代　号　10701　　学　号　0625121814

分类号　TP391

西安电子科技大学

硕士学位论文

题（中、英文）目　基于机器学习的中文文本分类方法研究

Research on the Method of Chinese Text Categorization Based on Machine Learning

作者姓名　刘依璐　指导教师姓名、职称　焦艺　教授

学科门类　管理学　学科、专业　情报学

提交论文日期　二〇〇九年一月

西安电子科技大学

学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，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中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它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做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切的法律责任。

本人签名：刘依璐　　日期：2009.1.15

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文（保密的论文在解密后遵守此规定）。

本人签名：刘依璐　　日期：2009.1.15

导师签名：[illegible]　　日期：2009.3.1

摘 要

随着信息技术的迅速发展，特别是 Internet 的普及，信息容量呈爆炸性趋势增长，人们迫切需要一种技术来高效地组织和管理信息。文本分类技术可以准确高效地定位信息，为信息获取提供有力的支持。基于机器学习的文本分类方法，在分类效果和灵活性上都比传统的文本分类模式有所突破，成为相关领域研究和应用的经典范例。

论文首先介绍了文本分类技术的一般过程和相关技术，分析了文本分类技术在国内外的研究发展现状，在结合机器学习理论的基础上提出了论文的主要研究内容。针对现有文本分类系统存在的局限性，设计出一个关于 IT 领域文本分类的模型，包括构建 IT 领域的文本语料库、提出 WDP 特征处理方法和基于支持向量机和朴素贝叶斯的组合分类器，有效地提高了特征向量的准确度和分类方法的精确度，弥补了现有文本分类方法的不足。实验结果表明，采用 IT 领域模型的文本分类系统在查全率和查准率上都有显著地提高。

关键词：文本分类　机器学习　特征处理　组合分类器

Abstract

With the development of information technology and the prevalence of Internet, the information capacity increases explosively. There is a great desire to develop a technology which can organize and manage the information availably high-qualitily. Text categorization technology is necessary for locating the information accurately and rapidly, it can support the information extracting effectively. Basing on machine learning, the text categorization method has shown the better performance than the traditional text categorization model, and it has become the classic example of the relevant field of research and application.

This paper firstly introduces the general process and key technology of text categorization, then analyzes current research status, puts forward the main research content based on machine learning theory. Aiming at the limitation of the current text categorization system, the paper designs a text categorization model of IT field, including the construction of a IT-field text corpus, the proposal of the WDP feature processing approach and a combined classifier of SVM and NB. The model effectively improves the accuracy of the feature vector and the categorization methods, and overcomes the shortage of current categorization methods. The experiment results show that the recall and the precision of the system which adopts IT field model are promoted prominently.

Keywords: Text Categorization Machine Learning Feature Processing Combined Classifier

目 录

第一章 引 言

1.1 研究背景与意义

随着信息技术的快速发展，特别是网络的普及，以文本形式表示的信息越来越多，网络上电子文本的信息量成爆炸趋势。面对着互联网上呈指数级增长的海量信息，人们已经不能简单地依靠人工迅速有效地提取出所需的信息。如果计算机能够在信息的辨识和处理方面，为用户提供适当的支持和帮助，那将能够极大地改善目前用户面临的困境，提高信息的使用效率。

基于这种需求，人们对利用计算机进行智能化信息处理做了大量研究，包括信息检索、信息抽取、文本分类、文本摘要等研究领域。这些研究都旨在帮助用户对互联网上的大量信息加以辨识、分类，按用户兴趣加以筛选、排序，甚至提炼出要点形成摘录。这些研究成果在电子商务、数据库、web页面分类管理等领域增强了用户搜寻信息的能力，有效地提高了信息服务的质量。

其中，文本分类技术是信息检索和文本挖掘的重要基础，很多相关的研究都可以归结为分类问题。文本分类技术在预先给定的类别标记集合下，根据文本内容对文本集进行有序组织，把它划分到相关联的类别中去。最初的文本分类是依靠专家手工进行的，它对领域知识要求较高且花费巨大，不能满足大规模文本处理的要求。文本自动分类能较好地解决大量文本信息归类的问题，在自然语言理解与处理、信息组织与管理、内容信息过滤等领域都有着广泛的应用。20世纪90年代逐渐成熟的基于机器学习的文本分类方法，更注重分类器的模型自动挖掘、生成及动态优化能力，相比之前基于知识工程和专家系统的文本分类模式，在分类效果和灵活性上均有所突破，成为目前相关领域的研究热点之一[1]。

中文在构词成句上比英文复杂得多，理论和技术上也还不够成熟，但是中文是世界上使用人数最多的语言，而且随着信息时代的到来和知识经济的全球化，互联网上中文信息急剧增加，中文信息的利用率越来越大，其作用已经变得举足轻重。因此，对中文文本进行分类和研究，提高中文文本自动分类的效率已经成为促进我国经济发展和国际知识交流的迫切要求，具有重要的现实意义。

1.2 中文文本分类的研究基础

中文文本分类研究的基础是中文信息处理(Chinese information processing)[2]。中文文本分类中的许多技术都是基于中文信息处理的研究。所谓中文信息处理，

是用计算机对汉语（包括口语和书面语）进行转换、传输、分析等加工的科学。中文信息处理是自然语言信息处理的一个分支，需要以大量的语言知识、背景知识为依据，对中文信息的处理过程进行模拟。中文信息处理技术的发展与文本分类的发展过程相对应，具体分为如下 6 个阶段：

（1）理论探索的萌芽阶段，以介绍国外计算语言学领域的理论方法为主；

（2）汉字信息处理为主的早期阶段；

（3）字、词等表层处理为特征的初级阶段；

（4）句法和语义等深层处理为代表的中期阶段；

（5）语料库统计方法兴起的近期阶段；

（6）以 Internet 为主要应用对象，对大规模真实文本、智能信息访问的现阶段。

在现阶段，中文信息处理的特征主要表现为：机器学习方法与统计方法相结合、基础理论研究与实用系统并重、面向 Internet 的大规模真实文本的智能信息访问。

1.3 文本分类研究进展

自动分类研究始于20世纪50年代，美国IBM公司的H.P.Luhn在这一领域进行了开创性的研究，他首先将词频统计的思想用于文本分类中。1961年，Maron发表了有关自动分类的第一篇论文[3]，1962年H.Borko等人提出利用因子分析法进行文本的自动分类。随后许多著名的情报学家如Sparck、Salton等都在这一领域进行了卓有成效的研究[4]。文本分类从开始的基于知识的途径发展到基于机器学习的途径[5]。80年代末之前，有效地建立自动分类系统的方法大多是知识工程的方法，即利用专家规则来进行分类；到了90年代以后，统计方法和机器学习的方法被引入到文本自动分类中，取得了丰硕的成果并逐渐取代了知识工程方法。国外的自动分类研究经历了从理论研究到实用化的过程，大体上可以分为三个阶段：第一阶段(1958—1964年)主要进行自动分类的可行性研究；第二阶段(1965—1974年)自动分类的实验研究；第三阶段(1975—现在)自动分类的实用化阶段[6]。

国内的文本分类研究始于20世纪80年代，1981年南京林业大学的候汉清教授首次对自动分类进行探讨，从计算机管理分类表等方面介绍了国外的发展概况。国内文本分类研究所使用的方法也比较单一，基本上是在英文文本分类的基础上，结合中文文本的特点采用相应的策略，形成针对中文文本的分类系统。目前我国陆续出现的一批自动分类系统，整体上可以分为基于词典法的自动分类系统和基于专家系统的自动分类系统两大类。总体而言，我国的自动分类的发展阶段大体和国外相同，但我国在这个研究领域的起步较晚，并且我国的自动分类系统研究

离社会化、商品化还有一定的距离。

目前在实际应用方面，国外的已经有SAS、SPSS、KXEN等公司开发的大型商业挖掘软件，可以从事文本分类方面的研究应用；国内比较突出的有TRS公司开发的文本挖掘软件CKM等等。在国内很多高校及研究机构中，如中科院、清华大学、北京大学、复旦大学、哈尔滨工业大学等，也建立了实验室从事相关领域的研究，并且取得了突出的成绩。如北京大学的天网、复旦大学的文本分类、中科院计算所的基于聚类粒度原理的智多星中文文本分类器等。目前，我国在中文文本分类领域已经取得了令人瞩目的研究成果。

1.4 论文研究工作概述

论文对基于机器学习的中文文本分类方法进行了全面的研究，对文本分类过程中文本预处理、文本特征项的处理、现有文本表示模型以及文本分类方法进行了系统地探讨，设计了IT领域的文本分类系统，并对其进行了实验分析。论文的主要研究内容是：

（1）特征处理算法

论文对特征提取及权重确定算法进行了讨论，在现有特征处理算法的基础上进行改进，提出了一套考虑特征项离散度、专业程度和所处位置的特征处理算法，全面综合的考虑特征词在文本中的各项信息，以使特征向量能够更准确地表示文本内容。

（2）支持向量机和朴素贝叶斯的组合分类器

论文在对几种主要的文本分类方法进行研究分析的基础上，构建了支持向量机和朴素贝叶斯的组合分类器，并且进行了实验分析，得出的结论有利于深入地理解文本分类技术，从而更合理地应用它们，也可以作为研究新算法的基础。

（3）构建IT领域的语料库

由于目前中文文本分类没有统一的语料库，因此，论文构建了IT领域的中文语料库进行研究。构建的语料库力求能全面准确地代表当今互联网上的各类IT领域文本，使论文的研究具有实用价值。

1.5 论文的组织结构

论文中所有的内容都是围绕研究课题展开，对文本分类的一般过程进行规范化描述，提出特征处理和分类的改进方法，通过文本分类实验进行效果评估。具体来说，论文的组织结构如下：

第一章为“引言”，主要介绍了论文的研究背景与意义，理论基础及当前文

本分类技术的研究现状，并给出论文的主要研究内容和组织结构。

第二章为“文本分类相关技术”，主要介绍了文本分类的一般过程及机器学习思想，并描述了文本分类的相关技术。

第三章为“IT领域文本分类模型”，主要介绍了IT领域文本分类的理论模型，提出了新的文本特征提取及权重处理方法，并针对IT领域语料库构建了一个组合分类器。

第四章为“文本分类系统实验及结果分析”，介绍了一个能够实现基本功能的文本分类系统原型，该系统将作为文本分类的实验平台，帮助实现文本分类技术的测试研究工作。最后列出实验取得的各种实验结果，包括各种特征权重算法和分类算法的评估结果。通过对实验结果的分析，总结各种算法的优缺点。

第五章为“总结与展望”，主要对论文所作的工作做了简要总结，并提出下一步工作的展望。

第二章 文本分类相关技术

2.1 文本分类的一般过程

一般来说，文本是语言的实际运用形态。在具体场合中，文本是根据一定的语言衔接和语义连贯规则组成的语句系统。人们平常所接触的具有一定内容含义的文字段落都可以统称为文本。文本分类是在给定的分类体系下，根据文本内容或属性将待定文本划分到一个或者多个预先定义的类别中的方法[7]。从数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本映射到已有的类别中，该映射可以是一一映射，也可以是一对多的映射，因为通常一篇文本可以同多个类别相关联。用数学公式表示如下：

$$f: A \to B \qquad \text{其中：} \quad A = (D_1, D_2, \cdots, D_n) \qquad B = (C_1, C_2, \cdots, C_m)$$

即：A为所有待分类的文本的集合；B为给定分类体系下，所有类别的集合。A可以为无限集合，而B必须为有限集合。

文本分类的映射规则也就是通常所说的分类方法f是文本分类系统的关键，它是系统根据训练集的样本信息总结出来的分类规律，来建立判别公式和判别规则；遇到新文本后，根据总结出来的文本分类的映射规则，确定该文本相关的类别。

文本分类一般包括了文本的预处理、文本的表示、特征提取、分类器的选择与训练、分类结果的评价与反馈等过程，文本分类系统的主要功能模块为：

（1）文本预处理：将原始语料格式化为规范格式，便于后续的处理；

（2）文本模型表示：将文本分解为基本处理单元，用数学模型来表示；

（3）特征处理：从文本中抽取出反映主题的特征，并确定特征项的权重；

（4）分类器：选用分类算法，得到分类器的训练模型；

（5）效果评价：分类器的测试结果分析。

文本分类系统的主要功能模块如图2.1所示：

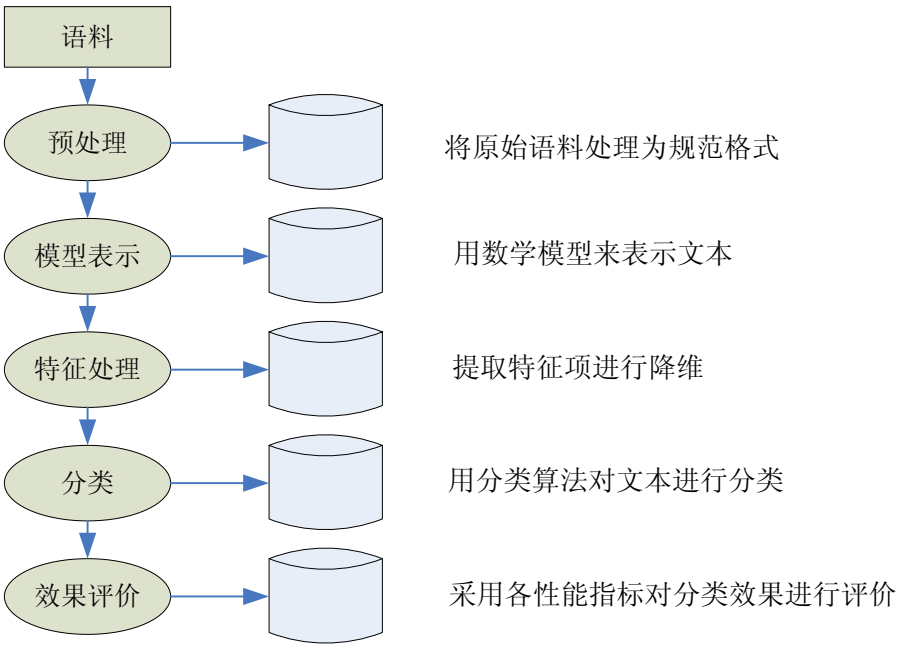

图2.1 文本分类系统的主要功能模块

2.2 机器学习思想描述

要实现自动的文本分类，通常需要让计算机对已存在的样本进行分析和学习，掌握类别知识，并能够应用于新文本的分类过程中，这就是一个机器学习的过程[8]。

关于机器学习，一直没有一个统一的定义，而且很难给出一个公认和准确的定义。一般而言，机器学习是研究如何使用计算机来模拟人类学习活动的一门学科。较为严格的提法是：机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。这里所说的“机器”指的就是计算机。

如何让机器学习尽可能地贴近人类的学习，让机器尽可能地掌握人类的智能是机器学习领域需要解决的最终问题。从20世纪50年代起，机器学习的发展过程大体上可以分为4个时期[9]：

第一阶段是从50年代中叶到60年代中叶。在这个时期，所研究的是“没有知识”的学习，即“无知”学习，其研究目标是各类自组织系统和自适应系统。主要的研究方法是不断修改系统的控制参数以改进系统的执行能力，不涉及与具体任务有关的知识。由此形成了机器学习的两种重要方法：判别函数法和进化学习。此阶段所取得的学习结果都很有限，远不能满足人们对机器学习系统的期望。

第二阶段是从60年代中叶到70年代中叶。研究目标是模拟人类的概念学习过程，并采用逻辑结构或图结构作为机器内部描述。机器能够采用符号来描述概念(符号概念获取)，并提出关于学习概念的各种假设。这种学习系统取得了较大的成功，但只能学习单一概念。

第三阶段是从70年代中叶到80年代中叶。在这个时期，人们从学习单个概念扩展到学习多个概念，搜索不同的学习策略和各种学习方法。机器的学习过程一般都建立在大规模的知识库上，实现知识强化学习。尤其令人鼓舞的是，该阶段已开始把学习系统与各种应用结合起来，并取得了很大的成功，促进了机器学习的发展。1980年，在美国CMU大学召开的第一届机器学习国际研讨会，标志着机器学习研究已经在全世界兴起。

最新的阶段是从80年代中叶至今。机器学习的研究已在全世界范围内出现新的高潮。机器学习与人工智能各种基础问题的统一性观点正在形成；各种学习方法应用范围不断扩大，一部分已形成了商品，在诊断分类型专家系统、声图文识别系统及工程控制等技术中得到广泛使用。机器学习已成为综合数学、自动化和计算机科学等学科的交叉性学科，并逐渐进入各大高校，成为一门新兴的课程。

