# 摘　要

随着 Internet 的迅速发展，网络已经成为各行各业获取信息的主要来源。如何从浩瀚的 Internet 海洋中，快速的获取有价值的信息呢？搜索引擎的出现很好的解决了这个问题。目前网络中以通用搜索引擎服务居多，它能够很好的满足人们对于普通信息的需求，然而，对于一些专业性比较强的领域，通用搜索引擎具有查不准、查不全、垃圾信息多的缺点。因此，许多行业都需要一种能够结合本专业特点的、专门应用于本行业的搜索引擎系统，也就是专业搜索引擎。

本文主要阐述了一个应用于水产渔业领域的中文专业搜索引擎的设计与实现过程。研究探索了适用于专业搜索引擎的网络爬行技术、中文分词技术、信息检索技术、网页的排序及消重等技术。该引擎的开发，采用了面向对象的设计方法，开发语言选用 Microsoft 的 Visual C++ 6.0，数据库管理系统采用 SQL Server 2000。

本文所述的搜索引擎的实现，能够满足与水产专业相关的学校、企业及其他单位通过网络，快速、准确地获得专业信息，具有广泛的应用价值。

**关键词：渔业信息搜索引擎；网络爬行；中文分词**

# Design and Realization of Fishery Information Search Engine

## Abstract

With the fast development of Internet, Internet has already been main source that every trade obtains the information. How to obtain the worthy information quickly from in the extensive ocean of Internet? The emergence of searching engine is good resolved this problem. Currently network in serve with the in general searching engine mostly,it can be good to satisfy the people'sneed for common information,however,the general searching engine has a lot of weakness to compare with some profession realm, for example checking not accurate,checking not whole,the garbage information Therefore,many professions all need a kind of can combine this professional characteristics of ,be apply in the profession industry exclusively of searching engine,is also the profession searching engine.

This thesis mainly elaborated an apply in the marine pruducts fishery realm of Chinese profession searching engine the design and the realization process.The network that studies to investigate to be applicable to the profession searching engine crawls technique, Chinese word segmentation technique, the information index technique, web page and eliminates heavy etc. The system is achieved by utilizing VC++6.0 of Microsoft with the concept of object-oriented technology, the SQL Server 2000.of the database management system

The realization of the searching engine that this thesis dissertated, can satisfy to pass the network with professional related school, business enterprise and other unitses of marine products, fast, acquire the professional information accurately, have the extensively applied value.

**Key Words: Fishery Information Search Engine; Network Crawl; Chinese Word Segmentation**

# 独创性说明

作者郑重声明：本硕士学位论文是我个人在导师指导下进行的研究工作及取得研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得大连理工大学或者其他单位的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

作者签名：郑重 日期：2007.5.10

# 大连理工大学学位论文版权使用授权书

本学位论文作者及指导教师完全了解“大连理工大学硕士、博士学位论文版权使用规定”，同意大连理工大学保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许论文被查阅和借阅。本人授权大连理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，也可采用影印、缩印或扫描等复制手段保存和汇编学位论文。

作者签名：

导师签名：

2017年5月11日

# 引　言

随着互联网的不断发展，网络成为人们获取信息的主要来源之一。目前人们上网获得信息的方式主要有两种：第一，直接输入所关心的网址，浏览网页，根据网页的内容及包含的超链接的引导，得到所需要的内容；第二，并不知道所需的信息的具体位置，而是通过某个网站的搜索引擎，在网络中搜索相关的信息。可见，搜索引擎已成为网络服务的一个重要的组成部分。然而，互联网上的搜索引擎，以通用搜索引擎居多，这样的搜索引擎可以很好的满足人们对普通信息的需求，但是，对于一些专业性比较强的领域，通用搜索引擎并不合适，主要体现在不能对专业信息很好的解读，要么查不到，要么查到了但垃圾信息过多，因此，需要一种专门应用于某一特定行业的搜索引擎，也就是所谓的专业搜索引擎。本人所在单位是一所水产特色院校，其渔业相关技术的研究在国内处于前列，为了能够对相关的教学单位、研究机构、生产企业提供准确的信息、快速的信息服务，因此，需要设计一个应用于渔业领域的专业搜索引擎，实现基于 WEB 的中文渔业信息检索服务。

