4.2 节，我们可知文本特征提取可分为基于 Filter 模型的文本特征提取与基于 Wrapper 模型的文本特征提取。基于 Filter 模型的文本特征提取不依赖分类器，根据某种评价准则对文本特征进行计算，得出评价值，然后对评价值进行排序，选择出排序靠前的文本特征子集；基于 Wrapper 模型的文本特征提取依赖分类器，将使用当前文本特征子集后分类器的性能好坏作为特征提取的评价准则，具有准确率高的优点。由于文本数据具有高维特性，而且基于 Wrapper 模型的文本特征提取本身计算复杂度高，故不适用与文本分类领域。由于基于 Filter 模型的文本特征提取简单易用计算量较小，因此文本领域中得到了广泛的应用。本文所讨论的特征提取的方法均属于过滤法。

基于 Filter 模型的文本特征提取有很多具体方法，但是基本过程一致，可将其具体步骤概括为：

1）通过对训练文本集合进行分词预处理，从而获得训练文本集合包含的所有文本特征项（特征项不重复），构成文本集合的原始特征集合 T。

2）用某种特征评价函数对集合 T 中的每一文本特征项评分。将集合 T 中各个文本特征项得分按分值由高到低进行排序。

3）按照实验需求进行向量降维设定，若需将原始的向量空间维数降为 N，则选择集合 T 中文本特征项排名最高的前 N 个特征项，这 N 个特征项便是最终用来表示文本的特征向量中的特征项，构成最终特征表示集合 T_f。T_f 为 T 的子集，T_f 将作为训练文本与测试文本的表示向量中的特征项，应用于文本分类的训练与测试过程。特征提取时具体选用多少维特征，目前还没有具体统一的方法。一般选取特征维数基于具体问题而定。不同的数据集、不同的特征评价函数选用的维数是不同的。

根据特征评价函数不同，产生了各种特征提取方法，因此特征提取方法以特征评价函数的名称来命名。目前一般的特征评估函数基本上都是基于数理统计理论与信息论。现有的特征提取方法包括文档频数(Document Frequency，DF)、信息增益（Information Gain，IG）、互信息（Mutual Information，MI）、x2 统计（Chi-square statistic，CHI）和 CC（Correlation Coefficient）[35-39]等。目前 CHI 和 IG 是性能较好的特征提取方法，DF 和 MI 性能较差[38]。下面将对这些方法的评价函数构造原理进行分析，并且讨论其优缺点。

3.2 常用的特征提取方法

3.2.1 文本频数

训练集合文本可以看成是一个庞大特征词集合，文本频数提取方法是基于稀有特征对分类几乎无帮助的假设，文本频数依据统计训练集合中包含当前特征词的文档数目进行特征词提取。若包含当前词条的文档数目超过预先定义的阈值，则该特征词被认为对分类预测有贡献，故被作为特征词选取保留。若低于预先定义的阈值，则从特征词集合中剔除，能有效降低向量空间的维度。文献[42]表明由于稀有特征很多时候表现为噪声数据，稀有特征的剔除有助于分类的准确性。文本频数是最简单特征提取方法之一，计算复杂度低，适合进行大规模的文本分类情况。

3.2.2 信息增益

① 信息论

信息论[27]是人们在长期通讯工程的视线中，由通讯技术，概率论，随机过程和数理统计相结合逐步发展起来的一门新兴学科。信息论的奠基人美国科学家香农（C.E.Shannon）在其著名的论文《通信的数学理论》中给出了信息的定量表示方法“香农信息”。“香农信息”反映的事物的不确定性。

设 i 元信源 X 的概率空间为：

X = a_1， a_2 ，... ， a_i

P(X)= $p(a_1)$， $p(a_2)$，… ， $p(a_i)$

则 X 中符号 a_i 的香农信息定义如公式 3.1 所示：

$$I(a_i)=\log\frac{1}{P(a_i)} \tag{3.1}$$

I(a_i)成为 a_i 的自信息，I(a_i)描述的是随机事件 a_i 出现的先验不确定性，也表示事件 a_i 出现所携带的信息量。

相对于自信息 I(a_i)描述信源单一事件 a_i 的信息量，信息熵是对整个信源 X 中所有事件做自信息并计算期望值，如公式 3.2 所示：

$$E\left(\log\frac{1}{P(a_i)}\right)=\sum_{i=1}^{n}P(a_i)\log\frac{1}{P(a_i)} \tag{3.2}$$

通常记作为，如公式 3.3 所示：

$$H(X)=\sum_{i=1}^{n}P(a_i)\log\frac{1}{P(a_i)} \tag{3.3}$$

$H(X)$ 称为信息源 X 的信息熵，$H(X)$ 是信息源 X 中每个事件出现的平均信息量，反映的是整个信息源的不确定性。

条件熵：在联合集合 XY 上的条件自信息，如公式 3.4 所示：

$$H(X|Y)=E\left[I(a_i|b_j)\right] \quad (3.4)$$

信息增益是对不确定的消除，衡量一个变量的出现与否对信源不确定性的消除，也就是信源从这个变量上获得的信息量，信息增益如公式 3.5 所示：

$$IG(Y)=H(X)-H(X|Y) \quad (3.5)$$

② 信息增益的数学模型表示

现在从广义信息论的角度出发，现在进行一种假设，在一个文档集合中，已知文档类别种类存在 i 种可能，将文档类别集合表示为$C=(c_1,c_2,...,c_i)$。文档可能包含特征 t,也可能不包括，将文档是否包含特征 t 表示为集合$T=(t,\bar{t})$。任取一个文档，在求解特征集合$T=(t,\bar{t})$为决定整个文档类别系统带来的信息量时候，可以将以上问题描述成两个随机事件集合：文档类别集合和特征词集合。用文档类别集合描述在这个文档类别集合中任何一个文档属于某个类别的概率事件的集合。

事件集合 C 概率空间表示如下：

C = c_1， c_2 ，... ， c_i

P(C)= $p(c_1)$， $p(c_2)$， … ， $p(c_i)$

$p(c_i)$表示随机取一个文档属于类别c_i的概率，$\sum_{i=1}^{n}p(c_i)=1$。

事件集合 T 概率空间表示如下：

T = t， $\bar{t}$

P(T)= $p(t)$， $p(\bar{t})$

$$p(t)+p(\bar{t})=1 \quad (3.6)$$

公式 3.6 中 $p(t)$ 表示文档包含特征 t 的概率， $p(\bar{t})$ 表示文档不包含特征 t 的概率。

基于上述定义我们可以得出以下信息，如公式 3.7 所示：

$$H(C)=-\sum_{i=1}^{n}P(c_i)\log P(c_i) \quad (3.7)$$

公式 3.7 表示文档类别归属情况的不确定性。

$$\begin{aligned}H(C|T)&=p(t)H(C|t)+p(\bar{t})H(C|\bar{t})\\&=P(t_k)\sum_{i=1}^{n}P(c_i|t_k)\log P(c_i|t_k)+P(\bar{t_k})\sum_{i=1}^{n}P(c_i|\bar{t_k})\log P(c_i|\bar{t_k})\\&=\sum_{i=1}^{n}P(c_i)\log P(c_i|t_k)+\sum_{i=1}^{n}P(c_i)\log P(c_i|\bar{t_k})\end{aligned} \quad (3.8)$$

公式 3.8 中 H(C|T)表示以文档包含特征集合 T 为前提下，文档类别归属情况的不确定性。

$$IG(C;T)=H(C)-H(C|T)=\sum_{i=1}^{n}P(c_i,t)\log\frac{P(c_i,t)}{P(t)P(c_i)}+\sum_{i=1}^{n}P(c_i,\bar{t})\log\frac{P(c_i,\bar{t})}{P(\bar{t})P(c_i)} \quad (3.9)$$

则 IG(C;T)=H(C)-H(C|T)表示为特征集合 T 为决定文档类别归属情况消除的不确定性，熵只是对不确定的描述，不确定的消除正式特征集合 T 给文档类别归属情况带来的信息量，这个信息量越大，说明该特征对文档类别归属贡献越大。应该作为重要特征被保留下来。

3.2.3 互信息

互信息是信息论中作为衡量两个信号关联程度的一种尺度,后来引申为描述两个随机变量统计相关性的测度。文本的特征t_k和类别c_i的互信息如公式 3.10 所示:

$$MI(t_k,c_i)=\log\frac{P(t_k\mid c_i)}{P(t_k)} \tag{3.10}$$

一般取互信息的期望值如公式 3.11 所示：

$$MI(t_k,c_i)=\sum_{i=1}^{n}p(c_i)\log\frac{P(t_k\mid c_i)}{P(t_k)}=\sum_{i=1}^{n}p(c_i)\log\frac{P(c_i,t_k)}{P(t_k)} \tag{3.11}$$

从概念上我们把互信息与信息增益进行比较，会发现实际上信息增益即是平均互信息。所不同的是概率空间的不同，信息增益概率空间基于$T=(t,\bar{t})$，互信息的概率空间基于$T=(t_1,t_2,...,t_n)$。

互信息衡量的是某个词和类别之间的统计独立关系。互信息的不足之处在于得分非常受词条的边缘概率的影响，且低频词具有较大的互信息。

3.2.4 x^2 统计

CHI 统计方法，在数理数理统计中一种常用的检验两个变量独立性的方法。CHI 最基本的思想就是通过观察实际值与理论值的偏差来确定理论的正确与否。运用在文本特征选择中，假设t_k与类别c_i之间是独立的[25]。这种独立关系类似于具有一维自由度的x^2分布，t_k对于c_i的 CHI 统计量，如公式 3.12 所示：

$$\chi^2(t_k,c_i)=\frac{N(P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(\overline{t_k},c_i)P(t_k,\overline{c_i}))^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})} \tag{3.12}$$

CHI 统计方法用$\chi^2(t_k,c_i)$来度量特征项t_k和类别c_i之间的相关程度。特征性选择的过程就是计算特征项t_k和类别c_i的$\chi^2(t_k,c_i)$值，然后按值大小排序，根据实际需要选取 CHI 值大的特征项。由于 CHI 方法是基于分布假设的，如果特征词和文本类别间的这种分布假设被打破，那 CHI 方法会更倾向于选择低频特征词

3.3 基于类别相关的新文本特征提取方法

本节提出了一种基于类别正相关和类别强相关的新特征提取方法 SP (Strong Correlation and Positive Correlation，正相关与强相关)，SP 只选取与类别正相关和强相关的特征。首先给出特征与类别相关性和相关度的概念，通过对 CHI 不能区别正相关特征与负相关特征重要程度的原因进行分析，将文本特征与类别正、负

相关特征的重要性差异度作为正相关性因子，用来区别特征与类别正负相关性。通过对文本特征在文本集合中分布规律的分析，将文本特征类内和类间的相关度指标结合成为强相关度因子，用来区别特征与类别的强弱相关度。实验结果表明该方法的可行性和有效性。

3.3.1 特征与类别的相关性

特征与类别的相关性包括正相关性和负相关性。为了更好地描述特征与类别的正负相关性概念，这里将通过对文本集合的一般特性的分析，具体地描述特征与类别的正负相关性，文本集合这种特性可以由表 1 中四种基本元素 A，B，C，D 表示。为进一步解释，现做一下假设：在一个文本集合 H 中，假设存在一个类别c_i，集合中除c_i以外的类别表示为$\overline{c_i}$，则按照文本所属类别划分，该集合 H 可以表示为{c_i，$\overline{c_i}$}。假设存在一个文本特征词t_k，集合中除t_k外的特征词表示为$\overline{t_k}$，则按照是否包括特征词t_k划分，该集合 H 可以表示为{t_k，$\overline{t_k}$}。按照文本是否包括特征词t_k，是否属于类别c_i，可将文本集合 H 表示为{c_i，$\overline{c_i}$}和{t_k，$\overline{t_k}$}的笛卡尔乘积。该集合由以下四个信息元素 A，B，C，D 构成：

表 3.1　信息基本元素表

Table3.1　Fundamental information elements

元素	c_i	$\overline{c_i}$
t_k	A_i	B_i
t_k	C_i	D_i

其中$A_i=(t_k,c_i)$表示不仅包含特征t_k而且属于类别c_i的文本的数量，$B_i=(t_k,\overline{c_i})$表示包含特征t_k但是不属于类别c_i的文本的数量，$C_i=(\overline{t_k},c_i)$表示不包含特征t_k但是属于类别c_i的文本的数量，$D_i=(\overline{t_k},\overline{c_i})$表示不包含特征$t_k$并且不属于类别$c_i$的文本的数量。文本总数量 N=A+B+C+D。

基于表 3.1 的描述，将特征选择方法根据特征与类别的相关性进行分类。特征与类别的相关性包括正相关性和负相关性[41]：

定义 1（若特征项t_k当且仅当只出现在与类别c_i相关的文本中，则称特征项t_k与类别c_i正相关。）

定义 2（若特征项t_k当且仅当只出现在与类别c_i不相关的文本中，则称特征项t_k与类别c_i负相关。）

表 3.1 中$A_i=(t_k,c_i)$和$D_i=(\overline{t_k},\overline{c_i})$描述的是特征项$t_k$和类别$c_i$的正相关程度，$B_i=(t_k,\overline{c_i})$和$C_i=(\overline{t_k},c_i)$描述的是特征项$t_k$和类别$c_i$的负相关程度。

在接下来的新特征方法提出部分，我们将寻求一种区别特征与类别间的正负相关性的表现形式。该表现形式将作为特征与类别正相关因子在新特征提取方法中呈现。

3.3.2 特征与类别的相关度

特征与类别的相关度可分为强相关和弱相关。在此我们将通过结合本文第 2 部分的文本集合一般性描述来进一步阐述特征与类别的相关度概念。特征的相关度是衡量一个特征代表一个类别的程度，特征与类别的相关度的评价可以基于以下两种指标：

① 若存在一个特征词，该特征词越集中出现在当前所在类别文档中，在其他类别文档中越少出现，那么这个特征词越能区别所在类别与文本集合中的其他类别。即该特征词越能代表当前类别，对分类贡献越大。这里将衡量类别间特征分散度的指标称为相关度的类间指标。

② 若存在一个特征词，在当前所在类别内，包含该特征词的文档在该类别中比例越高越能代表当前类别，对分类贡献越大。将衡量类内特征分散度的指标称为相关度的类内指标。

在此我们采用两种比值表示相关度的类间指标和类内指标

① $\frac{A_i}{B_i}$，如果特征词t_k只与类别c_i相关度高，$\frac{A_i}{B_i}$值越高那么说明特征词t_k能很好的代表类别c_i，那么$\frac{A_i}{B_i}$的值就高。

② $\frac{A_i}{C_i}$，在类别c_i中有两个特征词t_k和t_m，两个特征词各自对应一个$\frac{A_i}{C_i}$值，那么$\frac{A_i}{C_i}$值越大的特征词越能代表类别c_i。

鉴于以上概念的提出，我们结合类间指标和类内指标来刻画特征与类别的相关程度，区别强、弱相关特征。将特征与类别的强相关度因子 SCD(Strong Correlation Degree)表示为，如公式 3.13 所示：

$$SCD(t_k, c_i) = \frac{A_i}{B_i}\frac{A_i}{C_i} \tag{3.13}$$

若一个特征与一个类别相关度越强，则 SCD 值越大。SCD 优先选择与类别强相关的特征，从而能够有效地避免弱相关特征产生的噪声干扰，SCD 作为特征与类别的强相关度因子，将在新的特征提取方法中呈现。

3.3.3 SP 文本特征提取方法

基于以上文本特征与类别的相关性和相关度的概念的提出，我们采用了强相关度因子 SCD 来区别特征与类别相关度的强弱，接下来将对 CHI 无法体现正相关

特征与负相关特征的重要性差异度的原因进行分析，寻求特征与类别间的正负相关性的具体表示方法。

基于 3.2.4 节对 CHI 统计方法的介绍，可知 CHI 方法是在数理数理统计中一种常用的检验两个变量独立性的方法。CHI 统计方法用 $x^2(t_k,c_i)$ 来度量特征项 t_k 和类别 c_i 之间的相关程度。现在我们将 CHI 概率表现形式公式 3.12，重新以信息四大基本元素 A，B，C，D 诠释，公式如 3.14 所示：

$$x^2(t_k,c_i)=\frac{N(A_iD_i-B_iC_i)^2}{(A_i+C_i)(B_i+D_i)(A_i+B_i)(C_i+D_i)} \tag{3.14}$$

