

北京化工大学

硕士学位论文

中文文本分类系统的研究与实现

姓名：甘立国

申请学位级别：硕士

专业：计算机应用技术

指导教师：董小国

20060606

中文文本分类系统的研究与实现

摘 要

随着信息技术的迅速发展，特别是 Internet 的普及，网页数量呈海量增长。由于网页中的内容大部分是文本信息，因此如何根据网页中的文本信息自动分类成为目前研究的重要课题。文本自动分类是信息检索中的一个重要环节，它是指在给定的分类体系下，根据文本的内容自动判定文本类别的过程，以便于信息的检索。本文首先介绍了文本自动分类在国内外的研究现状；其次对文本自动分类所涉及的关键技术，包括信息检索模型、中文分词方法、特征抽取、特征项权重方法以及关键的分类算法，分别进行了研究和探索；再次在特征项权重方面，我们分析了传统特征项权重方法的缺点，提出使用句子的重要度对特征项的权重进行加权，实验证明这种方法能有效地反映文本的内容；接下来介绍了基于向量空间模型的中文文本分类系统的总体框架，系统流程和功能模块；最后对分类系统中实现的各种特征抽取算法、权重算法和分类算法分别进行了实验对比。

关键词：文本分类，向量空间模型，特征抽取，特征项权重

The Research and implementation of Chinese text categorization system

ABSTRACT

With the development of Information technology and the prevalence of Internet, the amount of web page increase explosively. Because the content of web page is mostly text, how to categorize web page automatically by its text information became an important research subject. Text categorization, the automated assigning of natural language texts to predefined categories based on their contents, is an important part of Information retrieval. This paper firstly introduce the research status of text categorization, secondly we study and discuss the key technique of text categorization, including Information retrieval model, Chinese word segment, Feature Selection, Feature Weight and Classify Methods. Considering the disadvantage of tradition Feature Weight, we use sentence's importance to compute feature's weight and experiment prove that this method is good for Categorization. Thirdly, we introduce the frame, system flow and function module of Chinese text categorization system based on vector space model. Finally, we list the result of experiment on feature selection, feature weight and classify

method.

KEY WORDS: text categorization, vector space model, Feature Selection, Feature Weight

北京化工大学学位论文原创性声明

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。本人完全意识到本声明的法律结果由本人承担。

作者签名： 杨立强 日期： 2006-06-06

关于论文使用授权的说明

学位论文作者完全了解北京化工大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京化工大学。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磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。

保密论文注释：本学位论文属于保密范围，在 2 年解密后适用本授权书。非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名： 杨立强 日期： 2006-06-06
导师签名： 董士国 日期： 2006-06-06

第一章 绪论

1.1 文本分类的研究背景和意义

在当今的信息社会,随着Internet网的应用不断普及深入,使人们从信息缺乏的时代过渡到了信息极大丰富的时代。当今社会的信息突出表现为:信息量急剧增加,各种电子文本形式的情报源所提供的信息量正以惊人的速度递增;信息结构更加复杂,WWW网上包含的信息以文本、图像、视频等多媒体格式存在;信息的全球化,要求处理与传递信息的速度加快。面对Internet上日益膨胀的信息,如何快速、准确地从浩瀚的信息资源中寻找出所要的狭小领域内的相关内容就成了一项十分有意义的课题。正是在这样的背景之下,基于机器学习^[1](Machine learning)的文本分类^[2-4](Text Categorization)正逐渐成为一个日益重要的研究领域。特别是Internet上在线信息的增加,文本分类显得越来越重要。它对于网上资源有效共享、提高工作效率、更进一步地普及Internet等网络通信都具有极其现实的意义。由于分类可以在较大程度上解决目前网上信息杂乱的现象,方便用户准确地定位所需的信息和分流信息。因此,文本分类已成为一项具有较大实用价值的关键技术,是组织和管理数据的有力手段。近年来,文本分类技术已经逐渐与搜索引擎^[5](Search engine)、信息推送^[6-7](Information push)、信息过滤^[8-9]

(Information Filtering)等信息处理技术相结合,有效地提高了信息服务的质量。文本分类同时还可被用于抽取符号知识、新闻分发、排序电子邮件以及学习用户兴趣。它是信息检索^[10]、机器翻译、自动文摘、信息过滤等技术的基础。

一个优秀的检索系统必须建立在良好的文本分类上,许多WWW索引系统(如Yahoo)在对下载的Web文本进行索引前,需要对文本分类处理,以便于用户的查找和提高检索的性能和效率。其实这点很容易理解,如果被检索的文本已经分类,而后在与用户要求相关的文本类别内进行检索,则可大大降低检索空间,从而提高检索的速度和检索系统的性能。传统的文本分类建立在手工分类的基础之上的,这种手工分类的做法存在着许多弊端:一是耗费大量的人力,物力和精力。二是存在分类结果与要求的不一致。即使分类人的语言素质较高,对于不同的人来分类,其分类结果仍然不尽相同。甚至同一个人,在不同时间做分类也可能会来分类,其分类结果仍然不尽相同。甚至同一个人,在不同时间做分类也可能会

有不同的结果。手工分类由于效率太低而面临越来越多的困难,为了提高文本分类的速度和精度,基于机器学习的文本分类自然地变成了发展方向。

文本分类其实也是一种文献检索的手段,与普通的文本检索不同的是文本分类技术预先设定了一个类别集合,对检索的文本,根据一定的判别法则,判断其是否属于这个集合中的某个类。文本分类技术在文本检索、信息过滤、数据组织、符号知识抽取、新闻分发、排序电子邮件、学习用户兴趣等方面都具有相当的实际应用价值。例如,(1)随着Internet的快速普及,人们越来越多的通过互联网络查找各种文献资料,诸如书刊、论文、科研资料、会议记录等等。但是使用过互联网络的人都知道,要想在网上找到自己需要的资料不是一件容易的事,即使借助于专门的搜索引擎(目前大部分的搜索引擎是按关键词搜索的),信息检索的精度也往往不能够令人满意,检索得到的结果经常包含大量无关的资料。笔者曾经网上检索关键词为“文本分类”的文献,其检索的结果令人感到尴尬:有数百篇文献被检索到,而事实上真正符合笔者的要求的文献只有数十篇。其原因在于检索文献的范围太宽,只要含有关键词“文本分类”的文献均被检索出,但笔者实际想检索的文献是那些与“文本分类”这一主题相关的文献。显然,如果能够使用文本分类技术对检索结果进行过滤,剔除掉无关的文献,无疑将有效提高网上文献检索的精度。(2)图书馆的文本资料管理,也需要对文本进行分类。众所周知,图书馆的整个索引系统都是建立在分类的基础之上的。(3)此外,文本分类对于语料库语言学的发展也将有很大的推动作用。随着语料库语言学的发展要求语料库的规模越来越大;同时,电子出版业的迅速发展也使得获取大量的电子文本建立大规模语料库已成为可能。但语料处理的速度却相对落后于语料收集的速度。因为收集来的粗语料通常是杂乱无章的,在加工整理前必须进行分类处理,而目前对粗语料的分类处理过程仍然是以手工为主,不但效率低,而且对从事分类的工作人员水平要求较高。如果能够代之以自动分类,无疑将大大加快语料处理的速度。

1.2 文本分类的定义

文本分类就是对一篇文本,根据其内容,从预先定义好的标记集中找出一个

或者多个最适合于该文本的标记。文本分类技术从开始出现到现在,经历了从基于规则到基于统计分类,再到规则和统计相结合的一个过程。

简单地说,文本分类系统的任务是:在给定的分类体系下,根据文本的内容自动地确定文本关联的类别。从数学角度来看,文本分类是一个映射的过程,它将未标明类别的文本映射到已有的类别中,该映射可以是一一映射,也可以是一对多的映射,因为通常一篇文本可以同多个类别相关联。用数学公式表示如下:

$$f:A \rightarrow B$$

其中, A 为待分类的文本集合, B 为分类体系中的类别集合, f 是建立 A 和 B 映射的函数或模型。

文本分类的映射规则是系统根据已经掌握的每类若干样本的数据信息,总结出分类的规律性而建立的判别公式和判别规则。然后在遇到新文本时,根据判别规则,确定文本相关的类别。

1.3 文本分类技术的研究现状

1.3.1 文本分类技术的分类

按照人们解决文本分类问题的切入点的不同,文本分类可分为基于自然语言理解的文本分类和基于统计数学的文本分类。当然在实际应用中使用的方法大都是这两种方法的结合。文本分类的处理对象是文本,而文本是以自然语言的形式出现的,因此,试图把文本分类建立在自然语言理解的基础之上是很自然的事情。然而事实并非如此,由于自然语言的复杂性,想把自然语言中的一切规则用知识的形式表示出来几乎不可能,这条途径事实上困难是相当大的。因此,一方面,人们继续跟随着自然语言处理技术的发展,试图建立更为理想的基于知识库的分类系统;另一方面,人们另辟蹊径开拓了一条新的文本分类的途径,即把文本分类建立在统计数学的基础之上,用统计的方法从文本的字频、词频等相关元素中提取文本的特征,再建立相应的数学模型以实现分类。尽管此类方法的分类效果不太理想,但作为一种辅助工具,仍然具有积极的意义。

按文本语料的性质和应用需求的不同,文本分类可分为基于分类体系的分类和基于信息过滤和用户兴趣的分类。基于分类体系的分类一般要经

过抽取主题词, 计算权值, 根据分类体系对主题词分析定类几个步骤, 目前国内对分类的研究多是基于分类体系的系统。网上信息多如牛毛, 人们没有必要而且也没有足够的时间和精力去阅读所有的信息, 因此, 越来越多的用户只希望看到自己感兴趣的信息, 基于信息过滤的分类显得越来越重。

1.3.2 文本分类在国外的发展

国外对文本分类的研究始于20世纪50年代末, H.P.Luhn首先将词频统计思想用于分类, 在该领域进行了开创性的研究。1960年, Maron在Journal of ASM上发表了有关自动分类的第一篇论文《On relevance, probabilistic indexing and information retrieval》, 其后许多学者在这一领域进行了卓有成效的研究工作。

从20世纪60年代直到20世纪80年代末, 这期间最有效的文本分类系统一直是由专家人工构建的基于知识工程技术的分类系统。其典型应用就是卡内基集团为路透社开发的Construe系统^[11], 它主要是由专业人员编写了一些分类规则来指导分类, 在Reuters的部分语料库上它的效果非常好, 平均准确率和召回率大约都可达到90%, 但是在其他的应用领域采用Construe系统将会消耗大量的人力和物力。这种自动分类器构造方法的缺点是知识获取瓶颈的存在。它必须要为领域专家获取的知识和知识工程师的知识表示之间架起桥梁, 二者缺一不可, 如果这种分类器被转到完全不同的领域, 工作必须得重新开始。90年代初期, 基于机器学习的分类技术开始取代基于知识工程的方法成为文本分类的主流技术。这种算法通过归纳文本集的特征自动创建一个分类器, 这些文本集合事先被领域专家人工地分类到类集 $C = (c_1, c_2, \dots, c_m)$ 的各个类 c_i 中, 分类器可作为一个规则决定文本 d_j 是否属于类 c_i 。如果类集C被更新, 或者系统要应用于其他不同的领域, 只需要重新构造一个人工分类文本集合, 通过机器学习, 自动地构造一个分类器。显然由于这种分类方法不再需要知识工程师和领域专家的介入, 节约了大量的专家人力资源, 同时加快了分类系统的建立速度。

近年来, 研究者们机器学习的技术进行了大胆的探讨, 提出了多种分类模型和分类算法, 如基于向量空间模型的Rocchio分类算法^[12]及其一系列的改进算法, K近邻法(KNN)^[13-16], 决策树(Decision Tree)^[17], 朴素贝叶斯(Naive Bayes)^[18-19],

