海南大学

硕士学位论文

文本自动分类的研究与实现

姓名：石敏

申请学位级别：硕士

专业：通信与信息系统

指导教师：康耀红

20060501

摘要

随着 Internet 与 Intranet 的迅猛发展，电子文本的信息量呈指数增长，人们越来越渴望拥有能帮助其查找、过滤以及管理如此海量信息的工具，文本自动分类就是这样的工具之一。利用文本自动分类，文本信息可以自动地被分配到一个或多个已经定义好的类别中，这在很大程度上就解决了信息杂乱的问题，方便了用户快速、准确、全面地查找信息。而且作为信息过滤、信息检索、搜索引擎、数字化图书馆等领域的技术基础，文本自动分类有着广泛的应用前景。

本文对文本自动分类中的几项关键技术，如文本预处理、文本表示模型、特征选择、分类算法等进行了研究。从提高系统分类性能的角度出发，提出了几种有效的解决方法和改进技术。本文的主要研究内容和创新工作包括以下几点：

⑴特征选择方法

特征选择就是在不降低分类性能的前提下，提取能够区分不同类别的特征子集合，从而达到删除冗余特征项，缩减文本特征空间维数，减轻分类器学习负担的目的。目前采用较多的特征选择方法有文档频次、信息增益、χ^2 统计、互信息等，本文从这些方法的基本原理和分类性能入手，着重分析了 χ^2 统计和互信息这两种特征选择方法的优缺点以及它们之间存在的互补性，并在此基础上提出了一种联合的特征选择方法。在中文文本分类实验中，该方法取得了较高的微平均查全率和微平均查准率。

⑵ 基于广义向量空间模型的文本自动分类的研究

目前，很多分类方法都是基于传统向量空间模型和布尔模型的。然而在这两种文本表示模型中，特征项之间都被假设为是相互独立的，因此在该前提下讨论文本分类的问题显然不能令人满意。在广义向量空间模型中，不仅特征项之间相互独立的假设被剔除了，而且在该模型中文本能更加准确的表示出来，因此本文在文本自动分类中引入了广义向量空间模型，并在此基础上提出了基于广义向量空间模型的 KNN 和 TFIDF 文本分类方法。

⑶对广义向量空间模型下布尔交运算的修正

本文在对基于广义向量空间模型的文本自动分类问题进行研究时，发现广义向量空间模型下的布尔运算定律存在着不能满足吸收律、德.摩根法则的缺陷。本文通过对该定律中布尔交运算进行修正弥补了这一缺陷，并从理论的角度证明了改进后的布尔交运算的有效性。

关键词： 文本自动分类 文本表示模型 特征选择 分类器 广义向量空间模型 联

合特征抽取方法

ABSTRACT

With the rapid developments of Internet and Intranet, electronic text information greatly increases. So there is a growing need for tools helping people better find, filter and manage the large mount of information. As one of these tools, text classification can assign free text documents to one or more predefined categories , and it also offers convenience for users to find the required information quickly, fully and exactly . Moreover, text classification has the broad applied future as the technical basis of information filtering, information retrieval, search engine, text database, digital library and so on .

Research on text classification and its related technologies such as text expressed model, feature selection, and classification method are did in the paper. From the angle of improving the text classification performance, several methods and techniques are presented. Our primary works are as follow:

(1) Feature Selection

In order to reduce the number of features and the burden of text classification without compromising the classification performance, parts of the features are extracted . Usually, the text classification system uses some methods to do the feature election such as information gain, χ^2 statistic, mutual information and so on . In this paper, the performance of χ^2 statistic and mutual information is investigated, and the relation between χ^2 statistic and mutual information is pointed. Moreover, a new combined feature selection method based on χ^2 statistic and mutual information is proposed .Our experimental results also indicate that the combined feature selection function can improve the performance of text classification.

(2) Research on text classification based on Generalized Vector Space Model

Currently, there are many text classification methods which are based on Vector Space Model and Boolean Model. Because the two text expressed models have the same shortcoming that features are assumes dependent, the classification results can not be trusted by peoples. In Generalized Vector Space Model , the assumption above is taken off and texts are expressed more exactly ,so the model is used for the text classification in this paper. Meanwhile, two new classification methods named KNN Based On Generalized Vector Space Model and TFIDF Based On Generalized Vector Space Model are proposed .

(3) Modification of AND operation definition under Generalized Vector Space Model.

The shortcoming of boolean operation under Generalized Vector Space model is investigated in the paper. In order to reduce the shortcoming, a new definition of AND operation under the Generalized Vector model is made. Meanwhile, it is proved in theory that the application of new definition can make the traditional Generalized Vector model solve some problem which it owns.

Key words: Text Classification, Text Expressed Model , Feature Selection ,Classification Machine, Generalized Vector Model ,Combined Feature Selection Function.

海南大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：石敏 日期：2006年 6月15日

学位论文版权使用授权说明

本人完全了解海南大学关于收集、保存、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。本人授权海南大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。本人在导师指导下完成的论文成果，知识产权归属海南大学。

保密论文在解密后遵守此规定。

论文作者签名：石敏 导师签名：

日期：2006年 6月15日 日期：2006年6月15日

本人已经认真阅读“CALIS高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS 高校学位论文全文数据库”中全文发布，并可按“章程”中规定享受相关权益。同意论文提交后滞后：□半年；□一年；☑二年发布。

论文作者签名：石敏 导师签名：

日期：2006年 6月15日 日期：2006年 6月15日

1 绪 论

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

1.1 研究背景和研究意义

我们正处在一个信息迅猛增长的时代！1998 年的统计结果表明：全球每年出版的刊物大约有 156000 种，并且这一数字正以每年 12000 种的速度递增。于此同时，互联网已经成为人们获取信息的一个主要渠道，Internet 上也充斥着大量的文本信息。1999 年的统计结果表明，Internet 上约有 3.5 亿个的静态H T M L 页面，并且以每天大约 100 万的惊人速度增长[1]。

面对这庞大纷杂而不断膨胀的信息，我们急需解决的问题是：如何有效的将这些信息组织和管理起来，以便用户快速、准确、全面的找到自己所需的信息。文本分类技术作为处理和组织大量文本数据的关键技术，可以在较大程度上解决这个信息杂乱的问题，方便用户准确地定位和分流信息。

早期的文本分类主要是由人手工完成的，这种做法存在着许多弊端：一是费时费力，工作效率低下；二是分类结果的一致性不高，对于不同的分类人员，其分类结果可能不同。甚至同一个分类人员在不同时间进行分类，其结果可能也不相同；三是难以保证分类结果的准确性，准确率一般都在 40%左右。

文本自动分类弥补了人工分类的不足，它作为一项具有较大实用价值的技术被广泛得应用到智能缓存、信息过滤、信息检索、搜索引擎、文本数据库、数字图书馆、互联网信息监控（如“垃圾”邮件的过滤）等领域。以下是文本自动分类在搜索引擎领域中的应用：

目前的搜索引擎产品很多，如 Yahoo,Google,Baidu 等等，它们所能提供的信息查询方式可以归结为以下两种：分类浏览和关键词检索。分类浏览一般是基于网站分类目录的浏览，它浏览的对象是网站。分类浏览的质量较高，检索效果较好，但信息更新慢，维护的工作量比较大；关键字检索的对象不是网站，而是符合条件的网页，其检索的信息量大，更新及时，但是返回信息量过多，质量比较低。

目前，很少有搜索引擎能够提供对网页的分类浏览，其中一个重要原因就是由人工进行网页的分类几乎是不可能的。因此，如果能够实施对网页的自动分类，搜索引擎就能兼有对网页的分类浏览、检索和关键词检索的功能，而用户也能迅速地判断出返回的网页是否符合自己的检索要求。例如用蝴蝶作为关键词进行检索，在返回的结果中，作为动物的蝴蝶、作为一种手表的蝴蝶和作为一种牌子自行车的蝴蝶等内容是夹杂在一起的，用户要对这些蝴蝶的具体所指进行判断和分析，才能确定出哪些网页是自己所需要

的。如果采用了自动分类技术，就可将返回的网页按照其属于动物、手表还是自行车分到不同的类目中，用户只需根据自己的需求直接到某个类别中去寻找和浏览网页就可以了，从而节省了用户判断时间，提高了检索的效率。

此外，文本自动分类在其它领域的研究也有着重要的意义，它们都已成为信息处理的前沿课题，在得到很多机构重视的同时，也成为目前商用领域的应用焦点。

1.2 问题描述

文本自动分类(Text Categorization)是一个有监督的学习过程。其系统任务就是：通过学习一个已被标注类别（这些类别是事先确定的）的训练文档集合，找出文档和类别之间的关系模型，并利用这种学习得到的关系模型对新文档进行类别判断，将其分配到一个或多个类中。

从数学的角度看，文本自动分类的系统任务可以描述为如下过程：

设训练文档集合为 D＝$\{d_1,d_2,...d_n\}$，类别集合为 C＝$\{c_1,c_2,...c_k\}$，则客观上已经存在一个映射函数 H 使得：

$$H:D\to C \tag{1-1}$$

由于现有的手段很难精确地找到这个函数 H．因此，文本自动分类的一般做法是通过学习一个训练文档集合，找出一个和 H 最近似的函数 f 来代替 H，该函数 f 应最大程度上满足：

$$f:D\to C \tag{1-2}$$

接着利用这个函数 f 将待分文本 d 映射到类别集合 C 的一个或多个类别中．该映射可以是一对一的映射，也可以是一对多的映射，用数学公式表示如下：

$$f:d\to C',\ C'\in C \tag{1-3}$$

1.2.1 文本自动分类的关键技术

一般来讲，构建一个文本自动分类系统需要解决五个关键问题：

[1] 获取训练文本集与测试文本集

训练和测试文本集是一组已被标注和确定类别的文本集合，在整个分类过程中将起到训练分类系统和测试分类结果的作用。由于训练和测试文本集选择地合适于否对整个系统的分类性能影响很大，因此我们应该选择能广泛地代表各个类别的文档来组成训练和测试文档集。一般来讲，训练和测试文档集是一些公认的并经人工分类的语料库。

国外对于自动分类的研究起步比较早，到目前为止已经开发了许多标准、开放的英文分类语料库，具有代表性的有路透社的 Reuters 分类语料库，Usenet 的 20NewsGroups 分类语料库。

就中文语料库而言，目前使用较多的有 TREC 的人民日报分类语料库，复旦大学的中文文本分类语料库等等。

[2] 建立文档表示模型

即采用什么样的语言要素（文档特征）表征文档以及用怎样的数学形式来组织这些特征。文档特征可以是词、短语、句子或段落。但是，随着这些特征所处的语法层次的增高，组合出来的特征将呈指数级增长，分析所付出的代价也就越来越大，所以基于句子和段落的特征在文本分类中比较少见，大多数系统都是以词或词组作为文档特征的。此外，词性、标点符号也被作为文档特征项[2,3,4] 。

分类系统通常采用布尔模型和向量空间模型这两种模型来组织特征、表征文档。基于决策树等分类方法的分类系统一般采用布尔模型来组织文档特征；而基于 KNN 、支持向量机、回归模型的分类系统一般采用向量空间模型来组织文档特征。

[3] 文档特征选择

语料库中往往包含着大量的文本特征项，特别是语料库比较庞大时，如此众多的特征项给分类系统带来了沉重的计算负担。另外，还有一些特征对文档分类的作用并不大。因此应该尽可能选取少量且最能代表文档类别的特征项进行分类，以便减轻文本分类系统的计算负担。

[4] 选择分类方法

也就是说用什么方法建立从文档到文档类别的映射关系，这是文本自动分类的一个核心问题。常采用的方法有 Naïve Bayes 、KNN 、支持向量机、类中心向量、回归模型、决策树等。实际中采取较多的是 KNN 方法和支持向量机方法，这两种分类方法的效果不错，而且具有较强的稳定性。

[5] 性能评估模型

即如何评估系统分类性能的好坏。在文本自动分类中，我们一般采用如下两种指标对系统的分类性能进行评估：

①空间和时间代价。即分类系统在训练和分类两个阶段所付出的时间和空间代价。一般来讲，在训练阶段，高昂的空间和时间待价是可以忍受的。但是，分类阶段则不然。

②查全率和查准率。对于不同的分类问题如单类分类和多类分类，查全率和查准率的计算方法有所不同，本文在第二章对其进行了详细的介绍。

图 1.1 为文本自动分类主要过程：

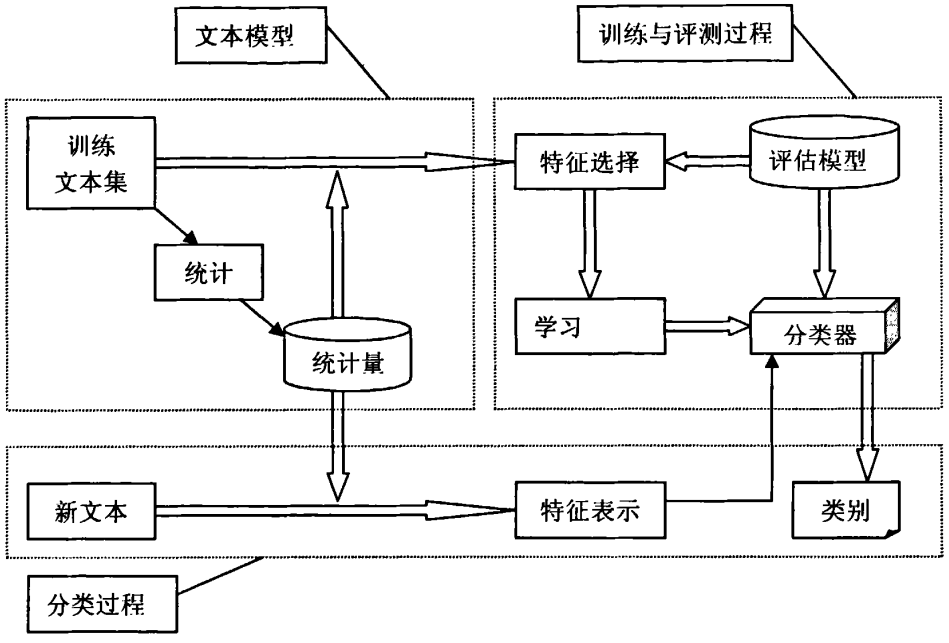

图 1.1　文本分类过程图

从图 1.1 中我们可以看出，文本的自动分类过程可以概括为以下几个步骤：首先将训练文本集向量化，得到一个特征的集合；然后从这个特征集合中抽取一个最优的特征子集，并用这个最优的特征子集将训练文本表示出来；接着利用这些训练文本对分类器进行训练，并根据测试的结果不断的调节相关参数，直至分类器具备最佳的分类效果为止；最后将新文本也用这个特征子集进行表示，并经分类器分类，得到新文本所属的类别。

1.3 研究现状

1.3.1 国内外研究状况

有关文本自动分类的研究始于上个世纪 50 年代末，H.P.Luhn 在这一领域进行了开创性的研究[5]，提出了词频统计思想用于文本自动分类；1960 年，Maron 在《Journal of ACM 》上发表了有关自动分类的论文“On Relevance, Probabilistic indexing and Information Retrieval”，随后许多著名的学者如 G. Salton 等在这领域进行了卓有成效的研究。

直至八十年代末，在自动文本分类领域占据主导地位的一直是基于知识工程的分类方法，即首先由领域专家根据自己的经验或知识归纳出分类规则，然后再根据这些规则建立专家分类系统，并对文本进行分类；九十年代以来，随着信息技术的不断发展，基于机器学习的自动文本分类逐渐取代了基于知识工程的分类方法，成为文本分类的主流技术。基于机器学习的自动文本分类存在着两个基本假设：一是文本类别仅仅是符号标识而已，它实际的含义是什么对分类器的构造没有任何的影响；二是构造分类器所使用

的信息必须是从文档内部抽取出来的，而与文本以外的信息，诸如出版日期、文档类型、出版来源等等是无关的[6]。

近些年，文本自动分类技术取得了很大的进展，主要表现在以下几个方面：分类使用的特征，不再仅仅限于词和短语，诸如词性、标点符号等词法特征以及基于文本语法层次的特征也开始被应用；提出了多种特征选择和分类的方法，如希望交叉熵、文本证据权、支持向量机模型、决策树等；研究开发出了一些相当成功的分类系统，如卡内基集团为路透社开发的 construe 系统；建立了 OHUMED、Reuters 等开放的分类语料库；成立了专门的研究机构，如著名的文本检索会议（TREC）以及主题检测和跟踪会议(TDT)，这些机构通过提供规范的大规模语料库(GB)对文本分类系统的性能进行客观、公正的评测，在很大程度上促进了文本分类技术的交流与发展。

国内对文本自动分类的研究起步相对较晚，1981 年，侯汉清教授首次对计算机在文本分类工作中的应用作了探讨。此后，李晓黎使用概念推理网进行了文本分类的研究[7]，李荣陆提出了一种基于密度的 KNN 分类器训练样本裁剪方法[8]，黄萱菁提出了一种基于机器学习的、独立于语种的文本分类模型[9]。王怡提出了一种基于潜在语义分析的中文文本层次分类方法[10]。该方法的核心思想是利用奇异分解将原始的文档空间转换成一个能反映特征项之间潜在语义的文档空间，并在这个新的空间中逐层地将那些相似的类别进行合并，从而得到一个类间区分度更为明显的类别集合。刁力力用 Boosting 来组合决策树的方法进行文本分类[11]。

1.3.2 中文文本分类研究的现状

汉语与英文、德语、法语等一些语言不同，存在词间无间隙，词性没有明显标记等特性，这给中文文本自动分类带来了一些困难。但是随着中文自然语言理解技术特别是中文自动分词和词性标注技术的日渐成熟，中文文本分类技术在短短 20 年中，已经逐步从可行性探索阶段向实用化阶段转变。直至今天，我国也陆续开发出了一些计算机辅助分类系统和自动分类系统。比较有代表性的有东北大学的图书分类专家系统，上海交大研制的基于神经网络优化算法的中文自动分类系统，清华大学的自动分类系统，中科院的自动分类系统。

1.3.3 当前研究重点

随着技术的不断发展与成熟，近年来文本自动分类的研究主要集中在以下的几个方面：

⑴多语种文本的自动分类，即如何使用一种分类器就可以对不同语种的文本进行分类。

⑵海量文本的快速分类。信息技术的不断发展使得待分文本的数量成指数级增长，因而对文本自动分类的速度和精度提出了更高的要求。如何在不降低分类准确度的前提下，提高对海量文本的分类速度是目前文本自动分类研究的一个重要问题。

