分类号：　　单位代码：10019

密　级：　　学　　号：S02543

中国农业大学

硕士学位论文

Web 挖掘技术在东亚植物遗传资源管理系统中的应用研究

Research on Application of Web Mining Technology on EA-PGR Management System

研　　究　　生：张海龙

指　导　教　师：王莲芝 副教授

合 作 指 导 教 师：

申请学位门类级别：工学　硕士

专　业　名　称：计算机应用技术

研　究　方　向：计算机网络及其应用

所　在　学　院：信息与电气工程学院

2005 年 6 月

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国农业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：　　时间：　　年　　月　　日

关于论文使用授权的说明

本人完全了解中国农业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。同意中国农业大学可以用不同方式在不同媒体上发表、传播学位论文的全部或部分内容。

(保密的学位论文在解密后应遵守此协议)

研究生签名：　　时间：　　年　　月　　日

导师签名：　　时间：　　年　　月　　日

摘　要

为促进东亚地区植物遗传资源的保护和利用，国际植物遗传资源研究所（IPGRI）与东亚各国有关研究机构决定建立“东亚植物遗传资源协作网”（EA-PGR）Web 信息管理系统。

EA-PGR 的 Web 信息管理系统的信息是分类进行管理的，管理员要处理大量来自 IPGRI 的静态 Web 文本，然后把它们按类上传到 Web 信息管理系统中，这些 Web 文本的组织往往处于混乱的状态，采用人工分类，工作量既大，效率又低。正是出于需要对 Web 文本进行分类管理的目的，作者研究了 Web 文本的自动分类技术。本论文研究结果如下：

（1）分析了 Web 文本分类的三个重要技术：特征词提取、特征赋权、特征选择方法的 IG、CHI、期望交叉熵等 6 种评估函数。对来自 IPGRI 的 Web 文本集进行了系统测试，分析了各种评估函数对不同分类器的优劣。

（2）研究了 Web 文本分类算法：类中心向量、KNN、朴素贝叶斯、SVM 等几种分类器，并对 KNN 和 SVM 两种分类器在标准语料库和来自 IPGRI 的 Web 文本集进行了实验比较分析，得出 SVM 是比 KNN 更好的分类器。

（3）作为 Web 文本自动分类技术研究的结果，采用 VC++设计与实现了基于内容的中英文 Web 文本自动分类系统。该系统具有支持 KNN 和 SVM 两种分类器、多种特征选择方法、兼类分类、自定义特征空间维数和分类结果评测曲线、直方图显示等特点。

（4）IPGRI 为了在成员国之间开展多个领域的学术交流和合作活动，决定建立“东亚植物遗传资源协作网”的网站。作者用 ASP.NET 和 ADO.NET 技术，结合 SQL Server 2000 数据库系统，用 C＃语言开发了基于 ASP.NET 的信息管理系统。在此基础上，利用 Web 文本自动分类技术研究结果把分好类的 Web 文本和其他信息进行有效的发布和共享，实现了 EA-PGR 相关信息和数据的动态管理与发布。

关键词：Web 挖掘，Web 文本分类，向量空间模型，SVM，ASP.NET

Abstract

In order to promote the conservation and use of plant genetic resource in East Asia, International Plant Genetic Resource Institute（IPGRI） and related institutes of five member countries of East Asia decided to develop the website of Regional Network for Conservation and Use of Plant Genetic Resources in East Asia（EA-PGR）.

Information of EA-PGR web site is categorized to manage. Administrator of EA-PGR will manage large amount of web documents from IPGRI, which are usually in a state of disorder, then upload them to Web management system. Because manually categorizing will waste a lots of time and is in low efficiency. There is need for developing a technology to categorize webpage automatically. Base on their need, the study on webpage management was carried out from 2003 to 2004. The following are main results for this study:

（1）Three technologies of webpage categorization was analyzed, including features distill, features weight and features selection of IG、CHI、expected cross entropy etc., Web corpus from IPGRI were tested for evaluating function of KNN and SVM categorization machines.

（2）Four categorization arithmetic of web page, namely Cancroids of vector, KNN, Naive Bayes and SVM were analyzed. And a comparison analysis between KNN and SVM on standard data collections from fudan university and web pages from IPGRI was done, SVM was proved better than KNN.

（3）Based on above results item（1） and（2）, Chinese and English automatic web page categorization system was designed with the tool of VC++6.0. This system supports KNN and SVM, multi-features selections, multi-categorization of a text documents, custom feature space quantity and curve and straight showing of categorization result evaluation etc.

（4）The IPGRI website of “EA-PGR” has been established for strengthening regional collaboration sharing and exchanging information. ASP.NET, ADO.NET, SQL Server 2000 database system and C# language were used to design and develop the management system of EA-PGR. Information and data of EA-PGR can be dynamically managed and released through this system.

Key words: web mining，web pages categorization，vector space model，SVM，ASP.NET

目　录

摘　要 I
ABSTRACT II
目　录 III
第一章　前　言 1
1.1　研究背景和意义 1
1.2　国内外研究现状 3
1.3　研究目标和内容 5
1.4　论文组织 5
第二章　WEB 文本分类过程概述 7
2.1　文本分类的概念 7
2.2　文本自动分类的实现过程 7
2.3　文本的预处理与分词 9
2.4　分类器的设计与训练 11
2.5　文本分类的性能评价 13
2.6　实验用到的文本集 14
第三章　WEB 文本分类的特征提取、特征选择与实验比较 15
3.1　WEB 文本的特征提取 15
3.2　WEB 文本分类中的特征赋权 17
3.3　WEB 文本分类中的特征选择 18
3.4　特征选择实验比较与结论 20
第四章　WEB 文本分类算法与实验比较 23
4.1　类中心向量分类法 23
4.2　KNN 分类法 24
4.3　NAÏVE BAYES 分类方法 24
4.4　SVM 分类法 25
4.5　KNN 和 SVM 的算法比较与分析 27
第五章　WEB 文本分类系统的设计与实现 28
5.1　系统结构简介 28
5.2　各功能模块 29
第六章　东亚植物遗传资源管理系统的设计与实现 34

6.1 系统设计 34
6.2 工作原理和主要技术 41
第七章 结论与展望 43
7.1 结论 43
7.2 展望 44
参考文献 45
致谢 48
作者简介 49

第一章 前 言

1.1 研究背景和意义

东亚是世界上重要的作物多样性起源中心之一，大约有 300 多种栽培植物起源于该地区。很多重要作物如大豆、水稻、小麦、燕麦、大麦、荞麦、大白菜、红小豆、柑桔、茶等均起源位于该地区的中国。因此，加强东亚国家栽培植物遗传资源的保护和利用是极为重要的。为促进本地区植物遗传资源的研究工作，国际植物遗传资源研究所（IPGRI）与东亚各国的有关研究机构合作，于 1991 年成立了“东亚植物遗传资源保护和利用协作网”，有 5 个成员国，包括中国、日本、朝鲜、韩国和蒙古。该协作网自成立以来，在成员国之间开展了多个领域的合作活动，其中包括对共同感兴趣的作物开展联合考察收集、鉴定、多样性研究，共同组织和参加地区会议，开展人员培训等。起到了把各个国家的有关单位有效地组织在一起，充分利用现有人力物力资源，联合开展相关研究，促进共同发展的作用。信息共享是协作网的重要目标之一。只有实现信息共享，才能使合作伙伴及时了解相关领域的进展情况，获取相关技术信息，避免不必要的重复。因此，为完成由成员国制定的合作行动计划，确保协作网的有效运行，有必要采取措施，促进相关领域的信息共享。网络是传播和共享信息的必要和有效途径。网站具有信息量大、更新快、传播广的特点。通过 Internet 可以使信息在更广范围内传播，使更多的人在同一时间内共享到最新信息。因此，建立网站是促进信息共享的一条有效途经。在东亚国家，除朝鲜外，其它国家都具备良好的网络设备，具备较强的上网获取信息的能力。这些国家的有关单位也积极支持建立一个协作网中心站点，以此为平台进行信息和学术交流，并通过该网站，连接到其它与植物遗传资源保护和利用有关的网站，获取相关信息。

实践证明，协作网是促进东亚各国在植物遗传资源领域合作的有效机制。为了进一步加强成员国之间合作研究，促进学术交流，实现信息共享和快速传播，根据“东亚植物遗传资源保护和利用协作网国家协调员会议”的建议，IPGRI 决定建立“东亚植物遗传资源协作网”（EA-PGR）Web 信息管理系统。

EA-PGR 的 Web 信息管理系统的许多数据是来自 IPGRI 的，Web 信息管理系统的管理员每天处理大量的属于关系型数据库中的结构化信息，但更多的属于文档、网页或视频等非结构化信息，而来自 IPGRI 的大量 Web 文本是缺少组织的，因此需要对其进行基于内容的分类管理，若采用人工分类，需要较多的经验和专业知识，然而分类质量有时得不到保证、分类周期长、费用高、效率低、不易满足需要。

虽然，现在有比较成熟的信息检索技术，如搜索引擎 Yahoo、Google、Baidu 等，它们部分地解决了信息获取的问题，但并未完全解决信息组织和知识获取的问题。Web 上特有的大量异质非结构信息包括网页内容信息、结构信息、用户访问模式、网站日志等促使一系列新的课题产生，发现和管理这些知识是搜索引擎力所不能及的。为此，提出了 Web 挖掘（Web Mining）的概念，Web 挖掘是指应用一定的技术手段，从 Web 文档和 Web 服务中发现和抽取知识。Web 挖掘借鉴

了数据挖掘的思想，都是为了发现有用的知识和模式，但两者研究和处理的对象不同。数据挖掘研究的对象主要是数据库等结构化信息，而 Web 挖掘主要针对 WWW 上的半结构化或者无结构的内容信息、结构信息、用户访问模式和网站日志信息。

论文要实现的系统流程如图 1-1 所示。

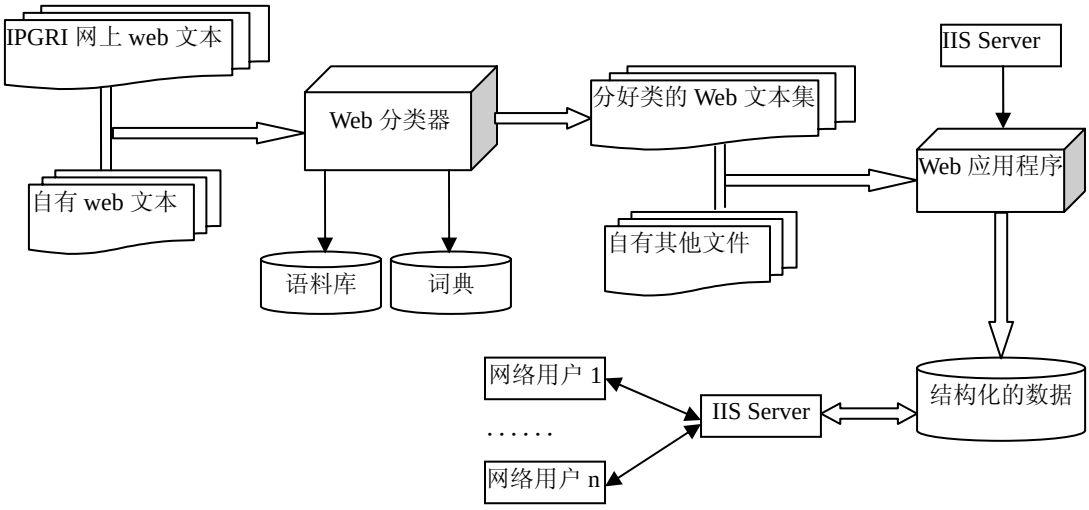

图 1-1 系统总体流程图

按照处理对象的不同，我们将 Web 挖掘分为三大类：Web 内容挖掘、Web 结构挖掘和 Web 用户行为挖掘。Web 内容挖掘指的是从 Web 文件的内容中抽取知识，又分为对文本文件（包括 text，HTML 等格式）和多媒体文件（包括 image，audio，video 等媒体类型）的挖掘。Web 结构挖掘指的是从 Web 文件的结构信息中推导知识，不仅仅局限于文件之间的超链接结构，还包括文件内部的结构、文件 URL 中的目录路径结构等。Web 用户行为挖掘试图从用户和互联网交互时产生的二级数据中发现知识，包括来自 Web 服务器的访问日志数据，代理服务器日志，浏览日志，用户登陆数据，用户查询请求等交互结果数据。Web 挖掘的分类体系如图 1-2。

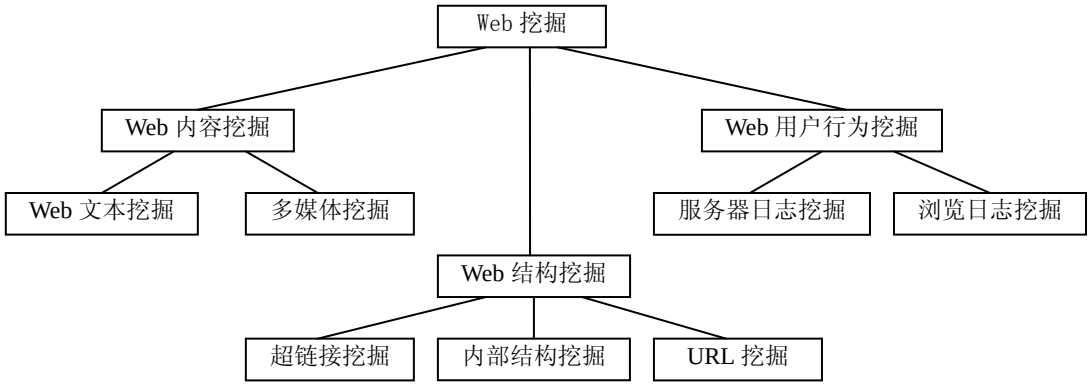

图 1-2 Web 挖掘分类

其中作为 Web 挖掘的重点，Web 文本挖掘研究的内容包括基于内容的 Web 文本分类、文本聚类，关联分析，以及利用 Web 文档进行趋势预测等研究。

网页分类是文本分类应用于 Web 文档的情况，Web 文本分类和文本分类的唯一不同在于网

页/文本在特征提取时存在差别。因为网页中包含许多<title>、<head>、<B>等 tag 标记显示出了一定的结构信息，在经过特征提取后，网页或者文本通过特征来建模，此后的处理包括特征选择，分类学习算法、分类测试和性能评价等各个环节完全一致。在本论文的下述部分，除了在特征提取部分外，Web 文本分类和文本分类两个术语不再加以区分，除非特殊说明，文中的文本即是指 Web 文本。

文本自动分类技术在实际生活中有相当好的应用价值和广泛的应用前景：（1）搜索引擎。通过搜索引擎可以检索到包含用户输入的关键词的大量 Web 文本，但搜索的精度往往不能令人满意，其搜索结果包含很多无关的资料，采用文本自动分类技术则可以大大提高查全率和查准率。（2）电子出版业。在电子出版业中文本的处理速度相对落后于文本的收集速度。目前对电子文本的分类处理仍然以手工为主，如果能采用自动分类，无疑大大加快了对电子文本处理速度。（3）电子图书馆。随着图书馆文本资料管理电子化的普及，也要求对电子图书进行自动分类处理。（4）网络安全。文本自动分类在防火墙技术中也有广泛的用途，利用文本自动分类技术可以有效地过滤掉诸如不健康的信息等。（5）电子邮件分捡。随着电子邮件的普及，电子邮件的数量剧增，其中包含着大量垃圾邮件，采用文本自动分类技术可以对电子邮件进行过滤和分类。（6）新闻网站。根据专家建立好的新闻类别，文本分类器可以对新闻进行实时分类。（7）猎头公司：猎头公司存储了各种人才的简历，手工分类费时费力，文本分类器可以对简历进行自动分类。

本研究采用 Web 挖掘技术中的 Web 文本自动分类技术，对来自国际植物遗传资源研究所（IPGRI）的组织比较混乱的 Web 文本进行分类，并对分好类的 Web 文本和其他资料信息通过 Web 应用程序和数据库的控制，让终端用户达到信息浏览、查询等目的。可进一步加强成员国之间合作研究，促进学术交流，实现信息共享和快速传播，具有重要应用价值。

1.2 国内外研究现状

文本分类是机器学习理论在文档信息获取方面的一个重要应用。文本分类是依据文本内容，由计算机通过某种自动分类算法将文本（集）判分到某一个或几个预定义的类别中去。文本分类技术综合运用了自然语言处理、信息检索、机器学习、计算语言学、统计数据分析、线性几何、概率理论，甚至还有图论以及模式识别等多个领域知识，可以在较大程度上解决目前本地以及网络上信息杂乱的问题，方便用户准确地定位信息和分流信息。

文本的人工分类从很早以前就已经开始了。例如，图书馆的工作人员按照一定的分类体系将各种图书按照内容分到不同的类别中。人工分类需要大量工作，并且要求分类人员具有较多经验和专门知识。在人工分类文本中存在大量问题，主要体现在精确度和代价上。在当前 Internet 快速发展，网上电子文本不断涌现的情况下，手工分类显然不能满足准确、及时的分类需求。文本自动分类技术的实现，为用户提供了一个有力的工具，来帮助用户进行知识发现与内容管理。

根据肖明对文本自动分类的发展历史所做的考证[1]，文本自动分类的研究始于 20 世纪 50 年代，H.P.Luhn 在这一领域进行了开创性的研究，它提出了词频统计思想，用于自动分类。1961 年，Maron 发表了有关文本自动分类的第一篇论文。1962 年，H.Borko 等人提出了因子分析法进行文献的自动分类。随后许多著名的学者如 Sparck、G.Salton 等人在这一领域进行了卓有成效的研究。其中，G.Salton 提出了向量空间模型（Vector Space Model），成为文本分类中文本表示的常

用方法。

80 年代，文本分类用到的主要方法之一是知识工程的方法，它主要思想是手工建造一个能进行分类决策的专家系统，这类专家系统包括了一些形式如 If<DNF 布尔表达式>then <class>的规则。这种手工方法的缺点是建立自动分类器时，需要领域专家的知识和知识的表示。基于知识库的分类方法在特定领域具有良好的应用前景，但是由于该方法主要缺陷在于分类过程需要领域专家手工建立由一系列规则构成的知识库，这个过程相当费时费力，而且不具备可移植性。当应用环境发生变化时，必须重新构建知识库。