神经网络 (Neural network)^[20], 支持向量机(Support Vector Machine)^[21-25]等等这些方法在英文以及欧洲语种文本分类上有广泛的研究, 均取得了不错的效果。国外很多研究人员对英文文本分类领域的各个问题都有相当深入的研究, 对几种流行的方法进行了大量的对比研究。还有一些研究人员研究表明结合不同的分类器能够提高分类的精度。

目前, 国外的分类系统已经从最初的可行性研究经历了实验研究进入了实用化阶段。并在邮件分类、电子会议、信息过滤等方面得到了较为广泛的应用。

1994年, AT&T实验室的David D. Lewis等人对基于非确定性的分类技术做了研究。两年后, 该实验室将分类的技术应用于了电子邮件领域。1997年, 德国Dortmund大学计算机系的Torsten Joachims等人研究了基于向量空间模型的自动分类系统。同年, 美国Stanford大学计算机系的Daphne Koller等人提出了基于很少语料词汇的层次自动分类方法。1998年, 美国Carnegie Mellon大学计算机系的Yiming Yang等人将决策树等聚类算法应用于在线自动分类。1999年, 美国Just Research公司的Andrew McCalum等人运用信息熵理论、Bayes理论等实现了多类号的自动分类。随后, 美国Massachusetts大学计算机系专门针对文本库开发了自动分类系统, 美国IBM和Oracle公司为推广电子商务而研制了基于文本内容的电子邮件自动分类系统, Microsoft公司也为其浏览器开发了基于内容属性分类的插件。

1.3.3 文本分类在国内的发展

自从文本分类的概念在国内出现以来, 该技术在国内得到了长足的发展。然而和国外的发展状况相比, 发展水平仍相对滞后。一方面由于国内起步较晚, 另一方面则由于国内的工作主要是针对中文文本。由于汉语有许多不同于英语的特点, 使得中文文本分类的难度更大。比如, 汉语的书面形式是连续书写的, 词与词之间没有自然的界限, 在进行文本分类之前, 首先要对文本进行分词。另外, 在不同的语言的研究工作中, 句法分析和语义分析所占的比例是不同的。在英语中, 句法分析比语义分析的比例要大, 而汉语是一种分析型语言, 语义分析在汉语研究中起着举足轻重的作用, 其所占的比例比句法分析要大得多。这使得在中文文本分类中, 通过句法分析等基于语法的手段把握文本的内容变得更加困难。就

发展历史而言,国内的文本分类的发展经历了三个阶段:国外研究成果引进阶段、分类技术完善阶段以及面向汉语分类技术的发展阶段;而就发展方向来看,则有基于外延的分类方法和基于概念的分类方法之分。中国科学院、清华大学、上海交通大学、复旦大学、南京大学、一些大学的著名学者在该领域做出了一些研究成果,研制出一批基于词典法和基于专家系统的分类系统。由于中文与英文存在较大的差异,不能照搬国外的研究成果,中文文本分类的研究基本上是在英文文本分类的研究策略上,结合中文文本的特点,继而形成中文文本分类研究体系。

1.4 本文工作

本课题的主要研究内容是:(1)在熟悉中文信息处理、文本分类技术的基础上,实现了一个基于向量空间模型、多类别的中文文本分类系统,为研究文本分类的相关技术提供了实验平台。(2)同时对文本分类技术进行了深入的研究和分析,使用句子的重要度计算特征项的权重。(3)此外还完成了对各主要算法的评测,以此来分析总结它们不同的特点及使用场合。得到的结论不仅有利于我们深入的理解文本分类技术,从而更为合理的应用它们,也将作为研究新算法的基础。

1.5 本文安排

论文的组织如下:

第一章首先介绍了论文的研究背景,接着介绍当前分类技术的研究现状,给出本文的主要内容和组织结构。

第二章介绍信息检索模型及相关技术,并提出了利用句子的重要度计算特征项的权重。

第三章介绍目前常用的分类算法。

第四章将介绍一个相对完整的文本分类系统的实现过程,以及系统的功能、运作方式等等。该系统将作为我们的实验平台,帮助我们完成文本分类技术的测试研究工作。通过对该系统地讲解,有助于深入了解文本分类的运作模式。

第五章列出我们取得的各种实验结果,包括各种特征抽取算法、权重算法和

分类算法的测试结果。通过对实验结果的分析，我们总结了各种算法的优缺点。

第六章对课题所作的工作做了简要总结，并提出下一步工作的展望。

第二章 文本分类相关技术研究

2.1 信息检索模型

信息检索是指将信息按一定的方式组织和存储起来,并根据用户的需要找出相关信息的过程和技术。由于网上大部分信息为文本信息,因此我们这里只考虑文本信息检索。文本分类与文本检索是相辅相成、密不可分的,一方面文本分类技术的提高使得文本检索的精确度和速度得以提高,另一方面文本分类是建立在文本检索的基础之上的,因为它借鉴了许多检索的表示方法和技术。文本检索技术的发展已经有 30 年的历史,取得了很大的成就,产生大批实用的检索系统,积累了许多成熟的技术,这对文本分类技术的形成和发展起了很大推进作用。根据文本的表示方法和检索的实现方式,目前主要有三种信息检索模型^[26]: (1)布尔模型(Boolean Model) (2)向量空间模型(Vector Space Model, 简称 VSM) (3)概率模型(Probabilistic Model)。

在传统的信息检索模型中,认为每个文本是由一组具有代表性的关键字或词来描述,这些关键字或词被称为特征项。对于某个文本中的一组特征项来说,它们在描述文本内容方面的作用是不相同的。因此,决定一个特征项对文本内容描述的贡献程度是一个十分重要的问题。这个问题可以利用一些容易度量和评估的属性,来评价特征项对文本内容的描述重要程度。例如,我们考虑一个具有数十万份文本的文本集。如果一个词在每份文本中都出现,那么用它作为特征项就没有意义,因为它不能描述每份文本之间的差别,也就不能告诉用户哪份文本是他感兴趣的。另一方面,如果一个词语仅仅出现在其中的某几份文本中,那么利用它作为特征项就非常适合,因为它所能描述的文本空间相当窄,能确切地告诉用户这份文本是否是他所关心的。因此,用来描述文本内容的特征项应该是与文本内容密切相关的词语。我们可以为文本中的特征项定义一个权重来描述这种相关程度。假设 d_i 表示一个文本, t_j 表示一个特征项, w_{ij} 是文本 d_i 中特征项 t_j 的权重

2.1.1 布尔模型

布尔模型(Boolean Model)是基于集合论和布尔代数的一种简单检索模型,用布尔表达式表示用户的查询串,查询串通常以语义精确的布尔表达式^[27]的方式输入,如 $q = t_1 \wedge (t_2 \vee \neg t_3)$,通过对文本与查询串的逻辑比较进行搜索,是一种简单常用的严格匹配模型。不过,布尔模型存在着一些缺陷:

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

2、虽然布尔表达式具有精确的语义,但常常很难将用户的信息需求转换为布尔表达式,实际上大多数查询用户发现在把他们需要的查询信息转换为布尔表达式时并不那么容易。

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

$$Sim(d_i, q) = \begin{cases} 1, q \in d_i \\ 0, q \notin d_i \end{cases} \quad (2-1)$$

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

Boolean 模型的主要优点在于具有清楚和简单的形式、速度快,而主要的缺陷在于完全匹配会导致太多或太少的结果文本被返回。针对特征项权重的选择,引出了向量空间模型。

2.1.2 向量空间模型

向量空间模型(Vector Space Model, 简称 VSM),是由 G.Salton^[28]等人在 20 世纪 60 年代提出的,是效果较好、近些年来得被广泛应用的一种方法。在向量空间模型中,文本 d_i 被看作为由一组特征项 (t_1, t_2, \dots, t_n) 组成的 n 维向量空间,文本 d_i 简化为以特征项的权重为分量的向量表示 $(w_{i1}, w_{i2}, \dots, w_{in})$,权重 w_{ij} 表示

特征项 t_j 对文本 d_i 分类的贡献程度, 取值范围是 $[0, 1]$ 。查询串 q 同样可以表示成向量 $(w_{q1}, w_{q2}, \dots, w_{qm})$ 。在对所有文本和查询串 q 进行向量化之后, 检索过程从而简化为空间向量的运算, 文本信息的匹配问题转化为向量空间中的向量匹配问题, 大大减小了问题的复杂性。文本与查询串的相关度一般用两个向量的夹角余弦值来计算, 其公式如下:

$$\text{sim}(d_i, q) = \frac{W_i \cdot W_q}{\sqrt{\sum_{j=1}^n w_{ij}^2} \sqrt{\sum_{j=1}^n w_{qj}^2}} \quad (2-2)$$

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

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

- 1、特征项权重的算法提高了检索的性能;
- 2、部分匹配的策略使得过滤得到的结果文本集合更接近用户的查询需求;
- 3、根据结果文本对于查询串的相似度通过余弦公式对结果文本进行排序。

2.1.3 概率模型

概率模型是基于以下的基本假设 (概率原则): 给定一个用户查询 q 和文本集

中的一个文本 d_i ，概率模型试图估计用户找到其感兴趣的文本 d_i 的概率，模型假设这个相关概率只是依赖于查询和文本的表示。

对于概率模型，特征项的权重都是二值的，即 $w_{ij} \in \{0,1\}$ ， $w_{iq} \in \{0,1\}$ 。设 R 是已知的相关文本集， \bar{R} 是 R 的补集，即已知的不相关文本集。设 $P(R|W_i)$ 是文本 d_i 与查询 q 相关的概率， $P(\bar{R}|W_i)$ 是文本 d_i 与查询 q 不相关的概率。文本 d_i 与查询 q 的相关度 $sim(d_i, q)$ 可以定义为一个比值

$$sim(d_i, q) = \frac{P(R|W_i)}{P(\bar{R}|W_i)} \quad (2-3)$$

使用贝叶斯定理，比值写成

$$sim(d_i, q) = \frac{P(W_i|R) \times P(R)}{P(W_i|\bar{R}) \times P(\bar{R})} \quad (2-4)$$

其中， $P(W_i|R)$ 代表从相关文本集 R 中随机选择文本 d_i 的概率。 $P(R)$ 代表从整个文本集中随机选择一个文本是相关的概率。 $P(W_i|\bar{R})$ 代表类似的从补集中选择文本的概率， $P(\bar{R})$ 代表从整个文本集中随机选择一个文本是不相关的概率。因为对文本集中的所有文本来说， $P(R)$ 和 $P(\bar{R})$ 都是一样的，于是相关度可以写为

$$sim(d_i, q) \sim \frac{P(W_i|R)}{P(W_i|\bar{R})} \quad (2-5)$$

假设特征项是独立的，那么

$$sim(d_i, q) \sim \frac{(\prod P(t_j|R)) \times (\prod P(\bar{t}_j|R))}{(\prod P(t_j|\bar{R})) \times (\prod P(\bar{t}_j|\bar{R}))} \quad (2-6)$$

其中， $P(t_j|R)$ 是特征项 t_j 在 R 集合中某个文本中出现的概率， $P(\bar{t}_j|R)$ 是特征项 t_j 不在在 R 集合中某个文本中出现的概率。对于集合 \bar{R} ，相似的概率具有相似的含义。而

$$P(t_j|R) + P(\bar{t}_j|R) = 1 \quad (2-7)$$

$$P(t_j|\bar{R}) + P(\bar{t}_j|\bar{R}) = 1 \quad (2-8)$$

对式(2-6)取对数, 有

$$\text{sim}(d_i, q) \sim \sum_{j=1}^n w_{q,j} \times w_{i,j} \times [\log \frac{P(t_j | R)}{1 - P(t_j | R)} + \log \frac{1 - P(t_j | \bar{R})}{P(t_j | \bar{R})}] \quad (2-9)$$

在公式 i , 关键是要知道概率 $P(t_j | R)$ 和 $P(t_j | \bar{R})$ 值。概率模型就是采用相关反馈的方法, 从假设两个初始的概率开始, 不断调整概率估计值, 直到得到一个满意的概率排序。