$\chi^2(t_k,c_i)$ 值越大说明特征项 t_k 和类别 c_i 相关程度越高，此时特征项 t_k 所包含的与类别 c_i 相关的信息就越多。$\chi^2(t_k,c_i)=0$ 表示特征项 t_k 与类别 c_i 相互独立。特征项 w 与类别 c_i 的相关性越强，$\chi^2(t_k,c_i)$ 的值就越大。

由定义 1 和 2 可知 A_i 和 D_i 描述的是特征项 t_k 和类别 c_i 的正相关程度，B_i 和 C_i 描述的是特征项 t_k 和类别 c_i 的负相关程度。由概率论可知假设 A_i 的值越大，D_i 的值越大，即特征 t_k 出现在 c_i 类的文本里面的概率越大，则 t_k 与 c_i 正相关度越大。假设 B_i 的值越大，C_i 的值越大，那么特征 t_k 出现在类别 $\overline{c_i}$ 的概率越大。

$A_iD_i-B_iC_i$ 体现了特征 t_k 与类别 c_i 的正、负相关特征重要性差异度。若 $A_iD_i-B_iC_i>0$，则特征 t_k 与类别 c_i 体现出正相关性。若 $A_iD_i-B_iC_i<0$，则特征 t_k 与类别 c_i 体现出负相关性。由公式（3.14）可知，分子中 $(A_iD_i-B_iC_i)^2$ 使得 $A_iD_i-B_iC_i$ 呈现非负性，若特征 t_k 与类别 c_i 体现出正相关度越大，那么 $A_iD_i-B_iC_i$ 必然为正值，且值将随正相关度的增加而增大，CHI 得出的特征值也将增大。若特征 t_k 与类别 c_i 体现出负相关性，那么 $A_iD_i-B_iC_i$ 必然为负值，而且值随着负相关度的增加而减小，但是 $(A_iD_i-B_iC_i)^2$ 的值将随之变大，CHI 得出的特征值也将变大。这样正、负相关特征的重要性差异度便无法体现。

通过以上分析，正、负相关特征重要性差异度能很好地刻画文本特征与类别的正负相关性。假定用正相关性因子 PC（Positive Correlation）来区别特征与类别间的正负相关性，公式如 3.15 所示:

$$PC(t_k,c_i)=A_iD_i-B_iC_i \tag{3.15}$$

新文本特征提取方法的目地是选择与类别正相关并且强相关的特征词，在综合以上所述，我们提出新的文本特征提取方法 SP（SCD&PNC），如公式 3.16 所示：

$$SP(t_k,c_i)=SCD*PC=\frac{A_i}{B_i}\frac{A_i}{C_i}(A_iD_i-B_iC_i)=(\frac{A_iD_i}{B_iC_i}-1)A_i^2 \tag{3.16}$$

若一个特征词与类别正相关并且强相关，那么 SP 值一定越大，故 SP 优先选择与类别正相关并且强相关的特征词。

由 2.4.3 节可知文本特征提取策略可以分为：局部特征提取和全局特征提取。

在均衡数据中全局特征提取优于局部特征提取，单独利用局部特征提取的缺点是局部特征提取不能有效地选取到能代表所有类别的特征集合，忽视了全体训练样本和类别的整体性。本文使用均衡数据集，采用全局特征提取策略。

基于全局特征提取策略的特征提取方法，需获取特征项t_k对整个文档集的全局特征值时，由于 SP 是基于类别正相关性与强相关性的特征提取方法，采用公式 3.17 基于最大值的方式计算全局特征，度量各个特征对于分类的重要性，能够选择出对某一个类具有较好标识作用的特征项。

$$SP_{Max}(t_k) = \max_{i=1}^{m}\{SP(t_k, c_i)\} \quad (3.17)$$

SP 特征提取方法的优点在于：计算复杂度小，只选择与类别正相关并且强相关的特征，避免了弱相关特征和负相关特征的干扰。实验表明结果该方法表明具有较好的降维效果，通过构造低维特征向量能有效表示文本，并且 SP 在各项分类评价指标上优于 CHI、CC、DF 提取方法。

3.4 本章小结

本章首先对特征提取一般过程的介绍，然后对常用特征提取方法进行详细介绍。通过研究分析各个常用特征提取方法的优缺点，发现弱相关特征与负相关特征在文本特征提取过程中会产生干扰，从而影响特征提取的质量，使得分类性能下降。为了能够有效避免提取过程中弱相关特征与负相关特征产生的干扰，3.3 节提出了本文提出一个新的基于类别正相关和类别强相关的特征提取方法 SP (Strong Correlation and Positive Correlation，正相关与强相关)，SP 方法中正相关性因子通过区别特征与类别正负相关性，优先选择正相关特征，避免了负相关特征的干扰。强相关度因子通过区别特征与类别的强弱相关程度，优先选择强相关特征，避免了弱相关特征的干扰。SP 通过对两种因子的结合，能够有效地提取高质量的文本特征。接下来将在第四章中通过多组对比实验来验证新特征提取方法的可行性和有效性，同时表明新特征提取方法具有强降维能力和良好的分类效果。

4 中文文本分类系统的设计与实现

4.1 中文文本分类系统的总体设计

4.1.1 系统需求

中文文本分类系统是为满足本文实验需求，支撑实验顺利进行而开发的实验平台。本文基于开源全文检索引擎 Lucene 和开源汉语词法分析系统 ICTCLAS 两个开源工具包进行中文文本分类系统开发。该系统通过对训练文本以及测试文本集合的处理，提取文本特征，并用提取获得的文本特征表示训练文本与测试文本，从而作为输入数据通过分类算法进行分类，最后通过对分类效果的评价来作为特征提取方法好坏的依据。

根据本文第二章“文本分类关键技术及分析”对详细地对文本分类过程中关键文本预处理、文本表示、文本特征降维、文本特征加权、分类方法的选择和分类性能评价五个部分进行的研究与分析，可知基于空间向量模型构造的中文文本分类系统需要满足实验过程会产生的以下具体需求：

① 文本预处理

1）中文文本分词

中文分词技术是中文信息处理的基础。由于本文利用VSM空间模型表示文本，因此需要将中文文本表示成为文本特征向量形式，特征项的确定需要从训练文本中提取。中文中的特征项便是字和词，因此需要对中文文本进行分词。

2）词条选择与停用词删除

停用词（StopWords）指的是虽然在文本集合中出现频度很高，但是对分类毫无贡献，存在只会增大特征空间维度，增加分类运算复杂度的无用词。

在向量空间模型中，空间向量由文本的特征项构成，未做特征选择之前文本集合中所有词条都可以作为表示文本的特征项。由于文本集合中的词汇量庞大，导致文本特征空间经常高达几万维，甚至更高。不能体现文本内容信息的虚词对文本分类无贡献，若作为文本特征用于文本分类将产生负面影响，因此应被视为是噪声数据。

由于停用词删除和词条选择，对文本特征空间进行压缩，能够减少存储空间，提高文本分类准确度，降低运算复杂度，提高程序效率。所以在预处理中进行用词删除和词条选择作为特征空间初降维最有必要性。

② 文本特征提取

文本特征提取是一种特征降维技术，是文本分类过程中最核心的环节。由于文本特征空间具有“高维性”和“高稀疏性”的特点，导致了很多分类算法无法应用，

即使能够应用，也严重影响了分类的准确度和时间。通过文本特征提取对特征空间的降维，不仅提高了分类速度，而且过滤了噪声数据，提高了精度的同时还有助于解决过拟合问题，所以文本特征提取具有必要性。提取所得的文本特征需要保存，以便在通过特征权重计算后对训练文本、测试文本进行表示。

③ 文本特征加权

在文本特征提取后，通过对文本特征进行加权，使得文本特征向量表示文本时能够最大限度的区分不同文本。通过文本特征权重计算对训练文本、测试文本进行表示。作为分类算法的输入数据，因此文本特征加权是必要环节。

④ 文本分类算法

通过某个特征提取方法和特征加权方法对训练文本、测试文本进行文本表示，作为文本分类算法的输入数据。文本分类算法的结果的好坏是对特征提取方法以及特征加权方法好坏的评价依据。本文使用特定的 KNN 算法，因此文本分类算法实现过程中需要解决如何选择 K 值的问题。

⑤ 分类性能评估

前面提过文本分类算法的结果的好坏是对特征提取方法以及特征加权方法好坏的评价依据。分类性能评估将实现如何评价分类算法的结果的好坏。

⑥ 大量文本的处理

由于文本分类数据集相对较大，少则数千篇文本，多则上万篇文本。在文本特征加权实现与文本特征提取实现中涉及到权重公式以及特征评价函数的计算。本文采用的特征评价函数与权重公式都是基于数量统计理论，于是需要对特征在文本中的分布情况，在类别中的分布情况，在整个数据集合中的分布情况进行统计。在这个统计过程中，计算量较大，复杂度较大，加上文本数据集合具有大量文本，如何快速准确的进行统计亦是系统需要解决的问题。

4.1.2 开发平台

① 系统开发语言：Java，Matlab

② 架构：C/S

③ 开发环境: MyEclipse 6，JDK 1.6 及以上版本，Matlab7.0

④ 操作系统：Windos7

⑤ 关键工具包: 开源全文检索引擎 Lucene 工具包 lucene-core-3.0.3.jar，汉语词法分析系统 ICTCLAS 2011 工具包 ICTCLAS2011_Windows_32_jni。

其中 ICTCLAS[13]是由中国科学院计算技术研究所通过多年研究工作积累，研制开发出的汉语词法分析系统。ICTCLAS 主要功能包括中文分词；词性标注；命名实体识别；新词识别；同时支持用户词典；支持繁体中文；支持 gb2312、GBK、UTF8 等多种编码格式。 ICTCLAS 分词速度单机 500KB/s，分词精度 98.45%，API

不超过 100kb，各种词典数据压缩后不到 3M，是世界上最好的汉语词法分析器。

Lucene[12]是 Apache 软件基金会 Jakarta 项目组开发的成员项目，Lucene 是一个开源的全文检索引擎工具包。它包含部分文本分析引擎，完整的索引引擎与查询引擎。开发人员通过引擎工具包提供的函数接口，便可将 Lucene 快速地与各种应用整合，实现在该应用中的全文检索功能。从功能上来说，Lucene 有建立索引、优化索引结构、处理查询返回结果集等基本功能。如图 4.1 所示，从结构上来说，Lucene 包括文本分析引擎、索引引擎、查询引擎以及二次应用开发接口等。如若想了解更多 Lucene 的信息详见其官方网站，网址为：http://lucene.apache.org。

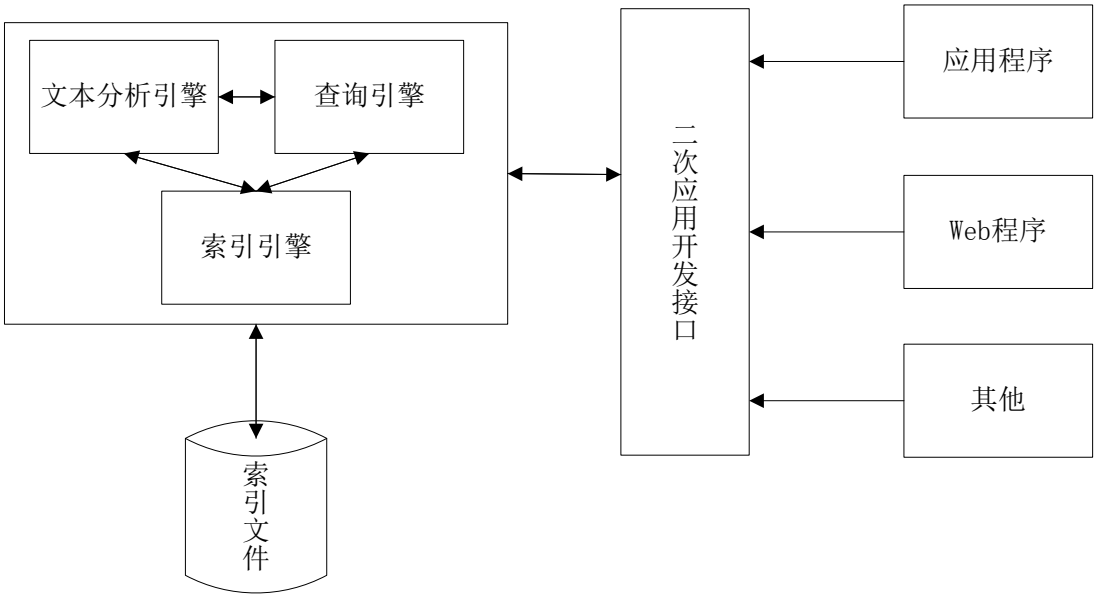


图 4.1　基于 Lucene 的全文检索系统结构

Fig.4.1　Architecture Information Retrieval System Based on Lucene

4.1.3 系统关键问题解决方案

针对文本分类系统中将会遇到的问题，特别是中文文本分词和大量文本统计两个问题，通过大量的调查研究工作，制定了适合本文需求的中文文本分类系统的解决方案。该解决方案以 Lucene 为基础，将 ICTCLAS 整合进 Lucene 文本分析引擎中，对 Lucene 中文分词模块进行扩展。该方案结合了 ICTCLAS 与 Lucene 二者的优点，充分利用各自的特点，进行优势互补。该解决方案具体内容包括对文本分析引擎进行中文分词扩展，自定义查询引擎相关索引信息并创建文本索引，自定义查询引擎中具体查询。接下来本节将结合 Lucene 搜索、索引工作流和源码结构进行解决方案的介绍：

本节将通过 Lucene 索引结构，Lucene 的源码结构分析，Lucene 索引过程和搜

索过程中工作流的分析，具体介绍如何对 Lucene 进行二次开发以满足文本分类系统需求。

① Lucene 总体架构

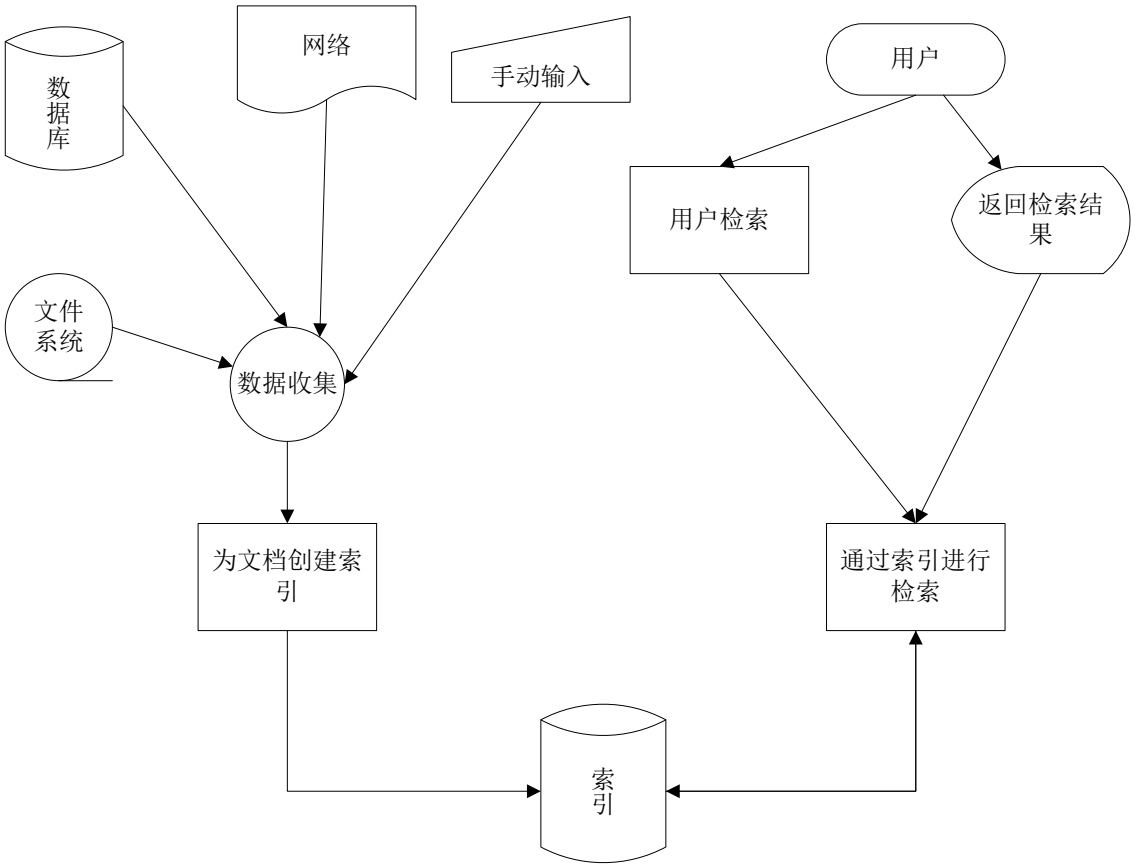


图 4.2 Lucene 总体架构

Fig.4.2 Architecture of Lucene

Lucene 总体结构图表明了 Lucene 工作方式主要分为:索引的建立,索引建立是通过数据收集器，收集各种数据，通过分析处理数据建立索引，然后存入索引数据库。检索，数据的检索从用户的输入文本出发，通过对文本分析处理，把输入的文本转化为关键字、词，搜索索引数据库，返回数据给用户。