⑶新的分类方法。传统的文档分类方法有决策树、贝叶斯方法和 KNN 法等，近些年来出现了一些新的、分类性能更好的方法，如支持向量机分类方法、基于语义的分类方法、人工神经网络分类方法等等。

⑷ Web 文档的分类。相对于纯文本文档的自动分类来说，Web 文档的自动分类不仅能利用文本自身的内容和段落结构信息，还可以充分利用网页中的字体、颜色、超链接、标记等结构信息进行分类，这就为进一步提高分类性能提供了有利的条件。

⑸标准分类语料库的建立。由于文本分类的结果受训练语料库和测试语料库的影响非常大，根据对不同语料库进行分类的结果，很难比较出分类系统性能的好坏，因此需要建立一个标准的分类语料库。

尽管这些年来文本自动分类技术取得了很大的进展，也出现了一些比较成功的系统。但从整体来说，自动文本分类技术在分类的查准率和效率方面仍然不能满足人们的现实需要。这也就是我们要继续研究这一课题的原因所在。

1.4 本文研究的内容

我们一般从稳定性、快速性和准确性三个方面衡量一个文本自动分类系统的性能。其中稳定性是指当文本中存在较多词法和语法错误，或训练文档集发生了较大变化时，分类系统仍能保持稳定的分类精度。快速性是指在文档信息急剧增加时，分类系统仍然能够实时地对海量文本进行分类。准确性是指当类别之间的边界不是很明显时，分类器仍能对处于类边界部分的样本作出比较准确的判断。

本文皆在对影响文本自动分类性能的各种问题进行探讨，并从提高文本自动分类性能的角度出发，对特征选择和分类算法这两大关键技术进行了深入的研究和有效的改进：

1) 深入分析和比较了各种特征选择方法的性能，针对 χ^2 统计和互信息两种特征选择方法存在的不足展开研究，根据这两者之间存在的互补性提出了一种联合的特征选择方法。并用实验证明了该特征选择方法能在一定程度上提高分类系统的微平均查全率、微平均查准率。
2) 深入分析了基于向量空间模型分类算法存在的不足，并针对这一不足之处提出了基于广义向量空间模型的 KNN 和 TFIDF 分类算法.
3) 对广义向量空间模型中的布尔交运算进行了修正，并从理论的角度证明了改进后的交运算能在一定程度上弥补广义向量空间模型下布尔运算定律不满足吸收律、德.摩根法则的缺陷。

1.5 本文结构

本文共分 5 章，文章结构及各章主要内容组织如下：

第 1 章介绍了文本自动分类的基本概念、研究背景及研究意义，分析了国内外文本分类的研究现状、研究重点，介绍了本文主要的研究工作。

第 2 章介绍了文本自动分类的过程和评价方法，对文本自动分类中每一个环节用到的主要技术进行了描述和总结，为后面章节的讨论做了概念和技术上的铺垫。

第 3 章对 χ^2 统计和互信息两种特征选择方法进行了研究。首先指出了它们各自存在的优缺点以及两者之间存在的互补性，接着在此基础上提出了一种联合的特征选择方法，并用实验证明了该方法的有效性。

第 4 章对广义向量空间模型下的自动文本分类进行了研究。首先指出了基于向量空间模型的 KNN 和 TFIDF 分类方法存在的不足，并在此基础上提出了基于广义向量空间模型的 KNN 和 TFIDF 分类方法。然后对广义向量空间模型下的布尔交运算进行了修正，并从理论的角度证明了改进后的交运算的有效性。

第 5 章总结了本文的研究工作，并对今后的研究做出了展望。

2 文本分类技术

本章按照文本自动分类的过程，对文本预处理、文本表示模型、特征选择、分类算法、分类性能评价进行了总结和归纳。

2.1 文本预处理

文本自动分类首先要做的工作就是将通常以字符串表示的文档转化为适合于学习算法以及分类任务的表现形式，这一过程我们称之为文本的预处理。预处理通常包括以下具体的内容：

⑴分词。由于中文文本是按词连写的，词与词之间没有明显的界限，因而在处理中文文本时，首先要解决的问题就是词的切分。中文文本中存在着大量的词组，而词组对短语和句子又有着依赖性，若把一个词组中的词分开来分析则会损坏这个词组原有的意义。因此必须进行词的正确切分，这也是中文文本处理的必要条件。现有的分词方法很多，主要分为 3 大类：基于字符串匹配分词方法、基于理解的分词方法、基于统计的分词方法。

⑵去除停用词。停用词是指诸如介词、冠词等语义内容很少的词，也指那些在文档集的每个文档中都可能出现的高频词。停用词由于出现在很多文档中，所以对区分文档的类别贡献不大。因此，这些词通常在预处理阶段就被从文章词汇列表中过滤掉了。目前，解决去除停用词问题的方法就是把这些停用词组织成一个禁用词表，依照禁用词表过滤掉文章词汇列表中包含的停用词。附录 A 给出了一个由 161 个高频词构成的禁用词表。

⑶词性标注。词性的标注就是给文档中的每一个词选择一个最有可能的词类。自然语言中词存在着大量的兼类现象，举例来说，单词“cook”有两个意思，一个指某种人（厨师，如 a cook is in the kitchen），另一个指的是动作（烹调 ，如 he can cook delicious food）。当 cook 做第一种意思时，它是一个名词，而作为第二种意思时，它是一个动词。因此我们要根据词语所在的上下文对该词的词性进行标注，以便帮助系统更好的理解词语的含义。词性标注在一定程度上排除了由于词的兼类而造成的歧义。

⑷词还原。词还原指的是把一些变形词复原为该词原来的表示形式。主要包括以下一些内容：①名词去复数②动词时态转换③动词第三人称的转换④词根还原，包括去掉词的前后缀。⑤简写词复原。词还原的主要目的是将具有同样概念的词能统一处理，以便减轻分类系统的负担。

⑸权重的计算。对于文档中的特征项常常被赋予一个权重，表示它们在文档中的重要程度，权重越大则该特征项的重要程度就越大。计算特征项的权重一般采用两种方法：第一种是由专家或者用户根据自己的经验和领域知识人为的负上权重。这种方法随意性

很大，而且效率也很低，很难适用于大规模真实文本的处理；另一种方法是运用文本的统计信息（如词频，词之间的同现频率等）来计算特征项的权重。目前被广泛采用的权重计算公式是 TF-IDF 公式：

$$W_{ij} = tf_{ij} * idf_i \tag{2-1}$$

其中 tf_{ik} (Term Frequency)表示项 t_i 在文本 d_j 中的文本内频数，idf_i (Inverse Document Frequency)表示 t_i 的反比文本频数，它有多种计算方法，目前较为常用的公式为：

$$idf_i = \log(\frac{N}{n_k} + 0.01) \tag{2-2}$$

其中 N 表示全部训练集的文本数，n_k 表示训练文本中出现 t_i 的文本频数。在香农信息学理论中，如果项在所有文本中出现频率越高，那么它所包含的信息熵就越少；如果项的出现较为集中，且只在少量文本中有较高的出现频率，那么它就会拥有较高的信息熵。上述公式就是基于这个思想的一种实现。

考虑到文本长度对权值的影响，还应该对相权值公式作归一化处理，将各特征项的权值规范到［0,1］之间：

$$W_{ij} = \frac{tf_{ij} * idf_i}{\sqrt{\sum_{k=1}^{n} tf_{ij}^{\ 2} * idf_i^{\ 2}}} \tag{2-3}$$

从上述 TF-IDF 的计算方法可以看出，我们在统计 tf_{ij} 的值时并没有区别不同位置的文本特征对表达文本内容的不同能力。基于此，陈治刚等人提出了一种改进的权重计算方法[12]。该方法是将文本划分为几个区域，如标题区域、摘要区域、正文区域等，那么出现在标题区域的特征项往往要比出现在摘要区域的特征项更能代表文本的内容，同样出现在摘要区域的特征项也要比出现在正文区域的中的特征项更能代表文本的内容。这样在统计每个区域的特征项频率得到 tf_{ik} 后，再乘以一个反映其重要程度的比例系数来加以修正和调整。则特征项 t_i 在文本 d_j 中出现的频数就被定义为：

$$tf_{ij} = \alpha \times tf_{ij1} + \beta \times tf_{ij2} + \gamma \times tf_{ij3} \tag{2-4}$$

其中，tf_{ijr} 为在第 r 个区域的频率（r 为 1,2,3 时分别对应标题区域，摘要区域，正文区域），$\alpha > \beta > \gamma \geq 1$ 为比例系数。

权重的计算只能视具体情况而定，至今仍没有普遍使用的“最优”公式。另外，在前面的讨论中项的权值一般为正，其实权值也可以取负值，用来描述某用户厌弃某特征。TF-IDF 公式是一种经验公式，并没有坚实的理论基础。但是，多年的实验表明，上述公式是文本处理中的一个有效工具。事实上，这一公式不仅在文本自动分类中得到了成功的应用，它对于其他文本处理领域，如信息分发，信息过滤和信息检索也有很好的借鉴作用。

2.2 文本的表示模型

要实现文本的自动分类，首先需要将文本表示为计算机所能理解的形式。用于文本自动分类的文本表示模型通常有布尔模型和向量空间模型。这两种模型都来源于信息检索领域，只是在文本自动分类中它们更关心文档与类别之间的关系。

2.2.1 向量空间模型

在向量空间模型中，我们假设组成文档的各个特征项之间是相互独立的。因此各个特征项对应的向量$\vec{t}_i$都被看成是两两正交的，则由这些正交的向量可以生成一个欧氏空间。

在该空间内，任一文档d_j都可用一个向量$\vec{d}_j$来表示，称之为文档向量。该向量在各个轴上的分量就是相应的各个特征项在各个文档中的权重，一般采用 TFIDF 方法计算。文本的向量可以d_j表示为：

$$\vec{d}_j = \sum_{i=1}^{n} w_{ij}\vec{t}_i \tag{2-5}$$

其中ω_{ij}为特征项t_i在文档d_j中的权重。

文档之间的相似程度可以借助于向量之间的某种距离来衡量，常用向量之间的内积进行计算：

$$Sim(d_1, d_2) = \sum_{i=1}^{n} w_{i1} * w_{i2} \tag{2-6}$$

或夹角余弦值来表示

$$Sim(d_1, d_2) = \cos\theta = \frac{\sum_{i=1}^{n} w_{i1} * w_{i2}}{\sqrt{\sum_{i=1}^{n} w_{i1}^{\ 2} \sum_{i=1}^{n} w_{i2}^{\ 2}}} \tag{2-7}$$

向量空间模型的优势在于它的简单性，同时功能非常强大。能将非结构化的文档表示成向量的形式，使得各种数学处理成为可能，通过对向量的操作，就能够有效地处理

非常大的文献集合。然而，向量空间模型依然存在着明显的不足之处：该模型是建立在特征项之间相互独立的假设下的，即认为特征向量之间是相互垂直、正交的。

2.2.2 传统布尔模型

传统布尔模型可以看作是向量空间模型的一个特例。在该模型中，根据特征项是否在文档中出现，特征的权值被赋予“1”或“0”，即

$$w_{ij}=\begin{cases}1(\text{如果文献}d_j\text{包含该特征项}t_i)\\0(\text{如果文献}d_j\text{不包含该特征项}t_i)\end{cases} \tag{2-8}$$

采用布尔模型还是向量空间模型进行文本表示，要根据具体的分类方法而定：决策树、Boosting 等分类方法是基于布尔模型的；而 KNN、SVM、LISF 等方法是基于向量空间模型的。

2.3 文档特征选择

2.3.1 特征选择的目的

文本自动分类的一个核心难题就是特征空间的高维性，如一个包含 1000 篇文档的文本训练集对应的特征向量空间就高达几万维。如果直接在这样一个高维的特征空间中进行分类器的训练与分类，很可能带来两个棘手的问题：一是很多在低维空间中具有良好性能的分类算法在高维空间中变得不可行，如贝叶斯、神经网络等；二是过多的特征项使得对样本统计特性的估计变得非常困难，降低了统计分类器的推广或范化能力。

因此很有必要在不影响分类性能的前提下，把特征维数压缩到与训练样本相适应的程度，从而降低分类器训练和分类阶段的工作量，提高统计分类器的识别能力。

特征选择就是选出那些最优的且具有代表性的特征项来构成特征空间，从而达到维数缩减的目的。根据特征选择在实现过程中是分别在单独的类别上还是在所有的类别上完成，可以分为“局部特征选择”和“全局选择”. 所谓“局部特征选择”，就是依据给定的特征评分函数，对每个特定的类别选择一组最优的特征[13][14]项。而“全局特征选择”则是对分类涉及到的所有类别选择一组共同的“最优”特征[15][16] 。

2.3.2 特征选择的步骤

特征选择的具体步骤为：

⑴从训练文档库中提取所有特征项，构成文档特征集合 F；

⑵对集合 F 中的每一个特征项用下列特征评估函数进行打分。当 F 中的所有的特征项都打分完成后，按份值由高到低进行排序；

⑶假设需要选取 N 个特征项，从 F 中选取分值最高的前 N 个特征项，构成最终的分类特征集 F_s。F_s 将用于文档分类的训练和测试。

2.3.3 特征选择的方法

由于特征评估函数的不同，因而产生了不同的特征选择方法：文档频次(DF)、信息增益 (IG)、互信息(MI)、χ^2统计 (CHI)、期望交叉熵(ECE)和几率比(OR)等。其中文档频次、信息增益、互信息和χ^2统计在实际应用中都是非常有效的评估函数。

(1)文档频率(DF)

某个特征项的文档频率一般定义为文档集合里包含该特征项的训练文档的个数。因此，文档频率通常用于全局特征选择。如果将文档频率用于局部特征选择，就将其定义为属于某一类别且包含该特征项的训练文档个数。

采用文档频率进行特征选择是基于如下假设的：文档频次太低或太高的特征项含有较少的鉴别信息，对分类的贡献不是很大。因此在用此方法进行特征选择时一般都会设置两个阈值，将大于高阈值和小于低阈值的特征项从原始空间中移除。

文档频率是一种简单而行之有效的特征选择方法。但是由于缺乏必要的理论基础，文档频率一直被认为是一种用来改善文本分类器效率的权宜之计，而不能算是严格意义上的特征选择方法。

(2) 信息增益 (IG)

信息增益在机器学习中经常被用作特征项评判的标准。它是一种基于熵的评估方法，被定义为特征项在文档中出现前后的信息熵之差。计算公式如下：

$$\begin{aligned} IG(t) = &-\sum_{i=1}^{m} p(c_i)\log_2(p(c_i)) + p(t)\sum_{i=1}^{m} p(c_i \mid t)\log_2(p(c_i \mid t)) \\ &+ p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log_2(p(c_i \mid \bar{t})) \end{aligned} \tag{2-9}$$

其中c_i表示类别，t为特征项。$p(c_i)$为类别c_i出现的概率，$p(t)$为特征项t出现的概率，$p(\bar{t})$表示特征项t不出现的概率，$p(c_i \mid t)$表示已知特征项t出现的前提下属于类别c_i的条件概率，$p(c_i \mid \bar{t})$表示已知特征项t不出现的前提下属于类别c_i的条件概率。

根据各个特征项的信息增益值，从大到小依次选择信息增益较大的特征项构成特征项的子集。

(3) 互信息 (MI)

互信息是用于表征两个变量之间相关性的一种方法。在分类中它用于度量特征项和类别之间的相依关系，互信息越大，特征项和类别之间的关联程度也就越大。计算公式如下：

$$MI(t,c) = \log\frac{p(t,c)}{p(c) * p(t)} = \log\frac{p(t \mid c)}{p(t)} \tag{2-10}$$

其中 $p(t|c)$ 为在类别 c 出现的前提下特征 t 出现的概率，$p(t,c)$ 为特征 t 和类别 c 共现的概率，$p(t)$ 的定义和信息增益的相同。在实际计算中我们采用它的近似公式：

$$MI(t,c)=\log\frac{A\times N}{(A+E)\times(A+B)} \tag{2-11}$$

其中 N 为训练集中所有的文本数，A 为包含特征 t 且属于类别 c 的文档数，B 为包含特征 t 且不属于类别 c 的文档数，E 为属于类别 c 而不包含特征 t 的文档数。如果定义 U 为不包含特征 t 且不属于类别 c 的文档数，则有 $N=A+B+E+U$。

为了更好的评价一个特征项的好坏，我们结合特征项在不同类中的得分对特征进行评价，并使用以下两种交替的方法：

$$MI_{avg}(t)=\sum_{i=1}^{m}p(c_i)MI(t,c_i) \tag{2-12}$$

$$MI_{\max}(t)=\max_{i=1}^{m}\{MI(t,c_i)\} \tag{2-13}$$

从互信息的定义公式(2-10)来看，如果某特征项和某一类别是相互独立的，那么 $p(t,c)=p(c)*p(t)$，即 $MI(t,c)=0$，也就是说该特征项的出现对预测这一类别没有什么贡献。从公式(2-10)还可以看出，在条件概率 $p(t|c)$ 相等情况下，低频词比常用词更易被选出。

(4) χ^2 统计(CHI)

χ^2 统计(CHI)度量的是特征 t 和类别 c 之间的非独立程度，并认为两者之间的非独立关系类似于一维自由度的 χ^2 分布。计算公式如下：

$$\chi^2(t,c)=\frac{r[p(t,c)p(\bar{t},\bar{c})-p(t,\bar{c})p(\bar{t},c)]^2}{p(t)p(\bar{t})p(c)p(\bar{c})} \tag{2-14}$$

其中 $p(\bar{t},\bar{c})$ 为 c 不出现且不包含特征项 t 的概率，$p(\bar{c})$ 为类别 c 不出现的概率，$p(t,\bar{c})$ 为包含特征项 t 且 c 不出现的概率。其它概率的定义与信息增益和互信息相同。r 为训练集中的文本数，只起到微调的作用[17]。在实际应用中我们常采用它的近似公式：

$$\chi^2(t,c)=\frac{N\times(AU-EB)^2}{(A+E)\times(B+U)\times(A+B)\times(E+U)} \tag{2-15}$$

其中 N、A、B、E、U 的定义与上述互信息中的定义相同。

直观的看，χ^2统计值越大，特征项与类别之间的独立性就越小，对分类的贡献就越大。我们结合特征项在不同类中χ^2统计的得分对特征进行综合评价，并使用以下两种交替的方法：

$$\chi^2_{avg}(t)=\sum_{i=1}^{m}p(c_i)\chi^2(t,c_i) \tag{2-16}$$

$$\chi^2_{\max}(t)=\max_{i=1}^{m}\{\chi^2(t,c_i)\} \tag{2-17}$$

(5)特征词强度（TS）

该方法是通过计算特征项在相关文档中出现的概率来评估特征项对分类的重要性。对一个相关文档对(x,y)而言，特征项的强度定义为：

$$s(t)=P_r(t\in y\mid t\in x) \tag{2-18}$$

该特征选择方法是基于这样的一个思想：如果一个特征项在一些相关的文档中多次出现，那么这个特征项就能很好得表征这些相关文档所属的类别。为了寻找这些相关文档，得对文档集进行聚类分析