概率模型的主要优点是: 从理论上, 文本按照其相关概率的降序排列。其主要缺点包括: 需要最初将文本分为相关和不相关的集合; 这个方法不考虑特征项在文本出现的次数, 即所有的权重都是二值的。

由于空间向量模型的简单性和高效性, 本文的分类系统采用向量空间模型。

2.2 中文分词

我们进行分类的文本一般都是中文文本。与英文文本不同, 英文文本利用空格作为词的分隔符, 而中文文本一般是无分隔符的字符串, 词与词之间没有分隔标志。为了对文本进行表示, 我们需要对中文文本进行分词^[29]。

分词就是将连续的字序列按照一定的算法划分成词序列的过程。在英文文本中, 单词之间是以空格作为自然分隔符的, 而中文只是字、句和段可以通过简单的分界符来划界, 唯独词没有一个形式上的分隔符, 虽然英文存在短语的划分问题, 但是在词这一层上, 中文比英文要复杂得多, 困难得多。

分词是汉语自然语言处理的第一步。目前, 汉语自然语言处理的应用系统处理对象越来越多的是大规模语料, 因此分词的速度和分词算法的易实现性变得相当关键。目前研究中文分词的大多是科研院校, 清华、北大、中科院东北大学、IBM研究院、微软亚洲研究院等都有自己的研究队伍。

2.2.1 中文分词技术

现有的分词算法分为三类: 基于字符串匹配的分词算法、基于统计的分词算法和基于理解的分词算法。

1、基于字符串匹配的分词算法

它是按照一定的策略将待分析的汉字串与一个“充分大的”机器中的词条进行匹配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。按照扫描方向的不同,串匹配分词方法可以分为正向匹配和逆向匹配;按照不同长度优先匹配的情况,可以分为最大(最长)匹配和最小(最短)匹配;按照是否与词性标注过程相结合,又可分为单纯分词方法和分词与标注相结合的一体化方法。

常用的机械分词方法有:正向最大匹配、逆向最大匹配和双向匹配。也可以是上述各方法的相结合。

2、基于理解的分词算法

基于理解的分词方法就是在分词的同时进行句法、语义分析、利用句法信息和语义信息来处理歧义现象。它通常包括三个部分:分词子系统、句法语义子系统、总控部分。在总控部分的协调下,分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断,即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。

3、基于统计的分词算法

从形式上看,词是稳定的字的组合,因此在上下文中,相邻的字同时出现的次数越多,就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计,计算它们的互现信息。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时,便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进行统计,不需要切分词典,因而又叫无词典分词法。

2.2.2 分词中的难题

有了成熟的分词算法,是否就能容易的解决中文分词的问题呢?事实远非如此。中文是一种十分复杂的语言,让计算机理解中文语言更是困难。在中文分词过程中,有两大难题一直没有完全突破。

1、歧义识别

歧义是指同样的一句话,可能有两种或者更多的切分方法。例如:表面的,

因为“表面”和“面的”都是词，那么这个短语就可以分成“表面的”和“表面的”。这种称为交叉歧义。像这种交叉歧义十分常见，前由于没有人的知识去理解，计算机很难知道到底哪个方案正确。

交叉歧义相对组合歧义来说是还算比较容易处理，组合歧义就必需根据整个句子来判断了。例如，在句子“这个门把手坏了”中，“把手”是个词，但在句子“请把手拿开”中，“把手”就不是一个词；在句子“将军任命了一名中将”中，“中将”是个词，但在句子“产量三年中将增长两倍”中，“中将”就不再是词。这些词计算机又如何去识别？

如果交叉歧义和组合歧义计算机都能解决的话，在歧义中还有一个难题，是真歧义。真歧义意思是给出一句话，由人去判断也不知道哪个应该是词，哪个应该不是词。例如：“乒乓球拍卖完了”，可以切分成“乒乓球拍卖完了”、也可切分成“乒乓球拍卖完了”，如果没有上下文其他的句子，恐怕谁也不知道“拍卖”在这里算不算一个词。

2、未登录词识别^[30]

未登录词就是那些在字典中都没有收录过，但又确实能称为词的那些词。最典型的是人名，人可以很容易理解句子“王军虎去广州了”中，“王军虎”是个词，因为是一个人的名字，但要是让计算机去识别就困难了。如果把“王军虎”做为一个词收录到字典中去，全世界有那么多名字，而且每时每刻都有新增的人名，收录这些人本身就是一项巨大的工程。即使这项工作可以完成，还是会存在问题，例如：在句子“王军虎头虎脑的”中，“王军虎”还能不能算词？新词中除了人名以外，还有机构名、地名、产品名、商标名、简称、省略语等都是很难处理的问题，而且这些又正好是人们经常使用的词，因此对于搜索引擎来说，分词系统中的新词识别十分重要。目前新词识别准确率已经成为评价一个分词系统好坏的重要标志之一。

2.3 特征项抽取

文本分类问题的最大特点和困难是特征空间的高维性和文本表示向量的稀疏性。在汉语处理中，通常采用词条作为最小的独立语义载体，原始的特征空间

由可能出现在文章中的全部词条构成。而一个中等规模的语料库就常常可以包含上十万条不同的词，这样高维的特征空间对于几乎所有的分类算法来说都偏大。在不降低分类器的准确性的前提下寻求一种自动高效的特征抽取方法，降低特征空间的维数，提高分类器的效率，成为文本分类中需要面对的重要问题。

近年来在中文文本分类中经常采用的特征抽取^[31-33]方法包括最简单的停用词移除、互信息MI、信息增益IG和 χ^2 统计等。特征抽取方法的选取主要依据Y.Yang的实验结果。由于中文文本分类问题与英文文本分类相比具有相当大的差别，体现在原始特征空间的维数更大，文章表示更加稀疏，词性变化更加灵活等多个方面，也更困难。因此，在英文文本分类中表现良好的特征抽取方法未必适合中文文本分类。对中文文本分类中的特征抽取方法进行系统的比较研究显得十分重要。

1、文档频率

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

文档频率是最简单的特征抽取技术，由于其具有相对于训练语料规模的线性计算复杂度，它能够容易地被用于大规模语料统计。但是在信息抽取(Information Retrieval)研究中却通常认为DF值低的词条相对于DF值高的词条具有较多的信息量，将这些词条从特征空间中移除会降低分类器的准确率。

Y. Yang否定了这一说法，同时还指出当IG和CHI等特征抽取方法的计算“费用”太高而变得不可用时，DF可以安全的代替它们被使用。

2、信息增益

信息增益(Information Gain)在机器学习领域被广泛使用。在文本特征抽取中，对于词条 t 和类别 c ，IG考察 c 中出现和不出现 t 的文档频数来衡量 t 对于 c 的信息增益。信息增益的计算公式如下：

$$IG(t) = - \sum_{i=1}^m P(c_i) \log P(c_i) + P(t) \sum_{i=1}^m P(c_i | t) \log P(c_i | t) + P(\bar{t}) \sum_{i=1}^m P(c_i | \bar{t}) \log P(c_i | \bar{t}) \quad (2-10)$$

其中 $P(c_i)$ 表示 c_i 类文档在语料中出现的概率, $P(t)$ 表示语料中包含词条 t 的文档的概率, $P(c_i | t)$ 表示文档包含词条 t 时属于 c_i 类的条件概率, $P(\bar{t})$ 表示语料中不包含词条 t 的文档的概率, $P(c_i | \bar{t})$ 表示文档不包含词条 t 时属于 c_i 类的条件概率, m 表示类别数。

3、CHI统计

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

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (2-11)$$

对于存在多个类别的应用, 分别计算 t 对于每个类别的CHI值, 再用下式计算词条 t 对于整个语料的CHI值:

$$\chi_{\max}^2(t) = \max_{i=1}^m \chi^2(t, c_i) \quad (2-12)$$

其中 m 为类别数。

4、互信息

互信息 (Mutual Information) 在统计语言模型中被广泛采用。词条 t 和文档类别 c 的互信息定义为:

$$MI(t, c) = \log \frac{P(tc)}{P(t) \times P(c)} \quad (2-13)$$

其中 $P(tc)$ 表示语料中属于 c 且包含 t 的文档频数, $P(t)$ 表示语料中包含词条 t 的文档的概率, 表示语料中 $P(c)$ 类文档出现的概率。如果用 A 表示包含 t 且属于 c

的文档频数, B 为包含 t 但是不属于 c 的文档频数, C 表示属于 c 但是不包含 t 的文档频数, N 表示语料中文档总数, 则可用下式来近似 t 与 c 之间的互信息:

$$MI(t, c) = \log \frac{P(tc)}{P(t) \times P(c)} \quad (2-14)$$

如果 t 和 c 无关 (即 $P(tc) = P(t) \times P(c)$), $MI(t, c)$ 值自然为零。为了将互信息应用于多个类别, 与 CHI 统计的处理类似, 我们首先对各个词条计算它与每个类别的互信息, 再用下式计算它对整个语料的互信息。

$$I_{\max}(t) = \max_{i=1}^m I(t, c_i) \quad (2-15)$$

5、期望交叉熵 (ECE, Expected Cross Entropy)

$$ECE(t) = P(t) \sum_i P(c_i | t) \log \frac{P(c_i | t)}{P(c_i)} \quad (2-16)$$

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

2.4 特征项权重算法

2.4.1 传统的权重算法

词是组成文本的基本元素。在所有的词中抽取出能够表征文本特征的词组成文本的特征项, 并按某一方法赋予特征项相应的权重。特征项的权重综合反映了该特征项对标识文本内容的贡献度和文本内容之间的区分能力。特征项在不同文本中出现的频率满足一定的统计规律, 因此可以通过特征项的频率特性计算其权重^[34]。一个有效的特征项集合必须具有以下两个特征:

- 1、完全性: 特征项能够反映目标文本的内容;
- 2、区分性: 特征项具有将目标文本和其他文本相区分的能力。

根据以上两个特征, 特征项权重的计算满足以下两个原则: 一是正比于特征项在文本中出现的频率; 二是反比于文本集中出现该特征项的文档频率。

常见的特征项权重算法有以下几种:

1、布尔函数

$$w_i = \begin{cases} 1, TF_i \geq 1 \\ 0, TF_i = 0 \end{cases} \quad (2-17)$$

2、开根号函数

$$w_i = \sqrt{TF_i} \quad (2-18)$$

3、TFIDF 函数

$$w_i = \frac{TF_i \times \log(N/DF_i)}{\sqrt{\sum_{i=1}^n [TF_i \times \log(N/DF_i)]^2}} \quad (2-19)$$

其中 w_i 为第 i 个特征项在文本 d 中的权重, TF_i 是特征项 t_i 在文本 d 中出现的频数, N 表示全部训练文档的总数, DF_i 表示包含特征项 t_i 的文档频数, n 表示特征向量的维数。

2.4.2 改进的特征项权重算法

传统的权重算法只是简单的考虑特征项出现的频率, 而没有考虑到特征项在文本中出现的位置。每个句子在文本中的作用是不同的, 重要的句子反映了文本的内容并且可以和其他文本相区分。既然句子对于分类效果有很大影响, 如果在计算特征项权重时考虑特征项所属句子的重要度^[35]会提高分类的准确度。目前主要有两种方法在计算特征项权重时考虑了句子的重要度。第一种方法认为在文本每个段落的第一句或者最后一句话具有较高的重要度, 这种方法只适合于结构化或半结构化的文本。第二种方法认为标题是文本主要内容的总结, 因此具有较高的重要度, 缺点是如果标题不能反映文本的内容, 会增加分类的模糊性。为了解决以上问题, 我们采用了文本统计的方式计算句子的重要度, 一是基于标题计算句子的重要度, 二是基于句子中各个特征项的统计特性来计算句子的重要度。最后结合这两种统计方式, 计算句子总的重要度。