从 90 年代初以来，用机器学习的方法进行文本自动分类吸引了广大研究者，并逐渐发展为文本分类领域的主流。机器学习的方法主要是用带有类别标记的训练样本集来训练分类器，通过有监督的学习过程来逼近真实的分类机制，从而达到抽取蕴含分类模式的目的。显然，与基于知识工程的方法相比，机器学习的方法具有领域无关性，无须人工干预，节省人力的优点。文本自动分类的研究主要集中在几个关键技术上：分类模型、分词算法、文本特征的提取和选择、分类算法等。

机器学习方法较少考虑文本的语义信息，随后又出现了将语义分析与概念网络与机器学习方法相结合的分类模型。Web 文档及其之间的超链接信息提供了多于传统文档的有用信息。如标题、段落标题、超链接文字等，以及所用的字号等辅助信息为文档分类提供了有用的信息。利用这些超链接特征和文档向量特征一起进行 Web 文档的挖掘和分类也是目前研究的热点之一。Google 对超文本信息的全面合理的利用，使得它在搜索性能上较以前的网页搜索引擎有了很大的提高。

国外在文本分类的研究中，常用的方法有 Recchio 方法，Naïve Bayes[Joachims,97]，KNN [Yang,99]，决策树[Quinlan,93]，SVM[Dumais,98] [Joachims,98]以及 Boosting [Schapire,00] [Weiss,99]等。其它具有代表性的工作包括 D.Lewis 用线性分类器进行文本分类，Yiming Yang 的 LLSF 方法，Thorsten Joachims 的 SVM 方法，Cohen William 的 Sleeping Expert 方法，Hwee Tou Ng 等人的 Perceptron 方法等。在英文文本分类中，标准数据集包括来自路透社的新闻语料 Reuters-21578，新闻组语料 20-NewsGroup，以及医疗文献集合 OHSUMED 等。在中文文本分类没有标准的数据集，研究者通常拿新闻语料来作为文本分类研究的数据集。例如从新闻网站上下载各类新闻网页，作为分类训练和测试的语料。

在文本分类、信息过滤领域中较为成功的系统有 MIT 为白宫开发的邮件分类系统，卡内基集团为路透社开发的 Construe 系统，以及 Internet 上众多的个性化新闻定制服务等。Construe 系统利用手工构造的规则为基础的专家系统对新闻进行分类，在 750 个测试集上取得了超过 90％的准确率。Yahoo！搜索引擎采用人工将文本分类，构成文本分类的树状结构，从而在其基础上提供网页搜索服务和个性化的新闻定制服务。近年来很多研究者对辅助用户进行 Web 浏览的助手（Agent）进行了研究，如卡内基梅隆大学（CMU）的 WebWatch，明尼苏达大学（UMN）的 WebACE 等。为用户提供个性化的新闻服务也成为近年来的一个研究方向。

在文本挖掘软件中，IBM 的 Text Miner 很有代表性，其主要功能是特征提取、文档聚集、文档分类、支持 16 种语言的多种格式文本的数据检索。Text Miner 的特征抽取器能从文档中抽取人名、组织名和地名以及由多个字组成的复合词。此外，特征抽取器还能抽取表达数字的词汇，例如，“钱”、“百分比”、“时间”等。抽取完特征以后，有相似特征的文档就被自动聚集成一个集

合。利用这一功能，知识管理系统可以从大量文档中找到相关文档。Text Miner 还可以对文档进行自动分类。

国内在中文文本分类领域也进行了大量研究，如中国科学院的李晓黎、史忠植等人应用概念推理网进行文本分类[2]，中国科学技术大学的范焱等人在 KNN、Bayes 和文档相似性研究的基础上提出了一个超文本协调分类器[3]，清华大学的解冲锋提出一种补偿性的 Sleeping Expert 方法进行文本分类[4]，清华大学的张义忠用自组织特征映射网络（SOFM）来做文本分类[5]等等。各种分类方法在应用于具体问题时，都取得了比较好的结果。

1.3 研究目标和内容

1.3.1 研究目标

采用 Web 挖掘技术中的 Web 文本自动分类技术对来自国际植物遗传资源研究所（IPGRI）的组织比较混乱的 Web 文本进行分类，然后对分好类的 Web 文本和其他资料信息通过 Web 应用程序和数据库的控制，让终端用户达到信息浏览、查询等目的。进一步加强成员国之间合作研究，促进学术交流，实现信息共享和快速传播。

1.3.2 研究内容

（1）分析 Web 文本分类的重要技术，即特征词提取、特征赋权、特征选择方法的 IG、CHI、期望交叉熵等 6 种评估函数，对来自 IPGRI 的 Web 文本集进行系统测试，分析各种评估函数对不同分类器的优劣。

（2）研究 Web 文本分类算法，即类中心向量、KNN、朴素贝叶斯、SVM 等几种分类器，并对 KNN 和 SVM 两种分类器在标准语料库和来自 IPGRI 的 Web 文本集进行实验比较分析。

（3）设计并实现基于内容的中英文 Web 文本自动分类系统。

（4）设计基于 ASP.NET 和 ADO.NET 技术的信息管理系统。在此基础上，利用 Web 文本自动分类系统把分好类的 Web 文本和其他信息进行有效的发布和共享，实现 EA-PGR 相关信息和数据的动态管理与发布。

1.4 论文组织

在本章中首先介绍了论文的研究背景和意义，分析了东亚植物遗传资源协作网管理系统建立的必要性和意义，然后从 Web 文件管理面临的问题，引出了 Web 挖掘的一个重要研究方向 Web 文本自动分类研究。接着介绍了 Web 文本分类的概念和实际的应用价值。最后介绍了 Web 文本分类研究在国内外的研究现状。

第二章介绍 Web 文本分类技术的基本过程，包括 Web 文本分类的概念、文本的预处理与分词、特征选择方法、各种文本分类的数学模型和经典算法，文本分类的性能评价以及实验用到的数据集。

第三章介绍了 Web 文本分类的特征提取与选择，包括网页的特征提取与赋权、特征选择方法，

并在实际文本集上对各种特征选择方法进行了实验比较。

第四章介绍了几种 Web 文本分类算法，综合不同规模的训练文档集，以信息增益（IG）、期望交叉熵、互信息（MI）、χ^2 统计（CHI）、文本证据权和 Right half of IG 6 种评估函数进行权重计算，计算公式分别采用词频型和文档型两种方法，用 K 近邻算法（KNN）、SVM 分类算法进行交叉实验和评估。

第五章介绍了作为文本分类研究平台的文本自动分类系统的设计原理与实现过程。

第六章介绍了东亚植物遗传资源 ASP.NET 管理系统的设计方法、总体结构、前后台功能模块设计、数据库设计、系统工作原理以及程序设计主要技术。

第七章对目前工作做了总结并对未来的工作进行了展望。

第二章 Web 文本分类过程概述

Web 文本分类过程涉及到 Web 文本预处理、特征赋权、特征选择、分类算法等几个环节的关键技术，本章概述了 Web 文本分类的总体过程，对设计 Web 文本分类器的思路和采用的技术方法作了分析。

2.1 文本分类的概念

Web 文本的自动分类，是文本自动分类在各种 Web 文档上的应用，是指按照预先定义的主题类别 $C=\{c_1,c_2,...,c_l\}$，根据 Web 文档的内容或属性，将 Web 文档集合 $D=\{d_1,d_2,...,d_n\}$ 中的每个文档 $d_i, i=1,2,...,n$，归到一个或多个类别 $c_k, k=1,2,...l$ 的过程。其中 c_k 可以是并列也可以是分层次组织起来的。自动分类一般考察被分类对象的特征，使之与各种类别中的对象所具有的共同特征（或一定的分类标准、分类参数）进行比较，然后将对象划归为特征最接近的一类，并赋予相应的分类号。

Web 文本分类是一种有监督的学习过程。通过对有类别标记的网页集进行训练，得到分类结果最好时的分类器参数，然后就可以用训练好的分类器对未知类别的网页进行自动分类了。分类的首要问题是采取何种方法表示文件，从而得到计算机可以识别和处理的输入资料。

对于文本的自动分类而言，按照文本的表示方法可以将现有的分类方法分为三类[6]：基于词的分类技术，基于知识的分类技术和基于信息的分类技术。

文本的自动处理是以概念为基本单元，而词是概念的基本组成部分，是信息的载体。所以基于词的分类技术是利用那些可以代表文档主题内容的词汇对文档进行类别判定的。

基于知识的文本自动分类方法主要依赖一个明确的知识库。知识的表示方法主要有规则库、语义模型或者框架。基于知识的分类技术的显著特点是需要手工建造的知识库，且建造的知识库领域性极强，移植困难。有研究表明，在一定的领域内，基于知识的系统能进行快速准确的分类。

基于信息的分类技术是介于基于词的分类技术和基于知识的分类技术之间的方法，该方法对上下文敏感，是一种有选择的概念抽取。该方法用于文本自动分类时，只抽取那些对文本分类有用的信息。它抽取短语及短语周围的文本和潜在的语义信息进行文本类别的确定。

由于基于词的分类方法是一种基于统计的方法，它借鉴了机器学习的理论，把待分类的对象看作是特征向量来进行处理，机制相对简单，在处理大规模真实文本方面取得了令人满意的效果，所以我们把基于词的方法作为解决网页分类问题的主要途径。

2.2 文本自动分类的实现过程

向量空间模型（Vector Space Model ，VSM）是文本分类中广泛使用的模型，它由 Gerard Salton 和 McGill1969 年提出的。向量空间模型（VSM）的基本思想是把文件表示成特征向量，通过相似度来确定文档的类别。在向量空间模型中，每个文档都被表示成形如 $V(d_i)=(t_{i1},w_{i1};t_{i2},w_{i2};...t_{i3},w_{ik};...t_{in},w_{in})$，其中，$w_{ik}=f(t_k,c_j)$，为权值函数，反映特征 t_k 决定文档 d_i 属于类 c_j 的重要性。特征项是从文档中提取的特征词。这样，给定两个文本，它们

的相关程度，就可以用它们在 N 维空间中的相对位置决定，通常取两个向量间夹角的余弦表示相似度。两个向量间的夹角越大，余弦值越小，说明相似程度越小，说明这两个文档分属不同类的概率就越大。反之，若两个向量间的夹角越小，余弦值越大，则这两个文档的相似程度越大，极有可能属于一个类别。两文档的相似程度用公式表示如下：

$$Sim(V_i,V_j)=\cos(V_i,V_j)=\frac{V_i \bullet V_j}{|V_i|\times|V_j|}=\frac{\sum_{k=1}^{N} w_{ik}\times w_{jk}}{\sqrt{\sum_{k=1}^{N} w_{ik}^{\ 2}}\sqrt{\sum_{k=1}^{N} w_{jk}^{\ 2}}} \qquad (2\text{-}1)$$

Web 文本自动分类的过程可以用图 2-1 表示如下：

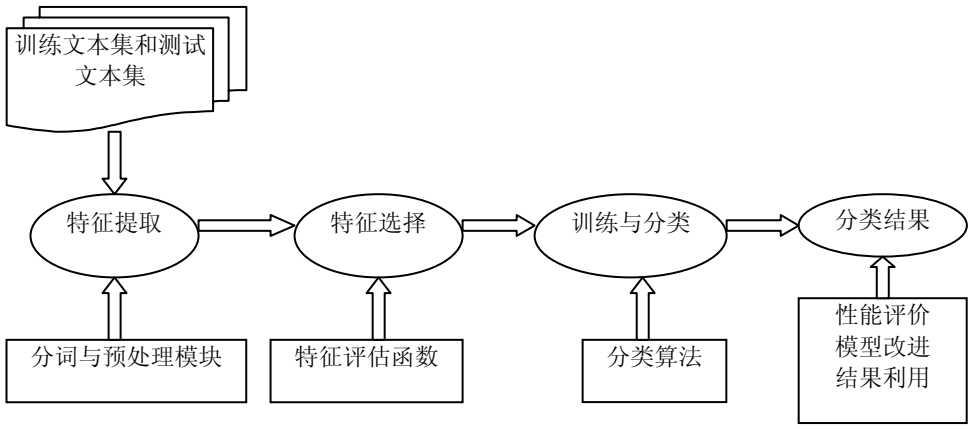

图 2-1　Web 文本分类过程

一个基于 VSM 的分类系统，至少应该包括以下几个部分：

（1）　训练文档集：即由人工分好类的文档组成的集合。这是系统学习分类知识的基础。训练文档集里的文档应该具有广泛的覆盖性和正确的类别标记。训练文档集的质量将直接影响到分类的性能和分类系统的推广程度。

（2）　文本集的分词和预处理：为了把文件表示成计算机能理解的形式，我们要采用一定的模型来表示文件。比如 VSM 模型中，把文件表示成 N 维向量。特征可以用词或者短语来表示。为此，我们必须首先做分词处理。经过分词处理，意义连贯的文档变成词的列表，同时统计的还有各个词在文档中出现的次数。其次还要去掉停用词，英文的如：“the、we、then”等，中文的如：“的、是、我们”等，还可能有同义词处理等语义方面的处理过程。

（3）　特征的选择和压缩：实验证明以未经处理的简单词频作为特征来表示文档，对于分类的意义不大，因此我们必须寻找新的特征评估函数，来给特征对文档分类的贡献打分。这种特征评估函数可以从词的统计信息方面着手，也可以从语义的层次着手，还可以考虑网页这种特殊文本本身所蕴含的一些结构信息作为辅助特征。在文本分类问题中，特征空间的维数往往相当高。例如，在我们的系统中，切词后经过去高频词和去平凡词处理后，还剩下十几万个词。如果简单地把这些词在文件中的出现频率作为特征，那么在这么高的特征维数下，后续计算的复杂度非常高，分类器的训练几乎成了不可能的事。因此需要根据特征评估函数进行特征选择，压缩特征空间的维数，选取那些对分类最有意义的特征。

（4） 分类器的设计与学习：设计适用于文本分类问题的分类器。分类器的输入是向量表示的带有类别标记的训练集文档，输出是该文档被判归属的类别。把分类器的分类结果和正确类别进行对照，选择适合文本分类问题的分类器，并反复调整分类器的参数，直到获得可以接受的分类结果。这是一个典型的有监督学习过程。通过学习，选定适用的分类器，作为下一步分类测试的基础。

（5） 输入文档分类：对于输入的未知类别文档进行分类并给出分类结果和分类评价。

2.3 文本的预处理与分词

我们知道基于统计的方法是把文档当作特征向量来处理的，把文档转换成特征向量一种简单的方法是用词袋表示法（bag-of-words），这种表示法假定文档是由一个个相互独立的特征词组成的，特征词的出现位置无关紧要。为了把文档表示成特征向量需要对文档进行预处理和分词。

2.3.1 文本的预处理

预处理过程对分类效果的影响至关重要，数据是否准备好将直接影响到文本挖掘的效率和准确度以及最终模式的有效性。文本的预处理过程可能占据整个系统的 70％~80％的工作量。

为了对文本建模，必须进行文档预处理过程，即特征提取的过程。对中文网页来说，预处理过程包括网页去标记过程，分词过程和关键词提取过程。

这里需要提出的是，中文文本和英文文本的分词过程有些不同。如果对中文文本进行分类，那么预处理过程较英文文本的预处理过程更为复杂，因为中文的基元是字而不是词，句子中各词语间没有固有的分割符（如空格），因此对中文文本还需要进行词条切分处理，一般需要汉语基本词典来进行词的匹配。

网页与普通文档不同，其所含信息包括内容信息和结构信息，体现在三个方面：网页正文、网页所含的超文本标记和网页间的超链接。其中网页正文包含了大部分的信息，是预处理的重点对象。为了对网页进行分类，首先要将超文本去掉 HTML 标记，转化成普通文本，然后对得到的文本进行包括分词在内的特征提取。具体的网页特征提取方法将在 3.1 中介绍。

在文档的预处理过程中还要统计特征词的词频和文档频率，以便计算特征词的权重。特征词的权重计算方法将在 3.2 中介绍。

2.3.2 中文文本的分词

中文的基元是字，字的信息量比较低，句子中各词语间没有固有的分隔，词能表达完整的语义对象，所以通常选择词作为表示文档的特征，为此首先要对文档进行基于词典的自动分词。

目前汉语分词主要有两大类方法：基于词典与规则的方法和基于统计的方法。基于词典与规则的方法应用词典匹配、汉语词法或其它汉语语言知识进行分词，如：最大匹配法、最小分词法等。这类方法简单、分词效率较高，但对词典的完备性、规则的一致性等要求比较高。基于统计的分词方法则将汉语基于字和词的统计信息，如相邻字间互信息、词频及相应的贡献信息等应用于分词，由于这些信息是通过训练集动态获得，因而具有较好的鲁棒性能，但是完备性相对较差。

系统实现时采用了最大匹配分词法（Maximum Match Method，简称 MM），因为最大匹配分词法是所有方法的基础，它的性能也并不比其它方法差。现在把最大匹配法介绍如下：

首先，机器中应有一个已知的词表，这里称为词典。从要被切分的语句中，按给定的方向截取一定字长的字符串，称这个字符串的长度为最大词长。然后将这个字符串与词典中的词进行匹配，如匹配成功，则确定这个字符串为一个词，将指向被匹配语句的指针按给定的方向移动，移动的距离为该字符串的长度，继续进行下一次匹配；若匹配不成功，则将字符串的长度逐次减去一个字长，再进行匹配，直到成功为止。

我们知道，最大匹配分词法的难点是最大词长的确定，如果最大词长太大，则时间复杂度大大增加，实用性不强；如果最大词长太小，则超过此词长的词将匹配不到，降低了分词的正确率。为了确定最大词长，需要先对汉语词汇的长度进行统计分析。根据对词典的统计，汉语词汇中单个字的词占 5%，两个字的词占 75%，三个字的词占 14%，四个字的词占 6%，其中六个字以上的词约占 0.06%，八个字以上的词小于十万分之一，如果把最大词长定在 6－8 之间，则能保证对汉语中常用词汇的正确切分。考虑到一个重要的常用词“中华人民共和国”，最终决定将最大词长定为 7。