(6)希望交叉熵

希望交叉熵和信息增益相似，不同之处在于信息增益中同时考虑了特征在文本中发生与不发生两种情况，而希望交叉熵只考虑特征在文本中发生的情况，对于特征t，其希望交叉熵记为$CE(t)$，计算公式如下：

$$CE(t)=p(t)\sum_{i=1}^{m}p(c_i\mid t)\log\frac{p(c_i\mid t)}{p(c_i)} \tag{2-19}$$

在只考虑单个类的时候，则有：

$$CE(t)=p(c\mid t)\log\frac{p(c\mid t)}{p(c)} \tag{2-20}$$

(7)文本证据权

文本证据权值反映的是类概率与在给定某一特征值下的类概率差。对于特征t，文本证据权值记为$wet(t)$，计算公式如下：

$$wet(t)=p(t)\sum_{i=1}^{m}p(c_i)\left|\frac{p(c_i\mid t)(1-p(c_i))}{p(c_i)(1-p(c_i\mid t))}\right| \tag{2-21}$$

在只考虑单个类的时候，则有：

$$wet(t) = p(t)p(c)\left|\frac{p(c|t)(1-p(c))}{p(c)(1-p(c|t))}\right| \tag{2-22}$$

(8)优势率

优势率原本用于二元分类器，定义如下：

$$OR(w) = \log\frac{p(w|c_{pos})(1-p(w|c_{neg}))}{(1-p(w|c_{pos}))p(w|c_{neg}))} \tag{2-23}$$

其中：c_{pos} 表示正例集的情况，c_{neg} 表示负例集的情况。

2.3.4 几种特征选择方法的比较

1997 年，Yiming Yang 和 Jan O.Pedersen 两位学者对 DF,IG,TS,MI,CHI 这五种特征选择方法进行了比较试验。实验采用 KNN 和 LLSF 两种分类器对 Reuters22173 中的 13272 篇文档进行多类分类（其中训练样本为 9610 篇和测试样本为 3662 篇）。图 2.1、图 2.2 依次为在 KNN 和 LLSF 中五种特征选择方法的性能曲线图，其中，横坐标为特征项的数量，纵坐标为 11 点插值平均查全率：

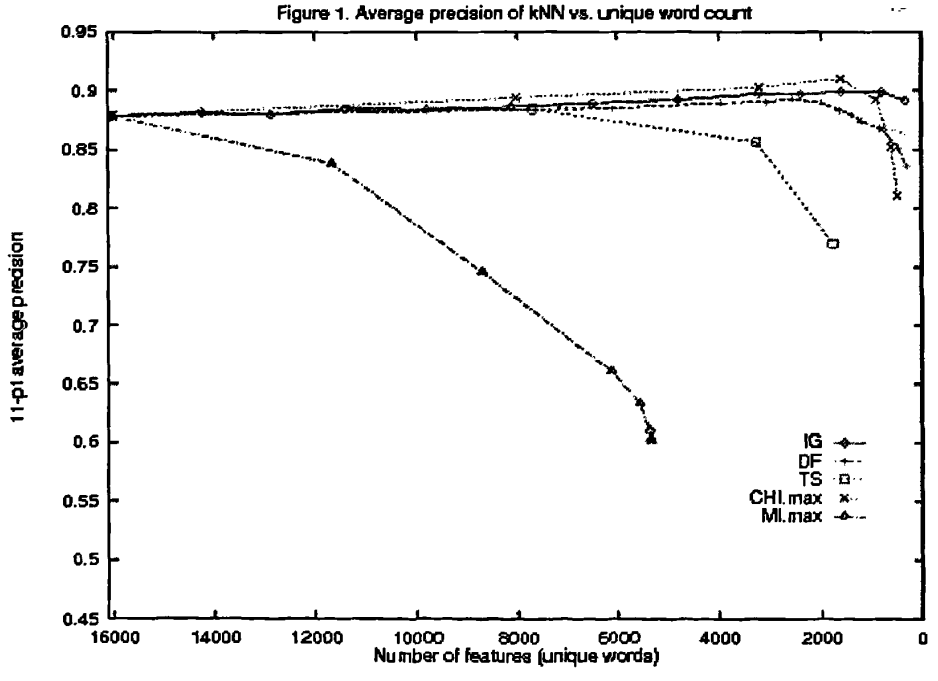

图 2.1　KNN 分类器上的分类性能

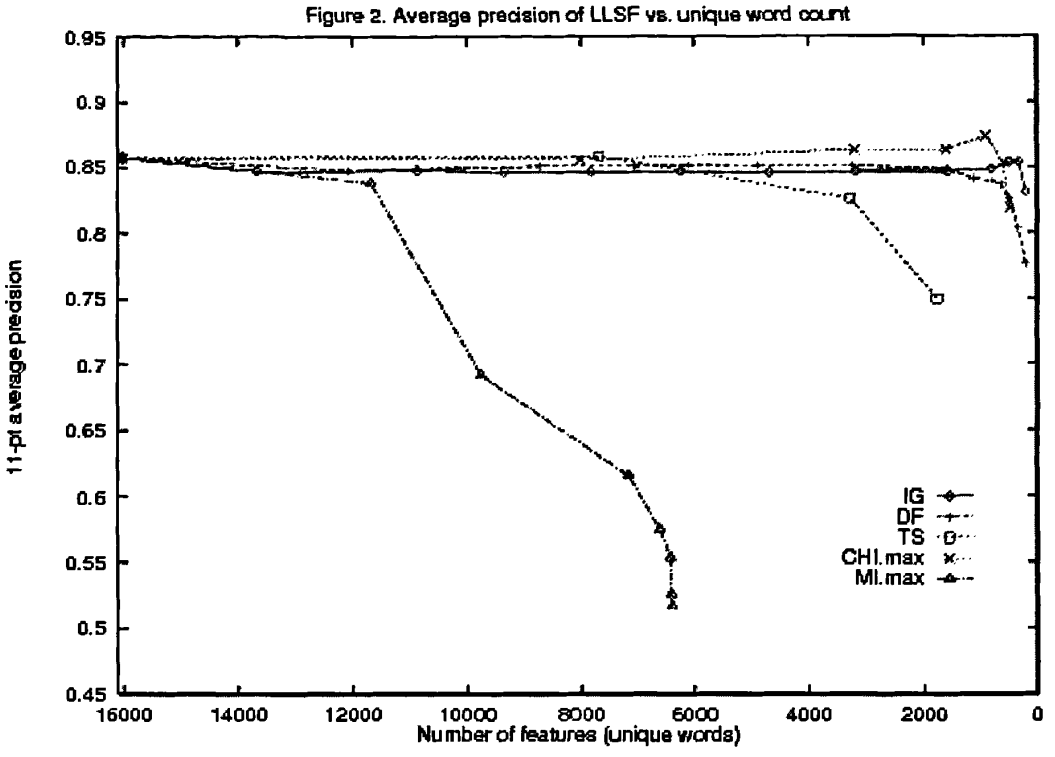

图 2.2　LLSF 分类器上的分类性能

由图 2.1、2.2 我们可以看出 IG，DF 和 CHI 这三种特征选择方法的性能都比较好，它们都能在保持良好分类性能的前提下去掉 90%或更多的特征项。例如 IG 方法可以将特征项的数量从 16039 减少至 321，而 11 点插值平均查全率从 87.9%提升到了 89.2%；但当特征项的数量降至一定程度时（本实验是 2000 维左右），三种分类方法的性能都开始下降，其中 CHI 和 DF 的性能下降得比较快，而 IG 下降得相对较慢一些。

由图 2.1、2.2 我们还可以看出 MI 的分类性能比较差，一般认为是由低频词得分过高造成的。但这一推测一直没有得到证实，Yang 在该实验[18]中对这一点验证了；TS 方法的分类性能介于 MI 和 IG、DF、CHI 之间，在特征数降到 11000 维左右以后，分类性能急剧下，大约能去掉 50%～60%的特征项。

此外，由图 2.1、2.2 我们还可以观察到 IG、DF、CHI 三者的性能曲线图非常相似性，这引起了 Yang 的高度关注，于是他对 Reuters22173 语料库中特征项的 IG、CHI、DF 值进行了比较，结果如下图 2.3、图 2.4 所示：

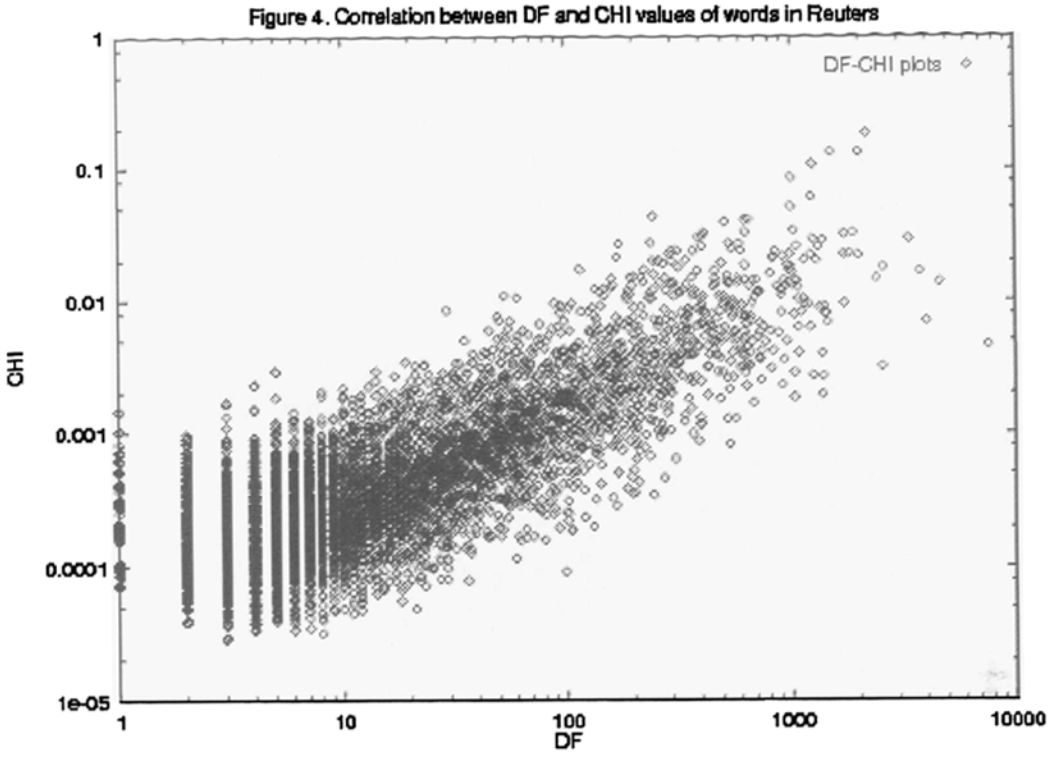

图 2.3　DF 和 CHI 的关系图

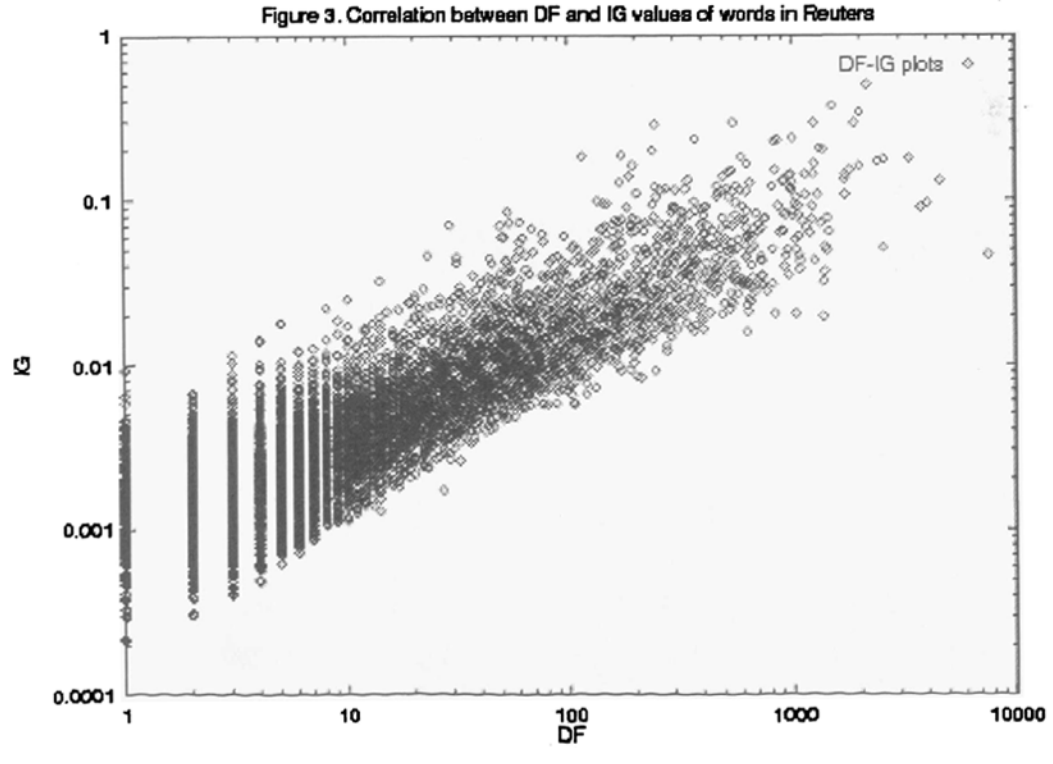

图 2.4　DF 和 IG 的关系图

由图 2.3、图 2.4 我们可以看出样本点大多集中在对角线的位置；如果将图 2.4 中 CHI 的值乘以 10000，则特征项的 CHI 值和该特征项的 DF 值将非常接近。如果将图 2.4 中 IG 的值乘以 1000，则特征项的 IG 值和该特征项的 DF 值也非常接近。因此，IG 和 DF 以及 CHI 和 DF 之间存在着高度的相关性。由于文档频率方法实现简单，时间花费少，因此在信息增益和 χ^2 统计计算开销过大的时候，可以考虑用文档频率代替。

2.3.5 文档特征选择中的概率值估算

$$p(c) \approx \frac{N_c}{N} \tag{2-24}$$

$$p(t \wedge c) \approx \frac{N_{tc}}{N} \tag{2-25}$$

$$p(t) \approx \frac{N_t}{N} \tag{2-26}$$

这里 N_c 为类 c 中包含的文档数，N_{tc} 为类 c 中特征 t 出现的频数，且

$$N_t = \sum_{c \in C} N_{tc} \tag{2-27}$$

又有：

$$p(c \wedge \bar{t}) = p(c) - p(t \wedge c) \tag{2-28}$$

$$p(\bar{t}) = 1 - p(t) \tag{2-29}$$

$$p(\bar{c}) = 1 - p(c) \tag{2-30}$$

$$p(\bar{c} \wedge t) = p(t) - p(t \wedge c) \tag{2-31}$$

$$p(\bar{c} \wedge \bar{t}) = p(\bar{t}) - p(c \wedge \bar{t}) \tag{2-32}$$

2.4 分类算法

自动文本分类的方法大部分来自于模式分类，基本上可以分为三大类：基于统计的方法，基于连接的方法和基于规则的方法[19]。

基于统计的方法是一种基于概率的分类方法，因此其必然掩盖了小概率事件的发生。它的优势在于各类别模板都是通过对大规模语料库进行训练和分析得到的，因而对语言的处理提供了比较客观的数据依据和可靠的质量保证。它的不足在于：①当类别之间交叉现象比较严重时（两类之间的特征重复较多），分类的精度会大大降低。②对训

练语料库的数量和质量均有较严的要求，如果语料库不全面，代表性不强或类中夹杂着不属于该类的文档都会影响分类的效果。常用的基于统计的方法有朴素贝叶斯法、k 邻近法、类中心向量、回归模型、支持向量机等。

基于连接的方法，即人工神经网络，它是一个通过模拟人脑神经系统的基本组织特性而构成的新型信息处理系统，该系统期望其本身能像大脑一样运作、学习、产生智慧。人工神经网络能根据输出结果和实际结果之间的误差自动的调节系统，具有很好的自适应性、较强的容错性以及运算全局并行、处理的非线性等特点。但是使用它学习所形成的知识结构是人所难以理解的，系统本身对于人来说就像是一个变魔术的黑盒子，根据输入给出输出，答案正确但不知道是怎么算出来的。

基于规则的方法是一种唯物主义方法，本质上是一种确定性的演绎推理方法，优点在于其能根据上下文对确定性事件进行定性地描述。它根据大量的知识总结和归纳出一些分类的规则，这些规则具有可读性强、容易理解的优点。不过，在对不确定性事件的描述中，规则之间的想容性等方面存在一些缺陷和限制。常用的基于规则的方法有决策树等等。

在众多分类算法中，我们重点讨论以下几种算法：

2.4.1基于统计的分类方法

⑴朴素贝叶斯算法

贝叶斯算法是一种简单而非常有效的分类算法，它基于这样的一个假设：在给定的文档类别情况下，特征项是相互独立的。假设 d_i 为一任意文档，它属于文档类 $C=\{c_1,...,c_k\}$ 中的某一类 c_j。根据贝叶斯分类方法有：

$$p(c_j \mid d_i)=\frac{p(c_j)p(d_i \mid c_j)}{p(d_i)} \tag{2-33}$$

$$p(d_i)=\sum_{j=1}^{k} p(c_j)p(d_i \mid c_j) \tag{2-34}$$

对文档 d_i 进行分类，就是按照公式(2-33)计算所有文档类在给定 d_i 情况下的概率，概率值越大文档 d_i 属于该类的可能性越大。

由公式(2-33)可知，在给定分类背景和训练文档的情况下，用贝叶斯方法分类的关键就是计算 $p(c_j)$ 和 $p(d_i \mid c_j)$。计算 $p(c_j)$ 和 $p(d_i \mid c_j)$ 的过程就是建立分类模型的过程。

根据 $p(d_i \mid c_j)$ 计算方法的不同，可以将贝叶斯方法分为最大似然模型、多项式模型、泊松模型等。

(2)支持向量机

支持向量机方法由Vapnik于1995年提出的，用于解决二分类模式识别问题。它试图在空间中找到一个决策面，这个面能“最好”地将两个类别中的数据进行分割。何为“最好”，我们以下进行介绍：

