中山大學
SUN YAT-SEN UNIVERSITY

本 科 生 毕 业 论 文（设计）

题　　目：中文文本分类中的特征选择研究

院　　系：信息科学与技术学院

专　　业：网络工程

学生姓名：张广腾

学　　号：05374039

指导教师：衣杨 副教授

（职 称）

二〇 〇九 年 五 月

附表一、毕业论文开题报告

论文（设计）题目：中文文本分类中的特征值选取研究

目的：

随着信息技术不断前进和互联网技术的迅猛发展和普及，信息呈近乎爆炸的形式急速膨胀。无论网络上、企业中或是个人系统上，都有海量的信息需要处理。文本作为计算机系统中信息的最重要表现形式之一，其增长速度更为惊人。如何在海量文本库中搜寻、过滤和管理这些文本成为一个亟待解决的问题。作为数据挖掘技术的重要手段之一，基于机器学习的文本分类技术可以在较大程度上解决文本库杂乱无章的现象，帮助人们将大量的文本自动分门别类，从而更好地把握文本信息，使信息的价值最大化。

在采用向量空间模型对文本进行表示的情况下，文本分类的最大特点和困难之一是特征空间的高维性和文档表示向量的稀疏性。中文的词条总数有二十多万条，寻求一种有效的特征抽取算法，降低特征空间的维数，提高分类的效率和精度，成为文本自动分类中需要首先面对的重要问题。特征选择是解决这个问题的有效方法。

本选题的核心目的就在于研究如何进行特征项的选取，使得分类的效率和效果最好。

思路：

首先需要理解中文文本分类技术以及应用的框架，熟悉中文文本分类技术的各个组成部分，然后搭建一个中文文本分类的辅助平台（包括分词组件、分类器、测试文档集、训练文档集，大部分都可以从开源软件或公开资料里获得），在辅助平台的基础上研究特征相的提取并用实验检验之。

方法：

通过阅读大量的资料或文档学习所要用的知识和技术，并通过实验验证自己的想法和理论。

相关支持条件：

PC、java或C++开发环境、中文文本分类辅助平台（自己搭建）、Internet

进度安排：

第一个两月：完成相关理论的理解并搭建中文文本分类辅助平台。

第二个两月：对特征值选取技术做针对性研究，并期待能提出效果更好的选取方案。

第三个两月：延续上两个月的工作并完成毕业论文的整理和编写。

学生签名：　　年　月　日

指导教师意见：

1、同意开题（ ） 2、修改后开题（ ） 3、重新开题（ ）

指导教师签名：　　年　　月　　日

附表二、毕业论文过程检查情况记录表

指导教师分阶段检查论文的进展情况（要求过程检查记录不少于 3 次）：

第 1 次检查

学生总结：

初步完成了对中文文本分类的论文综述，包括学习了中文文本分类的问题原型，明确了中文文本分类中特征选择的研究意义，了解了中文文本分类在国内外的大体发展，并研究了当前系统存在的缺陷，确定了本文研究的切入点。上完成了论文的开始工作。

指导教师意见：

第 2 次检查

学生总结：

深入学习了中文文本分类中使用的相关技术，完善了本文综述。学习了包括中文分词、特征权重赋值、文本分类器及其评价方法，并重点研究了特征选择算法，完成本文的第二章与第三章的一部分。在基本掌握了中文文本分类过程所用技术后，开始了实验平台的搭建，并完成部分搭建工作。

指导教师意见：

第 3 次检查

学生总结：

继续深入学习中文文本分类的相关技术以及本文的研究重点特征选择方法，同时完成了本文实验平台的搭建工作。在上述所有工作的积累下，提出了多项特征选择方法的改进设想，并利用已搭建完成实验平台对改进的特征选择方法进行验证，并确定本文算法的该进方案。

指导教师意见：

第 4 次检查

学生总结：

继续对最后确定的特征选择的改进算法进行实验，并展开论文的整理与完善工作，根据实验结果完成了论文的第四章并对了论文进行总结，按进度计划完成了初稿。

指导教师意见：

学生签名：　　年　　月　　日

指导教师签名：　　年　　月　　日

总体完成情况	指导教师意见： 1、按计划完成，完成情况优（ ） 2、按计划完成，完成情况良（ ） 3、基本按计划完成，完成情况合格（ ） 4、完成情况不合格（ ） 指导教师签名：　　年　　月　　日

附表三、毕业论文答辩情况

答辩人		专 业	
论文（设计）题目			
答辩小组成员			

答辩记录：

记录人签名：　　年　　月　　日

学术诚信声明

本人所呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。本毕业论文的知识产权归属于培养单位。本人完全意识到本声明的法律结果由本人承担。

本人签名：　　日期：

摘　要

信息技术的迅猛发展与互联网的快速普及引发了信息的爆炸性增长。文本作为最重要的电子数据形式之一，增长速度更为惊人。为了从规模庞大的文本集里获取有用的信息，需要快速有效的方法。基于机器学习的文本分类技术可以在较大程度上解决文本库杂乱无章的现象，帮助人们将大量的文本自动分门别类。文本自动分类技术有广阔的应用前景，因此得到了广泛的关注，对其相关方面的研究也取得很大的进展。

特征选择是文本自动分类中最重要的环节之一，是本文研究的重点。特征选择是一个从原始特征集中抽取出它的一个由重要词汇组成的真子集的过程。通过一个评估函数给原始特征集里的每个特征打分，选取分值高于阈值的特征。

高效的文本分类分类器要求组成其向量空间的特征应该带有较强的分类信息，同时向量空间能很好的兼顾各个类别里的文本的信息。本文分析了典型的特征选择算法 DF 和 MI 的优点和不足，并以此为基础提出了基于二类信息差值的特征提取方法（IDTC），IDTC 强调特征对任意两个类别的分类作用，理论上能很好的满足文本分类中对特征集的要求。实验结果表明，使用 IDTC 选择方法的分类器，比使用 DF 和 MI 选择方法的分类器的效果要好得多，并有不逊于使用其他特征选择算法的分类器的表现。

关键词：　中文；文本分类；特征选择；二类信息差值

Abstract

Along with the swift and violent development of information technology and rapid popularization of the Internet technology, the amount of the information that is stored in computer systems increases explosively. Electronic text is one of the most important form of data in computer systems, and the growth of it is more astonishing, so does the growth of text that written in Chinese. In order to gain useful information from the large scale text set, fast and effective methods are needed. As one of important instrument of the data mining technology, automatic text classification technology, which is based on machine learning technology, can help people solve the problem of information disorder to a great extent. It can assign a text to one of the predefined categories automatically, so as to help people to index texts more conveniently, and find the useful information easier. Automatic text classification technology can be used in many areas because it is very useful, therefore, it becomes a hot point people focus on in research areas and big progress has been made in research of it.

In the progress of automatic text classification, there are two most important factors that can influence the classification effect mostly. One is the classification algorithm; the other is the feature selection method. After introducing the key technologies that will be used in automatic text classification, this paper lay great effort on research in the feature selecting methods. Feature selection is a progress that making up a subset of the original feature set, with picking up the most important words in the text set. There is a evaluation function that grade each feature of its effect, and then picking up features of which scores are higher than the giving threshold.

If a text classifier can be called a high effective one, it requires that the features making up the vector space should be with strong information of differences among categories, and the vector space can represent the texts well. DF algorithm and MI algorithm are both typical methods of feature selection, but not good ones. This paper analyzes that both the two methods above cannot meet the request in the progress of text classification in detail,

this paper also introduce their strengths.

Integrating the strengths and the weaknesses of DF algorithm and MI algorithm, this paper presents a new method that bases on the information difference between two categories (IDTC), IDTC algorithm emphasize the classification effect on every two categories of each feature, and can meet the request in the progress of text classification in theory. The experimental results show that, the classifier which used IDTC algorithm as its feature selection method achieved good classification results, much better than that used DF algorithm or MI algorithm as its feature selection method. Moreover, the classification results of the classifier which used IDTC algorithm are no less favorable than those of classifiers which used other algorithm especially IG or CHI, which is considered as one of best, as their feature selection method. In a word, the IDTC algorithm is one of the best algorithms.

