密　级：　　学校代码：10075
分类号：　　学号：106081203001

工学硕士学位论文

基于朴素贝叶斯方法的中文文本分类研究

学位申请人：李　丹
指 导 教 师 ：袁　方　教授
学 位 类 别 ：工学硕士
学 科 专 业 ：计算机应用技术
授 予 单 位 ：河北大学
答 辩 日 期 ：二〇一一年六月

Classified Index： **CODE：10075**

U.D.C： **NO：106081203001**

A Dissertation for the Degree of M.Engineering

The Study of Chinese Text Categorization Based on Naïve Bayes

Candidate: **Li Dan**

Supervisor: **Prof. Yuan Fang**

Academic Degree Applied: **Master of Engineering**

Specialty: **Computer Applied Technology**

University: **Hebei University**

Date of Oral examination: **June, 2011**

河北大学

学位论文独创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知， 除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得河北大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了致谢。

作者签名：李丹 日期：2011年6月8日

学位论文使用授权声明

本人完全了解河北大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

本学位论文属于

1、保密 □ ，在_____年_____月_____日解密后适用本授权声明。

2、不保密 ☑。

（ 请在以上相应方格内打“ √ ” ）

保护知识产权声明

本人为申请河北大学学位所提交的题目为（基于朴素贝叶斯方法的中文文本分类研究）的学位论文，是我个人在导师（袁方）指导并与导师合作下取得的研究成果，研究工作及取得的研究成果是在河北大学所提供的研究经费及导师的研究经费资助下完成的。本人完全了解并严格遵守中华人民共和国为保护知识产权所制定的各项法律、行政法规以及河北大学的相关规定。

本人声明如下：本论文的成果归河北大学所有，未经征得指导教师和河北大学的书面同意和授权，本人保证不以任何形式公开和传播科研成果和科研工作内容。如果违反本声明，本人愿意承担相应法律责任。

声明人：李丹　　日期：2011年6月8日

作者签名：李丹　　日期：2011年6月8日

导师签名：袁方　　日期：2011年6月8日

摘 要

计算机与网络技术自出现以来，发展迅速，并日趋完善，互联网已成为获取信息的主要来源。由于网络中大部分信息是文本数据，作为有效组织与管理文本数据重要基础的文本自动分类已成为具有重要应用价值的研究领域。基于贝叶斯理论的朴素贝叶斯分类方法具有简单、有效、速度快的优点，成为文本分类算法的重点研究内容之一。

本文首先对文本分类涉及到的中文分词、文本向量表示及特征权重计算等关键技术做了比较详细的分析研究；然后针对朴素贝叶斯文本分类的模型以及常用特征选择方法对朴素贝叶斯文本分类的性能影响进行了详细的研究与分析；最后，设计并使用 Java 在 MyEclipse 平台上实现了基于朴素贝叶斯方法的中文文本分类系统。

本文重点分析了多变量伯努利模型与多项式模型，通过实验对比得出在中文文本分类中多项式模型优于多变量伯努利模型。为了进一步提高分类精度，本文对多项式模型的平滑因子进行了改进，实验表明具有良好的分类效果。由于朴素贝叶斯分类模型是建立在属性之间条件独立性假设之上，因此特征选择的好坏与否对分类精度有较大影响。本文通过实验表明信息增益和 χ^2 统计量是朴素贝叶斯文本分类较好的特征选择方法。

关键词 文本分类 朴素贝叶斯分类 多变量伯努利模型 多项式模型 特征选择

Abstract

Since the technology of computer and network appeared, it had been developed very rapidly. Network has becoming one of the most mainly-used information source. Because most of the information in the network is text data type, automatic text categorization which is the important basic of effective organization and management text data has become an important study field. Naive Bayes classification method is based on the Bayesian theory，which is accepted as simple and effective probability classification method and has become one of the important contents in the text categorization.

Firstly, the paper studies key technologies of the text categorization that includes Chinese text segmentation, representation of text vector and feature weighting. After that, Naive Bayes text classification model and the affect of feature selection method on performance of Naive Bayes text classification is studied. At last, java on MyEclipse to realize Chinese text categorization system based on Naive Bayes method is accomplished.

This paper mainly analyzes Multi-variate Bernoulli Model and Multinomial Model. By experiment, the effect of Multinomial Model is better than Multi-variate Bernoulli Model in the Chinese text categorization. In order to increase classification accuracy，smoothing factor of Multinomial Model is improved. The experiment shows excellent classification performance. Due to Naive Bayes text classification model based on conditional independence assumptions of attributes, feature selection is important to classification accuracy. By means of the experiments, the paper shows information gain and χ^2 statistic value are the preferably feature selection methods of Naive Bayes text classification.

Keywords Text Categorization Naive bayes classification Multi-variate Bernoulli Model Multinomial Model Feature Selection

目 录

第 1 章　绪　论

1.1 研究背景及意义

随着计算机技术与网络技术的快速发展，互联网得到了广泛应用。中国互联网络信息中心（CNNIC）在 2011 年 1 月 19 日发布的《第 27 次中国互联网络发展状况统计报告》表明：截至到 2010 年 12 月底，我国网民规模达 4.57 亿人，较 2009 年底增加 7330 万人。互联网普及率持续上升增至 34.3%，与 2009 年底相比提高了 5.4 个百分点。全国域名数 866 万个，全国网站数 191 万个。互联网成为人们信息获取的重要来源。网络的大部分信息是文本数据，面对如此巨大的信息海洋，如何有效地组织和管理，进行自动分类，并快速、准确、全面地从中找到用户所需的信息已成为一个重要用途的研究课题。

文本自动分类简称文本分类（Text Categorization，TC）是信息检索和文本挖掘的重要基础。分类任务就是通过学习得到一个目标函数，即分类模型 f，通过此分类模型把每个属性集 x 映射到一个预先定义的类标号 y[1]。文本分类问题与其他分类问题相似，文本分类是在预定义的分类体系下，根据文本的特征，即文本的内容，将给定文本与一个或多个类别相关联的过程[2]。文本自动分类能较好地解决大量文档信息归类的问题并可以应用到很多方面，如文献组织、文本识别、智能搜索、邮件过滤等。因此，对文本分类的研究具有重要的理论意义和实用价值。

朴素贝叶斯分类器是贝叶斯分类器中最常用的方法，是一种基于概率统计的方法。朴素贝叶斯分类方法是基于条件“独立性假设”，因此它适合于处理属性个数较多的分类任务，而文本分类正是这种多属性的分类任务，因此朴素贝叶斯成为文本分类的一种常用分类方法。它是目前公认的一种简单有效的概率分类方法，其性能可以与决策树、神经网络等算法媲美，在某些领域中表现出很好的性能[3][4]，成为文本分类算法的重点研究对象之一。

1.2 文本分类研究现状

文本分类任务从数学的角度来看就是一个映射过程。可以使用如下的数学模型来描述文本分类任务[5][6]：给定文档集合 $D=\{D_1,D_2,...,D_n\}$，D_i 表示第 i 篇文档。D 由 n 篇

文档组成；预先定义的文档类别集合$C=\{C_1,C_2,...,C_{|c|}\}$。假设文档集合与类别存在一个未知的目标函数：

$$\Phi：\mathrm{D}\times\mathrm{C}\rightarrow\{\mathrm{True}，\ \mathrm{False}\} \tag{1-1}$$

文本分类任务可以描述为要努力找到一个函数：

$$\hat{\Phi}：\mathrm{D}\times\mathrm{C}\rightarrow\{\mathrm{True}，\ \mathrm{False}\} \tag{1-2}$$

使$\hat{\Phi}$尽量逼近未知的目标函数Φ。$\hat{\Phi}$称为分类器（Classifier）或者模型（Model）。如果$\Phi(D_i,C_j)=True$，则称文档D_i属于类别C_j；$\Phi(D_i,C_j)=False$，则称文档D_i不属于类别C_j。也就是说，文本分类任务的最终目的就是要找到一个有效的映射函数，准确地实现$D\times C$到值 True 或 False 的映射。

国外文本分类的理论研究可以追溯到 20 世纪 50 年代末，它的发展大致可以分为如下四个阶段[7]：

第一阶段（1958-1964）：主要进行文本自动分类的可行性研究；

第二阶段（1965-1975）：进行文本自动分类的实验研究；

第三阶段（1975-1989）：文本自动分类进入实用化阶段；

第四阶段（1990 年至今）：面向互联网的文本自动分类研究阶段。

1961 年，Maron 发表了关于自动分类的第一篇论文[8]；1975 年 Salton 提出了向量空间模型[9]，在信息检索、人工智能和机器学习技术的推动下，文本自动分类技术在诸多领域取得了应用成果。如今文本分类技术在邮件自动分类、信息过滤、电子会议、图书馆的数字化管理等多方面取得了较为广泛的应用。

相对于国外来说，国内在文本自动分类技术方面的研究起步较晚，始于 20 世纪 80 年代初期。1981 年，侯汉清[10]对文本自动分类技术进行了概述性报告。此后，国内逐渐开始了对文本自动分类技术的研究。由于中文和英文的差别，国外的技术并不能完全适应于中文语料库。中文文本不像英文文本那样单词与单词之间有空格，因此中文文本分类需要进行中文分词。如今，中文分词的技术已趋于成熟，主要有中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS。结合中文文本的特点，逐步形成了中文文本数据的分类研究体系。

根据分类知识获取方法的不同，文本自动分类系统可以分为两大类：基于知识工程（Knowledge Engineering，KE）的分类系统和基于机器学习（Machine Learning，ML）的分类系统。在 20 世纪 80 年代，文本分类系统以知识工程的方法为主，根据领域专家对给定文档集合的分类经验，人工提取出一组逻辑规则，作为计算机文本分类的依据。进入 90 年代后，基于机器学习的文本分类方法日益受到重视，这种方法在准确率和稳定性方面具有明显的优势。系统使用训练样本进行特征选择和分类器参数训练，根据选择的特征对待分类的样本进行形式化，然后使用分类器进行类别判定，最终得到待分类样本的类别。

近几年来，随着互联网技术的迅速发展和普及，对网络内容管理、监控和有害信息过滤的需求越来越大，网络信息的主观倾向性分类受到越来越多的关注。这种分类与传统的文本分类的区别在于：传统的文本分类所研究的对象是文本的客观内容，而倾向性分类所关注的是作者所表达的观点，即“主观倾向性”。分类的结果是要获取特定文本是否支持某种观点的相关信息。这种倾向性文本分类又称为情感分类。

目前基于统计机器学习的文本分类技术相对成熟，被广泛应用于很多系统。其中包括基于概率方法的朴素贝叶斯分类器、基于实例的 k 近邻分类器、基于统计学习理论和结构风险最小原理基础上的支持向量机方法。还有其他的分类方法，包括线性分类器、回归模型、神经网络、决策树方法等。基于机器的学习方法很少考虑文本语义信息，因此出现了基于语义的分类方法。由于自然语言处理研究进展较慢，基于语义的文本分类的发展也受到影响。目前研究者大多是把语义分析、概念网络和机器学习方法相结合，从概念级来获取文本的语义，进而提高文本分类的效果[11]。