系统采用的中文自动分词算法是在改进了的 MM 算法基础上实现的，并且考虑了标号切分，即遇到非汉字时，开始新的切词。算法的流程如图 2-2 所示。

1．如果 source[pos]>=0xb0 and source[pos+1]>=0xa1，则说明 pos 指向的是一个汉字，于是进行匹配；

2．如果 source[pos]<0xb0 and source[pos+1]>=0xa1，则说明 pos 指向的是一个两个字节表示的符合，于是 pos=pos+2；

3．如果 source[pos]<0xa1，则说明 pos 指向的是一个 ASCII，于是 pos=pos+1；

4．保留未匹配字符串的工作是这样进行的：在遇到匹配或 pos 指向的是符号或 ASCII 码时，如果分词成功标志数 is_seg 为 0，而是否汉字标志数 is_word 为 1，则从上次切分的未知开始到 pos-1 为止，保留这段字符，作为未匹配字符串。

5．异常处理是在 source[pos]>=0xb0 and source[pos+1]<0xa1 的情况下进行的，因为对于正常的 GB 编码的汉字文本是不会出现上述情况的。在遇到上述异常时，系统退出分词进程。

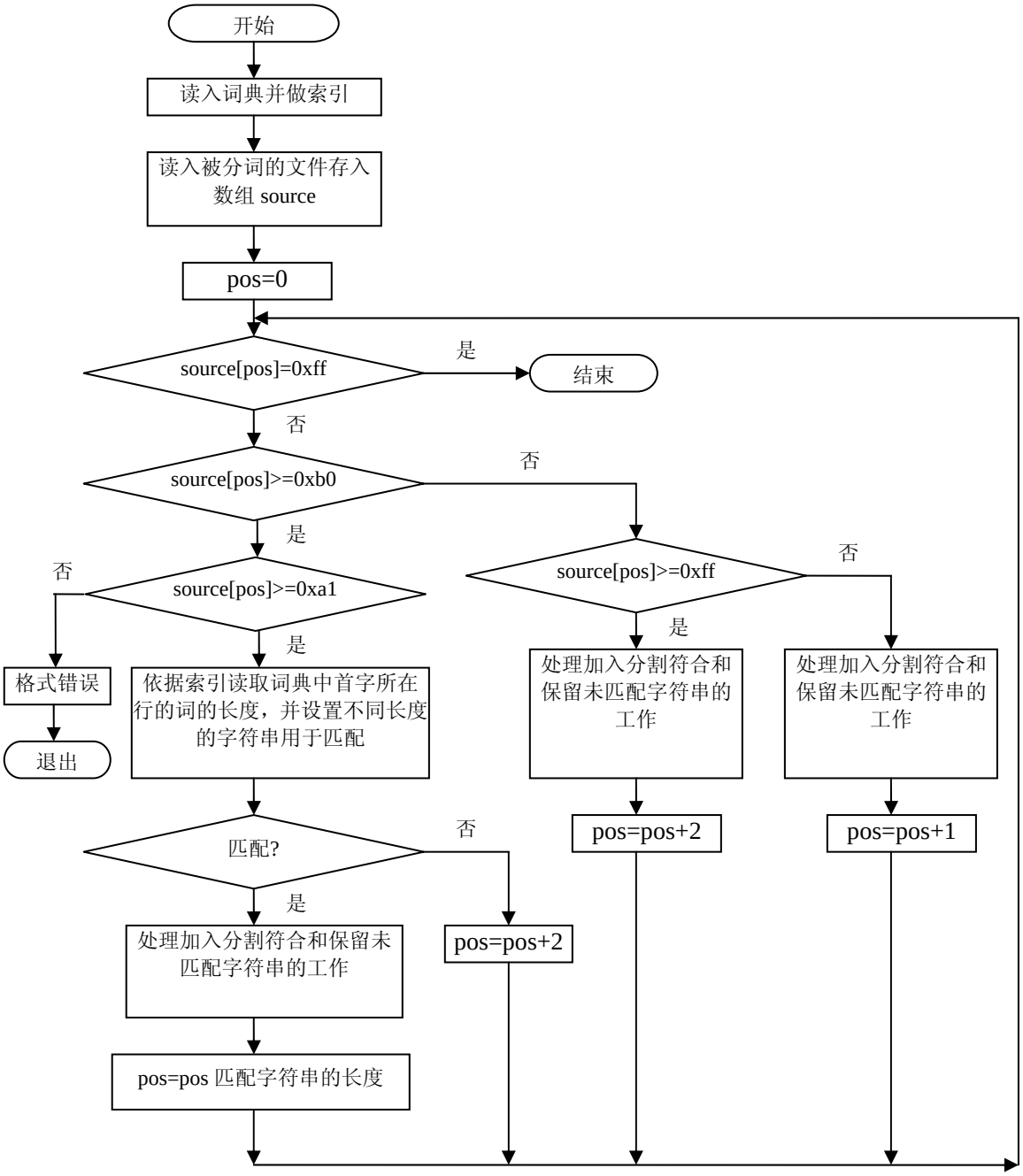

图 2-2 中文自动分词算法流程

2.4 分类器的设计与训练

文本分类一般分为训练和分类两个阶段，具体过程如下：

训练阶段：

（1） 定义类别集合$C=\{C_1,C_2,...C_m\}$，共分成 m 个类别，这些类别可以是层次的，也可以是平面的。

（2） 给出训练文本集合$S=\{s_1,s_2,...s_n\}$，每个训练文档s_j被标上所属的类别标识C_i。

（3）　统计 S 中所有文本的特征向量$V(s_j)$写成输入文件，根据特定的分类机制来训练分类参数，用训练文本集的类别标签来监督学习过程，直到达到满意的分类精度为止，此时分类器训练完成。

分类阶段：

（4）　对于测试文本集合$T=\{d_1,d_2,...,d_l\}$中的每个待分类文档d_k，将特征向量$V(d_k)$输入文本分类器进行判分。

目前存在多种基于向量空间模型的分类机制，包括基于文本距离（或者说相似度）的方法，如基于文本相似度的类中心向量法，Rocchio 算法、K-最近邻算法（K－Nearest Neighbor，简称 KNN），线性核函数的支持向量机算法（SVM）和基于独立性假设的概率模型方法，如 Naïve Bayes 方法，以及基于映射机制的方法如最小二乘法，神经网络方法等。

在向量空间模型中，文档与文档之间，或者文档与类别之间的相似度用代表文档或者类别的向量V_i和V_j之间的夹角余弦来表示，计算方法如公式 2-1 所示。

基于文本相似度的类中心向量法，简称为质心向量法，由于其分类机制简单，速度较快的优点，它又是其他分类算法的基础。为此，我们介绍文本相似度方法的分类机制。

根据输入的训练样本及相应的类别标记，得到各类的类中心向量，例如第 j 类的类中心向量为$V_j=(w(t_{j1}),w(t_{j2}),...,w(t_{jn}))$，n 为特征空间维数。若第 j 类共有$N_j$，第 m 个样本的向量表示为$d_{jm}=(w(t_{m1}),w(t_{m2}),...,w(t_{mn}))$，那么类中心向量$V_j$计算如下：

$$V_j=\frac{\sum_{m=1}^{N_j}d_{jm}}{N_j} \tag{2-2}$$

于是第 j 类中心向量V_j的第 k 个分量$w(t_{jk})=\frac{\sum_{m=1}^{N_j}w(t_{mk})}{N_j}$。

若待分类的测试文本d_i的向量表示为$(w(t_{i1}),w(t_{i2}),...w(t_{in}))$。那么计算测试样本$d_i$和各类中心向量的余弦相似度，把该测试样本判分为相似度最大的那一类。即有：

$$C=\max_j\cos(d_i,V_j)=\frac{d_i\bullet V_j}{|d_i|\times|V_j|}=\frac{\sum_{k=1}^{N}w_{ik}\times w_{jk}}{\sqrt{\sum_{k=1}^{N}w_{ik}^{\ 2}}\sqrt{\sum_{k=1}^{N}w_{jk}^{\ 2}}} \tag{2-3}$$

分类算法是文本分类的核心问题。在机器学习和模式识别领域，人们设计出了很多种分类器，但不是每种分类器都对我们的问题适用。因此寻找最合适文本分类问题的分类算法是研究的关键所在。

如同任何其他模式识别问题一样，特征提取和特征选择同样是关系到最终文本分类性能的重要环节。特征中蕴含了模式信息，表征了文本内容。提取强类别表征意义的特征将对分类极为有利。任何分类器顺利工作的前提是给它输入一定数量的好的样本特征，特征好坏的评价标准主要看该特征是否表达了样本的类别属性，即该特征是否有利于样本的分类。

举例来说，为了把 A 类样本$\{S_1\}$和 B 类样本$\{S_2\}$区分开来，我们取了某种特征度量 Measure。

如果对于不同类别的样本，Measure 的取值范围相距很远，而同类样本的 Measure 值在一个比较小的变化范围内，比如对于 A 类样本，Measure 取值在 0 和 1 之间，而对于 B 类样本，Measure 取值在 10 和 11 之间，那么我们有理由相信 Measure 是一个很好的特征。反之，如果对于 A 类样本 Measure 取值在 0 和 1 之间，而对于 B 类样本，Measure 取值在 0.3 和 1.3 之间，那么 Measure 用来做区分类别特征的有效性就值得怀疑。可见，提取和选择好的特征对于分类有着显著的帮助作用。

2.5 文本分类的性能评价

评价自动分类的性能指标有分类准确率、查准率、查全率以及综合考虑查准率和查全率的 F 值，为了对以上几个指标进行说明，建立如表 2-1 所示的二值分类问题的列联表。

表 2-1 二值分类列联表

	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*

（1）准确率（Accuracy）：

分类准确率是指在所有类别中，准确分类的样本数占样本总数的比率。它是衡量文本分类性能的重要指标。数学公式表示如下：

$$A=\frac{a+c}{a+b+c+d} \tag{2-4}$$

对于每一类样本来说，查准率和查全率衡量了分类系统对于该类样本的分类性能。

（2）查准率（Precision）：

查准率是所有判分的文本中与正确类别吻合的文本所占的比率，其数学公式表示如下：

$$p=\frac{a}{a+b} \tag{2-5}$$

（3）查全率（Recall）：

查全率是分类系统正确分类的文本在该类别应有的文本中所占的比率，其数学表示如下：

$$r=\frac{a}{a+c} \tag{2-6}$$

（4）F 方法：

查准率和查全率反映了分类质量的两个不同方面，两种必须综合考虑，不可偏废。但一般来说，随着查准率的提高，查全率反而下降。因此除了准确率和查全率这两个最基本的评价标准之外，还有其它的一些评价指标，比如 F 方法，两者相对重要性用一个参数 β 来刻画。其数学公式如下：

$$F_{\beta}=\frac{(1+\beta^2)\cdot p\times r}{\beta^2\times p+r} \tag{2-7}$$

当 $\beta=1$ 时，即准确率和查全率在评估函数中有着同样的重要性，称为 F1 度量：

$$F1=\frac{p\times r\times 2}{p+r} \tag{2-8}$$

当$\beta<1$时强调 precision 的作用，$\beta>1$时强调 recall 的作用。

2.6 实验用到的文本集

Web 文本分类系统是用来对东亚植物遗传资源信息进行分类的，所以实验数据集必然要用东亚植物遗传资源信息作为训练文本集和测试文本集。同时 Web 文本分类还处于研究阶段，为了比较分析分类算法的好坏，必然要用到其它的 Web 文本数据集作为参照。

实验分析的英文 Web 文本集来自于国际植物遗传资源研究所的网上资料（www.ipgri.org），作者在 IPGRI 的文本里随机挑选了 700 篇文档，500 篇作为训练文档集，200 篇作为测试文档集，分成五类存放，每类 Web 文本的划分都是来自领域内的专家划分的，所以具有一定的权威性。训练集和测试集是按照 Lewis 的划分标准（训练集与测试集没有交集）进行划分的[21]，训练 Web 文本存在放在文件夹 entrain500，测试 Web 文本存放在文件夹 entest200 下，训练样本的数量和测试样本的数量的比是 2.5：1。类别和文本数如表 2-2 所示。

表 2-2 国际植物遗传资源训练和测试 Web 文本类别和数量

	类别	entrain500	entest200
1	collaborative-activity	100	40
2	country-profile	100	40
3	crop	100	40
4	Human-resource	100	40
5	general	100	40

设计的是中英文文本分类系统，为了对 Web 中文文本进行测试和实验，中文实验数据选用了自复旦大学的文本分类语料库语料，训练集和测试集也是按照 Lewis 的划分标准进行划分的，是经过相关专家整理并分类的，具有一定的权威性。中文训练文本集 chtrain2000 和测试文本集 chtest800，各包括 10 个类别，类别和文本数如表 2-3 所示。

表 2-3 中文训练文本集类别和文本数

	类别	chtrain2000	chtest800
1	计算机	134	66
2	交通	143	71
3	环境	134	67
4	经济	217	108
5	医药	136	68
6	军事	166	83
7	政治	338	167
8	体育	301	149
9	艺术	166	82
10	教育	147	73
	共计	2025	866

第三章　Web 文本分类的特征提取、特征选择与实验比较

对于任何模式识别问题来说，提取和选择合适的特征来表征输入样本，直接关系到模式学习和分类器的设计成败，Web 文本分类也不例外。本章主要介绍了 Web 文本分类中特征提取、特征选择和特征赋权方法，并结合实验结果就各种特征选择方法和特征赋权方法的性能做了比较和分析。

3.1　Web 文本的特征提取

绝大多数现有的文本分类器都使用所谓“词袋表示法”（bag-of-words）来表示文本。这种表示法有一个关键的假设，就是文章中词条出现的次序是无关紧要的，所以可以把文本看做一系列词条的集。在标准的词袋表示中，不考虑文本中单词的位置信息。

基于内容的网页分类试图在理解网页内容的基础上作分类，但由于计算机尚未发展到像人那样真正理解自然语言的地步，只能通过离散化的特征建模，来近似表示语义。此外，网页的结构信息如 HTML 语言中的 tag 标记起到了对重要信息的提示和强调作用，对确定网页的内容所属类别也有帮助。因此，如何提取网页特征是进行网页分类遇到的第一个重要问题。

网页分类的特征提取主要包括网页 tag 特征的提取和正文内容的特征提取。前者利用网页的 HTML 语法特点，提取特定的标记信息，后者主要是自动分词过程。

3.1.1　Web 文本的 Tag 特征提取

由 HTML 语言创作的网页是由一系列元素组成，HTML 中用不同的元素实例来表示文本、显示图像、超文本链接和描述文字等。元素的描述一般由开始标记、内容和结束标记组成，也有个别元素没有内容或者结束标记。开始标记在 HTML 中标记为<元素名称>，对应的结束标记为</元素名称>，内容出现在开始标记和结束标记之间。

在 HTML 语言的标准中大约定义了 180 多个元素，从功能上可以大致分为 7 类：网页框架元素、字符风格控制元素、版面控制元素、标题分级控制元素、锚元素、表格元素和交互元素。各类元素的实例如表 3-1 所示。

表 3-1　HTML 语言标记

网页框架元素	<HTML>、<HEAD>、<TITLE>、<META>等
字符风格控制元素	<H1>、<H6>、<B>、<I>、<U>等
版面控制元素	<PRE>、<ALIGN>、<P>、<HR>等
标题分级控制元素	<UL>、<OL>、<LI>等
锚元素	超文本链接元素<a href=“链接地址”>
表格元素	<TABLE><CAPTION><TR><TH>等
交互元素	<INPUT TYPE=“TEXT”>

在这 7 类元素中，网页框架元素和字符风格控制元素对于提取网页的结构信息尤为有用。例如超文本标题标记<HEAD>和</HEAD>之间的内容是概括网页内容的标题，而<TITLE>和

</TITLE>之间的内容是网页窗口的标题栏显示文字，通常对网页内容起概括作用。<META>和</META>之间的内容是描述网页 HTML 文档的元信息。字符风格控制元素，如定义字体大小的<H1></H1>到<H6></H6>标记，粗体标记<B></B>，斜体标记<I></I>等往往是为了标记网页上的重要内容。在这些标记之间的文字对于理解网页内容，进而按内容分类有重要意义，因此，根据这些标记类提取 Tag 特征，反映了网页的结构信息。在特征提取步骤，提取网页结构信息是网页分类与普通文本分类的唯一不同之处。

我们利用网页中特征的词频信息及特征间的相关信息、网页标记信息，为每个特征赋予一个可调的加权参数。网页分类与文本分类同样需要降维，纯文本分类中使用信息增益、互信息、期望交叉熵、文本证据权、χ^2 统计等特征评估函数，以及特征选择方法和权值调整技术，也可用于网页分类。但是 Web 网页的特征的结构信息在特征权重计算方面与纯文本又不同。

如前面所述，在<title>中出现的特征是最重要的，它概括和总结了整个网页的内容，因此在分类中起关键作用。基于网页的结构特征的权重函数 $SWF(t_{ik})$ 定义如下：

$$w_{ik} = SWF(t_{ik}) = \sum_{s}(\alpha_s * Fweight(t_{ik,s})) = \sum_{s}(\alpha_s * TF(t_{ik,s}) * TEF(t_{ik,s})) \quad (3\text{-}1)$$

其中，$s \in S$，α_s 表示出现在网页 d_i 中带有标记 s 的特征 t_{ik} 的权重调整因子，根据标记 s 的不同而异，如下表。

表 3-2　tag 权重系数表

tag	α_s	tag	α_s
<Title>	4	<Head>	2
<H1>	4	<B>	2
<H2>	2	<I>	2
<H3>/<H4>/<H5>/<H6>	1	<U>	2

$TF(t_{ik,s})$ 表示带有标记 s 的特征 t_{ik} 出现的词频数，$TEF(t_{ik,s})$ 表示使用权值评估函数打的分值。$Fweight(t_{ik,s})$ 是权值函数。

3.1.2　Web 文本的内容特征提取

如果忽略 HTML 网页的各种标记，将正文内容作为普通文本来进行特征提取，提取的就是内容特征，显然这部分是网页信息的主体，是进行基于内容的网页分类所必须考虑的主要特征。对于英文网页来说，进行内容特征提取，主要包括去 HTML 标记操作，分词操作以及去平凡词操作等。去 HTML 标记把网页变成普通文本。平凡词是指那些起结构作用而没有语义，在所有文档中都频繁出现的词，如：定冠词、不定冠词和代词等。去除平凡词一般通过建立平凡词表来对分词提取的特征词集合进行过滤操作。内容特征提取过程如下图：