设给定的训练集为$(x_1,y_1),(x_2,y_2),...,(x_l,y_l)$，$x\in R^d, y\in\{+1,-1\}$，其中$y_i$是对$x_i$的分类（+1表示它是正例，-1为反例）

且可被一个决策面$(w^T x)+b=0$线性得分割为两类。

将该决策面进行左右平移（这种平移不会造成数据的分割错误）并使之分别通过两类中与此决策面最近的样本点，我们就可以得到两个与之平行的平面。这两个平行的平面之间的距离我们称之为分类间隔(margin)。如果一个训练集中的样本能被一个决策面无错误的进行线性分割，且能使该决策面对应的分类间隔最大，我们就称该决策面为“最好”的决策面，即图中的H

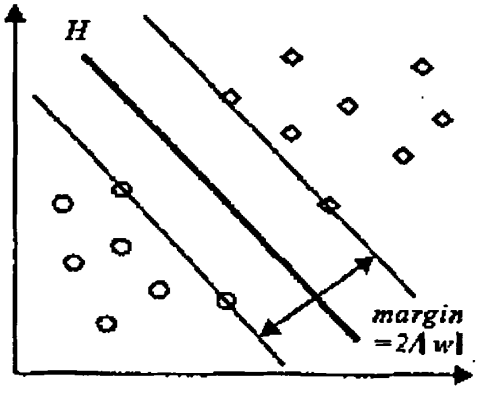

图2.5 支持向量机分类图

因此，使用支持向量机进行分类的关键就是寻找使得分类间隔最小且满足以下条件的$\bar{t}$和b：

$$\begin{cases} w^T\cdot x_i+b\geq 1 & \text{当 } y_i=+1 \text{ 时} \\ w\cdot x_i+b\leq -1 & \text{当 } y_i=-1 \text{ 时} \end{cases} \tag{2-35}$$

在空间R^d中样本点$x=(x_1,...,x_d)^T$到决策面的距离d可由$d=|w^T\cdot x+b|/\|w\|$来计算，其中$\|w\|=\sqrt{w^T\cdot w}$。当存在x使得$w^T\cdot x_i+b=\pm 1$时，则图2.5中的分类间隔$margin=2/\|w\|$，由此寻找满足上述条件的w和b的问题就转化为求如下二次规划问题：

$$\min \phi(w) = \min(\frac{1|}{2} \| w \|^2) \tag{2-36}$$

且满足约束条件：

$$y_i(w^T \cdot x_i + b) \geq 1 \quad i = 1,....,l \tag{2-37}$$

采用Lagrange乘子将其转换为一个对偶问题，形式如下：

$$\max W(a) = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j a_i a_j (x_i, x_j) \tag{2-38}$$

且满足约束条件：

$$0 \leq a_i, i = 1,...,l \text{ 和 } \sum_{i=1}^{l} a_i y_i = 0 \tag{2-39}$$

其中 a_i 为每一个样本对应的Lagrange乘子，根据Kuhn_Tucker条件，这个优化的解必须满足：$a_i(y_i[w^T \cdot x_i + b] - 1) = 0, i = 1,..l$，因此多数样本对应的 a_i 将为0，少部分不为0的 a_i 对应的样本就是支持向量。最后得到分类判别函数为：

$$g(x) = \text{sgn}(f(x)) = \text{sgn}(\sum_{\text{支持向量}i} y_i a_i * (x_i, x_j) + b^*) \tag{2-40}$$

b^* 是分类阈值，可以通过两类中任意一对支持向量去中值求得。

根据上述易知，对于空间 R^d 中样本 $x = (x_1,...,x_d)^T$ 当 $|f(x)| < 1$ 时，表示此时 x 在超平面的分类间隔内，$|f(x)|$ 越趋于 0，则当前分类超平面对于 x 的区分能力越差。而 $|f(x)| \geq 1$ 时 x 能被超平面正确的分类。

支持向量机方法的一个有趣特性是决策面只是由那些刚好和决策面距离为 $1/\|w\|$ 的数据点来决定的，这些点称之为支持向量，它们是训练集合中仅有的对分类有效的样本点，删除其他的数据点不会影响该算法的分类结果。因此，该方法是一种将降维和分类结合在一起的分类方法。

对于线性不可分的问题，可以通过引入松弛因子来推广最优分类超平面的概念，更一般的方法是用满足Mercer条件的核函数 $K(x_i, x_j)$ 代替上式(2-38) 和 (2-40)中的内积，就是通过一个非线性映射，将输入空间转换到一个高维特征空间，然后在这个高维特征空间中给出一个最优分类超平面。

(3)回归模型

近年来，各种统计回归模型在自动文本分类领域中得到了广泛运用，取得了良好的

分类效果，回归模型中最为典型的就是 LLSF 模型[20]。给定训练文档集和文档类集，LLSF 将其表示为两个矩阵 A 和 B。A 代表原始空间，矩阵的第 i 行、第 j 列的元素代表第 j 个特征项在第 i 个文档中的权值。B 代表目标空间，矩阵的每一个元素只能取 0 或 1。如果文档 T_i 属于类别 C_j，那么矩阵的第 i 行、第 j 列的元素取值为 1，否则取值为 0。例如：给定如下 4 篇文档，4 个类别。

Doc1	Neuropathy and Guillain Barre Syndrme	⟷	Category set 1={C_3 ,C_4}
Doc2	Neuropathy and Guillain Barre Syndrme	⟷	Category set 2={ C_2,C_3 ,C_4}
Doc3	AIDS and Guillain Barre Syndrme	⟷	Category set 3={C_1,C_4}
Doc4	AIDS and Neuropathy	⟷	Category set 4={C_3,C_1}

统计文档中所有词出现频率，使用 IDF 计算词的权重值，结果如下所示：

Word	Weight(IDF)
T1. AIDS	5.0
T2. and	1.0
T3. barre	8.1
T4. guillain	4.9
T5. Neuropathy	4.2
T6. Syndrme	3.1

则我们可以得到词关系矩阵 A：

$$\begin{array}{c} \\ Doc1 \\ Doc2 \\ Doc3 \\ Doc4 \end{array} \begin{array}{c} \begin{array}{cccccc} T_1 & T_2 & T_3 & T_4 & T_5 & T_6 \end{array} \\ \begin{bmatrix} 0.0 & 1.0 & 8.1 & 4.9 & 4.2 & 3.1 \\ 0.0 & 1.0 & 8.1 & 4.9 & 4.2 & 3.1 \\ 5.0 & 1.0 & 8.1 & 4.9 & 0.0 & 3.1 \\ 5.0 & 1.0 & 0.0 & 0.0 & 4.2 & 0.0 \end{bmatrix} \end{array}$$

统计每一篇训练文档的所属类别，我们可以得到目标空间矩阵 B 为：

$$\begin{array}{c} \\ Category\_set\_1 \\ Category\_set\_2 \\ Category\_set\_3 \\ Category\_set\_4 \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_1 & c_1 & c_1 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

这样文本分类问题就转换为求一个满足条件 $(B_{m\times l})^T = F_{l\times n}.(A_{m\times n})^T$ 的矩阵 $F_{l\times n}$ 的问

题，其中 $F_{l\times n}$ 是一个特征项与类别的关联矩阵，行代表类别，列代表特征项。LLSF 就是寻找矩阵 $F_{l\times n}$，使得下式的值最小：

$$\sum_{i=1}^{m}\left\|\vec{e}_i\right\|^2=\left\|FA^T-B^T\right\|^2 \tag{2-41}$$

解决 LLSF 的方法之一是将矩阵 A 进行奇异分解，F 最后表示为：

$$F=B^T(A^{-1})^T=B^TUS^{-1}V^T \tag{2-42}$$

对于上例中的矩阵 A 和 B，矩阵 F 的计算结果为：

$$\begin{array}{c} \\ c1 \\ c2 \\ c3 \\ c4 \end{array}\begin{array}{c} \begin{array}{cccccc} T_1 & T_2 & T_3 & T_4 & T_5 & T_6 \end{array} \\ \begin{bmatrix} 0.198 & 0.20 & -0.001 & -0.001 & -0.002 & 0. \\ -0.050 & 0.003 & 0.20 & 0.012 & 0.059 & 0.008 \\ -0.003 & 0.028 & -0.001 & -0.001 & 0.234 & 0.000 \\ 0.001 & 0.005 & 0.082 & 0.049 & -0.001 & 0.032 \end{bmatrix} \end{array}$$

给定一篇文档 d，可以通过计算 $y=(Fd^T)^T$，将其映射到目标空间。向量 $\vec{y}=\{y_1,....,y_l\}$ 中每一个 y_i 的取值在-1～1 之间，表示文档 d 与每一类的相关度。

实验结果表明，线性最小二乘拟合的文本分类效果并不十分理想[21]，Tong Zhang 通过引入正则化参数对其进行了有效的改进[22].

(4)TFIDF 算法

设每篇文档对应的向量可表示为 $\vec{d}=(w_{i1},w_{i2},...,w_{in})$，其中 w_{ik} 为特征项 t_k 按 TFIDF 方法计算得出的权重。则在 TFIDF 文本分类算法中，我们将所有属于类别 c_j（$c_j\in C$）的文档对应的向量进行相加，便可得到类别的向量表达式：

$$\vec{c}_j=\sum_{d\in c_j}\vec{d}_i \tag{2-43}$$

分类时只须计算新文档向量与每个类向量之间的相似度，并根据相似度的大小将新文档分到值最大的类中：

$$class(d_i)=\arg\max sim(\vec{d}_i,\vec{c}_j)=\arg\max\frac{\vec{d}_i\cdot\vec{c}_j}{\|\vec{d}_i\|\cdot\|\vec{c}_j\|}\quad c_j\in C \tag{2-44}$$

TFIDF 算法是一种有监督的学习分类算法，它的训练集是已经标注的文档，它对训练集规模很敏感，随着训练集规模的增大，分类精度显著提高。而聚类算法在这点上有着特殊的优势，它可以在未标识的数据训练集上学习，体现了机器学习自动、快速的特点，其缺点是分类精度较低，初始化过程需要定义大量的经验参数。因此，能否将 TFIDF 和聚类算法的优点结合起来也是值得我们研究的一个课题。

(5)KNN 算法

KNN 算法是由 Cover 和 Hart 提出的一种基于统计的懒惰学习算法。它的主要思想如下：对于一篇待识别的文档，系统在训练集中找到与之最相似的 K 个训练文档。在此基础上，给这 K 个文档打分，分值为 K 个文档中的文档与待识别文档之间的相似度. 然后将这 K 个文档中属于同一类别的文档的分值相加，得到每个类别的得分，分值最高的类别就是该待识别文档应属于的类别。此外，还应当选定一个阈值，只有分值超过这个阈值的类才予以考虑。形式化表示为：

$$scode(d,c_i)=\sum_{d_j\in Knn} Sim(d,d_j)y(d_j,c_i)-b \tag{2-45}$$

其中

$$y(d_j,c_i)=\begin{cases}1 d_j\in c_i\\ 0 d_j\in c_i\end{cases} \tag{2-46}$$

b 为阈值，是一个待优化的值. 一般根据实验结果进行调整. $Sim(d,d_j)$ 为文档 d 和 d_j 的相似度， $scode(d,c_i)$ 为文档 d 属于 c_i 类的分值.

KNN 算法原理简单，易于实现。但在文本自动分类中，该算法的缺陷也突出的暴露了出来，主要有：首先 KNN 算法是懒散分类算法，对于分类所需的计算都推迟到分类时才进行。在其分类中存贮有大量的样本空间向量，在未知类别样本需要分类时，再计算所有存贮样本和未知类别样本之间的相似度，对于高维文本向量或样本集规模较大的情况其时间和空间复杂度比较高；其次，该算法受 K 值的影响比较严重，K 值选择过小，得到的近邻数过小，会降低分类精度，同时也会放大噪声数据的干扰。而如果 K 值选择过大，并且待分类文本属于训练集中包含文档较少的类，则在这 K 个近邻样本中往往包含了实际上并不相似的文本，从而引起了噪声的增加导致了分类效果的降低。

2.4.2 人工神经网络

人工神经网络是通过模仿人类大脑来处理信息的，它是一个并行的分布式信息处理结构。它通过称为连接的单向信号通道将一些处理单元（神经元）互连而组成，每个处理单元都有一个单一的输出连接，这个输出连接可以根据需要被分支成希望个数的许多并行连接。且这些并行连接都输出相同的信号，信号的大小也不因分支的多少而变化。人工神经网络具有信息分布存放、运算全局并行、处理的非线性等特点，适用于学习一个复杂的非线性映射。根据网络结构和学习算法的不同，人工神经网络分为前馈网络、多层感知器和 BP 网络等。

下面我们以 BP 网络为例来说明人工神经网络在文本分类中的应用。BP 网络由三个基本层构成，即输入层、输出层和隐层，每个层都包含若干个处理单元。输入层的处理单元个数通常为输入向量的维数，输出单元层的个数为输出向量的维数，隐层处理单元

的个数可以认为与问题相关，目前的研究结果还难以给出其与问题的类型和规模之间的函数关系。输入层和隐层之间、隐层和输出层之间的每个连接都有一个可以调节的权，权值的大小是在神经网络的训练阶段得到的。研究结果表明，增加隐层的层数不一定能够提高网络的精度和表达能力，一般情况下，选用一个隐层就足够了。当给定一段文本及其特征集时，输入层神经元的个数就设定为特征集的大小，输出层神经元的个数就设定为类别集的大小，输入层的第i个分量的输入值为：

$$t_i = \begin{cases} 1, & \text{文本中存在特征集中的第}i\text{个特征项} \\ 0, & \text{文本中不存在特征集中的第}i\text{个特征项} \end{cases} \tag{2-47}$$

输出层的第j个分量的输出值为：

$$c_j = \begin{cases} 1, & \text{文本属于类别集中的第}j\text{个类} \\ 0, & \text{文本不属于类别集中的第}j\text{个类} \end{cases} \tag{2-48}$$

用于文本分类的BP网络的结构图如下所示：

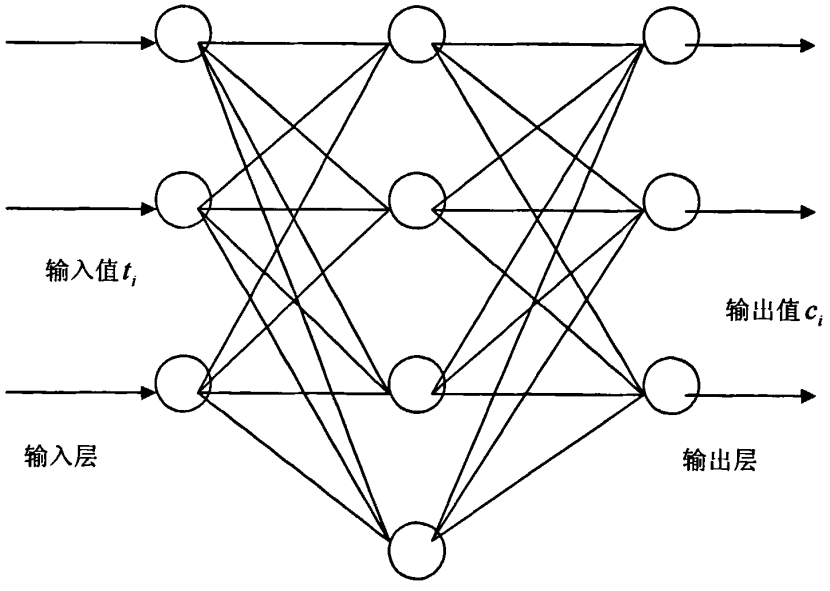

图2.6　神经网络分类图

使用BP网络进行文本分类主要包括网络训练阶段（或称学习阶段）和网络工作阶段。网络训练阶段由正向传播和反向传播组成，在正向传播阶段，输入值经过非线性变化从输入层经隐单元逐层处理，并传向输出层，每一层神经元的状态将影响到下一层神经元的状态．如果输出层的结果不是预计的结果，则认为是由连接权的错误造成的。因此在反向传播阶段，输出结果与预计结果之间的误差将被分摊给各神经元的连接权，并利用相对误差不断地修正各连接权，直到输出结果与预计结果之间的误差减至最小为

止。

使用 BP 算法进行训练，当网络稳定下来后，各神经对应的权值就作为文本分类时的知识，利用它完成在网络工作阶段对文本进行分类的任务。

2.4.3 基于规则的分类方法

基于规则的分类方法很多，这里我们主要介绍决策树方法。决策树分类方法是一种多级分类方法，利用树把一个复杂的多类别分类问题转化为若干个简单的分类问题。它不是企图用一种算法、一个决策规则把多个类别一次分开，而是采用分级的形式，使分类问题逐步得到解决。另外，决策树可以很方便地转化为分类规则，是一种非常直观的分类模式表示形式。

决策树一般是根据训练文本集构造的一个类似于二叉树或多叉树的树结构. 树中的每个分枝都对应于特征集中的一个特征项对训练集或训练子集的一次划分，每个非叶节点(包括根节点) 都对应于一次划分的结果，每个叶节点都对应于一个类或类分布. 从根节点到叶节点的一条路径就是一条分类规则. 图 2.7 给出了一个决策树分类器.

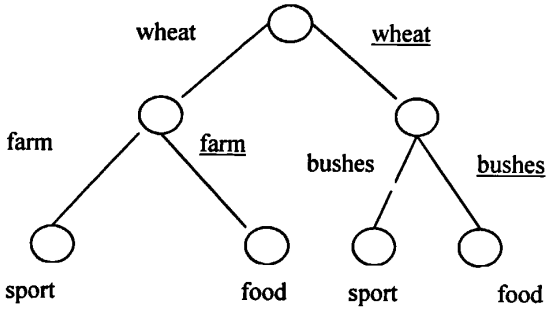

图 2.7　决策树分类图

则由这个分类器共能形成 4 条分类规则：

规则 1　如果某文档包含特征 wheat 而包含特征 farm，则该文档属于 sport 类.

规则 2　如果某文档包含特征 wheat 且不包含特征 farm，则该文档属于 food 类.

规则 3　如果某文档不包含特征 wheat 且包含特征 bushes 则该文档属于 sport 类.

规则 4　如果某文档不包含特征 wheat 且不包含特征 bushes 则该文档属于 food 类.