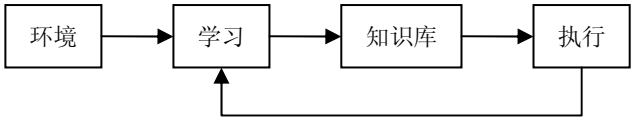

图2.2 机器学习一般过程

图2.2表示学习系统的基本结构。环境向系统的学习部分提供某些信息，学习部分利用这些信息修改知识库，以增进系统执行部分完成任务的效能，执行部分根据知识库完成任务，同时把获得的信息反馈给学习部分。在具体的应用中，环境、知识库和执行部分决定了具体的工作内容，学习部分所需要解决的问题完全由上述三部分确定。

2.3 文本表示

至今，计算机还不能像人类那样阅读完文章之后，根据自身的理解能力对文章的内容产生一定的认识。要使计算机能够高效率、高性能地处理自然文本，就需要有一个文本表示的过程，文本表示是将非结构化的文本文档表示为机器易于处理的形式的过程。文本表示通常包括文本预处理和文本模型表示等步骤，其中文本预处理为建立文本表示模型做必要的准备工作。

2.3.1 文本预处理

文本预处理主要是从文本中提取关键词来表示文本的处理过程。在预处理过程中，视具体文本来源的不同而采用不同的步骤。文本预处理的效果直接影响到文本分类的准确度，是文本分类过程中的关键因素之一。

一般情况下，计算机处理的文本不仅包含表达内容的文字，而且还包含一些

功能性的标签，如控制外观及显示样式的网页标签等。这些标签对于文本内容和属性的判断是没有实际意义的，属于文本分类的噪音，应在分类操作之前就予以删除。

对比在文本分类领域发展相对先进的英文文本分类，中文文本分类方法有很大的不同，主要区别就体现在文本的预处理部分。中文文本与英文文本最大的不同在于英文文本利用空格作为词的分隔符，而中文只是字、句和段可以通过简单的分界符来划分，唯独词没有一个形式上的分隔符。虽然英文文本也存在短语划分的问题，但在词这一层上，中文比英文要复杂得多，困难得多[10]。一般英文文本预处理的步骤包括处理与文本内容无关的标记，去除禁用词和单词的词根化等步骤。下面主要介绍中文文本预处理的步骤和方法：

1．处理文本标记

文本根据来源不同，一般还带有与内容无关的标记。这些标记可能是控制显示外观的记号；也可能是一些功能性符号，如标点符号等；还可能是一些其他媒体信息，如图像、声音、动画等；也有可能是一些乱码。它们都有一个共同的特点：都是不包含文本内容的标记，它们对文本内容没有任何实际的意义，无法对分类起到帮助作用。相反，由于文本分类系统的处理对象是具有内容的纯文本，这些标记都属于噪音，所以应该去除掉。

由于当今文本分类的研究主要采用的数据集是网页形式的数据集，所以在预处理模块需要被“过滤”的部分主要包括：HTML网页的标签、脚本语言、tag标记，以及其他媒体信息等与具体文本内容无关的部分。完成这一步骤后再进行找出标题、篇章切分为段落、段落切分为句子等相关的处理，将文本处理为统一的格式，便于分类的进行。

2．中文分词

在文本信息处理过程中，一般可以选择字、词或词组作为文本的特征项。根据实验结果，普遍认为选取词作为特征项要优于字和词组[7]。这是因为字所代表的信息量太少，且存在很多多义字，字与字之间的界限模糊，用字作为特征项将导致特征向量庞大，分类器学习时容易造成特征空间的维灾难。词组虽然携带足够的信息量，但词组在文本中出现的机率不多，用词组作为特征项会导致特征向量稀少，损失很多重要信息。因此，为了提取中文词条，需要对中文文本进行较为复杂的分词，称为中文分词。分词目的是将连续的字序列按照一定的规范重新组合成词序列。比如将“文本分类系统”切分成“文本/分类/系统”。

自20世纪70年代末80年代初以来，汉语自动分词方面的研究受到了人工智能、自然语言处理(NLP)、信息检索等多个领域的重视，各种分词方法不断涌现。根据研究出发点的不同，分词技术可以归纳为以下三类[11]：

（1）基于词典匹配的分词算法

基于词典匹配的分词方法又称机械分词方法，是按照一定的策略将待分析的字符串与词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功，即识别出一个词。按照扫描方向的不同，匹配分词方法可以分为正向匹配、逆向匹配和双向匹配；按照不同长度优先匹配的情况，可以分为最大匹配和最小匹配；按照切分出的词的多少，可以分为最小切分（使每一句中切出的词数最小）和最大切分；按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。这种基于词典匹配的分词方法简单、分词效率较高，但它完全依靠词典，没有歧义判断能力，并且由于汉语中语言现象复杂、词典不完备、规则不一致等问题的存在，也使得这种机械分词方法难以适应大规模文本的分词处理。

目前常用的词典匹配分词方法有：正向最大匹配、逆向最大匹配和双向匹配。也可以是上述各方法的相结合。

（2）基于统计的分词算法

从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。基于统计的分词方法就是把字与字相邻共现的频率作为成词的可信度评价标准。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词，否则，认为它们不能组合。这种方法只需对语料中的相邻字的组合频度进行统计，不需要切分词典，因而又叫无词典分词法。由于共现信息是通过调查真实语料库而取得的，因而基于统计的分词方法具有较好的实用性，可以根据共现信息识别歧义词和未登录词，但这种方法也有一定的局限性，它会经常切出一些共现频度高、但并不是词的常用字组合，例如“这一”、“我的”等等。

（3）基于理解的分词算法

基于理解的分词方法也称为基于知识的分词方法。这种方法将分词看作是知识推理的过程，需要进行句法、语义分析，因此它需要用大量的语言知识和信息来指导分类算法，使其可以通过上下文内容所提供的信息对词进行界定。由于汉语语言知识的多义性和复杂性，难以将这些信息组织成机器可以直接理解的形式，因此目前基于知识的分词系统还处于研究阶段。

3．过滤文本中的停用词

一篇文本的内容主要通过名词、动词和形容词等实词来体现，虚词以及在各种文本里经常出现的部分高频词对分类并无意义，这些无意义的字或词即被称为停用词(Stopword)。

通常意义上的停用词大致有如下两类：

（1）这些词应用广泛，但就是因为它随处可见，在所有类中都频繁出现，使

得它不含有不同于其它词的特性，因此没有区分类别的能力。比如“公告”一词几乎在每个网站上均会出现，这样的词无法保证能够给出真正相关的分类信息，反而还会降低分类的效率；

（2）包括语气助词、副词、介词、连接词等一些虚词，通常自身并无实际意义，和类别信息没有关联，如常见的“的”、“在”等词。适当地减少文本处理中虚词出现的频率，可以有效地提高关键词密度，更突出实词的分类信息。

为了节省存储空间，提高分类效率，需要将停用词从文本词集中剔除掉，使其不参与分类，从而减少分类噪音[12]。去停用词的方法是构建停用词表依次对分词得到的文本关键词集中的词与停用词表进行匹配，如果词存在于表中，表明该词为停用词，则从文本词集中删除；如果不在表中，则保留。

2.3.2 文本表示模型

在自然语言信息处理领域中，文本分类的表示模型与信息检索模型是相辅相成、密不可分的。一方面文本分类技术的提高使得信息检索的精确度和速度得以提高，另一方面文本分类是建立在信息检索的基础之上的，因为它借鉴了很多信息检索的表示方法和技术。信息检索技术的发展已经有30年的历史，取得了巨大的成就，产生了大批实用的检索系统，积累了许多成熟的技术，这对文本分类技术的形成和发展起了巨大的推动作用。从文本表示的角度来看，信息检索模型同样适用于文本分类系统。当今常用的信息检索的表示模型主要包括[13]：布尔模型（Boolean Model）、概率模型（Probabilistic Model）、向量空间模型（Vector Space Model）和统计语言模型（Statistical Language Model）等，下面一一进行介绍。

在传统的信息检索模型中，通常认为每个文本是由一组具有代表性的关键字或词来描述，这些关键字或词被称为特征项。用来描述文本内容的特征项应该是与文本内容密切相关的词语，我们可以为文本的特征项定义一个权重来描述这种相关程度。假设d_i表示一个文本，t_j表示一个特征项，w_{ij}是文本d_i中特征项t_j的权重。

1．布尔模型

布尔模型(Boolean Model)是基于特征项的严格匹配模型。首先，建立一个二值变量的集合，这些变量对应于文本的特征项。文本用这些特征项变量来表示，如果出现相应的特征项，则特征变量取值为“True”；否则，特征变量取“False”。由于布尔模型是基于集合论和布尔代数的一种简单检索模型，所以在具体的应用中，可以用“1”来表示特征项在文本中出现的情况，用“0”来表示特征项没有在文本中出现的情况。查询由特征项和逻辑运算符“AND”、“OR”、“NOT”组成。文本与查询的匹配规则遵循布尔运算的法则。在具体应用中，查询串通常以语义精确的布尔表达式的方式输入[14]，如$q = t_1 \wedge (t_2 \vee \neg t_3)$，通过对文本与查询

串的逻辑比较进行搜索。

布尔模型在20世纪六七十年代取得了较大的发展，出现了许多基于布尔模型的商用检索系统，如DIALOG，STAIRS，MEDLARS等。其主要优点是：速度快，具有清楚和简单的形式，易于表达一定程度的结构化信息：如同义关系（电脑OR微机OR计算机）或词组（文本AND分类AND系统）。不过，布尔模型存在着一些缺陷：

（1）它的匹配策略是基于二元判定标准(binary decision criterion)，对于一篇文本的检索来说，只有相关和不相关两种状态，缺乏对文本相关性排序(ranking)的概念，限制了检索功能。

（2）虽然布尔表达式具有精确的语义，但常常很难将用户的信息需求转换为布尔表达式，在实际应用中大多数查询用户在把他们需要的查询信息转换为布尔表达式时并不那么容易。

Boolean定义特征项只有两种状态，出现或不出现在某一篇文本中，这样就导致了特征项权重都表现为二元性，例如w_{ij}={0,1}。查询串q是一个传统的布尔表达式，文本与查询串相关度定义为：

$$sim(d_i,q)=\begin{cases}1, q\in d_i\\ 0, q\notin d_i\end{cases} \qquad \text{式(2-1)}$$

如果$sim(d_i,q)$=1，布尔模型表示查询串q与文本d_i相关，否则就表示与文本d_i不相关。

布尔模型主要的缺陷在于完全匹配会导致太多或太少的结果文本被返回。针对特征项权重的选择，引出了向量空间模型。

2．向量空间模型

向量空间模型(Vector Space Modal，简称VSM)[15]，是由G.Salton教授等人在20世纪60年代提出的，是效果较好、近些年来被广泛应用的一种方法，一直以来都是信息检索领域最为经典的计算模型。它使用向量表示文本，最早用于信息检索领域，成功应用于著名的SMART系统中[16]，这项技术后来又在文本分类领域得到了广泛的应用。该模型现已经成为最简便、最高效的文本表示模型之一。

向量空间模型的出发点是：每篇文本和查询都包含一些用特征项表达的揭示其内容的独立属性，而每个属性都可以看成是向量空间的一个维数，那么文本和查询就可以表示为这些属性的集合，从而忽略了文本的结构中段落、句子及词语之间的复杂关系。这样，文本和查询可以分别用空间的一个向量来表示，文本与查询之间的相似度可以用向量间的距离来衡量。相似度的计算方法有很多种，常用的方法有内积、dice系数、Jaccard系数和余弦系数[17]。通常采用余弦系数法，即用两个向量之间的夹角余弦来表示文本与查询间的相似度。夹角越小，说明文本

和查询间的相似度越大。向量空间模型和机器学习算法在自动文本分类领域中的紧密结合和成功运用，使得基于向量空间模型的文本表示方法迅速成为文本分类研究领域中文本表示的主流方法。

在向量空间模型中，文本d_i被看作为由一组特征项(t_1，t_2，…，t_n)组成的n维向量空间，文本d_i简化为以特征项的权重为分量的向量表示(w_{i1}，w_{i2},…，w_{in})，权重w_{ij}表示特征项t_j对文本d_i分类的贡献程度，取值范围是[0,1]。查询串q同样可以表示成向量(w_{q1},w_{q2},…，w_{qm})。在对所有文本和查询串q进行向量化之后，检索过程简化为空间向量的运算，文本信息的匹配问题转化为向量空间中的向量匹配问题，大大减小了问题的复杂性。用两个向量的夹角余弦值来计算文本与查询串的相关度，公式如式(2-2)：

$$sim(d_i,q)=\frac{Wi\square Wq}{\sqrt{\sum_{j=1}^{n}w_{ij}^2}\quad\sqrt{\sum_{j=1}^{n}w_{qj}^2}} \qquad \text{式(2-2)}$$

VSM只是提供了一个理论框架，项的权重评价、相似度的计算没有统一的规定，可以使用不同的权重评价函数和相似度计算方法，使得此模型有广泛的适应性，在多种系统中得到了成功地应用。

向量空间模型具有较强的可计算性和可操作性，特别是随着网上信息的迅速膨胀，它的应用已经不仅仅局限于文本检索、自动文摘、关键词自动提取等传统问题，还可以应用到搜索引擎、个人信息代理、网上新闻发布等信息检索领域中。在向量空间模型中，文本的内容被形式化为多维空间的一个点，把文本以向量的形式定义到实数域中，能够使用模式识别和其它领域中各种成熟的计算方法，极大地提高了自然语言文本的可计算性和可操作性。知识表示(knowledge representation)始终是知识处理(knowledge processing)的主要瓶颈之一，特别是在以自然语言为研究对象的知识处理和知识获取(knowledge acquisition)问题中更是如此。向量空间模型的最大优点在于知识表示方法上的巨大优势。向量空间模型的优点具体表现在：

（1） 特征项权重的算法提高了检索的性能；

（2） 部分匹配的策略使得过滤得到的结果文本集合更接近用户的查询需求；

（3） 可根据文本与查询串的相似度计算结果对文本进行排序。

3．概率模型

在信息检索中，由于文本信息相关性判断的不确定性和查询信息表示的模糊性，促使人们使用概率的方法解决这方面的问题。信息检索的概率模型是基于概率排序原则，对于给定用户查询q，计算所有文本概率，并依照计算结果将文本从大到小排序。概率模型试图估计用户找到其感兴趣的文本d_i的概率，模型假设这个

相关概率只是依赖于查询和文本的表示。

在概率模型中，概率公式为P(*R*|D,*q*)，其中P表示文本D与用户查询*q*相关的概率。另外，用 *R′* 来表示文本D与用户查询*q*不相关的概率，这样，就有P(*R*|D,*q*)+P(*R′*|D,*q*)=1，即用二值形式判断相关性[18]。

把文本用特征项来表示，即$d_i=(t_1, t_2, \cdots, t_n)$，在概率模型中，用特征向量来表示，$d_i=(w_{i1}, w_{i2}, \cdots, w_{in})$，查询串*q*也用向量来表示，$q=(w_{q1}, w_{q2}, \cdots, w_{qm})$。在概率模型中，特征项的权重都是二值的，即 $w_{ij} \in \{0,1\}$, $w_{qj} \in \{0,1\}$，若特征项出现了则权重取值为1，若没有出现权重取值为0。

在信息检索中，估计参数是比较困难的，一般不直接计算P，而是把计算P($R|d_i,q_k$)换为计算P($R|t,q_k$)，这样处理略去了公式中与文本无关的特征项，便于计算处理。假设包含相同特征项的文本，经过计算后，它们的相关概率是相同的。将所有文本按照相关概率P进行排序，等价于将所有文本按照特征向量排序。任一文本*d*的概率相关性的计算为

$$P(R \mid D,q) = \sum_i d_i \times \lg \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad \text{式(2-3)}$$

其中， $p_i = P(t_i = 1 \mid R,q)$ ， $q_i = P(t_i = 1 \mid R',q)$ 。

参数p_i，q_i主要是通过相关反馈进行估计，简单的方法如：

$$p_i = r_i / r \ , \quad q_i = (n_i - r_i)/(n-r) \qquad \text{式(2-4)}$$

其中，*n*为反馈文本集所含文本总数，*r*为与用户查询相关的文本数，n_i为特征t_i出现的样本个数，r_i为特征t_i出现且与用户查询相关的文本个数。概率模型就是采用相关反馈的方法，从两个初始的概率开始，不断调整概率估计值，直到得到一个满意的概率排序。