Keywords: Chinese; text classification; feature selection; Information Difference between Two Categories

目录

第一章 前言

1.1 研究背景

信息技术在今天的世界环境里扮演着越来越重要的角色。随着信息技术不断前进和互联网技术的迅猛发展和普及，信息呈近乎爆炸的形式急速膨胀。无论网络上、企业中或是个人系统上，都有海量的信息需要处理。文本作为计算机系统中信息的最重要表现形式之一，其增长速度更为惊人。第 23 次中国互联网络发展状况统计报告显示，截至 2008 年 12 月，中国互联网上网页个数超过 160 亿个，这些网页的总数据量更是接近 400TB[1]，而这些仅仅是中国互联网络上的数据。伴随 Internet 的成长，越来越多的信息以电子文档的形式存在。

要想在极为庞大的信息空间里获取想要的信息，没有快速高效的方法几乎是不可能的事情。如何在海量文本库中搜寻、过滤和管理这些文本成为一个亟待解决的问题。数据挖掘技术正是为了解决海量数据的管理、组织并从中获取有效信息的数据处理新技术。作为数据挖掘技术的重要手段之一，基于机器学习的文本分类技术可以在较大程度上解决文本库杂乱无章的现象，帮助人们将大量的文本自动分门别类，更好地把握文本信息。因此，文本自动分类技术已成为一项具有较大实用价值的关键技术，得到了广泛的关注，对其相关方面的研究也取得很大的进展。

特征选择是文本自动分类中的重要一环，在中文文本分类中更显得重要。

中文文本数据属于非结构化数据，在没进行任何处理的情况下，表示文本的特征空间的维数高达几万甚至是几十万。即便是在经过了文本预处理（停用词过滤、低频词过滤等），特征空间依然有很高的维数。在一定的分类算法下，过高的特征维数不但不能够提高分类的精度，反而可能在降低分类精度的同时导致效率的地下。因此，在文本分类的过程中，对特征进行选择显得至关重要。

特征选择是排除特征空间里认为不重要或者对分类很少或几乎没有贡献的特征。一般的做法是构造一个评估函数，然后用其对原始特征空间里的每一个特征进行打分，然后按打分值对所有特征进行大小的排序，最后用给定的阈值选定所要个数的特征，用这些特征构造一个精简的特征空间。

特征选择的动机有两个[2]：1）提高分类的精度；2）提高分类的效率。一个高

效的特征选择算法都应该在高的效率上达到高的精度。一个特征算法是否高效可以用达成这两个目标的程度来评判。要达到这两个目标，就要求选取出来的特征带有较强的类别信息，同时由这些特征组成的向量空间模型能更好更准确的表示文本，而这两个要求一般情况下是相互矛盾的。

鉴于特征选择在文本分类过程中的重要作用，本文的研究重点就在于特征选择算法，本文首先在对常用的特征选择算法进行比较研究，然后再在上述研究的基础上提出一种新的特征选择算法 IDTC。

1.2 文本分类的应用领域

文本呢分类技术可以用在许多领域，它可以作为信息过滤、信息智能识别、邮件分类、数字化图书馆等领域的技术基础，拥有广泛的应用前景。

（1） 信息过滤

信息量大是信息爆炸时代最大的特点，这给人们对信息进行处理带来了极大的困难。为了快速的获取用需要的信息，对源信息进行过滤成为一种必然。信息过滤技术可以用来解决这个问题。信息过滤的本质是一个两类分类的问题，既可以用来过滤掉对用户没价值的信息，如大量的广告、反动的信息等，也可以用来将用户感兴趣的信息过滤出来，主动的推送给用户，方便用户以最快速的方式获取最精确的信息。

（2） 信息智能识别

这实际上还是一个文本分类问题，主要针对实时性较强的信息，即在信息输入后输出的反馈结果对用户而言有实时性价值。对财经类消息进行正负面判断并及时反馈就是典型应用。该技术还可以用在诸如天气领域等多方面。

（3） 邮件分类

对用户收到的电子邮件进行分类。麻省理工大学曾经为白宫开发过一个专用的邮件分类系统，该系统能自动地把每天发给总统的大量的电子邮件分门别类，比如将邮件归为外交、税收、环保、家庭等类别，然后针对不同类别的邮件安排专门的人员对其进行回复。更为典型的，文本自动分类技术可以用于垃圾邮件筛选和过滤，减少用户对邮件系统的意见。

（4） 数字图书馆

对图书进行归类，一直是图书馆馆管理工作的重点。然而，图书管理员不可能对

每个学科都有深入了解，通过计算机系统对图书进行数字化管理是当前流行的趋势，使用自动文本分类技术，可以帮助图书管理员正确的对图书资料进行归类。

除此之外，自动文本分类技术还在信息检索、文本数据库、文档组织、语言识别、流派识别等方面有广泛应用。可以预见，文本分类技术有着广阔的应用前景。

1.3 中文文本分类概述

1.3.1 文本分类问题描述

文本分类是一个有指导的学习过程，它对于带有类别标识的文档集合，根据每一个文档类别的文档子集中文档的共同特征，找出一个分类规则或分类模型，根据该模型可以把一个未知类别的新文档映射到已知的某个文档类别中。该过程可以更形式化的描述，假设有一组已定义的文档类别 C和一组训练文档 D。文档集 D中的每一个文档对应一个 C中的类别，即在客观上，存在着一个映射[3]：

$$T：D \rightarrow C \tag{1.1}$$

在这里，T表示在客观事实上 D中的一个文档实例属于且只属于 C中的某一个已知类别，即对 D中的每一个文档 d，$T(d)$是已经确定的。通过对训练文档有指导的学习，可以找出一个近似于 T的映射 H[3]：

$$H：D \rightarrow C \tag{1.2}$$

对于一个新文档d'，$H(d')$表示对d'的分类结果。事实上，这样的近似映射可以有很多个，文本分类学习的目的就在于寻找一个与映射 T最相近的映射 H。对给定的评估函数 f，满足上述条件的 H应使得公式 1.3 取得最小值[3]。

$$\sum_{i=1}^{|D|} f\big(T(d_i) - H(d_i)\big) \tag{1.3}$$

1.3.2 文本分类过程

文本分类使用大量的训练文档对分类规则或分类模型进行训练，产生一个符合设计要求的分类器，使用该分类器就可以对未知类别的新文本进行自动分类。整体过程如图 1-1 所示[4]。

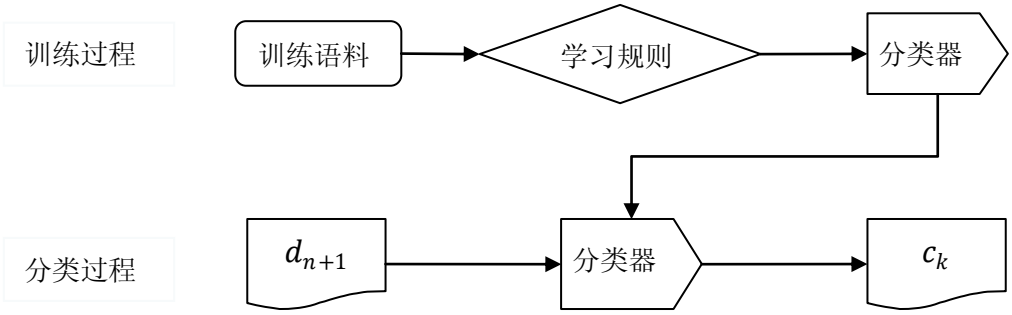

图 1-1 文本分类过程

对于基于向量空间模型的中文文本分类中的训练过程，主要由以下几个步骤完成[5]：

1) 分词

2) 停用词过滤

3) 特征选择

4) 特征权重赋值

5) 生成文档向量表示

6) 生成分类模型

上述步骤中用到的关键技术将在后续章节进行详细介绍。

1.4 国内外研究现状

1.4.1 国外研究现状

文本分类研究始于 50 年代末，H.P.Luhn 率先提出了词频统计的思想，可用于文本分类[6]，是这个领域的开创式研究。而 Maron、Kuhn 于 1960 年在 Journal of ASM 上表的论文 On relevance，probabilistic indexing and information retrieval[7]是第一篇对文本自动分类进行探讨的著作。到 1963 年，学者 H.Borko 和 M.Berniek 提出了采用“椅子分析法”对文献进行自动分类的思想[8]。其后，许多著名学者如 K Sparch、G. Salton 及 R. M. Needham 等在这个领域进行了卓有成效的研究。文本分类在国外大体经历了三个发展阶段[9]：

第一阶段是 20 世纪 80 年代前。在这一时期，模式识别和信息检索相继发展成一门学科。这一阶段主要集中在分类理论的研究，应用方面主要是用于信息检索。

第二阶段是 20 世纪 80 年代到 90 年代。这一阶段的技术特点是采用传统的知识工程技术，依赖专家提供的知识以形成规则，需要手工建立分类器。这一阶段分类器的缺点是很明显的：一是依赖专家，规则形成过程需要人为的大量介入；二是针对性过强，只适用于特定的领域，一旦要应用到其他领域，得重新人为介入生成规则；三是建立分类器过于耗费时间和精力，而且分类质量也难以保证。

第三阶段是 20 世纪 90 年代以后。伴随着 Internet 的兴起，呈爆炸般增长的文本数据量使得耗时、难以应用的知识工程方法已越来越不能满足迅速增长的实际应用的需求，基于机器学习的文本分类方法在这个时候兴起并逐渐取代了基于知识工程方法的文本分类方法。基于机器学习的文本分类方法比后者更具优势：一是不再依赖于专家，分类知识和规则都来源于机器对文本训练集的自主学习；二是实现了学习和分类过程的全自动化，不需要人为的介入，分类效率和分类结果的提高程度也不是后者所能比拟的。