根据规则对待分文本进行分析，并把该文本划分到与之最相近的一个规则对应的类中去。

总的来说，决策树的构造是一种自上而下、分而治之的归纳过程：从根节点开始，用特征集中的特征项对样本集进行划分，根据文档集中的文本是否包含该特征项将训练集分成若干个子样本集，每个子样本集构成一个新节点，对新节点再重复上述划分过程，这样不断循环，直至达到每一个叶节点包含的文档都属于同一个类别为止。因此，选用哪个特征项对文本集进行测试是构建决策树的一个关键环节，不同的决策树算法对此使

用的技术也不尽相同。

目前对于决策树分类已进行了大量的研究，这些研究工作涉及决策树推导、决策属性选择、决策树剪裁、由决策树抽取分类规则、提高决策效率等等，并开发了很多基于决策树的分类方法和系统，如 ID3[23]、C4.5[24] 、CART[25] 等。

2.5 分类性能的评价

随着文本自动分类技术的不断发展，各种分类方法如雨后春笋般层出不穷。因此，如何客观的评估和比较它们的分类性能就成了一个不容忽视的问题。用于文本分类性能评估的指标很多，究竟采用什么指标进行评估这和具体的应用有关。目前使用较多的分类性能评价指标为查全率和查准率，这是来源于信息检索领域的两个术语。但对于不同的分类问题，这两个指标的计算方式有所不同：

2.5.1 单类分类

文档分类中普遍使用的性能评估指标有查全率和查准率。对于文档类中的每一个类别，使用列联表来计算查全率和查准率。表 2.1 为一个列联表示例：

	真正属于该类的文档数	真正不属于该类的文档数
判断为属于该类的文档数	a	b
判断为不属于该类的文档数	c	d

表2.1　列联表示例

这时查全率和查准率分别定义为：

$$查全率 r = \frac{a}{a+c} \tag{2-49}$$

$$查全率 p = \frac{a}{a+b} \tag{2-50}$$

查全率和查准率只能反映分类算法单个类的分类性能，如果要对分类的整体性能进行评价我们一般采用如下的几种指标：

(1)宏平均查全率、查准率和微平均查全率、查准率

宏平均查全率、宏平均查准率是先对每一个类统计 r, p 值，然后对所有的类求平均值，即

$$宏平均查全率 \bar{r} = (\sum_{c \in C} r_c) / |C| \tag{2-51}$$

$$宏平均查准率 \bar{p} = (\sum_{c \in C} p_c) / |C| \tag{2-52}$$

微平均查全率、查准率是先建立了一个全局列联表，然后根据这个全剧列联表对文档进行平均，即：

$$\text{微平均查全率}\bar{r}=\frac{\sum_{c\in C}a}{\sum_{c\in C}a+\sum_{c\in C}c} \tag{2-53}$$

$$\text{微平均查准率}\bar{p}=\frac{\sum_{c\in C}a}{\sum_{c\in C}a+\sum_{c\in C}b} \tag{2-54}$$

(2)平衡点

对于分类系统来说，r,p值是相互影响的。提高r会引起p的减小，反之亦然。因此，为了更全面的反映分类系统的性能，一种做法是选取r,p的相等时的值来表征系统性能。这个值叫做平衡点值。当然，有时通过测试可能得不到r,p相等的值。这时取最接近的r,p值的平均值作为平衡点值。

(3)F 值

另一种常用的将查全率和查准率结合起来的性能评价方法是 F 测量，其计算公式为：

$$F_\beta=\frac{(\beta^2+1)*p*r}{\beta^2*p+r} \tag{2-55}$$

其中，β是一个用来调节查全率和查准率权重的参数。β一般取值为 1，这时公式转化为：

$$F_1=2rp/(r+p) \tag{2-56}$$

显然，平衡点是F_1的特殊形式，因为当$r=p$时，两者之值相等的。

2.5.2 多类分类

在多类分类问题中，查全率和查准率的定义如下：

$$\text{查全率}=\frac{\text{找到的该文档所属的正确类别数目}}{\text{判断为该文档所属类的类别数目}} \tag{2-57}$$

$$\text{查准率}=\frac{\text{找到的该文档所属的正确类别数目}}{\text{该文档所属的所有类别数目}} \tag{2-58}$$

整个分类器的评估应该是所有测试文档的这两个指标的统计平均值。通常使用的统计值为 11 点差值平均查准率，具体做法如下：在查全率分别达到 0%,10%,20%.....100%时分别测出它们对应的查准率，并取这 11 个值的平均值作为衡量这五种特征选择方法

性能的标准。

2.6 本章小结

本章对文本自动分类涉及到的几项关键技术，如文本预处理、文本表示模型、特征选择、分类算法、分类性能评价进行了具体的介绍。

3 一种改进的联合特征选择方法

本章分析了互信息和χ^2统计各自的优缺点以及两者之间存在的互补性，提出了一种联合的特征选择方法，最后用实验证明了该方法能提高分类系统的查全查准率。

3.1 问题的提出

问题 1

Yang 的试验表明，在 KNN 和 LLSF 分类器中χ^2统计是一种性能较好的特征选择方法。然而从理论上分析，该特征选择方法依然存在着如下的缺陷：提高了在指定类中出现频率较低而普遍存在于其它类的特征项在该类中的得分。

在指定类中出现频率较低而普遍存在于其它类的特征项，显然不能很好的表征该类别，应该被滤去。但在公式(2-15)中，当$A \to 0$且$B \to N$，而E不$\to 0$时，χ^2统计的计算公式近似等于[26]：

$$\chi^2(w,c) = \frac{(A+E)\times(A+B)}{N-A-B} \tag{3-1}$$

代入A的值可知，此时χ^2统计的值相对较大。于是由于上述类型的特征项得分较高，所以χ^2统计并不能去掉此类不合适的特征项。

问题 2

我们知道在训练语料库达到一定规模的时候，特征空间中必然存在相当多数量的低频特征项。由于它们出现的频率较低，必然只属于少数特定的类别，因此这些特征项携带了较为强烈的类别信息。但经过仔细观察后发现，低频特征项中只有不到 20%的特征项确实带有较强的类别信息，大多数的特征项都是噪音，不应该被选入到特征子集中[27]。然而在使用互信息进行特征选择时，由于该方法不能判别出哪些低频特征项是属于噪音的，并且在该方法中低频特征项往往比其他特征项更易被选入到特征子集中[28]，因此难免会有一些本属于噪音的低频特征项也被选入到了特征子集中，从而导致了互信息分类性能的下降。Yang 的实验也表明低频特征项是降低互信息分类性能的一个重要原因[29]。

3.2 问题的解决

针对问题一，根据互信息的计算公式(2-12)可知，对于在指定类中出现频率较低，而在其他类出现频率较高的特征项，当$A \to 0$且$B \to N$，而E不$\to 0$时，互信息的值将趋于负无穷大。因此互信息能很好得过滤掉那些在指定类中出现频率较低而普遍存在于其它类的特征项。

针对问题二，根据公式(2-15)可知，当$A \to 0$且$B \to 0$时，而E不$\to N$时，χ^2统计的计算公式近似等于[30]：

$$\chi^2(w,c) = \frac{(A+E)\times(A+B)}{N-A-E} \tag{3-2}$$

代入A、B的值可知，此时χ^2统计的值趋于 0。因此在χ^2统计中，低频特征项的得分往往很低，很容易被淘汰掉。

基于以上的分析，我们惊喜的发现互信息和χ^2统计之间存在着一定的互补性。由此，我们提出了如下的联合特征选择方法，并希望该方法能弥补互信息和χ^2统计存在的不足，提高它们的分类性能：

$$unit(t,c) = MI(t,c)\times\chi^2(t,c) \tag{3-3}$$

根据该联合特征选择方法的公式(3-3)我们可以看出：

⑴当$A \to 0$且$B \to N$，而E不$\to 0$时，根据公式(3-3)可知，χ^2统计的值被乘以了一个趋近于负无穷大的数，即此时联合特征的值是一个非常小的数。因此使用联合特征方法进行特征选择时，那些在指定类中出现频率较低而普遍存在于其它类的特征项将被优先过滤掉，这就弥补了χ^2统计方法存在的不足。

⑵当$A \to 0$且$B \to 0$，而E不$\to N$时，根据公式(3-3)可知，低频特征项对应的互信息值被乘以了一个近接于零的数。这就在很大程度上抑制了低频特征项的得分，降低了低频特征项被选入到特征子集的机率，从而也在一定程度上弥补了互信息方法存在的不足。但由于低频特征项中有 20%的特征项的确带有较强的鉴别信息，而联合特征方法并不能很好的将它们与其他低频特征项区分对待，可能将它们的得分一并抑制了，因此该方法存在着对低频特征项不公的缺点，这同时也是χ^2统计存在的不足。

⑶对于那些普遍出现于某个特定的类而在其他类中几乎不出现的频特征项来说（即当$B \to 0$，且$A \to A+E$时），我们认为它们能很好的表征这个特定的类，应该被选入到特征子集中。在联合特征选择中，这些特征项也很容易被选入到特征子集中.

根据公式(2-11)、 公式(3-3) 、公式 (2-15)，有：

$U = N-A-E-B \approx N-A-E$，

$$unit(t,c) = MI(t,c)\times\chi^2(t,c)$$

$$\approx [\log\frac{A\times N}{(A+E)\times(A+B)}]\times[\frac{N\times(AU-EB)^2}{(A+E)\times(B+U)\times(A+B)\times(E+U)}]$$

$$\approx [\log\frac{N}{(A+E)}]\times[\frac{N\times A^2U^2}{(A+E)\times U\times A\times(E+U)}]$$

$$\approx [\log\frac{N}{(A+E)}]\times[\frac{N\times AD}{(A+E)\times(E+U)}]$$

$$\approx [\log\frac{N}{(A+E)}]\times[\frac{N\times A\times(N-A-E)}{(A+E)\times(N-A)}] \tag{3-4}$$

①当某个特定的类包含的文档数不趋近于整个文档集包含的文档数时（即当$(A+E)$不$\to N$时），由于特征项普遍出现于这个特定的类中，所以有：

$$\frac{N\times A\times(N-A-E)}{(A+E)\times(N-A)}\approx N\times\frac{(\frac{N}{A+E}-1)}{(\frac{N}{A}-1)}\approx N$$

$\log\frac{N}{(A+E)}$为一个大于零，且不会趋近于零的数。

故

$unit(t,c)\approx[\log\frac{N}{(A+E)}]\times[\frac{N\times A\times(N-A-E)}{(A+E)\times(N-A)}]$是一个大于零且不会趋近于的数。例如当$(A+E)\to\frac{1}{2}N$，且$A\to(A+E)$时，$unit(t,c)\approx N$。因此在联合特征中，当特定的类包含的文档数不趋近于整个文档集包含的文档数时，那些普遍出现于这个特定的类而在其他类几乎不出现的特征项很容易被选入到特征子集中。

此外，当$A+E=\alpha$为一个定值时公式(3-4)近似等于：

$$unit(t,c)\approx[\log\frac{N}{(A+E)}]\times[\frac{N\times A\times(N-A-E)}{(A+E)\times(N-A)}]$$

$$\approx\log\frac{N}{\alpha}\times[\frac{N\times A\times(N-\alpha)}{\alpha\times(N-A)}]$$

$$\approx(\log\frac{N}{\alpha})\times N\times\frac{(\frac{N}{\alpha}-1)}{(\frac{N}{A}-1)} \tag{3-5}$$

根据上式(3-5)可知，A的值越趋近于α，$unit(t,c)$的值就越大。即当类中包含文档数一定时，特征项在这个类中出现的频率越高就越能表征这个类，这一点非常符合我们的常识。

②由于在训练过程中我们很少使用几乎所有样本都属于同一个类的训练集，因

此在$(A+E)\rightarrow N$的情况下，那些普遍存在于某个类而在其它类中几乎不出现的特征项能否被选入到特征子集的问题就显得不重要了，我们在这里也不予讨论.

3.3 实验设置

3.3.1 实验系统

使用了中国科学院计算所研制开发的自动文本分类系统作为实验系统。该系统采用向量空间模型来表示文本，使用 KNN 分类器对文本进行单类分类。另外，在特征选择之前，针对中文语料进行了必要的切词处理，并利用停用词表过滤了一些对分类贡献不大虚词和功能词。

由于该系统源代码公开，用户可以针对自己的需求对其进行修改，在本实验中，为了便于实现和测试联合的特征选择方法分类性能，我们对该系统进行了必要的修改。

3.3.2 测试集

在实验中，使用了复旦大学提供的中文文本分类语料库作为训练集和测试集。该语料库分为 10 大类：环境、计算机、交通、经济、军事、教育、体育、医药、艺术、政治，共包括 1882 篇训练样本和 937 篇测试样本。以下为训练文本集和测试文本集文本分布情况：

主题类	计算机	艺术	经济	环境	政治	军事	交通	医药	体育	教育
训练集	134	166	217	134	338	166	143	136	301	147
测试集	66	82	108	67	167	83	71	73	68	149