概率模型的主要优点是：理论上，文本按照其相关概率的降序排列，综合全面地考虑了文本集的整体情况。主要缺点包括：文本向量只采用简单的二值形式，没有利用到文本中的更多信息，比如特征项在文本中出现的频率等。这种模型对所处理的文本集依赖过强，而且处理问题过于简单。

4．统计语言模型

如果把文本看成词的序列（或字的序列），那么这些词的出现与否及其出现的次序可以看成是一种语言结合模式。这些结合模式信息可以被用来进行文本类别判定，也就是说不同类别的文本，其语言结合模式是不同的。即设t_i是文本中的任意一个词，如果已知它在该文本中的前两个词 $t_{i-2}t_{i-1}$，便可以用条件概率P($t_i|t_{i-2}t_{i-1}$)来预测t_i出现的概率。这就是统计语言模型的概念[19]。

统计语言模型中，常用于文本分类任务的是N元语言(N-gram)模型。N-gram模型不考虑组成文本的语义单位是字、词还是词组，而是将整个文本看成是由不同

字符组成的字符串[20]，因而可以方便地表示各种语言文本文档。一般来说，如果用变量*d*代表文本，它由顺序排列的n个词组成，即$d=t_1t_2\cdots t_n$,则根据条件概率的乘积公式[21]，有：

$$P_c(d)=P_c(t_1t_2\cdots t_n)=P_c(t_1)P_c(t_2\mid t_1)P_c(t_3\mid t_1t_2)\cdots P_c(t_n\mid t_1t_2\cdots t_{n-1}) \qquad \text{式(2-5)}$$

其中，$P_c(t_n|t_1\ t_2...t_{n-1})$表示在给定历史信息$t_1\ t_2...t_{n-1}$的条件下，词$t_n$在类别C中出现的概率。所有这些信息构成了一条Markov链，也就是说N元语言模型就是N-1阶的Markov模型。$P_c(t_n|t_1\ t_2...t_{n-1})$的计算量是非常巨大的，尤其当n取值较大时，在实际应用中，为简化计算，往往只考虑一个或两个历史信息，形成二元语言模型和三元语言模型。表面上看，N取值越大，计算出来的概率的准确性越高。但是，这种准确性的提高是以计算量的级数上升为代价的，同时，高阶模型的数据稀疏问题也比低阶模型的要严重得多，从而会降低估计值的可靠性，这会对分类的性能起到负面的影响。

一般来说，N元模型就是假设当前词的出现概率只与它前面的N-1个词有关。论文考虑任意一个词t_i的出现概率只与它前面的两个词有关的情况，即三元语言模型(tri-gram)：

$$P_c(d)\approx P_c(t_1)P_c(t_2\mid t_1)\prod_{i=3}^{n}P_c(t_i\mid t_{i-2}t_{i-1}) \qquad \text{式(2-6)}$$

公式中这些概率参数可以转化为特定词序在语料库C类训练文本集中的出现频率来替代，即：

$$P_c(t_i\mid t_{i-2}t_{i-1})\approx count(t_{i-2}t_{i-1}t_i)/count(t_{i-2}t_{i-1}) \qquad \text{式(2-7)}$$

式中count(…)表示特定词序列在语料库C类训练文本集中出现的累计次数。

新文本的分类过程是：将待分类的新文本*d*先预处理成连续的汉字串集合。对每个连续的汉字串在进行分词、停用词和非实词过滤后得到的一个词串T，分别计算其属于各个类别C的概率：

$$P_c(d)=P_c(t_1t_2\cdots t_m)=P_c(t_1)P_c(t_2\mid t_1)\prod_{i=3}^{m}P_c(t_i\mid t_{i-2}t_{i-1}) \qquad \text{式(2-8)}$$

即可表示为：

$$P_c(d)\propto \log P_c(t_1)+\log P_c(t_2\mid t_1)+\sum_{i=3}^{m}\log P_c(t_i\mid t_{i-2}t_{i-1}) \qquad \text{式(2-9)}$$

新文本*d*属于类别C的概率为$\sum_i P_c(d)$，拥有最大概率的那个类别被判别为文本*d*所属的类别，即$\arg\max\sum_i P_c(d)$。

事实上，N元语言模型在自然语言处理的许多领域上都取得了相当的成功，但是由于存在着数据噪声大、特征生成复杂、计算量大、模型参数存储空间需求较大，且训练过程也相对复杂缓慢，以其作为文本分类模型较为少见。然而随着计算机硬件水平的迅速提高，存储空间和计算量的要求已经越来越可以满足，考虑到N-gram表示模型方式具有语言无关性这样一个显著优点，在文本分类领域，还是具有相当大的研究潜力的。相关学者研究表明：在分类准确率及稳定性上，N元语言模型效果较好。

5．各种模型的比较

布尔模型由于其简洁性一直受到商业搜索引擎的青睐，但功能也是最弱的。向量空间模型和概率模型由于其形式化备受学者们的推崇，如果从学术研究状况及商业运用模式上看, 向量空间模型则更受欢迎。统计语言模型是发展时间最短的模型，虽然能更为合理地表达文本信息，但目前实现技术及方法尚未成熟，还有待进一步的研究[22]。四种模型的区别见表2.1：

表2.1 四种模型的区别

经典模型	布尔模型	向量空间模型	概率模型	统计语言模型
提出时间	20世纪50年代	20世纪60年代	20世纪80年代	20世纪90年代末
理论基础	集合论	代数理论	概率论	概率论、随机过程
相关文本判断	二元无序	非二元有序	非二元有序	非二元有序
系统实现难度	简单	简单	较难	简单
部分匹配支持	不支持	支持	支持	支持
文本表示方法	词	词向量	词	N-gram
学术代表系统	无	SMART	INQUERY	LEMUR
商业运用情况	采用	常采用	采用	未采用

综上，由于向量空间模型的简单性和高效性，论文的研究采用向量空间模型。

2.4 特征处理

训练样本集经过预处理得到的关键词的集合构成了初始特征项集合，简称特征集，这个集合中特征数目过多是制约分类的重要因素。即使一个小规模的样本集，经过预处理也会得到几万个特征词，其中有些词在文本中出现次数极少，无明显作用，甚至可能成为噪音，通常认为这些词意义不大，称为低频弱关联词；有些词则出现频率较高，蕴含了大量和类别有关的信息，称为高频强关联词。特征处理，就是对初始特征集中的初始特征进行特征提取，将弱关联词去掉，抽取强关联词构成用于学习的特征集，并通过特征权重函数给这些特征赋予不同的权重，来表示特征对文本的重要程度[23]。

2.4.1 特征提取方法

文本分类系统应该选择尽可能少、准确并且与文本主题密切相关的文本属性进行文本分类，但由于构成文本的词汇数量是相当大的，表示文本的向量空间的维数也相当大，导致文本分类的最大问题是特征空间的高维性和文本表示向量的稀疏性。通常标准的分类方法很难处理如此大的特征集，计算开销很大，而且分类结果不可靠。为了解决这一问题，一般要进行维数缩减的工作。在降维操作后，不仅可以将向量空间的维数大大降低，而且可以减少“过度拟合”的问题，使分类模型更具有普遍性，有效的达到提高程序效率和分类精度的目的。

在不降低分类器的准确性的前提下寻求一种自动高效的特征抽取方法，降低特征空间的维数，提高分类器的效率，成为文本分类中需要面对的重要问题。近年来在中文文本分类中经常采用的特征抽取方法包括最简单的停用词移除、互信息MI、信息增益IG和CHI统计等。特征抽取方法的选取主要依据Y.Yang的实验结果[24]。由于中文文本分类问题与英文文本分类相比具有相当大的差别，体现在原始特征空间的维数更大、文章表示更加稀疏、词性变化更加灵活等多个方面。因此，在英文文本分类中表现良好的特征抽取方法未必适合中文文本分类。对中文文本分类中的特征抽取方法进行系统的比较研究显得十分重要。

1．频率统计

文本分类中常用到的频率统计包括：词频和文档频率。

词频(Term Frequency，简称TF)，是一种词汇分析研究方法，它通过统计一定长度的语言材料中每个词出现的次数，分析统计结果，以便描绘词汇规律。分类中一个词在文本中出现的次数对这篇文本和所属类别都有着很重要的作用，一个词在文本中出现的次数越多，表明这个词与文本关系越紧密。词频统计在文本分类中扮演着重要角色，在早期的特征提取研究中，曾使用词频作为特征提取的依据指标，而现在常用它来计算预处理得到的关键词的类别信息的含量。

文档频率(Document Frequency，简称DF)，是指在训练语料中出现该词条的文档数。采用DF作为特征抽取是基于如下基本假设：DF值低于某个阈值的词条是低频词，它们不含有或含有较少的类别信息，将这样的词条从原始特征空间中移除能够降低特征空间的维数，不会对分类器的性能造成影响。如果低频词恰好是噪音词，还有可能提高分类器的性能。在实际操作中，一般统计每个词条在训练语料中的文档频率，从原始特征空间中移除文档频率低于某一阈值的词条，保留文档频率高于该阈值的词条作为特征。

文档频率是最简单的特征抽取技术，由于其具有相对于训练语料规模的线性计算复杂度，它能够容易地被用于大规模语料统计，通常被认为是一个提高效率

的有效方法。

2．信息增益

信息增益(Information Gain)在机器学习领域被广泛使用，它是一种基于熵的评估方法，涉及较多的数学理论和复杂的熵理论公式，被定义为某特征在文本中出现前后的信息熵之差。根据训练数据，计算出各个特征词的信息增益，并按信息增益从大到小排序，删除信息增益很小的特征词。在文本特征抽取中，对于词条*t*和类别C，IG考察C中出现和不出现*t*的文档频数来衡量词条*t*对于C的信息增益。信息增益的计算公式如下：

$$IG(t)=-\sum_{i=1}^{m}P(c_i)\log P(c_i)+P(t)\sum_{i=1}^{m}P(c_i\mid t)\log P(c_i\mid t)+P(\overline{t})\sum_{i=1}^{m}P(c_i\mid \overline{t})\log P(c_i\mid \overline{t}) \quad \text{式 (2-10)}$$

其中 $P(c_i)$ 表示C_i类文本在语料库中出现的概率，$P(t)$ 表示语料中包含词条*t*的文本的概率，$P(c_i\mid t)$ 表示文本包含词条*t*时属于C_i类的条件概率，$P(\overline{t})$ 表示语料中不包含词条*t*的文本的概率，$P(c_i\mid \overline{t})$ 表示文本不包含词条*t*时属于C_i的条件概率，m表示类别数。

3．互信息

互信息(Mutual Information)是信息熵的引申概念，它是对两个随机事件相关性的度量，在统计语言模型中被广泛采用。词条*t*和文本类别C的互信息定义为：

$$MI(t,c)=\log\frac{P(tc)}{P(t)\times P(c)} \quad \text{式(2-11)}$$

其中 $P(tc)$ 表示语料中属于C且包含*t*的文档概率，$P(t)$ 表示语料中包含词条*t*的文本的概率，$P(c)$ 表示语料中C类文本出现的概率。从概率上说，如果某个词与某一类别在分布上是统计独立的，即P(*tc*)=P(*t*)×P(*c*), MI(t,c)值自然为零，也就是说词*t*的出现对于预测类别C没有什么信息量。

在实际计算中，互信息表达式可用训练集中相应的出现频数予以近似。如果用X表示包含*t*且属于C的文档频数，Y为包含*t*但不属于C的文档频数，Z表示属于C但不包含*t*的文档频数，N表示语料中文本总数，则有：

$$MI(t,C)=\log\frac{X\times N}{(X+Y)\times(X+Z)} \quad \text{式(2-12)}$$

对于存在多个类别的应用，分别计算*t*对于每个类别的MI值，再用式(2-13)计算词条*t*对于整个语料的MI值：

$$MI(t)=\max_{i=1}^{m}MI(t,C_i) \quad \text{式(2-13)}$$

互信息计算有一个时间复杂度，类似于信息增益。互信息的不足之处在于得分非常受到词条出现频数的影响，而且前期的计算量比较大。

4．CHI统计

CHI统计方法度量词条*t*和文本类别C之间的相关性，并假设*t*和C之间符合具有一阶自由度的 χ^2 分布。词条对于某类的 χ^2 统计值越高，它与该类之间的相关性越大，携带的类别信息也越多。令N表示训练语料中的文本总数，C为某一特定类别，*t*表示特定的词条，*A*表示属于C类且包含*t*的文档频数，*B*表示不属于C类但包含*t*的文档频数，*C*表示属于C类但不包含*t*的文档频数，*D*是既不属于C也不包含*t*的文档频数。则词条*t*对于C的CHI值由下式计算：

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad \text{式(2-14)}$$

为了将CHI统计应用于多个类别，与互信息的处理类似，可以首先对各个词条计算它与每个类别的CHI值，再用式(2-15)计算它对整个语料的CHI值：

$$\chi^2(t)=\max_{i=1}^{m}\chi^2(t,c_i) \qquad \text{式(2-15)}$$

其中m为类别数。

5．期望交叉熵(ECE, Expected Cross Entropy)

$$ECE(t)=P(t)\sum_i P(c_i\mid t)\log\frac{P(c_i\mid t)}{P(c_i)} \qquad \text{式(2-16)}$$

如果词条*t*和类别C_i强相关，那么 $P(c_i\mid t)$ 就大，若$P(c_i)$又很小，则说明该词对该类的影响大。ECE反映了文本类别的概率分布和出现了某种特征的条件下文本类别的概率分布之间的距离。

2.4.2 特征词权重确定

词是组成文本的基本元素。在所有的词中抽取出能够表示文本特征的词组成文本的特征项，并按某一方法赋予特征项相应的权重。特征项的权重综合反映了该特征项对标识文本内容的贡献度和文本内容之间的区分能力。特征项在不同文本中出现的频率满足一定的统计规律，因此可以通过特征项的频率特性计算其权重[25]。一个有效的特征项集合必须具有以下两个特征：

（1）完全性：特征项能够完整反映目标文本的内容；

（2）区分性：特征项具有将目标文本和其他文本相区分的能力。

根据这两个特征，特征项权重的计算满足以下两个原则：一是正比于特征项在文本中出现的频率；二是反比于文本集中出现该特征项的文档频率。

常见的特征项权重算法有以下几种。设tf_{ik}表示词t在文本d中出现的频率（词频），N表示所有文本数量，M表示所有词的数量，df_i表示包含词t的文本数（文档频率），w_{ik}表示词t在文本d中的权重。

1．布尔权重法：

布尔权重法是最简单的取权重方法，如式(2-17)所示，如果某个词在文本中出现，其权重则为1，否则为0。

$$w_{ik}=\begin{cases}1 & df_{ik}>0\\ 0 & df_{ik}=0\end{cases} \qquad \text{式(2-17)}$$

2．词频权重法：

词频权重法如式(2-18)所示，是简单的用词在文本中出现的频数作为权重。

$$w_{ik}=tf_{ik} \qquad \text{式(2-18)}$$

3．tf×idf权重法

tf×idf权重法是一种常用的权重方法[26]，这种权重法在词频权重法的基础上引入了对文档频率的考虑，文档频率越大的词的权重应该越低。

$$w_{ik}=tf_{ik}\times\log\left(\frac{N}{df_i}\right) \qquad \text{式(2-19)}$$

4．tfc权重法

*tfc*权重法[27]是在*tf*×*idf*权重法的基础上用文本长度做正规化，如式(2-20)所示：

$$w_{ik}=\frac{tf_{ik}\times\log\left(\frac{N}{df_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[tf_{jk}\times\log\left(\frac{N}{df_i}\right)\right]^2}} \qquad \text{式(2-20)}$$

5．ltc权重法

*ltc*权重法[28]与*tfc*权重法的区别在于使用词频的对数取代词频，以减少词频差异过大的影响，其详细形式如式(2-21)所示。

$$w_{ik}=\frac{\log(tf_{ik}+1)\times\log\left(\frac{N}{df_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[\log(tf_{jk}+1)\times\log\left(\frac{N}{df_j}\right)\right]^2}} \qquad \text{式(2-21)}$$