1.3 朴素贝叶斯与文本分类

文本分类的一个关键问题就是分类器的设计，朴素贝叶斯分类器是文本分类中常用的分类方法。它是贝叶斯学习方法中最常用的方法，是一种简单而又非常有效的概率分类方法。朴素贝叶斯文本分类方法的一个前提假设是：在给定的文档集中，文档属性是相互独立的。其实质是首先利用贝叶斯条件概率公式，计算出已知文档属于不同文档类别的条件概率（即后验概率）；然后根据最大后验假设将该文档归结为具有最大后验概率的那一类。在自然语言中，贝叶斯假设在实际情况中是不成立的，但是基于该假设的概率分类器仍然具有较好的分类效果。贝叶斯假设一方面大大简化了贝叶斯分类器的计

算量，但导致了贝叶斯分类器的分类质量常常不太理想。为了提高分类的准确性，许多学者提出了相应的改进方法，来提高分类质量。本文针对朴素贝叶斯方法在中文文本分类中的应用问题进行相关研究。

1.4 本文的工作

本文主要是使用朴素贝叶斯分类方法进行中文文本分类的研究，工作如下：

1. 首先讨论了文本分类的关键技术，包括文本分类的整个过程：中文分词、向量空间模型、特征权重计算、文本分类方法等。

2. 详细阐述了贝叶斯理论，在此基础上对朴素贝叶斯文本分类的模型与算法进行了比较详细的分析研究。

3. 实验比较朴素贝叶斯方法中常用的多变量伯努利模型和多项式模型的分类效果，并对多项式模型中的平滑因子进行了改进，来提高分类质量。

4. 通过对比实验分析了常用的特征选择方法对朴素贝叶斯中文文本分类性能的影响，实验表明信息增益和 χ^2 统计量是朴素贝叶斯分类较好的特征选择方法。

5. 根据对朴素贝叶斯文本分类的研究，使用 Java 在 MyEclipse 平台上设计并实现了一个基于朴素贝叶斯方法的中文文本分类系统。

1.5 本文的组织结构

本文的内容组织如下：

第 1 章为绪论，主要介绍了本文的研究背景，文本分类的研究现状，介绍了朴素贝叶斯文本分类方法，同时介绍了本文的研究工作，最后给出了本文的章节安排。

第 2 章论述了文本分类的关键技术。按照文本分类系统的几个主要阶段进行了介绍。文本的向量表示包括文本的预处理、分词、向量空间模型、特征权重；常用的文本分类方法；文本分类性能的评估方法。

第 3 章为朴素贝叶斯分类模型。首先介绍了贝叶斯基础理论，包括贝叶斯定理，极大后验假设和极大似然假设，事件独立性。然后详细阐述了朴素贝叶斯文本分类算法思想，重点介绍了多变量伯努利模型和多项式模型，并对多项式模型的平滑因子进行了改进。通过实验比较两种模型的中文文本分类效果，并实验验证了平滑因子改进的效果。

第 4 章为选择性朴素贝叶斯。主要阐述了常用的特征选择方法，并通过实验比较了

常用的特征选择方法对朴素贝叶斯中文文本分类性能的影响。

第 5 章朴素贝叶斯文本分类的设计与实现。根据对朴素贝叶斯文本分类的研究，使用 Java 在 MyEclipse 平台上设计并实现了一个基于朴素贝叶斯方法的中文文本分类系统。

第 6 章全文总结，并对今后的研究工作进行展望。

第 2 章　文本分类技术

2.1 文本分类的过程

一个完整的中文文本分类系统通常由如下几个功能模块[12][13]组成：

（1）文本预处理：文本预处理是对文档进行分词，去除停用词，其中中文分词是文本预处理的首要步骤。

（2）文本表示：文本表示是文本分类的基础。要将计算机技术应用到文本分类上，必须把文档转化为计算机容易处理的表示形式。目前使用最普遍的文本表示方式是向量空间模型。

（3）文本特征选择：特征选择的目的是为了维数约简，从文档中抽取出若干最有利于文本分类的特征项。特性选择的相关内容将在第 4 章阐述。

（4）特征权重计算：特征权重是用于衡量某个特征项在文档表示中的重要程度或者区分能力的强弱。

（5）分类器学习训练：分类器学习训练的目的是建立分类器，是文本分类的核心问题。利用一定的学习算法对训练样本集进行统计学习，估算出分类器的各个参数，从而建立出对训练集进行学习训练的自动分类器。

（6）测试与评价：利用学习训练阶段建立的分类器，对测试集文档进行分类测试。在完成训练和测试后，选择合适的评价指标对分类器的性能进行评价。如果分类性能不符合要求，需要返回前面步骤，重新再做。

按照文本分类的工作顺序，文本分类可以分为三大阶段，如图 2-1 所示：

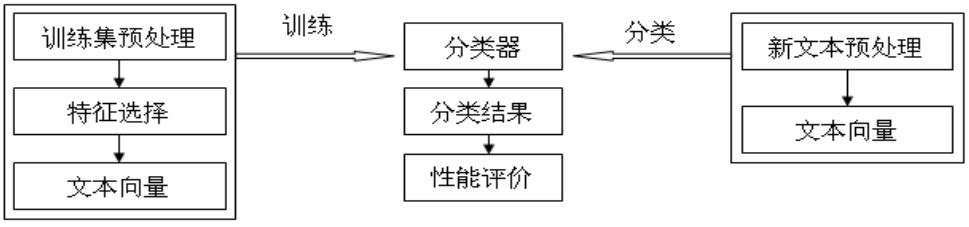

图 2-1　文本分类系统工作过程

第一阶段：将文本表示成文本向量。这个阶段需要完成的工作是先对文本进行预处理，然后进行特征选择和特征权重计算后，将文本转换成向量空间模型的形式。

第二阶段：学习训练阶段。选择分类方法，使用已经表示成文本向量的训练集来建立分类模型。

第三阶段：测试与评价。将第二阶段建立好的分类模型运用于测试集来检验分类效果，并使用评价指标对分类模型的性能进行评价。

2.2 文本向量表示

要将文本表示成文本向量，必须先对文本进行预处理、维数约简即降维、特征权重计算。下面分别介绍这几个步骤。

2.2.1 文本预处理

在文本分类实验中，语料库（Corpus Base）的选择至关重要，语料库的选择有可能影响分类的最终效果[14]。然而由于语料库中存储格式不同、文档的不完整、存在重复文档等问题，为了提高分类性能，避免这些问题影响文本分类系统的后续工作，因此必须对语料库进行预处理工作，去除其中的噪音信息，将内容进行规范化，使其符合文本分类的数据输入要求。

中文分词是文本预处理的一个重要步骤。词是最小的能够独立运用的语言单位，而中文不像英文，中文的词与词之间没有空格，因此计算机处理中文面临的首要问题就是中文的自动分词问题。简单的说，中文自动分词指的是将一个汉字序列切分成一个一个单独的词，也就是让计算机系统在文本的词与词之间自动加上空格或其它边界标记。目前中文自动分词的主要困难是：分词规范、歧义切分和未登录词的识别。现有的中文自动分词方法大致可以分为如下三类[15][16]：

（1）基于字符串匹配的机械分词方法

这种分词方法是根据分词词表按照字符串匹配的原理进行。根据字符串切取方向不同，可将字符串匹配法分为正向匹配和逆向匹配。根据长词优先还是短词优先，可分为最大匹配和最小匹配。根据匹配不成功时重新切取的策略，又可分为增字法和减字法。其中使用较多的最大匹配法是根据一个基本的“长词优先”的切分原则和一个给定的分词词表进行分词。该方法程序实现简单，开发周期短，是一个简单实用的方法[17]。最大匹配法又可分为正向最大匹配法和逆向最大匹配法。

（2）基于统计的分词方法

中文的词是由字组合而成的，文本中相邻的字同时出现的频率越高就越有可能是一

个词。因此，可以统计字与字在语料库中相邻共现的频度，频率越高构成词的可信度就越高。基于统计的分词所使用的统计模型主要有：互信息、N元文法模型、神经网络模型、隐马尔可夫模型和最大熵模型等。这些统计模型主要是利用字与字的联合出现概率作为分词的依据。基于统计的分词方法的优点是：不受处理文本的领域限制；不需要一个分词词典。但这种方法经常会抽取出一些不是词的常用字的组合，如“之一”、“我的”、“有的”等。因此在实际使用的统计分词系统中需要根据一个常用分词词典按照字符串匹配的原理进行分词，同时利用统计的方法识别一些新的词。通过将字符串统计和字符串匹配相结合的方式，既利用了匹配分词的优点：速度快、频率高；又发挥了统计分词方法能根据上下文识别生词并能自动消除歧义的优点。但是解决歧义的方法很大程度上取决于统计语言模型的精度和决策算法，需要有大量的标注语料，并且分词速度也因搜索空间的增大而有所减缓。

（3）基于理解的分词方法

这种分词方法的基本思想是在进行分词时使计算机模拟人脑对句子的理解，进行句法、语义的分析，从而达到分词的效果。因为它要模拟人对句子的理解，需要利用大量的语言知识和信息，而汉语是世界上最复杂的语言，很难统一规划出机器可以直接读取的信息资源。因此，这种分词方法还处于试验阶段。

目前常用的中文分词系统是 ICTCLAS[18][19](Institute of Computing Technology, Chinese Lexical Analysis System)，它是由中国科学院计算技术研究所研制的汉语词法分析系统。ICTCLAS 获得了由国内 973 专家组组织的评测第一名，并获得了第一届国际中文处理研究机构组织的多项测评第一名。其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 采用了层叠隐马尔可夫模型，全部采用 C/C++编写，支持 Linux、FreeBSD 及 Windows 系列操作系统，支持 C/C++/C#/Delphi/Java 等主流的开发语言。ICTCLAS 分词精度达 98.45%，API 不超过 100KB，各种词典数据压缩后不到 3MB，是目前世界上最好的汉语词法分析器。

2.2.2 向量空间模型

对文本进行预处理后，需要使用一种形式化表示方法，使计算机能够高效地处理文本。目前文本表示通常采用向量空间模型（Vector Space Model，VSM）。VSM 是 20

世纪 60 年代末期由 G.Salton 等人提出的，最早用在 SMART 信息检索系统中[20]，现在已经成为自然语言处理中常用的模型。使用 VSM 来进行文本表示，涉及到特征项和特征项权重两个概念。

特征项：特征项是向量空间模型中最小的不可分的语言单元，可以是字、词、词组或短语。一般采用词作为特征项。一个文档的内容可以看成是它所含有的特征项所组成的集合，表示为 $D=(t_1,t_2,...,t_m)$，其中 t_k 是特征项，$1\le k\le m$。