1.4.2 国内研究现状

国内中文文本数据分类研究起步比国外晚很多，直到20世纪80年代初期才开始。1981 年，侯汉清先生对自动分类进行了探讨，向国内业界介绍了文本技术在国外的发展概况，第一次从国外引进了自动分类技术。我国文本分类的研究大体上经历了可行性探讨、辅助分类、自动分类系统三个发展阶段。早期研究主要将英文文本分类的研究成果套用到中文文本分类上。到 20 世纪 90 年代后期，随着中文文本数据处理需求的大量增加，国内才开始注重对中文文本分类的研究，结合中文自然语言的特点，形成了中文文本数据分类研究体系[9]。

1.4.3 现有分类系统的缺陷

在中文自然语言处理领域，各个分类系统的分类正确率都在 80%左右，离实用化、商品化尚有一定的距离，其主要原因有：

（1） 信息源不足

现有中文自动分类系统的信息源主要来自文献的提名或者文摘，其依据是：社会科学文献的题名与内容的平均符合率为 84%，自然科学的符合率为 89.3%。显然，必

然有百分之十几的文献有可能被错分。

（2） 分词算法的不足。

中文文本处理需要对句子进行分词，这点不同于英文各个单词由空格隔开。而到目前为止各种分词算法都不能很好的解决中文二义性切分问题，这直接影响了分类的效果。

（3） 分类算法的不足

现有的文本分类算法多是基于统计的。每一个分类算法都有它的局限性，也就是它都容许有一定的错分率，事实上，没有一个分类算法能够做到 100%准确分类。分类算法的不完善很大程度上影响了分类的效果。

（4） 特征选取算法的不完善

为了降低文本向量的维数，需要对有效的特征进行提取。目前的特征提取算法都存在不可克服的局限性，选取出来的特征集不能保证完整同时还存在噪声。正是由于特征选取算法存在着不足，又非常重要，成为了文本分类领域中热门的研究方向之一。

（5） 知识库规模小

人工智能技术尚未能从根本上解决知识学习的问题，这就直接导致了知识库的更新慢，跟不上知识的增长。这是目前文本自动分类系统不实用的原因之一。

1.5 本文组织

本文后续章节组织如下：

第二章，介绍中文文本分类中使用到的关键技术，首先介绍了自动分词技术，紧接着依次介绍了 VSM 向量空间模型、停用词过滤、单词权重的计算、常用文本分类算法，最后介绍了文本分类的性能评估方法。

第三章，重点介绍了文本分类过程中的特征选择的问题，包括介绍了各种常用特征选择算法，然后通过实验对比各个特征选择算法并对实验结果进行了分析。

第四章，通过对第三章的总结提出新的特征选择算法，介绍了算法的流程，并通过实验验证该算法的有效性。

第五章，总结了本文的工作，并对未来的研究做出了展望。

第二章 中文文本分类的关键技术

中文文本分类的过程要使用到很多技术，其中主要包括自动分词技术、空间向量模型、停用词过滤、特征项选择、单词权重赋值、自动分类器以及分类器评估方法等。本章节将介绍除特征选择外的关键技术，而作为本文的研究重点，特征想选择将放在第三章详细介绍。

2.1 自动分词技术

中文自然语言处理与英文自然语言处理存在一个较大的差异：英文文章中词是独立的，词与词之间用空格隔开，而中文文章中词与词之间是相互连续的无间隔的，同时中文词汇存在较大的二义性。为了提取文章中的关键词语，对中文文章进行分词是必需的处理，中文自动分词技术是中文自然语言处理领域最为重要的技术之一，国内国外的研究机构都投入了巨大的人力物力加以研究，并取得了重大进展。

2.1.1 中文分词方法

中文自动分词是把输入计算机中的汉语语句自动切分成词的序列的过程。建立一个自动分词系统，一般的做法都需要一个自动分词算法和一个自动分词词库或者汉语字典。自动分词算法是最影响分词效率和效果的因素，也是目前研究的重点，而自动分词词库是否完善也从很大程度上影响了分词的效果。

目前国内分词系统多采用的或者正在研究的方法基本上分为以下几类：词典匹配法、设立标志法、词频统计法、联想词群法、语义语用法、知识与规则法、人工智能法[4,11,12]。这些方法基本上都基于词典，进行字符串匹配，再结合词法、语法和语义规则进行分词。

由于中文自动分词不是本文研究的重点，本文也不对其中使用到的方法与技术进行详细研究。

2.1.2 中科院 ICTCLAS 分词组件简介

中文词法分析是中文信息处理的基础与关键。中国科学院计算技术研究所在多年研究工作积累的基础上，研制出了汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)，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。经过中科院多年的打造，至今内核升级 6 次，目前已经升级到了 ICTCLAS3.0。ICTCLAS3.0 分词速度单机 996KB/s，分词精度 98.45%，API 不超过 200KB，各种词典数据压缩后不到 3M，是当前世界上最好的汉语词法分析器[10]。

2.2 VSM 向量空间模型

文本的自动分类首先要涉及到的是文档的表示的问题。在自然语言处理领域，文本的表示主要采用向量空间模型（VSM)。VSM 的基本思想是：每个文本都包含一些用概念词表达的解释其内容的独立属性，而每个属性都可以看成是概念空间一个维度，这些独立属性成为文本特征项（如字、字串、词、短语等，现有研究表明以词为单位进行处理最为合理)，这样文本就可以表示成这些特征项的集合。形式化的，可以把文本表示成形如 $d=(t_1, w_1; t_i, w_2; \cdots\cdots t_n, w_n)$，其中$t_i$为特征项，$w_i$是其对应特征项的权值。

VSM 表示会带来一个问题，即使对于普通规模的文本集，通常也含有一万个以上的单词，空间的维数很高，而通常文本只包含相对很少量的单词，这将引发严重的高维稀疏问题。形象的，即对一个文本 T 在高维稀疏的情况下将表示成<0，0，0，……0，1，0，……0，1，0，>，文本向量上大部分的维度对应的值为 0。高维稀疏导致文本分类的性能急剧下降，使分类活动在耗费很长时间的同时达不到令人满意的结果。一般采用给特征空间降维的方式缓解高维稀疏带来的困难，而特征空间降维最直接的手段是特征选择。

2.3 停用词过滤

停用词，也称为禁用词，是指在文本中出现频率很高但是对文本分类几乎没用作

用的词条。停用词主要包括数字、虚词、量词、介词、连词等，它们几乎在任何类别文本中都会出现，在分类中基本不具价值。无论是英文、中文还是其它语言，都存在很多停用词。在普通的英文文本中，“the”、“of”、“and”、“to”、“for”等单词几乎都会出现，而且几乎会出现在一个完整的句子中。中文也不例外，像“的”、“了”、“把”、“被”、“啊”等词条在中文文本中非常常见。停用词的存在，增加了文本向量空间的维数，同时淹没了一些有用的词汇，整体上降低了计算效率的同时也降低了准确率。

停用词的过滤一般通过构造停用词表，把停用词表中的词条从特征集中剔除。停用词过滤是提高分类效果的有效手段，不过也该注意到，如果不小心过滤掉了有效的单词，反而削弱了分类效果，所以把那些词作为停用词过滤需要谨慎的考虑。

2.4 单词权重的计算

在文本由 VSM 表示并以单词作为文本的特征项之后，还要考虑特征项的权重评价问题，它将作为直接决定文档分类效果的词——文档矩阵的数据项。

一个单词在越多的其他文本中出现，它把这个文本区别于其他文本的能力越低，越不具有特殊性。一个单词，在同一个文本中出现次数越多，在其他越少的文本中出现，其区分度越高。

典型的权重计算方法有布尔权重、TF 权重、IDF 权重和 TF-IDF 权重等[2,3,8,9,13-19]。

2.4.1 布尔权重

布尔权重是最简单的权重赋值方法，根据特征项是否在文本中出现来计算特征项的权重。这种方法将所有的特征项都等同对待，不突出也不抑制任何特征的作用。布尔权重值为 1 或 0，对于一个文本d_j，其特征项t_i权重计算公式如下：

$$w_{ij} = \begin{cases} 1, & t_i \in d_j \\ 0, & t_i \notin d_j \end{cases} \qquad (2.1)$$