6．熵权重法

熵权重法[29]是基于信息论的方法，根据词所具备的信息量的多少来决定词的权重。这是这些权重方法中最繁琐也是效果最好的方法，如式(2-22)所示。公式中

的Z表示词条t的熵，如果词条t在每篇文章中都出现，Z的值则为-1；如果词条t仅出现在一篇文章中，Z的值则为0。

$$w_{ik} = \log(tf_{ik} + 1) \times (1 + Z) \quad \text{式(2-22)}$$

$$Z = \frac{1}{\log(N)} \sum_{j=1}^{N} (\frac{tf_{ij}}{df_i} \log(\frac{tf_{ij}}{df_i})) \quad \text{式(2-23)}$$

2.5 文本分类方法

文本分类的核心问题是如何根据语料库构造一个分类函数或分类模型，并利用此分类模型将未知类别的文本映射到指定的类别空间。目前存在多种基于向量空间模型的分类算法，主要包括简单向量距离法[30]、K最近邻居算法[31]、贝叶斯算法[32]、决策树法[33]和支持向量机算法[34]等传统技术方法以及基于软计算的神经网络[35]、粗糙集、模糊逻辑和遗传算法等。其中，基于软计算的方法通过协同工作提供了一种灵活的数据处理能力，其目标是实现对不精确、不确定、部分信息的处理能力和近似推理能力，以求能方便、稳健、低代价地逼近人类的分析判断能力。模糊逻辑提供处理由于模糊而不是随机产生的不精确、不确定的算法，粗糙集则处理由于不可分辨关系导致的不确定性，神经网络用于模式分类与聚类，而遗传算法则用于优化和搜索。本论文主要研究了文本分类领域的一些传统技术，它们也是目前文本分类领域广泛应用的技术，下面做详细的介绍。

2.5.1 简单向量距离算法

简单向量距离算法也被称为Rocchio算法，该方法的分类思路十分简单，根据算术平均为每类文本集生成一个代表该类的中心向量，然后在新文本来到时，确定新文本向量，计算该向量与每类中心向量间的相似度，最后判定文本属于与之距离最近的类。具体步骤如下：

（1）计算每类文本集的中心向量，方法为所有训练文本向量的算术平均；

（2）新文本到来后，首先进行分词，然后将新文本表示为特征向量；

（3）计算新文本特征向量与每类中心向量间的相似度，公式为：

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{m} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{m} w_{ik}^2)(\sum_{k=1}^{m} w_{jk}^2)}} \quad \text{式(2-24)}$$

其中，d_i为新文本的特征向量，d_j为第j类的中心向量，m为特征向量的维数；

（4） 比较每类中心向量与新文本的相似度，将文本划分到与文本相似度最大的那个类别中。

简单向量距离算法简单易行，分类速度较快。

2.5.2 KNN（K 最近邻居）算法

该算法的基本思路是：在给定新文本后，考虑在训练文本集中与该新文本距离最近（最相似）的K篇文本，根据这K篇文本所属的类别判定新文本所属的类别，具体的算法步骤如下：

（1）根据特征项集合重新描述训练文本向量；

（2）在新文本到达后，根据特征词对新文本进行分词，确定新文本的向量表示；

（3）在训练文本集中选出与新文本最相似的K个文本，计算公式如式(2-24)所示。其中，K值的确定目前没有很好的方法，一般采用先定一个初始值，然后根据实验测试的结果调整K值，一般初始值定为几十到几百之间；

（4）在新文本的K个邻居中，依次计算每类的权重，计算公式如式(2-25)：

$$W(C_j)=\sum Sim(x,d_j)y(d_j,C_j) \qquad \text{式(2-25)}$$

其中，x为新文本的特征向量，$Sim(x,d_j)$为相似度计算公式，与上一步骤的计算公式相同，而$y(d_j,C_j)$为类别属性函数，即如果d_j属于类C_j，那么函数值为1，否则为0；

（5）比较各类的权重，将文本划分到权重最大的那个类别中。

KNN的训练过程较快，而且可以随时添加或更新训练文本来调整。但由于需要较大的空间来保存训练文本，分类开销很大，因此在大规模数据集上的分类效果并不十分理想，而在小数据集上的表现优异。

2.5.3 贝叶斯方法

贝叶斯分类是一种统计学分类方法，它基于贝叶斯定理，可以用来预测类成员关系的可能性，给出文本属于某特定类别的概率。分类时根据预测结果将该文本分到概率最高的类别中去即可。

朴素贝叶斯(Naive Bayes)假定在一个具有许多属性的事例中，文本的一个属性对于分类的影响独立于其他属性，即文本的属性之间是不相关的。这是为了降低计算开销而引入的类条件独立的假定。文本d用其包含的特征词表示，即$d=(t_1,t_2,\cdots,t_j,\cdots,t_n)$，n是$d$的特征词个数，$t_j$是第j个特征词，由特征独立性假设，得到：

$$P(d \mid C_i) = P((t_1, t_2, ...t_n) \mid C_i) = \prod P(t_j \mid C_i) \quad \text{式(2-26)}$$

式中：$P(t_j|C_i)$表示分类器预测单词t_j在类C_i的文本中发生的概率。因此式(2-26)可以转换为：

$$P(C_i \mid d) \propto P(C_i) \times \prod P(t_j \mid C_i) \quad \text{式(2-27)}$$

朴素贝叶斯分类模型训练的过程其实就是统计每一个特征在各类中出现规律的过程，具有较好的速度和准确度，它在多领域的成功使它得到了非常广泛的应用。

2.5.4 决策树法

决策树是一种树状结构，它从根节点开始，对数据样本（由实例集组成，实例有若干属性）进行测试，根据不同的结果将数据样本划分为不同的数据样本子集，每个数据样本子集构成一个子节点。生成的决策树每个叶节点对应一个分类，构造决策树的目的是找出属性和类别间的关系，用它来预测将来未知类别的记录类别。这种具有预测功能的系统叫决策树分类器。

一般来说，决策树算法主要围绕两大核心问题展开：第一，决策树的生长问题，即利用训练样本集，完成决策树的建立过程；第二，决策树的剪枝问题，即利用检验样本集对形成的决策树进行优化处理。

决策树的构建是一种自上而下、分而治之的归纳过程，本质是贪心算法。各种算法建树的基本过程相似，是一个递归的过程。

设数据样本集为S，算法框架如下：

（1）如果数据样本集S中所有样本都属于同一类或者满足其他终止准则，则S不再划分，形成叶节点；

（2）否则，根据某种策略选择一个属性，按照属性的各个取值，对S进行划分，得到n个子样本集，记为S_i，再对每个S_i迭代执行步骤（1）。

经过n次递归，最后生成决策树。从根到叶节点的一条路径对应着一条规则，整棵决策树就对应着一组析取表达式规则。

为了防止决策树和训练样本集的过度拟合，特别是存在噪声数据或不规范属性时更为突出，需要对决策树进行剪枝。剪枝的算法通常利用统计方法决定是否将一个分支变为一个节点。通常采用两种方法进行决策树的剪枝，即在决策树生长过程完成前就进行剪枝的事前修剪法和在决策树生长过程完成后才进行剪枝的事后修剪法。

决策树分类算法自提出以来，出现了很多种，早期的是CLS学习算法和CART算法，最有影响的是1986年Quinlan提出的ID3算法。ID3算法体现了决策树分类的

优点：算法的理论清晰、方法简单，学习能力较强。缺点是：只对比较小的数据集有效，且对噪声比较敏感。在ID3算法的基础上，Quinlan又发展了具有重要影响的C4.5算法，它继承并改进了ID3算法，使用非常广泛。为了适应处理大规模数据集的需要，后来学者又提出了若干改进的算法，如SLIQ和SPRINT等，也都取得了较好的效果。

决策树文本分类法分类精度较好，并且可以很好的抵抗噪声，但是在处理大规模数据集的情况下效率不高。

2.5.5 支持向量机算法

支持向量机[32](Support Vector Machines，SVM)是由Vapnik在1995年提出的，用于解决二分类模式识别问题[36]。Joachims最早将SVM方法用于文本分类[37]，它将文本分类问题变成了一系列二分类问题。

支持向量机算法是建立在统计学习理论的VC维理论和结构风险最小原理基础上的，它将降维和分类结合在一起，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，这里模型的复杂性代表对特定训练样本的学习精度，而学习能力代表无错误的识别任意样本的能力。支持向量机算法的目的在于寻找一个超平面H，该超平面可以将训练集中的数据分开，且与类别边界的沿垂直于该超平面方向的距离最大，故SVM法也被称为最大边缘(Maximum Margin)算法。样本集中的大部分样本不是支持向量，移去或者减少这些样本对分类结果没有影响，这样只用各类别边界样本的类别来决定分类结果的做法，具有较强的适应能力和较高的准确率。

支持向量机方法适合大样本集的文本分类，而且由于SVM算法不受样本趋于无穷大理论的限制，它对小样本的自动分类同样有着较高的精度。SVM分类器的文本分类效果很好，具有其他机器学习技术难以比拟的优越性。其缺点在于难以针对具体问题选择合适的函数；另外SVM训练速度受到训练集规模的较大影响，计算开销较大。

2.6 文本分类效果评价

对文本分类结果主要从三个方面评价：有效性、计算复杂度和描述的简洁度。有效性衡量的是一个分类器正确分类的能力；计算复杂度包括时间复杂度和空间复杂度，即通常所说的分类速度和占用硬件资源大小；描述的简洁度即算法的描述越简单越好。在这三个方面中，有效性最为重要。文本分类系统的任务就是对文本进行正确的分类，保证有效性是判定该系统是否合格的最重要因素，也是进

行其他指标评价的基础。

评价分类结果有效性主要用3个指标：查全率(Recall)、查准率(Precision)和F-测量(F-Measure)[38]，这几个指标均来源于信息检索领域。

2.6.1 查全率和查准率

查全率(Recall) *R*衡量的是所有实际属于类别C的文本被分类器分到该类别中的比率；查准率(Precision)*P*衡量的是所有被分类器分为类别C的文本中正确文本的比率。用公式表示如式(2-28) 和式(2-29)：

$$R = \frac{TP}{TP + FN} \quad \text{式(2-28)}$$

$$P = \frac{TP}{TP + FP} \quad \text{式(2-29)}$$

其中，*TP*指的是被分类器正确分到类别C的文本数；*FN*指的是实际属于类别C但分类器没有将其正确分到类别C的文本数；*FP*指的是实际不属于类别C却被分类器错误的分到类别C的文本数。

2.6.2 F-测量

查全率和查准率是两个互相矛盾的衡量指标。一般情况下，查全率会随着查准率地升高而降低，两者不可兼得。所以很多情况下要将它们综合在一起考虑。最常用的综合方法就是F-测量(F-Measure)，定义如式(2-30)：

$$F_\beta(P,R) = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{式(2-30)}$$

其中β是一个调整参数，用于以不同权重综合查全率和查准率。当β=1时，表示查全率和查准率被平等的对待，此时F-测量(F-Measure)又被称为F_1指标，定义如式(2-31)：

$$F_1(P,R) = \frac{2PR}{P + R} \quad \text{式(2-31)}$$

2.6.3 微平均和宏平均

查全率、查准率及F-测量方法，都是针对单个类别的分类情况而言的，当需要评价某个分类方法时，还需要将所有类别的结果综合起来得到平均的结果。综合的方法有两种，微平均(micro-averaging)和宏平均(macro-averaging)。

微平均计算所有类别中正确分类和错误分类的实例总数，再求查全率R和查准

率P，宏平均先计算各个类别的查全率和查准率，然后取算术平均。据相关文献论述，目前还没有关于哪种方法好的定论。当数据集间的差异比较大时，两者值的差异也比较大。当某类别具有较低通用性的时候，宏平均能更客观的评价分类效果。

2.7 研究现状

综上，文本分类的一般过程可以用图2.3表示，其中训练阶段和分类阶段是文本分类过程的两个重要阶段。

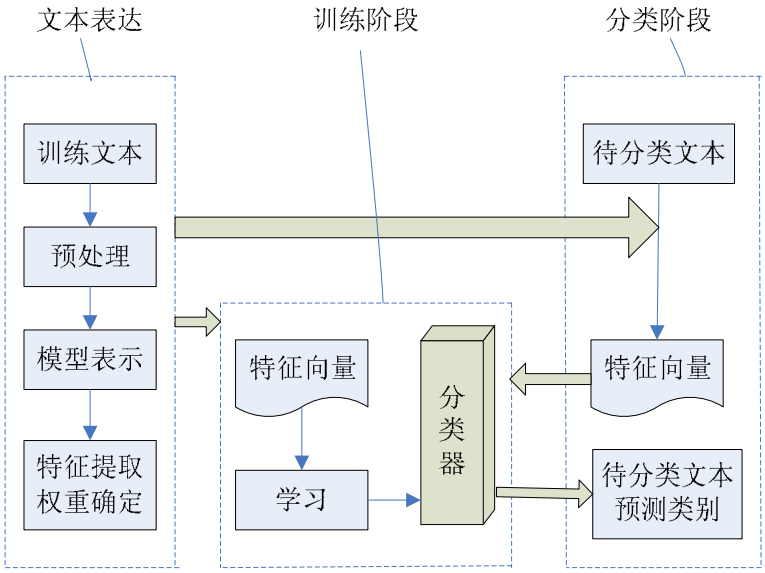

图2.3 文本分类的一般过程

目前，国内文本分类的研究取得了一定的效果，但与国外先进水平比较，相对落后，主要存在以下几个问题：

（1）缺少统一的中文语料库

至今尚无标准的用于文本分类的中文语料库，各个研究者分头收集自己的训练文本集，并在此基础上开展研究。因此语料库基本上都是针对自己的系统而规划的，不具有普遍性。

（2）特征向量形成方法有待改进

特征向量的形成包括特征提取和权重确定两个方面，是文本分类中十分重要的一个环节，对文本分类正确率有着决定性的影响。在目前适用的方法中，普遍采取与词频和倒文档频率相关函数确定权重的方法，文本中很多其他的信息没有用上，造成了特征词权重的片面性。

（3）分类方法的准确度

目前文本分类方法主要以机器学习方法为主，取得了较好的效果。但单一的

分类方法往往在保证分类准确度和高效率之间难以取得平衡，实际需要建立一个即能保证分类准确度又能取得高效率的文本分类系统。

2.8 本章小结

本章全面系统地介绍了文本分类的一般过程和文本分类的相关技术，并总结了当前的研究现状。

文本分类的主要步骤依次为文本预处理、文本模型表示、特征处理、文本训练及分类和分类效果评价，要进行文本分类研究以上步骤缺一不可。通过对文本分类一般过程和相关技术知识的简单介绍，就能从总体上了解文本分类，同时也为进一步的研究工作打下了良好的基础。

第三章 IT 领域文本分类模型

3.1 IT 领域文本分类模型

本章主要针对 IT 领域的文本构建了一个文本分类模型。模型首先设计了 IT 领域的文本语料库，力求能够全面代表 IT 领域中文本的特点，为文本分类提供研究素材。然后提出了特征处理中特征提取和特征权重确定的新方法，论文称之为 WDP 特征处理方法。在对语料库中的文本进行预处理之后，采用 WDP 特征处理方法对文本进行特征提取和特征权重计算的处理，形成特征向量。最后构造了一个基于支持向量机和朴素贝叶斯的组合分类器。文本特征向量采用组合分类器进行分类之后，得到具体的分类结果。在整个分类过程完成之后，需要对分类结果进行评价，以判定模型的分类效果。该文本分类模型整体框架如图 3.1 所示：

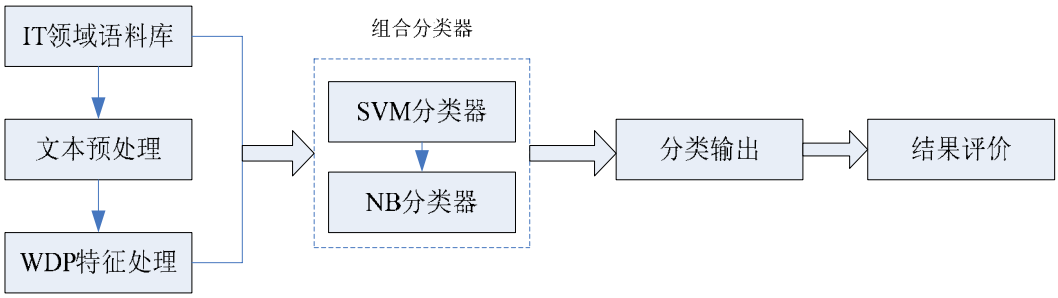

图 3.1 IT 领域文本分类模型框架图

3.2 IT 领域语料库设计

自然语言处理的研究必须以语言事实作为根据，必须详尽地、大量地占有材料，才有可能在理论上得出比较可靠的结论，而这种语言材料的集合就是所谓的语料库。严格来说，语料库就是指在随机采样的基础上收集的有代表性的真实语言材料的集合，是语言运用的样本。现代的语料库是指按照一定的语言学原则收集的具有一定容量的大型电子原始语料文本库或者经过加工后带有语言学信息标注的语料文本。文本分类首先需要做的就是对语料库的采集和维护。