1、基于标题的句子重要度计算

文本的标题一般归纳了文本的主要内容, 出现在标题中的特征项应该具有更高的权重。这种方法的有效性取决于标题质量的好坏。在多数情况下, 文本

的标题并不能很好反映文本的内容，因此我们通过句子与标题的相似度来对计算特征项的权重，避免直接使用标题的重要度。例如“怎么解决这个问题”，这个标题并不能反映文本的任何内容，但是文本中带有“问题”这个特征项的句子应该具有更高的重要度，因为这个句子有可能包含关于“问题”的关键特征项。我们通过计算句子与标题的相似度，与标题具有高相似度的句子具有更高的重要度。标题与句子分别表示为特征项的向量，向量间的夹角越小，标题与句子的相似度越高。句子与标题的相似度的计算公式如下：

$$Sim(S_i, T) = \frac{S_i T}{\|S_i\| \|T\|} \quad (2-20)$$

T 代表标题的向量， S_i 代表句子的向量。

2、基于特征项的句子重要度计算

使用句子相似度计算权重的方法取决于标题质量的好坏，当文本中的标题没有意义或者文本没有标题时，这种方法的作用就很小。如果句子与标题具有很高的相似度，但是并没有包含区分性好的特征项，那么这个句子中对于分类的影响就很小；反之，如果句子与标题的相似度小，但是包含了类区分性好的特征项，这样的句子应具有高的重要度。考虑到这些因素，我们先衡量句子中每个特征项重要性，然后根据特征项计算句子的重要度。其计算方法是首先统计特征项的 TF 和 IDF 值，然后把句子中每个特征项的 TF 和 IDF 的乘积值进行加权并进行归一化，得到的结果就是句子的重要度。

$$Imp(S_i) = \frac{\sum_{t \in S_i} tf(t) \times idf(t)}{\max_{t \in T} \{\sum_{t \in T} tf(t) \times idf(t)\}} \quad (2-21)$$

其中， $tf(t)$ 代表词的频率， $idf(t)$ 代表倒转文档频率

3、句子重要度

结合上述两种重要度的计算方法，句子总的重要度计算公式如下：

$$Total(S_i) = 1.0 + Sim(S_i, T) + Imp(S_i) \quad (2-22)$$

4、特征项权重的计算

权重函数的一个基本原则是权重正比与特征项在文本中出现的频率。我们使用文本统计的方式计算了各个句子的重要度，在计算特征项的 TF 值时可以根据句子的重要度对其 TF 值进行调整，其计算公式如下：

$$WTF(d,t) = \sum_{S \in d} tf(S,t) \times Total(S) \quad (2-23)$$

其中 $tf(S,t)$ 代表特征项 t 在句子 S 中的出现频率

由以上公式可以看出，基于句子重要度特征项的 TF 值比传统的 TF 值有所增大，但是在重要句子中出现的特征项 TF 值比在次要句子中出现的特征项 TF 值增加的幅度大，这样可以区分文本中重要特征项和非重要特征项，使特征项的文本区分能力更强，分类效果更好。文本 d 中特征项 t 的权重函数采用传统的 $TFIDF$ 计算公式，其中 TF 值是根据句子的重要度进行计算而来，权重公式如下：

$$w(d,t) = \frac{WTF(d,t) \times \log(\frac{N}{DF})}{\sqrt{\sum_{i=1}^T [WTF(d,t) \times \log(\frac{N}{DF})]^2}} \quad (2-24)$$

第三章 训练和分类算法

文本分类的核心问题是如何根据语料库构造一个分类函数或分类模型，并利用此分类模型将未知类别的文本映射到指定的类别空间。目前存在多种基于向量空间模型的分类算法，例如支持向量机算法、神经网络算法、最大平均熵算法、最近K邻居算法和贝叶斯算法等等。

3.1 简单向量距离算法

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

- (1) 计算每类文本集的中心向量，计算方法为所有训练文本向量简单的算术平均；
- (2) 新文本到来后，分词，将文本表示为特征向量；
- (3) 计算新文本特征向量和每类中心向量间的相似度，公式为：

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^m w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^m w_{ik}^2)(\sum_{k=1}^m w_{jk}^2)}} \quad (3-1)$$

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

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

3.2 KNN (K最近邻居) 算法

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

具体的算法步骤如下：

- (1) 根据特征项集合重新描述训练文本向量；
- (2) 在新文本到达后，根据特征词分词新文本，确定新文本的向量表示；
- (3) 在训练文本集中选出与新文本最相似的K个文本，计算公式如式(3-1)。其中，K值的确定目前并没有很好的方法，一般采用先定一个初始值，然后根据实验测试的结果调整K值，一般初始值定为几十到几百之间；
- (4) 在新文本的K个邻居中，依次计算每类的权重，计算公式如下：

$$W(C_j) = \sum Sim(x, d_j) y(d_j, C_j) \quad (3-2)$$

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

- (5) 比较类的权重，将文本分到权重最大的那个类别中。

3.3 朴素贝叶斯算法

3.3.1 贝叶斯定理

贝叶斯方法的特点是使用概率去表示所有形式的不确定性，学习或其他形式的推理都用概率规则来实现。贝叶斯学习的结果表示为随机变量的概率分布，它可以解释为我们对不同可能性的信任程度。

设 x 是类标号未知的数据样本。设 h 为某种假定，如数据样本 x 属于某特定的类 c 。对于分类问题，我们希望能确定 $P(h|x)$ （给定观测数据样本 x ，假定 h 成立的概率）。 $P(h|x)$ 是后验概率，条件 x 下 h 的后验概率。贝叶斯定理是：

$$P(h|x) = \frac{P(x|h) \times P(h)}{P(x)} \quad (3-3)$$

其中， $P(x|h)$ 是条件 h 下， x 的后验概率。 $P(x)$ 是样本具有某些属性值的概率， $P(h)$ 是 h 的概率。

3.3.2 朴素贝叶斯分类法

基于贝叶斯定理的朴素贝叶斯分类是统计学分类方法。将训练实例分解成特征向量 x 和决策类别变量 c 。它可以预测类成员关系的可能性，如给定文本样本属于一个特定类的概率。朴素贝叶斯分类假定一个属性值对给定类的影响独立于其他属性的值。这一假定称作类条件独立。在实际应用中，这一假定以指数级降低了其复杂性。其工作过程如下：

(1) 给出训练文档集合 D ，对集合 S 中的每个训练文本样本 d ，用一个 n 维特征向量 $d = \{t_1, t_2, \dots, t_n\}$ 表示。

(2) 给定一个未知类别的文本样本 x ，分类法将归类 x 于具有最高后验概率（条件 x 下）的类。即是说，朴素贝叶斯分类将未知分类的样本分配给类 c_i ，当且仅当 $P(c_i | x) > P(c_j | x)$ ，其中 $1 \leq j \leq m, j \neq i$ 。 $P(c_i | x)$ 最大的类 c_i 称为最大后验假定。根据贝叶斯定理，得

$$P(c_i | x) = \frac{P(x | c_i) \times P(c_i)}{P(x)} \quad (3-4)$$

(3) 由于(式(3-5))中， $P(x)$ 对于所有类为常数，所以只需要 $P(x | c_i)P(c_i)$ 最大即可。如果类的先验概率 $P(c_i)$ 未知，则通常假定这些类是等概率的，即 $P(c_1) = P(c_2) = \dots = P(c_m)$ 。类的先验概率也可以用 $P(c_i) = N_i / N$ 计算，其中 N_i 是类 c_i 中的训练样本数，而 N 是训练样本总数。如果语料库中每一类 c_i 中的训练样本数相等，可以认为每一类是等概率的。

(4) 给定具有许多属性的训练集，计算 $P(x | c_i)$ 的开销可能非常大。为降低 $P(x | c_i)$ 的开销，可以做类条件独立的朴素假定。给定样本的类标号，假定属性值相互条件独立，即在属性间不存在依赖关系。这样

$$P(x | c_i) = \prod_{k=1}^n P(t_k | c_i) \quad (3-5)$$

概率 $P(t_1 | c_i)$ ， $P(t_2 | c_i)$ ， \dots ， $P(t_n | c_i)$ 可以由训练样本估值，其中 $P(t_k | c_i)$ 是 c_i 类文档中，特征 t_k 出现的条件概率。通常采用拉普拉斯概率估计

$$P(t_k, C_i) = \frac{1 + N(t_k, C_i)}{2 + N_i} \quad (3-6)$$

其中, $N(t_k, c_i)$ 是 c_i 类文本中特征 t_k 出现的文档数, N_i 为 c_i 类文档所包含的文档数目。

(5) 对未知类别的文档样本 x 分类, 对每个类 c_i , 计算 $P(x|c_i)P(c_i)$ 。文本样本 x 被指派到类 c_i , 当且仅当 $P(x|c_i)P(c_i) > P(x|c_j)P(c_j)$ 。 ($1 \leq j \leq m, j \neq i$)。于是, 文本 x 将被归类到其 $P(x|c_i)P(c_i)$ 最大的类 c_i 。

3.4 神经网络算法

神经网络模式识别方法是近几年兴起的模式识别领域的一个新的研究方向。由于神经网络的高速并行处理, 分布存储信息等特性符合人类思维系统的基本工作原则, 具有很强的自学习性、自组织性、容错性、高度非线性性、高的鲁棒性、联想记忆功能和推理意识功能等, 能够实现目前基于计算理论层次上的模式识别理论所无法完成的模式信息处理工作, 所以采用神经网络进行模式识别, 突破了传统模式识别技术的束缚, 开辟了模式识别发展的新途径。同时, 神经网络模式识别也成为神经网络最成功和最有前途的应用领域之一。在这种模型中, 分类知识被隐式地存储在连接的权值上, 使用迭代算法来确定权值向量。当网络输出判别正确时, 权值向量保持不变, 否则进行增加或降低的调整, 因此也称为奖惩法。

神经网络文本分类器可采用一种三层前馈型BP网络, 来进行自动知识获取, 如图3-5所示。BP网络有三个基本层, 即输入层、隐含层和输出层。每个层都包含若干个节点(神经元), 输入层的节点数通常为矢量的个数, 输出层节点数为输出矢量的个数。

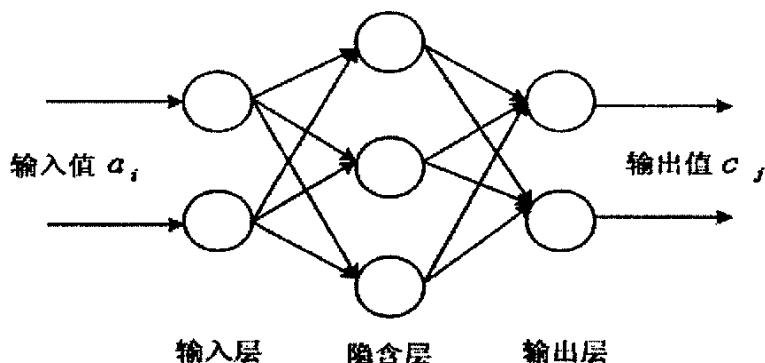


图3-1 三层前馈BP网络

给定一段文本及其特征集，输入层神经元的个数设定为特征向量的维数，输出层神经元的个数设定为类别集的大小。可以定义该神经网络的输入层第 i 个分量的输入值为：

$$a_i = w_i \text{ 文本中第 } i \text{ 个特征项的权值}$$

在训练神经网络时，定义输出层的第 j 个分量的输出值为：

$$c_j = \begin{cases} 1, & \text{文本属于类别集的第 } j \text{ 类} \\ 0, & \text{反之} \end{cases}$$

3.5 支持向量机算法

统计学习理论^[36](Statistical Learning Theory)是一种专门研究小样本情况下机器学习规律的理论。V.Vapnik等人从六七十年代开始致力于此方面研究，到90年代中期，随着其理论的不断发展和成熟，统计学习理论开始受到越来越广泛的重视。统计学习理论是建立在一套较坚实的理论基础之上的，为解决有限样本学习问题提供了一个统一的框架。同时，在这一理论基础上发展了一种新的通用学习方法-支持向量机(Support Vector Machine, 简称SVM)，它已初步表现出很多优于已有分类算法的性能。