2.4.2 TF 权重

TF（Term Frequency）权重，也称为绝对词频，指的是特征项在文档中出现的频数。一般而言，一个单词在一个文本中出现越多次，则它与这个文本的主题联系越密切，其应具有大的权重值，代表其对这个文本很重要，如果单词在文本出现的频率很低，那么它可能跟文本的主题无关，其权重值应为很小，代表它对这个文本有作用但是很小。不同的文本，在特征项的出现频率上都会有不同程度的差异。这种方法认为特征项出现的频率跟它在文本中的作用成正比，并以该频率作为特征项的权重，其计算公式为：

$$w_{ij} = tf_{ij} \tag{2.2}$$

其中tf_{ij}表示单词t_i在文档d_j中出现的频率数。实际应用时考虑到各个文本的长度不一，一般还要进行归一化处理。

2.4.3 IDF 权重

IDF(Inverse Document Frequency)权重，即反比文档权重。IDF 的思想是一个出现在文本中的单词在越多的其他文本中出现，它把这个文本区别于其他文本的能力越低，越不具有特殊性，应赋予较小权值。一个出现在文本中的单词在其他越少的文本中出现，其区分度越高，应赋予较大权值。IDF 权重常用的计算公式：

$$w_{ij} = log\left(\frac{N}{n_i} + 0.01\right) \tag{2.3}$$

其中 N 表示全部训练集的文本数，n_i表示训练文本中出现t_i的文本频数。

2.4.4 TF-IDF 权重及其变体

TF-IDF 是 TF 方法和 IDF 方法的组合。TF 方法忽略了特征项在文档间的分布情况，而 IDF 方法则忽略特征项在文档中的分布情况，二者各有所失。TF-IDF 方法利用了词频和文档频率两种信息，是目前采用最多的权重赋值方法，在实践中也取得良好的应用效果。TF-IDF 典型的计算公式为

$$w_{ij} = tf_{ij} * log\left(\frac{N}{n_i} + 0.01\right) \tag{2.4}$$

通常采用归一化后的 *TF*IDF* 公式：

$$w_{ij} = \frac{tf_{ij} * log\left(\frac{N}{n_i}+0.01\right)}{\sqrt{\sum_{t_i \in d_j}[tf_{ij} * log\left(\frac{N}{n_i}+0.01\right)]^2}} \quad (2.5)$$

其中w_{ij}表示词t_i在文本 j中的权重，而tf_{ij}为词t_i在文本d_j中的词频，N为训练文本的总数，n_i为训练文本中出现词t_i的文本数，分母为归一化因子。

通过修改 TF-IDF 公式某个部分，可以得到其他不同的公式变体，这里不加以详述。

2.5 文本分类算法

现有文本分类模型和方法大多基于机器学习方法。常用的文本分类算法众多，主要有k近邻算法（KNN）、线性最小方差、朴素贝叶斯算法（NB）、决策树方法、支持向量机（SVM）、神经网络、Rocchio算法等[2,3,8,9,13-19]。

因为各个方法面向的对象不同，所以很难综合比较这些算法性能的差异。SVM可以达到很高的分类精度但对计算条件要求很高，实现起来也比较困难。NB方法在分类效果上的表现并不是最好的，但是它的模型简单、稳定性好、分类的效率比较高。KNN方法稳定性好，分类效果出色，也是最常用的文本分类算法之一。

2.5.1 朴素贝叶斯分类算法

贝叶斯分类器由于其简单性及计算的有效性，一直在文本分类领域中占有很重要的地位。朴素贝叶斯分类器假设特征对于给定类的影响独立于其他特征，即特征独立性假设。给定一个类c_j以及文档$D(a_1, a_2, a_3 \cdots\cdots a_k)$,其中$a_i$表示该文档出现的第$i$个特征项，$n$为训练集中包含特征项$i$的文档总数，朴素贝叶斯的定义为：

$$P(c_j|D) = \frac{P(c_j)P(D|c_j)}{P(D)} \quad (2.6)$$

$$P(D) = \sum_{j=1}^{N} P(c_j)P(D|c_j) \quad (2.7)$$

朴素贝叶斯分类器将未知样本归于类c_j的依据如下：

$$P(c_j|D) = arg\,max\{P(c_i|D)\},\ i = 1,2,\ldots,m. \quad (2.8)$$

即将文档D归于计算所得$P(c_i|D)$值最大的类别中。

在具体的实现时，一般又分成两种情况：

（1）文档D采用DF向量表示方式，文档向量V的分量为一个布尔值，0表示对应的特征没有在文档中出现，而1表示对应的特征在文档中出现。这时

$$P(D|c_j) = \prod_{i=1}^{k} P(a_i|c_j) \quad (2.9)$$

$$P(D) = \sum_{j=1}^{N} P(c_j) \prod_{i=1}^{k} P(a_i|c_j) \quad (2.10)$$

因此，

$$P(c_j|D) = \frac{P(c_j)\prod_{i=1}^{k} P(a_i|c_j)}{\sum_{j=1}^{N} P(c_j)\prod_{i=1}^{k} P(a_i|c_j)} \quad (2.11)$$

其中，$P(c_j)$为c_j类文档的概率，$P(a_i|c_j)$是对c_j类文档中特征a_i出现的条件概率的拉普拉斯估计：

$$P(a_i|c_j) = \frac{1+N(a_i|c_j)}{2+|D_{c_j}|} \quad (2.12)$$

$N(a_i|c_j)$表示出现特征a_i且属于类c_j的文档个数，$|D_{c_j}|$表示类c_j中文档的个数。

（2）文档采用TF向量表示方式，文档向量V的分量为对应特征在该文档中出现的频度，这时

$$P(c_j|D) = \frac{P(c_j)\prod_{i=1}^{k} P(a_i|c_j)^{TF(a_i,D)}}{\sum_{j=1}^{N} P(c_j)\prod_{i=1}^{k} P(a_i|c_j)^{TF(a_i,D)}} \quad (2.13)$$

其中，$TF(a_i,D)$是文档D中特征a_i出现的频度，$P(a_i|c_j)$是对c_j类文档中特征a_i出现的条件概率的拉普拉斯估计：

$$P(a_i|c_j) = \frac{1+TF(a_i,c_j)}{|V|+\sum_{i=1}^{k} TF(a_i,c_j)} \quad (2.14)$$

这里的$TF(a_i,c_j)$是文档D中特征a_i出现的频度，$|V|$为特征集的大小，即文档表示模式中特征项的总个数。

虽然贝叶斯算法没有考虑到文档内容中特征项之间存在着的语义联系，但其在多个实例中显示了良好的应用效果[12]。

2.5.2 KNN 分类算法

K-Nearest Neighbor(KNN)分类方法是一种稳定而有效的文本分类方法，因其在

准确率和召回率上表现出众，被广泛的应用于文本自动分类研究。

KNN的基本思想是：对于给定的测试文档d，通过相似度在训练文档中查找离它最近的k个邻近文档，并根据这些邻近文档的分类来给该文档的每一个候选分类打分。最简单的打分方式是在这k个文档中包含越多实例的类别分值越高。最常用的打分方式是为这k个训练文档中属于该类的文档与测试文档之间的相似度求和，并将和作为该类别和测试文档之间的相识度。然后通过对候选分类评分值的排序。还应当给出一个阈值，只有分值超过这个阈值的类别才予以考虑。测试文档属于超过阈值的所有分类。决策规则可以描述为：

$$score(d,c_i)=\sum_{d_j\in kNN} sim(d,d_j)y(d_j,c_i)-b_i \tag{2.15}$$

其中，

$$y(d_j,c_i)=\begin{cases}1 & d_j\in c_i\\ 0 & d_j\notin c_i\end{cases} \tag{2.16}$$

$sim(d,d_j)$表示文档 d和d_j的相似度，b_i为二元决策的阈值，$score(d,c_i)$为测试文档 d属于c_i的分值。

对于某一特定的类来说，b_i是一个待优化的值。一般的，b_i可以通过一个验证文档集来进行调整，可取训练文档集的一部分作为训练文档集。

2.6 文本分类器性能评价

针对不同的目的，人们提出了多种文本分类器性能评价方法，包括召回率、正确率、F-测度值、微平均和宏平均、平衡点、11 点平均正确率等[9,11]。以下对最常用的正确率、召回率、F-测度值进行描述。

表 2-1 文本分类器输出的结果

文本与类别的实际关系 / 分类器对二者关系的判断	实际属于	实际不属于
认为属于	a	b
认为不属于	c	d

假设一个文本分类器关于类别c_j的输出结果如表 2-1 所示：

在表 2-1 所示共有四个项，其中：

a 表示被分类器正确分到类别c_j的文本数；

b 表示实际不属于类别c_j但是却被分类器分到类别c_j的文本数；

c 表示实际属于类别c_j但是却没有被分类器分到类别c_j的文本数；

d 表示实际不属于类别c_j且没有被分类器分到类别c_j的文本数。

分类器的召回率、正确率和 F-测度值分别采用以下公式计算：

召回率

$$r = \frac{a}{a+c} * 100\% \tag{2.17}$$

正确率

$$p = \frac{a}{a+b} * 100\% \tag{2.18}$$

召回率和正确率是两个相互矛盾的标准，一般情况下，正确率会随着召回率的上升而下降，两者不可兼得。F-测度值是综合以上两种度量方法的最常用方式。

F-测度值

$$F_\beta = \frac{(\beta^2+1) * p * r}{\beta^2 p + r} \tag{2.19}$$

其中，β是调整正确率和召回率在评价函数中所占比重的参数，通长取$\beta = 1$，这时的评价指标变为

$$F_1 = \frac{2 * p * r}{p+r} \tag{2.20}$$

正确率、召回率和 F-测度值都是针对单个类别的分类情况而言的。如果要评价某个分类算法的分类效果，则需要综合所有类别的上的分类结果得到平均结果。综合的方法通常用微平均和宏平均。