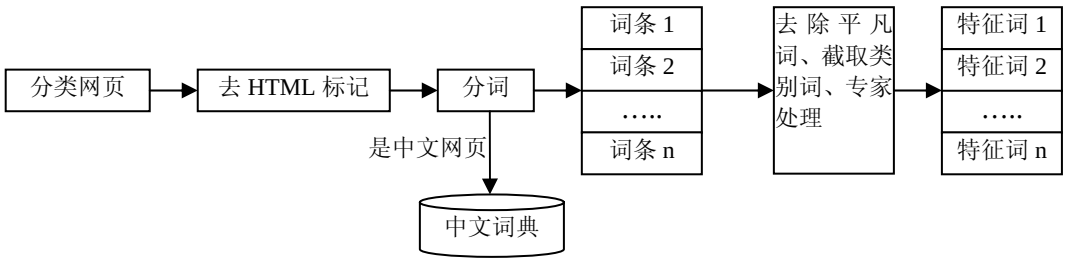

图 3-1　网页特征提取过程

在分词过程中得到各个特征词的出现频率，包括在每篇网页中的出现频率和在整个网页集中的多少篇网页中出现，前者被称为特征词对于网页的词频（Term Frequency，简称 TF），后者被称为特征词的文档频率（Document Frequency，简称 DF），汇总分词结果中所有特征词的频率形成的原始词频特征文件是下一步进行特征选择和计算各个特征权值的重要依据。

3.2　Web 文本分类中的特征赋权

向量空间模型把文档表示成向量形式 $D_i=(w(t_{i1}),w(t_{i2}),...w(t_{ik}),...w(t_{in}))$，其中 t_{ik} 表示第 k 个特征，$w(t_{ik})$ 对应于文档 D_i 的第 k 特征项的权重，权重一般取做词频的函数，有布尔型、简单词频、TFIDF 等几种形式，特征赋权试图通过词频和文档频率的函数来突出重要特征在整个向量表示中的影响力，而通过减少权重来弱化另外一些特征的影响力。

有很多种方法可以用来计算特征项 t_{ik} 在文档 D_i 的权重 $w(t_{ik})$，这些方法大部分都基于这样的经验知识：特征项 t_{ik} 在文档中出现的次数越多，它和文档主题的相关程度越高；特征项 t_{ik} 在整个文档集中出现的文档数越多，它对于文档主题的区分能力越弱。

经过特征提取，得到第 k 个特征词 t_k 在第 i 个网页中的词频值为 $TF(t_{ik})$，和在整个网页集中出现的网页数，即文档频率 $DF(t_k)$，那么 t_k 在第 i 个网页向量中的权重 $w(t_{ik})$ 计算如下：

（1）　最简单的布尔型：$w(t)=\begin{cases}0 & TF(t_{ik})=0\\ 1 & TF(t_{ik})=1\end{cases}$ 文本向量由 0，1 组成，

（2）　词频型：$w(t_{ik})=TF(t_{ik})$

（3）　TF-IDF 公式：$w(t_{ik})=TF(t_{ik})*\log(\frac{N}{DF(t_k)})$，其中 N 为文本集的文本总数。

（4）　归一化 TF-IDF 公式：$w(t_{ik})=\frac{TF(t_{ik})*\log(\frac{N}{DF(t_k)})}{\sqrt{\sum_{k=1}^{m}[TF(t_{ik})*\log(\frac{N}{DF(t_k)})]^2}}$，其中 N 为文本集的文本总数，m 为特征总数。

（5）　熵加权：

$$w(t_{ik})=\log(TF(t_{ik})+1)*(1+\frac{1}{\log N}\sum_{j=1}^{N}[\frac{TF(t_{jk})}{n_k}\log(\frac{TF(t_{jk})}{n_k})])$$

特征赋权方法影响了文本向量的表示方式，必然对文本分类的结果有一定影响，归一化 TF-IDF 方法综合考虑了特征的词频、文档频率，以及文档长度三个方面，是文本分类和信息检

索领域应用最为广泛的特征赋权方法。

词频型和文档型的特征赋权方法比较：

训练数据集：entrain500（国际植物遗传资源网站资料信息）；测试数据集：entest200.

实验目的：比较不同特征赋权方法在 SVM 分类器中的优劣。

实验方法：采用支持向量机（SVM）的分类方法

表 3-3　不同特征赋权方法的比较

F1 度量 / 特征选择方法	特征词数：1000		特征词数 2000	
	词频型	文档型	词频型	文档型
IG	88.50	88.00	90.00	88.50
MI	59.50	58.00	71.50	72.00
CHI	88.50	88.00	88.50	87.00
Right half of IG	87.50	87.50	88.50	89.00
文本证据权	89.50	88.50	89.00	88.00
交叉熵	88.50	88.00	90.00	88.50

实验分析：在其它条件相同的情况下，不管采用何种特制选择方法，基于词频统计的特征赋权方法始终比基于文档统计的特征赋权方法要好，因为文档频率法只简单的统计词是否在文档中出现，所以精度没有词频统计法的高。

3.3　Web 文本分类中的特征选择

用词袋表示法来表示网页时，特征向量会达到数十万维的大小，即使经过删除停用词表中的词以及应用 ZIP 法则删除低频词，仍会有数万特征留下，如此高维的特征空间对于分类问题来说极为不利，不仅增加了分类器运算的时间和空间复杂度，而且使得某些分类器的训练无法进行。实际上，最后一般只选择 2%～5%的最佳特征来作为分类依据。所以有必要进一步对特征进行精选。

在文本分类中，常用的特征选择方法有基于阈值的统计方法，如文档频率方法（DF），信息增益方法，互信息方法，CHI 方法，期望交叉熵，文本证据权；基于词频覆盖度的特征选择方法等以及由原始的低级特征（比如词）经过某种变换构建正交空间中的新特征的方法，如主量分析方法和潜在语义索引等。基于阈值的统计方法具有计算复杂度低，速度快的优点，尤其适合做文本分类中的特征选择。

KNN 和 SVM 是当前最好的两种分类算法，分类速度较快，分类精度较高。为了研究和比较特征选择方法的优劣，因此选用 KNN 和 SVM 这两种分类器来验证特征选择的效果。

根据文本分类的特点，用机器学习中的评估函数来进行文本分类的特征选择，它通过计算各个特征词的评估分数，来衡量词条与各类别之间的相关程度。常用的文本特征选择方法有文本频率法、信息增益法、期望交叉熵法、互信息法、χ^2 统计量法、文本证据权法和几率比法等。已经有很多研究者在特征选择方法方面做了大量的研究工作，其中以美国的卡内基梅隆大学的 Yang Yiming 和斯坦福大学的 Mechra Sahami 的研究最具有代表性和总结性，国内外的其它研究者也研究出了很多新的算法和方法，但都是以这几种方法作为基础的。所以很有必要介绍这几种方法。

下面，为了便于我们对文本分类中常用的特征选择方法进行讨论，先引入如下符号：

A—属于类别c且包含特征w的训练文本个数，

B—不属于类别c（或属于类别$\bar{c}$）且包含特征w的训练文本个数，

C—属于类别c且不包含特征w的训练文本个数，

D—不属于类别c且不包含特征w的训练文本个数，

M—属于类别c文本个数，

N—训练文本总数。

显然有$A+C=M$，和$A+B+C+D=N$。

（1）文本频率（Document Frequency，DF）

$$TEF_{df}(w)=\text{在训练文本集中特征}w\text{出现的文本个数}=A+B \quad (3\text{-}2)$$

利用 DF 特征选择方法降维是一种最简单的方法，求特征词的 DF 值计算简单，效果也不错，但是它的缺点是，有时 DF 值很小的特征词在某一类文件中，包含着重要信息，这些特征词对分类作用很大。

（2）信息增益（Information Gain，IG）

信息增益（Information gain，IG）是机器学习中的概念，用在决策树中来计算特征的权值，信息增益被定义为类特征向量的平均值。

$$TEF(w)=I(c,w)=I(w,c)=$$

$$P(w)\sum_t P(c_t\mid w)\log\frac{P(c_t\mid w)}{P(c_t)}+P(\bar{w})\sum_t P(c_t\mid \bar{w})\log\frac{P(c_t\mid \bar{w})}{P(c_t)} \quad (3\text{-}3)$$

在基于词频的特征向量表示法中，其中，$P(w)$表示特征词w在文本集中出现的概率，可以用w在训练文本集中出现的总词频除以训练文本集中所有特征的词频总和来估计。$P(c_t\mid w)$表示w出现，文本属于c_t的概率，可以用属于c_t且包含w的文本中w的词频总数除以集合中属于c_t类的概率。

另外，在基于文档型（也称布尔型）的方法中，$P(w)$可以用文本集中包含该词的文本数目除以文本集合中的所有文本数目来估算，所以，信息增益可以用下公式来计算

$$TEF(w)=IG(c,w)=\frac{A+C}{N}\log\frac{A+C}{N}+\frac{1}{N}[(A+D)\cdot\frac{A}{M}\cdot\frac{1}{A+B}\cdot\log(\frac{A}{M}\cdot\frac{1}{A+B})] \quad (3\text{-}4)$$

（3）互信息（Mutual Information，MI）

互信息是信息论中的概念，它用于度量一个消息中两个信号之间的相互依赖程度。在特征选择领域中人们经常利用它来计算特征w与类别 c 之间依赖程度，将特征w与各个类的互信息融合起来作为特征的权重。特征w与类c的互信息$MI(w,c)$计算公式如下：

$$MI(w,c)=\log\frac{p(w,c)}{p(w)\times p(c)} \quad (3\text{-}5)$$

$p(w,c)$定义为w和c的同现概率，$p(w)$定义为w出现的概率，$p(c)$定义为类别c的概率。

用词频来代替概率，互信息$MI(w,c)$也可以用下面的公式来计算

$$MI(w,c)=\log\frac{A\times N}{(A+C)\times(A+B)} \quad (3\text{-}6)$$

（4）χ^2统计量（Chi-square Statistic，CHI）

χ^2 统计量特征选择方法又被称作开方拟合检验（CHI， χ^2-test），这个概念来自列联表检验(Contingency Table Test)，它可以用来衡量特征 x 与类别 c 之间的统计相关性。其计算公式如下：

$$CHI(w,c)=\frac{N\cdot[p(w,c)\cdot p(\overline{w},\overline{c})-p(w,\overline{c})\cdot p(\overline{w},c)]^2}{p(w)\cdot p(\overline{w})\cdot p(c)\cdot p(\overline{c})} \tag{3-7}$$

用各个事件的频率代替其相应的概率，χ^2 统计量 $CHI(x,c)$ 可以用下式来近似计算：

$$CHI(w,c)\approx\frac{N\cdot(A\cdot D-B\cdot C)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} \tag{3-8}$$

考虑到 N， $A+C$ 和 $B+D$ 均是常数，上式可以进一步简化为

$$CHI(w,c)\approx\frac{(A\cdot D-B\cdot C)^2}{(A+B)\cdot(C+D)} \tag{3-9}$$

当特征 w 与类别 c 相互独立时，$CHI(w,c)=0$，此时特征 w 不包含与类别 c 有关的鉴别信息。$CHI(w,c)$ 的值就越大，此时特征 w 包含的与类别 c 有关的鉴别信息就越多。

（5）期望交叉熵（expected cross entropy, ECE）

期望交叉熵所衡量的是在获知一个特征文本中出现时所获得的信息量。词 w 的期望交叉熵定义为：

$$ECE(w,c)=P(w)\sum_t P(c_t\mid w)\log\frac{P(c_t\mid w)}{P(c_t)} \tag{3-10}$$

（6）文本证据权（the weight of evidence for text）

文本证据权是一种新的特征选择方法，它衡量类的概率和给定特征时类的条件概率之间的差别。文本处理中，不需要计算 w 的所有可能值，而只考虑 w 在文本中是否出现。

$$WET(w,c)=P(w)\sum_t P(c_t)\left|\log\frac{P(c_t\mid w)(1-P(c_t))}{P(c_t)(1-P(c_t\mid w))}\right| \tag{3-11}$$

（7）几率比（Odds Ratio，OR）

几率比的定义如下：

$$OR(w,c)=\log\frac{odds(w\mid pos)}{odds(w\mid neg)}=\log\frac{P(w\mid pos)(1-P(w\mid pos))}{P(w\mid neg)(1-P(w\mid neg))} \tag{3-12}$$

其中，pos 代表正类，neg 代表负类，$P(w\mid pos)$ 表示特征 w 在正类 pos 中出现的概率，$P(w\mid neg)$ 表示 w 在负类 neg 中出现的概率。

分析 Odd Ratio 的定义，它不是象前面介绍的评估函数那样将所有类同等对待，而是只关心目标类值，所有几率比特别适合用于二元分类器。在二元分类中，我们希望能识别出尽可能多的正类，而不关心识别出负类。

3.4　特征选择实验比较与结论

为了比较各个特征选择评估函数的优劣，选择适合 Web 文本分类器的特征选择评估函数，我们用文本分类器上进行了比较实验。

SVM 分类器上不同特征选择方法实验比较与结论：

训练数据集：entrain500（国际植物遗传资源网站资料信息）；测试数据集：entest200。

实验目的：比较不同特征选择评估函数在 SVM 分类器中的优劣。

实验方法：采用支持向量机（SVM）的分类方法

表 3-4　不同特征选择方法的比较

F1 度量 / 特征选择方法	特征词数：1000		特征词数 2000	
	基于词频统计	基于文档统计	基于词频统计	基于文档统计
IG	88.50	88.00	90.00	88.50
MI	59.50	58.00	71.50	72.00
CHI	88.50	88.00	88.50	87.00
Right half of IG	87.50	87.50	88.50	89.00
文本证据权	89.50	88.50	89.00	88.00
交叉熵	88.50	88.00	90.00	88.50

实验结果分析：

1.不论特征词数取多少，MI 的结果最差，测试的 F1 度量都不超过 75%。YangYiming 曾对此进行了解释，她认为这是由于 MI 方法在选择特征时，偏爱那些出现频率低的词。

2.期望交叉熵和信息增益（IG）相比，两者对 SVM 的分类精度基本相等。

3. CHI 统计量法的结果始终比信息增益和期望交叉熵的差。

4.由于 Right half of IG 只取信息增益的右半部分，所以它的分类结果也比信息增益和期望交叉熵的分类结果差。

5.在其它条件相同的情况下，基于词频统计的特征赋权方法比基于文档统计的特征赋权方法要好，因为文档频率法只简单的统计词是否在文档中出现，所以精度没有词频统计法的高。

6.在其它条件相同的情况下，提取的特征词数多比特征词数少的分类精度高，但是以牺牲训练的时间为代价的。

实验结论：

对 SVM 分类器而言，最好的特征选择评估函数是 IG 和期望交叉熵法，其次是文本证据权、CHI、Right half of IG、MI。基于词频统计的特征赋权方法比基于文档频率的特征赋权方法好，提取的特征词数量越多越有利于分类精度的提高。

KNN 分类器上不同特征选择方法实验比较 (表 3-5)：

训练数据集：entrain500（国际植物遗传资源网站资料信息）；测试数据集：entest200。

实验目的：比较不同特征选择评估函数在 KNN 分类器中的优劣。

实验方法：采用 KNN 的分类方法

K 的取值：35

实验结果分析：

1.不论特征词数取多少，MI 的结果最差，测试的查准率都不超过 55%，可以认为它不适合应用在 KNN 分类器中。

2.期望交叉熵和信息增益（IG）相比，两者对 KNN 的分类精度也基本相等。

3. CHI 统计量法的结果始终比信息增益和期望交叉熵的强，这一点正好和 SVM 分类器相反。

4.由于 Right half of IG 只取 IG 的右半部分，所以它的分类结果也比 IG 和期望交叉熵的分类结果差。

表 3-5　不同特征选择方法的比较

F1 度量 / 特征选择方法	特征词数：1000		特征词数 2000	
	基于词频统计	基于文档统计	基于词频统计	基于文档统计
IG	80.00	82.50	77.00	80.00
MI	23.40	18.30	39.90	51.51
CHI	80.00	83.00	78.50	81.00
Right half of IG	79.50	82.00	76.00	80.00
文本证据权	79.50	82.50	76.00	82.50
交叉熵	80.0	83.00	77.00	80.00

5.在其它条件相同的情况下，基于文档统计的特征赋权方法比基于词频统计的特征赋权方法要好，这一点也和 SVM 分类器相反。

6.在其它条件相同的情况下，单纯提高提取的特征词数量，未必能提高分类的精度，这一点也和 SVM 分类器相反。

实验结论：

对 SVM 分类器而言，最好的特征选择评估函数是 CHI，然后依次是：IG 和期望交叉熵法、文本证据权、Right half of IG、MI。基于文档统计的特征赋权方法比基于词频频率的特征赋权方法好，提取的特征词数量多并不一定有利于分类精度的提高。

第四章 Web 文本分类算法与实验比较

分类算法是文本分类的核心，近年来的大量统计学习的方法越来越多地应用到文本分类这一领域，常用的分类算法有：决策树、K-最近邻、朴素贝叶斯、神经网络、支持向量机、归纳学习等算法，但是到底哪种分类算法适合我们的网页文本分类问题，能够取得比较好的分类性能呢？在各种现有分类算法的基础上，是否能设计出新的适合文本分类问题的分类算法呢？这些都是值得我们研究的问题，也是本章讨论的重点。本章首先介绍了文本分类领域最常用的几种分类算法，包括类中心向量法，Naïve Bayes 方法，KNN 方法和 SVM 方法，并在网页集上作了实验，根据实验结果对各种分类器的性能作了分析和比较。

4.1 类中心向量分类法

类中心向量分类算法是基于向量空间模型和最小距离法，是一种最简单的有导师的学习方法，也可以看成是其它分类算法的基础。在向量空间表示法中，每个文本由一个特征向量表示。于是很自然的想到用两个特征向量之间的距离来衡量两个文本的近似程度，这也是矢量相似度法的基本思想。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，然后在测试文本到来时，确定新文本向量，计算该向量与每类中心向量的距离（相似度），最后判定文本属于与该文本距离最近的类。