表3.1 训练文本集

3.3.3 实验系统界面

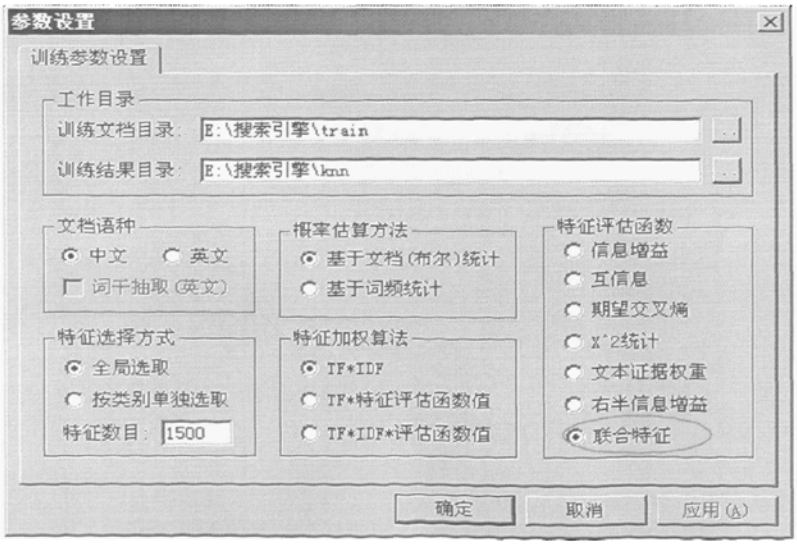

图 3.1 训练参数设置界面

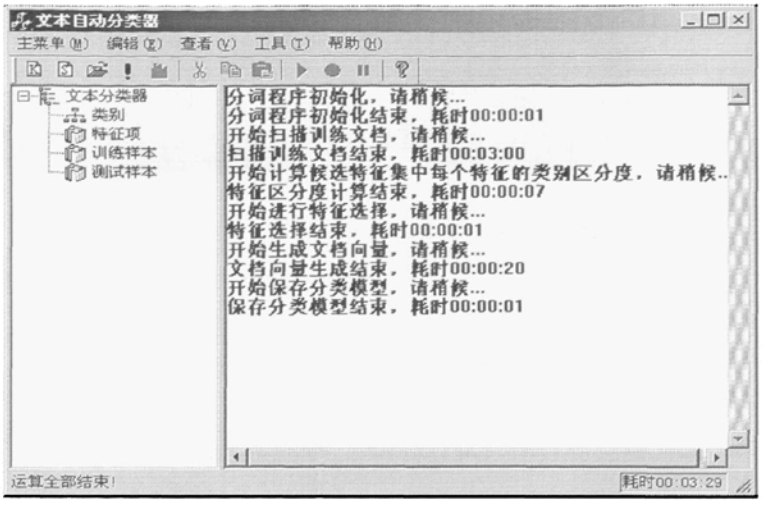

图 3.2　训练分类器界面

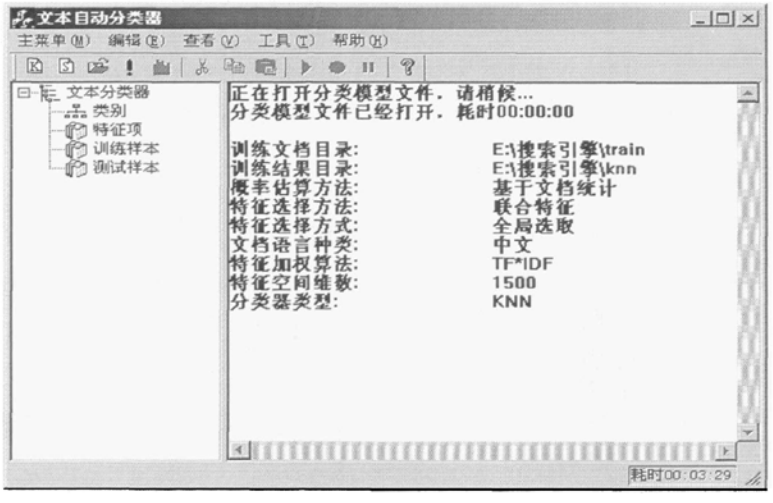

图 3.3　打开分类器界面

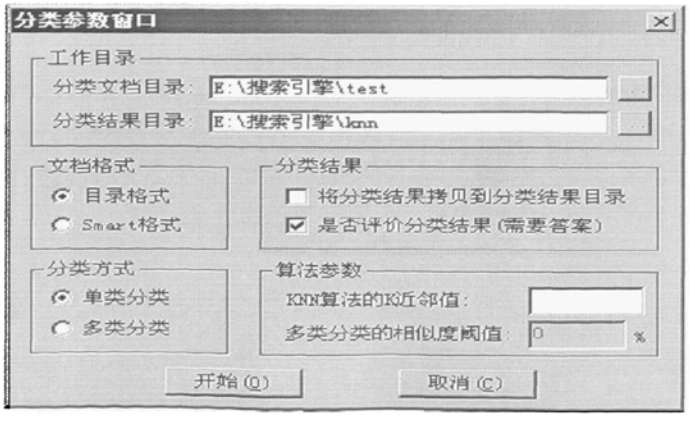

图 3.4 分类参数界面

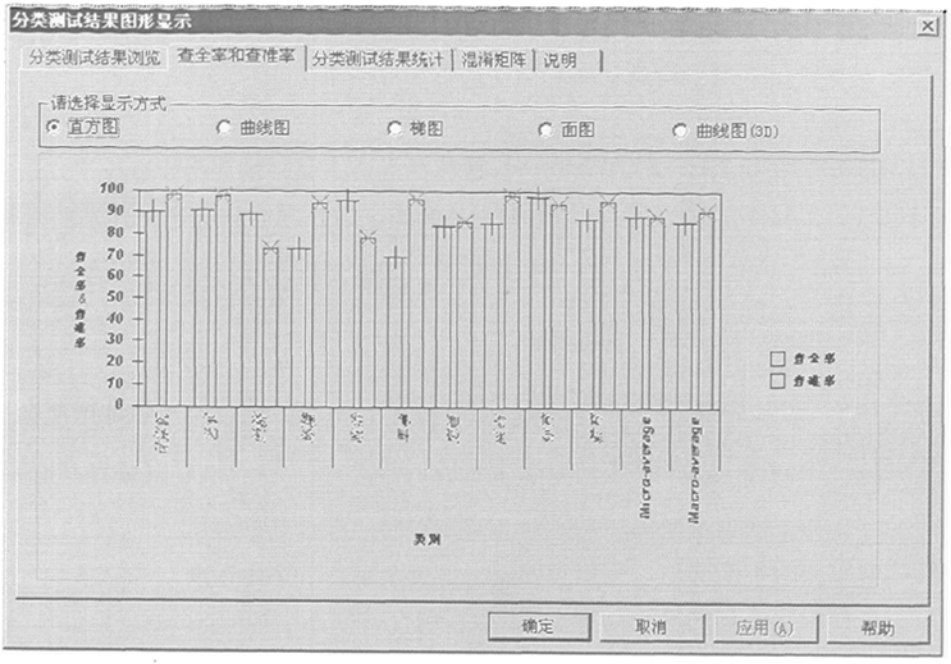

图 3.5 实验结果评估界面

3.3.3 实验结果及分析

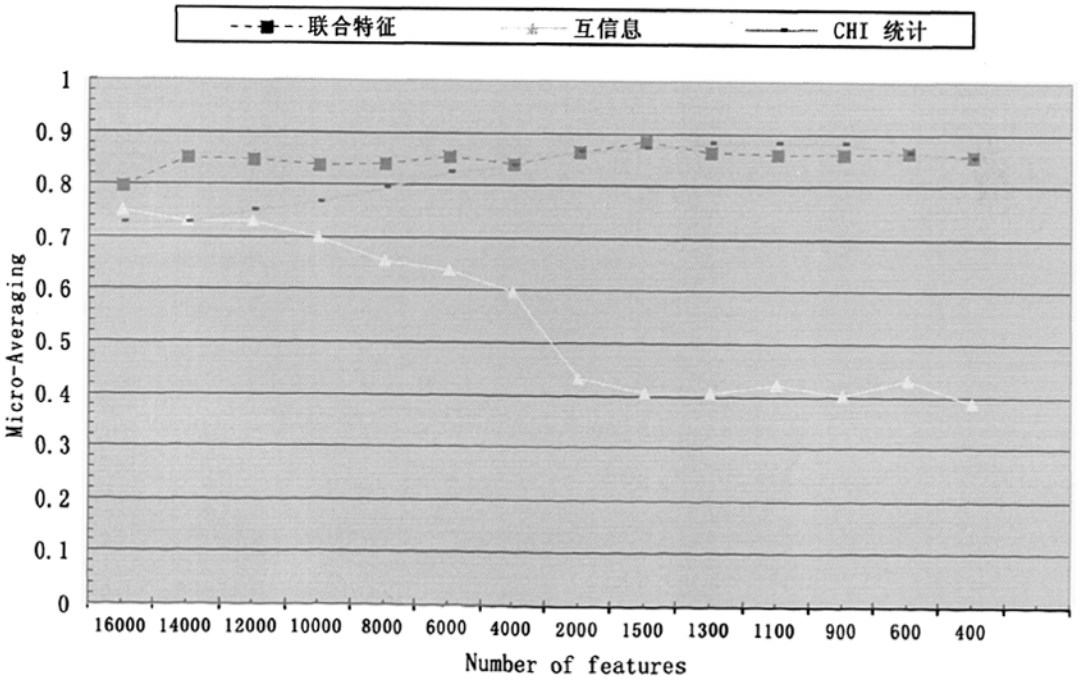

图 3.6　文本自动分类曲线图

在图 3.6 中横坐标为特征的维数，纵坐标为分类产生的微平均查全率、微平均查准

率。由于在该分类系统中得到的微平均查全率和微平均查准率是相等的，所以这两幅图的纵坐标同时代表了微平均查全率和微平均查准率两个值。

从图3.6我们可以看出：

⑴当特征维数减至1500维时，联合特征选择方法的微平均查全率、微平均查准率达到最高，其值为88.437%；当特征维数减至1500维以后，联合特征选择方法的微平均查全率、微平均查准率开始下降。当特征维数减至1300维时，χ^2统计方法的微平均查全率、微平均查准率达到最高，其值为88.33%；当特征维数减至1300维以后，χ^2统计方法的微平均查全率、微平均查准率开始下降。因此在本实验中，使用联合特征选择方法得到的最佳微平均查全率、微平均查准率要比使用χ^2统计特征选择方法得到的最佳微平均查全率、微平均查准率高一些。

⑵随着维数的不断降低，互信息的微平均查全率、微平均查准率明显下降，并且在整个分类过程中互信息的微平均查全率、微平均查准率都比较低；在整个降维的过程中，联合特征选择方法的微平均查全率、微平均查准率都明显得高于互信息的微平均查全率、微平均查准率。因此在本实验中，联合特征选择方法的分类性能在很大程度上优于互信息的分类性能.

表3.2 给出了维数减至1500时，使用三种特征选择方法得到的每个类别的具体分类结果

类别	互信息		χ^2统计		联合特征	
	查全率	查准率	查全率	查准率	查全率	查准率
计算机	84.848%	17.391%	83.333%	94.828%	90.909%	98.361%
艺术	18.293%	50.000%	92.683%	93.827%	91.463%	97.403%
经济	52.336%	52.336%	95.370%	72.028%	89.815%	73.485%
环境	11.940	29.630%	71.642%	88.889%	73.134%	94.231%
政治	58.182%	59.259%	92.814%	78.680%	95.808%	78.818%
军事	21.687%	21.687%	60.241%	90.909%	69.880%	96.667%
交通	8.451%	33.333%	83.099%	93.651%	84.907%	86.957%
医药	26.471%	48.649%	83.824%	98.276%	85.294%	98.305%
体育	55.479%	79.412%	98.658%	96.078%	97.315%	94.156%
教育	31.507%	38.333%	89.041%	90.278%	87.671%	95.552%
微平均	40.625%	40.625%	87.259%	87.259%	88.437%	88.437%

表 3.2 文本自动分类试验结果

根据上表我们可以看出：在所有类别中，联合特征选择方法的微平均查全率、微平均查准率都明显得高于互信息方法的微平均查全率、微平均查准率；在计算机、环境、政治、军事、医药这五个类别中，联合特征选择方法的微平均查全率、微平均查准率都高于χ^2统计方法的微平均查全率、微平均查准率；在教育这个类别中，联合特征选择方法的微平均查全率、微平均查准率略低于χ^2统计方法的微平均查全率、微平均查准率；但从总体看来，联合特征选择方法的微平均查全率、微平均查准率还是高于χ^2统计方法的微平均查全率、微平均查准率。

3.4 本章小结

本章对互信息和χ^2统计各自存在的优缺点以及两者之间存在的互补性进行了分析，并在此基础上提出了一种联合特征选择方法. 实验证明该方法能在一定程度上弥补了互信息和χ^2统计的不足，提高系统分类的微平均查全率、微平均查准率。从联合特征选择方法的公式中我们还可以直观的看到，该方法增加了系统训练阶段的计算量，提高了系统训练阶段的时间复杂度，不过在训练阶段高额的待价我们认为是可以忍受的，而且这些计算可以通过并行的方式计算完成，这一点在未来的工作中我们将着重研究。

4 基于广义向量空间模型的文本自动分类研究

本章对广义向量空间模型（Generalized Vector Space Model ,简称 GVSM)下的文本分类进行了研究，提出了基于广义向量空间模型的 KNN、TFIDF 分类方法，并对广义向量空间模型下的布尔交运算进行了修正。

众所周知，传统向量空间模型的一个缺陷是，在文档和查询的向量表示中假定了特征项是相互独立的，即假设了特征向量是相互正交的。在此前提下讨论特征项与特征项之间、文档与文档之间、以及文档与查询之间的相互关系显然不能令人满意。Wong 在[31]中建立了一种新方法，把特征向量用一组经挑选的正交基向量来表示。由此，特征项之间的关系可以直接由其向量表示给出较为精确的计算。我们称他所建立的模型为广义向量空间模型.

4.1 广义向量空间模型的基本原理

Wong 于 1985 年提出了这样的一个观点：特征向量之间是线性但不正交的，这就意味着特征向量不能看作是构成空间的正交基向量。基于这样的一个观点，Wong 建立了广义向量空间模型。在该模型中空间将由更小的分量构成，这些分量源于如下特定的集合。

设包括 m 篇文档的文档集合 D ={d_1, d_2,..., d_m}（d_j 为 D 中文献），被 n 个特征项 t_i(i=1,2,...，n)表示成一个 $m\times n$ 的文档矩阵 A：$A=(a_{ji})_{m\times n}$，其中 A 的元素 a_{ji} 为特征项 t_i 在文档 d_j 的权重 w_{ij} 。则由 D 的这 n 个特征项可生成一个包含 2^{2^n} 个布尔查询的集合 Q_{2^n}。Q_{2^n} 中同时包含了 2^n 个互不相同的最小项．在每个最小项中，t_i 和 $\neg t_i$ 其中之一出现且只出现一次。显然这些最小项再不能被进一步简化，它们被称为 Q_{2^n} 的基本元素，并用 m_K 来表示，且 Q_{2^n} 中的其它任何元素都可以由基本元素的析取范式表示。令 $\{m_K\}_{2^n}$ 表示 Q_{2^n} 的基本元素的集合，则其每一个元素都可由一个布尔向量 $(\delta_1,\delta_2,...,\delta_n)$ 唯一确定，即

$$m_k=t_1^{\delta_1}\wedge t_2^{\delta_2}\wedge...\wedge t_n^{\delta_n}\quad (k=1,2,...,\ 2^n),\ 此处\ t_n^{\delta_n}=\begin{cases} t_i 当 \delta_i=1 \\ \neg t_i 当 \delta_i=0 \end{cases} \tag{4-1}$$

因为 m_K 表示的是 Q_{2^n} 的最小项，而各个最小项之间是不相关的，所以各 m_K 之间是

相互独立的。因此集合$\{m_K\}_{2^n}$可以表示为一组2^n维正交向量的集合$\{\vec{m}_k\}$，即

$$\vec{m}_1 = (1,0,0,...,0),$$

$$\vec{m}_2 = (0,1,0,...,0),$$

$$\vdots$$

$$\vec{m}_{2^n} = (0,0,0,...,1) \tag{4-2}$$

这些向量就是广义空间模型的基向量，也是我们上文所提到的更小的分量。

由于Q_{2^n}中的任何一个布尔表达式的向量都可以通过这些基向量的向量和得到，而特征项t_i也是Q_{2^n}中的一个布尔表达式：

$$t_i = m_{k_1} \vee m_{k_2} \vee ... \vee m_{k_r} \tag{4-3}$$

其中m_{k_j}要满足$m_{k_j} \vee t_i = t_i$。由此我们可以得到特征项在广义向量空间模型中的向量表达式：

$$t_i = \frac{\sum_{\forall, g_i(m_k)=1} c_{i,k} m_k}{\sqrt{\sum_{\forall, g_i(m_k)=1} c_{i,k}^2}} \tag{4-4}$$

其中$c_{i,k} = \sum_{d_j | g_i(d_j)=g_i(m_k).for.all\ l,} w_{ij}$，被称为关联因子，用于计算特征项$t_i$和文档$d_j$的权重$w_{ij}$之和（文档$d_j$中词语的出现模式和最小项$m_K$的出现形式完全一致）。因此，只有当与词语形式相匹配的集合至少包含一篇文献时，最小项才是所需的目标。$g_i(m_k)$为一个二值函数，它返回最小项m_K中特征t_i的权值$\{0,1\}$。例如，当$g_1(m_2)=0$则表示最小项m_2中不包含特征相t_1。

给定了特征项的向量表示形式，特征项之间的相关性就可以由内积函数精确地计算出来，计算公式如下：

$$t_i \bullet t_j = \sum_{\forall r | g_i(m_k)=1 \wedge g_j(m_k)=1} c_{i,k} \times c_{i,k} \tag{4-5}$$

文档也可以表示为基本正交向量$\vec{m}$的线性组合，只需将各特征的向量表达式带入公式(2-5)便可．下面举例说明广义向量空间模型的原理。

例 4.1 设文献集合为$D=\{d_1,d_2,d_3,d_4,d_5,d_6\}$，特征项集合为：$T=\{$偶数$(t_1)$,素数$(t_2)$,殆素数$(t_3)\}$，设各文献表示所构成的矩阵为：

$$A=\begin{pmatrix}0 & 0.6 & 0.4\\ 0.7 & 0 & 0.2\\ 0 & 0.8 & 0\\ 0.7 & 0.1 & 0\\ 0.2 & 0.5 & 0.7\\ 0.2 & 0.8 & 0\end{pmatrix}$$

不失一般性，设$T=\{t_1,t_2,t_3\}$所生成的查询语言布尔代数Q^{2^3}所包含的基本元素为：

$$m_1=t_1\bar{t}_2t_3, \qquad m_2=t_1\bar{t}_2\bar{t}_3$$

$$m_3=\bar{t}_1t_2t_3, \qquad m_4=t_1t_2t_3$$

$$m_5=t_1t_2\bar{t}_3, \qquad m_6=\bar{t}_1\bar{t}_2t_3$$

$$m_7=\bar{t}_1t_2\bar{t}_3, \qquad m_8=\bar{t}_1\bar{t}_2\bar{t}_3$$

将特征项t_1,t_2,t_3用上述基本元素的析取范式表示如下：

$$\begin{aligned}t_1&=t_1\wedge(t_2\vee\bar{t}_2)\wedge(t_3\vee\bar{t}_3)=[(t_1\wedge t_2)\vee(t_1\wedge\bar{t}_2)]((t_3\vee\bar{t}_3)\\&=(t_1\wedge t_2\wedge t_3)\vee(t_1\wedge t_2\wedge\bar{t}_3)\vee(t_1\wedge\bar{t}_2\wedge t_3)\vee(t_1\wedge\bar{t}_2\wedge\bar{t}_3)\\&=m_4\vee m_5\vee m_1\vee m_2\end{aligned}$$

同理可得：

$$t_2=m_4\vee m_3\vee m_5\vee m_2$$

$$t_3=m_4\vee m_1\vee m_3\vee m_6$$

由公式(4-4)可将t_1,t_2,t_3用基本向量表示如下：

$$\begin{aligned}\vec{t}_1&=\frac{0.7\vec{m}_1+0.2\vec{m}_4+(0.7+0.2)\vec{m}_5}{\sqrt{0.7^2+0.2^2+0.9^2}}\\&=0.6\vec{m}_1+0.17\vec{m}_4+0.77\vec{m}_5;\end{aligned}$$

$$\begin{aligned}\vec{t}_2&=\frac{0.6\vec{m}_3+0.5\vec{m}_4+(0.1+0.8)\vec{m}_5}{\sqrt{0.6^2+0.5^2+0.9^2}}\\&=0.5\vec{m}_3+0.42\vec{m}_4+0.75\vec{m}_5\end{aligned}$$

$$\begin{aligned}\vec{t}_3&=\frac{0.2\vec{m}_1+0.4\vec{m}_3+0.7\vec{m}_4}{\sqrt{0.2^2+0.4^2+0.7^2}}\\&=0.24\vec{m}_1+0.48\vec{m}_3+0.84\vec{m}_4.\end{aligned}$$

将上述结果代入(2-5)式的文献的向量表示如下：

$$\vec{d}_1 = 0.6\vec{t}_2 + 0.4\vec{t}_3 = 0.1\vec{m}_1 + 0.5\vec{m}_3 + 0.57\vec{m}_4 + 0.38\vec{m}_5$$

$$\vec{d}_2 = 0.7\vec{t}_1 + 0.2\vec{t}_3 = 0.42\vec{m}_1 + 0.1\vec{m}_3 + 0.2\vec{m}_4 + 0.7\vec{m}_5$$

$$\vec{d}_3 = 0.8\vec{t}_2 = 0.4\vec{m}_3 + 0.33\vec{m}_4 + 0.6\vec{m}_5$$

$$\vec{d}_4 = 0.7\vec{t}_1 + 0.1\vec{t}_2 = 0.42\vec{m}_1 + 0.1\vec{m}_3 + 0.16\vec{m}_4 + 0.62\vec{m}_5$$

$$\vec{d}_5 = 0.2\vec{t}_1 + 0.5\vec{t}_2 + 0.7\vec{t}_3 = 0.28\vec{m}_1 + 0.58\vec{m}_3 + 0.84\vec{m}_4 + 0.53\vec{m}_5$$

$$\vec{d}_6 = 0.2\vec{t}_1 + 0.8\vec{t}_2 = 0.12\vec{m}_1 + 0.4\vec{m}_3 + 0.36\vec{m}_4 + 0.75\vec{m}_5$$

4.2 基于广义向量空间模型的文本分类方法

Wong 在 SMART 系统中对广义向量空间模型和传统向量空间模型作了比较试验，结果表明广义向量空间模型的检索效率在很大程度上优于传统向量检索模型[32]。

通过对传统向量空间模型和广义向量空间模型进行比较，我们发现使得广义向量空间模型的检索效率优于传统向量空间模型的检索效率的原因在于以下两个方面：一是在广义向量空间模型中，我们剔除了特征项之间相互独立的限制；二是在广义向量空间模型中，特征项、文档以及查询由更小的分量更为准确的表示了出来，因而利用内积或夹角余弦计算出的相似度就能更客观、准确地度量查询向量与文本向量之间的匹配程度，检出的文本也就更与查询相匹配。

从上述的实验结果和原因分析中我们得到了启示，希望将广义向量模型也应用到文本自动分类中，从而达到提高系统分类性能的目的。

在目前的文本自动分类中，很多分类方法都是基于传统的向量空间模型的。因而在使用这些方法进行分类时，我们同样也假设了特征项之间是相互独立的。但在实际中，特征项之间往往存在着各种各样的关系，诸如同义、蕴含等等。这些关系往往会造成同类文档分类的差异。例如，在两篇关于 IT 业的文档中，一篇是训练文档，在该篇文档中出现了较多的特征项“计算机”，而另一篇是待分文本，在该篇文档中特征项“电脑”出现的概率很高，但很少出现特征项“计算机”。在利用内积或夹角余弦度量时，这两篇文档的相似性将较小。而在自然语言中同义现象是普遍存在的，这种类型特征项的积累将会导致不精确的相似度计算，从而降低了分类的正确率；又例如，在法律类中，“警察”、“案件”、“暴力”、“公安局”、“纠纷”、“犯罪”、“派出所”、“动乱”等等特征项是存在某种联系的。其中一个特征的存在往往在某种程度上具有代替其他特征项的作用，而在相似度计算时却无法考虑这种影响，所以同样会导致分类结果的差异。因此，在特征项之间相互独立的假设下讨论待分文本与类别之间的关系显然不能令人满意，分类的结果也不能让人信服。