宏平均是先计算每一个类的召回率 r 和正确率 p，然后针对所有类求 r，p 的平均值，其计算公式如下：

$$Macro - r = \frac{\sum_{j=1}^{|C|} r_j}{|C|} \tag{2.21}$$

$$Macro - p = \frac{\sum_{j=1}^{|C|} p_j}{|C|} \tag{2.22}$$

其中 C 为类别，|C|是类别总数，则$\mathrm{Macro} - \mathrm{F_1}$可以通过下面的公式计算：

$$Macro - F_1 = \frac{2 * Macro - r * Macro - p}{Macro - r + Macro - p} \quad (2.23)$$

微平均是直接根据各个类的分类情况直接计算 r 和 p 值，计算公式如下：

$$Micro - r = \frac{\sum_{j=1}^{|C|} a_j}{\sum_{j=1}^{|C|} a_j + \sum_{j=1}^{|C|} c_j} \quad (2.24)$$

$$Micro - p = \frac{\sum_{j=1}^{|C|} a_j}{\sum_{j=1}^{|C|} a_j + \sum_{j=1}^{|C|} b_j} \quad (2.25)$$

其中 C 为类别，$|C|$ 是类别总数，a、b 和 c 的定义如表 2.1，则 $Micro - F_1$ 可以通过下面的公式计算：

$$Micro - F_1 = \frac{2 * Micro - r * Micro - p}{Micro - r + Micro - p} \quad (2.26)$$

对比宏平均和微平均的计算公式可以看出，宏平均的关注点在于每个类别都一样重要，而微平均的关注点则在于每个文本都一样重要，这两个指标宜结合在一起使用[9]。

2.7 中文文本语料库

为了对各种文本分类算法进行实验性的评价、比较和分析，采用同一个文本集合是最基本的要求，这样的文本集合被称为基准测试数据集（Benchmark Test Collection）。采用基准测试数据集不仅可以减少建设数据集的费用，也使得分类结果具有可比性。

不幸的是，在中文文本分类领域，至今还没有形成一个开放的、相对标准的语料库可供使用，研究者们一般自己建立文本集，进行训练和测试，测试结果可比性较差。

复旦大学李荣陆收集整理的中文文本数据集是一个较多学者引用的数据集[9]。该数据集共有 19637 篇文本，分为 20 个类。其中训练文本 9804 篇，测试文本 9833 篇。

第三章 基于二类信息差值的特征提取方法

3.1 特征选择概要

在采用向量空间模型对文本进行表示的情况下，文本分类的最大特点和困难之一是特征空间的高维性和文档表示向量的稀疏性。中文的词条总数有二十多万条[13]，每天还有新词汇涌现。对于普通的文本集，即使是在去掉了停用词及低频词之后，特征空间的维数也在数万。特征空间维数过大，不仅在计算的时候要耗费更多的资源，许多噪声词汇的存在也降低了分类的效果。寻求一种有效的特征抽取算法,降低特征空间的维数,提高分类的效率和精度,成为文本自动分类中需要首先面对的重要问题。特征选择是解决这个问题的有效方法。

特征选择是一个从原特征集中抽取出它的一个由重要词汇组成的真子集的过程。一般通过一个特征选择算法的评估函数给原特征集里的每个特征打分，选取分值高于阈值的特征。实现过程可以表述如下：

1) 初始状态下，特征集包括所有特征；
2) 使用特征选择算法的评估函数特征集里的每一个特征打分；
3) 将特征集里的每个特征按分值由大到小排序；
4) 选取前n个特征组成新的特征集。n值不宜过大，过大将引入过多的噪声特征；n值也不宜过小，过小将使新的特征集失去很多有效信息。n值的取定一般通过实验测试选择分类效果最佳的大致值。

3.2 常用的特征选择算法

常用的文本特征选择方法有:文档频率(DF) 、信息增益(IG) 、互信息(MI) 、χ^2统计量(CHI) 、期望交叉熵(CE)等[3,4,9,13-21]。这些方法的基本思想都遵循小节3.1的特征提取流程，关键在于评估函数的不同。

3.2.1 文档频率

文档频率（DF）一般可以定义为：

$$DF(t) = \text{出现特征t的文档数} \quad (3.1)$$

文档频率是最简单的特征抽取技术。这种方法基于这样一个假设:在整个训练语料中DF值低于某个阈值的词条不含或只含有较少类别信息，将这样的词条从原始特征空间中移除，在降低特征空间的维数的同时还可以提高分类的精度和效率。但这一假设显然是不全面的。因此，在实际运用中一般并不直接使用DF，而是把它作为评判其他评估函数的一个标准。

3.2.2 信息增益(IG)

信息增益(IG)在机器学习领域被广泛使用。该方法依据特征t_i为整个分类所能提供的信息量的多少来衡量该特征项的重要程度，并以此来决定是否选择该特征项。对于词条t_i和类别C_j，IG考察C_j中出现和不出现t_i的文档频数来衡量t_i对于C_j的信息量的差别，其中信息量多少由熵来衡量。因此，信息增益可以定义为不考虑特征t_i和考虑特征t_i文档熵的差值：

$$IG(t_i) = -\sum_{j=1}^{|C|} P(C_j) log\, P(C_j) + P(t_i) \sum_{j=1}^{|C|} P(C_j|t_i) log\, P(C_j|t_i) + P(\sim t_i) \sum_{j=1}^{|C|} P(C_j|\sim t_i) log\, P(C_j|\sim t_i) \quad (3.2)$$

其中，$P(C_j)$表示C_j类文档在语料中出现的概率，$P(t_i)$表示训练语料文本集中包含t_i的文档的概率，$P(C_j|t_i)$表示文档包含t_i时属于类C_j的条件概率，$P(\sim t_i)$表示训练语料文本集中不包含t_i的文档的概率，$P(C_j|\sim t_i)$表示文档不包含t_i时属于类C_j的条件概率，$|C|$为预定义的类别总数。

在进行特征选取时，从原始特征空间中移除低于给定阈值的词条，保留高于阈值的词条作为表示文档的特征。

3.2.3 互信息(MI)

互信息(MI)法的基本思想是：互信息越大，特征t_i和类别C_j共现的程度越大。

如果令N表示训练语料中文档的总数，A表示属于C_j类且包含t_i的文档频数，B表示不属于C_j类但包含t_i的文档频数，C表示属于C_j类但不包含t_i的文档频数，D表示不属于C_j类且不包含t_i的文档频数，特征t_i和类别C_j关系如表3-1所示，那么t_i和C_j的互信息可以由下面的公式表示：

表 3-1 特征类别关系

特征项 \ 类别	C_j	$\sim C_j$
t_i	A	B
$\sim t_i$	C	D

$$MI(t_i, C_j) = log\frac{P(t_i,C_j)}{P(t_i)P(C_j)} = log\frac{P(t_i|C_j)}{P(t_i)} \approx log\frac{A*N}{(A+C)*(A+B)} \tag{3.3}$$

如果t_i和C_j无关，则$P(t_i, C_j) = P(t_i)P(C_j)$，那么，$MI(t_i, C_j) = 0$。对于多类问题，基于MI的特征提取方法可以采用两种实现方法：一种是计算每一个t_i对各个类别C_j的MI值，然后在整个训练语料中计算最大值：

$$MI_{max}(t_i) = max_{j=1}^{|C|} MI(t_i, C_j) \tag{3.4}$$

另一种方法是计算每一个t_i对各个类别C_j的MI值，然后在整个训练语料中计算平均值：

$$MI_{avg}(t_i) = \sum_{j=1}^{|C|} P(C_j) MI(t_i, C_j) \tag{3.5}$$

其中，$|C|$为类别总数。然后从原始特征空间中去除MI值低于给定阈值的特征，保留MI值高于给定阈值的特征作为文档特征。

3.2.4 χ^2 统计量(CHI)

χ^2统计量(CHI)衡量的是特征t_i和类别C_j之间的相关联程度，并假设t_i和C_j之间符合具有一阶自由度的χ^2分布。特征对于某类的χ^2统计值越大，它与该类之间的相关性越大，携带的信息也越多，反之则越少。

如果A、B、C、D、N的含义依然如表3-1所示，那么t_i和C_j的CHI值为：

$$\chi^2(t_i, C_j) = \frac{N*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \tag{3.6}$$