特征项权重：对于含有 m 个特征项的文档 $D=(t_1,t_2,...,t_m)$，每个特征项 t_k 都根据一定的原则被赋予一个权重 w_k，表示它们在文档中的重要程度。这样一个文档 D 可用它含有的特征项及其特征项所对应的权重表示：$D=D(t_1,w_1;t_2,w_2;...;t_m,w_m)$,可以简单记为 $D=D(w_1,w_2,...,w_m)$，其中 w_k 就是特征项 t_k 的权重，$1\le k\le m$。称 $D=D(w_1,w_2,...,w_m)$ 为文档 D 的向量空间模型。

采用向量空间模型进行文本表示，必须经过两个关键步骤：首先根据训练样本集将文本表示成特征项序列 $D=(t_1,t_2,...,t_m)$；然后根据文本特征项序列，对训练样本集中的各个文档进行权重赋值、规范化等处理，将其转化为所需的向量。图 2-2 为向量空间模型下的文本表示，其中每行表示一个文本向量，每列表示一个特征项，w_{ij} 表示第 j 个特征项在第 i 个文档的权重。

	t_1	…	t_j	…	t_m
D_1	w_{11}	…	w_{1j}	…	w_{1m}
…	…	…	…	…	…
D_i	w_{i1}	…	w_{ij}	…	w_{im}
…	…	…	…	…	…
D_n	w_{n1}	…	w_{nj}	…	w_{nm}

图 2-2　向量空间模型下的文本表示

2.2.3 特征权重

文档使用向量空间模式表示后，为了权衡不同的特征项对文档的重要程度和区分影响能力的强弱，需要对特征项进行权重计算。权重的调整一般从两方面考虑：一个词在

某篇文档中出现的次数越多，则对识别文档的贡献越大；一个词在不同文档中出现的次数越多，则它区分不同文档的能力越弱。权重计算的一般方法是利用文本的统计信息，主要是词频。常用的特征权重计算方法有布尔权重、绝对词频（TF）、TF-IDF 权重、TFC 权重、LTC 权重、熵权重等。下面分别介绍常用的计算特征权重的方法，其中使用的变量说明如下：w_{ij} 表示特征项 t_j 在文本 D_i 中的权重；tf_{ij} 表示特征项 t_j 在文本 D_i 中出现的频数；n_j 是训练样本集中出现特征项 t_j 的文档数；N 是训练样本集中总的文档数。

（1）布尔权重

布尔权重是最简单的权重计算方法。如果特征项在文档中出现过，那么文本向量中该特征项的权重值为 1；如果特征项在文档中没有出现，则为 0。由于布尔权重计算方法无法体现特征项在文本中的作用程度，因而在实际运用中 0、1 值逐渐地被更精确的特征项的频率所代替。

（2）绝对词频

使用特征项在文档中出现的频率作为权重，特征权重 w_{ij} 等于 tf_{ij}，即 $w_{ij}=tf_{ij}$。在绝对词频方法中，无法体现低频特征项的区分能力，因为有些特征项频率虽然很高，但分类能力很弱，如很多常用词；而有些特征项虽然频率较低，但分类能力却很强。

（3）TF-IDF 权重

TF-IDF（Term Frequency-inverse Document Frequency）权重[21]是一种非常常用的计算权重的方法，即“词频与倒排文档频数”。权重与特征项在文档中出现的频率成正比，即特征项在文档中出现的次数越多就越重要；同时与语料库中含有该特征项的文档数成反比，即认为特征项在不同文档出现的频率越大，该特征项的重要性就越低。

$$w_{ij}=tf_{ij}\times\log\frac{N}{n_j} \tag{2-1}$$

当 N=n_j 时，权重为 0，为此进行平滑处理，如下式所示：

$$w_{ij}=\log(tf_{ij}+1.0)\times\log(\frac{N+1.0}{n_j}) \tag{2-2}$$

（4）TFC 权重

TFC 权重[6]是 TF-IDF 权重的变形，TF-IDF 没有考虑文档长度对权重的影响，TFC

权重对文档长度进行归一化处理。

$$w_{ij}=\frac{tf_{ij}\times\log(N/n_j)}{\sqrt{\sum_{t_j\in D_j}[tf_{ij}\times\log(N/n_j)]^2}} \tag{2-3}$$

（5）LTC 权重

LTC 权重[6]也是 TF-IDF 权重的变形，可以看做是对式（2-2）的归一化处理。

$$w_{ij}=\frac{\log(tf_{ij}+1.0)\times\log(\frac{N+1.0}{n_j})}{\sqrt{\sum_{t_j\in D_j}[\log(tf_{ij}+1.0)\times\log(\frac{N+1.0}{n_j})]^2}} \tag{2-4}$$

2.3 文本分类方法

文本分类方法是文本分类系统的核心部分。目前许多统计学习、机器学习的算法都被广泛应用于文本分类中，基于统计和机器学习的文本分类技术已成为主流技术。常用的分类算法有：朴素贝叶斯、k 近邻法、决策树、支持向量机、神经网络，Rocchio 分类法、关联规则和组合分类法等。下面介绍几种常用的分类方法。

2.3.1 决策树分类器

决策树（Decision Tree，DT）分类器是一种常用的简单的并广泛使用的分类方法，是一种基于规则的分类器，同样适用于文本分类。决策树在分类的过程中，将数据按树状结构分成若干分枝形成决策树，每个分支包含数据类别归属共性，从每个分枝中提取有用信息，形成规则。决策树算法有多种，目前主要包括：基于信息增益的启发式算法 ID3；基于信息增益率的解决连续属性分类的算法 C4.5；基于 Gini 系数的算法 CART；针对大样本集的可伸缩算法 SLIQ；可并行化算法 SPRINT 等[22]。

文本分类的主要特点是属性很多，这就导致了决策树的结构非常复杂，规模极其巨大，限制了决策树应用于大规模文本的分类问题。决策树的构造主要分为两步：决策树的生成，是指由训练数据生成决策树的过程；决策树的剪枝，是对上一阶段所生成的决策树进行检验、校正和修正的过程。

2.3.2 k 近邻分类器

k 近邻（k-Nearest Neighbor，kNN）分类器[23]是一种经典的分类方法，是基于实例

的学习方法，它使用具体的训练实例进行预测。k 近邻分类器的原理简单：给定一个待分类的测试文档，系统在训练集中查找和待分类文档最相似的已知 k 个文档，然后根据这 k 篇文档的分类来判定待分类文档的类别。k 个邻近文档中多数属于哪一类，测试文档就可以判定归为哪一类。k 近邻分类器是消极的学习方法，不需要建立模型，但 k 近邻分类器是基于局部信息进行预测，因此对噪声非常敏感。因为需要分别计算测试文本和训练文本之间的相似度，所以使用 k 近邻分类一个测试样例开销较大。

在 k 近邻分类器中，参数 k 值的选取至关重要。如果 k 值选取太大，可能会包含与测试文档不相关的文档，造成噪声增加，影响分类精度。相反，如果 k 值选取太小，不能充分体现测试文档的特点，同样影响分类准确性。

2.3.3 朴素贝叶斯分类器

朴素贝叶斯（Naïve Bayes，NB）分类器是基于贝叶斯概率公式的一种分类方法，本文将在第 3 章中做详细介绍。

2.3.4 支持向量机分类器

支持向量机（Support Vector Machine，SVM）目前已经成为一种倍受关注的分类技术。它是 Vapnik 于 1995 年首先提出的，建立在统计学习理论和结构风险最小原理基础之上的，根据有限的训练样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力[24]。

支持向量机的基本思想是：在向量空间中找到一个决策平面，这个平面能“最好”地分割两个分类中的数据点。支持向量机分类方法就是要在训练集中找到对于最大的类间界限的决策平面。

基本的支持向量机的算法是针对两类分类问题的，为了实现对多个类别的识别，需要对支持向量机进行扩展，建立多个两类分类器。

2.4 性能评估方法

对分类性能进行评估是在文本分类器完成了学习训练和测试之后非常重要的一个步骤。常用的文本分类性能评价方法包括精确率、召回率、F1-测试值[25]。

精确率（Precision）是指在所有被判断为正确的文档中，有多大比例是确实正确的。

$$p_{cj} = \frac{\text{被正确分类到 } C_j \text{类的文档数}}{\text{实际分类到 } C_j \text{类的文档总数}} \tag{2-5}$$

召回率（Recall）是指在所有确实正确的文档中，有多大比例被确实判为正确。

$$R_{cj} = \frac{\text{被正确判断为} C_j \text{类的文档数}}{\text{所有应归为} C_j \text{类文档数}} \tag{2-6}$$

F1 测试值是为了综合考虑精确率和召回率，也称为综合分类。

$$F_1 = \frac{2 \times \text{召回率} \times \text{精确率}}{\text{召回率} + \text{精确率}} \tag{2-7}$$

2.5 本章小结

本章主要介绍了文本分类的关键技术。包括文本分类过程、文本向量表示、文本分类方法。在文本向量表示中介绍了预处理中文分词、向量空间模型、特征权重的计算。文本分类方法包括决策树、k 近邻、支持向量机等。最后对文本分类性能的评估方法进行了分析。

第 3 章　朴素贝叶斯分类模型

朴素贝叶斯分类（Naïve Bayes Classification，NBC）是贝叶斯学习方法中最常用的方法，是一种简单而又非常有效的分类方法。朴素贝叶斯分类是建立在经典的贝叶斯概率理论基础之上，其基本思想是利用特征项和类别的条件概率来估算给定文档的类别概率，是一种基于概率统计的分类方法。

3.1 贝叶斯基础理论

3.1.1 贝叶斯定理

假设两个事件 A 和 B，且 $P(A)>0$，在事件 A 已经发生的条件下，事件 B 发生的概率，称为事件 B 在给定事件 A 的条件概率（也称为后验概率），条件概率表示为 $P(B|A)$。相对于条件概率，$P(B)$ 可称为无条件概率（也称为先验概率）。条件概率的公式为：

$$P(B|A)=\frac{P(AB)}{P(A)} \tag{3-1}$$

由条件概率可得到乘法公式：

$$P(AB)=P(A)P(B|A)=P(B)P(A|B) \tag{3-2}$$

假设 S 是试验 E 的样本空间，A 为 E 的事件，B_1，B_2，…，B_n 为 S 的一个划分，且 $P(B_i)>0$ (i=1,2,…,n)，则全概率公式为：

$$\begin{aligned}P(A)&=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+...+P(A|B_n)P(B_n)\\&=\sum_{i=1}^{n}P(A|B_i)P(B_i)\end{aligned} \tag{3-3}$$

由条件概率公式和全概率公式可得如下的贝叶斯公式[26]：

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{P(A)}=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}P(A|B_j)P(B_j)} \tag{3-4}$$