② 索引结构

Lucene 在传统全文检索引擎的倒排索引的基础上，实现了分块索引，能够针对新的文件建立小文件索引，提升索引速度。然后通过与原有索引的合并，达到优化的目的。索引文件格式独立于应用平台。Lucene 定义了一套以 8 位字节为基础的索引文件格式，使得兼容系统或者不同平台的应用能够共享建立的索引文件。

顺序扫描和全文检索区别，所谓顺序扫描，顺序扫描运用正向索引技术，顺序扫描即在一个文件集合中寻找内容包含某一个字符串的文件，若当前文档包含特定字符串，则此文档即是所需文档，需要对每个文档挨个扫描查找，直到扫描完所有的文件。这种查找方法耗时太长。全文检索运用反向索引技术，倒排索引源于实际应用中需要根据属性的值来查找记录。这种索引表中的每一项都包括一个属性值和具有该属性值的各记录的地址，倒排索引是从字符串到文件的映射，是文件到字符串映射的反向过程，因此搜索速度大大提升。反向索引的所保存的信息一般如下：

假设文档集合中有 100 篇文档，编号为 1-100。则反向索引结构如下：

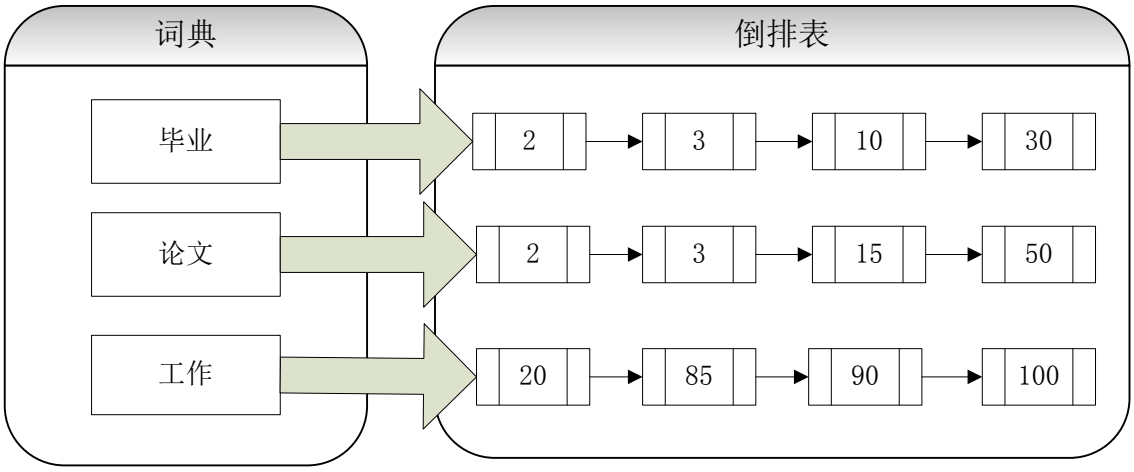


图 4.3　反向索引结构图

Fig.4.3　Architecture of Inverse Index

左边保存的是一系列字符串，称为词典。

每个字符串都指向包含此字符串的文档(Document)链表，此文档链表称为倒排表(Posting List)。

在数据量小的情况下，由于反向索引技术需要消耗时间创建索引，故全文检索不一定比顺序扫描快。但是顺序扫描是每次都要扫描，而创建索引的过程仅仅需要一次，之后搜索操作直接对创建好的索引进行搜索。全文搜索相对于顺序扫描的优势之一便是：创建一次索引，可多次使用。

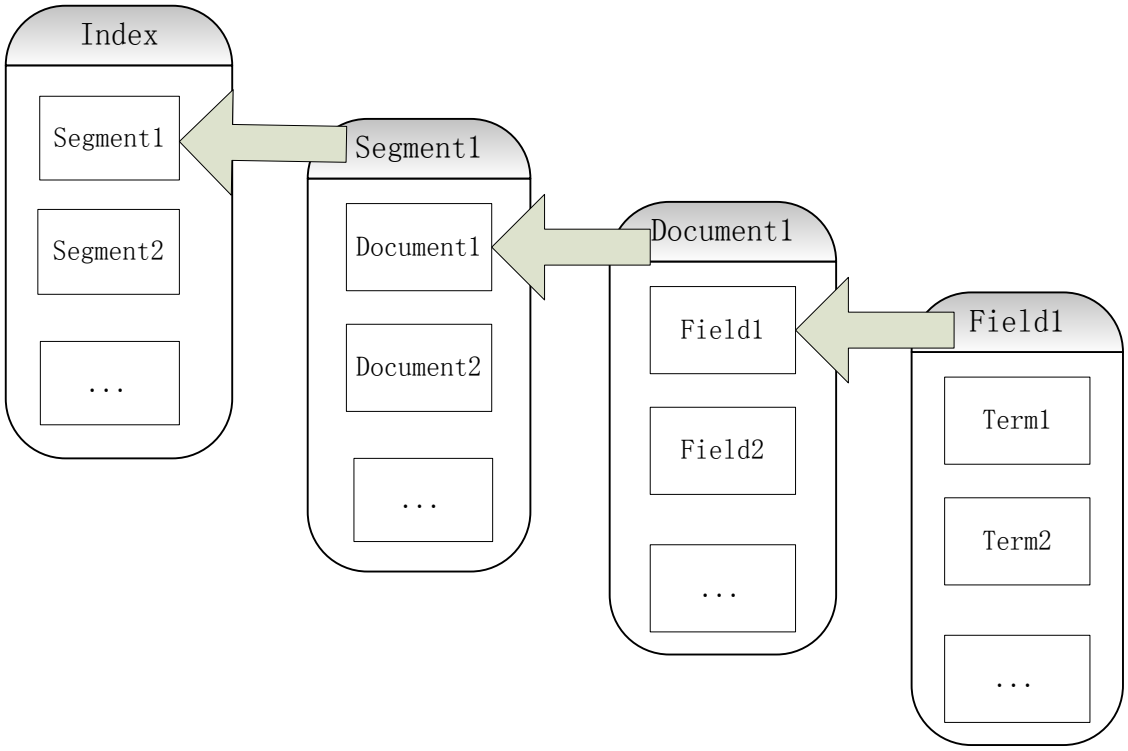


图 4.4　Lucene 索引文件的概念结构

Fig.4.4　Concept Architecture of Lucene's Index File

如图 4.4 所示，Lucene 索引文件结构中，一个域(field)包含多个项（term)，一个文档(document)包括多个域，一个段(segment)包含个文档，多个段最终组成了 Lucene 索引 index。项是最小的索引概念单位，它包含了一个字符串的内容以及其在文件中的出现次数等信息。域是是一个“域名-域值”对，其中域名代表文档的一个属性，如“标题”，域值是一个项，如实际标题的项，二者组成了一个域。文档包含了各个域的内容。如上图所示，Lucene 的索引结构就是传统的倒排索引结构。

③ Lucene 内部工作流

在此将通过 Lucene 内部工作流的分析，明确为了更好中文文本分类系统，需要进行的具体工作，以便结合 Lucene 的源码结构，对相关问题解决进行具体的设计。

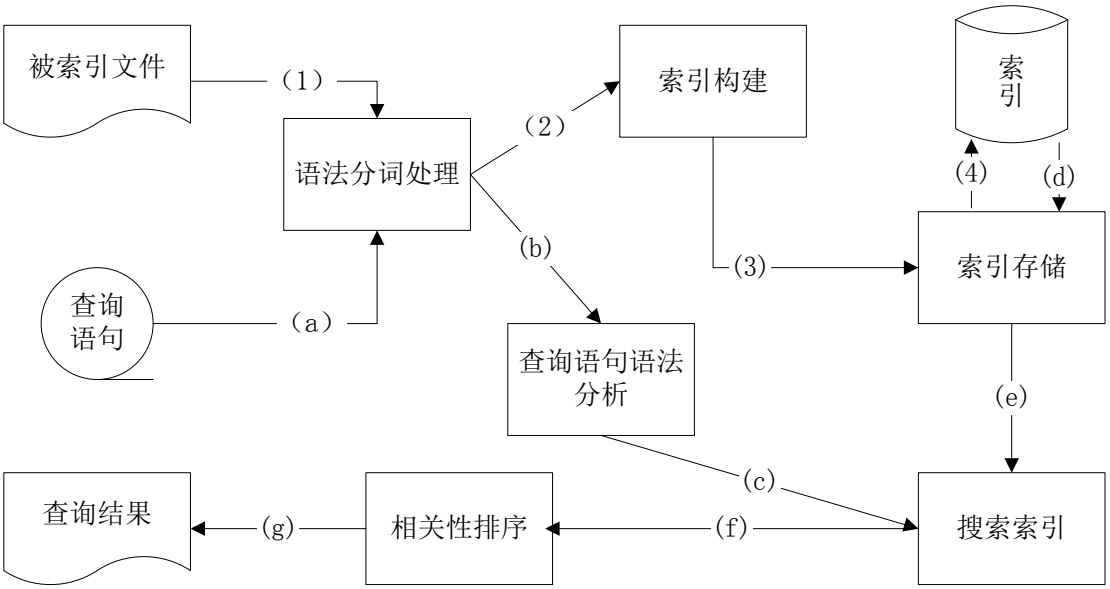


图 4.5　Lucene 的索引和搜索过程示意图

Fig.4.5　Index and Search Process of Lucene

如图 4.5 所示，索引的过程为 1-4，搜索的过程为 a-g。

1) 索引过程：被索引的文件通过文本分析引擎的具体分析处理形成一系列词(Term)，接着经过索引引擎创建形成词典和反向索引表，最后通过索引存储将索引存储在硬盘上。

2) 搜索过程：文本分析引擎通过对查询语句具体分析出来形成一系列词(Term)和一个查询树。通过索引存储将索引读入到内存，利用查询树搜索索引，得到每个词(Term)的文档链表，在对文档链表进行交，差运算后到结果文档（Hits）。将搜索到的结果文档按照查询的相关性进行排序并返回。

通过索引和搜索过程的分析，可知为了适应文本分类系统，需要做如下工作：

1) 对文本分析引擎进行中文分词扩展，从而更好地支持中文文本分析处理。

2) 自定义查询引擎相关索引信息并创建文本索引，通过自定义域值信息指定索引中要包括的文本的相关信息。

3) 自定义查询引擎中具体查询。实现具体查询过程，统计相关数据。

④ Lucene 源码结构

在此将结合 Lucene 索引、搜索过程和源码结构，介绍相关工作的具体实现。

如图 4.6 所示，Lucene 的系统由基础结构封装、索引核心、对外接口三大部分组成。索引核心包的代码涉及到索引文件的具体操作是索引引擎的实现的关键代码。Lucene 的将所有源码分为了 7 个包，各个包所属的系统部分也如上图所示。

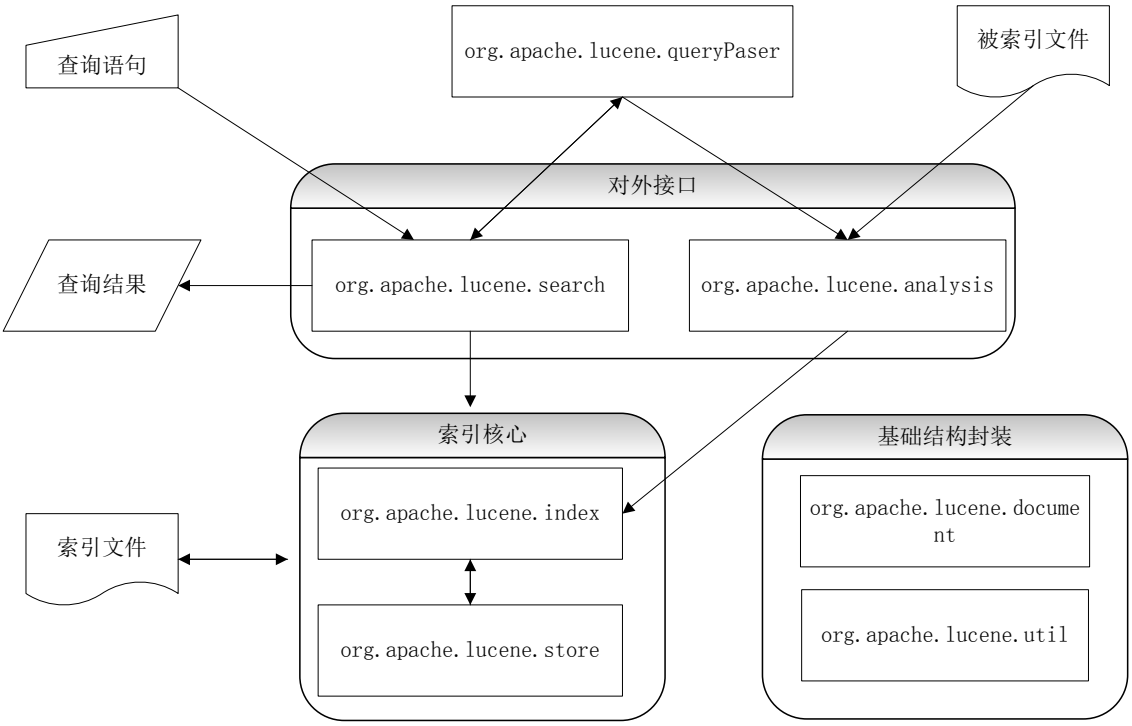


图 4.6 Lucene 源码结构图

Fig.4.6 Architecture of Lucene's Source Code

如上图所示，Lucene 源码中共包括 7 个子包，每个包完成特定的功能：

表 4.1 Lucene 包结构功能表

Table4.1 Function of Lucene's Packets

包名	功能描述
org.apache.lucene.analysis	语言分析器，主要用于的切词，支持中文主要是扩展此类
org.apache.lucene.document	索引存储时的文档结构管理
org.apache.lucene.index	索引管理，包括索引建立、删除等
org.apache.lucene.queryParser	查询分析器，实现查询关键词间的运算
org.apache.lucene.search	检索管理，根据查询条件，检索得到结果
org.apache.lucene.store	数据存储管理，主要包括一些底层的 I/O 操作
org.apache.lucene.util	工具包，包括高亮关键字等

Lucene 的核心类包中 org.apache.lucene.analysis 是分析引擎的具体代码实现，org.apache.lucene.index 是索引引擎的具体代码实现， org.apache.lucene.search 是查询引擎的具体代码实现。

⑤ 解决方案

在前面分析中，我们已经明确了需要解决的具体工作内容，现在将具体制定对相应的工作制定措施。

1) 文本分析引擎进行中文分词扩展的具体实现思路

Lucene 设计了独立于语言和文件格式的文本分析接口，索引器通过接受 Token 流完成索引文件的创立，用户扩展新的语言和文件格式，只需对 org.apache.lucene.analysis 进行扩展。