类中心向量分类算法分训练和分类两个阶段，具体如下：

（1） 训练阶段

step1：定义类别集合$C=\{c_1,c_2,...,c_m\}$。这些类可以是层次型的，也可以是并列型的。

step2：给出训练文档集合$D=\{d_1,d_2,...,d_n\}$，每个训练文档d_j被标上所属类别标识c_i。

step3：统计 D 中所有文档，确定特征矢量$V(d_j)$，再根据$V(d_j)$确定代表每个类别c_i的特征矢量$V(c_j)$。

（2） 分类阶段

step1：对于测试文档集$T=\{d_1,d_2,...,d_n\}$中的每个待分类文档d_k，计算其特征矢量$V(d_k)$与每个$V(c_i)$之间的相似度$sim(d_k,c_i)$。

$$C=\max_j \cos(d_i,V_c)=\frac{d_i \bullet V_c}{|d_i|\times|V_c|}=\frac{\sum_{k=1}^{N} w_{ik}\times w_{ck}}{\sqrt{\sum_{k=1}^{N} w_{ik}^{\ 2}}\sqrt{\sum_{k=1}^{N} w_{ck}^{\ 2}}} \qquad (4\text{-}1)$$

Step2：选取相似度最大的一个类别$\max_{c\in C} sim(d_k,c_i)$作为$d_k$的类。

有时只要d_j与这些类别间的相似度超过某个阈值，可为d_j指定多种类别。但若这种情况发生得太频繁，则说明预定义类别$C=\{c_1,c_2,...,c_m\}$不当，应加以修改。当文档 d 与所有类的相似度都低于该阈值，则将其标注为“其它”类。

4.2 KNN 分类法

KNN（K-Nearest Neighbor）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，由这 K 篇文本所属的类别来判定新文本应该属于哪一类。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，根据特征集对新文本分词，确定新文本的向量表示。

3） 在训练文本集中选出与新文本最相似的 K 个文本。

4） 依次计算新文本属于各个类的加权相似度，计算公式如（4-3）所示：

$$P(x,C_j)=\sum_{i=1}^{K} sim(x,d_i)y(d_i,C_j) \qquad (4\text{-}3)$$

其中 x 表示新文本的特征向量，$P(x,C_j)$ 表示新样本属于第 j 类的相似度，而 $y(d_i,C_j)$ 为类别属性函数，当 d_i 属于 C_j 时 $y(d_i,C_j)$ 为 1，否则 $y(d_i,C_j)$ 为 0。

5） 比较新文本同各类的相似度，将文本分到相似度最大的那个类别中。

目前，K 的确定还没有一个公认的标准，一般通过实验的方法确定比较合适的 K 值，表 4-1 描述了 KNN 分类查准率和 K 的关系，实验训练文本集采用 chtrain2000，测试文本集采用 chtest800，特征选择方法采用 IG 法，权重采用文档统计方法。

表 4-1 KNN 分类查准率和 K 的关系

K	1	5	10	15	16	20	30	35	40	50
查准率（%）	85.0	78.5	77.5	76.0	77.0	76.5	78.5	80.0	81.0	81.5

从实验结果可以看出，K 的取值对分类性能有一定的影响，但 K 的值对分类准取率的影响只在 4%的范围内变化。一般认为当 K 取 40 左右时，KNN 的分类准确率达到最高。Yiming Yang 在文献[23]里指出 K 的取值通常可以选 30 或 40。

4.3 Naïve Bayes 分类方法

朴素贝叶斯学习模型（Simple Bayes 或 Naïve Bayes）是一种最常用的基于概率的有导师分类算法。它将训练文档集 D 分解成文档特征向量和决策主题类向量，也就是各分量（各个特征）是独立地作用与决策类别变量。换句话说，就是假设各个特征词分布独立，所有的词条节点只有唯一的父节点（类节点）。虽然这一假定在一定程度上限制了朴素贝叶斯模型的适用范围，然而在实际应用中，不仅以指数级降低了贝叶斯网络构建的复杂性，而且在许多领域，在违背这种假定的条件下，朴素贝叶斯学习模型也表现出相当的健壮性和高效性。它已经成功地应用到分类、聚类及模型选择等数据挖掘任务中。

Naïve Bayes 分类器的参数由先验类概率值 $P(c_j)$ 和基于类的词条的条件概率 $P(w_t \mid c_j)$ 组成，完全由已标注的训练集文档确定。每个类 c_j 的先验类概率值 $P(c_j)$ 的计算公式如（4-4）：

$$P(c_j)=\frac{1+\sum_{d_i\in D}P(c_j \mid d_i)}{|C|+|D|} \qquad (4\text{-}4)$$

其中 $|C|$ 为类数目，$|D|$ 为训练集合中的文本数目。

基于类的词条的条件概率 $P(w_t \mid c_j)$ 由下面公式估计：

$$P(w_t \mid c_j) = \frac{1+TF(w_t,c_j)}{|V|+\sum_k TF(w_k,c_j)} = \frac{1+\sum_{d_i \in D} N(w_t,d_i)P(c_j \mid d_i)}{|V|+\sum_{k=1}^{|V|}\sum_{d_i \in D} N(w_k,d_i)P(c_j \mid d_i)} \quad (4\text{-}5)$$

$TF(w_t,c_j)$ 为特征词条 w_t 在 c_j 类中出现的频度，$N(w_t,d_i)$ 为特征词条 w_t 在文本 d_i 中出现的次数。$|V|$ 代表文本集合中全部不同特征词条的数目。对于训练文本集合的文本 d_i，定义当 d_i 属于类别 c_j 时，$P(c_j \mid d_i)=1$，否则 $P(c_j \mid d_i)=0$。

对于测试文本集中的文本，利用已经训练好的分类器，我们可以求出文本 d 属于类别 c_j 的后验概率 $P(c_j \mid d)$，用 $w_{d,k}$ 表示文本 d 中的第 k 个特征词，$P(c_j \mid d)$ 的计算公式如（4-6）。

$$P(c_j \mid d) \propto P(c_j)\prod_{w \in V} P(w \mid c_j) \propto P(c_j)\prod_{k=1}^{|d|} P(w_{w,k} \mid c_j) \quad (4\text{-}6)$$

Step1：训练阶段，在训练文本集 $D=\{d_1,d_2,...,d_n\}$ 和类别集合 $C=\{c_1,c_2,...,c_m\}$ 上计算每个类的先验概率 $P(c_j)$，计算特征词属于每个类的条件概率 $P(w_t \mid c_j)$。

Step2：测试文本达到，生成特征向量，计算文本 d 属于每个类 c_j 的后验概率 $P(c_j \mid d)$。

Step3：比较测试文本属于每个类别的概率，将其分到最大的那个类别。

有时只要 d 的后验概率 $P(c_j \mid d)$ 超过某个预定阈值，可为 d 指定多种类别。但若这种情况发生得太频繁，则说明预定义类别 $C=\{c_1,c_2,...,c_m\}$ 不当，应加以修改。也可以设置阈值，当文档 d 对所有类的后验概率都低于该阈值，则将其标注为“其它”类。

4.4 SVM 分类法

支持向量机（Support vector machine，SVM）是一种近年来发展很快的机器学习方法，它在多种分类问题中表现出了优异的推广性能，其基本思想是基于统计学习理论的结构风险最小化。如果给出两类线性可分样本，在给出线性分类面的时候，人们直观的趋向于将分类面取在离两类的样本点都距离较远的地方，因为感觉上这种做法比较保险。Vapnik 从数学理论上给出了这种做法的理论依据，并推导出了这种方法风险性能的衡量，以及一整套求解的步骤。针对两类分类问题，在高维空间寻找一个超平面作为两类的分割，以保证最小的分类错误率，而且 SVM 的一个重要的优点是处理线性不可分的情况。

在线性可分的情况下，假设存在训练样本 $(x_1,y_1),...(x_n,y_n), y \in \{+1,-1\}, n$ 为样本数，x_i 为训练样本的特征列向量，在线性可分的情况下就会有一个超平面使得这两类样本完全分开，设该超平面的形式为：

$$g(x) = \omega \cdot x + b = 0 \quad (4\text{-}7)$$

公式中的圆点“.”表示向量点积，分类如下：

$$g(x) \geq 0, \qquad \text{对应于} y_i = +1$$
$$g(x) < 0, \qquad \text{对应于} y_i = -1$$

如果训练样本可以无误差地被划分，以及每一类样本与超平面距离最近的向量与超平面之间的距离最大则称这个超平面为最优超平面。

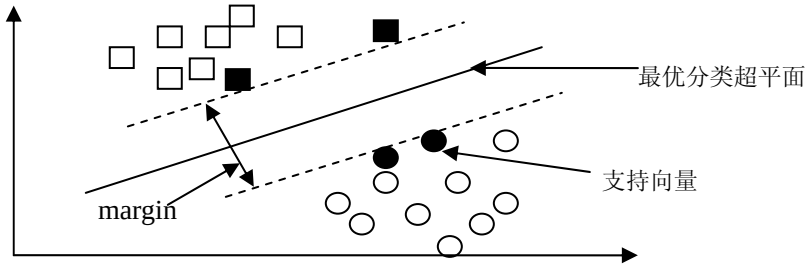

图 4-1　最优超平面

我们需要寻找有最小 VC 容量的超平面，所得正样本和负样本之间的距离被最大化。这样，当面临一个未知测试样本时，如果它在此超平面上方，则判断其类标为 $y_i = +1$，反之 $y_i = -1$。它可以通过最大化 margin 来求得。此处的 margin $\rho = \frac{2}{\|w\|^2}$ 被定义为某训练样本点与分类超平面的最小距离，我们所说的支持向量（SV）就是满足 $g(x) = \omega \cdot x + b = 1$ 的点 (x_i, y_i)。因此分类问题就被转化为一个二次规划问题。

将判别函数归一化，使得两类所有样本都满足 $|g(x)| \geq 1$，即

$$y_i[(\omega \cdot x_i) + b] - 1 \geq 0, i = 1,2,...,n \tag{4-8}$$

据此可以定义 Lagrange 函数：

$$L(\omega, b, \alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i \{y_i[(\omega \cdot x_i) + b] - 1\} \tag{4-9}$$

其中 $\alpha_i > 0$ 为 Lagrange 乘数，对 ω 和 b 求偏微分并令其为 0，原问题转换成为如下对偶问题：在约束条件

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \alpha_i \geq 0, i = 1,...,n \tag{4-10}$$

下对 α_i 求解下列函数的最大值：

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{4-11}$$

如果 α_i^* 为最优解，那么

$$\omega^* = \sum_{i=1}^{n} a_i^* y_i x_i \tag{4-12}$$

对于线性不可分的情况，可以引入松弛因子，在求最优解的限制条件中加入对松弛因子的惩罚函数。完整的支持向量机还包括通过核函数的非线性变换将输入空间变换到一个高维空间，然后在高维空间中求取线性分类面。常见的核函数包括多项式核函数、径向基函数，Sigmoid 函数等。值得指出的是，最终判别函数只包括与支持向量的内积的求和，所以识别时计算复杂性只取决于支持向量的个数。

SVM 很大的一个优点是它不受问题维数的限制，所以特别适宜在高维空间中工作，这对于文本分类来说，是非常适宜的。但是，当样本很大时，用标注的数学方法解决二次规划，在计算量上是不可行的。所以人们又提出了种种算法，使 SVM 能够实用化。由于具有较好的泛化性能，支持向量机被用于多个模式识别领域。在文本分类方面亦优于多种研究试验结果。在多个实验结

果中，SVM 均取得了较原有多种分类方法更高得分类精度。

4.5 KNN 和 SVM 的算法比较与分析

训练数据集：entrain500（国际植物遗传资源网站资料信息）；测试数据集：entest200.

实验目的：比较 KNN 和 SVM 分类器对英文 Web 文本的分类优劣。

特征词数量：2000

表 4-2 KNN 和 SVM 分类器英文网页分类结果

F1 度量 / 类别	KNN(K=35)		SVM	
	基于词频统计	基于文档统计	基于词频统计	基于文档统计
collaborative activity	64.79	69.45	77.78	71.64
country profile	81.01	81.48	92.50	88.89
human resource	85.39	71.57	97.50	96.38
crop	74.67	74.36	93.03	95.24
general	76.74	81.40	87.80	87.06
总体结果	77.00	80.00	90.00	88.50

实验分析：对于英文 Web 文本来说，不论采用基于词频统计赋权方法还是基于文档频率统计赋权方法，SVM 分类器分类效果都要优于 KNN 分类器。

中文 Web 文本分类实验：

训练数据集：chtrain2000（复旦大学中文语料）；测试数据集：chtest800。

实验目的：比较 KNN 和 SVM 分类器对中文 Web 文本的分类优劣。

特征词数量：2000。

表 4-3 SVM 中文网页分类结果

F1 度量 / 类别	KNN		SVM	
	基于词频统计	基于文档统计	基于词频统计	基于文档统计
交通	90.77	87.69	97.10	97.10
体育	97.35	94.80	98.35	98.35
军事	75.36	78.09	98.82	88.66
医药	85.95	88.89	96.30	97.02
政治	86.10	85.00	93.84	92.76
教育	93.24	94.52	96.55	95.77
环境	86.18	80.34	95.38	95.38
经济	78.46	78.74	97.17	95.81
艺术	91.36	93.33	98.78	96.97
计算机	87.18	86.21	99.25	98.49
总体结果	87.47	86.94	96.15	95.50

实验分析：对于中文 Web 文本来说，不论采用基于词频统计赋权方法还是基于文档频率统计赋权方法，SVM 分类器分类效果都要优于 KNN 分类器。

第五章　Web 文本分类系统的设计与实现

为了研究文本自动分类问题，建立一个标准的实验平台，这样研究者可以把主要精力放在理论和算法的研究上从而可以达到节省时间，提高效率的目的。

利用李荣陆提供的二元 SVM 分类器源代码[42]的基础上在 Windows2000 平台上用 VC++开发了 Web 文本分类系统，实现了两种分类方法 KNN 和 SVM 分类算法，对网页文件或 txt 文件的自动分类和结果评价。用户使用时可以开始新的训练，也可以打开以前训练好的分类模型文件 model.prj 直接进行分类。

5.1　系统结构简介

系统结构图如图 5-1 所示。

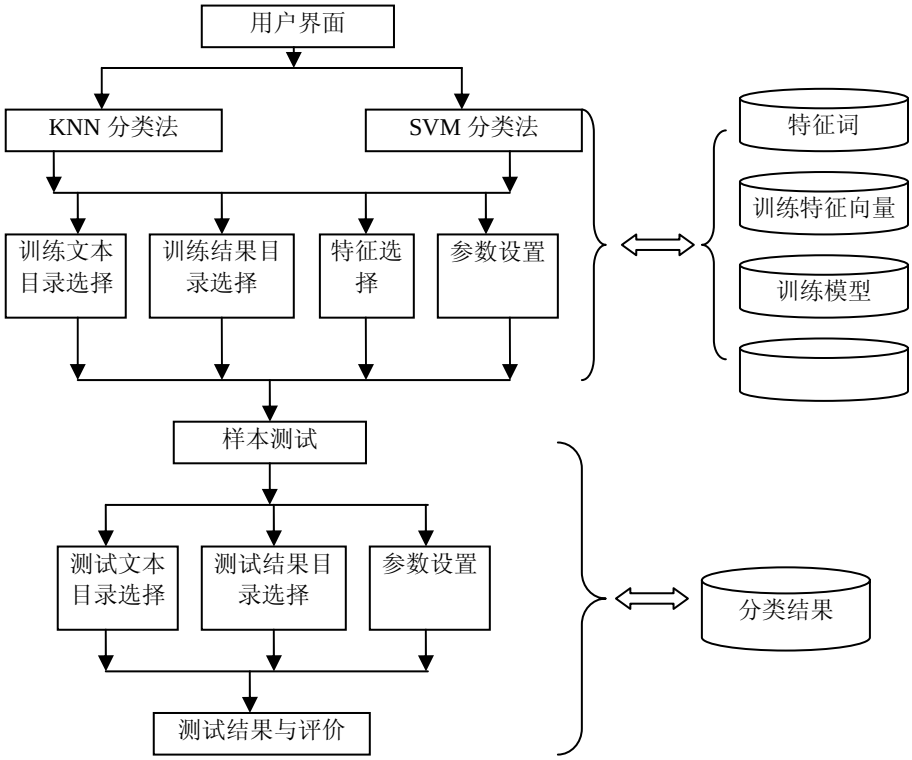

图 5-1　系统结构图

作为训练过程的结果，根据训练样本生成了三个比较有用的文件：train.txt，feature.txt，model.prj。train.txt 作为训练样本的特征向量文件，包含了训练样本所属的类别、特征词编号、特征词的权重等重要信息，方便用户查看、分析和比较；feature.txt 文件包含了从训练文本集中提取的特征词和特征词的权重，在特征提取过程中已经进行了停用词的处理，停用词保存在 stopwords.txt 中，用户可以自行扩充必要的停用词；model.prj 是训练模型文件，包含了训练的参数和模型，是进行分类的基础。

作为分类的结果，根据分类模型和测试样本产生的比较重要的文件有：classes.txt，result.txt。

classes.txt 文件给出了测试样本所包含的类别，方便用户查看所分出来的类别；result.txt 文件包含测试样本集中各个文件所属的类别。

分类结算后，用户还可以查看分类结果评测，系统给出分类的宏平均准确率、宏平均召回率、微平均准确率、微平均召回率分类评价指标。而且，还可以曲线、直方图等方式查看文档在各个类别中的准确率、查全率，以及文档在各个类别中的分布情况。系统主界面如图 5-2 所示。