3.1.2 极大后验假设与极大似然假设

贝叶斯定理提供了一种基于假设 h 的先验概率来检验假设 h 概率的方法。用 $P(h)$ 表示没有训练数据前假设 h 的先验概率；用 $P(D)$ 表示将要观察的训练数据 D 在没确定某一假设成立时 D 的概率；$P(D|h)$ 表示假设 h 成立的情况下 D 的条件概率。在机器学习中，感兴趣的是 $P(h|D)$，即给定一个训练样本数据 D 时 h 成立的概率，即 h 的后验概率，它反映了在看到训练样本数据 D 后假设 h 成立的置信度。由贝叶斯公式求得后验概率为：

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)} \tag{3-5}$$

在许多场合，都需要考虑对于给定训练样本数据 D，在候选假设集合 H 中寻找可能性最大的假设 h，$h\in H$。这种具有最大可能性的假设称为极大后验假设（maximum a posteriori），记为 h_{MAP}。

$$h_{MAP}=\arg\max_{h\in H}P(h|D)=\arg\max_{h\in H}\frac{P(D|h)P(h)}{P(D)} \tag{3-6}$$

因为 $P(D)$ 是不依赖于 h 的常量，式（3-6）可以简化为：

$$h_{MAP}=\arg\max_{h\in H}P(D|h)P(h) \tag{3-7}$$

在特定情况下，可以假设 H 中的任意假设 h_i 和 h_j，都有 $P(h_i)=P(h_j)$，即它们的先验概率相同，这样式（3-7）就可以进一步进行简化，只考虑 $P(D|h)$ 来寻找极大可能假设。$P(D|h)$ 被称为给定 h 时训练样本数据 D 的似然度（likelihood），而 $P(D|h)$ 最大的假设被称为极大似然假设 h_{ML}。

$$h_{ML}=\arg\max_{h\in H}P(D|h) \tag{3-8}$$

3.1.3 事件的独立性

假设 A，B 是试验 E 的两个事件，一般情况下事件 A 的发生对事件 B 的发生是有

影响的，也就是条件概率 $P(B|A)$ 与无条件概率 $P(B)$ 不相等， $P(B|A) \neq P(B)$ 。但是在有些情况下，当这种影响不存在时就会有 $P(B|A) = P(B)$ ，这时称 A，B 为相互独立事件，则：

$$P(AB) = P(A)P(B|A) = P(A)P(B) \tag{3-9}$$

同理，对于 n 个事件 A_1，A_2，…，A_n，如果

$$P(A_1A_2...A_n) = P(A_1)P(A_2)...P(A_n) \tag{3-10}$$

则称 A_1，A_2，…，A_n 为相互独立事件。

3.2 朴素贝叶斯分类器

基于贝叶斯理论的分类方法的实现主要有朴素贝叶斯分类器和贝叶斯信念网络。本文主要研究朴素贝叶斯分类器[27][28][29][30]。

当给定训练样本集，对新样本最有可能的分类可以使用极大后验假设 MAP 来解决。假设每个训练样本都用一个 n 维的向量 $X = (x_1, x_2, ..., x_n)$ 表示，分别描述对 n 个属性 A_1，$A_2, \ldots, A_n$ 的度量；类标号的集合为 $\{C_1, C_2 ..., C_m\}$。当给定一个测试样本 X，应用极大后验假设 MAP 进行分类，得到最可能的类标号 $c(x)$ 为：

$$c(x) = \arg\max_{c \in C} P(x_1, x_2, ..., x_n | c)P(c) \tag{3-11}$$

其中，c 为类集合中的某一类别。从上式可以看出，主要是要估计两个概率值。$P(c)$ 的值容易估计，只要计算每个类标记 c 出现在训练样本集中的频率即可。但估计每个 $P(x_1, x_2, ..., x_n | c)$ 就不太容易。为此朴素贝叶斯分类器假定：对于样本的 n 个属性之间相互条件独立。根据事件的独立性， $P(x_1, x_2, ..., x_n | c)$ 可以表示为每个单独属性的概率的乘积，如下式：

$$P(x_1, x_2, ..., x_n | c) = \prod_{j=1}^{n} P(x_j | c) \tag{3-12}$$

朴素贝叶斯分类器公式为：

$$c(x) = \arg\max_{c \in C} P(c) \prod_{j=1}^{n} P(x_j \mid c) \tag{3-13}$$

有了条件独立性假设，就不必计算 X 的每一个组合的类条件概率，只需对于给的类别，分别计算每个 x_j 的条件概率。

朴素贝叶斯分类器中的条件独立性假设可用图 3-1 表示，其中：$A_1, A_2, \cdots, A_n$ 表示 n 个属性结点，C 表示目标类结点。

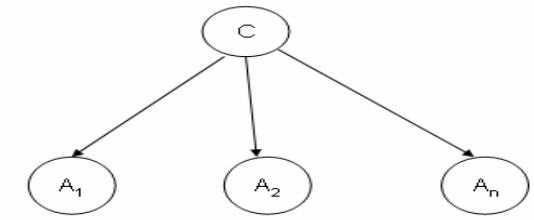

图 3-1　朴素贝叶斯分类条件独立性假设示意图

3.3 朴素贝叶斯文本分类

3.3.1 朴素贝叶斯文本分类算法

朴素贝叶斯文本分类的任务就是将表示成为向量的待分类文档 $D_i(x_1, x_2, ..., x_n)$ 归类到与其关联最紧密的类别集合 $C = \{C_1, C_2 ..., C_m\}$ 中的某一类。其中 $D_i(x_1, x_2, ..., x_n)$ 为待分类文档 D_i 的特征向量，$C = \{C_1, C_2 ..., C_m\}$ 为给定的文档类别集合。也就是说，求解向量 $D_i(x_1, x_2, ..., x_n)$ 属于给定类别 $C_1, C_2 ..., C_m$ 的概率值 $(p_1, p_2, ..., p_m)$，其中 p_j 为 $D_i(x_1, x_2, ..., x_n)$ 属于 C_j 的概率，则 $\max(p_1, p_2, ..., p_m)$ 所对应的类别就是文档 D_i 所属的类别。假设 D_i 为一任意文档，根据贝叶斯分类器，文档 D_i 属于 C_j 的概率为：

$$P(C_j \mid D_i) = \frac{P(C_j)P(D_i \mid C_j)}{P(D_i)} = \frac{P(C_j)P(x_1, x_2, ..., x_n \mid C_j)}{P(x_1, x_2, ..., x_n)} \tag{3-14}$$

其中 $P(x_1, x_2, ..., x_n)$ 对应所有类均为常量，所以只需估算 $P(C_j)P(x_1, x_2, ..., x_n \mid C_j)$，求解式（3-14）的最大值可转化为如下公式：

$$c(D_i) = \arg\max_{c_j \in C} P(x_1, x_2, ..., x_n \mid C_j) P(C_j) \tag{3-15}$$

朴素贝叶斯文本分类的一个前提假设是：在给定的文档类别下，文档属性即特征项是相互独立的。即：

$$P(x_1, x_2, ..., x_n \mid C_j) = \prod_{i=1}^{n} P(x_i \mid C_j) \tag{3-16}$$

所以式（3-15）可简化为

$$c(D_i) = \arg\max_{c_j \in C} P(C_j) \prod_{i=1}^{n} P(x_i \mid C_j) \tag{3-17}$$

朴素贝叶斯文本分类的关键就是计算 $P(C_j)$ 和 $P(x_i \mid C_j)$。计算 $P(C_j)$ 和 $P(x_i \mid C_j)$ 的过程就是建立分类模型（或者说学习）的过程，通过计算 $P(C_j)$ 和 $P(x_i \mid C_j)$，求出后验概率，返回后验概率最大的类别。

根据 $P(D_i \mid C_j)$ 计算方式的不同，朴素贝叶斯分类方法可分为最大似然模型（Maximum Likelibood Model，MLM）、多变量伯努利模型（Multi-variate Bernoulli Model，MBM）、多项式模型（Multinomial Model，MM）、泊松模型（Poison Model，PM）等。本文主要讨论多变量伯努利模型和多项式模型。

3.3.2 多变量伯努利模型

多变量伯努利模型[31][32]，只考虑特征项在文档中是否出现，权重为 1 表示特征项在文档中出现，0 表示未出现。这种方法不考虑特征词出现的顺序，也不考虑特征项在文档中出现的次数。设特征项为 n 项，将文档当做一个事件，这个事件是通过 n 重伯努利实验产生的，即某个特征项出现或者不出现。设 B_{xt} 表示特征项在文档中的出现情况，表示出现或不出现，则有：

$$P(D_i \mid C_j) = \prod_{i=1}^{n} (B_{xt} P(x_i \mid C_j) + (1 - B_{xt})(1 - P(x_i \mid C_j))) \tag{3-18}$$

$P(x_i \mid C_j)$ 表示属于 C_j 类的情况下 x_i 出现的概率。从公式可以看出，在多变量伯努利模型中，文档是所有特征项的类条件概率之积，如特征项在文档中出现，乘以的是 $P(x_i \mid C_j)$,若不出现，乘以的是 1- $P(x_i \mid C_j)$。

给定类别C_j，特征项x_i的类条件概率，$P(x_i \mid C_j)$的估算采用文档频数：

$$P(x_i \mid C_j) = \frac{n_{ij}}{n_j} \tag{3-19}$$

其中，n_{ij}为类C_j中包含特征项x_i的文档数，n_j为类C_j中包含的总文档数。为了避免出现零概率，通常需要加入平滑因子，一般采用 m 估计（m-estimate）方法来估计条件概率：$P(x_i \mid y_j) = \frac{n_c + mp}{n+m}$，其中 m 称为等价样本大小的参数，p 是用户指定的参数。

使用 m 估计方法，m=2，p=1/2，这样$P(x_i \mid C_j)$的计算可以表示为如下所示：

$$P(x_i \mid C_j) = \frac{1+n_{ij}}{2+n_j} \tag{3-20}$$

而先验概率$P(C_j)$的计算如下：

$$P(C_j) = \frac{\text{类}C_j\text{中的训练文本总数}}{\text{训练样本总数}} \tag{3-21}$$

3.3.3 多项式模型

多变量伯努利模型[33][34]忽略了每个特征项在文档中出现的次数，然而决定一个文档的类别时，这些信息具有重要的价值。在多项式模型中，考虑了特征项在文档中出现的次数。文档被看作是一系列单词序列，并且假定文档的长度与类别无关，而且文档中的任何一个词与它在文档中的位置以及上下文无关。文档属于类C_j时特征词x_i出现一次的概率为$P(x_i \mid C_j)$，出现n_k次的概率为$P(x_i \mid C_j)^{n_k}$，假定共有 n 个词，则$n = n_1 + n_2 + ... + n_k$,则有[34]：

$$P(D_i \mid C_j) = n! \prod_{i=1}^{n} \frac{P(x_i \mid C_j)^{n_k}}{n_k!} \tag{3-22}$$