支持向量机主要思想是建立一个超平面作为决策曲面，使得正例和反例之间的隔离边缘被最大化。如图1所示，实心点和圆圈分别代表两类样本，H为分类线， H_1, H_2 分别代表各类中离分类线最近的样本且平行于分类线H的直线，它们之间的距离称为分类间隔(margin)。所谓最优分类线就是要求分类线不但能将两

类正确分开且分类间隔最大。同理，在多维空间假定训练数据可以被一个超平面分开，如果这个向量集合能被超平面没有错误地分开，并且离超平面最近的向量与超平面之间的距离最大，则称这个向量集合被这个最优超平面(Optimal Separating Hyperplane, OSH)最大分现开。

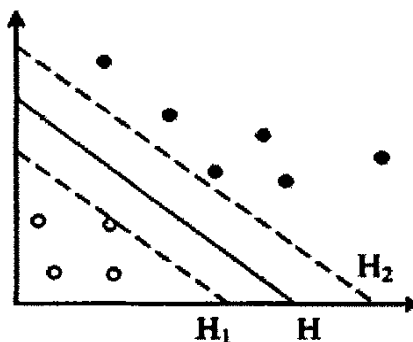


图3-2 线性可分条件下的最优分类线

给定训练数据: $(x_1, y_1), \dots, (x_l, y_l), x \in R^n, y \in \{-1, 1\}$ 可以被一个超平面 $(w \cdot x) - b = 0$ 分开。如果这个向量集合被超平面没有错误地分开，并且离超平面最近的向量与超平面的距离是最大的，则我们说这个向量集合被这个最优超平面分开。

为了描述分类超平面，我们使用下面的形式：

$$y_i[(w \cdot x_i) - b] \geq 1, i = 1, \dots, l \quad (3-7)$$

最优超平面就是满足式(3-8)并且使得 $\Phi(w) = \|w\|^2$ 最小化的超平面

要找到这个超平面，我们需要求解下面的二次规划问题：最小化泛函

$$\Phi(w) = \frac{1}{2}(w \cdot w) \quad (3-8)$$

约束条件为不等式类型：

$$y_i[(w \cdot x_i) - b] \geq 1, i = 1, \dots, l \quad (3-9)$$

这个优化问题的解由下面的拉格朗日函数的极值点给出的

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^l \alpha_i \{[(x_i \cdot w) - b]y_i - 1\} \quad (3-10)$$

其中， α_i 为拉格朗日乘子。我们需要把对拉格朗日函数关于 w, b 求其最小值和关于 $\alpha_i > 0$ 求其最大值。

在极值点上，解 w_0, b_0, α^0 必须满足以下条件：

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial b} = 0 \quad (3-11)$$

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = 0 \quad (3-12)$$

通过上述两个公式，我们得到最优超平面的下列特性：

1、对最优超平面，系数 α_i^0 必须满足约束

$$\sum_{i=1}^l \alpha_i^0 y_i = 0, \alpha_i^0 \geq 0, i = 1, \dots, l \quad (3-13)$$

2、最优超平面（向量 w_0 ）是训练集中的向量的线性组合

$$w_0 = \sum_{i=1}^l y_i \alpha_i^0 x_i, \alpha_i^0 \geq 0 \quad (3-14)$$

我们可以利用Lagrange优化方法将求最优分类面问题转换为其对偶问题，即在约束条件下 $\sum_{i=1}^l \alpha_i^0 y_i = 0, \alpha_i^0 \geq 0, i = 1, \dots, l$ 下对 α^0 求解下列函数的最大值：

$$W(\alpha^0) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (3-15)$$

基于最优超平面的分类规则就是下面的指示函数：

$$f(x) = \text{sgn}(\sum_{i=1}^l y_i \alpha_i^0 (x_i \cdot x) - b_0) \quad (3-16)$$

其中 x_i 是支持向量， α_i^0 是对应的拉格朗日系数， b_0 是常数

$$b_0 = \frac{1}{2}[(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))] \quad (3-17)$$

其中， $x^*(1)$ 表示属于第一类的某个支持向量， $x^*(-1)$ 表示属于第二类的某个支持向量。对于非线性可分问题，可以通过非线性变换转化为某个高维空间中的线性问题，在变换空间最优分类面。选用合适的核函数 $K(x_i, x_j)$ 满足Mercer条件变换到高维空间，相应的分类函数变为

(4) 新的决策函数只通过改变定义特征空间的核函数即可实现。

由于SVM只能解决二分类问题，所以需要将其推广以便解决多分类问题。目前主要有两种策略，第一种是将多分类看作二分类的组合，最终将多分类问题转化为二分类问题，第二种是通过修改目标函数，从根本上解决SVM处理多分类问题。由于后者代价过高，只适用于小规模问题，目前多采用第一类方法。对于k类问题，给定样本集 $(x_1, y_1), \dots, (x_l, y_l), x_i \in R^n, i = 1, \dots, l, y_i \in \{1, \dots, k\}$ 。

目前的用SVM解决多分类问题的方法有：

(1) 一对多SVM分类(One-against-the rest)

一对多SVM分类是最为简单，也是最为普通的实现方案。对于 $k(k \geq 2)$ 类SVM分类问题，把类1作为一类，其余 $k-1$ 类视为一类，自然地将k分类问题转化为二分类问题。这种分类方法在训练过程中，每个分类函数都需要所有的样本参与。分类函数为：

$$f^j(x) = \sum_{i=1}^l (\alpha_i^j y_i^j K(x, x_i^j) - b^j) \quad j = 1, \dots, k$$

上标表示第j个SVM分类器的决策函数， α_i^j 和 y_i^j 分别为第i个支持向量的参数和类别标号， b^j 为偏移量。对待测样本，若：

$$f^j(x) = \max(f^j(x)) \quad j = 1, \dots, k$$

则输入样本属于第l类。这种方法的训练时间与类别的数量成正比，并未考虑多个分类器对测试错误率的影响，当训练样本较大时，训练较为困难。

(2) 一对一SVM分类

一对一的解决办法是在k类问题中进行两两组合，构造 C_k^2 个分类器，该方法也称作Pairwise Method。这种方法的确定是对于类别k过大时，产生的子分类其过多，相对于一对多分类，分类器明显增加，训练时间更长。由于测试时要对任意两类进行比较，训练速度随着类别的增加成指数倍降低。

(3) 有向无环图SVM分类在(Directed Acyclic Graph)

有向无环图SVM分类在训练阶段也是采用一对一SVM的任意两两组合训练方式，但那样也需要构造 C_k^2 个分类器，但是在分类过程中，DAG将所用分类器构造

成有向无环图如图2, 包括 C_k^2 个节点和 k 个叶子, 其中每个节点是一个分类器。

当对未知样本训练时, 从根节点开始分类, 只需 $k-1$ 步即可完成分类。和一对一 SVM 分类相比, 在分类过程中减少了重复操作, 大大提高了分类速度。这种分类方法的缺点是未考虑样本不平衡数据对分类速度的影响, 而且没有考虑错误传递对后续产生影响。

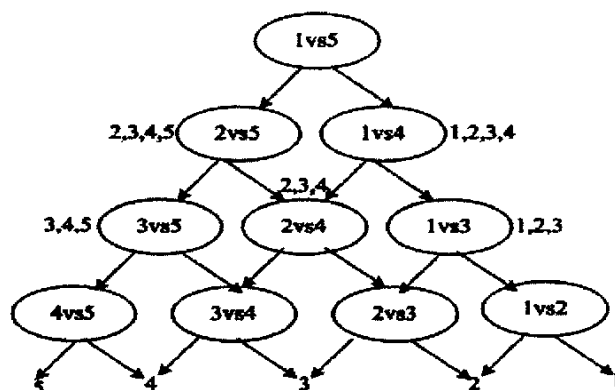


图3-3 采用DAG5分类决策过程

(4) 多分类SVM分类(Multi-class SVM)

该方法通过修改目标函数, 把多分类问题转换为解决单个优化问题, 从而建立 k 分类支持向量机。这种方法可以一步完成, 训练方法多采用块算法或SMO算法。缺点是训练时间相对前面几种方案较慢, 适用于样本数量规模较小的问题的求解。

对于在当前特征空间中线性不可分的模式, 则使用一个核函数把样本映射到一个高维空间中, 使得样本能够线性可分。

以上介绍了常用的文本分类系统的主要步骤和方法, 不同的方法有不同的优缺点, 在实际的系统实现中, 根据系统对精度、算法复杂度、算法实现的难易程度的不同要求, 选择不同的算法。

第四章 中文文本分类系统的设计与实现

4.1 系统实现的目的和意义

本文实现了一个完整的中文文本分类系统，为研究文本分类相关技术提供了实验平台，为研究和完善分类技术打下基础。在实验平台的基础上，验证有关算法的有效性，对提高文本分类系统的准确率和效率的提高具有一定参考价值。在系统的实现过程中，我们综合考虑到了应用和研究的需要，对系统进行了细致的模块化设计，使系统具有很好的扩展性和可替换性。

4.2 系统框架

从系统实现的角度来说，文本分类系统包含四大模块：建立特征库、训练、测试和分类。系统框架如图4-1。

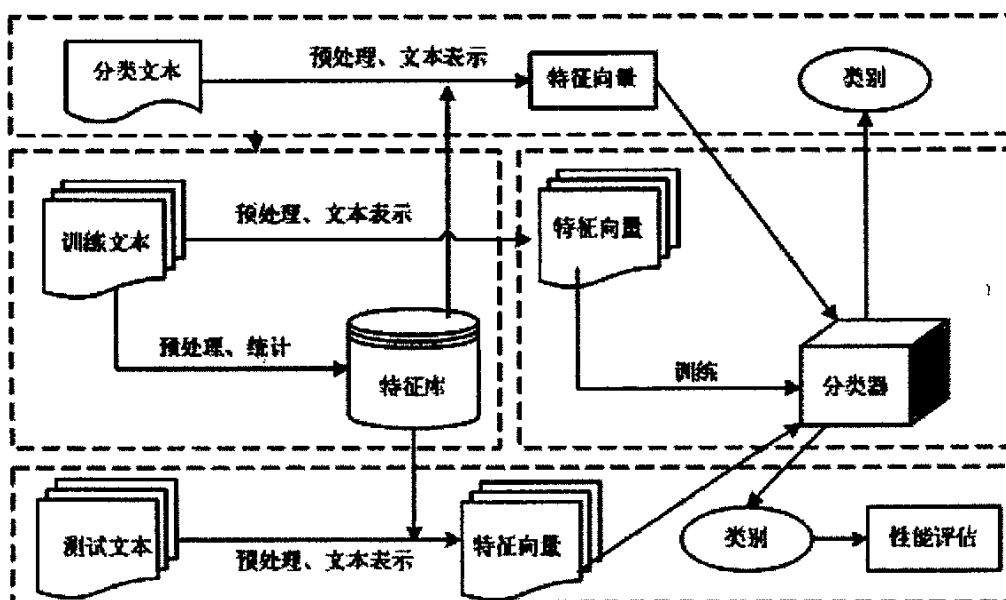


图4-1 网页分类系统框架

系统的处理过程如下：

首先，系统对训练文本的特征项进行评估，建立一个关于特征项信息的特征

库；然后系统使用训练样本根据特征库中特征项的信息进行文本表示，用于分类器的训练；然后对系统分类的效率进行评估；最后系统进入应用阶段。

建立特征库的目的在于减少文本特征向量的维数，去除冗余特征项，保留有区分能力的特征项。同时具有区分能力的特征可以提高系统的效率和精度。建立特征库的过程是根据大量样本的统计，选择一个评价函数对所有特征项进行评估。这样每个特征项就有一个评估分。根据所有特征项的评估分排序，我们保留评估分高的特征项构成特征库，舍弃评估分低的特征项。我们认为，评估分高的特征项是表现文章的主干，对文章的理解和表示起关键作用，而评估低的特征项对文章内容的贡献很小，去掉这些特征项将不影响文章意思的表达和分类的结果。目前常用的评价函数主要有信息增益、互信息、期望交叉熵、CHI统计。由于各个评价函数考虑的侧重点不同，有可能出现各个评价函数对同一个特征项的评估分会相差很大，所以我们对于每个评价函数建立相应的特征库。我们构造的特征库规模约为2万个词左右，已足够满足系统的需要。