在本文 2.3.1 节中对中文分词的详细分析表明中文分词具有相当的复杂性，因此考虑自己研究中文分词问题很大程度上无法满足中文分词准确率的需求，容易对后续分类实验造成严重影响，不能准确判断提出的特征提取方法的有效性和可行性。因此我们使用现今成熟的开源工具包汉语语法分析系统 ICTCLAS 2011 版。

ICTCLAS 对中文分词的处理是在中文词条之间加入分隔符，并且在分隔符后加上词性标注，使之转化为以词为单位的形式。通过词性标注，我们可以方便选取具有实际意义的名词等，过滤无用的虚词等，实现对文本特征空间的初降维。例如，句子“我今天拿到礼物好开心”，经过 ICTCLAS 分词后得到：“我/r 今天/t 拿/v 到/v 礼物/n 好/a 开心/a”，每个词后面都带有词性标注，/r 表示代词，/t 表示时间，/v 表示动词，/n 表示名词，/a 表示形容词。词性标注方便对词条的选择操作，本文采用正则表达式完成对特定词性词条的选择。

综上所述运用该系统工具包进行中文分词处理可以满足稳定性，准确性，有效性等需求，能够很好地解决系统中的中文分词难题。最后整合 ICTCLAS 自定义 Analyzer，解决文本分析引擎中文文本分词问题。

2) 自定义查询引擎相关索引信息并创建文本索引的具体实现思路

自定义索引中文档结构信息，通过对 org.apache.lucene.document 进行扩展，制定构成文档的域信息。考虑到对于每个特征项需要统计如下数据：

a. 特征词出现次数相关统计。特征词在当前文本中出现的次数，特征词在当前类别内出现的次数，特征词在整个数据集合中出现的次数。

b. 包含特征词文本数的相关统计。当前类别内包含特征词的文本数，文本集合中包含特征词的类别数。

c. 包含特征词类别数的相关统计。文本集合中包含特征词的类别数。

因此设计每个文档将包含三种类型的域，域名分别为：文件名，内容，文件类别，所有特征词将被表示成 Term 对象包含在“内容”域中。

3) 自定义查询引擎中具体查询，实现具体查询过程，统计相关数据

Lucene 的查询实现中默认实现了布尔操作、模糊查询(FuzzySearch)、分组查询等操作。实现具体查询过程，通过对 org.apache.lucene.search 进行扩展，查询语句的自定义通过自定义 Term 对象，经过查询获取包含 Document 对象的 Hits 对象等，通过统计 Hits 对象中包含的所与 Document 对象信息，可完成对包含某个特征词的文档数，包含某个特征词的类别数等统计，从而完成特征提取和特征加权等技术环节。

4.1.4 系统整体设计

为了保证实验的顺利进行，本人开发了一套基于开源全文检索引擎 Lucene 和汉语词法分析系统 ICTCLAS 中文文本分类系统。该系统包括文本预处理、特征词提取、特征加权、分类器训练和分类性能评估等功能模块。各个模块之间相互独立，便于实验中各个部分的工作分别开展。其中文本预处理模块使用汉语词法分析系统(Institute of Computing Technology,Chinese Lexical Analysis System)，简称 ICTCLAS，来解决中文分词问题。特征提取和特征加权模块使用开源全文检索引擎 Lucene 进行特征词提取和特征加权工作。中文文本分类系统结构如图 4.1 所示：

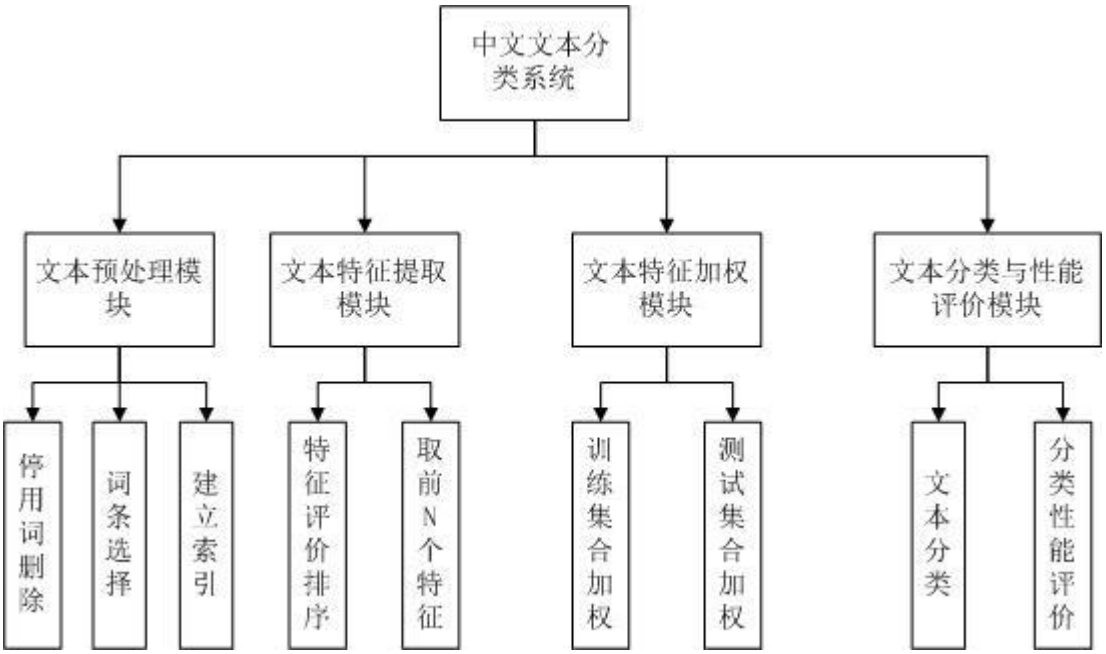


图 4.7　中文文本分类系统结构

Fig.4.7　Chinese text categorization system architecture

由图 4.7 可以看出，中文文本分类系统主要分为四大模块文本预处理模块，文本特征提取模块，文本特征加权模块，文本分类与性能评价模块。其中文本预处

理模块包括了停用词删除，词条选择，建立索引三个部分。特征提取包括特征评价排序和取前 N 个特征两个部分。文本特征加权模块包括训练集合加权和测试集合加权两个部分。文本特征与性能评价模块包括文本分类和分类性能评价两个部分。具体四大模块的工作流程如图 4.8 所示：

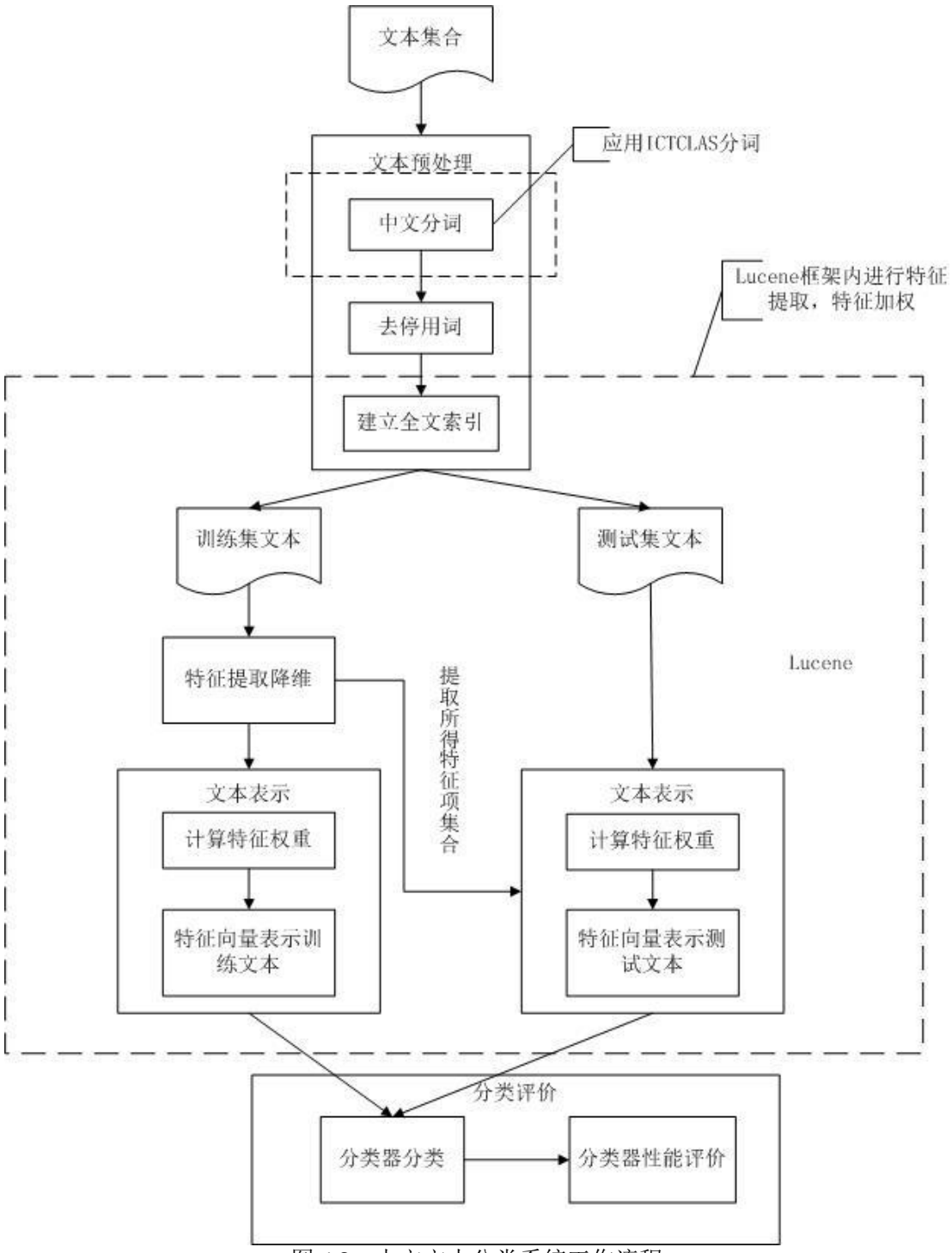


图 4.8　中文文本分类系统工作流程

Fig.4.8　Chinese text categorization experiment system work flow

本文的中文文本分类系统工作的流程如下：

1）进行文本预处理，通过 Lucene 对训练文本集合和测试文本集合分别建立全文索引，获得训练集合索引和测试集合索引，索引文件存储所有特征词和文档的相关信息。在建立索引过程中，通过 ICTCLAS 对文本集合中的训练文本集合和测试文本集合进行分词，去除停用词，词条选择，分别获得训练、测试文本集合的原始特征词集合。

2）利用特征提取方法进行特征降维，通过自定义 Lucene 对索引的查询功能，对特征提取方法中评价函数所需数据进行统计，代入计算后，按特征评价值的高低进行降序排列，选取出前 N 个具有最好类别效果的特征，构成表示文本的特征向量中的特征项集合；

3）对训练集合和测试集合中的文本进行特征加权，由于训练集合和测试集合的文本需要由空间特征向量表示，特征向量中的特征项为文本特征提取处理的特征项集合。对个每个文本，通过自定义 Lucene 对索引的查询功能，获得特征项对应该文本和相应类别中的分布情况统计，将统计的数据代入特征加权公式，完成空间特征向量对每个文本的表示。

4）进行分类和分类性能的评价，将表示训练集合和文本集合中所有文本的空间特征向量作为输入数据，运行相应的分类算法，对测试集中的文本进行分类。完成分类后，利用性能评价指标对分类结果进行评价，本文不仅将评价结果作为分类效果好坏的依据，还根据评价结果的好坏对文本分类系统前几个步骤的参数进行调整，提高分类系统的性能。

4.2 中文文本分类系统模块设计

4.2.1 文本预处理模块设计

用户可以通过文本预处理模块对训练文本以及测试文本进行预处理。文本预处理在本系统中实际上是对训练文本和测试文本建立全文索引的过程。详细设计表如下：

表 4.2　文本预处理模块详细设计表

Table 4.2　Detailed Design of Text Pretreatment module

模块名称：文本预处理模块		**使用者：用户**
基本流一：训练文本或测试文本集合建立索引		
输入部分　I	处理描述　P	输出部分　O
1.在主页面点击“建立文本索引“按钮。 2.输入建立索引所需相关信息。 3.点击“建立索引”。	1.读入数据 2.进行建立索引处理，首先进行分词处理，再进行词条选择和停用词删除，对剩余的词条和文本建立索引关系。 3.显示建立索引的结果	1.索引建立成功后结果显示栏中输出提示信息。 2．输出了索引文件

流程图：

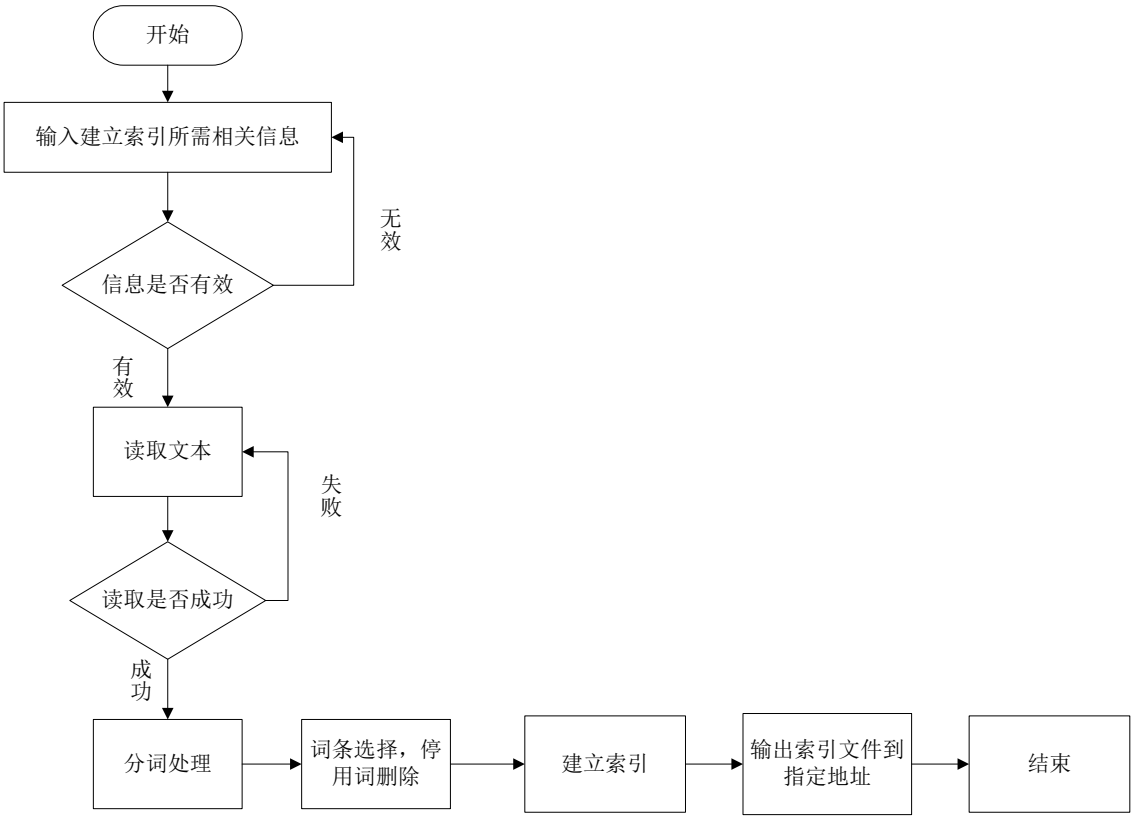


图 4.9　文本预处理模块流程图

Fig.4.9　Flow Diagram of Text Pretreatment module

4.2.2 文本特征提取模块设计

文本特征提取需要利用特征评价函数对每个特征进行打分，本文中采用的评价函数基于数理统计理论。文本特征提取模块中，用户通过导入训练文本集合的索引文件，通过查询得到特征词与文本、类别之间的关系映射值，代入特征评价函数为每个特征词打分，当为所有特征词打分之后，按得分对特征词进行排序，导出存储已排好序的特征词。因实验中需要对 N 值进行调整，观察实验结果，所以导出存储已排好序的特征词将避免重复的评价得分计算，减少运算复杂度和时间。用户通过指定提取特征词数 N，导入已排好序的特征词，取前 N 个特征词作为表示文本的特征向量中的特征项。文本预处理在本系统中实际上是对训练文本和测试文本建立全文索引的过程。详细设计表如下：