和互信息的处理方法类似，CHI也有最大值和平均值两种计算方法：

$$\chi^2{}_{max}(t_i) = max_{j=1}^{|C|} \chi^2(t_i, C_j) \tag{3.7}$$

$$\chi^2{}_{avg}(t_i) = \sum_{j=1}^{|C|} P(C_j)\chi^2(t_i, C_j) \tag{3.8}$$

3.2.5 期望交叉熵

期望交叉熵(CE)和信息增益(IG)相似，不同之处在于IG同时考虑了特征在文本中出现和不出现时的两种情况，而期望交叉熵(CE)则只考虑特征t_i在文本中出现一种情况。对于特征t_i，其相对于类C_j的期望交叉熵为：

$$CE(t_i, C_j) = P(C_j, t_i) log \frac{P(C_j, t_i)}{P(C_j)P(t_i)} \tag{3.9}$$

特征t_i对于整个预料的期望交叉熵为：

$$\begin{aligned} CE(t_i) &= \sum_{j=1}^{|C|} P(C_j, t_i) log \frac{P(C_j, t_i)}{P(C_j)P(t_i)} \\ &= P(t_i) \sum_{j=1}^{|C|} P(C_j|t_i) log \frac{P(C_j|t_i)}{P(C_j)} \end{aligned} \tag{3.10}$$

如果特征词和类别强相关，也就是$P(C_j, t_i)$大，且相应的类别出现概率小，则说明该词对分类的影响大相应的*CE*值就大。交叉熵大小反映了文本归属类别的概率分布和在出现了某个特定词的条件下文本类别的概率分布之间的距离，词t_i的交叉熵越大，对文本归属的类别分布的影响也越大，其重要程度越高。

以上是文本分类中比较经典的一些特征提取方法，实际上还有很多其他的文本特征选取方法，例如文本证据权法、优势率方法等，这里就不加以详述。

3.3 改进的基于二类信息差值特征提取方法

本节将吸纳特征提取方法DF和MI的思想，在将多类别问题细化成二类问题的基础上提出基于二类信息差值的特征提取方法IDTC(Information Difference between Two Categories)。

3.3.1 DF 与 MI 的思想与不足

文本分类过程中对特征提取的要求有两个，一个是特征要带有较强的类别信息，一个是特征集组成的向量空间模型能更好的表示文本，即特征更具代表性。

文档频率（DF）提取方法认为一个特征在训练文档集中出现在越多的文档之中，该特征具有更多的类别信息因而剔除在训练文档集中出现次数少的特征。一方面，这种思想能提取在训练文档集中高频率出现的词汇，这种词汇在文档中更具普遍性，能准确的表示文本的内涵，然而如果这种词汇在各类文档中出现的几率几乎一致的时候，这种词汇是基本不具有分类信息的。虽然是这样，大部分高频出现的词汇在各类的分布一般都有所差别，这是DF特征提取算法能运作的原因，问题在于用DF特征提取算法提取出来的占比不小的一部分特征只具有很少的分类信息。

MI特征提取方法通过计算特征与类别的相关度实现。MI方法能提取出带有强类别信息的词汇作为特征。然而，如果一个特征只在某个类别中的几个文本中出现，根据MI的计算方法，这个特征将很有可能被提取出来。而这个特征在该类别中并不具有普遍性，不能代表该特征，将该特征用于表示该类别的其它文本的时候，该项的词频权值将为0。如果这类不具代表性的强类别信息词汇过多，将导致无法正确的表示大部分文档的文本信息，从而导致在分类过程中文档被错分可能性过大。事实上，现有的研究大多支持MI特征提取方法过分依赖低频词汇，而高频词汇一般在分类中的作用要高于低频词汇的观点。同时，以特征与所有类别的互信息值求和也不能突出特征对类之间的区分能力，这也从一定程度上削弱了分类信息。

DF和MI特征提取方法都不能很好的满足使提取出来的特征具有强类别信息同时更具代表性更能正确表示文本信息的要求，使得使用这两种特征提取方法的分类器的分类效果不尽如人意。

3.3.2 基于二类信息差值的特征提取方法 IDTC

一个类别与其他多个类别的区别都要考虑这个类别与其他的每个类别之间的区别。因此，多类别关系问题可以转化成多个二类别关系问题。IDTC的核心思想就是强调特征对于一个类别与其他的每个类别之间的区分度（能力），并将所有二类关系的区分度的累积作为特征的分类左右的评估值。

IDTC改进自DF和MI，兼顾有两者的特点。IDTC同时使用特征出现概率和特征与类别的相关度两个指标来表示特征对于类别的信息值。与DF直接使用文档的频数不同的是，IDTC使用特征出现的概率来强调特征出现频率的作用。IDTC将依然使用互信息（MI）来表示特征与类别的相关度。IDTC将多类别关系细化为多个二类关系，对于每

一个二类关系，将其当作一个新的整体计算两个类别间的信息差值作为特征对这两个类别的区分度贡献，然后特征对所有二类关系的区分度贡献的和作为特征的所有类别的分类作用大小。

IDTC特征选择方法的流程依然遵循章节3.1中提到的特征选择流程。对于每一个二类关系，其区分度可以公式：

$$IDTC(t_i, C_j, C_k) = P(t_i) * |MI(t_i, C_j) - MI(t_i, C_k)| \quad (3.10)$$

计算。$P(t_i)$用以强调特征t_i的词频信息，$|MI(t_i, C_j) - MI(t_i, C_k)|$用以强调特征$t_i$的二类区分能力。特征$t_i$对于整个训练预料的IDTC值为：

$$\begin{aligned} IDTC(t_i) &= \sum_{j=1}^{|C|-1} \sum_{k=j}^{|C|} IDTC(t_i, C_j, C_k) \\ &= \sum_{j=1}^{|C|-1} \sum_{k=j}^{|C|} P(t_i) * |MI(t_i, C_j) - MI(t_i, C_k)| \\ &= \sum_{j=1}^{|C|-1} \sum_{k=j}^{|C|} P(t_i) * |log \frac{P(t_i|C_j)}{P(t_i)} - log \frac{P(t_i|C_k)}{P(t_i)}| \end{aligned} \quad (3.11)$$

公式中涉及到的所有概率运算都是相对于以二类C_j、C_k的“整体”而言，$P(t_i)$表示特征t_i在二类C_j、C_k的整体中出现的概率，$P(t_i|C_j)$表示类别C_j中含有特征t_i的文档出现的概率。

在进行特征选取时，从原始特征空间中移除低于给定阈值的词条，保留高于阈值的词条作为表示文档的特征。

IDTC算法可以用表3-2表示。

表 3-2　IDTC 算法流程

步骤1　初始状态下，特征集包括所有特征；
步骤2　使用公式3.11为每个候选特征打分；
步骤3　将特征集里的每个特征按分值由大到小排序；
步骤4　使用给定的特征数阈值n，选取按分值有序特征序列的前n个特征组成新的特征空间

IDTC算法强调了特征的强类别信息，同时也强调了特征满足表示文本信息的特性，理论上很好的满足了文本分类过程中对特征的要求。

第四章 实验结果与分析

4.1 实验目的

本文在第三章介绍了文本分类过程中常见的几种分类方法，包括文档频率(DF)、信息增益(IG)、互信息(MI)、χ^2 统计量(CHI)、期望交叉熵，并提出了改进与DF与MI的基于二类信息差的特征提取方法IDTC。本实验的目的在于通过对采用IDTC的分类器与才有其他方法的分类器的分类效果的比较，同时以对采用各种算法得出的实验结果的分析对本文改进算法的论点提供支持，验证IDTC算法的有效与可行性。

4.2 实验数据

实验用数据采用复旦大学李荣陆收集整理的中文文本数据集，该数据集共有19637篇文本，其中训练文本9804篇，测试文本9833篇，分为20个类，包括计算机、体育、政治、艺术、农业、能源等。本文测试用选择了太空、环境、经济、艺术、历史、计算机、农业和运动8个类别。实验数据构成如表4-1所示：

表 4-1 实验数据构成

类别 / 类型	太空	环境	经济	艺术	历史	计算机	农业	运动	全部
训练集	500	520	510	510	466	510	520	500	4036
测试集	402	418	401	392	368	418	342	404	3145

4.3 实验方案

实验文本分类器采用KNN分类器，特征项权重采用TF权重计算方法，最终的实验结果采用整体的正确率、召回率和F-测度值三个指标加以衡量。对于每一个特征选择方法，测试出其在特征数为500、1000、2000、3000、4000、5000、10000以及15000下的三项指标值。然后根据实验测试出来的结果分析各个特征选择方法在不同特征维数下的效果表现。

4.4 实验结果及分析

实验将采用整体的正确率、召回率和F-测度值对实验结果进行分析，通过测试各个特征选择函数在不同特征数下的结果：

表 4-2　使用 DF 方法的实验结果

指标 / 特征数	正确率		召回率		F1 值	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
500	0.67576	0.65246	0.65327	0.65246	0.66432	0.65246
1000	0.74515	0.72528	0.72585	0.72528	0.73537	0.72528
2000	0.79111	0.77266	0.77239	0.77266	0.78164	0.77266
3000	0.81542	0.80922	0.80784	0.80922	0.81162	0.80922
4000	0.81988	0.81431	0.81316	0.81431	0.81651	0.81431
5000	0.82249	0.81876	0.81775	0.81876	0.82011	0.81876
10000	0.82890	0.82671	0.82569	0.82671	0.82729	0.82671
15000	0.83329	0.83116	0.83020	0.83116	0.83174	0.83116