目前国外已经有比较成熟的语料库，如美国Brown大学建立的布朗语料库、路透社的Reuters语料库、医学方面的OHSUMED语料库以及Newsgroups等标准分类语料库，都得到了广泛的应用。国内学者也开发了一些中文语料库，主要有《人民日报》语料库、北京大学的中文语料库、复旦大学建立的中文语料库以及国家语言文字工作委员会建立的汉语语料库等等，这些语料库一般包含了当今文本可能涉及的所有类别，如汽车、财经、IT、健康、体育、旅游、教育、招聘、文化

和军事等等，全面地代表了各个学科的文本分布。但这些语料库大多是收费的，而且主要针对大型的研究项目制定。

考虑到笔者的专业方向以及相关项目经历，而且由于各个类别的文本本身就具有较大的差异性，为了体现文本分类方法的精度和有效性，论文的研究主要把IT领域文本作为分类实验的对象，针对IT领域的文本进行分类实验研究。

由于文本分类器的分类效果与训练文本的选择具有很大的关系，因此，论文的研究根据分类效果调整选入训练集的文本。语料库以搜狗网站上下载的SogouC.reduced 文本分类语料库为基础，搜狗文本分类语料库是由搜狐旗下独立品牌搜狗实验室开发出来的，文本分类语料库来源于 sohu 新闻网站保存的大量经过编辑手工整理与分类的新闻语料及其对应的分类信息，为各种从事中文文本分类工作的研究者提供一个标准的较大规模的研究平台。其分类体系包括十个分类节点，其中 IT 领域中文本规模约一万篇，较为全面地覆盖了所有 IT 领域内各方面的文本。由于搜狗文本分类语料库的文本收录时间为 2005 年，为了使实验文本更具有代表性，笔者在计世网、赛迪网、硅谷动力等其他知名网站的 IT 版块上也下载了一些 IT 类新闻文本作为研究的补充，确保语料库的全面性。论文在选择进入语料库的文本时基于以下原则：

（1）文本必须是知名网站的 IT 类新闻，保证文本确实可以作为用户获取相关信息的渠道；

（2）以提高语料库的代表性为目的，搜集的文本必须能全面的反映 IT 类文本的特点；

（3）对于内容基本相同的文本，只应保留一篇即可，其余的应予以删除。

针对 IT 领域文本，根据相关学者的研究并参考各大网站的分类习惯，论文把 IT 领域的文本集划分为互联网、通信、产业、产品和培训等五类，分别代表互联网、通信企业相关信息、产业新闻、具体产品相关信息和培训教学类文章。在研究这五大类文本下的各个小类时以产品类文本为例，将产品文本集继续划分为手机、相机、PC、家电、软件、硬件设备等六小类。以此作为基本的分类体系框架，进行文本分类的研究。对语料库的维护包括文本类别的添加、删除、索引等操作，语料库中的文本按指定目录结构存储。

3.3 WDP 特征向量的形成

3.3.1 特征提取

在文本分类中，常用的特征提取方法主要是基于阈值的统计方法[39]，因为这

种方法具有计算复杂度低、速度快的优点，特别适合做文本分类中的特征提取。Yang Yiming教授在文本分类的特征提取的研究中取得了非常有代表性的成果[24]。她针对平面文本分类问题，综合分析和比较了DF、IG、MI和CHI等方法，得出了IG和CHI方法效果相对较好的结论。我国的很多学者针对中文的文本分类也做了相应的研究，实验发现CHI方法针对中文的特征提取也有相对较好的效果[40]。因此，论文的研究就采用CHI统计作为文本特征提取的方法。

传统的CHI方法把每个特征项在各个类别中出现的频度与它在整个文本集中出现频度的比率作为该特征项对分类贡献的大小。这种方法是单纯的从频度指标的角度出发的，论文尝试从特征项在语料库中的分布情况来确定特征项的提取。

特征项的分布信息，又可称为特征项频率分布的离散度。离散度可分为类间离散度DI_{ac}(Distribution Information Among Classes)与类内离散度DI_{ic}(Distribution Information Inside a Class)，分别表示特征项在类间与类内文本间的分布差异。

特征项的类间离散度DI_{ac}的计算公式如下：

$$DI_{ac}=\frac{\sqrt{[\sum_{i=1}^{m}(tf_i(T_k)-\overline{tf(T_k)})^2]/(m-1)}}{\overline{tf(T_k)}} \quad \text{式(3-1)}$$

其中，$tf_i(T_k)$表示特征项T_k在第i类中出现的频度，m为类别总数。$\overline{tf(T_k)}$为T_k在各类中出现频度的平均值，计算公式如下：

$$\overline{tf(T_k)}=\frac{1}{m}\sum_{i=1}^{m}tf_i(T_k) \quad \text{式(3-2)}$$

特征项的类内离散度DI_{ic}的计算公式如下：

$$DI_{ic}=\frac{\sqrt{[\sum_{j=1}^{n}(tf_j(T_k)-\overline{tf'(T_k)})^2]/(n-1)}}{\overline{tf'(T_k)}} \quad \text{式(3-3)}$$

其中，$tf_j(T_k)$表示特征项T_k在第j篇中出现的频度，n为类内总文本数。$\overline{tf'(T_k)}$为T_k在各篇文本中出现频度的平均值，计算公式如下：

$$\overline{tf'(T_k)}=\frac{1}{n}\sum_{j=1}^{n}tf_j(T_k) \quad \text{式(3-4)}$$

由以上四式可知：

（1）当特征项只在一个类别中出现时，其DI_{ac}取最大值1，其分类能力最强；当特征值在每个类别中的出现频度都相同时，其DI_{ac}取最小值0，其分类能力最弱。由此可知：DI_{ac}的分布区间为[0,1]，同时特征项的DI_{ac}与分类能力成正比。

（2）当特征项只在一篇文本中出现时，DI_{ic}取最大值1，其分类能力最弱；当

特征项在每篇文本中出现的频度都相同时，DI_{ic}取最小值0，其分类能力最强。

因此，DI_{ic}的分布区间也为[0,1]，但特征项的DI_{ic}与分类能力成反比，故在实际的计算中，可以用(1-$DI_{ic}(T_k)$)来表示。

论文定义特征项分类能力参数：

$$\alpha = DI_{ac} \times (1 - DI_{ic}) \quad \text{式(3-5)}$$

由以上分析可知，当α越大时，表示特征项对文本分类越有用，分类能力也相对越强。

因此，论文提出了一种新的特征抽取算法，记为DI算法：在CHI统计的特征提取基础上，加以特征项分类能力参数的修正，即针对特征项T_k在C_j类中DI值的计算公式如式(3-6)：

$$DI(T_k, c_j) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \times \alpha \quad \text{式(3-6)}$$

再用式(3-7)确定特征项T_k对整个语料库的DI值：

$$DI(T_k) = \max_{j=1}^{m} DI(T_k, c_j) \quad \text{式(3-7)}$$

全部特征词的DI值计算出来之后，将各特征词按DI值从大到小排列，设定一个阈值S，所有DI值大于S的特征词都被选中，作为抽取出的特征词参与特征向量的构成。

3.3.2 特征权重的确定

1．传统方法的不足

根据论文前面所述，目前在中文文本分类领域，大多数研究是以词语作为特征项的。传统的特征权重算法主要考虑特征项的频率信息TF以及反文档频率信息IDF[41]。不同类别的文本在某些特征项的出现频率上有很大差异，因此频率信息是文本分类的重要参考之一。在最初的文本自动分类中，文本向量就是用TF来构造的。

然而单纯使用TF往往会导致一个问题，就是文本中大量出现的停用词或者对分类意义较小的高频词会干扰特征权重的计算。这些高频词在所有文本中出现的频率都比较高，对文本类别意义的贡献度却很小。为了处理这些高频词，有的系统采用了过滤高频词的方法，这样做需要依赖于一个专家构造的高频词词典。然而高频词的界定本身就是一个主观性很强的判断，而且词典在扩充和修改上都需要一定程度的人工干预。因此更为简单合理的处理办法是使用反文档频率。

反文档频率IDF(Inverse Document Frequency)是特征项在文本集分布情况的量化。IDF常用的计算方法为：

$$IDF = \log(N / n_k + L) \quad \text{式(3-8)}$$

其中，L为修正因子，它的取值通过实验来确定。N为文本集中的总文本数，n_k为出现特征项T_k的文本数。IDF算法的核心思想是：在大多数文本中都出现的特征项不如只在小部分文本中出现的特征项重要。IDF算法能够弱化一些在大多数文本中都出现的高频特征项的重要度，同时增强一些在小部分文本中出现的低频特征项的重要度。

一个有效的分类特征项应该既能体现所属类别的内容，又能将该类别同其它类别相区分。所以，在实际应用中，TF与IDF通常是联合使用的。TF与IDF的联合公式如式(3-9)：

$$weight(T_k) = tf(T_{ik}) \times idf(T_k) \quad \text{式(3-9)}$$

在很多情况下还需要将向量归一化，TF-IDF的归一化计算公式如式(3-10)：

$$weight(T_k) = \frac{tf(T_{ik}) \times \log(N / n_k + L)}{\sqrt{\sum_{k=1}^{s} (tf(T_k))^2 [\log(N / n_k + L)]^2}} \quad \text{式(3-10)}$$

传统的特征权重算法存在明显的不足。因为TF-IDF是将文本集作为整体来考虑的，特别是其中IDF的计算，并没有考虑到特征项在类间和类内的分布情况。如果某一特征项在某个类别大量出现，而在其它类别出现很少，这样的特征项的分类能力显然是很强的。但这在TF-IDF算法中是无法体现的。

另一方面，同样是集中分布于某一类别的不同特征项，类内分布相对均匀的特征项的权重应该比分布不均匀的要高，因为如果某一特征项只在某个类别的一两篇文本中大量出现，而在类内的其它文本中出现得很少，那么不排除这一两篇文本是该类别中特例的情况。因此这样的特征项不具备代表性，权重相对较低。对于这种情况，传统的TF-IDF算法也不能很好地处理。

这里通过一个很小的文本集来说明上述问题。假设有三个类别，每个类别各5篇文本，只考虑三个特征项T_1、T_2和T_3。

表3.1为特征项在各篇文本中出现的频率。表3.2为传统TF-IDF算法的权值计算结果。由表3.1中特征项的频数，在此基础上根据式(3-10)（其中L取0.1）得到归一化的TF-IDF权值。

表3.1 各个特征项的出现频率

文本 / 特征项	类1	类2	类3
	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
T_1	4 3 2 5 6	1 1 1 1 1	1 1 1 1 1
T_2	2 2 2 2 2	2 2 2 2 2	2 2 2 2 2
T_3	1 1 1 1 1	3 1 3 1 1	2 2 2 2 2

表3.2 特征权重计算结果（TF-IDF）

特征项 / 类别	T_1	T_2	T_3
	TF-IDF	TF-IDF	TF-IDF
类1	0.873	0.436	0.218
类2	0.348	0.696	0.627
类3	0.333	0.667	0.667

从表3.1中可以看出：T_1在各类中出现的*tf*差别很大，分类能力应该最强；而T_2在各篇文本中的*tf*都相同，对分类基本没有信息贡献，因此其分类能力应该最弱。但从表3.2的结果来看，T_2的权值却非常高。这是因为根据TF-IDF算法的定义，特征项的权重由TF和IDF决定。当文本集中包含特征项T_1、T_2和T_3的文本数相同时，这些特征项的IDF相同，特征项的权重由其TF唯一确定。所以导致表3.2得到了一个极不合理的结果，几乎没有分类能力的被赋予了很高的权值。由此可见，在没有考虑到特征项在类间和类内分布的比例情况时[42]，单纯使用TF-IDF算法会导致很大误差。

2．基于特征项分散度的权重改进方法

根据上述分析，论文研究结合上文提到的特征项分类能力参数参与特征项权重的确定。则权重计算公式用$WD(T_k)$来表示，具体表达式为：

$$WD(T_k)=weight(T_k)\times\alpha=\frac{TF_i\times IDF(t_i)}{\sqrt{\sum_{i=1}^{n}(TF_i\times IDF(t_i))^2}}\times\alpha \qquad \text{式(3-11)}$$

由式(3-11)结合表3.1和表3.2的数据可以计算出相应的DI_{ac}，DI_{ic}（表3.3）以及特征权重（表3.4）。

表3.3 特征项在类间及类内的分布离散度

分布离散度 / 特征项	DI_{ac} (3类间)	DI_{ic}		
		类1	类2	类3
T_1	0.866	0.605	1	1
T_2	0	1	1	1
T_3	0.331	1	0.711	1

表3.4 特征权重计算结果（TF-IDF-DI）

类别 / 特征项	TF-IDF-DI		
	类1	类2	类3
T_1	0.457	0.301	0.288
T_2	0	0	0
T_3	0.072	0.147	0.221

从表3.3和表3.4的结果来看，T_1在类1分布不均匀，因此权值得到削弱；T_2在各类的TF都相同,对分类没有贡献,因此权值为0；T_3在各个类别的TF相近，因此权值较低。这些都体现了DI算法的基本思想。

由于上述算法是将文本集作为整体来考虑的，对所有特征词都统一处理，所以并没有考虑到词语分类能力的强弱。一般来说，各个学科都具有本学科的专业术语，例如在互联网领域的文本集中，“局域网”、“交换机”、“网关”等术语显然要比“论坛”、“流量”等具有更强的分类能力。因此论文考虑将各个类别中的专业术语单独提取出来，人为赋予较高的权重。考虑到非专业术语的分类能力也互不相同，因此权重计算公式可以表示为：

$$W(t_i)=\begin{cases} a & T_k \in \mathrm{D} \\ \dfrac{TF_i \times IDF(t_i)}{\sqrt{\sum_{i=1}^{n}(TF_i \times IDF(t_i))^2}} \times \alpha & T_k \notin \mathrm{D} \end{cases} \quad \text{式(3-12)}$$

其中，D为IT领域的专业词表，a的分布区间为[0,1]，一般取值在0.8以上，具体取值与训练样本集有关。T_k是否属于专业术语需要事先通过一个专业词表判定，由于目前已经有较为成熟且使用广泛的专业词表，所以可以直接对照专业词表进行是否专业词语的判定。

一般来说，出现在文章标题和首末段中的词表达文章主题的能力要比其他正文中的词要强，国内有研究者抽样统计，国内中文期刊自然科学论文的标题与文本的基本符合率为98%，新闻文本的标题与主题的基本符合率为95%，美国一学者进行过统计，反映主题的句子，80%出现在段首，10%出现在段尾。这说明不同位置的词对文本的作用也是不一样的，有些词虽然出现频度不高，但却很能反映文本的特性。因此，对于出现在不同位置的词对分类的贡献也是不同的，在文本分类中有必要对处在标题和首末段的词作加权处理。

针对以上问题，论文引入了词项位置权值λ，对$W(t_i)$进行了改进，将权重计算公式设计为：$WDP(t_i)=W(t_i)\times\lambda$，特征词位置权重系数可根据经验制定：标题权重系数为2，出现在首末段的为1.5，其它部分为1。改进后的权重计算公式为：

$$WDP(t_i)=\begin{cases} \lambda a & T_k \in \mathrm{D} \\ \dfrac{\lambda TF_i \times IDF(t_i)}{\sqrt{\sum_{i=1}^{n}(TF_i \times IDF(t_i))^2}} \times \alpha & T_k \notin \mathrm{D} \end{cases} \quad \text{式(3-13)}$$