表 4.3　文本特征提取模块详细设计表

Table 4.3　Detailed Design of Text Feature Selection module

模块名称：文本特征提取模块		**使用者：用户**
基本流一：文本特征词评价及排序		
输入部分　I	处理描述　P	输出部分　O
1.在主页面点击“特征评价及排序“按钮。 2.输入特征评价及排序所需相关信息。 3.点击“开始评价及排序”。	1.读入索引文件，对每个特征词进行打分。 2.按分数对特征词进行排序 3.将已排好序的特征词写入文本中，然后导出。	1.操作成功后结果显示栏中输出提示信息。 2. 输出了包含已拍好序的特征词的文本。
基本流二：取前 N 个文本特征词		
输入部分　I	处理描述　P	输出部分　O
1.在主页面点击“取前 N 个文本特征“按钮。 2.输入取前 N 个文本特征所需相关信息 3.点击“开始提取”。	1.读入包含已排序的特征词文本。 2.将前 N 个特征词写入新文本 3.导出新文本	1.操作成功后结果显示栏中输出提示信息。 2．输出了包含前 N 个特征词的文本。

特征词评价并排序的流程图：

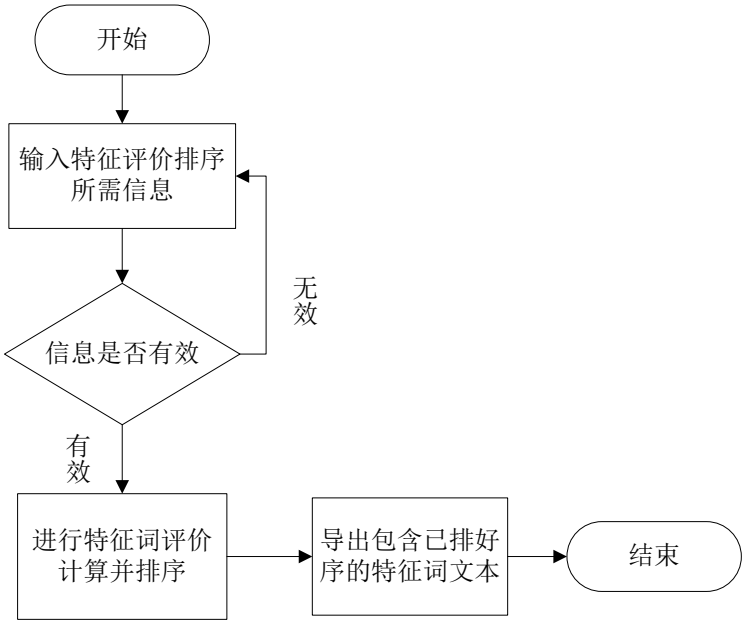


图 4.10 特征词评价及排序流程图

Fig.4.10 Flow Diagram of Feature Evaluation and Sort

选择前 N 个特征词：

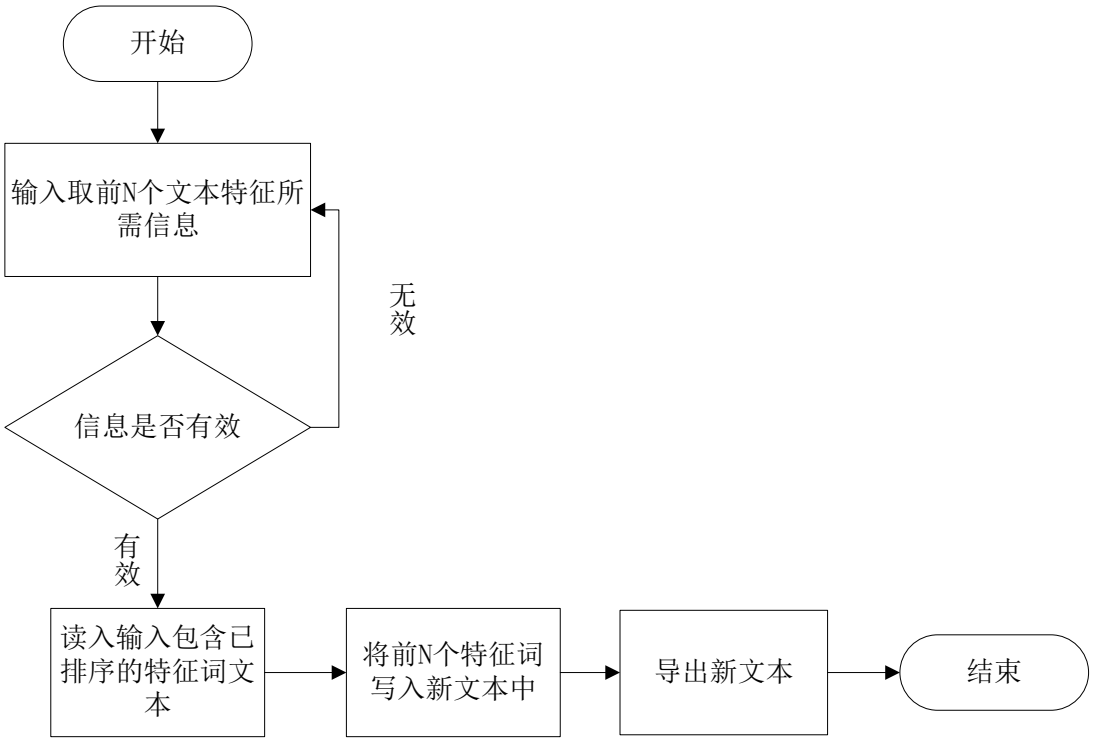


图 4.11 取前 N 个特征词流程图

Fig.4.1 Flow Diagram of Top N feature Selection

4.2.3 文本特征加权模块设计

文本特征加权模块为训练文本和测试文本加权而设计，文本特征加权模块按照特征加权方法为文本集合中的所有文本一一加权，使得每个文本能被唯一的特征向量表示。表示每个文本的特征向量将作为文本分类的输入数据。详细设计表如下：

表 4.4　文本特征加权模块详细设计表

Table 4.4　Detailed Design of Text Feature Weighted module

模块名称：文本特征加权模块		**使用者：用户**
基本流一：训练、测试文本集合中的文本加权		
输入部分　I	处理描述　P	输出部分　O
1.在主页面点击“文本特征加权“按钮。 2. 分别文本特征加权所需相关信息， 3.点击“开始加权”。	1.分别针对训练、测试集合中的每个文本进行权重计算方法。 2. 将训练、测试集合得出的权重分别以矩阵形式写入新文本中，然后导出到各自的权重文件存放地址。	1.操作成功后结果显示栏中输出提示信息。 2. 输出了包含训练集合文本权重的文本和保护测试集合文本权重的文本。

权重计算流程图：

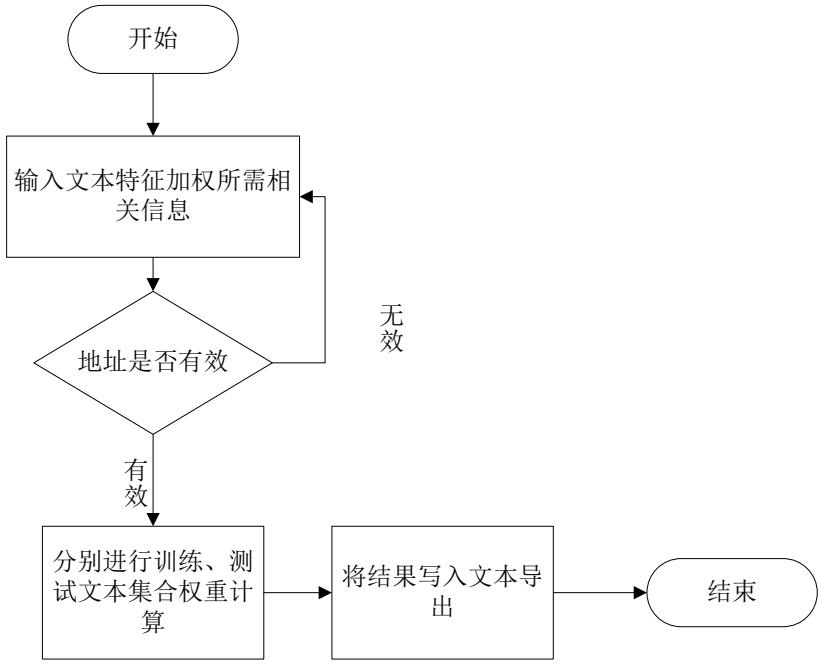


图 4.12　文本特征加权流程图

Fig.4.12　Flow Diagram of Text Featue Weighted

4.2.4 文本分类及性能评价模块设计

文本分类及性能评价模块以训练文本集合权重文件（包含训练文本集合类别信息）、测试文本集合中权重文件（包含测试文本集合类别信息）和 K 值为输入，在此说明训练和测试文本集合权重文件都是以矩阵方式表示，每行第二列至最后一列代表一个文本，每行第一列表示该文本对应的类别信息。分类结果计算出来以后，通过将分类结果与正确的类别信息进行一一对比，统计四个分类结果数据元素 TP、FP、TN 和 FN。然后进行文本分类准确率，查准率，查全率，F1 值以及宏平均值计算，因实验数据对比所需，最后将数据导出存储。详细设计图如下：

表 4.5　文本分类及性能评价模块详细设计表

Table 4.5　Detailed Design of Text Classification and Result Evaluation module

模块名称：文本分类模块		**使用者：用户**
基本流一：文本分类		
输入部分　I	处理描述　P	输出部分　O
1.在主页面点击“文本分类及性能评价“按钮。 2. 分别输入文本分类及性能评价所需相关信息 3.点击“开始分类及评价”。	1.读取全部文件 2.运行分类算法 3. 将分类结果与测试文本正确类别信息进行比较统计相关数据，进行分类性能评价。 4、根据不同 K 值循环运行 2-3 步，选择效果最好的一组数据，并导出结果。	1.操作成功后结果显示栏中输出提示信息 2. 输出了包含分类性能评价数据的文本。

文本分类流程图：

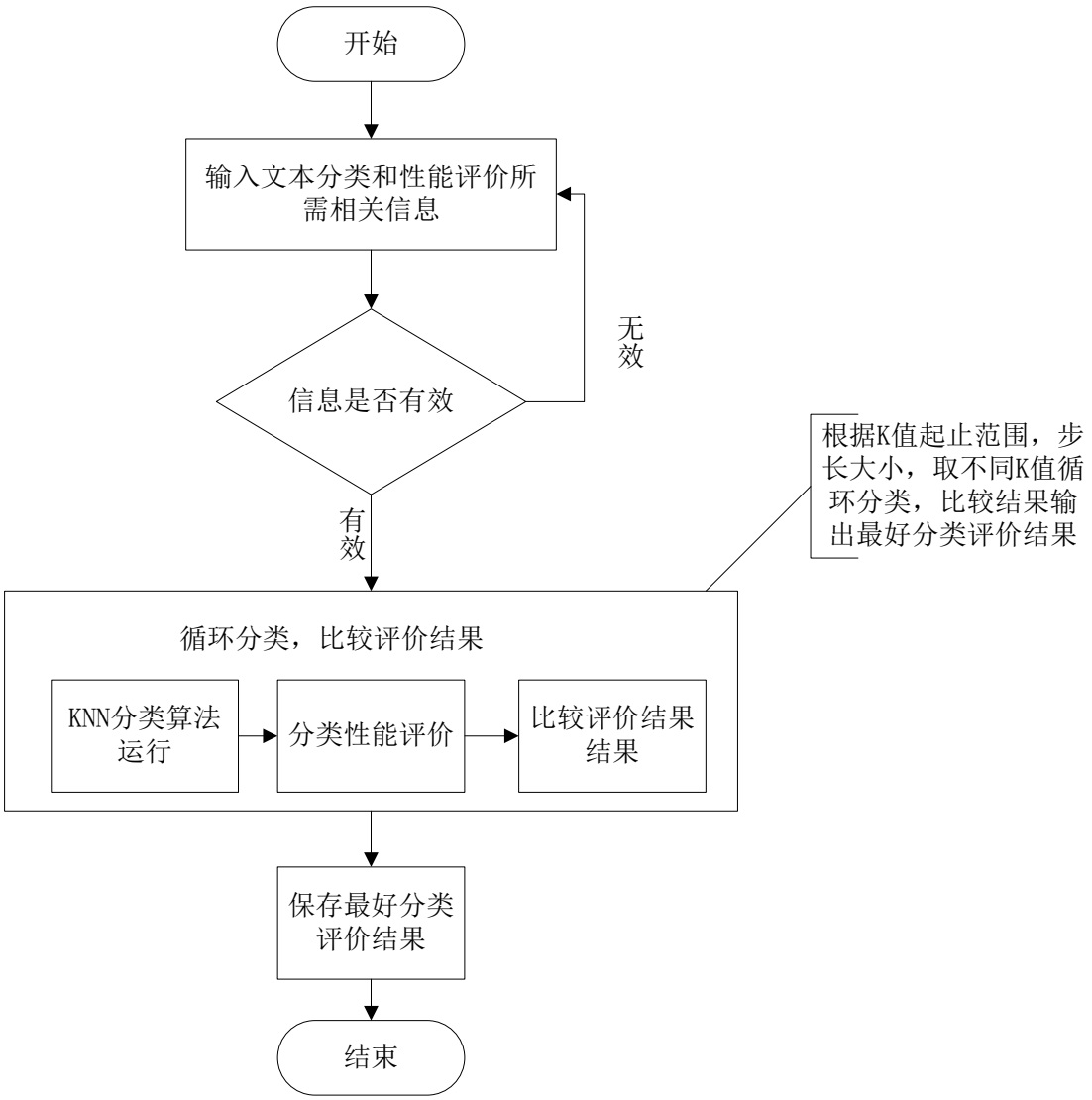


图 4.13　文本分类及性能评价流程图

Fig.4.13　Flow Diagram of Text Classification and Result Evaluation

4.3 中文文本分类系统的实现

本系统采用开源汉语词法分析系统 ICTCLAS 2011 工具包结合开源全文检索引擎 Lucene 工具包来实现特征预处理、特征提取和特征加权三大模块的功能。其中 ICTCLAS 主要帮助 Lucene 解决在完成对训练、测试集合文本建立索引过程中中文分词的问题。Lucene 完成对训练、测试文本集合建立索引后，我们通过自定义 Lucene 查询功能，统计特征提取与文本加权过程中所需的相关信息，完成特征提取与文本加权功能。