再者，很多基于传统的向量空间模型的分类方法的关键就是寻找与待分文本向量最为匹配的训练文本向量或类向量，这就和信息检索的关键问题——寻找与查询向量最为匹配的文本向量非常得相似。因而，如果待分文本、训练样本以及类能更为准确的表示成向量，则利用内积或夹角余弦计算出的相似度就能更客观地度量待分文本向量与训练样本或类向量之间的匹配程度，找出的训练文本或类也就更与待分文本相匹配。

如上所述，在广义向量空间模型中，我们不仅避免了特征项之间相互独立的假设，考虑了特征项之间的相关性，而且训练文本、类也能被更加准确地表示出来，因此本文将广义向量空间模型引入到了文本自动分类中。如下是两种基于广义向量空间模型的分类方法：

4.2.1 基于 GVSM 的 KNN 文本分类

传统的 KNN 方法是一种基于向量空间模型的分类方法，它的基本思想是：首先，对于一个待分文本，计算出它与训练样本集中每个文本之间的相似度，并依据相似度的大小找出 k 个与该待分文本最为相似的训练文本；接着，在此基础上给每一个类别打分，分值是这 k 个文本中属于该类的文本与待分文本之间相似度的和；最后，对类别的分值进行统计和排序，并把待分文档划分到那个得分最高的类别中去。

从 KNN 方法的基本思想中我们可以看出，该方法的分类结果在很大程度上依赖于这 k 个最相似文本的选择。由于在 KNN 中我们是利用相似度来选择这 k 个文本的，所以相似度越能客观、准确的度量待分文本与训练样本之间的匹配程度，选出的这 k 个训练文本就越与待分文本最为相似。如前所述，在广义向量空间模型中，由于文本能被更为准确的表示出来，因此待分文本与训练文本之间的相似度就能更为准确的反映出它们之间的匹配程度。近而选出的这 k 个训练文本也就越与待分文本相似，KNN 方法的分类性能也就越好。如下为使用广义向量空间模型进行 KNN 分类的具体过程：

(1)对训练样本进行预处理，计算特征权重，生成原始的训练样本向量；

(2)根据特征集计算 m_K 的集合，并生成训练样本在广义向量空间模型中的向量；

(3)输入待分样本，对其进行预处理，生成待分文本向量；

(4)根据 m_K 的集合，生成待分文本在广义向量空间模型中的向量；

(5)在广义向量空间模型中计算待分文本向量和训练样本向量的相似度，公式如下：

$$S(d_1,d_2)=\frac{\vec{d}_1\cdot\vec{d}_2}{\left|\vec{d}_1\right\|\vec{d}_2|} \tag{4-6}$$

(6)根据相似度找出与待分文本最相似的 k 个训练样本，并分别将这 k 个训练样本中属于同一类别的文本的相似度相加，得到该待分文本对不同类别的打分，对于这 k 个训练样本不涉及的类来说，其得分为零。

(7)将待分文本分到得分最高的那个类中。

经特征选择后，一个中等规模的训练文本集（Reuters-21578）大约包含 2000 个特征项，由这些特征项将生成大约 2^{2000} 个 m_K。如此众多的 m_K 造成了使用广义向量空间模型进行 KNN 分类时，相似度的计算和排序都非常耗时，需要付出的代价也相当高。这为我们通过实验证明广义空间向量模型能提高 KNN 方法的分类性能带来了巨大的困难。因此我们迫切需要在不降低分类性能的前提下缩减 m_K 的个数，究竟采用什么样的方法进行 m_K 个数的缩减以及传统的特征选择方法是否依然适用于广义向量空间模型，由于时间的关系我们在下一步的工作中再作研究。

4.2.2 基于 GVSM 的 TFIDF 文本分类

TFIDF 分类方法的一个关键是利用相似度寻找与待分文本向量最为匹配的类向量。由于在该方法中，类向量是由类中所有文档向量进行相加而得到的，因此如果能得到更为准确的文档向量表示形式，则由文档向量生成的类向量就能更加准确的表征这个类。近而利用内积或夹角余弦计算出的待分文本与类之间的相似度，就能更客观地度量它们之间的匹配程度，找出的类也就与待分文本更相匹配。由此，我们在 TFIDF 分类方法中引入了广义向量空间模型，希望通过提高文本以及类向量表示的准确度来提高 TFIDF 方法的分类性能。如下为使用广义向量空间模型进行 TFIDF 分类的具体过程：

⑴对训练样本进行预处理，计算权重，生成原始的训练样本向量；

⑵根据特征集计算 m_K 的集合，生成训练样本在广义向量空间模型中的向量；

⑶将训练样本中属于同一类别的文档对应的向量相加，生成各个类在广义向量空间模型中的向量。

⑷输入待分样本，对其进行预处理，并根据特征集合生成待分文本向量；

⑸生成待分文本在广义向量空间模型中的向量；

⑹在广义向量空间模型中计算待分文本向量和各个类向量之间的相似度，公式如下：

$$S(d_i,c_j)=\frac{\vec{d}_i\cdot\vec{c}_j}{\|\vec{d}_i\|\cdot\|\vec{c}_j\|} \qquad c_j\in C \tag{4-7}$$

⑺将待分文本分到相似度值最大的那个类中

4.3 广义向量空间模型中布尔运算的研究

决策树是一种基于布尔模型的分类方法。它通过不断地训练文本集从而构造一棵分类决策树，并生成由特征项及逻辑算符∧、∨、¬组成的分类规则，根据决策树和分类规则就可以判别出待分文本的类别。由于布尔模型也假设了特征项之间是相互独立的，因此，在使用决策树进行分类时，我们也忽略了特征项之间的关联性。那么，由决策树方法生成的分类规则也就不能让人信服。基于此，我们希望在决策树分类方法中也引入

广义向量空间模型。由于分类规则的生成还涉及到布尔运算，因此我们在引入广义向量空间模型的同时也要引入该模型下的布尔运算定律，这个定律是由 Wong 在 1986 年提出来的。

4.3.1 广义向量空间模型下的布尔运算定律

令 M 为广义向量空间模型中所有 m_k 的集合，M_i 为表示特征项 t_i 所需的 m_k 的集合，M_j 为表示特征项 t_j 所需的 m_k 的集合，$M_{i\wedge j}=M_i\cap M_j$，$M_{i\vee j}=M_i\cup M_j$ 则：

$$M=\sum_{i=1}^{n}M_i,\quad \vec{t}_i=\sum_{m_k\in M_i}c_{ik}\vec{m}_k,\quad \vec{t}_j=\sum_{m_k\in M_j}c_{jk}\vec{m}_k$$

广义向量空间模型下的布尔运算定义如下：

(1)特征项向量 $\vec{t}_i$ 的"¬"运算：

$$\neg\vec{t}_i=I-\vec{t}_i=I-\sum_{m_k\in M_i}c_{ik}\vec{m}_k=\sum_{m_k\in M_i}(1-c_{ik})\vec{m}_k+\sum_{m_k\in M-M_i}\vec{m}_k,\ \text{其中}\ I=\sum_{m_k\in M}\vec{m}_k \tag{4-8}$$

(2)特征项向量间的"∨"运算：

$$\vec{t}_i\vee\vec{t}_j=\sum_{m_k\in M_{i\vee j}}\max(c_{ik},c_{jk})\vec{m}_k \tag{4-9}$$

(3)特征项向量间的"∧"运算：

$$\vec{t}_i\wedge\vec{t}_j=\sum_{m_k\in M_i\wedge M_j}\max(c_{ik},c_{jk})\vec{m}_k \tag{4-10}$$

4.3.2 对广义向量模型中布尔运算的修正

通过广义向量模型下布尔运算定律的定义，任何布尔表达式都可以用相应的向量形式表示出来，从而布尔运算问题就转化为了向量之间的运算问题。在建立一个布尔运算体系时，我们应该遵守如下的基本原则：

(1)交换律：

$$t_1\vee t_2=t_2\vee t_1 \tag{4-11}$$

$$t_1\wedge t_2=t_2\wedge t_1 \tag{4-12}$$

(2)结合律：

$$(t_1\vee t_2)\vee t_3=t_1\vee(t_2\vee t_3) \tag{4-13}$$

$$(t_1\wedge t_2)\wedge t_3=t_1\wedge(t_2\wedge t_3) \tag{4-14}$$

(3)分配律：

$$t_1\wedge(t_2\vee t_3)=(t_1\wedge t_3)\vee(t_1\wedge t_2) \tag{4-15}$$

$$t_1 \vee (t_2 \wedge t_3) = (t_1 \vee t_3) \wedge (t_1 \vee t_2) \tag{4-16}$$

(4)等幂律：

$$t_1 \vee t_1 = t_1 \tag{4-17}$$

$$t_1 \wedge t_1 = t_1$$

(5)吸收律：

$$t_1 \vee (t_1 \wedge t_2) = t_1 \tag{4-18}$$

$$t_1 \wedge (t_1 \vee t_2) = t_1, \tag{4-19}$$

(6) 德·摩根法则：

$$\neg(t_1 \vee t_2) = \neg t_1 \wedge \neg t_2, \tag{4-20}$$

$$\neg(t_1 \wedge t_2) = \neg t_1 \vee \neg t_2 \tag{4-21}$$

然而，在利用 Wong 提出的布尔运算定律进行布尔运算时，我们发现该定律并不能完全地满足上述基本原则，以下通过举反例来说明：

例2.2　设 $M = \{\vec{m}_1, \vec{m}_3, \vec{m}_4\}, \vec{t}_1 = 0.3\vec{m}_1 + 0.09\vec{m}_4\ \vec{t}_2 = 0.2\vec{m}_3 + 0.17\vec{m}_4$，

则根据 GVSM 中的布尔运算定律有：

$$\vec{t}_1 \wedge \vec{t}_2 = (0.3\vec{m}_1 + 0.09\vec{m}_4) \wedge (0.2\vec{m}_3 + 0.17\vec{m}_4) = 0.17\vec{m}_4$$

$$\vec{t}_1 \vee (\vec{t}_1 \wedge \vec{t}_2) = (0.3\vec{m}_1 + 0.09\vec{m}_4) \vee \{(0.3\vec{m}_1 + 0.09\vec{m}_4) \wedge (0.2\vec{m}_3 + 0.17\vec{m}_4)\}$$

$$= (0.3\vec{m}_1 + 0.09\vec{m}_4) \vee (0.17\vec{m}_4) = 0.3\vec{m}_1 + 0.17\vec{m}_4$$

显然

$$\vec{t}_1 \vee (\vec{t}_1 \wedge \vec{t}_2) \neq \vec{t}_1$$

故　该布尔运算定律不能满足吸收律。

根据 GVSM 中的布尔运算的定律又有：

$$\neg(\vec{t}_1 \wedge \vec{t}_2) = \neg\{(0.3\vec{m}_1 + 0.09\vec{m}_4) \wedge (0.2\vec{m}_3 + 0.17\vec{m}_4)\}$$

$$= \vec{m}_1 + \vec{m}_3 + (1 - 0.17)\vec{m}_4$$

$$= \vec{m}_1 + \vec{m}_3 + 0.83\vec{m}_4$$

$$\neg\vec{t}_1 \vee \neg\vec{t}_2 = \neg(0.3\vec{m}_1 + 0.09\vec{m}_4) \vee \neg(0.2\vec{m}_3 + 0.17\vec{m}_4)$$

$$=[(1-0.3)\vec{m}_1+\vec{m}_3+(1-0.09)\vec{m}_4]\vee[(\vec{m}_1+(1-0.2)\vec{m}_3+(1-0.17)\vec{m}_4]$$

$$=\vec{m}_1+\vec{m}_3+0.91\vec{m}_4$$

显然

$$\neg(\vec{t}_1\wedge\vec{t}_2)\neq\neg\vec{t}_1\vee\neg\vec{t}_2$$

故　该布尔运算定律不能满足德.摩根法则。

从以上的两个反例我们可以看出，Wong 提出的布尔运算定律不能满足德.摩根法则和吸收律这两大基本原则。基于此，我们对造成这一现象的原因进行了分析和研究，并尝试通过修改该定律中的"∧"运算解决了这一问题，修改后的"∧"运算为：

$$\vec{t}_1\wedge\vec{t}_2=\sum_{m_K\in M_{1\wedge2}}\min(c_{1K},c_{2K})\vec{m}_K \tag{4-22}$$

以下为引入新的"∧"运算的定义后，GVSM 下的布尔运算定律能满足交换律，结合律，等幂律，分配律，对合律、吸收律和德.摩根法则这几大基本原则的证明过程：

设

$\vec{t}_1=\sum_{m_K\in M_1}c_1\vec{m}_K$，$\vec{t}_2=\sum_{m_K\in M_2}c_2\vec{m}_K$，$\vec{t}_3=\sum_{m_K\in M_3}\bar{c}_3\vec{m}_K$。

根据 GVSM 中定义有：

(1)交换律：

$$\vec{t}_1\vee\vec{t}_2=\vec{t}_2\vee\vec{t}_2;$$

$$\vec{t}_1\wedge\vec{t}_2=\vec{t}_2\wedge\vec{t}_1$$

由于

$$\vec{t}_2\wedge\vec{t}_1=\sum_{m_k\in M_{1\wedge2}}\min(c_1,c_2)\vec{m}_K,$$

且

$$\vec{t}_1\wedge\vec{t}_2=\sum_{m_k\in M_{1\wedge2}}\min(c_2,c_1)\vec{m}_K;$$

又由于

$$\sum_{m_k\in M_{1\wedge2}}\min(c_1,c_2)\vec{m}_K=\sum_{m_k\in M_{1\wedge2}}\min(c_2,c_1)\vec{m}_K,$$

故

$$\vec{t}_1\wedge\vec{t}_2=\vec{t}_2\wedge\vec{t}_1,$$

同理可证$\vec{t}_1\vee\vec{t}_2=\vec{t}_2\vee\vec{t}_2$

(2)结合律：

$$(\vec{t}_1 \vee \vec{t}_2) \vee \vec{t}_3 = \vec{t}_1 \vee (\vec{t}_2 \vee \vec{t}_3)$$

$$(\vec{t}_1 \wedge \vec{t}_2) \wedge \vec{t}_3 = \vec{t}_1 \wedge (\vec{t}_2 \wedge \vec{t}_3);$$

由于

$$(\vec{t}_1 \wedge \vec{t}_2) \wedge \vec{t}_3 = \sum_{m_K \in M_{1\wedge 2} \wedge M_3} \min\{\min(c_1, c_2), c_3\} \vec{m}_K ,$$

且

$$\vec{t}_1 \wedge (\vec{t}_2 \wedge \vec{t}_3) = \sum_{m_K \in M_{3\wedge 2} \wedge M_3} \min\{\min(c_3, c_2), c_1\} \vec{m}_K$$

又由于

$$\sum_{m_K \in M_{1\wedge 2} \wedge M_3} \min\{\min(c_1, c_2), c_3\} \vec{m}_K = \sum_{m_K \in M_{3\wedge 2} \wedge M_3} \min\{\min(c_3, c_2), c_1\} \vec{m}_K$$

$$= \sum_{m_K \in M_{3\wedge 2} \wedge M_3} \min(c_3, c_2, c_1) \vec{m}_K$$

故

$$(\vec{t}_1 \wedge \vec{t}_2) \wedge \vec{t}_3 = \vec{t}_1 \wedge (\vec{t}_2 \wedge \vec{t}_3)$$

同理可证 $(\vec{t}_1 \vee \vec{t}_2) \vee \vec{t}_3 = \vec{t}_1 \vee (\vec{t}_2 \vee \vec{t}_3)$

(3)分配律：

$$\vec{t}_1 \wedge (\vec{t}_2 \vee \vec{t}_3) = (\vec{t}_1 \wedge \vec{t}_3) \vee (\vec{t}_1 \wedge \vec{t}_2),$$

$$\vec{t}_1 \vee (\vec{t}_2 \wedge \vec{t}_3) = (\vec{t}_1 \vee \vec{t}_3) \wedge (\vec{t}_1 \vee \vec{t}_2)$$

由于

$$\vec{t}_1 \wedge (\vec{t}_2 \vee \vec{t}_3) = \sum_{m_K \in M_1 \cap M_{2\vee 3}} \min\{c_1, \max(c_2, c_3)\} \vec{m}_K ,$$

$$(\vec{t}_1 \wedge \vec{t}_3) \vee (\vec{t}_1 \wedge \vec{t}_2) = \sum_{m_K \in M_{1\wedge 3} \cup M_{1\wedge 2}} \max\{\min(c_1, c_2), \min(c_1, c_3)\} \vec{m}_K$$

又由于

$$M_{1\wedge 3} \cup M_{1\wedge 2} = M_1 \cap M_{2\vee 3}$$

且

对所有的 m_K 若 $m_K \in M_1 \cap M_{2\vee 3}$ 有

$$\min\{c_1, \max(c_2, c_3)\} = \begin{cases} c_3, \text{当} c_2 \langle c_3 \langle c_1 \\ c_2, \text{当} c_3 \langle c_2 \langle c_1 \\ c_1, \text{当} c_1 \langle \max(c_2, c_3) \end{cases}$$

$$\max\{\min(c_1,c_2),\min(c_1,c_3)\}=\begin{cases}c_3,\text{当}c_2\langle c_3\langle c_1\\ c_2,\ \text{当}c_3\langle c_2\langle c_1\\ c_1,\ \text{当}c_1\langle\max(c_2,c_3)\end{cases}$$

则

$$\sum_{m_K\in M_1\cap M_{2\vee3}}\min\{c_1,\max(c_2,c_3)\}\vec{m}_K=\sum_{m_K\in M_{1\wedge3}\cup M_{1\wedge2}}\max\{\min(c_1,c_2),\min(c_1,c_3)\}\vec{m}_K$$

故

$$\vec{t}_1\wedge(\vec{t}_2\vee\vec{t}_3)=(\vec{t}_1\wedge\vec{t}_3)\vee(\vec{t}_1\wedge\vec{t}_2),$$

同理可证$\vec{t}_1\vee(\vec{t}_2\wedge\vec{t}_3)=(\vec{t}_1\vee\vec{t}_3)\wedge(\vec{t}_1\vee\vec{t}_2)$