其中，D为IT领域的专业词表。

根据前文对文本分类技术中特征项提取与权重确定方法的研究，可以认识到：特征处理是文本分类的关键技术之一，提高特征向量表示文本的准确度，可以有效地提高文本分类的精确度。然而，传统的特征处理方法只利用了特征项在文本中出现的词频和倒文档频率等有限信息，忽视了特征项其他的相关信息，导致特征向量在表示文本时具有一定的片面性。因此，论文设计了一种综合考虑特征项分散度、专业程度和分布位置的特征处理方法，全面地利用特征项在文本中的信

息，得到的特征向量能更好的表达文本。论文把这种特征处理方法称为WDP特征处理方法。

权重确定具体算法如下：

算法3.1 特征提取算法
输入：特征词
输出：该特征词的权重

```
Float getWeight(word)   //计算特征值的权重
{
    If( IfIn(word) )      //判断特征词是否在专业词表内
    {
        //专业词表内的词认为赋予较高权重
        w(ti)=0.8;
        } else {
                //非专业词表内的词计算带有离散度的权重
                w(ti) = TF_i × IDF(t_i) / sqrt(Σ_{i=1}^{n} (TF_i × IDF(t_i))^2) × DI_ac × (1 − DI_ic);
        }

        If( InTitle(word) ) //判断特征词是否在标题位置
        {
        wdp=2*w(ti);
        } else if( InFrontEnd(word) ) {   //判断特征词是否在首末段位置
            wdp=1.5*w(ti);
        } else {
            wdp=w(ti);
        }

        Return wdp;
}
```

3.4 支持向量机与朴素贝叶斯的组合分类器

在传统的文本分类技术中，根据相关学者的研究，支持向量机算法和朴素贝叶斯算法具有较好的分类效果，在文本分类领域中得到了广泛的应用。论文也主要针对这两种分类方法进行研究，下面将详细地介绍论文对这两种分类方法的相关研究。

3.4.1 支持向量机

支持向量机以统计学习理论[43]为基础，避免了传统分类算法中样本无限大的问题，具有很好的泛化性能，精度方面也表现出明显的优势，目前已成功应用于模式识别领域。统计学习理论是一种小样本统计理论，它提出了VC维和结构风险最小化，为研究有限样本情况下的统计模式识别[44]和更广泛的机器学习问题[45]建立了较好的理论框架。下面先介绍VC维和结构风险最小原理。

1．VC维

为了研究学习过程一致收敛的速度和推广性，统计学习理论提出了VC维的概念。模式识别方法中的VC维的直观定义是：对一个指示函数集，如果存在h个样本能够被函数集里的函数按照所有可能的2^h种形式分开，则称函数集能把h个样本打散。函数集的VC维就是它能打散的最大样本数目h。若函数对任意数目的样本都能打散，则称函数集的VC维是无限的。

VC维反映了函数集的学习能力，VC维越大学习能力越强，但机器容量也随之增大，会导致学习过程更加复杂。

2．结构风险最小化原理

预测函数对训练集产生的实际风险主要在于训练模型对训练集和学习算法产生的过度拟合的情况。过度拟合会使模型的预测能力下降，所以机器学习过程中，要尽量避免这种情况的发生，取得更小的实际风险，这样样本会有更好的推广性。这就是结构风险最小化原理。

3．支持向量机

支持向量机(SVM)算法是建立在统计学习理论的VC维理论和结构风险最小原理基础上的，它将原始数据集合压缩到支持向量集合，然后用子集学习得到新知识，同时也给出由这些支持向量决定的规则，并且可得到学习错误的概率上界[46]。假设线性分类面的形式为：

$$g(D)=\omega\square D+b=0 \qquad \text{式(3-14)}$$

其中ω为分类面的权系数向量，b为分类阈值，可用任一支持向量求得，或者通过两类中任一对支持向量取中值求得。将判别函数归一化，使得所有样本都满足|g(D)=1|，即 $y_i[(\omega\square D_i)+b]-1\geq 0$,i=1,2,…,N，y_i是样本的类别标记，即当样本属于类C时y_i=l，否则y_i=-1；D_i是相应的样本。H_l平面表示$\omega\cdot x_1+b>=1$的样本集合，H_2平面表示$\omega\cdot x_2+b<=-1$的样本集合，H为最优超平面，它使得每类距离超平面最近的样本到超平面的距离之和最大。距离这个最优超平面最近的样本被称为支持向量(SupportVector，简称SV)，也就是H_l和H_2上的样本。这样样本的分类间隔就等于2/||w||，设计的目标就是要使得这个间隔值最小。支持向量机原理如图3.2所示：

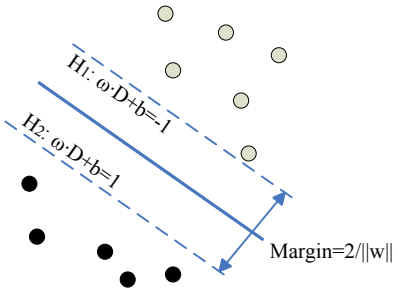

图3.2 支持向量机原理

支持向量机(SVM)方法的主要思想是针对两类分类问题寻找一个满足分类要求的最优超平面，使得这个分类超平面在保证分类精度的同时，能够使超平面两侧的空白区域最大化。目前在文本分类的实际应用中，大多数情况分类属于多类分类问题。多类分类问题和两类分类问题之间存在一定的对应关系，如果一个分类问题中有K类可分，则在K类中任意两类可分。因此，可以用支持向量机来处理多类分类问题，研究表明这种方法是可行的，并且取得了非常好的效果，在分类精度和效率上都取得了令人满意的结果[34]。

3.4.2 朴素贝叶斯算法

朴素贝叶斯分类器是基于贝叶斯学习方法的分类器，是一种基于概率的分类方法。其原理虽然比较简单，但在实际应用中很成功[47]。假定有m个类C_1，C_2,…，C_m，给定未知文本d，贝叶斯分类将给出对于文本d，判定类别C_i的最高后验概率，即最大化$P(C_i|d)$。根据贝叶斯定理可得：

$$P(C_i \mid d) = \frac{P(d \mid C_i) \times P(C_i)}{P(d)} \qquad \text{式(3-15)}$$

明显的，$P(d)$对于所有类是个常数，所以只需最大化$P(d|C_i)P(C_i)$即可。$P(C_i)$可以根据训练集中的类别分布来计算，为了避免$P(C_i)=0$，采用拉普拉斯概率估计有：

$$P(C_i) = \frac{1+|C_i|}{m+|D|} \qquad \text{式(3-16)}$$

其中$|C_i|$为类别C_i包含的文本数，$|D|$为训练集中的文本总数，并且简单的以各个属性在类别C_i上出现的概率来推算$P(d|C_i)$。

朴素贝叶斯分类器关于变量独立性的假设虽然大大减少了参数量，但在实际应用中，这种独立性假设通常是不满足的。经过分析得知，朴素贝叶斯分类器的本质是一种具有很强限制条件的贝叶斯网络分类器，由于它限制条件太强，不太适用于现实应用；然而，完全无限制的贝叶斯网络也是不现实的，因为学习这样的网络非常耗时，其时间复杂度为属性变量的指数级，并且空间复杂度也很高。

因此，论文选定朴素贝叶斯模型的多元模型(Multi-variate Bernoulli Model)[48]，它只考虑特征在文本中是否出现，若出现记为1，否则记为0。多元模型的计算式如式(3-17)：

$$P(X \mid C_i) = \prod_{t=1}^{|V|} (B_{xt} P(w_t \mid C_i) + (1 - B_{xt})(1 - P(w_t \mid C_i))) \qquad \text{式(3-17)}$$

其中 $P(w_t \mid C_i) = \frac{1+C_i\text{中包含特征}w_t\text{的文档数}}{2+C_i\text{所有文档数}}$，$w_t$代表第t个特征(即向量的第t分量)，|V|代表特征总数，B_{xt}表示w_t是否在文本X中出现（出现记为1，否则为0）。

Bayes分类方法在理论上的论证已经非常充分，在实际应用上也广泛使用[49]。

3.4.3 支持向量机与朴素贝叶斯的组合分类器

支持向量机和朴素贝叶斯算法是文本分类算法中比较经典的两种算法，在文本分类中都具有比较优秀的表现。由于它们简单易行，广泛应用于许多实际的分类系统中，并且常常被研究人员作为比较分类算法性能的基准，因此对它们进行研究是有现实意义的。同时，论文也希望本文的研究思路能够对其他的分类方法有所启发。实际上，文本分类领域的学者对这两种算法的研究一直都没有停步。

支持向量机的优点是它稳定的表现和良好的性能，朴素贝叶斯法的长处是分类速度快分类效率高。如何针对IT领域的文本语料库的特点，设计一种能够结合两种分类算法优点的分类器是论文需要思考的问题。首先分别用支持向量机和朴素贝叶斯对IT领域中的一小部分文本进行分类实验，检验它们各自针对IT领域文本语料库的分类效果。进行实验的文本是从IT领域的语料库中抽取的，能够代表实际分类实验的效果，同时在进行小类的分类研究时选取产品类文本进行实验。

表3.5 SVM和NB对IT领域文本分类结果

类别	查全率		查准率	
	SVM	NB	SVM	NB
互联网	0.72	0.68	0.84	0.70
通信	0.74	0.67	0.82	0.69
产业	0.83	0.73	0.79	0.73
培训	0.68	0.66	0.79	0.68
手机	0.72	0.72	0.72	0.80
相机	0.69	0.70	0.73	0.79
PC	0.71	0.73	0.69	0.72
家电	0.69	0.76	0.68	0.69
软件	0.70	0.74	0.72	0.76
硬件设备	0.73	0.75	0.73	0.78

通过表3.5可知，在五大类文本的分类效果上，支持向量机算法的分类效果较

好，查全率和查准率都要高于朴素贝叶斯方法；而在六个小类文本分类的效果上，朴素贝叶斯算法的分类效果较好。论文通过对IT领域中文本内容的分析比较可知：IT领域的五类文本相互之间内容的交叉性相对较大，而产品的六个小类文本之间内容独立性大些。支持向量机由于在抗噪声方面的良好性能，在内容交叉性和干扰性大的五大类文本分类实验中取得了较好的效果。由于支持向量机核函数的选择缺乏指导，难以快速针对具体应用问题选择最佳的核函数，从而造成了计算开销比较大的问题。针对具体的小类分类问题时，支持向量机在训练时难以对各个小类分类情况选择最佳的函数参数，表现并不如在大类分类中出色。而朴素贝叶斯由于有着坚实的理论基础和良好的分类性能，在产品类中六个小类的分类实验中取得了较好的效果。从提高分类精度的角度出发，同时针对IT领域文本分类体系的特点，论文认为可以将两种分类方法相结合，得到一种效果更好的文本分类器。

由于目前文本分类领域的研究都主要是针对分类体系只有一级的情况，即所要划分的类别都是对等的，因此通常情况下分类过程是一次性完成的。通过前文的研究可知：IT领域文本的分类体系是二级的，存在大小类的结构，即文本大类中可以继续划分小类的情况。而大类文本和小类文本具有各自不同的特点，在分类过程中需要有针对性的分别对大类和小类进行处理，此时如果仍旧使用传统的一次分类方法将两种类别的文本统一处理显然有些粗糙。因此，论文考虑分别对文本大类和小类进行处理。

如何在IT领域文本分类中将两种分类方法的优点结合在一起是论文需要考虑的问题。通过上面的实验可以知道：支持向量机在大类文本的分类表现较好，而朴素贝叶斯在小类文本的分类表现较好。因此，论文提出构造一种支持向量机与朴素贝叶斯的组合分类器：即待分类文本首先用支持向量机算法进行分类，若文本划分为可继续分类的类，则继续用朴素贝叶斯进行第二次分类，得到具体的文本类型；若第一次分类结果是不可继续分类的类，则记录所划分的类别，分类结束。具体的流程如图3.3所示。

组合分类器对于IT领域文本分类体系存在的大类和小类的情况分别有针对性地进行了处理，有效地解决了IT领域内大类文本内容交叉性大、而各个小类的文本内容相互独立且分类情况相对较多的问题。实际上，大小类的分类体系并不是IT领域文本所独有的，在大多数领域的文本体系中都有类似的类别结构。因此，组合分类器不仅仅可以运用在IT领域中，在其他领域也可以得到广泛的应用。

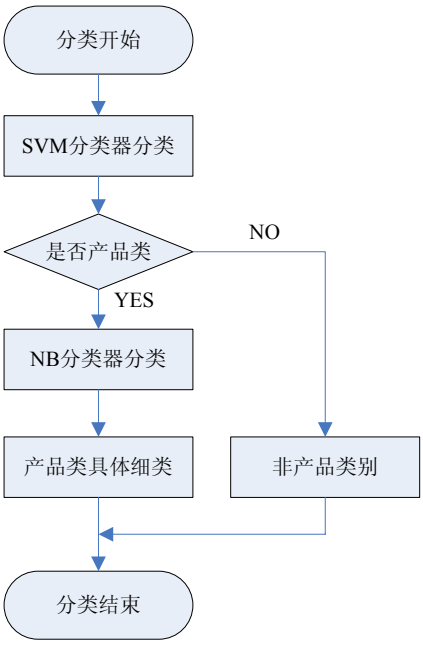

图3.3 SVM和NB组合分类器流程图

3.5 本章小结

本章重点对IT领域文本分类中的特征处理和分类方法进行了系统的研究。首先建立了一个IT领域的文本语料库，为进一步的研究工作提供分类素材；然后在分析传统特征处理方法不足的基础上，提出了改进的特征提取和权重确定方法，综合利用特征项在文本中的各项信息，从而使特征向量能够更好地表示文本的内容；最后针对IT领域的文本语料库构建了一个基于支持向量机和朴素贝叶斯算法的组合分类器，针对IT领域文本存在的大小类的情况做了有益的尝试。论文对IT领域文本分类模型在程序上进行了验证，这一章节是论文的核心部分。

第四章 文本分类系统实验及结果分析

4.1 系统原型的功能模块介绍

论文选择IT领域文本作为研究对象，设计了IT领域文本分类系统。IT领域是近几年来技术发展最快的领域之一，文本量十分巨大。与普通的语料库相比，IT领域语料库有自身的特点，如文本的专业术语更多，新生词汇更多，术语与一般词汇融合分布在文本中，各个类别的文本内容交叉更多等等。论文基本实现了一个完整的中文文本分类系统的主要功能，为研究文本分类相关技术提供了实验平台，也为进一步的研究和完善工作打下了基础。在系统原型的实验平台基础上，验证有关算法的有效性，对提高文本分类的准确率和效率具有一定参考价值。

论文设计的IT领域的文本分类系统原型功能模块如图4.1所示：

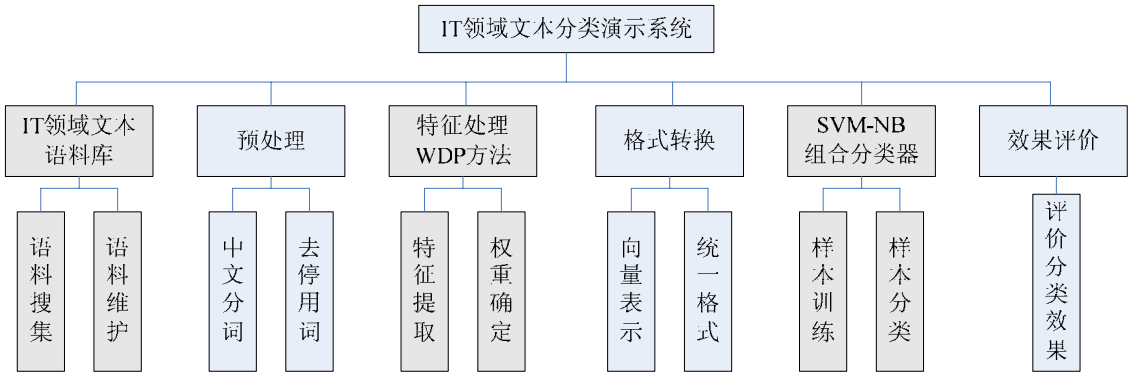

图4.1 IT领域文本分类系统原型功能模块

从图中可以看出，该系统共分六大模块：语料库模块、预处理模块、特征处理模块、格式转换模块、组合分类器模块和效果评价模块。其中各个模块又包含了具体的处理步骤。论文的工作主要集中在文本分类系统原型的整体设计，语料库模块、特征处理模块、格式转换模块和组合分类器模块的设计和实现，以及各个模块间相互集成等方面的工作。其中语料库模块构建了IT领域文本语料库，特征处理模块使用WDP方法实现了特征提取和特征权重的确定，组合分类器模块实现了支持向量机和朴素贝叶斯方法的组合分类器。所有这些模块互相配合，共同完成了IT领域文本分类的功能。