4.3.1 文本预处理模块实现

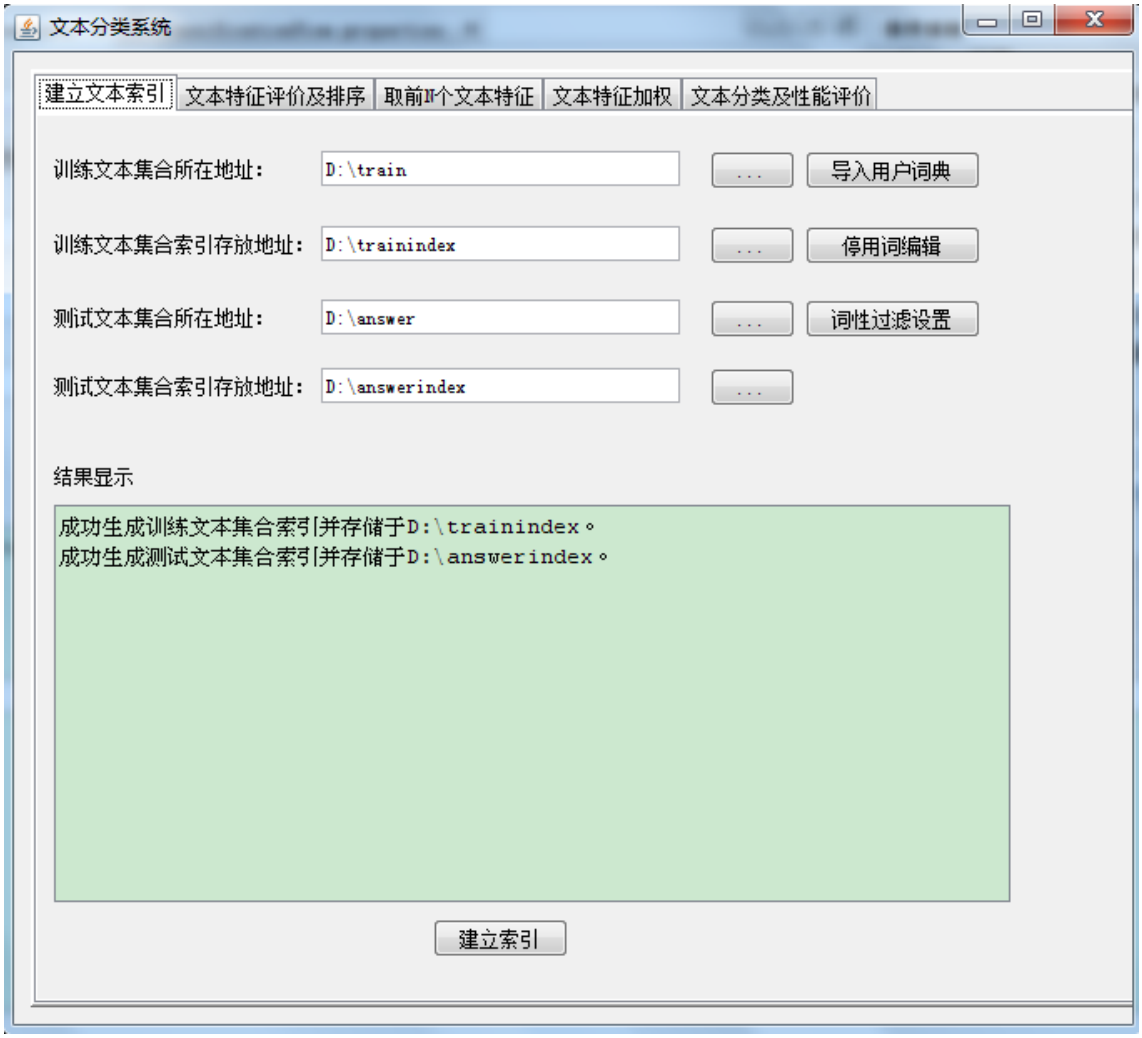


图 4.14　建立文本索引的页面

Fig.4.14　Page of Text Index creation

图 4.15　索引文件

Fig.4.15　Index's File

如图 4.14 所示，文本预处理模块的功能是在建立文本索引这个页面完成的。通过导入用户词典帮助提高分词效果，通过停用词编辑自定义停用词，达到有效删除停用词的目的。通过词性过滤设置有效过滤无实际意义的虚词等。在输入训练、测试文本集合所在地址，训练测试文本集合存放地址后进行索引建立，在索引建立的过程中将使用用户词典，用户自定的停用词和词性过滤设置进行分词，停用删除，以及词条选择，最后进行索引建立，建立好的索引中每个文档将不包括停用词，过滤掉的虚词等。

索引创建成功后，索引文件如图 4.5 所示。

4.3.2 文本特征提取模块实现

文本特征提取模块包含文本特征评价及排序和取前 N 个文本特征数两个部分。首先通过文本特征评价及排序部分对所有文本特征进行评价打分，并对评价结果进行排序。这样做的目的是为了提高实验的效率，由于实验中需要对 N 取不同值，若不保存评价排序结果，则每次 N 值变化需要重新评价排序，耗费时间做重复工作。文本特征评价及排序部分如图 4.16 所示：

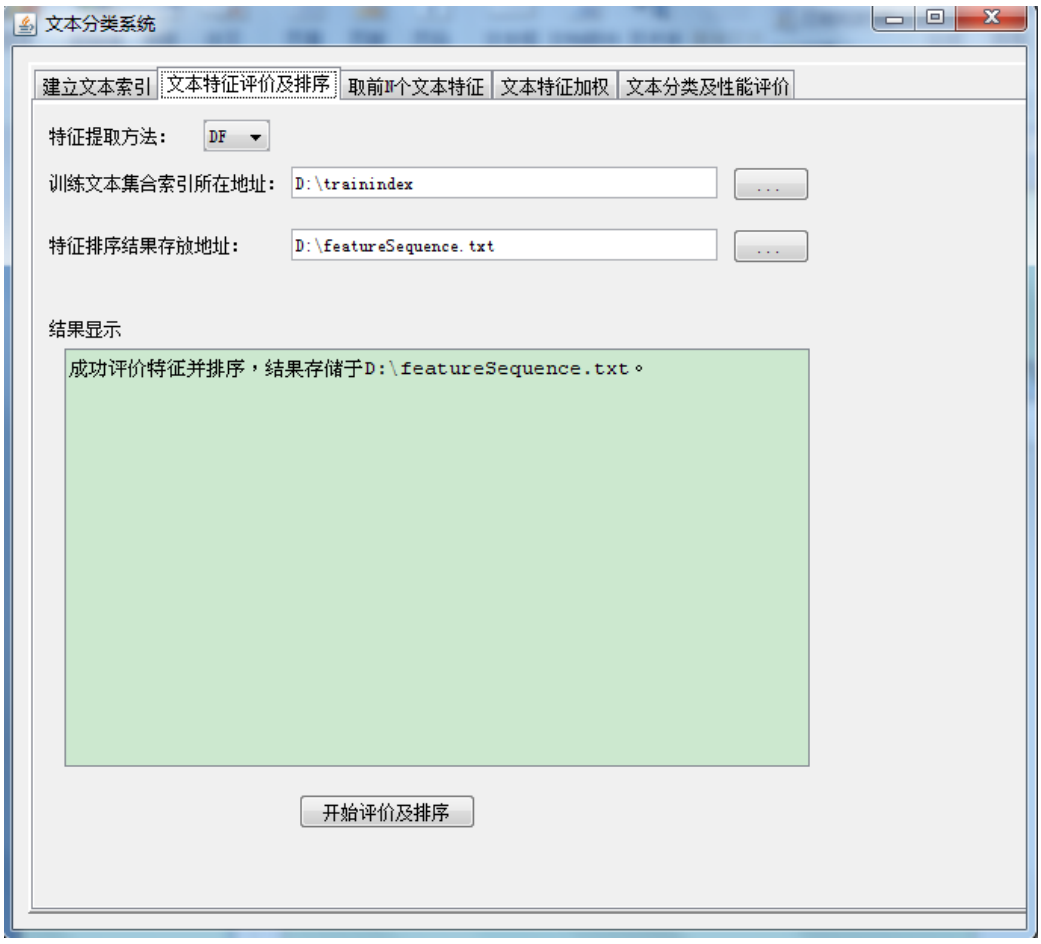


图 4.16　文本特征评价及排序界面

Fig.4.16　Page Of Feature Evaluation and Sort

如图 4.16 所示，在特征提取方法选择，输入训练文本集合索引所在地址以及特征排序结果存放地址过后，点击“开始评价及排序”按钮后就开始进行特征评价及排序操作，通过索引文件读取训练文本集合中包含的特征词，然后对每个特征词按照特征片提取方法进行评价打分，评价过程中评价函数计算涉及的相关数据统计通过重写 Lucene 的查询功能得到，然后对最后的评价结果按分数从高到底进行排序，并且将结果存储以便后续取前 N 个文本特征数使用，如图 4.17 所示：

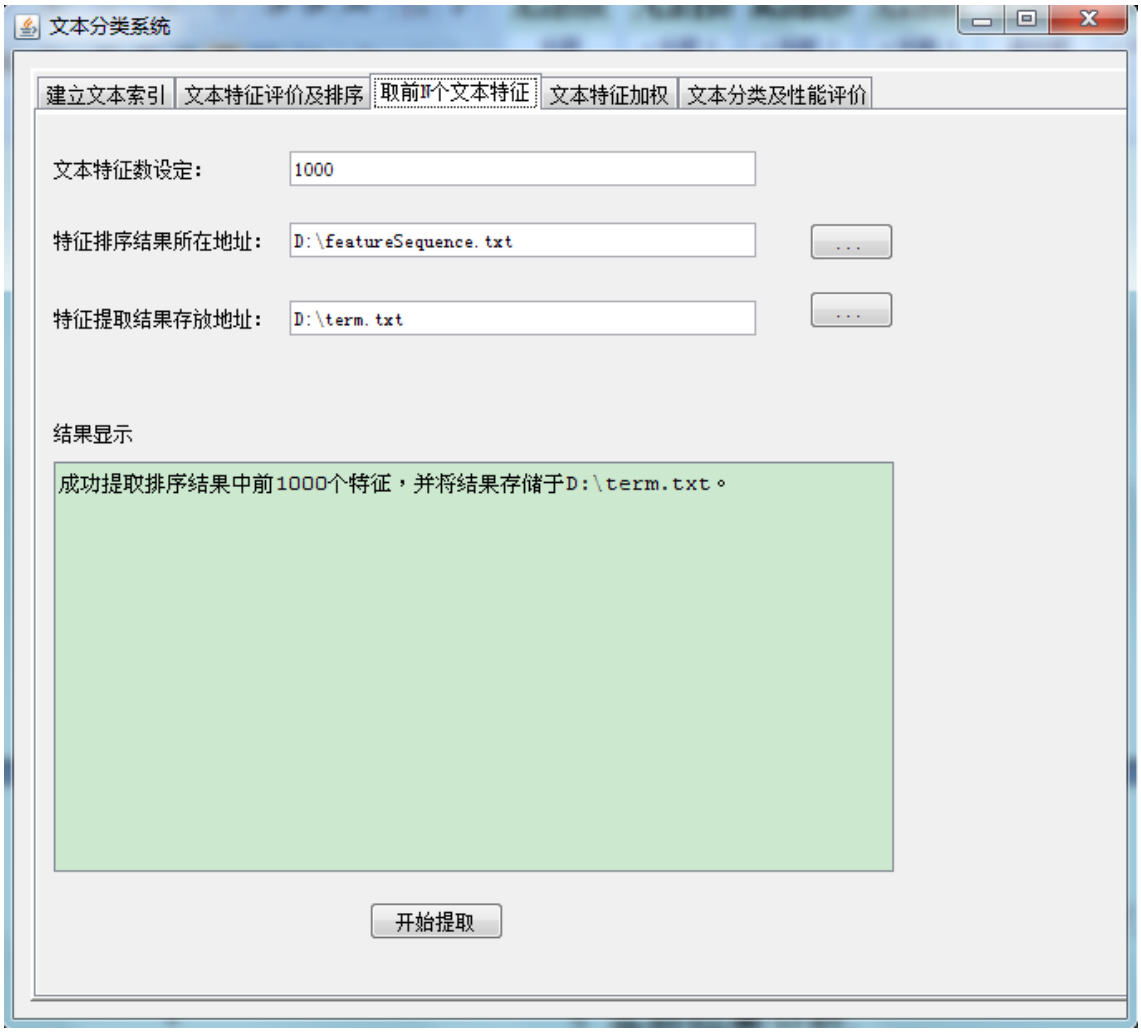


图 4.17　取前 N 个文本特征的页面

Fig.4.17　Page of Top N feature Selection

如图 4.17 所示，通过导入文本特征评价排序结果，设置要提取的文本特征数，输入特征提取结果存放地址，点击“开始提取”，便完成取文本特征评价排序结果中分数最高的前 N 个文本特征并存储。提取结果如图 4.18 所示：

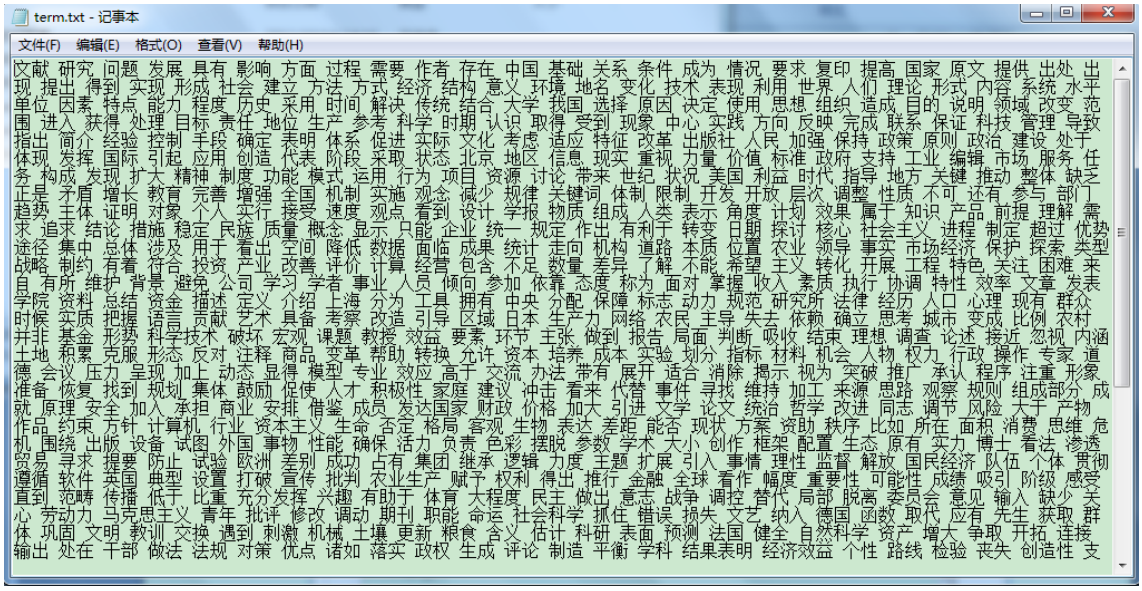


图 4.18　文本特征提取结果

Fig.4.18　Result of Text Feature Selection

4.3.3 文本特征加权模块实现

文本特征加权模块中，需要利用文本特征提取模块获得的特征词作为特征向量特征项来表示文本，文本特征加权就是计算每个特征词在每个文本中的权重，通过各个特征词权重的不同最大程度上区分不同类别的文本，以便进行分类。在该模块中，需要对训练文本集合和测试文本集合都进行权重计算。文本特征加权界面如图 4.19 所示：

图 4.19　文本特征加权界面

Fig.4.19　Page of Text Feature Weighted

如图 4.19 所示，通过选择特征加权方法，输入训练、测试文本集合的索引所在地址和权重存放地址，输入特征提取结果所在地址，点击“开始加权”按钮表开始文本权重计算，权重计算过程涉及的相关数据统计通过重写 Lucene 的查询功能得到。最后保存训练、测试文本集合的权重文件，权重文件中矩阵第一列包含了每个文本所属类别信息。

4.3.4 文本分类及性能评价模块实现

文本分类及性能评价模块包括了文本分类和性能评价两个部分。该功能模块中目前设定文本分类算法为 KNN 算法，性能评价通过结果利用查全率、查准率、正确率、F1 值和宏平均等指标加以衡量。文本分类及性能评价界面如图 4.20 所示：

图 4.20　分类与性能评价页面

Fig.4.20　Page of Text Classification and Result Evaluation

如图 4.20 过导入训练、测试文本集合的权重文件，输入分类评价结果存放地址，设置 K 值起止范围和步长，点击“开始分类及评价”按钮，便开始进行文本分类运算，KNN 运算过程中，由于需要计算训练文本集合与当前未知文本相似度或者距离，并按照相似度和距离进行升序排序，故在排序后进行 K 值选择性能最高。因此根据不同的 K 值，循环对排序结果进行选择，并通过最后分类效果比较觉得 K 值的选择。最后对分类性能进行评价，评价结果如结果提示栏中所表示。

4.4 本章小结

本章首先基于文本分类过程各个部分的研究，明确提出中文文本分类系统的要解决的问题，并通过 Lucene 和 ICTCLAS 的研究，对中文文本分词和大量文本数据统计两个待解决的关键问题提出解决方案。然后对系统进行整体设计和模块设计，最终实现中文文本分类系统。

5 实验结果分析

5.1 实验介绍

为了验证在第三章中提出的 SP 特征提取方法的可行性和有效性。本文采用通用且性能较好的 KNN 分类算法对 SP 和 CHI，CC，DF 等常用体征提取方法进行了多组对比实验。文本分类的实验结果利用查全率、查准率、正确率、F1 值和宏平均等指标加以衡量。