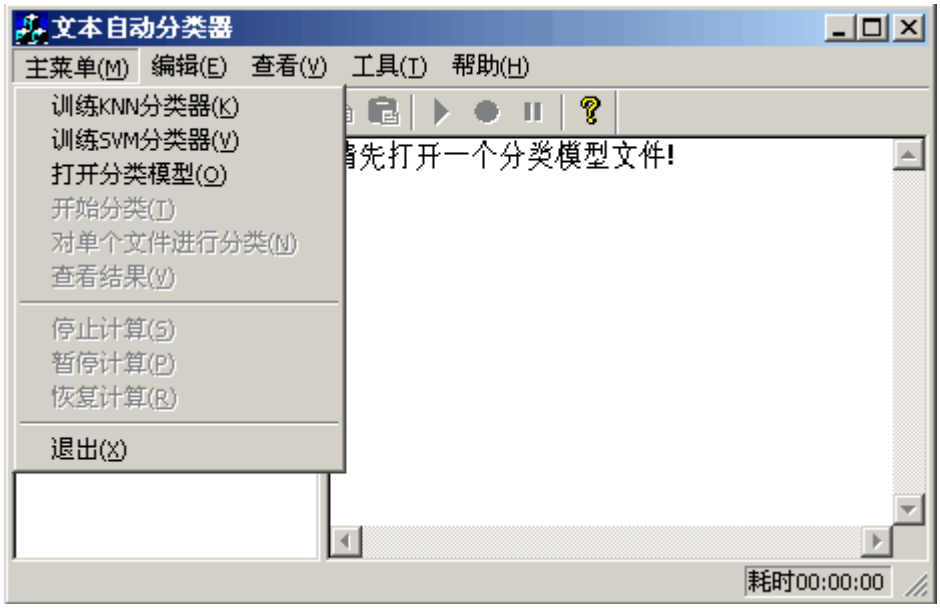

图 5-2　系统主界面

5.2　各功能模块

5.2.1　Web 文本处理模块

文本处理模块也就是网页特征提取模块，经过特征提取后生成原始的特征向量库。特征提取模块的功能如图 5-3 所示：

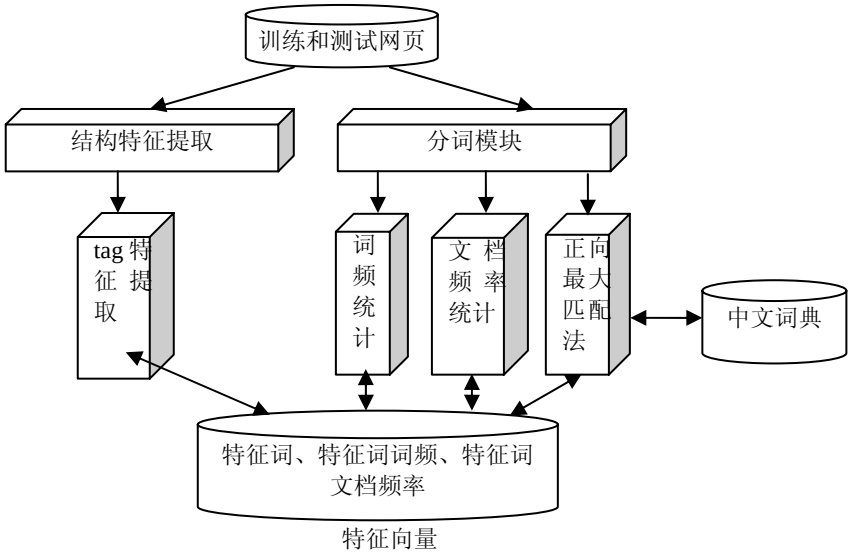

图 5-3　特征提取模块

该模块分为训练样本处理模块和测试样本处理模块，分别读取训练和测试网页集，进行特征提取和预处理，生成特征向量，保存到文本中，以便进行下面的特征选择和分类。

5.2.2 Web 文本特征选择模块

经过特征提取后的特征向量维数是很高的，其中有很多无用的特征，因此要进行特征选择处理，该模块实现了文本分类中的经典的特征选择方法，如文档频率，信息增益方法，互信息，CHI，期望交叉熵，文本证据权，Right half of IG。模块输入的是经过特征提取后的原始词频、文档频率特征向量，通过特征选择后，得到低维度的特征向量，以文件 train.txt 形式保存。

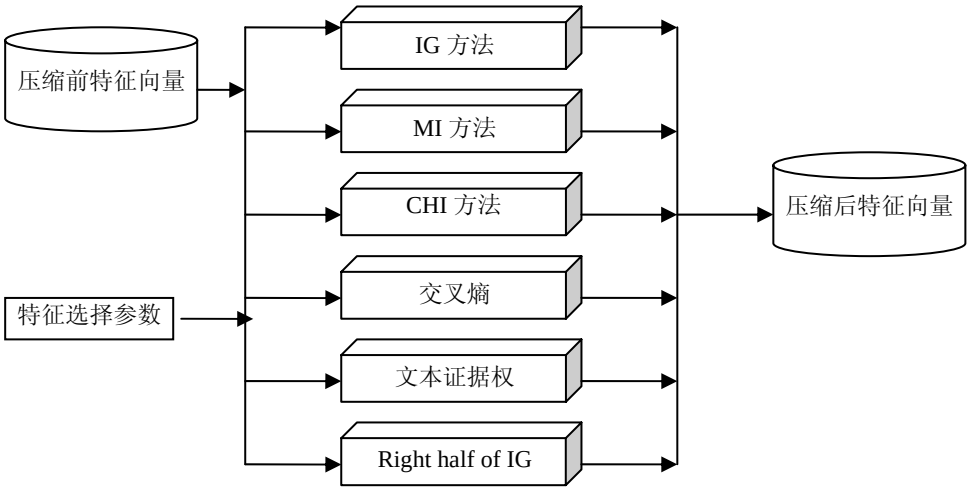

图 5-4 特征选择方法

文本分类系统的特征选择界面如图 5-5 所示。

图 5-5　特征选择界面

5.2.3　分类模块

该模块是网页分类系统的核心部分，完成网页的自动分类。用户选择一个分类方法和分类的阈值后，就可以进行自动分类了。并保存分类结果。

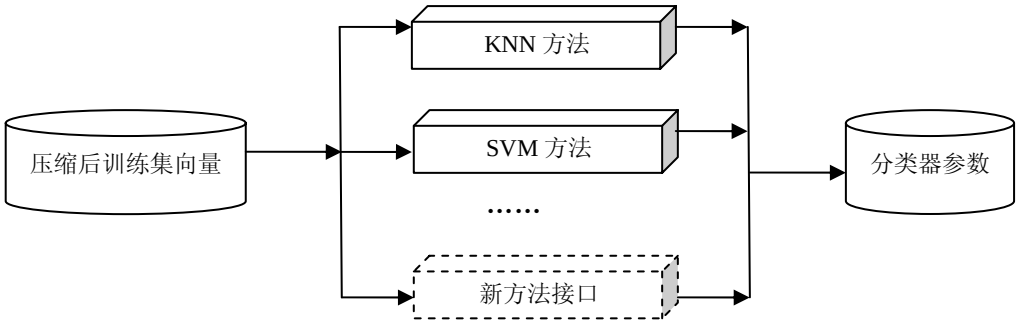

图 5-6　分类算法结构图

该模块的输入来自特征选择模块的输出，即压缩后的训练样本集的特征向量表示。该训练模块的结果，保存为对应各种算法的分类参数。在系统中分类模型保存在 model.prj 文件中。

5.2.4　测试模块

测试模块的主要功能是对训练好的分类器进行测试，将特征向量表示的测试样本集和已经训练好的分类模型作为分类算法的输入，输出分类结果，作为评估分类算法的分类性能的依据，以便比较和研究不同分类算法用于网页分类的性能优劣。

图 5-7　分类界面

5.2.5　分类结果与评价模块

该模块显示分类的结果和评价。显示采用的分类算法、采用的特征选择方法，特征赋权方法，以及分类的正确率，分类出错的文本数，特征提取后保留的特征数等。

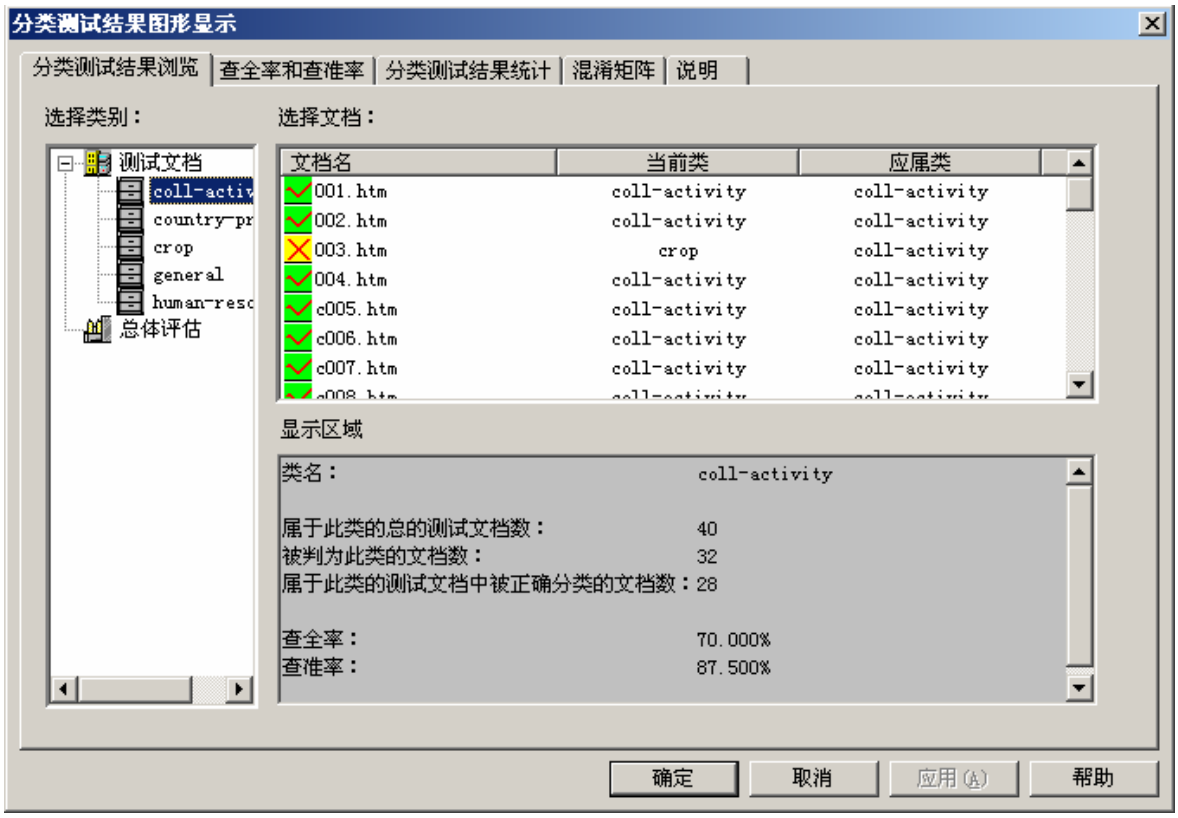

图 5-8　分类结果评价界面

第六章　　东亚植物遗传资源管理系统的设计与实现

为了把来自于国际植物遗传资源研究所（IPGRI）的已分好类的 Web 文本和东亚植物遗传资源协作网（EA-PGR）自有的信息资料在协作网的成员国内传播共享，促进东亚植物遗传资源保护和利用，构建了这个基于 ASP.NET 的动态信息管理平台。

基于 ASP.NET 的东亚植物遗传资源保护和利用管理系统具有网络化、易用性、开放性和分块管理的特点，使协作网内的成员能高效、稳定地发布，查询，浏览植物遗传资源信息、新闻事件、合作活动以及相关经验等。

6.1　系统设计

6.1.1　体系结构

系统主要以 ASP.NET 为开发平台，SQL Server 2000 为后台数据库，采用 Web 流行的 Browser/Server 模式，完成后台数据信息的管理和前台浏览、查询系统的构建。从结构和功能上，系统可以分为接口层、应用层和数据层三层体系结构，如图 6-1 所示。

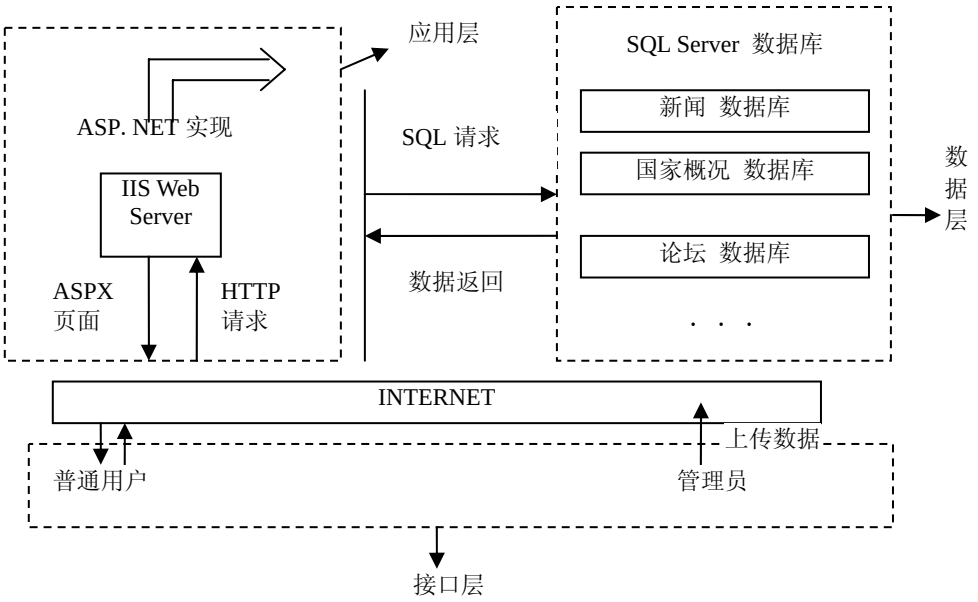

图 6-1　体系结构图

接口层位于客户端，相当于用户界面，即 Internet Explore 等 Web 浏览器；应用层是系统核心部分，担当主要的应用处理任务，包括处理接口层的 Http 请求及与数据库服务器的连接和交互；数据层位于底层，以 ADO.NET 为接口，Microsoft SQL Server 为架构，主要处理应用层对数据的请求。以 Windows2000+IIS5.0+.NET 框架作为平台，为了提高编程的效率和程序的可维护性，以 VS.NET 作为程序设计和编译环境，C#为程序设计语言。

客户端可以运行于各个版本的 Windows 操作系统，通过浏览器运行本管理系统，实现用户客户端的零安装。

6.1.2　后台管理模块设计

后台管理模块是面向管理员级别用户的，有权限的用户才能访问，是进行资源分配和数据管理的部分。

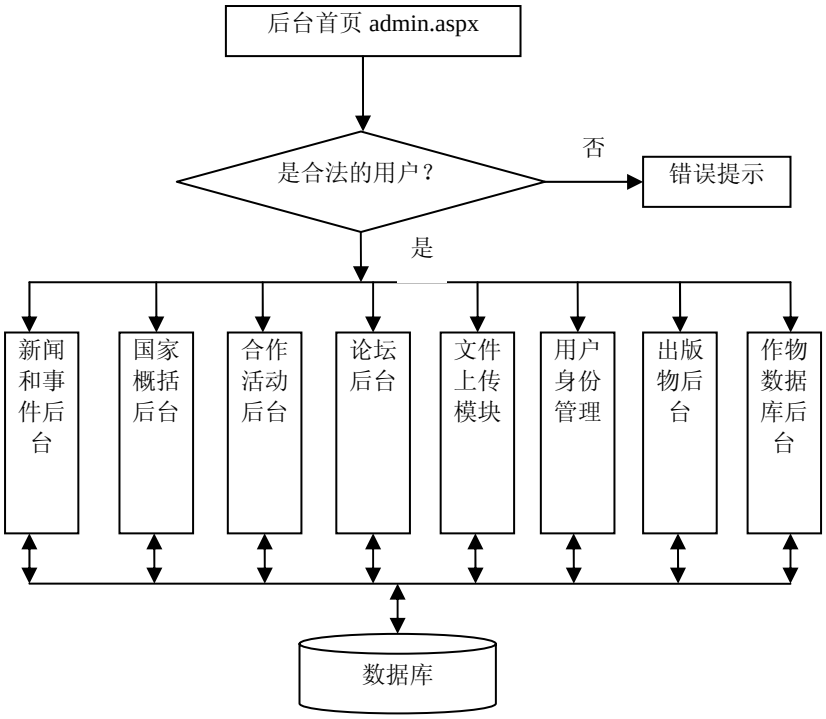

图 6-2　后台总体结构图

（1）身份认证模块

该模块属于窗体认证系统，进入到具体的某一个管理模块前都要通过用户身份认证，只有具有有效身份的管理者才能登录到各个管理模块。

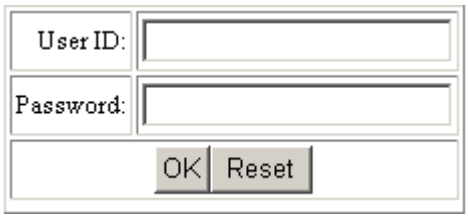

图 6-3　登录模块图

身份登录模块如图 6-3 所示，通过用户提供的 UserID 和 Password，在管理员数据表 Sysmanager 中进行搜索匹配，只有两者和数据表中的某条记录完全匹配并且 UserID 当前状态是有效才能进入各个管理模块，否则不能进入管理模块。UserID 保存在 Cookie 对象中，以便在各个管理模块中进行身份跟踪。

（2）新闻和事件管理模块

新闻和事件的文章是分两层管理的，即新闻文章组织在 Category（类目）下的，管理员通过维护该 Category，可以把所有的新闻条目和文章组织在不同的 Category 下。新闻条目和文章都

有批准状态即 approved，管理员提交的新闻条目默认批准状态 approved 都为 true，管理员可以任意修改批准状态，如果 approved 为 false，则该条新闻对前台用户是不可见的。任何一条新闻条目都有发布日期和过期日期，发布日期默认值是创建日期，即创建后立即发布，过期日期默认值是发布日期加上一百年，即有效期是 100 年。只有当前日期介于发布日期和过期日期之间的新闻条目才能为前台的用户所获得。每个新闻文章都有一个 User，系统能够跟踪提交该条新闻的人，了解新闻贡献者是否活跃，谁对不正确的新闻内容负责。管理员提交新闻时，系统自动通过读取 Cookie 对象获得 User 值和新闻文章作为整体保存到数据库，此过程管理员是不可见的。