从系统工作流程的角度来说，文本分类演示原型中系统流程如下：

1．将训练文本进行分词处理和停用词去除操作，得到初始的文本特征项信息的特征库；

2. 系统使用训练样本根据特征库中特征项的信息使用WDP方法进行特征处理，完成文本表示操作，获得文本特征向量，用于分类器的训练；

3. 将文本特征向量输入组合分类器，对文本进行分类，即首先用支持向量机

分类器对文本进行分类，再用朴素贝叶斯分类器对文本进行二次分类，从而得到分类结果；

4. 对系统分类的最终结果进行评价，判定系统原型的分类效果。

由此，可以得出系统原型的四大主要模块：文本预处理、WDP特征处理、训练分类和效果评价。系统主要的流程框架如图4.2所示。

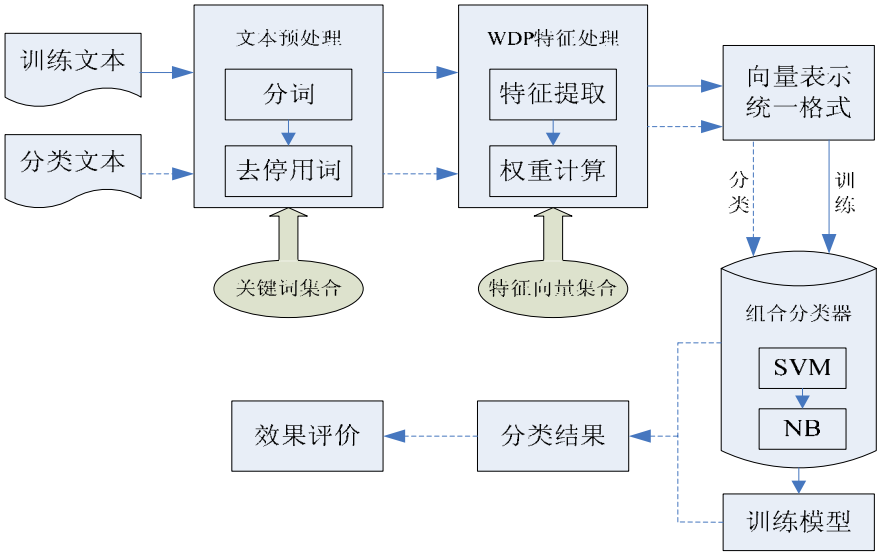

图4.2 文本分类框架流程图

由于论文的系统原型目的在于演示文本分类的一般过程，实现文本分类的基本功能，因此系统的功能相对简单，还有很多需要完善的地方。

4.2 文本分类演示原型

4.2.1 IT 领域的文本语料库

论文语料库采用以搜狗文本分类语料库为基础构建的IT领域的文本语料库[50]，具体的文本类别见表4.1。该语料库均是文本格式的，是由各大网站的IT板块的新闻组成，具有一定的代表性。同时，由于训练文本集的选择对最终分类结果有非常大的影响，因此在实验的过程中，需不断调整进入语料库的文本，以使文本分类系统原型具有较好的分类效果。

表4.1 IT领域文本语料库中的类别分布

文本类别	互联网	通信	产业	产品	培训	总数
实验文本数	50	50	50	100	30	280

其中，产品类别还可以再继续细分为各种小类，具体如下：

表4.2 IT领域语料库中产品类文本的类别分布

文本类别	手机	相机	PC	家电	软件	硬件设备	总数
实验文本数	20	10	20	10	20	20	100

根据文本测试过程所使用的测试语料的不同，测试被分为封闭性测试和开放性测试。当采用的测试语料是训练分类器所用训练语料的一部分或全部时，所作的测试被称为封闭性测试；如果用来测试的语料不曾用作训练语料，这时所作的测试就是开放性测试[51]。论文研究计划进行开放性测试。

4.2.2 文本的预处理

文本格式的中文语料库预处理一般包括分词和去除停用词两个步骤。

中文分词系统是中文信息处理的基础性工程，在每个信息应用领域为每个系统建立一个分词系统是没有必要也是不可能的，那样，不仅要耗费巨大的时间和精力，而且分词的效果也不一定比现有的好。论文研究的主要目的是中文文本的分类，在特征提取、文本表示等过程中都需要分词。鉴于前面所说的原因，论文研究直接采用中科院提供的分词系统。

中科院在多年的研究基础上，耗时一年研究出了汉语词法分析系统ICTCLAS(Institute of Computer Technology Chinese Lexical Analysis System)[52]，该系统提供的功能有：中文分词、词性标注和未登录词识别。分词正确率高达97.58%，未登录词识别查全率均高于90%，其中中国人名的识别查全率接近98%，处理速度为31.5Kbytes/s。论文研究使用ICTCLAS软件构建文本分类关键词库。

在进行中文分词操作之后，还需要去除停用词，以获得更为准确的关键词集。在这一步中，通常去除的词语一般包括介词、副词、感叹词等，还要去除标点和空格等冗余符号。

4.2.3 WDP 特征处理

构成文本的词汇，数量是相当大的，因此，表示文本的向量空间的维数也相当大，可以达到几万维。因此研究需要进行维数压缩的工作，这样做的目的主要有两个：第一，可以提高程序的效率，提高运行速度；第二，所有几万个词汇对文本分类的意义是不同的，一些通用的、各个类别都普遍存在的词汇对分类的贡献小，在某特定类中出现比重大而在其他类中出现比重小的词汇对文本分类的贡献大。

目前有一些软件可以实现特征项的提取，但大多数软件的抽取依据是特征词在文本出现的频次，即词频TF。用户可以在程序中设置特征项的出现次数，例如可以设置(1，5)即出现频次在1次和5次之间的词汇可以选入特征库。这种方法是初级的特征提取的方法，具有比较大的片面性，文本中出现次数较多但对分类意义较小的高频词会干扰特征项的选择。而且，这种特征项提取方法提取的特征项数

量巨大，动辄达到一万词以上，分类效率大大地降低了。有关研究表明特征项数量应该控制在1000左右分类效果较好。因此，为了提高分类精度，应去除那些表现力不强的词汇，筛选出针对该语料库的特征项集合。根据前文的研究分析，即保留那些类间离散度大，类内离散度小的特征项。

论文研究提出了一种新的特征处理的方法WDP方法，包括特征提取以及特征权重的确定。实验证明该方法是切实有效的。因此，论文使用该方法进行特征处理，参与特征库的建立。

4.2.4 格式转换

格式转换是要将文本格式转换为分类器可以识别的格式，主要包括两个步骤：向量表示和统一格式。由于论文研究采用使用广泛且发展较为成熟的向量空间模型，在向量空间模型中，文本由一组代表给定特征词项相关性的向量来表示，这个过程称为向量化，也可称为向量表示。由于各篇文本的格式不同，为了便于分类处理，需要将所有的向量统一格式，转换为分类算法可以直接处理的格式。

论文研究采用基于Java的开源软件Wvtool(word vector tool)[53]完成特征词的向量化工作，Word Vector Tool(Wvtool)是一种灵活的Java类库，它可以在向量空间模型构造文本文档词语的向量表达式。向量化是一个在文本处理应用中非常重要的环节，在文本分类和信息检索中都是必不可少的步骤。Wvtool有几个重要特征，如是开源的，提供API接口，易于配置和使用，支持文件形式多样等。

Wvtool的开发目的是提供一个易于使用和扩展的创建特征项向量的Java类库，它可以很方便地加入到Java应用程序中。Wvtool在相对复杂的语言材料与实际的文本信息处理之间找到了联接方法。论文使用Wvtool作为类库完成了中文文本向量化的工作，取得了非常好的效果。

在文本向量化后，还需要将各篇文本的向量文件进行处理，得到统一的格式。以便顺利完成训练和分类过程。

4.2.5 训练及分类过程

训练文本的过程是系统构造分类器的过程，是文本分类系统的核心。系统原型重点实现论文构造的支持向量机和朴素贝叶斯的组合分类器，因此需要分别构造支持向量机和朴素贝叶斯的分类器。首先将经过向量化的IT领域的训练集文本输入支持向量机分类器，经过分类之后得到各个大类的分类结果；再将划分为“产品”类的文本输入朴素贝叶斯分类器进行分类，从而得到“产品”类别下的各个小类的分类结果。整个训练文本的过程就是构造组合分类器的过程。文本训练的

结果是产生了文本的训练模型，训练模型是进行分类操作必不可少的条件，是指导对新文本进行分类的重要依据。分类器在依照训练模型对未知类型的新文本进行分类之后可以得到预测的分类结果，分类过程就此完成。

论文构造的组合分类器将支持向量机分类器和朴素贝叶斯分类器有机地结合起来，针对IT领域文本分类语料库的文本特点进行分类。以产品类文本为例，经过特征处理的向量化文本先使用支持向量机分类器进行分类，如果属于“产品”类别再使用朴素贝叶斯分类器分类，从而得到分类结果。整个分类过程分为两个步骤，根据两种分类器的性能分别加以有效的利用。

论文研究以开源的Weka软件为基础来实现文本分类[54]。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年由位于新西兰的怀卡托智大学开发，最初的软件采用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国际会议上，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已有11年的发展历史，目前每月下载次数已超过万次。

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。开发者可使用Java语言，利用Weka的架构开发出更多的数据挖掘算法。论文的研究正是在Weka的基础上实现了文本分类系统原型的分类模块。

论文研究所采用的支持向量机分类器利用LibSVM库文件构造。LibSVM是台湾大学林智仁教授等开发设计的一个简单、易于使用和快速有效的SVM模式识别与回归的软件包，它不但提供了编译好的可执行文件，而且还提供了源代码，方便改进。LibSVM对SVM所涉及的参数调节相对比较少，提供了很多的默认参数，利用这些默认参数就可以解决很多问题。因此LibSVM在实际的研究中得到了广泛的应用。在论文的研究中，笔者把LibSVM作为Weka的插件使用，取得了非常好的分类效果。

由于Weka本身就带有朴素贝叶斯分类器，因此论文直接在Weka的基础上进行朴素贝叶斯分类器的构造，与支持向量机分类器相结合，实现了文本分类系统原型的组合分类器。

Weka可以处理的数据文件格式为一种特有的数据文件结构arff格式，它是一种类似于二维表的结构形式。在用Weka进行数据处理的时候，首先需要将数据文件转化为arff格式。在文本分类研究领域中，一般是先将文本进行向量

化处理，把文本转化为一组特征向量。一组特征向量就是arff数据文件的一个实例，在一个arff文件中可以有若干个实例，对应有若干组特征向量，即可以包含若干篇文本。也就是说可以通过一个arff文件来表示一个文本集。因此，在用系统原型进行分类之前，需要将待处理的文本转化为arff格式，这一步骤应该在格式转换模块完成。

综上，训练分类模块是通过组合分类器对训练文本集进行训练，得到对应的训练模型，再用该训练模型对未知类别的文本集进行分类预测，从而得到文本分类的结果。

4.2.6 分类效果评价

文本分类本质上是一个映射过程，评价性能能够反映分类系统映射的准确程度，对于比较各种特征提取算法、权值计算方法和分类方法都有着重要的作用。如前文所述，文本分类主要的评价指标一般包括查全率、查准率和F_1指标，以及针对整体的评价指标微平均和宏平均。其中查全率、查准率和F_1指标是在经常使用的基础型指标，针对查全率和查准率相互制约、难以平衡的状况，有学者提出了F_1指标。通常情况下使用这三种评价指标就可以较为全面的评价文本分类系统的分类效果。

具体在IT领域的文本分类系统原型中，论文采用查全率、查准率和F_1指标进行分类效果评价。在训练过程中根据各评价指标可以调整训练模型的建立；而在分类过程中根据评价结果可以评估文本分类模型的分类效果。

4.3 实验结果与分析

论文使用了Java作为开发语言，Windows XP作为编程环境，在赛扬1.6GHz，512M内存的PC上实现了WDP特征处理，包括特征提取和特征项权重的确定，并且实现了支持向量机和朴素贝叶斯组合分类器。下面将列出相关实验结果，论文将特征处理和组合分类器区分开来分析。实验采用开放测试集，计划从预处理好的文本中提取20%作为测试文本集，剩余的80%作为训练文本集。

4.3.1 WDP 特征处理实验结果

WDP特征处理的实验分为两步：首先分析WDP算法得到的权重值；然后再用WDP算法处理得到的特征向量进行分类，检验分类效果。

实验一，测试WDP特征权重计算方法的效果。这里通过一个很小的文本集进

行WDP特征处理的实验，将经过改进的特征权重计算结果与传统的权重计算结果进行比较。这个小型的文本集是从IT领域的文本分类语料库中抽取出来的，并做了一定程度的抽象，以求能够尽量简洁地说明问题。假设有三个类别，每个类别各有3篇文章，只考虑五个特征项T_1、T_2、T_3、T_4和T_5，其中T_4为一个类别的专业词。

表4.3 各个特征项出现的频率

文本 / 特征项	类1			类2			类3		
	1	2	3	1	2	3	1	2	3
	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$	$P_1P_2P_3$
T_1	1 2 2	1 2 3	0 3 1	0 1 0	0 0 1	0 0 0	0 1 0	0 0 0	0 1 0
T_2	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1	0 1 1
T_3	0 0 1	0 1 1	0 4 1	1 0 1	0 1 1	1 2 2	0 0 1	0 3 2	0 1 1
T_4	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0	1 1 0	1 2 0	0 2 0
T_5	1 1 0	1 1 0	1 0 1	1 1 0	1 2 0	1 0 1	1 1 1	1 1 0	1 1 0

其中，P_1表示特征词位于标题位置，P_2表示特征词位于正文首末段位置，P_3表示特征词位于正文非首末段位置。若文章只有一段，则全部属于P_3部分；若文章有两段，则第二段属于P_3部分。

表4.3为特征项在各篇文章中出现的频率，具体到在文章各个位置的分布情况；表4.4为传统的TF-IDF算法的权值计算结果；表4.5为用WDP算法的特征权值计算结果。

表4.4 传统特征权重计算结果

文本 / 特征项	类1			类2			类3		
	1	2	3	1	2	3	1	2	3
T_1	0.985	0.986	0.970	0.154	0.090	0	0.063	0	0.034
T_2	0.114	0.095	0.140	0.371	0.324	0.485	0.152	0.055	0.082
T_3	0.057	0.095	0.140	0.557	0.487	0.728	0.152	0.028	0.082
T_4	0	0	0	0	0	0	0.512	0.397	0.478
T_5	0.114	0.095	0.140	0.371	0.487	0.485	0.228	0.055	0.082

表4.5 WDP特征权重计算结果

文本 / 特征项	类1			类2			类3		
	1	2	3	1	2	3	1	2	3
T_1	1.312	1.501	1.270	0.643	0.562	0	0.264	0	0.141
T_2	0	0	0	0	0	0	0	0	0
T_3	0.025	0.034	0.050	0.368	0.286	0.534	0.057	0.010	0.026
T_4	0	0	0	0	0	0	1.400	1.300	1.200
T_5	0.172	0.143	0.181	0.384	0.526	0.471	0.222	0.062	0.093

将表4.4和表4.5中特征权重计算结果进行比较，逐一分析各个特征词的权重值可得：与传统的特征权重计算方法相比，WDP算法确定的权重不同之处在于特征词T_1在各类别出现的频率差别较大，且T_1在各个类别中分布较为均匀，因此分类能

力较强，权重值得到提升；特征词T_2在各个类别中出现的频率相同，可以说基本没有分类能力，因此权重值为0；特征词T_3在各个类别的频率接近，但在类别中的分布不太均匀，分类能力较弱，权重值有所下降；特征词T_4属于专业词，具有较强的分类能力，T_4的权值也增大了；T_5由于处在文本的标题位置较多，因此也具有较强的分类能力，通过比较，WDP算法中T_5的权重也有所增加。

通过实验可知，WDP算法可以有效强化特征词的分类能力，表现在该算法可以增加具有较大分类能力的特征词权重，相应地减小分类能力较小的特征词权重。因此，使用WDP算法可以弥补传统权重计算方法忽视特征词项分布和专业程度等的不足，可以形成更为准确的特征向量，从而更好地表示文本。

实验二：测试三种特征权重确定算法，即传统的TF、TF-IDF算法以及WDP算法，然后使用支持向量机法分别对这三种方法处理得到的文本向量进行分类。实验从IT领域文本抽取实验样本，采用了几种不同数量的测试集测试各种特征加权算法的查准率。具体分类情况如下：