在多项式模型中，$P(x_i \mid C_j)$采用词频估算：

$$P(x_i \mid C_j) = \frac{\sum_{k=1}^{|D|} N(x_i, D_k)}{\sum_{i=1}^{|V|}\sum_{k=1}^{|D|} N(x_i, D_k)} \tag{3-23}$$

其中，$\sum_{k=1}^{|D|} N(x_i, D_k)$ 表示特征项 x_i 在类 C_j 的各文档中出现的次数之和；$\sum_{i=1}^{|V|}\sum_{k=1}^{|D|} N(x_i, D_k)$ 为类 C_j 中所有特征项的总次数。

为了避免出现零概率，通常需要加入平滑因子，其中 m=V，p=1/|V|，如下所示：

$$P(x_i \mid C_j) = \frac{\sum_{k=1}^{|D|} N(x_i, D_k) + 1}{\sum_{i=1}^{|V|}\sum_{k=1}^{|D|} N(x_i, D_k) + |V|} \tag{3-24}$$

其中，V 是训练样本的单词表（即抽取单词，单词出现多次，只算一个）。

在多项式模型中，先验概率 $P(C_j)$ 的计算如下：

$$P(C_j) = \frac{\text{类}C_j\text{中特征项总数}}{\text{训练样本的特征项总数}} \tag{3-25}$$

3.3.4 两个模型的区别

两个模型的计算粒度不一样，多变量伯努利模型以文件为粒度，多项式模型以单词即特征项为粒度，因此两者的先验概率和类条件概率的计算方法都不同。

计算条件概率时，对于一个文档 D，多项式模型中，只有在 D 中出现过的特征项，才会参与条件概率的计算；在多变量伯努利模型中，没有在 D 中出现的特征项也会参与计算，不过是作为“反方”参与的。

McCallum 和 Nigam[34]对两种模型作过比较，但都是对英文文本进行分类，在 3.5 小节设计了一个实验，使用这两种模型对中文文本进行分类，比较其分类精度。

3.4 朴素贝叶斯分类器的改进

由于文本分类的数据稀疏，会出现零概率问题。在多项式模型中，为了解决零概率问题，采用了 m 估计方法。在式（3-24）中，假设每个特征项出现的次数比实际出现的次数多一次，V 是训练样本的单词表。然而，如果训练集中一些类包含的单词量大，而

另一些类包含的单词量相对少，对于单词量少的类别来说，使用式（3-24）计算类条件概率时，当“训练样本的单词表”与“类 C_j 中的单词总和”相近或更大时，就使我们引入的虚拟样本数目超过训练集中的样本数目，这样就会降低训练样本的作用，从而造成较大的估计偏差。

为了解决这一问题，采用修改的平滑技术。p=1/|V|，m 取训练样本的单词表 V 的 α 倍，其中 α 取特征项 x_i 在所有类中出现的概率，采用新的平滑因子的 $P(x_i \mid C_j)$ 的公式为：

$$P(x_i \mid C_j) = \frac{\sum_{k=1}^{|D|} N(x_i, D_k) + \alpha}{\sum_{i=1}^{|V|}\sum_{k=1}^{|D|} N(x_i, D_k) + \alpha \mid V \mid} \tag{3-26}$$

其中，$\alpha = P(x_i / C)$。

3.5 实验设计与结果比较

3.5.1 实验 1：多项式模型与多变量伯努利模型比较

本节设计了一个实验来比较朴素贝叶斯的多变量伯努利模型 MBM 和多项式模型 MM 在中文文本分类中的分类精度。实验数据采用复旦大学自然语言处理实验室提供的基准语料库。选取了艺术、计算机、农业、经济、政治、体育 6 类，每类各取 500 篇共 3000 篇文档做实验，采用重复 10 折交叉验证法，将数据随机分割成 10 份，每次取 9 份进行训练，1 份进行测试，然后取其平均值。

实验使用了数据挖掘平台 Weka，Weka 是怀卡托智能分析环境（waikato envimnment for knowledge analysis）,是由新西兰怀卡托大学开发的基于 Java 环境的、开源的机器学习及数据挖掘软件[35][36]。Weka 作为一个公开的数据挖掘工作平台，汇集了当今前沿的机器学习算法及数据预处理工具，得到了广泛的认可，是现今最完备的数据挖掘工具之一。Weka 工作平台包含能处理所有的标准数据挖掘问题的方法：回归、分类、聚类、关联规则等，还提供了用于数据可视化和预处理的工具。同时，由于其源码的开放性，Weka 不仅可以用于完成常规的数据挖掘任务，也可以用于数据挖掘的二次开发。

本实验主要是利用 Weka 的二次开发功能，编写了朴素贝叶斯多变量伯努利模型的算法，与 Weka 中原有的朴素贝叶斯多项式模型进行比较实验。

实验步骤如下：

（1）将 Weka 源代码加载到 NetBeans 平台。

（2）将下载的语料库中的 6 大类 3000 篇文章通过自行编写的继承了 weka.core.converters.TextDirectoryLoader 的类进行加载。

（3）通过 weka.filters.unsupervised.attribute.StringToWordVector 类将文本转成向量。

（4）利用自行编写的多变量伯努利模型算法与 Weka 下现有的多项式模型算法进行比较。

通过精确率(Precision)、召回率(Recal1)、F1 测试值 3 项主要指标进行评估测试，实验结果如表 3-1 所示：

表 3-1 MM 与 MBM 分类结果比较

类别	多变量伯努利模型（MBM）			多项式模型（MM）		
	精确率	召回率	F1 值	精确率	召回率	F1 值
计算机	52.3%	100.0%	68.7%	97.8%	99.8%	98.8%
艺术	97.9%	84.2%	90.5%	97.7%	94.6%	96.1%
农业	98.7%	76.2%	86.0%	88.5%	90.4%	89.4%
经济	80.2%	80.4%	80.3%	90.6%	76.8%	83.1%
政治	97.3%	71.8%	82.6%	86.0%	93.2%	89.4%
体育	93.6%	67.0%	78.1%	89.8%	95.0%	92.3%
平均值	86.7%	79.9%	81.0%	91.7%	91.6%	91.5%

从上面的实验结果可以看出多变量伯努利模型 MBM 的分类效果低于多项式模型 MM。主要原因是多变量伯努利模型没有考虑特征项出现的次数。

3.5.2 实验 2：改进后的多项式模型与多项式模型比较

在实验 1 的基础上，对多项式模型的类条件概率的平滑因子进行改进，在原有的语料库的基础上进行实验，实验结果如表 3-2 所示：

表 3-2 改进后的 MM 与 MM 分类结果比较

类别	多项式模型（MM）			改进后的多项式模型		
	精确率	召回率	F1 值	精确率	召回率	F1 值
计算机	97.8%	99.8%	98.8%	97.7%	99.8%	98.7%
艺术	97.7%	94.6%	96.1%	99.8%	96.6%	98.2%
农业	88.5%	90.4%	89.4%	92.3%	93.4%	92.8%
经济	90.6%	76.8%	83.1%	90.3%	89.6%	90.0%
政治	86.0%	93.2%	89.4%	93.5%	94.0%	93.8%
体育	89.8%	95.0%	92.3%	97.2%	97.4%	97.3%
平均值	91.7%	91.6%	91.5%	95.1%	95.1%	95.1%

从实验结果可以看出，分类精确率提高了 3.4 个百分点，表明平滑因子的修改是有效的。

3.5.3 实验小结

为了直观的反映实验结果，将上面两个实验结果使用图来表示，如图 3-2 所示：

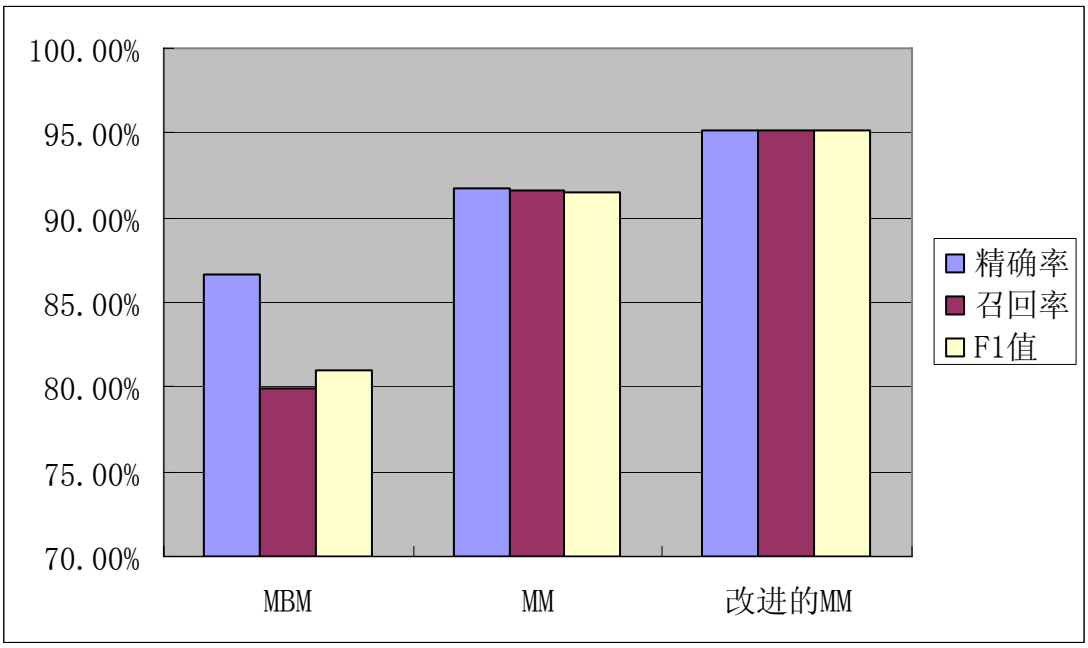

图 3-2　实验 1 和实验 2 结果对比

由上图可得出实验结论：

（1）在中文文本分类中，朴素贝叶斯多项式模型的分类效果优于多变量伯努利模型。

（2）改进后的朴素贝叶斯多项式模型的分类性能有所提高。

3.6 本章小结

本章首先介绍了贝叶斯基础理论，包括贝叶斯定理，极大后验假设和极大似然假设，事件独立性。然后详细阐述了朴素贝叶斯文本分类算法思想，重点介绍了多变量伯努利模型和多项式模型，为了进一步提高分类精度，对多项式模型的平滑因子进行改进，实验表明其具有良好的分类效果。