实验工作：

① 解决 KNN 分类算法中 K 取值问题。实验通过选定一种特征提取方法，表明 KNN 在取不用 K 值的情况下，分类准确率有较大影响，通过动态改变 K 值进行 KNN 分类，根据分类效果的好坏来决定采用适合的 K 值，实验还证明不同特征提取方法通过 KNN 取得最优值时 K 值也不同，需要区别对待，分别求对应最优 K 值。

② 解决特征提取中该提取多少文本特征数 N 的问题。实验通过选定一种特征提取方法，通过取不同的 N 值，并观察分类准确率的变化，得出使用不同的特征提取方法，当其达到分类准确率最优时，N 的取值是不同的。

③ 通过上述 1、2 实验的贡献，我们为试验中每种特征提取方法指定最优 K 值和最优 N 值，以使得实验中每种特征提取方法在分类环节都能达到性能最优，体现出实验结果可信以及可比性。在此基础上，重点对比 SP 特征提取方法和常用的特征提取方法的优劣，实验结果表明 SP 特征提取方法性能优于常用方法

5.1.1 实验数据集

本实验采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组的文本分类语料库[40]，该语料库共分为 20 个类别，测试集共 9804 篇文章，训练集 9804 篇文章。我们从中抽取九个类别作为训练集，包括艺术，历史，航空，计算机，环境，农业，经济，政治，体育。在原训练集中每个类别随机抽取 200 个训练文档，共计 1800 个训练文档。在原测试集中每个类别随机抽取 100 个测试文档，共计 900 个测试文档。训练文档和测试文档的比例为 2:1，通过预处理，测试文档包含 57771 个特征词。

表 5.1 实验数据介绍

Table 5.1 Data set of experiment

数据类型	艺术	历史	航空	计算机	环境	农业	经济	政治	体育
训练集文档数（篇）	200	200	200	200	200	200	200	200	200
测试集文档数（篇）	100	100	100	100	100	100	100	100	100

5.1.2 实验参数设定

5.2 实验结果及分析

本实验选取目前常用的 CHI、CC 和 DF 等方法与本文提出的 SP 特征提取方法作比较。本次实验中采用的特征加权方法为 TF-IDF，分类方法为 KNN。由于 KNN 算法不易确定，因此我们将通过一组实验数据来表明当特征提取方法为 CHI，提取特征为 100 维，K 取不同值时，分类系统分类准确率的波动情况。如图 5.2 所示：

表 5.2 当特征提取方法为 CHI 时，KNN 在不同 K 值情况下的分类效果(%)

Table 5.2 When Feature Selection method is CHI, result of classification by different parameter of K(%)

分类效果	K=5	K=8	K=11	K=14	K=17	K=20
分类准确率	83	83.22	83.22	82.56	82.22	82.44

表 5.3 当特征提取方法为 CC 时，KNN 在不同 K 值情况下的分类效果(%)

Table5.3 When Feature Selection method is CC, result of classification by different parameter of K(%)

分类效果	K=5	K=8	K=10	K=14	K=17	K=20
分类准确率	81.78	82.33	82.67	82.56	81.89	82.22

表 5.4 当特征提取方法为 DF 时，KNN 在不同 K 值情况下的分类效果(%)

Table5.4 When Feature Selection method is DF, result of classification by different parameter of K(%)

分类效果	K=6	K=8	K=11	K=14	K=17	K=20
分类准确率	72	71.11	70.11	69.67	68.67	68.22

表 5.5　当特征提取方法为 SP 时，KNN 在不同 K 值情况下的分类效果(%)

Table5.5　When Feature Selection method is SP, result of classification by different parameter of K(%)

分类效果	K=6	K=8	K=11	K=14	K=17	K=20
分类准确率	85.33	84.11	84	83.78	83.22	83.67

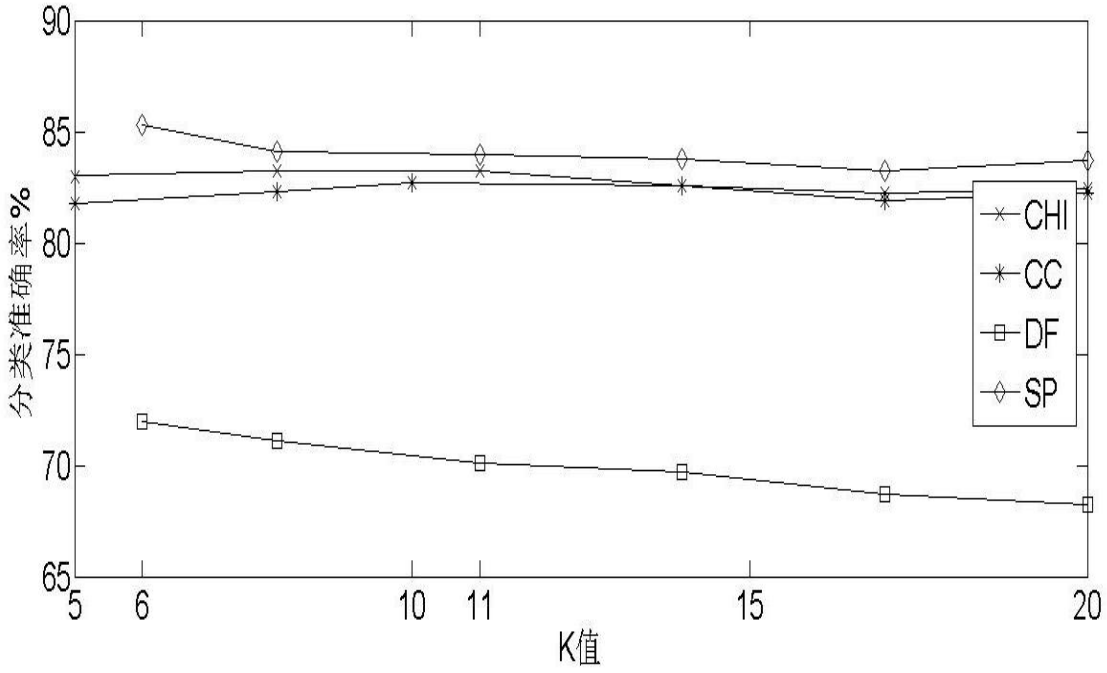


图 5.2　选用不同特征提取方法，KNN 在不同 K 值情况下的分类效果(%)

Fig.5.2　Using different feature selection method to show the result of classification by different parameter of K(%)

观察以上图表我们可以发现，当 K=6，6，8，10 时，将 DF，SP，CHI，CC 作为特征提取方法的分类系统的分类准确率分别达到最高，达到最佳效果。显然不同特征提取方法取得分类准确率最高时，K 值是不同的。并且当 K 取值超过最优 K 值时，分类准确率呈下降趋势。因此我们可以得知 K 值具有不易确定性，因此我们只能根据实验结果来调整 K 值，保证分类效果。具体措施是在文本分类过程中我们首先计算出待分类文本与所有测试文本的相似度，然后对相似度进行排序，取排列前 K 个最相似的测试文本进行类别统计，根据类别统计的结果决定待分类文本属于哪个类别，所以 K 值的大小并不影响分类算法的性能。因此我们可以为 K 值赋一个初值，一个最大值，同时规定 K 值在初值与最大值之间增加的步长。通过不同 K 值取得分类准确率的比较，来确定最优 K 值。

基于以上措施对取得最优 K 值的保证，我们开始利用 CHI、CC、DF 特征提

取方法与 SP 方法进行比较。在实验过程中，我们需要利用特征提取方法对各个特征词进行计算，将特征词按照得分大小进行降序排序，提取前 N 个特征词作为特征向量中的特征项。由于无法确定 N 值为多少时，分类效果最好，于是以 CHI 特征提取方法为例，我们对 N 值取值不用会对分类结果产生什么样的影响做了一下研究，如表 5.6 所示：

表 5.6 不同 N 值下的分类效果比较

Table5.6 Comparison of Macro-F1 with different N

特征词个数 N	分类准确率（%）	分类准确率增幅（%）
100	83.56	——
200	86.11	2.55
300	85.44	-0.67
400	87.33	1.89
500	87.44	0.11
1000	89	1.56
1500	89.44	0.44
2000	88.67	-0.77
2500	89	0.33
3000	87.89	-1.02

由表 6.1 知，特征词从 100 维到 1000 维，分类准确率值的上升幅度都大于 1.5%，到了 1500 维，上升幅度下降至 0.44，到了 3000 维甚至出现分类准确率下降的情况。由此说明，在一定范围内，随着特征向量维数的增大，分类效果会逐渐提高，但是当维数超过一定数值时，将趋于平稳，甚至出现下降，同时维数的增加不免付出计算代价。因此，从分类效率和分类精度二者进行综合考虑，我们对 CHI 方法取 N=1000。

接下来我们进一步通过实验表明，不同的特征提取方法对向量空间的降维能力是有强弱区别。

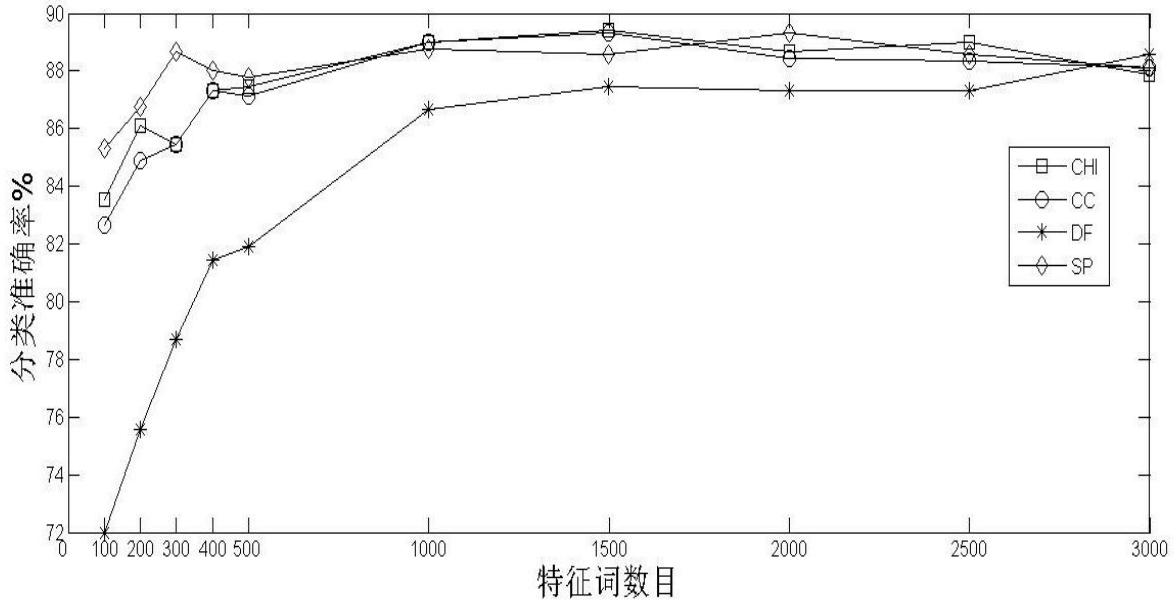


图 5.3 分类准确率效果图

Fig.5.3 Accuracy of Classification

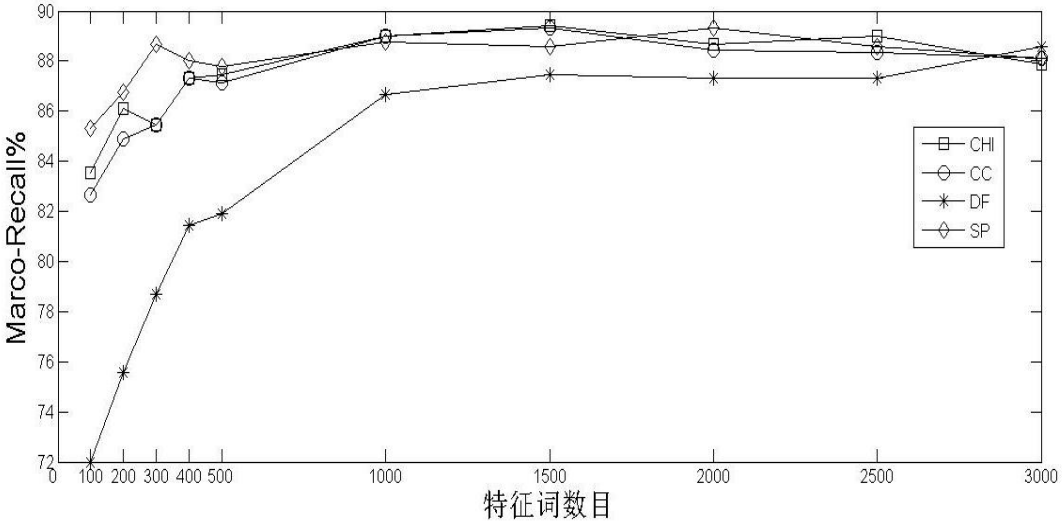


图 5.4 Marco-Recall 对比

Fig.5.4 Comparison results of Marco-Recall

如图 5.3 所示，CC 和 CHI 在提取的文本特征向量维数为 1500 维时，分类效果最优，SP 在提取的文本特征向量维数为 1500 维时，分类效果最优，DF 在提取的文本特征向量维数为 3000 维时，分类效果最优。通过观察，我们可知对 CHI、CC、DF、SP 来说，在一定范围内，随着特征向量维数的增大，分类效果都会逐渐提高，但是当维数超过一定数值时，都会趋于平稳，甚至出现下降。图 5.3 进一步表明了不同的特征提取方法具有不同的降维效果，只有选择合适的特征向量维

度才能保证分类准确率最优，并且降低计算代价。

通过图 5.3 初步观察可知，SP 具有强降维能力以及与 CHI 相当的分类准率，我们将通过 Marco-Recall 、Marco-Precision、Marco-F1 数据对 CHI、CC、DF 与 SP 进行对比实验，来进一步验证 SP 具有强降维能力。