表4.6 各种加权算法的查准率

加权算法 / 文章数量	分类查准率		
	A_1	A_2	A_3
56	72.6	80.1	82.4
112	75.9	82.3	84.2
168	79.3	82.9	84.7
224	80.4	83.5	85.5
280	81.3	85.0	86.2

其中，A_1：传统TF算法，A_2：传统TF-IDF算法，A_3：WDP算法

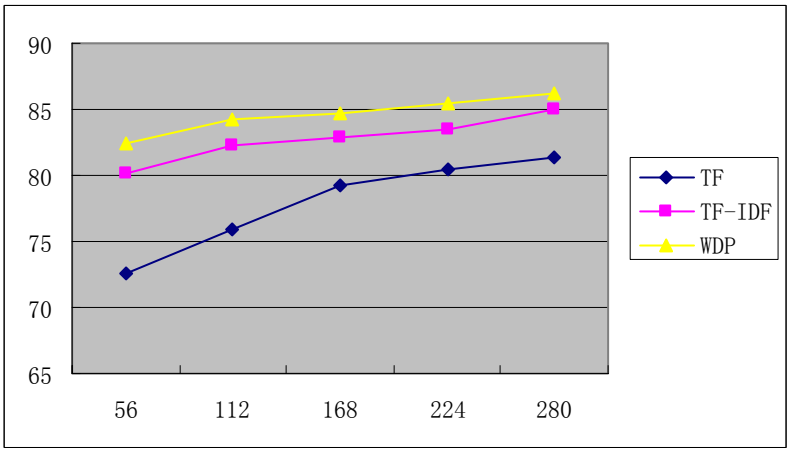

图4.3 表4.6中各种加权算法的查准率

通过表4.6可知，在所有的权重计算方法中，传统的TF算法的分类精确度最差，TF-IDF算法要优于TF算法，WDP算法的分类精确度最高，这个结果符合研究的预期估计。由于WDP算法考虑到特征词的离散度、在文本中所处的位置以及特征词的专业程度等，全面地利用了特征词在文本中的信息，可以更准确地表示文本，

同等条件下分类效果显然要更好。同时注意到，参与测试的文本数越多，分类精确度就越高。

特征加权算法的选择对文本自动分类系统的精度有很大的影响。通过比较传统的加权算法，并在分析其不足的基础上，论文提出了一种考虑类间及类内分布信息的特征加权改进算法。实验结果证明，改进的加权算法与传统算法相比，在分类的精度上有更好的表现。

4.3.2 组合分类器实验结果

实验三：测试组合分类器的分类效果，即分别测试支持向量机分类、朴素贝叶斯分类以及支持向量机与朴素贝叶斯分类的组合分类器，分类具体情况如下：

表4.7 三种分类方法的实验结果

类别	查全率			查准率			F1值		
	SVM	NB	SN	SVM	NB	SN	SVM	NB	SN
互联网	0.72	0.68	0.76	0.84	0.70	0.86	0.775	0.690	0.807
通信	0.74	0.67	0.78	0.82	0.69	0.87	0.778	0.680	0.833
产业	0.83	0.73	0.85	0.79	0.73	0.85	0.810	0.730	0.850
培训	0.68	0.66	0.72	0.79	0.68	0.83	0.731	0.670	0.771
手机	0.65	0.82	0.84	0.69	0.80	0.89	0.670	0.810	0.864
相机	0.75	0.82	0.85	0.73	0.85	0.88	0.740	0.835	0.865
PC	0.71	0.82	0.87	0.72	0.84	0.90	0.715	0.830	0.885
家电	0.66	0.76	0.83	0.69	0.80	0.86	0.675	0.779	0.845
软件	0.69	0.83	0.85	0.71	0.81	0.84	0.700	0.820	0.845
硬件设备	0.71	0.84	0.88	0.75	0.86	0.92	0.729	0.850	0.899

其中，SVM表示支持向量机法，NB表示朴素贝叶斯法，SN表示支持向量机和朴素贝叶斯的组合分类器。

通过表4.7可知，从整体上来看，三种分类方法的查准率要略高于查全率，其中支持向量机算法要比朴素贝叶斯算法的表现好。而支持向量机和朴素贝叶斯的组合算法的表现是三者当中最好的，不管在查全率还是查准率方面，都表现出较好的准确性和精度。通过对IT领域的文本分类语料库研究可知，各个大类之间的文本内容交叉较多，相互之间有较大的干扰性，因此使用抗干扰性较好的支持向量机分类器进行分类取得了较好效果。同时注意到：在使用支持向量机分类器对“产品”类别中的小类进行分类时，查全率和查准率都比对大的类别分类时有所下降；而朴素贝叶斯分类器在对小类分类时表现要比大类分类时更好；组合分类器对小类分类的效果也较好。因此，论文推断，在文本内容相对独立的情况下，朴素贝叶斯分类器的分类效果要更好。组合分类器结合了两种经典的分类器的优点，表现出非常好的分类效果。

文本分类技术的关键是文本分类方法的选择，论文主要研究了三种文本分类方法，并从实验上进行了比较。由实验结果可知：不管从查全率还是查准率上来看，论文提出的组合算法的分类效果都是三种方法当中最优的。

4.4 本章小结

本章是论文中的实证研究部分，在第三章提出的理论模型的基础上，设计并实现了一个IT领域的文本分类系统原型，基本实现了文本分类的主要功能。同时利用该文本分类系统原型进行了相关实验，对论文前面章节的理论研究进行了验证。实验结果证明论文研究提出的特征提取和组合分类器方法是切实有效的。

第五章 总结与展望

面对互联网上日益膨胀的信息，如何快速、准确地从浩瀚的信息资源中寻找到所需特定领域内的相关内容就成为一项非常有意义的课题。文本分类可以在很大程度上解决目前网上信息杂乱的现象，方便用户准确地定位所需的信息进而对信息进行处理。因此，文本分类已成为一项具有很大实用价值的关键技术，是组织和管理数据的有力手段。

5.1 论文总结

论文的研究主要进行了五个方面的工作：

（1）论文对中文文本分类涉及的技术进行了深入的研究，包括中文分词、向量空间模型、特征抽取、特征项权重算法、分类算法和分类效果评价。

（2）构建了IT领域的文本分类语料库，将语料库分为五大类，并以“产品”类文本为示例将其继续划分为六个小类，表现IT领域文本语料库的分类体系，为实验提供了分类素材。

（3）针对传统的特征权重计算方法只简单地考虑了特征项出现的词频，而忽略了特征词在语料库中的分布这一情况，论文引入了类内及类间离散度的概念，并对特征词的专业程度和分布的位置进行相应的加权处理，使得经过该方法计算的特征项权重可以更准确地表示文本的内容。实验证明，这种方法的分类效果要好于传统的特征项权重计算方法。

（4）构建了支持向量机和朴素贝叶斯的组合分类器，针对IT领域文本语料库进行了分类实验，取得了较好的分类效果。

（5）提出并实现了一套完整的基于向量空间模型的中文文本分类系统原型，为研究分类技术提供了实验平台。

5.2 研究前景展望

由于时间限制，论文的研究内容还有以下几个方面的工作值得进一步深入研究：

（1）文本分类系统原型的进一步完善，系统原型实现了文本分类的基本功能，但在页面和具体的功能方面还显得比较简单，应该进一步加以完善，提供一个功能更为全面的文本分类的研究平台。

（2）向量空间模型的进一步研究，需要考虑的不仅仅是向量空间模型如何计

算权重等传统问题，还应该考虑是否能够提出更好的文本表示方法，从根本上改进信息处理的性能。

（3）特征处理算法和文本分类算法是文本分类系统的核心，应进一步研究出更好的特征处理算法和文本分类算法，使分类系统具有更高的准确度和效率。

（4）引入概念空间，不仅能够大大降低特征维数，提高文本分类效率，还能有效的过滤噪声，提高文本分类的准确度。

致 谢

两年多硕士研究生的学习生活即将结束，在毕业论文完成之际，谨向在我攻读硕士学位期间指导、关心、帮助过我的老师、同学、家人和朋友致以诚挚的感谢。

首先要衷心感谢我的恩师焦艺教授。焦老师在学习上和生活上给予了我精心的指导和无微不至的关怀。在读研期间，焦老师帮助我树立了正确的研究方向和科学的学习方法，并为我提供了良好的研究环境和项目实践机会。焦老师渊博的知识、严谨的治学态度、平易近人的性格以及高尚的品格令我钦佩。尤其这篇论文的写作更是在焦老师的悉心指导下完成的，焦老师的谆谆教诲使我终生受益。在此向焦艺老师致以最诚挚的感谢！

其次特别要感谢王雁教授，她在我的学习、论文写作以及生活中都给与了很多的帮助。王老师渊博的知识、一丝不苟的治学态度、宽以待人的风格以及学习、生活中真诚的教诲，令我难以忘怀，受益匪浅！

还要感谢我在信息产业部电子情报所实习时的胡希敏老师、张建立老师、张海永老师、周伯行总工、陈忠总工、文静、孙立立、朱烨、胡彬等各位老师，你们在我实习期间都给予了大量的帮助和指导，谢谢你们！

感谢丁振国教授和邵月凤老师在百忙之中对我的深切关怀和悉心指导。在论文的选题与写作过程中，丁老师都给予了我很多的帮助。

还要感谢赵捧未教授和刘东苏教授，在硕士研究生的学习过程中，给予了我很多指导，并为我的论文提出了有益建议。

我要感谢在学习期间曾给予我无私帮助的同学们，尤其是陈荣、葛永娇、吴永亮、陈利玲、史超、胡志芳、刘觐、曹菲、邹涛、殷利梅等人，与同学们一起共度的研究生时光让我十分难忘。感谢一直都支持关爱我的家人、朋友们。

最后，衷心感谢论文评审委员会的各位老师对我论文评阅付出的辛勤工作。

参考文献

[1] 韩家炜,孟小峰,王静等.Web 挖掘研究[J].计算机研究与发展,2001,38(4):405-414

[2] 张华平.中文信息处理技术发展简史[EB/OL].http://www.nlp.org.cn,中国科学院计算技术研究所软件实验室.2002

[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] 甘立国.中文文本分类系统的研究与实现[D].北京化工大学,硕士论文,2006

[6] 王涛.文本自动分类研究[J].图书馆学研究,2007.12:40-44

[7] 庞剑锋,卜东波,白硕.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

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

[9] 陈治纲.基于向量空间模型的文本分类系统研究与实现[D].天津大学,硕士论文,2005

[10] 马玉春,宋瀚涛.Web 中文文本分词技术研究[J].计算机应用,2004,24(4):134-135

[11] WinterWen.中文搜索引擎技术揭秘：中文分词[EB/OL]. http://www.stlchina.org/twiki/bin/view.pl/Main/SESegment,2005.

[12] Zou F,Wang F L,Deng X T, etc.Stop Word List Construction and Application in Chinese Language Processing[J].WSEAS Transactions on Information Science and Application,2006,3(6):1036-1044.

[13] Ricardo Baeza-Yates,Berthier Riberiro-Neto.Modern Informaiton Retrieval[M]. 北京:机械工业出版社,2004,2.

[14] Verhoeff J, Goffmann W, Jack Belzer. Inefficiency of the use of Boolean functions for information retrieval systems[J].Communications of the ACM. 1961, 4(12):557-558.

[15] Salton G,Wong A,Yang. CS.A Vector Space Model for Automatic Indexing[J]. Communication of the ACM,1975,18(1):613-620.

[16] Gerald Salton. Automatic Information Organization and Retrieval[M]. Addison-Wesley,Reading PA,1968.

[17] 张毅波.中文结构化信息检索系统的研究与实现[D].中国科学院软件研究所,博士论文,2001

[18] Croft W B.Document Representation in Probabilistic Models of Information Retrieval[J]. JASIS,1981,32(6).

[19] 周强.基于语料库和面向统计学的自然语言处理技术介绍[J].计算机科学,1995,22(4):36-40

[20] Harman,D.K.Cavanar W B.N-Gram-Based Text Filtering For TREC-2[C], In Proceedings of The Second Text Retrieval Conference(TREC-2),1993

[21] 周新栋,王挺.基于 N 元语言模型的文本分类方法[J].计算机应用,2005,25(1):11-14

[22] 李纲,郑重.应用于信息检索的统计语言模型研究进展[J].情报理论与实践,2008,31(3):471-476

[23] 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(1):26-32

[24] Y.Yang.A comparative Study on Feature Selection in Text Categorization[C]. Proceeding of the Fourteenth International Conference on Machine Learning (ICML'97), 1997

[25] 蓝海洋,周杰韩,张和朋.文本索引词项相对权重计算方法与应用[J].计算机工程与应用,2003,15：68-70

[26] G.Salton, M.J.McGill. An introduction to modern information retrieval[J]. McGraw-Hill, 1983

[27] G.Salton, C.Buckley. Term weighting approaches in automatic text retrieval[J]. Information Processing and Management,24(5):513-523,1998

[28] C.Buckley,G.Salton,J.Allan and A.Singhal. Automatic Query Expansion Using SMART:TREC 3[C],In Proc.3rd Text Retrieval Conference,1994

[29] S.T.Dumais, Improving the retrieval information from external sources[J]. Behaviour Research Methods, Instruments and Computers,23(2):229-236,1991

[30] 庞剑锋.基于向量空间模型的自反馈的文本分类系统的研究与实现[D].中国科学院计算所,硕士论文,2001

[31] 于瑞萍.中文文本分类相关算法的研究与实现[D].西北大学,硕士论文,2007

[32] Han J,Kamber M.数据挖掘概念与技术[M].孟小峰等译.北京:机械工业出版社,2005

[33] 王煜.基于决策树和 K 最近邻算法的文本分类研究[D].天津大学,博士论文,2006

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

[35] 李斗,李弼程.一种神经网络文本分类器的设计和实现[J].计算机工程与应用,2005,41(17):107-109

[36] Vapnik V.Nature of Statistical Learning Theory(2nd edition)[M].New York:Springer Press,2000

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

[38] K.Van Rijsbergen.Information Retrieval[M].Butterworths, London.1979

[39] 寇苏玲,蔡庆生.中文文本分类中的特征选择研究[J].计算机仿真,2007,24(3):289-291

[40] 郑伟,王锐.文本分类中特征提取方法的比较和研究[J].河北北方学院学报（自然科学版）,2007,23(6):51-55

[41] James Auen.Natural Language Understanding[M].The Benjamin/Cummings Publishing Company, 1991

[42] 鲁松,李晓馨,白硕等.文档中词语权重计算方法的改进[J].中文信息学报,2000,14(6):8-20

[43] Vapnik V 著,张学工译.统计学习理论的本质[M].北京:清华大学出版社,2000

[44] Nello Cristianini, John Shawe-Taylor 著.李国正,王猛,曾华军译.支持向量机导论[M].北京:电子工业出版社,2004

[45] Ian H.Witten, Eibe Frank 著,董琳等译.数据挖掘实用机器学习技术[M].北京:机械工业出版社,2006

[46] 谭冠群,丁华福.支持向量机方法在文本分类中的改进[J].信息技术,2008,1:83-85

[47] 张丽霞，赵大宇.一种扩展的朴素贝叶斯分类器改进算法[J].计算机技术与发展,2006,16(5):28-30

[48] 罗海飞,吴刚,杨金生.基于贝叶斯的文本分类方法[J].计算机工程与设计,2006,27(24):4746-4748

[49] 蒲筱哥.自动文本分类方法研究述评[J].情报科学,2008,26(3):469-465

[50] 搜狗文本分类语料库[EB/OL].http://www.sogou.com/labs/dl/c.html

[51] 胡燕.Web 信息内容及其特征提取方法研究[D].河北农业大学,硕士论文,2008

[52] ICTCLAS 汉语分词系统[EB/OL]. http://ictclas.org/

[53] The Word & Web Vector Tool[EB/OL].http://wvtool.sourceforge.net/

[54] Weka Site[EB/OL]. http://www.cs.waikato.ac.nz/ml/weka

读研期间科研成果

在硕士研究生期间取得的研究成果如下：

[1] 刘依璐,陈利玲.Web 页面主题提取技术综述.现代图书情报技术,2008,第 166 期.

西安電子科技大學

地址：西安市太白南路2号
邮编：710071
网址：www.xidian.edu.cn