第 4 章　选择性朴素贝叶斯方法

在文本分类中，特征项应该具有如下特性：特征项要能够确实标识文本内容；特征项具有将目标文本与其他文本进行区分的能力；特征项的个数不能太多；特征项分离要比较容易实现。如果把所有的词都作为特征项，那么特征向量的维数将过于巨大，从而导致计算量太大，在这样的情况下，要完成文本分类几乎是不可能的。特征选择的目的就是在不改变文本核心信息的前提下尽可能的减少要处理的特征项数，从而降低向量空间维数，简化计算，提高文本处理的速度和效率。现存多种特征选择方法[37][38][39]，本章主要研究在中文语料库下，常用的特征选择方法对朴素贝叶斯分类性能的影响。

4.1 常用的特征选择方法

4.1.1 文档频率

文档频率（Document Frequency，DF）是指在训练语料库中出现了某个特征项的文档数，是最简单的一种特性选择方法。在使用过程中：首先从训练语料库中统计出包含某个特征项的文档频率，然后根据事先设定的阈值，从特征向量空间中去掉文档频率小于阈值的特征项，这是因为该特征项在文档中出现的频率太低，没有代表性。而当该特征项的文档频率大于另外一个阈值时，也需要从特征空间中去掉，这是因为该特征项在文档中出现的频率太高，没有区分度。

基于文档频率的特征选择方法可以降低向量空间的维数，去掉一部分的噪声。这种方法简单，计算量小。但有时频率低的词含有更多的信息，简单的删除会影响分类精度。

4.1.2 信息增益

信息增益[39]（Information Gain，IG）是基于信息论中熵的概念，是机器学习中普遍使用的特征选择方法。增益是指含有特征项 t 和没有特征项 t 时，为整个分类所能提供的信息量的差别。在信息增益中，重要性的衡量标准就是看特征能够为分类系统带来多少信息，带来的信息越多，该特征项就越重要。信息量的多少由熵来衡量。信息增益表示为：

$$IG(t_i)=-\sum_{j=1}^{n}P(C_j)\log P(C_j)+P(t_i)\sum_{j=1}^{n}P(C_j\mid t_i)\log P(c_j\mid t_i)+P(\bar{t_i})\sum_{j=1}^{n}P(C_j\mid \bar{t_i})\log P(C_j\mid \bar{t_i}) \quad (4\text{-}1)$$

其中，$P(C_j)$表示语料库中出现C_j类的概率，$P(t_i)$表示特征项t_i出现在语料库文档中的文档概率，$P(C_j \mid t_i)$表示当特征项t_i出现时，文档属于C_j类的条件概率，$P(\overline{t_i})$表示特征项t_i不出现在语料库中的文档概率，$P(C_j \mid \overline{t_i})$表示当特征项$t_i$不出现时，文档属于$C_j$类的条件概率。信息增益的方法是：首先对训练样本中出现的每个特征项计算其信息增益，然后指定一个阈值，从特征空间中删除那些信息增益低于此阈值的特征项，或者确定要选择的特征个数，按照增益值从高到底的顺序选择特征来组成特征向量。信息增益考虑了特征项未发生的情况，特征项不出现的情况有可能对文本类别具有贡献，但也有可能对特征分值带来干扰。

4.1.3 χ^2统计量

χ^2统计量[39]（CHI）用来衡量特征项t_i和类别C_j之间的相关联的程度。特征项t_i对类别C_j的 CHI 值越高，它与该类之间的关联性就越大。如果特征项t_i和类别C_j之间相互独立，那么特征项t_i对类别C_j的 CHI 值为 0。

假设 A 表示包含特征项t_i且属于C_j类的文档频数；B 表示包含特征项t_i但不属于C_j类的文档频数；C 表示不包含特征项t_i但却属于C_j类的文档频数；D 表示不包含特征项t_i也不属于C_j类的文档频数；N 表示训练语料库总的文档数。可以使用表 4-1 表示。

表 4-1　特征项t_i与类C_j关系示意图

类别 / 特征项	C_j	$\overline{C_j}$
t_i	A	B
$\overline{t_i}$	C	D

特征项t_i对C_j的 CHI 值如下所示：

$$\chi^2(t_i, C_j) = \frac{N \times (A \times D - C \times B)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{4-2}$$

由于最终只需获取特征项对某个类别的 CHI 值的大小顺序，而不需要具体的值，N、A+C、B+D 对于同一类别文档中的所有特征项来说都是一样的，因此在实际应用中通常将式（4-2）进行简化，如式（4-3）所示：

$$\chi^2(t_i, C_j) = \frac{(A \times D - C \times B)^2}{(A+B) \times (C+D)} \tag{4-3}$$

对于多类问题，分别计算t_i对于每个类别的 CHI 值，然后用下面的公式计算t_i在整个训练语料的 CHI 值。

$$\chi^2_{MAX}(t_i) = \max_{j=1}^{m} \chi^2(t_i, C_j) \tag{4-4}$$

其中 m 为类别数。将 CHI 值从大到小排序，选取前 k 个即可（k 值根据需要自行确定）。也可以设置一个阈值，从向量空间中移除 CHI 统计量低于给定阈值的特征项，将剩余的即高于阈值的特征项作为文档的表示特征。

4.1.4 互信息

互信息[40]（Mutual Information，MI）的基本思想是：互信息越大，特征项t_i和类别C_j同时出现的概率就越大。计算互信息的公式为：

$$MI(t_i, C_j) = \log \frac{P(t_i, C_j)}{P(t_i)P(C_j)} = \log \frac{P(t_i \mid C_j)}{P(t_i)}$$

$$\approx \log \frac{A \times N}{(A+C) \times (A+B)} \tag{4-5}$$

其中 A、B、C、D、N 的含义同上表 4-1。

如果特征项t_i和类别C_j无关，则$P(t_i, C_j) = P(t_i) \times P(C_j)$，那么$MI(t_i, C_j)$=0。为了将互信息应用于多个类别，与 CHI 统计的处理方法类似，可以使用最大值方法计算特征项t_i对于C_j的互信息，公式如下：。

$$MI_{\max}(t_i) = \max_{j=1}^{m} P(C_j) \times MI(t_i, C_j) \tag{4-6}$$

其中 m 为类别数。

以上是文本分类中比较常用的特征选择方法，实际还有许多其它特征选择的方法，

如：期望交叉熵法、文本证据权法等。文本分类特征选择的目的就是要降低特征向量的维数，避免“维数灾难”现象的发生，同时尽量减少噪声，来提高分类精度，减少分类所需时间。

4.2 实验设计与结果分析

本节设计了一实验来测试常用的特征选择方法对朴素贝叶斯文本分类性能的影响。实验数据与 3.5 小节中的实验一样，采用复旦大学自然语言处理实验室提供的基准语料库，选取了艺术、计算机、农业、经济、政治、体育 6 类做为实验数据，每类各取 500 篇共 3000 篇文档做实验，仍然采用重复 10 折交叉验证法。

实验步骤如下：

（1）采用中科院计算所汉语分词系统 ICTCLAS 对文档进行分词。

（2）对分词后的文本进行粗降维，即将停用词，低频词从文档中去除；停用词表包括词数 903 个。

（3）使用 WVTool 构建向量模型，编写程序将文本转化为文本向量形式，在利用 WVTool 时，使用了 WVTWordList 的 pruneByFrequency 方法，设置文档频率为 2 到 1000，这样将低频词和高频词过滤掉，然后采用 TFIDF 来计算特征项权重，形成文本向量。然后将文本特性向量转化为 WEKA 能识别的.arff 文件格式,并将数据集的正常格式进行稀疏矩阵转化。

（4）使用互信息、信息增益、χ^2 统计量三种特性选择方法分别在多变量伯努利模型 MBM、多项式模型 MM 和改进的多项式模型上进行实验，比较它们对朴素贝叶斯文本分类的性能影响。其中特征维数由 20 个特征项逐步增加到 10000 个特征项。

三种特性选择方法在三种模型上的分类精度结果如表 4-2, 表 4-3, 表 4-4 所示：

表 4-2　特征选择方法在 MBM 中的分类精度比较

特征数目		20	50	100	200	500	1000	2000	5000	10000
特征选择方法	MI	61.94%	62.54%	69.03%	69.80%	82.90%	87.77%	91.17%	92.32%	90.96%
	IG	63.56%	80.64%	86.88%	85.82%	87.41%	89.87%	89.87%	90.33%	90.96%
	χ^2	61.52%	83.38%	86.32%	86.43%	87.83%	89.42%	89.75%	90.24%	90.96%

表 4-3　特征选择方法在 MM 中的分类精度比较

特征数目		20	50	100	200	500	1000	2000	5000	10000
特征	MI	62.43%	64.77%	74.04%	74.21%	87.74%	90.33%	91.51%	93.44%	94.24%
选择	IG	71.15%	89.70%	91.97%	91.44%	92.56%	93.25%	93.49%	94.17%	94.24%
方法	χ^2	66.66%	89.90%	91.72%	92.03%	92.43%	92.84%	93.43%	94.02%	94.24%

表 4-4　特征选择方法在改进的 MM 中的分类精度比较

特征数目		20	50	100	200	500	1000	2000	5000	10000
特征	MI	39.18%	56.60%	78.43%	80.95%	91.13%	93.86%	94.92%	94.93%	94.51%
选择	IG	22.28%	90.56%	92.52%	91.51%	93.16%	93.78%	94.16%	94.43%	94.51%
方法	χ^2	65.08%	90.36%	92.22%	92.53%	92.86%	93.48%	94.01%	94.42%	94.51%