训练是文本分类系统的核心，一个文本分类系统做得是否成功，直接取决于分类模型的训练。训练的本质主要是通过已知类别的训练样本来得到分类模型，分类模型最后用来对未知样本进行分类。所有的训练样本都是经过人工已经分好类的，理论上是训练样本越多，所训练出来的分类模型就越精确，但是这样所花费的训练时间就越长，而且当训练样本到达一定的数目的时候，样本数目继续增加对分类效果没有明显的改进，因此需要一个折中的考虑。

测试的目的是通过专家与文本分类系统对于一些文章所做分类的对比，来衡量文本分类系统的效果。训练模块分为文本预处理、文本表示和训练分类器三个过程。文本预处理的目的是对文本进行分词和去除停用词。文本表示的目的是保留区分能力强的特征项并计算其权重，用特征向量的形式取代该文本。训练分类器的目的是对训练文本和其所属的类别利用机器学习方法建立一个函数映射关系。

根据测试模块所使用的测试语料的不同，测试被分为封闭性测试和开放性测试。当采用的测试语料是训练分类器所用训练语料的一部分或者全部时，所作的测试被称作封闭性测试；如果用来测试的语料不曾被用作训练语料，这时所作的测试就是开放性测试。测试所经历的过程有文本预处理、文本表示、分类和对分

类结果的评价这样四个步骤。文本预处理和表示与在训练模块中所作的工作完全一样。分类就是通过每个类节点的分类器判断文章是否属于这个类别。对于分类结果的评价是从所有参与测试的语料这样一个宏观的角度，以评估指标对分类系统的性能作出评价。

在研究性的实验中，分类模块的设计是没有意义的。然而从面向实用的角度来看，设计这样一个模块还是有必要的。分类模块的功能很简单，使用通过训练得到并且经过测试的分类系统，对随手取来的一篇文章进行分类的操作。分类模块所作的工作，同测试模块中的分类完全一样。当然在进行分类之前需要对待分类的文本进行预处理和表示，以得到该文章的特征向量。

4.3 总体结构

文本分类系统主要由语料库维护模块、文本预处理模块、建立特征库模块、文本表示模块、训练分类器模块、分类模块和性能评估模块组成。

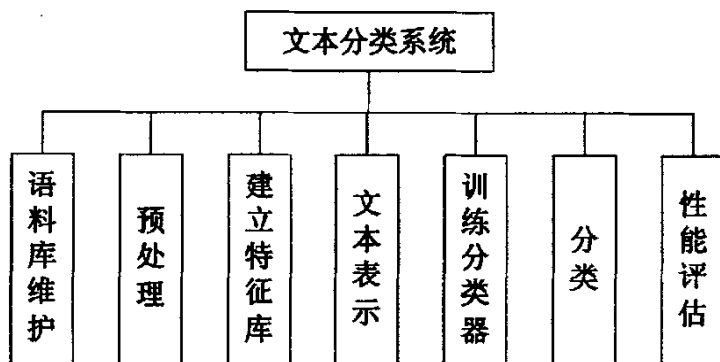


图 4-2 分类系统总体结构

4.4 功能模块说明

4.4.1 语料库维护

文本类管理负责文本类的建立和删除。文本管理负责训练文本的添加、删除、浏览和索引。语料库中的文本按照其所属类别以目录树的结构存储。

4.4.2 网页预处理



图4-3 网页预处理过程

1、网页内容提取

我们进行分类的对象是网页，所以需要将网页中的内容提取出来。互联网上的网页格式主要有HTML、XML、XHTML，本系统目前只限于对HTML网页进行处理。用HTML语言编写的源文件由标签（用“<>”括起来表示特定含义的标识）和文本组成。虽然HTML是一种半结构化的数据，但本质上是文本文件。其基本结构如下：

```

<html>
<head><title>标题</title></head>
<body>
...
超链接
图片
正文
...
</body>
</html>

```

网页标题被定义在标签<title>和</title>之间，网页的主体内容被定义在<body>和</body>之间，其间包括图片、超链接和正文。通过对大量中文网页进行分析，发现只要提取网页的正文、标题信息就可以反映网页的内容。网页的标题比较好提取，正文的提取稍显复杂。

2、中文分词

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

用中科院的提供的分词程序。

中科院在多年的研究基础上,耗时一年研制出了汉语词法分析系统 ICTCLAS(Institute of Computing Technology Chinese Lexical Analysis System),该系统提供的功能有:中文分词,词性标注,未登录词识别。分词正确率高达97.58%,未登录词识别召回率(即查全率)均高于90%,其中中国人名的识别召回率接近98%,处理速度为31.5Kbytes/s。ICTCLAS提供了调用接口,开发者可以忽略汉语词法分析,直接在自己的程序中调用ICTCLAS。

3、去除停用词

停用词(Stop word)是指没有区分意义的高频词,如文本中的“是”、“的”、“但是”等。从相关性的角度看,这些词会出现在各种类别的文本中,没有区分意义,会对分类算法造成干扰。从词性看,一般包括介词、副词、感叹词等。在实验中使用的停用词总数为538条。在进行分词之后,需要去除停用词,同时去除标点和空格等冗余符号。

以下两个图分别是原始网页和经过预处理的网页:

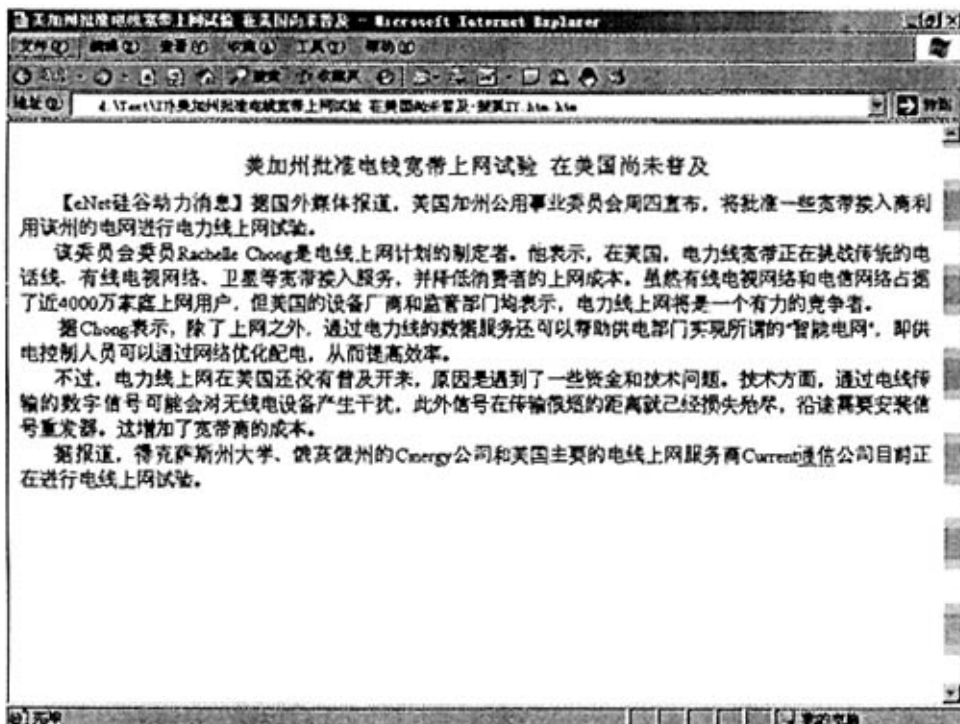


图4-4 原始网页

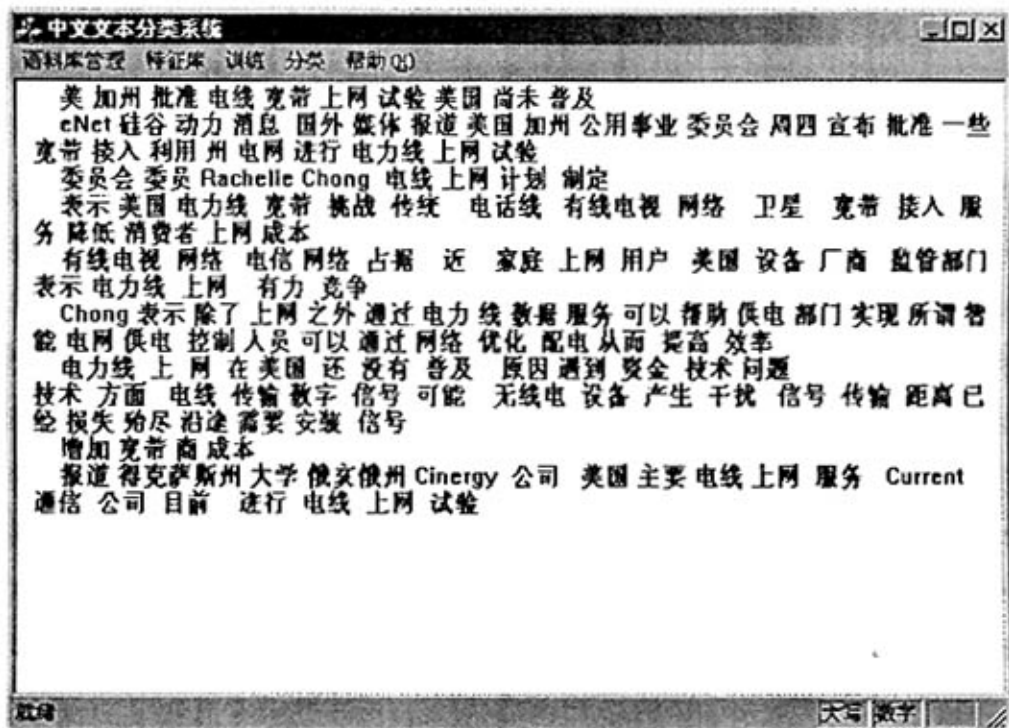


图4-5 经过预处理后的网页

4.4.3 建立特征库

构成文本的词汇，数量是相当大的，因此，表示文本的向量空间的维数也相当大，可以达到几万维，因此我们需要进行维数压缩的工作，这样做的目的主要有两个，第一，为了提高程序的效率，提高运行速度，第二，所有几万个词汇对文本分类的意义是不同的，一些通用的、各个类别都普遍存在的词汇对分类的贡献小，在某特定类中出现比重大而在其他类中出现比重小的词汇对文本分类的贡献大，为了提高分类精度。对于语料库，我们统计每一个特征项的信息，用特征抽取方法来评估该特征项，我们应去除那些表现力不强的词汇，筛选出针对该语料库的特征项集合。在中文文本分类中经常采用的特征抽取方法包括文档频率DF，互信息MI、信息增益IG和 χ^2 统计等。系统分别实现以上几种特征抽取，并针对各种特征抽取方法建立相应的特征库。系统取得了非常不错的实验效果。

4.4.4 文本表示

文本表示的主要目的是将文本表示成计算机可以识别的形式。目前有多种模型用于文本的表示，例如布尔模型、概率模型、空间向量模型(VSM)。在文本分类方面的研究，用得比较多的是空间向量模型。实践证明这是一种效果非常好的表示方法。关于向量模型，已经在第二章中作了详细的介绍。简要描述如下：