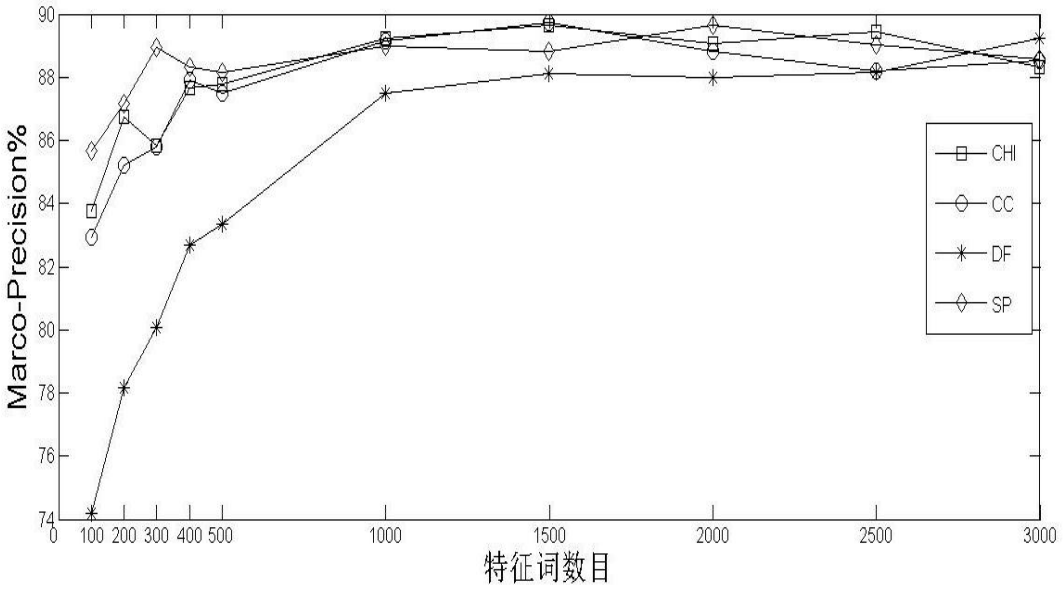


图 5.5 Marco-Precision 对比

Fig.5.5 Comparison results of Marco-Precision

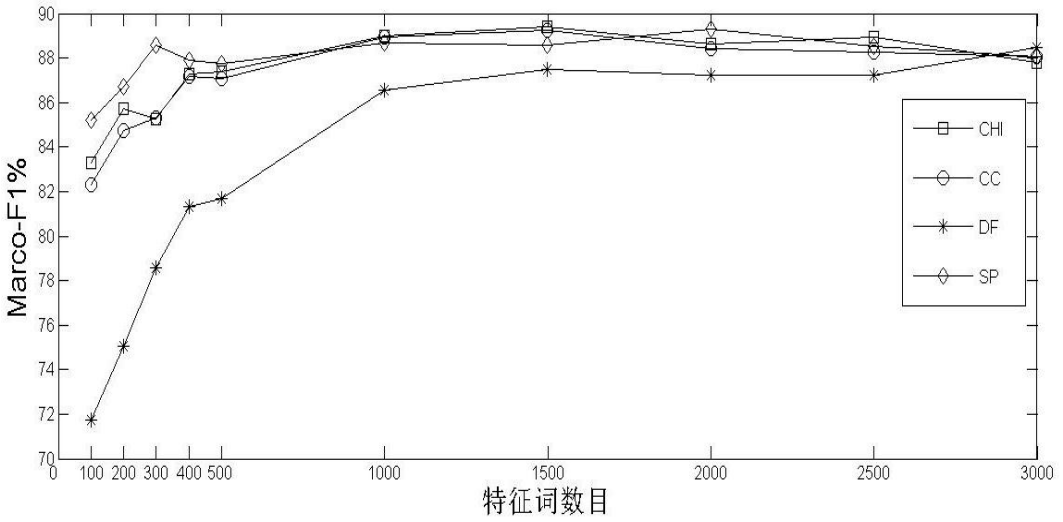


图 5.6 Marco-F1 对比

Fig.5.6 Comparison results of Marco-F1

如图 5.4-5.6 所示，CC 和 CHI 在提取的文本特征向量维数为 1500 维时，分类

效果最优， SP 在提取的文本特征向量维数为 1500 维时，分类效果最优，DF 在提取的文本特征向量维数为 3000 维时，分类效果最优。CC，CHI，SP 在分类效果达到最优时，Marco-Recall，Marco-Precision，Marco-F1 三项评价指标数值相当，DF 效果最差。

当提取的文本特征向量维数为 300 维时，SP 的分类效果远好于 CC、CHI 和 DF，此时 Marco-Recall= 0.8867，Marco-Precision= 0.8894，Marco-F1= 0.8858。通过观察我们可知，在一定范围内，随着文本特征向量维数的增加，分类效果会逐渐提高，但是当文本特征向量维数超过一定数值时将趋于平稳甚至下降，同时文本特征向量维数的增加不免付出计算代价。虽然 SP 在提取的文本特征向量维数为 2000 的时候，分类效果最好，此时 Marco-Recall= 0.8933，Marco-Precision= 0.8965，Marco-F1= 0.8932，但是通过与 SP 提取的文本特征向量维数为 300 维的分类效果作比较，发现其提高的分类性能很有限。由以上分析可知，SP 提取的文本特征向量维数为 300 维，即特征降维达到 99.48%时，已经使 KNN 具有一个良好的分类效果，可以有效对未知文本进行分类。

通过实验结果可知，SP 通过提取高质量的特征词，构造低维的特征向量，能够有效地降低特征空间维度，并且有效地表示各个类别的文本，反映了类别间的差异度。实验结果表明该方法应用在 KNN 分类算法上分类效果良好。

5.3 本章小结

本章首先在 4.1 节介绍了中文文本分类系统平台的实现方法，在 4.2 节中介绍了两个对比实验，用来验证本文在第三章中提出的新的文本分类特征提取方法 SP，在实验中，我们首先解决了 K 值的取值问题，以及特征词数目 N 取值问题，使得利用各个特征提取方法后，对应 KNN 的分类效果达到最好。在此基础上对 CDF 方法与其他常见的特征提取方法 CHI、CC、DF 进行对比实验，实验结果通过 Marco-Recall、Marco-Precision 和 Marco-F1 值等多项评价指标来衡量。结果表明本文提出一个新的基于类别正相关和类别强相关的新特征提取方法 SP，和传统方法相比，该方法能够充分利用特征项在文本集合中的分布统计信息选择与文本类别正相关并且强相关的特征。该方法在降维能力上有突出表现，实验中在保证良好分类效果的情况下，可对经过预处理后的特征向量进行高达 99.49%的降维。整体评价指标上表现出优于 DF 方法，降维能力、分类效果比传统特征提取方法 CHI，CC 性能上有所提升。

6 总结与展望

6.1 研究总结

随着社会的发展，我们处于一个信息爆炸的时代，文本分类技术能够更好地组织管理海量、异构的信息，并且从中快速，准确的查找自己所需的相关知识信息，帮助人们提高生活工作效率。

本文对文本自动分类过程包括的文本预处理、文本表示、文本特征降维、文本特征加权、分类方法的选择和分类性能评价五个部分内容进行详细的研究分析。研究分析结果表明文本自动分类大多建立在空间向量模型，“高维性“和”高稀疏性“成为了文本自动分类面临的两大难题，因此运用特征提取方法对向量空间进行降维成为文本分类关键的环节之一。特征提取作为特征降维的一种方法，成为本文研究的重点。

在对特征提取常用方法进行研究、对比、分析后，发现文本特征提取过程中负相关特征与弱相关特征对特征提取质量好坏产生很大的干扰。为了避免这种干扰。本文提出一个新的基于类别正相关和类别强相关的特征提取方法 SP (Strong Correlation and Positive Correlation，正相关与强相关)，SP 方法中正相关性因子通过区别特征与类别正负相关性，优先选择正相关特征，避免了负相关特征的干扰。强相关度因子通过区别特征与类别的强弱相关程度，优先选择强相关特征，避免了弱相关特征的干扰。SP 通过对两种因子的结合，能够有效地提取高质量的文本特征。

通过开发文本分类系统来验证 SP 方法的可行性与有效性。文本分类系统的开发建立在文本自动分类过程中关键技术的详细研究之上，根据研究结果明确了文本分类系统的需求，文本分类系统中各个关键模块的实现方法和文本分类系统实现过程中可能遇到的困难。将 ICTCLAS 与 Lucene 相结合作为中文文本分类系统搭建解决方案成功解决在文本分类系统中中文分词与大量文本数据统计两个难点。

在分类系统上对新特征提取方法与常用的特征提取方法进行了对比实验。最后通过实验结果可知，SP 通过提取高质量的特征词，构造低维的特征向量，能够有效地降低特征空间维度，并且有效地表示各个类别的文本，反映了类别间的差异度。实验结果表明该方法应用在 KNN 分类算法上分类效果良好。

6.2 下一步工作

文本分类是一项复杂的系统工程，文本分类技术，特别是中文文本分类技术

仍有待于进一步完善，目前仍有许多问题值得去研究和探索。下一步可以从以下几个方面进行研究：

① 文本分类数据集分为平衡数据集和偏斜数据集，进一步研究根据类别强弱相关度和类别正负相关两个指标如何需求更好的表现形式，使得特征提取方法适用于偏斜数据集合。

② 对文本特征加权进行深入研究，尝试通过类别强弱相关度和类别正负相关性两个指标，能否构建出文本特征加权公式。如果能够构建优于 TF-IDF 的特征加权公式，将在一定程度上大大提高文本分类的性能。

③ 通过语义信息角度出发，寻求一种能够有效表示文本的语义模型，使得模型能够完整包含文本的语言特向和潜在概念，然后通过与语义模型相对应的语义文本分类方法进行文本分类，相信能够在一定程度上提高文本分类系统的性能。

④ 对比国外文本分类研究，英文有具有多个公认、标准、开放的文本数据集合，可以方便且客观地评价研究者的工作成果，而中文目前还较缺乏此类公认、标准的文本数据集合，这也是以后将要研究的工作之一。

致 谢

在硕士论文即将完成之际，回顾硕士研究生学习以及论文写作期间。我得到了很多人的关心、鼓励和帮助，非常感谢他们。首先我要感谢我的导师杨丹教授。在本人硕士研究生期间，杨老师给予我细心的指导，帮助我打下了牢固的专业基础，扩展了我的学术眼界，注重培养我踏实的科研态度和创新的科研思维。本文在杨老师的关心帮助下顺利完成，在此向杨老师表示由衷的感谢！

另外要感谢徐玲讲师，她在对我的培养上也付出了很多心血，在硕士研究生学习期间，徐老师就注意引导我参与实际项目开发，帮助我积累实际项目经验。并且引导我参与实验室相关课题的研究和探讨，培养我学术研究的兴趣。在论文完成过程中徐老师也不迟辛劳给了我很多具体的指导，平常再累再忙只要我有学术方面的困扰，徐老师都会第一时间帮我解惑。在此向徐老师表示诚挚的谢意。

还要感谢行业信息化实验室的各位老师与同学。在实验室良好的学术氛围引导下，同学们、老师们互帮互助，相互照应，一人有难众人帮忙。与其说是在同一个实验室，不如说是在同一个家更为恰当。

从本科到研究生，七年的学习生涯，学院各位老师对本人的关怀无微不至，非常感谢各位老师，非常感谢学院对本人的培养。

在此特别感谢我的家人、朋友一直在我背后的默默支持，默默祝福。

最后衷心感谢在百忙之中评阅论文和参加答辩的各位专家。

林少波

二〇一一年十一月 于重庆

参考文献

[1] Sebastiani F. Machine learning in automated text categorization[C]. ACM Computing Surveys, 2002, 34(1): 1-47.

[2] H. L. Luhn. The Automatic Creation of Literature Abstracts[J]. IBM JOURNAL. 1958: 159-165.

[3] Maron, M. Automatic Indexing: An Experimental Inquiry[J]. Journal of the Association for Computing Machinery, 1961, 8(3): 404-417

[4] 成颖, 史九林. 自动分类研究现状与展望[J]. 情报学报, 1999, 1: 20-27

[5] 代六玲, 黄河燕, 陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报. 2004, 18(1): 26-32.

[6] 谭松波. 高性能文本分类算法研究[D]. 中国科学院计算技术研究所, 2006. 1.

[7] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006. 17(9): 1848-1859.

[8] Yiming Yang, Xin Liu. A re-examination of text categorization methods[J]. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999: 42~49.

[9] Hayes P. , Weinstein S. . Adding Value to Financial News by Computer[C]. In Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street. 1991: 2-8.

[10] 侯汉清. 分类法的发展趋势简论[M]. 北京: 中国人民大学出版社, 1981.

[11] 肖明, 沈英. 自动分类研究进展[J]. 现代图书情报技术, 2000. 5: 25-28.

[12] HATEHER E, GOSPODNETIE O. Lucene in Action[M]. Manning publications Co. , 2005.

[13] 刘群, 张华平, 俞鸿魁, 等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8): 1421-1429

[14] 徐燕, 李锦涛, 王斌, 等. 基于区分类别能力的高性能特征选择方法[J]. 软件学报, 2008, 19(1): 82-89

[15] Schubert Foo, Hui Li. Chinese word segmentation and its effect on information retrieval[J]. Information Processing and Management, 2004, 40(1): 161-190

[16] Christopher D. Manning, Hinrich Schutze. Foundation of Statistical Natural Language Processing[J]. Cambridge: MIT Perss, 2005: 225-250

[17] Ren Feiliang, Lv Xueqiang, Wu Honglin, Ma Yue. Searching the Best TranslationTemplate Based on Paradigm Similarity and Semantic Distance[C]. 20th International Conference on

Computer Processing of Oriental Languages(ICCPL03), 2003, 301-308

[18] 张旭. 一个基于词典与统计的中文分词算法[D]. 成都: 电子科技大学, 2007, 20-21

[19] Van Rijsbergen C. J. Information retrieval[M]. London: Butterworth's Scientific Publication, 1995

[20] Fox C. Lexical Analysis and Stoplists(including the'Brown Corpus'stoplist), Information Retrieva1: Data Structures and Algorithms[M]. Upper Saddle River, New Jersey: Prentice Hall, 1992

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

[22] YIMING YANG, XIN LIU. A re-examination of text categorization methods[C]. In: Proceedings, 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99), 1999: 42-49.

[23] Shlens, J. A tutorial on principal component analysis. Systems Neurobiology Laboratory , University of California at San Diego (December 2005)

[24] Dumais, S. T. . Latent semantic indexing (LSI)[C]: TREC-3 report. In D. K Harman, (Ed.), Proceedings of the Text Retrieval Conference (TREC-3), 1995: 219–230.

[25] BONG C H, NARAYANAN K. An empirical study of feature selection for text categorization based on term weightage[C]: Proceedings of the 2004 IEEE/W IC/ACM International Conference on Web Intelligence. Washington, DC: [s. n.]: IEEE Computer Society, 2004: 599-602.

[26] Soucy, P. &Mineau, G. W. Feature Selection strategies for text categorization[C]: Proceedings of the 16th Conference of Canadian Conference on AI, . Halifax, Canada: the Canadian Society for Computational Studies of Intelligence', Vol. 2671 of Lecture Notesin Computer Science, Springer-Verlag New York, Inc, 2003: 505–509.

[27] 雷菁, 信息论与编码基础[M]. 湖南: 国防科技大学出版社, 2011 年.

[28] LIU H, MOTODA H. Feature Selection for Knowledge Discovery and Data Mining [M]. Kluwer Academic Publishers, 1998: 66-67.

[29] JOHN G, KOHAVI R, PEGER P. Irrelevant features and the subset selection problem[C]. In: Cohen WW, Hirsh H, Eds. The 11th International Conference on Machine Learning, 1994: 121-129.

[30] Salton G, Buckley B. Term-weighting Approaches in Automatic Text Retrieval[J]. Information Processing and Management, 1998, 24(5) : 513-523.

[31] 王煜, 王正欧, 白石. 用于文本分类的改进 KNN 算法[J]. 中文信息学报. 2007, 21(3): 76-82.

[32] SANGBUM KIM, KYOUNGSOO HAN, HAECHANG RIM. Some Effective Techniques for

Naive Bayes Text Classification[J]. IEEE Transactions on Knowledge and Data Engineering , 2006(18): 1457-1466.

[33] Jochims T. Text categorization with support vector machines learning with many relevant features[J]. In: Proceedings of 10th European Conference on Machine Learning(ECML-98) , Chemnitz, DE, 1998: 4-15.

[34] 叶志刚. SVM 在文本分类中的应用[D]. 哈尔滨工程大学. 2006.

[35] S Li. , R. Xia, C. Zong, and C. Huang. A Framework of Feature Selection Methods for Text Categorization[C]: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). [S. l.]: [s. n.], 2009: 692-700.

[36] Liu Huawen, Sun Jigui, Liu Lei. Feature selection with dynamic mutual information[J]. PatternRecognition, 2009, 42(7) : 1330–1339.

[37] Xiao Ting, Tang Yan. Improved chi2 statistics method for text feature selection[J]. Computer Engineering and Applications. 2009, 45(14): 136-137, 140.

[38] Hwee Tou Ng, Wei Boon Goh, Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization[C]: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA: ACM, 1997: 67-72.

[39] Yang Yiming, Pedersen JO. A Comparative Study on Feature Selection in Text Categorization[C]: Proceedings of the 14th International Conference on Machine Learning. Nashville, USA: [s. n.], 1997: 412-420.

[40] Hiroshi Ogura, Hiromi Amano, Masato Kondo . Feature selection with a measure of deviations from Poisson in text categorization[J]. Expert Systems with Applications , 2009, 36(3): 6826–6832.

[41] Z. Zheng, X. Wu, R. Srihari. Feature Selection for Text Categorization on Imbalanced Data[J], SIGKDD Explorations, 2002, 6(1): 80-89.

附 录

作者在攻读硕士学位期间发表的论文目录

[1] 林少波，徐玲，杨丹. 基于类别相关的新文本特征提取方法[J]. 计算机应用研究,已录用.

中文文本分类特征提取方法的研究与实现

作者：林少波
学位授予单位：重庆大学