为了更清楚的反映实验结果，将上面三表的实验数据用下图 4-1，图 4-2，图 4-3 所示：

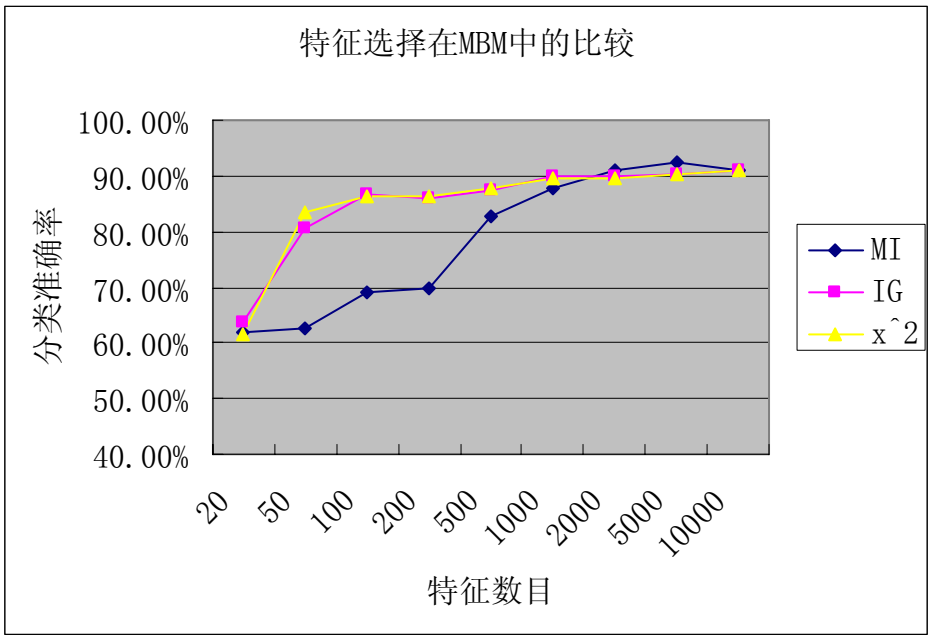

图 4-1　特征选择方法在 MBM 中的比较

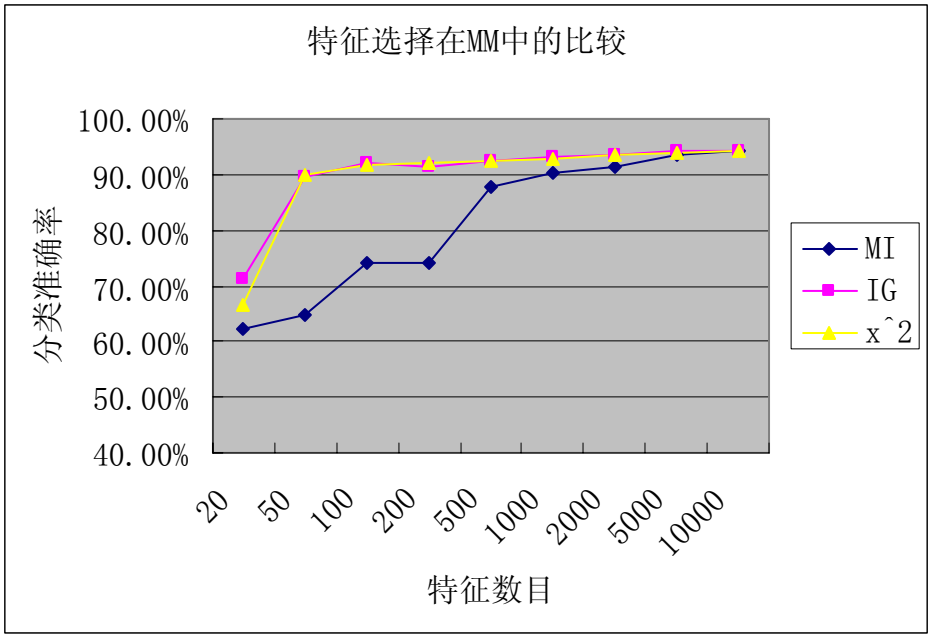

图 4-2　特征选择方法在 MM 中的比较

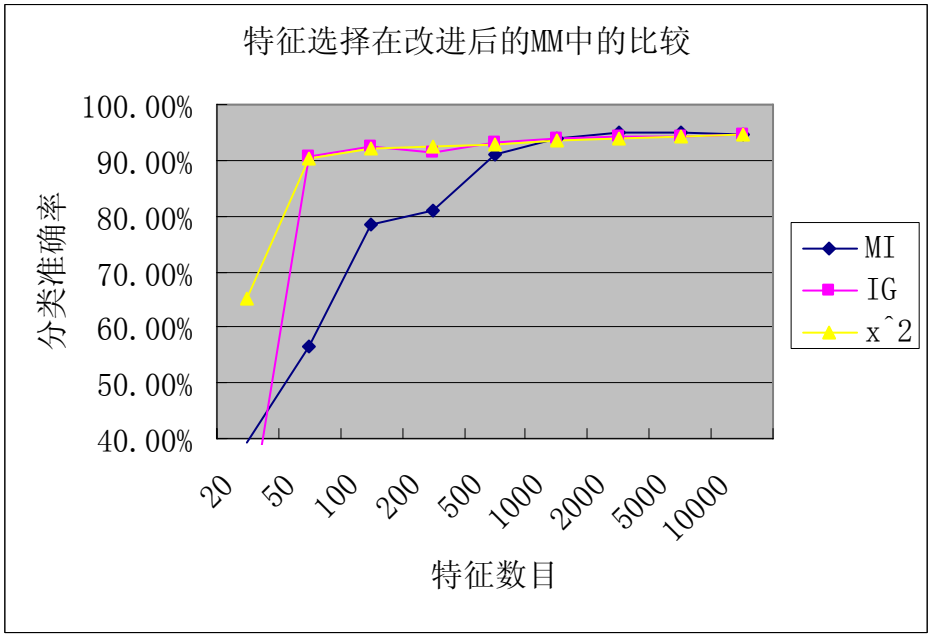

图 4-3　特征选择方法在改进的 MM 中的比较

从实验结果表 4-2，表 4-3，表 4-4，与图 4-1，图 4-2，图 4-3 中可以看出各种特征选择方法对朴素贝叶斯分类准确率的影响是随着特征数目的增加先增加后逐渐趋于相同的正确率。在将文本转成特征向量时，使用 WVTool 的 WVTWordList 类的

pruneByFrequency 方法将文档频率小于 2 和大于 1000 的特征项从文档中删除，因为低频词较低的出现频率，必然只属于较少的类别这样就尽可能的减少了低频词对分类的噪声；同样将文档频率大于 1000 的高频词去掉，因为这些词在不同的类别都出现，对分类的判断也没有意义。由于事先将这些低频词和高频词都删除了，所以在特征数目为 50 时，就表现出了较好的分类效果，除互信息的特征选择方法外，其他的特征选择方法分类准确率都达到了 80%以上。在实验中当特征数目为 10000 时，不管是什么特征选择，对于同种分类算法分类准确率都到达了相同。也就是说，这时基本上包含了所有出现的重要特征项，三种特征选择方法所选择出来的特征项都基本一样了，故最后出现了分类性能趋于一致的现象。

不论是朴素贝叶斯的多变量伯努利模型，还有多项式模型，特征选择中的信息增益和 χ^2 统计都具有相似的分类效果，当特征数目到达 10000 时，分类性能随着特征数目的改变没有很大的变化，这两种方法表现出较好的稳定性，具有较好的分类效果，优于互信息。

4.3 特征选择实验比较

将本章特征选择之后的实验结果与第 3 章没有使用特征选择的分类结果进行比较，比较结果如下图 4-4 和表 4-5 所示：

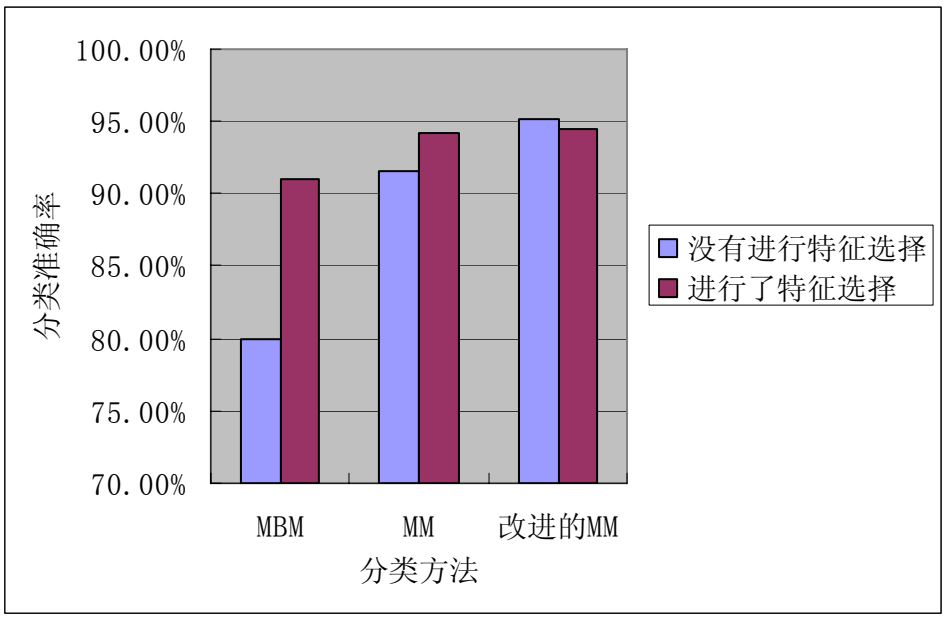

图 4-4　特征选择与否的分类效果比较图

表 4-5　特征选择与否的分类效果比较

	没有进行特征选择	进行了特征选择
多变量伯努利模型	79.93%	90.96%
多项式模型	91.63%	94.24%
改进后的多项式模型	95.13%	94.51%

从图中可以看出通过特征选择，分类准确率提高了，特别是对于多变量伯努利模型。因为多变量伯努利模型仅考虑一个词在文章中是否出现，出现则为 1，否则则为 0，而没有考虑词频，因此在使用所有的特征项进行文本分类时，准确率不高。然而通过了特征选择，不论是采用信息增益还是 χ^2 统计值，都使分类准确率提高了 10%。对于多项式模型分类准确率比没有使用特征选择时提高了 3%，而对于改进的朴素贝叶斯方法，进行特征选择前后的分类准确率没有多大的变化，进一步表明了改进方法的效果。

4.4 本章小结

朴素贝叶斯分类模型是建立在属性之间条件独立性假设之上，因此特征选择的好坏与否对分类精度有较大影响。现存多种特征选择方法，本章主要研究了在中文文本分类中，常用的特征选择方法对朴素贝叶斯分类性能的影响。通过实验表明信息增益和 χ^2 统计量是朴素贝叶斯文本分类较好的特征选择方法。

第 5 章　朴素贝叶斯文本分类的设计与实现

5.1 系统的实现

在前面的章节中，分别对朴素贝叶斯分类器和特征选择对朴素贝叶斯分类的影响进行了分析和实验。本章运用前面章节的关键技术，使用 Java 语言，MyEclipse 开发平台设计并实现了一个基于朴素贝叶斯方法的中文文本分类系统。本系统分为 4 大部分：分词、向量空间模型表示、训练和测试。系统的运行界面如下图所示：