向量空间模型的基本思想是以特征向量来表示文本： (w_1, w_2, \dots, w_n) ，其中 w_i 为第 i 个特征项的权值。

1、特征项抽取

经过预处理的文本根据特征库中各个特征项信息量的排序情况，保留区分能力强的特征项，去除区分能力弱的特征项，压缩文本表示的维数，提高系统分类效率。

2、特征向量

计算特征项在该文本中的权值。目前计算特征项权值最常用的一种方法是 *TFIDF*。特征项 t_i 在文档 d 中 *TFIDF* 值有下式定义：

$$TFIDF_i = TF_i \times \log(N/DF_i)$$

其中 TF_i 是特征项 t_i 在文档 d 中出现的频数， N 表示全部训练文档的总数， DF_i 表示包含特征项 t_i 的文档频数。为降低高频特征项对低频特征项的过分抑制，在实验中计算权值是还对 *TFIDF* 进行 L^2 规范化处理：

$$w_i = \frac{TFIDF_i}{\sqrt{\sum_{i=1}^n (TFIDF_i)^2}}$$

本系统分别实现*TFIDF*、开根号和基于句子重要度的权重算法，并在后面的实验中对这三种方法的分类效果进行了对比。

4.4.5 训练模块

训练的过程是构造分类器的过程，是分类系统的核心。系统实现了多种基于向量空间模型的训练算法：简单向量距离算法、K 最近邻居算法和支持向量机算法。训练完成后，保存分类模型的信息。各种分类模型需要保存的信息如

表 4.1

表 4-1 各分类模型需要保存的信息

分类算法	分类器需要保存的信息
简单向量距离法	保存每个类的类中心向量
K 最近邻居法	保存每个文本的特征向量
支持向量机法	(1) 每个文本的特征向量 (2) 每个向量机的对应的拉格朗日系数和偏移量

4.4.6 分类模块

待分类的文本经过预处理、文本表示后，选择相应的分类模型，最后得出分类结果

4.4.7 性能评估模块

因为文本分类本质上一个映射过程，评价^[37]性能能够反映分类系统映射的准确程度，对于比较各种特征抽取算法、权值计算方法和分类方法有重要的作用。信息检索常见的评价指标一般包括查准率、查全率和 F1。对于每类的查准率、查全率和 F1 值定义如下：

1、查全率(recall)

$$recall = \frac{\text{分类正确的文本数}}{\text{该类应有的文本数}} \times 100\%$$

2、查准率

$$precision = \frac{\text{分类正确的文本数}}{\text{实际分到该类的文本数}} \times 100\%$$

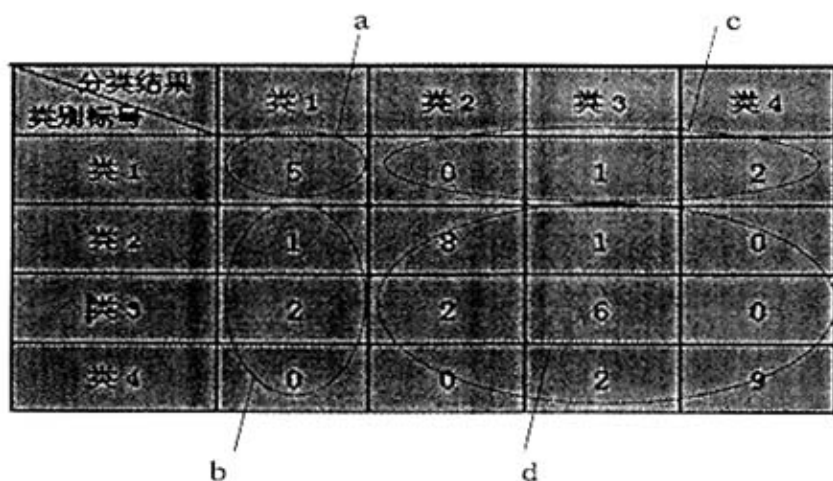
3、F1 值

F1 值是对查全率和查准率的综合考虑

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

为了计算以上指标，我们在分类结束后，对每一个类别，可以定义以下四个参数：

- a: 正确分到该类的文本数
- b: 不属于该类，但是分类程序将其分到该类的文本数
- c: 属于该类，但是分类程序将其分到其他类别的文本数



The diagram shows a 4x4 confusion matrix. The columns are labeled '分类结果' (Classification Result) and the rows are labeled '类别标号' (Category Label). The matrix contains the following values:

类别标号 \ 分类结果	类 1	类 2	类 3	类 4
类 1	5	0	1	2
类 2	1	8	1	0
类 3	2	2	6	0
类 4	0	0	2	9

Labels a, b, c, and d point to the following cells in the matrix:

- a points to the cell (类 1, 类 1) with value 5.
- b points to the cell (类 4, 类 1) with value 0.
- c points to the cell (类 1, 类 4) with value 2.
- d points to the cell (类 4, 类 3) with value 2.

图 4-6 计算性能指标的参数

根据性能参数，评价指标计算方法如下：

$$\text{precision} = \frac{a}{a+b}$$

$$\text{recall} = \frac{a}{a+c}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

以上三个评价指标只用于单个类别上准确率的评价，为了从总体上评价分类性能，有两种评价方法，即微平均和宏平均。微平均首先将每个类别的 a, b, c 分别相加，然后直接使用指标计算方法得到性能评价。宏平均首先计算每个类别的性能指标，然后对各个类别的指标求平均值得到总体上的性能评价。一般说来，微平均较易受大类的分类性能的影响，而由于宏平均是对全部类别取均值，故相对易受小类的影响。为了从整体上得到最全面的结论和简约起见，在实验中我们使用微平均作为分类性能的评价标准。

4.5 系统实现及运行环境

硬件环境:

Intel 2G CPU

256M 内存

40G 硬盘

软件环境:

Windows 2000 操作系统

Visual C++ 开发平台

第五章 实验结果与分析

本章研究特征抽取算法、权重算法、分类算法对分类的影响并进行实验和分析。

5.1 语料集

语料是从互联网上人工收集并进行分类的网页，其中包括军事、政治、经济、体育、艺术、环境、计算机、教育共 8 类，共计 2535 篇文章。每类包含的文本数见下表。

表 5-1 语料库情况

类别	军事	政治	经济	体育	艺术	环境	计算机	教育
文本数	198	315	336	391	102	404	283	506

5.2 实验一

该实验的目的是比较特征抽取方法及向量维数对分类的影响。对于一个中等规模的语料库，其特征项可能有数十万个。如果用所有的特征项进行训练和分类，不仅会降低系统的效率，而且系统的精度也不一定能保证。特征抽取的目的在于减少文本的特征向量维数，去除冗余特征，保留区分能力强的特征，同时能提高系统的分类效率和精度。下面将对四种常用的特征抽取算法进行比较和分析。为了避免特征抽取算法对分类器的依赖性，对每种特征抽取算法分别使用 KNN 和 SVM 算法进行训练和测试，其中 SVM 的核函数为线形核函数，KNN 中 K 的阈值设为 50，权重计算采用 TFIDF 公式。评价标准为宏平均 F1

下面列出各种抽取算法在 KNN 和 SVM 分类器上的实验结果

表 5-2 MI (Mutual Information) 对分类性能的影响

特征维数 分类算法	100	200	500	1000	2000	4000	8000
KNN	30.52%	31.72%	40.44%	43.28%	52.57%	59.8%	64.56%
SVM	24.83%	34.15%	49.25%	55.24%	65.31%	81.26%	86.61%

表 5-3 IG (Information Gain) 对分类性能的影响

特征维数 分类算法	100	200	500	1000	2000	4000	8000
KNN	78.37%	81.58%	85.43%	87.79%	87.36%	83.40%	76.76%
SVM	82.76%	90.25%	94.64%	95.28%	95.50%	95.71%	95.28%

表 5-4 ECE 对分类性能的影响

特征维数 分类算法	100	200	500	1000	2000	4000	8000
KNN	78.15%	80.83%	85.86%	87.15%	87.68%	82.97%	77.19%
SVM	82.86%	89.82%	94.43%	95.50%	96.14%	95.71%	95.82%

表 5-5 CHI 对分类性能的影响

特征维数 分类算法	100	200	500	1000	2000	4000	8000
KNN	78.37%	82.54%	85.43%	89.72%	86.93%	82.65%	78.37%
SVM	86.18%	89.82%	94.21%	95.93%	95.61%	95.28%	95.39%

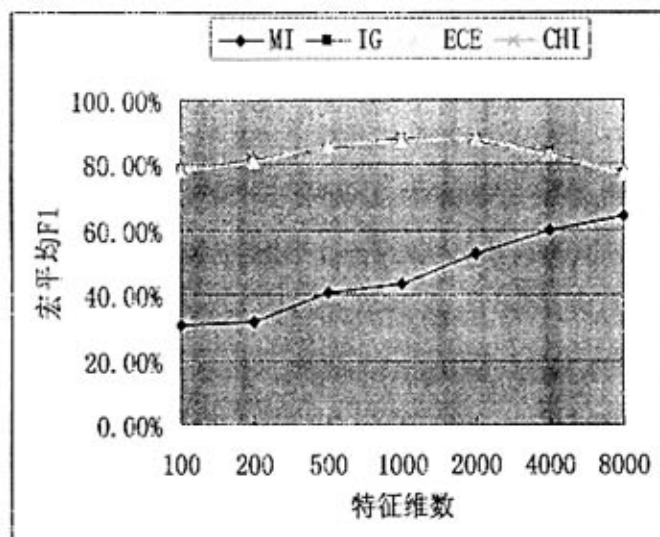


图 5-1 四种抽取算法在 KNN 分类器上的比较

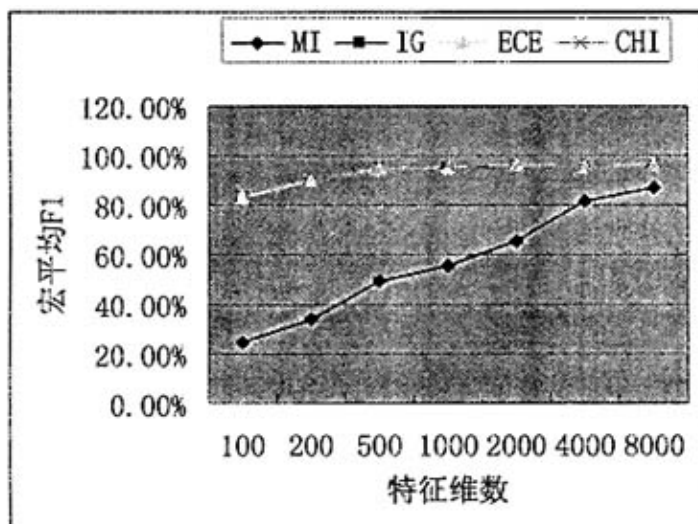


图 5-2 四种抽取算法在 SVM 分类器上的比较

由以上两个表可以得到如下结论：

(1) IG、CHI 和 ECE 抽取方法对分类效果的影响相当，当特征维数从 100 到 1000 变化时，F1 的变化比较显著，当特征维数 > 1000 时，F1 值趋于稳定，其变化是缓慢的；特征维数一般在 1000 左右比较合适，特征维数较小时，特征向量不能完全变现文本的内容，维数过大时不会改善分类性能，反而会降低分类性能。所以有必要通过特征抽取方法进行维数压缩。

(2) MI 的分类性能最差, 随着特征维数的增加, 分类性能变化明显, 究其原因, MI 函数不选择高频的有用词而选择稀有词作为特征向量。

5.3 实验二

该实验主要考察分类算法对分类的影响, 特征抽取算法采用 CHI, 权重算法采用 TFIDF 公式, 评价标准为宏平均 F1。

表 5-6 三种分类算法的宏平均比较

特征维数 分类算法	100	200	500	1000	2000	4000	8000
简单距离向量法	70%	72%	76%	75%	79%	77%	75%
KNN	78.37%	82.54%	85.43%	89.72%	86.93%	82.65%	78.37%
SVM	86.18%	89.82%	94.21%	95.93%	95.61%	95.28%	95.39%