表 4-3　使用 IG 方法的实验结果

指标 / 特征数	正确率		召回率		F1 值	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
500	0.79153	0.78887	0.78801	0.78887	0.78977	0.78887
1000	0.80129	0.79618	0.79530	0.79618	0.79828	0.79618
2000	0.81182	0.80477	0.80374	0.80477	0.80776	0.80477
3000	0.82558	0.82353	0.82227	0.82353	0.82392	0.82353
4000	0.83279	0.83148	0.83017	0.83148	0.83145	0.83148
5000	0.82744	0.82512	0.82404	0.82512	0.82573	0.82512
10000	0.83533	0.83307	0.83258	0.83307	0.83395	0.83307
15000	0.83700	0.83498	0.83398	0.83498	0.83549	0.83498

表 4-4　使用 MI 方法的实验结果

指标 特征数	正确率		召回率		F1 值	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
500	0.55802	0.53386	0.53257	0.53386	0.54499	0.53386
1000	0.59869	0.56630	0.56633	0.56630	0.58206	0.56630
2000	0.65842	0.61653	0.61763	0.61653	0.63737	0.61653
3000	0.68450	0.65024	0.65049	0.65024	0.66706	0.65024
4000	0.72952	0.71065	0.71083	0.71065	0.72005	0.71065
5000	0.74256	0.72591	0.72584	0.72591	0.73410	0.72591
10000	0.78777	0.77870	0.77782	0.77870	0.78276	0.77870
15000	0.80521	0.79936	0.79860	0.79936	0.80189	0.79936

表 4-5　使用 CHI 方法的实验结果

指标 特征数	正确率		召回率		F1 值	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
500	0.80049	0.79746	0.79670	0.79746	0.79859	0.79746
1000	0.8115349	0.80890	0.80793	0.80890	0.80973	0.80890
2000	0.8069039	0.79936	0.79859	0.79936	0.80272	0.79936
3000	0.82461	0.82289	0.82155	0.82289	0.82308	0.82289
4000	0.82938	0.82766	0.82662	0.82766	0.82800	0.82766
5000	0.83173	0.83052	0.82912	0.83052	0.83043	0.83052
10000	0.83594	0.83339	0.83278	0.83339	0.83436	0.83339
15000	0.83991	0.83816	0.83779	0.83816	0.83885	0.83816

表 4-6　使用 CE 方法的实验结果

指标 特征数	正确率		召回率		F1 值	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
500	0.78041	0.77870	0.77857	0.77870	0.77949	0.77870
1000	0.81409	0.81145	0.81050	0.81145	0.81229	0.81145
2000	0.81218	0.80890	0.80818	0.80890	0.81017	0.80890
3000	0.81940	0.81749	0.81638	0.81749	0.81789	0.81749
4000	0.82719	0.82671	0.82548	0.82671	0.82633	0.82671
5000	0.83096	0.83021	0.82870	0.83021	0.82983	0.83021
10000	0.83587	0.83561	0.83451	0.83561	0.83519	0.83561
15000	0.83720	0.83688	0.83555	0.83688	0.83637	0.83688

表 4-7　使用 IDTC 方法的实验结果

指标 特征数	正确率		召回率		F1 值	
	宏平均	微平均	宏平均	微平均	宏平均	微平均
500	0.77512	0.76979	0.76979	0.76979	0.77244	0.76979
1000	0.79101	0.78092	0.78023	0.78092	0.78558	0.78092
2000	0.81061	0.80509	0.80385	0.80509	0.80722	0.80509
3000	0.82689	0.82448	0.82329	0.82448	0.82509	0.82448
4000	0.82874	0.82734	0.82613	0.82734	0.82743	0.82734
5000	0.83449	0.83307	0.83170	0.83307	0.83309	0.83307
10000	0.83684	0.83434	0.83369	0.83434	0.83526	0.83434
15000	0.83636	0.83402	0.83313	0.83402	0.83474	0.83402

表4-2、表4-3、表4-4、表4-5、表4-6、表4-7展示了使用各个特征提取方法的实验效果。从这些表中可以看出，当特征空间的维数很高（10000以上），各个使用各个特征选择方法都能达到比较高的分类效果。从分类的效果上看，IG、CHI、CE、IDTC都能达到比较高的分类效果，DF的效果略差，MI的效果则是最差的。图4-1和图4-2更为直观地展示了各个使用各个特征提取方法的分类器的分类效果中的微平均F1值

和宏平均F1值的对比 ，这两个指标都能很好的衡量分类器的整体表现。

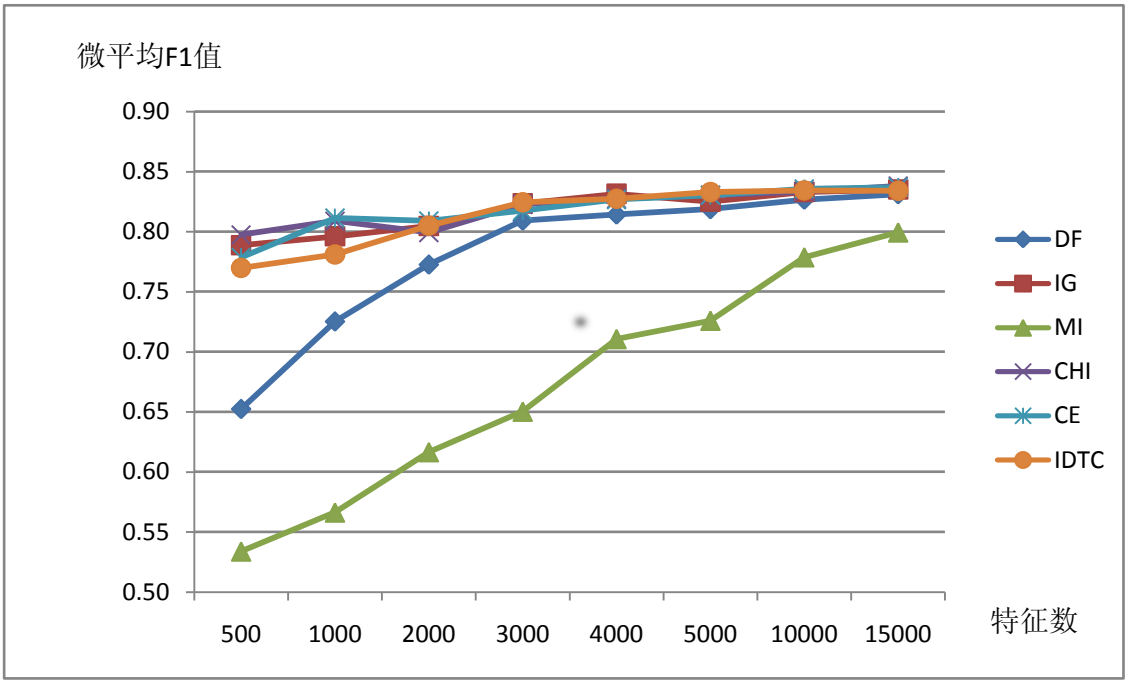

图 4-1　　各种方法的微平均 F1 值对比

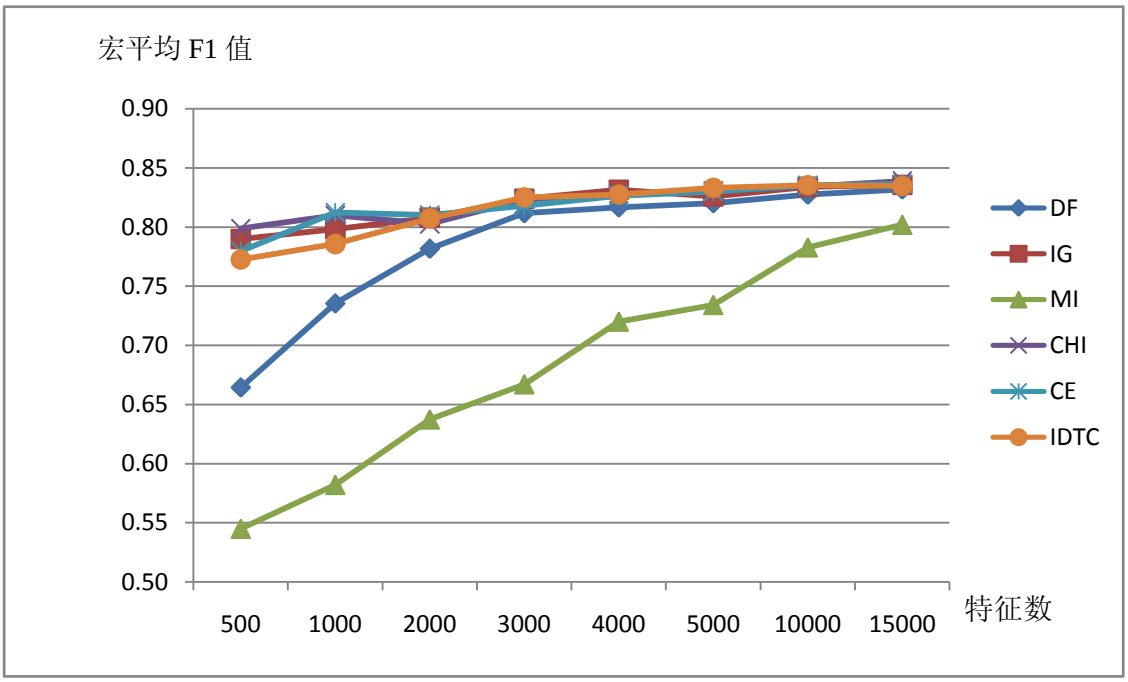

图 4-2　　各种方法的宏平均 F1 值对比

从图4-1与图4-2可以很直观的看到IDTC有不逊色于被普遍认为最好的特征选择算法的IG方法和CHI方法。IDTC方法虽然在特征数小于2000是比后两者的效果略差，但是在特征数高于2000时的效果确实要略强于后两者，从整体上来看，IDTC方法的效