(4)等幂律：

$$\vec{t}_1\vee\vec{t}_1=\vec{t}_1$$

$$\vec{t}_1\wedge\vec{t}_1=\vec{t}_1$$

由于

$$\vec{t}_1\wedge\vec{t}_1=\sum_{m_K\in M_{1\vee2}}\min(c_1,c_1)\vec{m}_K=\sum_{m_K\in M_1}c_1\vec{m}_K=\vec{t}_1$$

故

$$\vec{t}_1\wedge\vec{t}_1=\vec{t}_1$$

同理可证 $\vec{t}_1\vee\vec{t}_1=\vec{t}_1$

(5)吸收律：

$$\vec{t}_1\vee(\vec{t}_1\wedge\vec{t}_2)=\vec{t}_1,$$

$$\vec{t}_1\wedge(\vec{t}_1\vee\vec{t}_2)=\vec{t}_1,$$

由于

$$\vec{t}_1\vee(\vec{t}_1\wedge\vec{t}_2)=\sum_{m_K\in M_1\vee M_{1\wedge2}}\max\{c_1,\min(c_1,c_2)\}\vec{m}_K=\sum_{m_K\in M_1}c_1\vec{m}_K=\vec{t}_1$$

故

$$\vec{t}_1\vee(\vec{t}_1\wedge\vec{t}_2)=\vec{t}_1,$$

同理可证$\vec{t}_1\wedge(\vec{t}_1\vee\vec{t}_2)=\vec{t}_1,$

(6) 德·摩根法则：

$$\neg(\vec{t}_1\vee\vec{t}_2)=\neg\vec{t}_1\wedge\neg\vec{t}_2,$$

$$\neg(\vec{t}_1 \wedge \vec{t}_2) = \neg\vec{t}_1 \vee \neg\vec{t}_2$$

由于

$$\begin{aligned}\neg(\vec{t}_1 \vee \vec{t}_2) &= \sum_{m_K \in M_{1\vee 2}} \{1 - \max(c_1, c_2)\}\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K \\ &= \sum_{m_k \in M_{1\vee 2}} \min\{(1-c_1),(1-c_2)\}\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K \\ &= \sum_{m_K \in M_1 - M_{1\wedge 2}} (1-c_1)\vec{m}_K + \sum_{m_K \in M_2 - M_{1\wedge 2}} (1-c_2)\vec{m}_K \\ &\quad + \sum_{m_K \in M_{1\wedge 2}} \min\{(1-c_1),(1-c_2)\}\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K\end{aligned}$$

又由于

$$\begin{aligned}\neg\vec{t}_1 &= \sum_{m_K \in M_1} (1-c_1)\vec{m}_K + \sum_{m_K \in M - M_1} \vec{m}_K \\ &= \sum_{m_K \in M_1 - M_{1\wedge 2}} (1-c_1)\vec{m}_K + \sum_{m_K \in M_{1\wedge 2}} (1-c_1)\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K + \sum_{m_K \in M_2 - M_{1\wedge 2}} \vec{m}_K \\ \neg\vec{t}_2 &= \sum_{m_K \in M_2} (1-c_2)\vec{m}_K + \sum_{m_K \in M - M_2} \vec{m}_K \\ &= \sum_{m_K \in M_2 - M_{1\wedge 2}} (1-c_2)\vec{m}_K + \sum_{m_K \in M_{1\wedge 2}} (1-c_2)\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K + \sum_{m_K \in M_1 - M_{1\wedge 2}} \vec{m}_K\end{aligned}$$

$$\begin{aligned}\neg\vec{t}_1 \wedge \neg\vec{t}_2 = &\left(\sum_{m_K \in M_1 - M_{1\wedge 2}} (1-c_1)\vec{m}_K + \sum_{m_K \in M_{1\wedge 2}} (1-c_1)\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K + \sum_{m_K \in M_2 - M_{1\wedge 2}} \vec{m}_K \right) \wedge \\ &\left(\sum_{m_K \in M_2 - M_{1\wedge 2}} (1-c_2)\vec{m}_K + \sum_{m_K \in M_{1\wedge 2}} (1-c_2)\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K + \sum_{m_K \in M_1 - M_{1\wedge 2}} \vec{m}_K \right) \\ = &\sum_{m_K \in M_1 - M_{1\wedge 2}} \min\{(1-c_1),1\}\vec{m}_K + \sum_{m_K \in M_{1\wedge 2}} \min\{(1-c_1),(1-c_2)\}\vec{m}_K \\ &+ \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K + \sum_{m_K \in M_2 - M_{1\wedge 2}} \min\{(1-c_2),1\}\vec{m}_K \\ = &\sum_{m_K \in M_1 - M_{1\wedge 2}} (1-c_1)\vec{m}_K + \sum_{m_K \in M_2 - M_{1\wedge 2}} (1-c_2)\vec{m}_K \\ &+ \sum_{m_K \in M_{1\wedge 2}} \min\{(1-c_1),(1-c_2)\}\vec{m}_K + \sum_{m_K \in M - M_{1\vee 2}} \vec{m}_K\end{aligned}$$

故

$$\neg(\vec{t}_1 \vee \vec{t}_2) = \neg\vec{t}_1 \wedge \neg\vec{t}_2,$$

同理可证 $\neg(\vec{t}_1 \wedge \vec{t}_2) = \neg\vec{t}_1 \vee \neg\vec{t}_2$

从以上证明过程中我们可以看出，本文对广义向量空间模型中"∧"运算的修正使得 Wong 提出的布尔运算定律能更好地满足布尔运算的基本原则，这为我们在决策树分类方法中引入广义向量空间模型打下了坚实的基础。关于广义向量空间模型在决策树方法

中的具体应用过程，由于时间的关系我们在下一步的工作中再作研究。

4.4 本章小结

在本章中，我们将广义向量空间模型引入到了文本自动分类中，提出了基于广义向量空间模型的 KNN、TFIDF 分类方法。然后通过举反例说明了 Wong 提出的布尔运算定律存在着不能满足德.摩根法则和吸收律的不足，并针对这两点不足之处提出了相应的改进方法：对该定律下的布尔交运算进行了修正。此外，由于广义向量模型的向量维数较传统向量空间模型的向量维数成比例增长，这为我们把广义向量空间模型应用到实际中带来了很大的困难。鉴于此，我们今后工作的重点将放在如何解决广义向量空间模型维数缩减的问题上，并希望能构建一个基于广义向量空间模型的文本分类系统，从而为今后的研究提供一个实验平台。

5 总结与进一步工作

本章总结了本文所作的研究工作，概括了取得的研究成果，指出了今后需要进一步研究的问题和方向。

5.1 总结

随着信息时代的到来和 Internet 的日益普及，文本信息迅速膨胀，使得文本自动分类技术成为了信息技术领域的一个重要研究方向。文本自动分类技术的出现，使文档可以自动的按照类别的方式进行组织和处理，非常符合人类组织和处理信息的方式，方便了人们准确地定位所需要的信息和分流信息。

但是，文本分类器的稳定性、快速性和准确性还需要进一步提高，本文旨在对影响自动分类性能的各种问题进行深入研究，并提出了几种改进的方法。本文所取得的创新性成果主要体现在以下三个方面：

⑴提出了一种联合的特征选择方法。χ^2 统计和互信息是两个经常被采用的特征抽取方法，它们的不足之处在于：χ^2 统计提高了在指定类中出现频率较低而普遍存在于其他类的特征项在该类中的权重；互信息受低频特征项的影响，分类性能相对较差。然而，互信息对在指定类中出现频率较低而普遍存在于其它类中的特征项却有着很好过滤能力，χ^2 统计对低频词的过滤能力比较强。显然，互信息和 χ^2 统计存在着一定的互补性。基于此本文提出了一种联合特征选择方法，该方法在一定程度上弥补了 χ^2 统计和互信息各自存在的不足，实验证明该方法能在一定程度上提高分类系统的查准率与查全率

⑵基于广义向量空间模型下的文本自动分类研究。与传统向量空间模型相比，广义向量空间模型的优点在于：它剔除了特征项之间相互独立的假设，考虑了特征项之间的相关性；广义向量空间模型中的特征项由更小的分量较准确的表示了出来，近而文本的向量表示也更为准确.因此在广义向量空间模型中进行 KNN 和 TFIDF 分类时，利用内积或余弦函数计算出的相似度能更为客观地度量待分文本与训练样本以及类之间的匹配程度，从而也提高了这两种分类方法的性能.

⑶对广义向量空间模型下的布尔交运算进行了修正。广义向量空间模型中布尔定律的建立，将布尔运算问题转化为了向量之间的计算问题。然而在对该定律进行了深入分析后，发现该理论存在着无法满足吸收律和德.摩根法则的缺陷。因此，本文尝试对该布尔运算定律中的交运算进行了重新定义，并从理论的角度证明了新的交运算能有效性的弥补这一缺陷。

5.2 进一步工作

文本自动分类是一个充满机遇和挑战的研究领域，许多问题还有待进一步的探索和研究。本文今后的研究主要致力于以下几个方面：

⑴联合特征选择方法的进一步改进。本文提出的联合特征选择方法在一定程度上弥补了互信息和χ^2统计各自存在的缺陷，但依然存在着对低频词不公，增加了系统训练阶段计算量的缺点。因此如何克服这两个缺点，从而进一步提高联合特征选择方法的分类性能是我们下一步的主要工作之一。

⑵ 广义向量空间模型的降维。广义向量空间模型的维数较传统空间模型成指数增长，这给使用其进行的文本自动分类带来了相当大的计算量和时间空间复杂度．因此迫切需要一种行之有效的降维方法来对广义向量空间模型进行降维处理，具体采用何种方法进行降维也是一个有待我们深入研究的问题．

参 考 文 献

[1] S.Chakrabarti. Hypertext databases and data mining .In proceedings of SIGMOD99,1999.

[2] K.Dave,S.Lawrence,and D.Pennock.Mining the peanut gallery:opinion extraction and semantic classification of product reviews, In Proceedings of the 22th International World Wide Web Conference ,Budapest,Hungary,2003.

[3] J. Yi, T.Nasukawa, R.Bunescu, W, Niblack .Sentiment analyzer: Extracting Sentiments about a Given Topic using Natural language Processing Techniques ,Proceedings of the Third IEEE International Conference on Data Mining, November 19-22,2003

[4] 刘永丹，曾海泉，李荣陆，胡运发，基于语义分析的倾向性过滤[J]。通信学报

[5] 都云琪，中文文本自动分类的研究与实现。西安电子科技大学硕士论文 8~9,2002

[6] 宋枫溪，自动文本分类若干基本问题研究，南京理工大学博士论文,2004

[7] 李晓黎，刘继敏，史忠植。概念推理网及其在文本分类中的应用[j]。计算机研究与发展 2000,37(9):1033-1038.

[8] 李荣陆、胡运发。基于密度的 KNN 文本分类器训练裁剪方法[J].计算机研究与发展,2004,42(4):539~545.

[9] 黄萱菁，吴立德，石崎洋之等。独立于语种的文本分类方法[J].中文信息学报，2000,14(6):1~7

[10] 王怡，盖杰，武港山，王继成。基于潜在语义分析的中文文本层次分类技术[J]，计算机应用研究，2004,8:151~165

[11] 刁力力，胡可云，路玉昌，石纯一。用 Boosting 方法组和增强 Stumps 进行文本分类[J]软件学报 2002,13(8):1363~1367

[12] 陈治刚，何丕廉，孙越恒，郑小慎，基于向量空间模型的文本分类方法的研究与实现。计算机应用[J],2004,6(24)

[13] C.Apte, F.J. Damerau, S.M. Weiss, Automated learning of decision rules for text categorization, ACM Transactions on Information Systems, 12(3):223-251,1994.

[14] Y.H Li and A.K. Jain. Classification of text documents.The Computer Journal 41(8):537-546,1998.

[15] Y.Yang,J.O.Pedersen. A Comparative Study on Feature Selection in Text Categorization .In Machine Learning:Proceedings of the Fourthteenth International Conference (ICML'97),pp.412-420,1997.

[16] Y.Yang, X.Liu .A Re_examination of Text Categorization Methods.In Proceedings of SDAIR_99, 22nd ACM International Conference on Research and Development in Information Retrieval (Berkeley ,US,1999)pp.42_49.

[17] A.Berger. Error-correcting output coding for text classification. In Proceedings of International Joint Conference on Artificial Intelligence: Workshop on Machine Learning for Information Filtering,1999.

[18] Y.Yang,J.O.Pedersen. A Comparative Study on Feature Selection in Text Categorization .In Machine Learning:Proceedings of the Fourthteenth International Conference (ICML'97),pp.412-420,1997.

[19] 李荣陆，文本分类及相关技术研究,博士论文。

[20]Y.Yang and C.G.Chute. Alinear least squares fit mapping method for information retrieval from natural language texts In Proceedings of the 14th conference on Computational Liguistics ,1992.

[21] Y.Yang,X liu. A Re-examination of Text Categorization Methods.In Proceedings of SIGIR-99,22nd ACM International Conference on Research and Development in Information Retrieval (Berkeley, US,1999)

[22] T.Zhang and F.J.OLes,Text Categorization Based on Regularized Linear Classification Methods, Information Retrieval ,4,5-31,2001.

[23] J.R.Quinlan.Inducution of decision tree.Machine Learning.1986,1:81-106

[24] J.R.Quinlan.C4.5:Programs for Machine Learning.San Mateo,CA:Margan Kaufmann,1993.

[25] L.Breiman,J.Friendman, R.Qlshen ,and C.Stone. Classification and Regression Trees. Monterey CA: Wadsworth International Group,1984.

[26] 冯是聪，搜索引擎个性化查询服务研究，2006,6

[27] 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[J]，2004,18(1):26~32.

[28] Yiming Yang , Jan O. Pedersen .A Comparative Study on Feature Selection in Text Categorization [A] .Proc .of, the 14th International Conference on Machine Learning ,ICML'97[C],1997,412~420.

[296] Yiming Yang , Jan O. Pedersen .A Comparative Study on Feature Selection in Text Categorization [A] .Proc .of, the 14th International Conference on Machine Learning ,ICML'97[C],1997,412~420.

[30] 冯是聪，搜索引擎个性化查询服务研究，2006,6

[31] Wong, S.K.M.,Ziarko， W.,Wong ,P.C.N., Generalized Vector Space in Information Retrieval, Proceedings of the Seventh International Conference on Information Storage and Retrieval ,1985,pp 18-25.

[32] 康耀红，现代情报检索理论[M],北京：科学技术文献出版社，1990.

攻读硕士学位期间参与的科研项目及主要成果

参与的科研项目

1　Internet 信息检索理论与技术，教育部重点基金项目，课题编号 03144

2　广义信息检索理论，加拿大卡尔加利大学合作项目

3　语义图像信息检索的研究与实现，海南省科学基金项目

主要研究成果

1　石敏，康耀红．关于广义向量空间模型中布尔运算的修正．第二届大规模信息检索与安全会议，中科院，2005 年

2　石敏，康耀红．一种改进的特征抽取方法．海南大学学报（自然科学版）[J]，2005, 23(4):347~351.

致 谢

时光如梭，转瞬即逝，三年充实而愉快的硕士生活即将过去。在论文提交之际，我更加留恋我们美丽的学校——海南大学。留恋诲人不倦、知识渊博的老师们；留恋温暖而奋进的集体；留恋乐于助人的师兄弟、师姐妹……

值此之际，谨向三年来给我以指导、帮助、支持我的所有老师和同学们表示衷心的感谢。学位论文的顺利完成得易于众多师长、学友和亲人的鼎力支持，感激之情，难于言表。

首先，我要衷心的感谢我的导师康耀红教授，本文是在他的悉心指导下完成的。康耀红老师深厚的理论功底、渊博的学识、严谨认真的学术作风、积极勤奋的工作态度深深的影响和教育了我。从他身上，我不仅学到了知识和科学研究方法，更重要的是学到了很多做人道理；体会到了“吃苦耐劳”、“认真”等词语的真正含义！

其次，感谢李太君教授、雷京生博士、魏应彬教授、钟声教授、陈少凡老师、伍小芹老师、曾水香老师的帮助和中肯的意见。感谢同门师兄弟、师姐妹：张春元、温小斌、孙秉强、赵正文、王国金、曹聪聪、方磊坤，与他们的讨论拓宽了我的思路、开阔了我的视野。在生活上，大家相互关心，相互帮助，一起度过的海大三年的快乐时光，点点滴滴铭记心间，将成为人生美好的回忆。

特别地，我要感谢我的家人，感谢一直支持我的父母亲，给我创造了良好的学习和生活环境。我的每一滴成就的取得，都包含着他们辛勤的汗水。

海纳百川，大道致远。我将以饱满的热情迎接未来的工作和挑战，实现自己远大的理想！

再一次感谢所有关心、支持和帮助过我的人们！

石敏

二〇〇六年六月

附录 A

a	*can*	*having*	*must*	*resulted*	*they*
about	*could*	*here*	*nearly*	*resulting*	*this*
after	*did*	*how*	*neither*	*same*	*those*
again	*do*	*however*	*no*	*seem*	*through*
all	*does*	*if*	*nor*	*seen*	*thus*
almost	*done*	*in*	*not*	*several*	*to*
also	*due*	*into*	*now*	*should*	*under*
although	*during*	*is*	*obtain*	*show*	*up*
always	*each*	*it*	*obtained*	*showed*	*upon*
among	*either*	*its*	*of*	*shown*	*use*
an	*enough*	*itself*	*often*	*shows*	*used*
and	*especially*	*just*	*on*	*significant*	*using*
another	*etc*	*kg*	*only*	*significantly*	*various*
any	*followed*	*km*	*or*	*since*	*very*
approximately	*following*	*largely*	*other*	*so*	*was*
are	*for*	*like*	*our*	*some*	*we*
as	*found*	*made*	*out*	*such*	*were*
at	*from*	*mainly*	*over*	*suggest*	*what*
be	*furth*	*make*	*overall*	*than*	*when*
because	*er*	*may*	*per*	*that*	*whereas*
been	*give*	*might*	*perhaps*	*the*	*which*
before	*given*	*min*	*possible*	*their*	*while*
being	*giving*	*ml*	*previously*	*theirs*	*with*
between	*had*	*mm*	*quite*	*them*	*within*
both	*hardly*	*more*	*rather*	*then*	*without*
but	*has*	*most*	*really*	*there*	*would*
by	*have*	*mostly*	*regarding*	*these*	

文本自动分类的研究与实现

作者：石敏
学位授予单位：海南大学