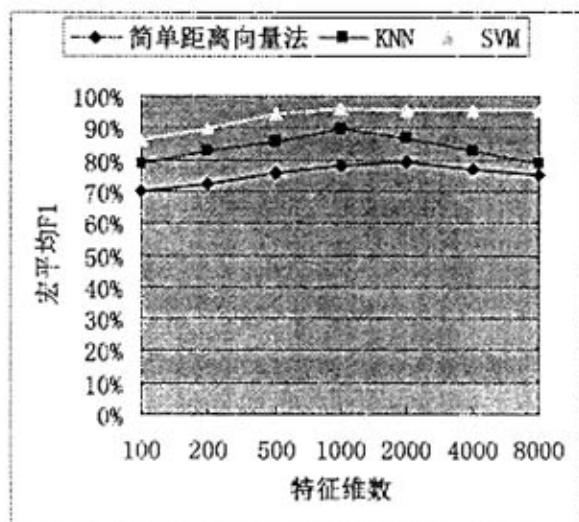


图 5-3 三种分类算法的宏平均比较

简单距离向量法效果最差, 原因是该方法是从每类的训练样本中总结出一

个代表该类的中心向量，但是在总结的过程中会损失原始类别信息。KNN 算法算法较好，但是当特征维数>1000 时，由于特征向量中类区分能力不强的特征项增多，分类性能有所下降。SVM 算法性能最好，随着特征维数的增加，分类性能趋于稳定。

5.3 实验三

该实验的主要目的是比较特征项权重计算方法对分类的影响，特征维数为 1000，分类算法采用简单距离向量法，特征抽取算法采用 CHI，评价标准为准确率。

表 5-7 权重算法对分类性能的影响

类别	开根号函数	TFIDF 函数	句子重要度函数
	准确率(%)	准确率(%)	准确率(%)
军事	62.1	65.2	67.2
政治	75.6	76.5	79.7
经济	74.7	75	80.1
体育	88.2	89	92.1
艺术	45.1	46.1	50
环境	79.5	82.4	85.9
计算机	71	74.6	80.1
教育	79.8	82.6	86.2
总计	76	78	82

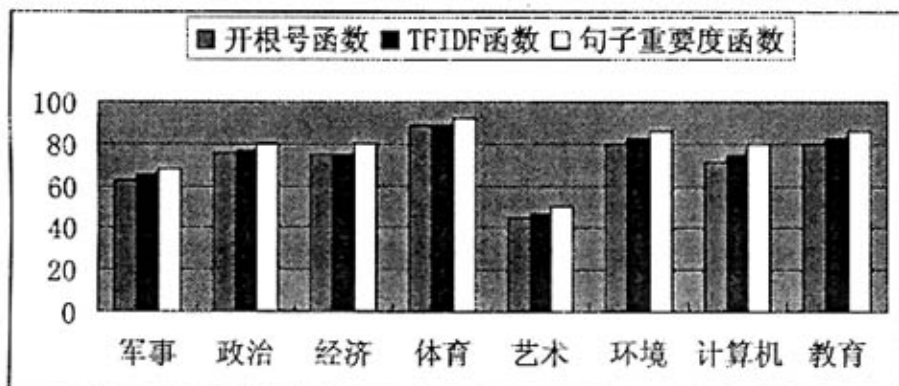


图 5-4 三种权重算法的在各类中准确率比较

从上图可以得出以下结论：

(1)开根号函数的只考虑了特征项出现频率，分类效果最差。TFIDF 函数考虑了特征项的出现频率和文档频率，分类效果比开根号函数要好。基于句子重要度的特征项权重算法不仅考虑特征项的出现频率和文档频率，还考虑了特征项所在的句子的重要度，使得特征项的权重携带了更多的信息，更能反映文本的内容，所以效果要比上面两种权重算法的效果好。实验证明基于句子重要度的权重计算是一种切实可行的算法。

(2)有些类别的分类效果比较差，其原因是该类别的文本数较少，不能覆盖该类别的各种题材，文本特征不能很好地对类别进行表示。

第六章 结束语

面对 Internet 上日益膨胀的信息,如何快速、准确地从浩瀚的信息资源中寻找到所要的狭小领域内的相关内容就成了一项十分有意义的课题。文本分类可以在较大程度上解决目前网上信息杂乱的现象,方便用户准确地定位所需的信息和分流信息。因此,文本分类已成为一项具有较大实用价值的关键技术,是组织和管理数据的有力手段。

6.1 总结

在本文的研究中,主要进行了四个方面的工作:

- 1、本文对中文文本分类涉及的技术进行了深入的研究,包括中文分词、向量空间模型、特征抽取、特征项权重算法和分类算法。
- 2、为了解决传统的特征项权重算法只是简单地考虑特征项权重出现的频率,而没有考虑特征项在文本中出现的位置的缺点,本文引入句子重要度的概念,用句子重要度对特征项权重进行加权,使得经过该方法计算的特征项权重可以准确地表现文本的内容,实验证明,这种方法的效果要好于传统的特征项权重方法。
- 3、提出并实现了一套完整的基于向量空间模型的中文文本分类系统,为研究分类技术提供了实验平台。
- 4、实验考查了特征抽取、特征项权重算法、分类算法对分类效果的影响并对其原因进行了分析。

6.2 下一步工作展望

本文需要进一步研究的工作有:

- (1)本系统的分类体系为平面体系,可以考虑层次分类体系
- (2)在分类系统的基础上,实现贝叶斯、神经网络等其它算法,测试分类效

果，选择最优的分类算法。

(3) 提高分类的时间效率，使系统具有更好的实用性。

参考文献

- [1] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. ACM computing Surveys, 2002, 34(1):11-12, 32-33
- [2] 黄萱菁, 吴立德. 基于向量空间模型的文档分类系统. 模式识别与人工智能. 1998, 11(2)
- [3] 李勇, 桑艳艳. 网络文本数据分类技术与实现算法. 情报学报. 2002, 21(1):21-26
- [4] Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-wesley, <http://citeseer.nj.nec.com/content/26897/0>, 1989
- [5] 王玲, 马文庆. 搜索引擎技术的现状与展望. 现代情报. 2004, (8):71-72
- [6] 王辉, 陈凌, 张丽娟. 信息推拉技术. 情报科学. 2004, 21(12):1440-1443
- [7] 徐险峰. 基于因特网的网络信息资源个性化服务研究. 图书馆建设. 2004(5):62-64
- [8] 张小兵, 李靖. 信息过滤技术发展趋势. 信息网络安全. 2004, (2):17-18
- [9] 徐小琳, 阙喜戎, 程时端. 信息过滤技术和个性化信息服务[J]. 计算机工程与应用. 2003, 39(9):182-184
- [10] 李国辉, 汤大权, 武德峰. 信息组织与检索. 北京: 科学出版社., 2003
- [11] Hayes PJ, Weinstein SP. Construe/Tis: a system for content Based indexing of a database of new stories. In: Second Annual Conference on Innovative Applications of Artificial Intelligence, 1990.
- [12] Thorsten Joachims. "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization". In Proceedings of ICML'97. 1997.
- [13] Eui-Hong(Sam) Han George Karypis, Vipin Kumar. "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification". In Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and data Mining. 2001
- [14] 孙健, 王伟, 钟义信. 基于 K-最近距离的自动文本分类的研究. 北京邮电大学学报. 2001, 24(1).
- [15] 李杨, 曾海泉, 刘庆华等. 基于KNN的快速WEB文档分类[J]. 小型微型计算机系统. 2004, 25(4):725-729.
- [16] 李永平, 程莉, 叶卫国. 基于隐含语义的 KNN 文本分类研究[J]. 计算机工程与应用. 2004, 40(6):71-73.
- [17] 韩家新, 何华灿. SVMDT 分类器及其在文本分类中的应用研究[J]. 计算机应用研究. 2004, 21(1):23-24.
- [18] 李静梅, 孙丽华, 张巧荣等. 一种文本处理中的朴素贝叶斯分类器[J]. 哈尔滨工程大学学报. 2003, 24(1):71-74.
- [19] 胡于进, 周小玲, 凌玲等. 基于向量空间模型的贝叶斯文本分类方法[J]. 计算机与数字工程. 2004, 32(4):28-32
- [20] 刘钢, 胡四泉, 范植华等. 神经网络在文本分类上的一种应用[J]. 计算机工程与应用. 2003, 39(36):73-74, 92.
- [21] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报. 2001, 26(1):32-42
- [22] Vojtech Franc, Vaclav Hlavac. Multi-class Support Vector Machine[C]. In: 16th International Conference on Pattern Recognition(ICPR 02), 2002:236-239
- [23] 郑勇涛, 刘玉树. 支持向量机解决多分类问题研究[J]. 计算机工程与应用.

- 2005,23:190-192
- [24] 张健沛, 徐华. 支持向量机 (SVM) 主动学习方法研究与应用[J]. 计算机应用. 2004, 24(1):1-3
 - [25] Vladimir N Vapnic. An Overview of Statistical Learning Theory[J]. IEEE Transaction on Neural Networks. 1999, 10(5):988-999.
 - [26] Ricardo Baeza-Yates, Berthier Riberiro-Neto. Modern Information Retrieval[M]. 北京: 机械工业出版社, 2004.2
 - [27] Verhoeff J, Goffmann W, Jack Belzer. Inefficiency of the use of Boolean functions for information retrieval systems[J]. Communications of the ACM. 1961, 4(12):557-558, 594.
 - [28] Salton G, Wong A, Yang CS. A Vector Space Model for Automatic Indexing[J]. Communication of the ACM, 1995, (1):2-8.
 - [29] 马玉春, 宋瀚涛. Web 中文文本分词技术研究[J]. 计算机应用. 2004, 24(4):134-135, 155.
 - [30] 秦浩伟, 步丰林. 一个中文新词识别特征的研究[J]. 计算机工程. 2004, 30(B12):369-370, 414.
 - [31] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报. 2004, 18(3):17-23.
 - [32] 代六玲, 黄河燕, 陈肇雄. 中文文本分类中的特征抽取方法的比较研究[J]. 中文信息学报. 2004, 18(1):26-32.
 - [33] Y. Yang. A comparative Study on Feature Selection in Text Categorization[C]. In: Proceeding of the Fourteenth International Conference on Machine Learning (ICML'97). 1997
 - [34] 蓝海洋, 周杰韩, 张和朋. 文本索引词项相对权重计算方法与应用[J]. 计算机工程与应用. 2003, 15:68-70
 - [35] Young joong Ko *, Jinwoo Park, Jungyun Seo. Improving text categorization using the importance of sentences[J]. Information Processing and Management 40 (2004) 65-79
 - [36] Vapni VN. 统计学习理论的本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000
 - [37] 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. 14(6):1-7.

致谢

本论文是在导师董小国教授悉心指导下完成的，在攻读硕士学位期间，导师给予了我严格、耐心的指导，帮助我树立了正确的研究方向和科学的学习方法，并为我提供了良好的学习、研究的环境，使我顺利地完成了硕士期间的学业。导师学习渊博、治学严谨、工作勤奋、为人谦逊随和。我在导师的指导下，既学到了科学的研究方法，也学到了踏实做人的原则，使我受益终身，在此，向导师致以衷心的感谢。

在这三年里，网络中心的各位老师在学习上给予了我很大的指导和帮助，在此向他们表示深深的谢意。

另外，还要对各位学友表示感谢，他们在三年来对我的帮助和支持使我永远难忘。

最后，感谢我的家人，他们的支持与鼓励是我前进的动力，使我能够安心顺利地完硕士学业。

研究成果及发表的学术论文

发表及已接受的论文

1. 董小国, 甘立国. 基于句子重要度的特征项权重计算方法. *计算机与数字工程*. 已收录.