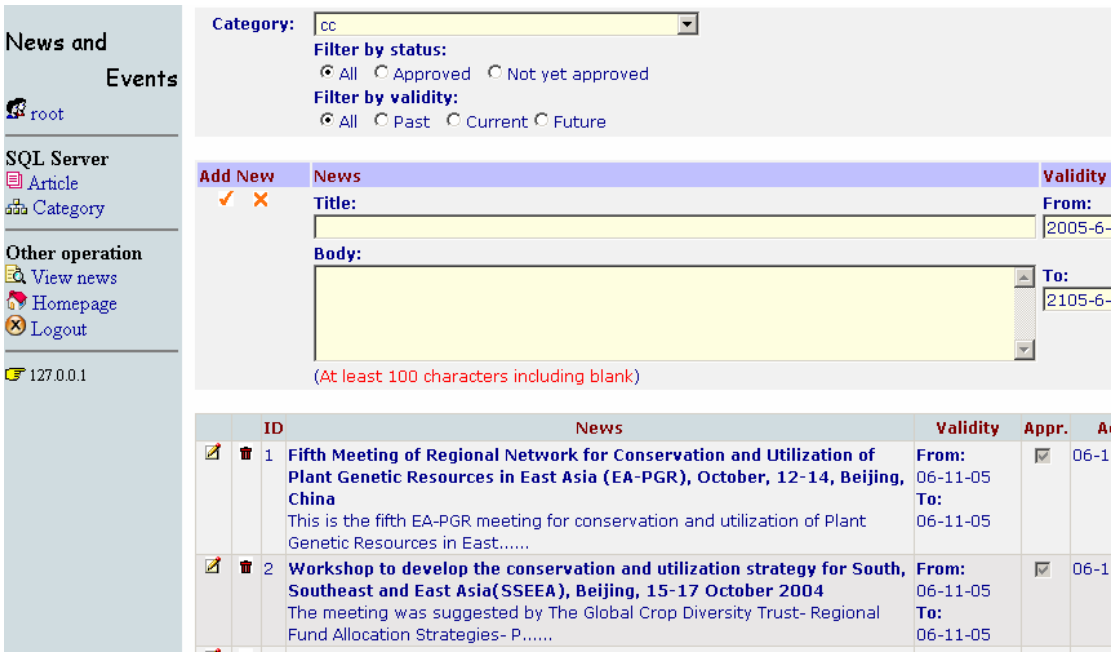

图 6-4　新闻和事件后台管理界面

（3）合作活动管理模块

合作活动模块和新闻模块的不同之处是，合作活动文章分三层管理的。首先是 Category（类目），其次是 Subcategory（子类目），最后是合作活动文章，这样把合作活动文章分类更细化，便于用户分类查询和浏览。

（4）国家概况管理模块

在本管理系统中，管理员权限分 super user 和 ordinary user，只有 super user 能增加、删除、修改国家，并且把一个国家的修改权限赋予系统内的任意一个管理员，管理员只可以管理得到授权国家的概况信息，不能管理未授权国家的概况。这样做的好处是某个国家的概况信息只能由本国的管理员来维护的。国家概况文章的管理采用四层目录树的结构，第一层是 country，第二层是 category，第三层是 subcategory，最后一层是国家概况信息。

（5）论坛管理模块

该模块的身份认证和其他管理模块的身份认证稍有不同。系统根据用户提供的 UserID 和 Password 判断是普通用户还是管理员，然后在 cookie 对象中做标记。如果是管理员，用户、主题、板块等论坛的管理都是有效的，如果是普通用户身份，则论坛管理功能是不可见的。管理员可以维护论坛的 Category（类目），在某个 Category 下可以维护 Forum（论坛板块），在 Forum

下可以具体的维护 Topic（主题帖）和 Reply（回复帖），也可以删除用户和修改用户的有效性等。

（6）出版物管理模块

在该模块中系统管理员可以上传东亚五国协作网内部的可以对外公布的出版物资料文件，文件的类型是可以任意的。上传的出版物相关信息如：文件名称，文件大小，文件类型等存储在数据库中，文件本身存储在该管理系统的某个子目录下，提高了文件的存储速度以及便于检索文件。

（7）作物数据库管理模块

该模块的作用是为用户提供可以对外公布的植物信息数据库，以便于普通用户浏览和查询，达到植物数据资源共享的目的。该模块暂时提供了 Adzuki（包括东亚五个国家的数据表）和 Saff 两个 Access 数据库，管理员还可以根据需要从外部导入新的 Access 数据库，供用户查询。

（8）文件管理模块

管理系统好坏的标准之一就是可维护性。为了便于系统的更新和维护以及在前面提到的任意一个模块的文章中插入图片、表格和各种类型文件，开发了本模块。服务器上运行的系统文件是经过在开发机器上编译的复本。若某个模块需要修改，则在开发机器上完成修改和编译后，直接上传到服务器上，覆盖原来的文件即可，不需要在 Web 服务器上操作。若要在某个管理模块的文章里引用图片文件，则可以把需要的图片上传到系统文件的某个子目录内，然后在文章中用 html 标记引用即可。例如，要在新闻文章内显示图片文件“1.gif”，则可以先在 news 目录下创建一个“img”目录，然后把要引用的图片文件上传至“img/”下，之后在新闻文章里加入 html 标记：<img src=“img/1.gif”>即可。为了能在服务器控件里使用 html 标记，需要在.aspx 文件开头处添加指令<%@ Page ValidateRequest=false %>。

图 6-5　文件管理界面

（9）后台用户管理模块

该模块是管理员用来维护自己的资料信息，超级管理员则可以增加、删除其他普通管理员和改变管理员的角色，如过期用户的删除，授权新的用户，把普通管理员升级为超级管理员等。

6.1.3　前台浏览查询模块设计

前台是面向普通用户的数据浏览、查询等操作的平台，要体现操作界面友好和操作方便的特点，图 6-6 是前台的总体结构。

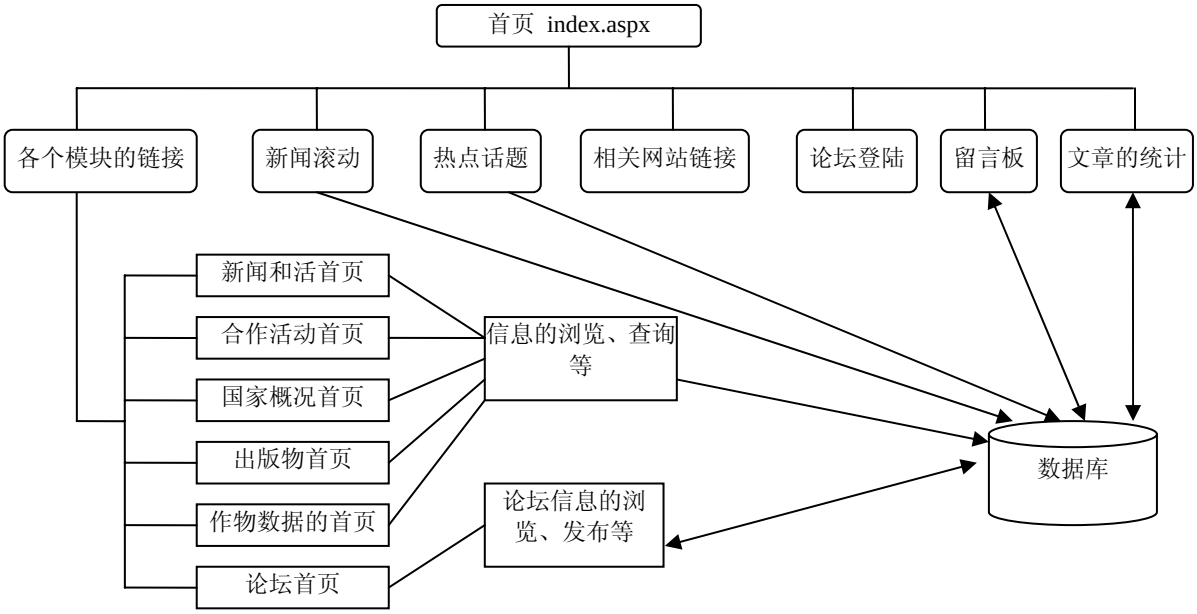

图 6-6　前台总体结构

前台系统主界面如图 6-7 所示：

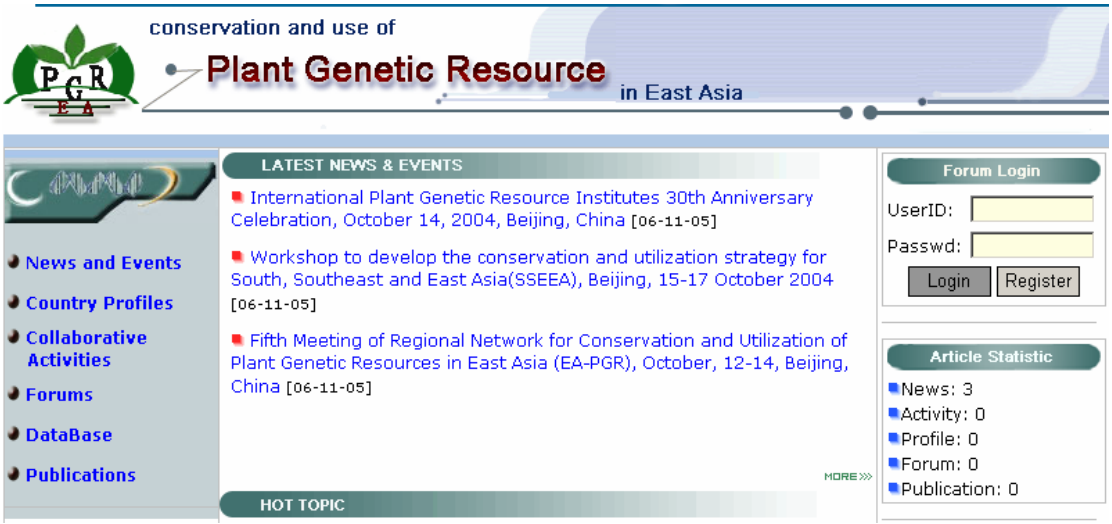

图 6-7 前台主界面

（1）新闻和事件浏览模块

为了便于用户浏览信息，该部分采用框架结构，用户点击左侧框架内的新闻类目，右侧框架内就列出了相应的新闻标题和新闻摘要。感兴趣某一条新闻可以单击标题进入。用户还可以按照关键字简单查询新闻，也可以按照类目分类查询新闻等达到快速浏览的目的。其他浏览查询模块与此模块结构和实现方法类似，不再赘述。

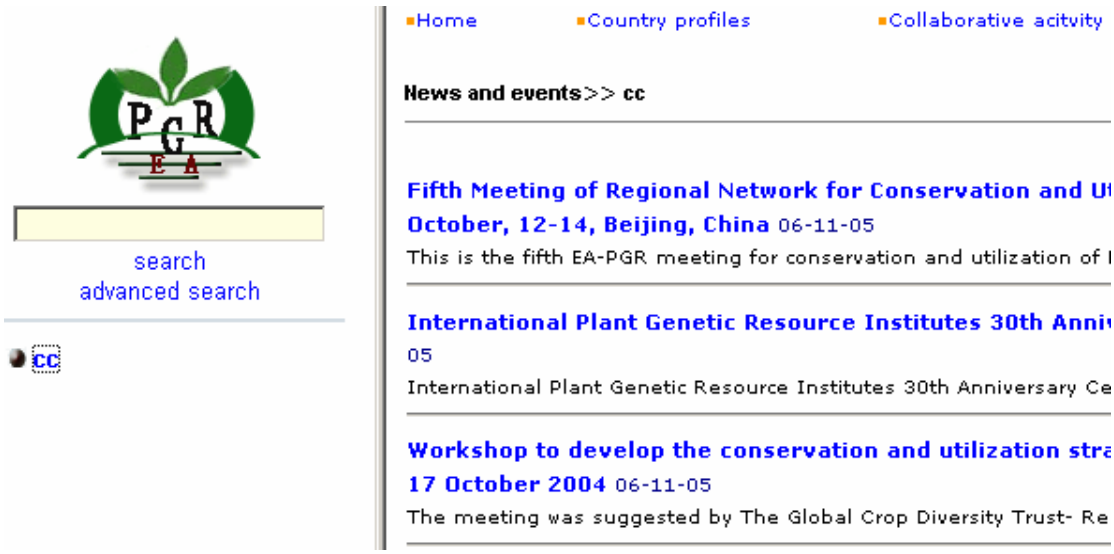

图 6-8　新闻和事件浏览界面

（2）作物数据库模块

在该模块中，用户可以在数据库 Adzuki 和 Saff 中进行任何一个或两个字段的查找，查找模式有精确查找、匹配查找和前方一致查找。

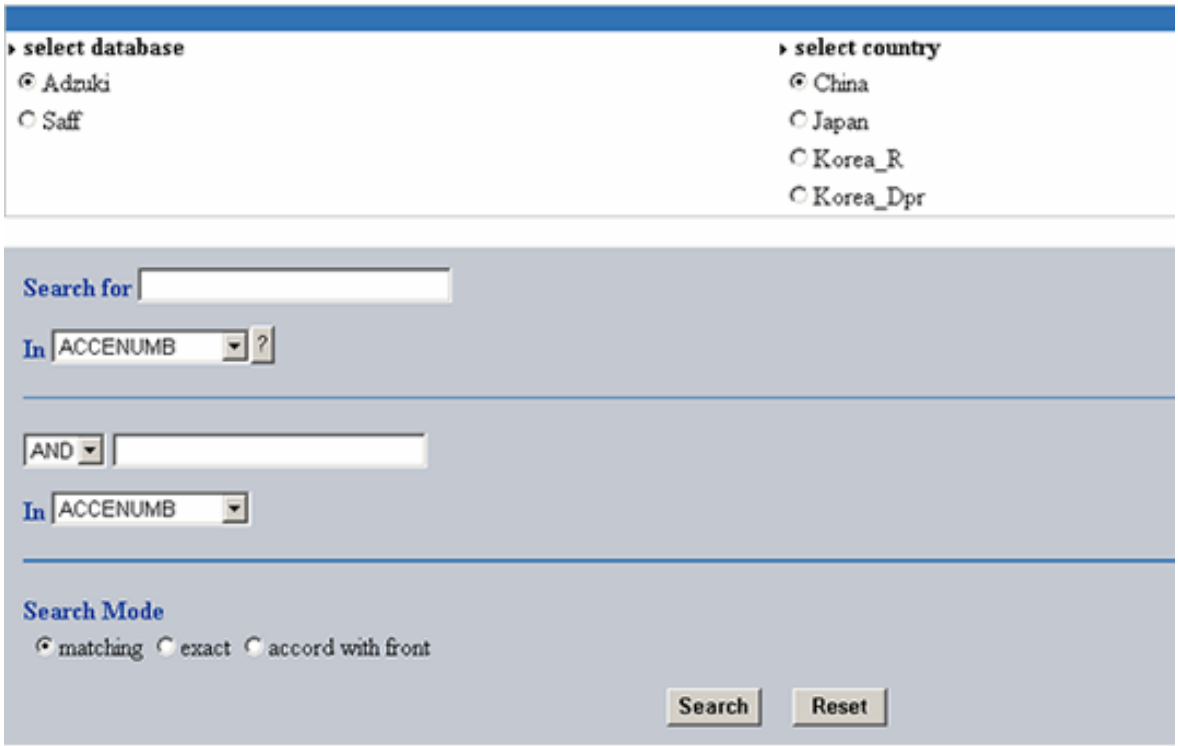

图 6-9　作物数据查询界面

6.1.4　数据库设计

要对一个庞大的数据库进行管理操作，实现查询的快速而无误，建库是首要工作。怎么样建库，采用什么样的库结构，关系到管理和查询的准确性和效率。为了便于建立数据表之间的关联

关系以及数据的备份和导入导出等操作，系统采用一库多表的形式，即只建立一个数据库，根据不同的功能模块分别设计数据表，然后组织成关系图。系统主要包括以下几个数据表和关系图：

临时数据表，记录查询过程中的中间数据，以供用户执行多步查询时进一步使用。该表生存期很短，随着某一次查询操作自动消失。

新闻数据关系图，记录新闻板块的相关信息。该图包含三个表：类目表 News_Category，文章表 News_Article，系统管理员表 Systemmanager。字段 NewsID（类型 bigint）是 News_Article 主键，自动增长，字段 body（text）存放新闻的内容；CategoryID（bigint）是 News_Category 的主键，同时是 News_Article 的外键，两者之间关系是对 insert 和 update 强制关系，即 News_Article 表中不能存在类目不属于 News_Category 的新闻条目。同样，systemmanager 中 UserID（varchar）是主键，同时是 News_Article 的外键，即新闻文章的插入者必须是系统管理员。

论坛数据关系图和新闻关系图有些不同。论坛数据表的关系图如图 6-10 所示。

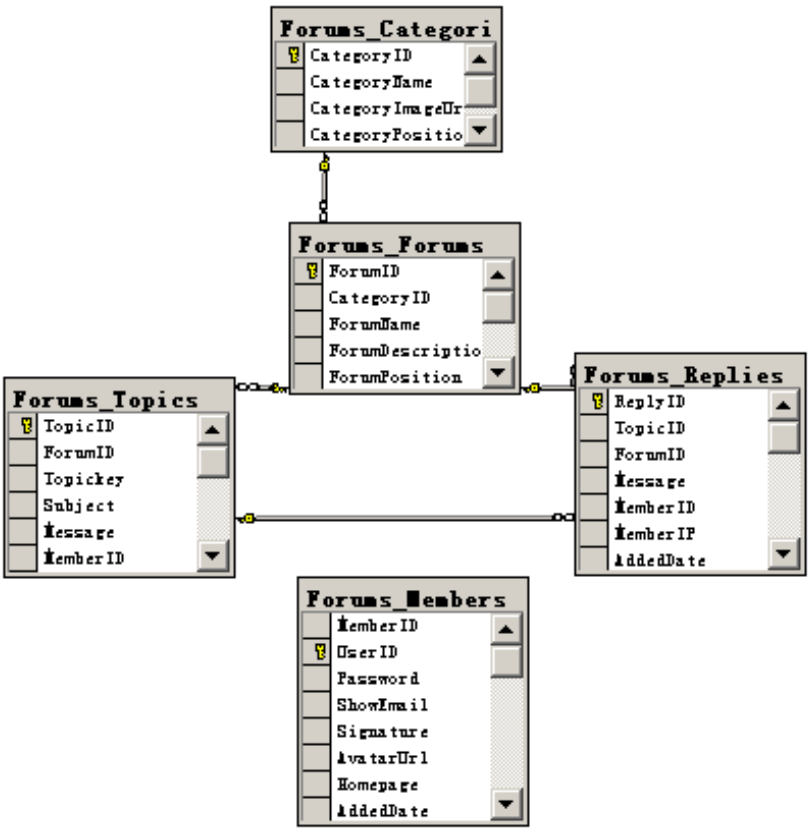

图 6-10　论坛数据关系图

Forum_Categories 是论坛的类目表，主键是 CategoryID（int）；Forums_Forums 是论坛的论坛板块表，主键是 ForumID（int），自动增长，外键是 CategoryID，即论坛板块只能是某个类目下的论坛板块；Forums_Topics 是论坛的讨论主题表，主键是 TopicID（bigint），自动增长，外键是 ForumID，即某个讨论主题只能是某个论坛板块下的主题；Forums_Replies 是论坛的回复表，主键是 ReplyID（bigint），自动增长，外键是 TopicID，即回复只能就已经存在的某个主题进行回复。Forums_Members 是论坛成员表，UserID（varChar）是主键。