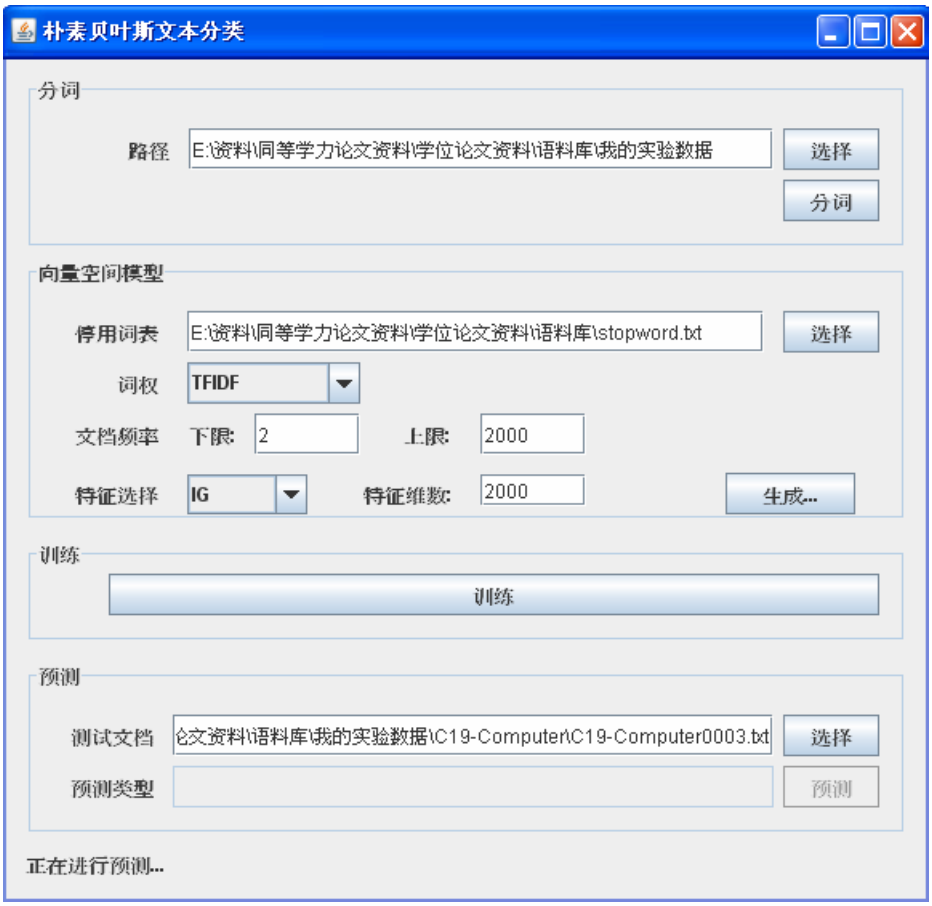

图 5-1　朴素贝叶斯文本分类系统界面

5.2 系统模块

分词：调用中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS 进行分词。

向量空间模型表示：首先将分词后的训练文本使用停用词表删除 900 多个停用词，进行粗降维；之后使用文档频率设置文档频率上限和下限去掉低频词和高频词，最后通过特征选择选取一定数量的特征项，从而达到降维的目的。本系统采用的特征选择方法是通过上章实验效果较好的信息增益与 χ^2 统计值，将文本表示成特征向量时使用的特征权重是 TFIDF。

训练：经过特征选择后，训练文本集就被表示成为了特征向量。训练阶段就是构造分类器。本系统采用的分类器是朴素贝叶斯分类方法的多项式模型，构造分类器的过程实际上就是估算先验概率和类条件概率的过程。朴素贝叶斯分类算法对每个类别的特征向量空间进行学习，归纳总结出分类信息，从而构建出文本分类器。

测试：通过构建出的分类器对新文本进行测试。测试的过程就是比较待分类文本在各个类别下出现的条件概率，按各个概率值进行排序，选出最大的概率值，即为待分类文本所属类别。

通过实验测试表明，系统的准确率可达 80%左右。

5.3 本章小结

本章根据对朴素贝叶斯文本分类的研究，使用 Java 在 MyEclipse 平台上设计并实现了一个基于朴素贝叶斯的中文文本分类系统。

第 6 章　结论与展望

6.1 工作总结

随着互联网技术的迅速发展和普及，人们进入了信息极大丰富的时代，面对如此巨大的信息海洋，如何有效地组织和管理，进行自动分类，并快速、准确、全面地从中找到用户所需的信息已成为一个具有重要用途的研究课题。文本分类是在预先定义好的分类体系下，根据文本的特性，将给定文本归到一个或多个类别中，从而对文本进行高效的组织和管理，满足人们快速准确的获取信息的需求。

本文着重研究了基于朴素贝叶斯方法的中文文本分类系统，主要进行了四个方面的工作：

1. 本文讨论了文本分类系统技术，分析了其中的关键技术问题，包括：文本的分类过程，文本的向量空间模型表示，特征项的权重计算方法，文本分类常用算法等问题。其中文本分类常用算法的介绍包括：决策树分类器、k-近邻分类器、支持向量机分类器等。

2. 深入研究了基于朴素贝叶斯分类模型的文本分类技术，包括：朴素贝叶斯分类模型的研究，多变量伯努利模型和多项式模型；在多项式模型的基础上对平滑因子进行了改进，实验表明，该改进取得了较好的分类效果。

3. 深入研究了常用特征选择方法对朴素贝叶斯分类性能的影响。实验表明，对于朴素贝叶斯分类方法，特征选择方法中的信息增益和 χ^2 统计值具有较好的分类效果。

4. 在本文研究的基础上，实现了一个基于朴素贝叶斯方法的中文文本分类系统。

6.2 后续工作

今后的研究工作，主要包括如下几个方面：

1. 由于大多数文本特征项之间存在一定的关联，完全独立是几乎不可能的，然而朴素贝叶斯分类方法是建立在特征项之间条件独立性假设之上。为了进一步提高朴素贝叶斯文本分类的性能，在今后的研究工作中应试图寻找衡量特征相关性的方法，从而最大程度的满足朴素贝叶斯分类方法的独立性假设。

2. 进一步对特征选择方法进行研究，在中文文本分类中，高维特征向量对分类器存

在不利的影响，如何降低特征向量的维数，提高特征选择的质量，仍是今后研究的一个重要问题。

3. 对设计的中文文本分类系统进一步改进，提高分类效率和分类准确率。

参考文献

[1] 范明, 范宏建. 数据挖掘导论. 北京: 人民邮电出版社, 2008.

[2] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2008.

[3] A. Retal. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, 1993(5): 914-925.

[4] N. Fried, D. Geiger, M. Goldszrnidt, et al. Bayesian network classifiers. Machine Learning, 1997, 29(2): 131-163.

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

[6] 陆旭. 文本挖掘中若干关键问题研究. 北京: 中国科学技术大学出版社, 2008.

[7] 肖明. www 科技信息资源自动标引的理论与实践研究. 北京: 中科院文献情报中心, 2004.

[8] M. Maron. Automatic indexing: an experimental inquiry. Journal of the Association for Computing Machinery, 1961, 8(3): 404-417.

[9] G. Salton. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(1): 613-620.

[10] 侯汉清. 分类法的发展趋势简论. 情报科学, 1981(1): 58-63.

[11] 罗远胜. 基于 PLS 的文本分类技术研究. 南昌: 江西师范大学, 2006.

[12] 唐春生, 张磊, 潘东, 等. 文本分类研究进展. http://epcc.sjtu.edu.cn/seminar/Categorization.pdf, 2001.

[13] 奉国和. 自动文本分类技术研究. 情报杂志，2007(12): 108-111.

[14] 尚文倩. 文本分类及其相关技术研究. 北京: 北京交通大学, 2007.

[15] 朱巧明, 李培峰, 吴娴, 等. 中文信息处理技术教程. 北京: 清华大学出版社, 2005

[16] 中文分词. http://baike.baidu.com/view/19109.htm.

[17] 李东. 汉语分词在中文软件中的广泛应用. 北京: 科学出版社, 2003.

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

[19] ICTCLAS 汉语分词系统. http://ictclas.org.

[20] G. Salton. the SMART retrieval system:experiments in automatic document processing. Prentice Hall, 1971: 115-411.

[21] 吴军. Google 黑板报数学之美系列. http://googlechinablog.com.

[22] 武森, 高学东. 数据仓库与数据挖掘. 北京: 冶金工业出版社, 2003.

[23] 张庆国, 张宏伟, 张君玉. 一种基于 k 最近邻的快速文本分类方法. 中国科学院研究生院学报, 2005,22(5): 554-559.

[24] 刘霞, 卢苇. SVM 在文本分类中的应用研究. 计算机教育, 2007(1): 72-74.

[25] 周文霞. 现代文本分类技术研究. 武警学院学报，2007, 23(12): 93-96.

[26] 陈方樱. 概率论与数量统计. 北京: 机械工业出版社, 2006.

[27] 王国才. 朴素贝叶斯分类器的研究与应用. 重庆: 重庆交通大学, 2010.

[28] 刘彧. 基于贝叶斯理论的文本分类技术的研究与实现. 长春: 吉林大学, 2009.

[29] 章舜仲, 王树梅, 黄河燕. 词间相关性在贝叶斯文本分类中的应用研究. 计算机工程与应用, 2009, 45(16):159-161.

[30] 史瑞芳. 贝叶斯文本分类器的研究与改进. 计算机工程与应用, 2009, 45(12): 147-148.

[31] Susana Eyheramendy, David D.Lewis, David Madigan. On the naive bayes model for text categorization. In Proceedings of the Ninth International Worlshop on Artificail Intelligence and Statistics, Key West, Florida, 2003.

[32] David D.Lewis. Naive(Bayes)at Forty: The independence assumption in information retrieval. In Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998: 4-15.

[33] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim. Some effective techniques for naive bayes text classification. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1457-1466.

[34] Andrew McCallum, Kamal Nigam. A comparison of event models for navie bayes text classification. In Proceedings of AAAI-98 Workshop on Learning for Text Categorization, 1998: 509-516.

[35] 董琳, 邱泉, 于晓峰. 数据挖掘实用机器学习技术. 北京: 机械工业出版社, 2006.

[36] wake 中文论坛. http://forum.wekacn.org/index.php.

[37] Yiming Yang, Jan Pedersen. A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning, l997: 412-420.

[38] 肖婷, 唐雁. 文本分类中特征选择方法及应用. 计算机科学, 2008, 34(7): 75-77.

[39] 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. 中文信息学报, 2004, 18(1): 26-32.

[40] 张龙飞. 基于互信息的朴素贝叶斯改进模型研究. 长春: 吉林大学, 2010.

致　谢

首先，衷心感谢我的导师袁方教授。本文的研究是在导师的悉心指导下完成的。从论文的选题、文献搜索、撰写到完成的每个阶段，导师都给予了我精心的指导。导师严谨的治学精神、一丝不苟的工作作风、平易近人的态度、渊博的专业知识以及丰富的实践经验将使我终生受益。在此再次向导师表示深深的谢意！

感谢王煜老师，在我论文遇到困难的时候，王老师的指点给予了我很大的启迪。

感谢在论文的完成过程中，给予我帮助的王亮老师。

感谢我的同学们，感谢你们的帮助和鼓励，让我一次次克服了学习的困难。

感谢我的家人，是他们无私的爱和默默的支持才让我能顺利的完成了学业。

感谢对论文进行评审并提出宝贵意见的各位专家和老师。另外，本文引用了许多专家、学者们的一些研究成果，在此对他们表示深深的谢意！

攻读硕士学位期间发表论文情况

[1] 李丹. 基于 Service Broker 的异步消息通信研究. 现代商贸工业, 2009(19): 271-273.