果并不逊色于后两者，也是效果最好的特征选择算法之一。

另外，从这两个表都可以直观的看随着特征数的增加，使用DF方法的分类器的效果也呈增长趋势，与表现最好的其他几个算法的效果差距也不断的缩小，DF特征选择选法优先挑选出现频率高的特征，这也证明了出现频率高的特征比出现频率低的特征具有更高的分类作用。

从图4.1与图4.2还可以看出使用MI方法的分类器表现与使用其他方法的分类短的效果差距比较大，这反映了MI方法选择出来的特征集里存在较多的噪声词汇。既然出现频率高的特征比出现频率低的特征具有更高的分类作用，这也从一定程度上反映出了MI特征选择方法更为依赖于低频词汇。

IDTC方法结合了DF和MI的思想，属于一种中庸的算法，但是分类效果确实要比后两者都要好，性能更为稳定。

以上实验结果都表明，IDTC方法是行之有效的，IDTC的改进是成功的。

4.5 实验总结

除了改进的特征选择方法，实验得出的结论与现有的研究并无二致，这也从侧面上证明IDTC的有效性。从整体上看，实验完成了目的。

然而本次实验也是有不足之处的。第一，实验数据的不完善，选择不同的数量的文本、不同的类别数得出来的实验数据有一定的差距。第二，基于计算条件有限，实验并不是采用分类效果更佳的实验手段（分类效果更好的分类器及特征权重赋值，需要更高要求的计算条件）进行的，这也从一定程度影响了实验的效果。

虽然实验有所不足，但这些不足并不影响各个分类特征选择算法的效果的对比。本次实验是有效的。

第五章　总结与展望

5.1 总结

随着信息技术的发展和Internet的快速普及，文本信息急剧膨胀，人类面临着日益严重的信息挑战。作为数据挖掘技术之一，文本自动分类技术能帮助人们自动的将各种文档分门别类，更好的组织和处理文档，给人们在信息处理和利用上提供了许多便利。因此，文本自动分类技术已成为一项具有广阔应用前景的关键技术，也成为了目前中文自然语言处理领域中应用最广的中文信息处理工具，对其在中文领域的应用的研究得到了广泛关注。

本文的主要工作概括如下：

1) 本文从文本分类技术的广阔应用前景出发，基于文本特征选择方面探讨了国内外文本分类研究的现状及其研究意义。
2) 总述了文本分类技术的框架体系和分类流程，同时阐述了文本分类技术应用在中文处理和英文处理方面的区别。
3) 详细的介绍了中文文本分类过程中使用到的关键技术，包括中文分词技术、VSM向量空间模型、停用词过滤、单词权重的计算、文本分类算法、文本分类效果的评价方法及中文预料库技术。
4) 重点研究了特征选择算法，详细介绍了特征选择的基本思想和基本流程，并介绍了常见的特征选择算法包括文档频率(DF) 、信息增益(IG) 、互信息(MI) 、χ^2 统计量(CHI) 、期望交叉熵(CE)。
5) 详细分析了DF和MI特征选择方法的优点和不足，在此基础上提出了改进的基于二类信息差值的特征提取方法IDTC。使用IDTC方法提取出来的特征理论上满足了分类过程对特征的要求。
6) 通过实验对比IDTC与其他算法，验证了IDTC的有效性和可行性。

5.2 进一步工作

本文结合原有特征选择算法，提出了改进的基于二类信息差值的特征选择方法，

并通过实验验证了得出其有不逊色于其他特征选择方法的效果。虽然研究基本上达到了预设的目标，但是在许多方面仍存在问题需要进一步的研究。

1) 虽然提出的IDTC有不逊色于其他包括最好的特征选择方法的表现，但是也看不出与最好的特征选择方法之间的差距，该算法还有继续改进的空间，如何通过改进二类间区分度的计算方法是值得挖掘的问题。
2) 目前国内没有统一标准的语料库，而现有的语料库又大多存在问题。因此，寻找更高质量、更加标准的预料库，以便与其他研究者的实验结果相比较也是个值得继续深入的工作、
3) 采用更好的实验方案进行实验。虽然现有的实验方案能得出较好的实验结果，使用更好的实验方案能更加精确，更能比较各种算法的优劣。本文进一步的工作将采取更高分类效果的支持向量机分类器，更好的权重赋值方案进行实验，以期更加准确的实验结果。

致谢

值此论文完成之际，首先谨向我的导师衣杨副教授表示衷心的感谢。在我做毕业论文期间，衣老师自始至终悉心指导。老师严谨的治学态度、实事求是的作风以及急学生之所急的精神给我留下了深刻的印象，激励我积极向上。毕业论文能够顺利的完成，跟衣老师负责任的监督与耐心的指导密不可分。

衷心感谢曾经实习时候的公司同事，是他们引发我研究该课题的兴趣。

感谢每一个在我论文期间给过我帮助的同学。

最后，向参考文献的作者表示衷心的感谢。

参考文献

[1] 中国互联网络信息中心,中国互联网络发展统计报告,统计报告,2009

[2] Monica Rogati、Yiming Yang，High-performing feature selection for text classification， In：Proceedings of the eleventh international conference on Information and knowledge management, McLean, Virginia, USA，2002

[3] 李荣陆，文本分类及相关技术研究，博士论文，复旦大学，2005

[4] 雷琼，中文文本分类和聚类中的特征值选择研究，硕士论文，中山大学，2006

[5] Ikonomakis M.、Kotsiantis S.、Tampakas V.，Text classification: a recent overview，In：Proceedings of the 9th WSEAS International Conference on Computers，2005

[6] Luhn H.P.，An Experiment in Auto abstracting，In：International Conference on Scientific Information，Washington D.C.，1958

[7] Maron M.E.，Kuhn J.L.，On Relevance Probabilistic Indexing and Information Retrieval. J.ACM，7(3):216-244，1960

[8] Borko H.、Berniek M.，Automatic Document Classification. Journal of the ACM，11(2):138-151，1964，

[9] 尚文倩，文本分类及其相关技术研究，博士论文，北京交通大学，2007

[10]百度百科， http://baike.baidu.com/view/1215398.htm

[11]宗成庆， 统计自然语言处理，北京：清华大学出版社，2008

[12]白若鹞、董渊、张素琴、徐大伟，研究中文文本分类技术的辅助平台，清华大学学报(自然科学版)，48(7):1149-1153，2008

[13]朱颖东、钟勇，一种新的基于多启发的特征选择方法，计算机应用,29(3): 849-851，2009

[14]彭时名，中文文本分类中特征提取算法的研究，硕士论文，重庆大学，2006

[15]FABRIZIO SEBASTIANI，Machine Learning in Automated Text Categorization， ACM Computing Surveys, 34(1):1-47，2002

[16]陈治纲、何丕廉、孙越恒、郑小慎，基于向量空间模型的文本分类系统的研究与实现，中文信息学报，19(1):36-41，2005

[17]陈集、樊兴华、王鹏，中文文本分类中的两步特征选择法，计算机辅助工程，17(3):76-80,2008

[18]Songbo Tan、Yuefen Wang、Xueqi Cheng， An Efficient Feature Ranking Measure for

TextCategorization，In：Proceedings of the 2008 ACM symposium on Applied computingFortaleza, Ceara, Brazil，2008

[19]Zhaohui Zheng、Xiaoyun Wu、Rohini Srihari，Feature selection for text categorization on imbalanced data ACM SIGKDD Explorations Newsletter,6(1):80-89 ,2004

[20]王维娜、康耀红、伍小芹，文本分类中特征选择方法研究，信息技术,(12):29-31,2008

[21]胡燕、吴虎子、钟珞，中文文本分类中基于词性的特征提取方法研究，武汉理工大学学报，29(4):132-135，2007

毕业论文成绩评定记录

指导教师评语：   成绩评定： 指导教师签名：　　年　　月　　日
答辩小组或专业负责人意见：   成绩评定： 签名（章）：　　年　　月　　日
院系负责人意见：   成绩评定： 签名（章）：　　年　　月　　日