6.2 工作原理和主要技术

6.2.1 工作原理

系统的工作原理与.NET框架工作原理一致，即：程序第一次在服务器端运行并编译成叫做MSIL的微软中间语言，并存储在可移植的可执行文件中，当用户查询时，MSIL被即时地转化成本机语言。这样，程序经编译成MSIL后，不仅提高了运行速度，而且巧妙地避免了传统程序面临的移植问题。

6.2.2 系统的实现过程

ASP.NET通过ADO.NET存取数据，ADO.NET是以离线的数据为基础的，可以在本地的机器上对数据集进行数据的添加、删除或修改，然后更新回真正的数据库。系统的整个实现可以概括为以下几步：

（1）在代码文件.CS文件开头部分用using方法导入System.Data和System.Data.SqlClient这两个命名空间，如下：

```
using System.Data;
using System.Data.SqlClient;
```

（2）建立与数据库的连接。为了程序维护方便，在Global.asax.cs文件中新增一个Application对象，保存连接字符串，如：

```
Application["sqlConnectionString"]="server=(local);database=ea-pgr;user id=sa;
password=ipgri";
```

然后在要与数据库操作的各页使用

```
SqlConnection Conn = new
SqlConnection((string)Application["sqlConnectionString"]);
Conn.Open();
```

即可建立连接对象。

（3）在与数据库连接的基础上，执行SQL功能语句，为了提高数据库的执行速度，该系统将常用的SQL语句如查询（Select）、插入（Insert）、更新（Update）、删除（Delete）等做成存储过程，直接保存在SQL数据库中。如下面根据关键字NewsID, 从新闻文章表（News_Article）查找符合要求的新闻，建立存储过程sp_News_GetNewsdetails，如下：

```
Create Procedure sp_News_GetNewsdetails
@NewsID int
AS
select * from News_Article where NewsID=@NewsID
GO
```

建立存储过程后，就可以在页面中利用Command对象直接调用，如：

```
SqlCommand Cmd = new SqlCommand（sp_News_GetNewsdetails, Conn)。
```

6.2.3 异常情况的捕获与控制

在系统运行的过程中，不可避免的会出现用户的误操作。为了提高系统的交互性和程序的可靠性，系统对各类异常操作进行了相应处理。异常情况的捕获与控制是利用 ASP.NET 的“Try…Catch…Finally…End Try”语句来实现的。例如，Try 块内可以执行更新操作，Catch 块内可以捕获错误信息，并给出提示信息，Finally 块可以做关闭数据库连接的操作等。

第七章 结论与展望

7.1 结论

Internet 的飞速发展为信息数据的管理、共享、传播以及数据分析提供了全新的途径。采用 Browser/Server 模式,充分利用 ASP.NET 的优点，实现了对数据进行集中管理维护。植物遗传资源管理系统将人工智能技术、网络技术及数据库技术有机集成，通过 Internet 为东亚地区植物遗传资源组织和植物遗传专家提供一个内容丰富、反应迅速、相互交流的信息化平台系统。

随着计算机网络的发展，大量电子文本的涌现，使得如何从海量信息中搜寻、过滤、管理这些电子文档，并对其有效地利用，成为一个值得研究的方向。对文本进行准确、高效的分类为文本管理、信息检索提供了有效的解决方法，是许多管理任务的重要组成部分。所谓分类，是对所给出的文本，给出预定义的一个或多个类标号。文本、电子邮件的内容实时辨识和过滤，结构化的搜索和浏览，提供个性化的服务等方面，均在一定程度上依赖于准确的文本分类技术。

网页分类技术涉及到模式识别、自然语言处理、统计学、机器学习等方面的内容。本论文对网页分类技术进行了研究。本文完成的研究工作主要有以下几个方面。

1．论文对 Web 文本分类进行了初步探讨，在理解前人工作的基础上，对 Web 文档的特征表示、特征提取和选择进行了归纳总结，并用来自国际植物遗传资源研究所的 Web 文本，在 KNN 和 SVM 分类器上对信息增益（IG）、互信息（MI）、文本证据权、CHI、期望交叉熵、右半信息增益 6 种特征选择方法进行了实验比较，分析了各种方法的优劣，找出了适合解决实际问题的特征选择方法。

2.论文对基于词频统计的特征赋权方法和基于文档统计的特征赋权方法分别在 KNN 和 SVM 分类器上进行了实验，基于词频统计的特征赋权方法比基于文档统计的特征赋权方法要好，因为文档频率法只简单的统计词是否在文档中出现，所以精度没有词频统计法高。

3．为了确定所取特征维数对分类的影响，分别在 KNN 和 SVM 分类器上用来自国际植物遗传资源研究所的 Web 文本进行了实验，得出提取的特征词维数多比特征词维数少的分类精度高，但是以牺牲训练时间为代价的。

4．对 Web 文档的分类算法：类中心向量、KNN、朴素贝叶斯、SVM 等几种分类器作了研究分析。并用中英文的不同规模的训练文本集和测试文本集对 KNN 和 SVM 两种分类器进行了实验比较和分析，认为在其他条件相同的情况下，SVM 是比 KNN 更好的分类器，更适合解决论文提出的具体问题。

5．为了确定 KNN 分类器中 K 的取值，在来自国际植物遗传资源研究所的 Web 文本进行了比较分析，一般认为当 K 取 40 左右时，KNN 的分类准确率达到最高。

6.用 VC++设计和实现了中英文本的自动分类系统,系统实现了 KNN 和 SVM 两种分类算法。该分类器实现了中英文文本自动分类，支持多种特征选择方法，兼类分类和分类结果评价显示。

7．采用最新的.NET 框架设计和实现了东亚植物遗传资源协作网管理系统平台，实现了信息和资源的及时发布与管理。

7.2 展望

由于时间有限，而且涉及的知识较多，还有许多问题有待进一步地深入、广泛的研究。中英文 Web 文本分类系统还存在着许多不足，其中主要包括以下几个方面。

1．对比已标注的分类训练样本资源匮乏，目前存在着大量的未标注样本，结合标注的和未标注样本进行分类学习，这样可以提高文本的分类范围。

2．Web 文档分类具有更加广泛的应用范围，可以尝试在其他领域应用该技术。另外，系统有待于进一步完善，进一步提高分类精度，逐渐把 Web 文档分类走向实用化。

3．文本自动分类系统只实现了 KNN 和 SVM 两种分类算法，还有许多其他有价值的分类算法还没有扩充到本系统中来，因此这些算法与 KNN 和 SVM 算法的优劣也无从比较，因此这是以后工作的一个内容。另外，改进已有的分类算法和寻找新的适合解决具体问题的分类算法也是今后的一个主要研究内容。

4．论文对 Web 文本分类采用的是有导师的学习方法，要对新文本分类，就要用标注好类别的文本对分类器进行训练，对一定量的文本进行标注工作量也是巨大的。而文本聚类技术是无导师的学习方法，具有相同或相似特征的文本自动聚集在同一类中，所以文本聚类更省时省力，因此将文本聚类技术应用到文本分类中也是今后工作的一个方向。

参考文献

[1] 肖明，沈英．自动分类研究进展．现代图书情报技术，2000，5，25~28

[2] 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．计算机研究与发展，2000，37（9）：1032~1038

[3] 范焱，郑诚，王清毅，等．用 Naïve Bayes 方法协调分类 Web 网页．软件学报，12（9）：1386~1392

[4] 解冲锋，李星．补偿型的 Sleeping expert 文本分类算法．清华大学学报（自然科学版），41（7）：39~42

[5] Zhang Yizhong，Zhao Mingsheng，Wu Youshou．The automatic classification of web pages based on neural networks．Neural information processing，ICONIP2001 Proceedings，November 14-18，2001 Shanghai，China，Vol.2，570~575

[6] 王梦云，曹素青．基于字频向量的中文文本自动分类系统．情报学报，2000，19（6）：644~649

[7] Dunija Mladenic，Marko Grobelnik．Feature selection on hierarchy of web documents．Decision Support Systems，2003，35：45~87

[8] Zhi-hua，Zhou KaiJiang，Ming Li．Multi-Instance Learning Based Web Mining．Applied Intelligence，2005，22：135~147

[9] K,Shima，M.Todoriki，A.Suzuki．SVM-based feature selection of latent semantic features．Pttern Recognition Letters，2004，25：1051~1057

[10] Oh-Woog Kwon，Jong-Hyeok Lee．Text categorization based on k-nearest neighbor approach for web site classification．Information Processing and Management，2003，39：25~44

[11] 李粤，李星，刘辉，等．一种改进的文本网页分类特征选择方法．计算机应用，2004，24(7)：119~121

[12] 庞剑锋，卜东波，白硕．基于向量空间模型的文本自动分类系统的研究与实现．计算机应用研究，2001，18(9)：23~26

[13] 卜东波．聚类/分类理论研究及其在大规模文本挖掘中的应用：[博士学位论文]．北京：中科院计算技术研究所，2000

[14] 刘敏．基于支持向量机的中文文本自动分类系统的设计和实现：[博士学位论文]．杭州：浙江大学，2001

[15] 黄萱菁．大规模中文文本的检索、分类与摘要研究：[博士学位论文]．上海：复旦大学，1998

[16] 宋枫溪．自动文本分类若干基本问题研究：[博士学位论文]．南京：南京理工大学，2004

[17] 崔伟东．基于支持向量机的自动文本分类算法研究与系统实现：[硕士学位论文]．北京：清华大学，2000

[18] 唐焕玲．文本自动分类方法的研究：[硕士学位论文]．北京：清华大学，2004

[19] 李凡，鲁明羽，陆玉昌．关于文本特征提取新方法的研究．清华大学学宝（自然科学版），2001，41（7）：98~101

[20] 王继成，潘金贵，张福炎．Web 文本挖掘技术研究，计算机研究与发展，2000，37（5）：

531~520

[21] Lewis, D. D.，Yang, Y.，Rose, T. ，*et al*． A New Benchmark Collection for Text Categorization Research . Journal of Machine Learning Research, 2004 ， 5:361-397 http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf

[22] Ciya Liao，Shamim Alpha，Paul Dixon．Feature Preparation in Text Categorization．Oracle Corporation．

[23] Yiming Yang．An evaluation of statistical approach to text categorization．In Technical Report CMU-CS-97-127，Computer Science Department，Carnegie Mellon University，1997

[24] 边肇祺，张学工．模式识别（第二版）．北京：清华大学出版社，2000，284~303

[25] 朱克斌，唐菁，杨炳惯．Web文本挖掘系统及聚类分析算法．人工智能及识别技术，30（13）：138~139

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

[27] 魏新，冯兴杰，刘山．基于支持向量机的多元文本分类研究．海军工程大学学报，16（5）：30~32

[28] 施洁斌．基于支持向量机的文本自动分类试验研究．现代图书情报技术，2004，7：27~29

[29] 刘良斌，王小平．基于支持向量机和输出编码的文本分类器研究．计算机应用，2004，24（8）：32~34

[30] G.Salton，A.Wong，C.S.Yang．A vector space model for automatic indexing．Communication of the ACM，1975，18（11）：613~620

[31] Yang Yiming，Pederson J O．A Comparative Study on Feature Selection in Text Categorization．Proceedings of the 14th International Conference on Machine learning，1997，412~420

[32] 李淑文．试论文本自动分类．现代计算机，2004，7：38~41

[33] 万志峰．搜索引擎的技术现状及其发展趋势．信息检索技术，2004，114：35~36

[34] 宋枫溪，高林．文本分类器性能评估指标．计算机工程，2004，30（13）：107~109

[35] 刘丽珍，宋瀚涛．文本分类中的特征选择．计算机工程，2004，30（4）：14~15

[36] 余芳．一个基于朴素贝叶斯方法的web文本分类系统：WebCAT．计算机工程与应用，2004，13：195~197

[37] 李亮，刘万春，徐泉清，等．一种基于支持向量机的专业中文网页分类器．计算机应用，2004，24（4）：58~61

[38] 李红莲，王春花，袁保宗，等．针对大规模训练集的支持向量机的学习策略．计算机学报，2004，24（5）：715~719

[39] 冯是聪，张志刚，李晓明．一种中文网页自动分类方法的实现及应用．计算机工程，2004，30（5）：19~20

[40] D.D.Lewis and M.Ringguette．Comparison of two learning algorithms for text categorization．In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval，1994

[41] Yiming Yang，Sean S.，Rayid G．A Study of Approaches to Hypertext Categorization

[42] http://www.cs.cornell.edu/people/tj/svm_light/

[43] 马荣邦．Web 技术发展的三个阶段综述．煤炭技术，2003，22（9）：128~130

[44] 朱　斌，宋先忠．动态网页开发技术探讨．计算机应用，2001，21（9）：55~56

[45] 宋晓勇．动态网页开发 Web 应用程序．　上海电机技术高等专科学校学报，2003，6（2）：19~22

[46] 桂思强．ASP.net 与数据库程序设计．北京：中国铁道出版社，2002

[47] 宫学庆，吴晓红，张龙，等．Web 数据管理研究现状与发展方向．计算机科学，2002，29（7）：19~24

[48] 魏应彬，黄健青，周星．PHP 技术及其应用．计算机与现代化，2000，5：86~89

[49] 李志华，孙荣胜．基于 JSP 技术的 Web 应用设计．电脑开发与应用，2002，15（3）：9~10

[50] 柳巧玲．JSP 运行环境及其应用．计算机工程，2002，28（8）：287~289

[51] 赵丽娜．如何利用 ADO.NET 技术访问数据库．计算机与现代化，2003，2：53~55

[52] 王剑钊，陈萍．数据库技术在网站设计中的应用．热力发电，2003，8，46~49

[53] 朱世顺. 基于 ontology 的土壤知识体系智能检索系统的设计与 Web 实现:[硕士学位论文]. 北京：中国农业大学，2003

[54] 李胜利，任军，任培民．建立企业信息门户的技术与实现．山东科技大学学报，2003，22（2）：99~101

[55] 王新伟．基于 Web 的远程教学系统的研究：[硕士学位论文]．北京：中国农业大学，2001

[56] Marco Bellinaso，Kevin Hoffman 著．ASP.NET Web 站点高级编程，康博译．北京：清华大学出版社，2002

[57] 鄂志国，王磊．基于 ASP.NET 的水稻品种数据库管理信息系统．计算机应用，2004，24(1)：140~142

[58] Stevenson, Ian．Adopting .NET platform environments．IEE Computing and Control Engineering，2003，14(5)：28~31

[59] Jason Price 著．C#数据库编程：从入门到精通，邱仲潘译．北京：电子工业出版社，2003

[60] 章立民．用实例学 ASP.NET．北京：电子工业出版社，2004

[61] 王国荣．ASP.NET 网页制作教程：从基本语法学起．武汉：华中科技大学出版社，2002

致谢

本论文是在导师王莲芝副教授的悉心指导下完成的。饮水思源，首先，给我的导师王莲芝送上最衷心的感谢。从论文选题开始到论文定稿整个过程中王老师倾注了大量的心血，付出了很多智慧和劳动。没有她的帮助我的课题和论文无法顺利完成。王老师严谨的治学态度、渊博的学识、孜孜不倦的敬业精神深深激励着我，那将是我一生的财富。她为人热情积极主动，在生活上给了我父母般的关爱，尤其在我生病住院期间，王老师更是无微不至的关心并在经济上帮助了我，使我深深地感动。

其次，感谢东亚植物遗传资源研究所东亚办事处的张宗文教授和白可喻博士，在这两位老师的帮助下使我有幸参与了国际合作项目的研究与开发。在张老师和白老师的严格要求和悉心指导下，使我的研究课题能顺利完成，并为以后的学习和工作打下了良好的基础。

在我完成课题和论文的过程中，当遇到一些问题时，向周围的老师和同学请教，他们都给予了很多好的见解和帮助，在此向所有帮助过我的老师和同学表示感谢！

最后，还要感谢我的父母。他们虽然在我的课题和论文中不能有什么直接的帮助，但是每当我遇到困难的时候，都是他们给我了克服困难的勇气和决心，所以使我完成了学业。

作者简介

姓　　名：张海龙

性　　别：男

出生年月：1976.12

籍　　贯：辽宁盘锦

最后学历：工学硕士

毕业院校：中国农业大学

在中国农业大学就读硕士研究生期间主要的研究工作经历和发表的论文如下：

[1] 国际合作项目：独立完成了东亚植物遗传资源协作网信息管理平台的设计和开发（项目基金编号：AP003/104），网址：http://211.155.251.240/webeapgr/index.aspx

[2] 科技部专项：独立完成了“农作物重大流行性病害病菌小种监测与数据库”中“病菌小种动态数据库应用系统的设计与开发”

[3] 教育部项目：参与了现代教育技术系统建设及运行机制的研究与实践和计算机文化基础课程网络教学课件的制作

[4] 独立进行了网页文本的自动分类的技术探索和方法研究，设计开发了基于 KNN 和 SVM 的中英文文本自动分类演示系统。

[5] 张海龙，王莲芝．基于 ASP.NET 的 EA-PGR 管理系统设计与实现．计算机工程与设计（已录用，2006 年 6 发表）