《渔业信息搜索引擎的设计与实现》属信息技术研究领域，课题来源于大连市科技计划项目《基于 WEB 的智能渔业信息检索系统》。

本系统实现了一个搜索引擎的主要部分：搜索器、索引器、检索器。并结合专业搜索引擎的特点，研究了适用于专业搜索引擎的分词方法、相关度计算方法、网页消重方法等内容。

系统开发所用的工具有：

操作系统：WindowsXp

实现软件：VC++6.0

数据库：SQL Server2000

支持浏览器：MS Internet Explorer 5.X 以上

本文的结构大致是这样的：

（1）搜索引擎综述。

这部分着重讲述了搜索引擎的种类、国内国外研究现状以及搜索引擎的基本原理。具体内容见第一章。

（2）渔业信息搜索引擎的系统分析。

这部分着重讲述实际项目的背景和该系统开发的可能性和必要性以及系统的需求分析。具体描述见第二章。

（3）渔业信息搜索引擎的系统设计。

这部分的研究工作主要包括该系统各部分的基本原理，设计方案，系统的应用。具体工作详见第三章。

（4）渔业信息搜索引擎的系统展望

正对系统的现有状况，提出系统今后扩展的方向，及研究内容。具体内容为本文第四章。

# 1 搜索引擎综述

## 1.1 搜索引擎的分类

搜索引擎按照工作原理的不同，可以分为两个基本类别：全文搜索引擎（FullText Search Engine）和分类目录（Directory）。

全文搜索引擎的数据库是依靠一个叫“网络蜘蛛（Spider）”的软件，通过网络上的各种链接自动获取大量网页信息内容，并按以定的规则分析整理形成的。Google、百度都是比较典型的全文搜索引擎系统。

分类目录则是通过人工的方式收集整理网站资料形成数据库的，比如雅虎中国以及国内的搜狐、新浪、网易分类目录。

全文搜索引擎和分类目录在使用上各有长短。全文搜索引擎因为依靠软件进行，所以数据库的容量非常庞大，但是，它的查询结果往往不够准确；分类目录依靠人工收集和整理网站，能够提供更为准确的查询结果，但收集的内容却非常有限。在网上，对这两类搜索引擎进行整合，还产生了其它的搜索服务，主要有这两类：

①元搜索引擎(META Search Engine)。这类搜索引擎一般都没有自己网络蜘蛛及数据库，它们的搜索结果是通过调用、控制和优化其它多个独立搜索引擎的搜索结果并以统一的格式在同一界面集中显示。元搜索引擎虽没有“网络机器人”或“网络蜘蛛”，也无独立的索引数据库，但在检索请求提交、检索接口代理和检索结果显示等方面，均有自己研发的特色元搜索技术。

②集成搜索引擎（All－in－One Search Page）。集成搜索引擎是通过网络技术，在一个网页上链接很多个独立搜索引擎，查询时，点选或指定搜索引擎，一次输入，多个搜索引擎同时查询，搜索结果由各搜索引擎分别以不同页面显示。

## 1.2 搜索引擎的工作原理

（1）全文搜索引擎原理

全文搜索引擎的“网络蜘蛛”是一种网络上的软件，它遍历 Web 空间，能够扫描一定 IP 地址范围内的网站，并沿着网络上的链接从一个网页到另一个网页，从一个网站到另一个网站采集网页资料。网络蜘蛛采集的网页，还要有其它程序进行分析，根据一定的算法建立网页索引，才能添加到索引数据库中。上网时看到的全文搜索引擎，实际上只是一个搜索引擎系统的检索界面，当你输入关键词进行查询时，搜索引擎会从庞大的数据库中找到符合该关键词的所有相关网页的索引，并按一定的排名规则呈现。不同的搜索引擎，网页索引数据库不同，排名规则也不尽相同，所以，当以同一关键词用

不同的搜索引擎查询时，搜索结果也就不尽相同。

搜索引擎的原理，可以看做三步：从互联网上抓取网页→建立索引数据库→在索引数据库中搜索排序。

①从互联网上抓取网页

利用能够从互联网上自动收集网页的 Spider 系统程序，自动访问互联网，并沿着任何网页中的所有 URL 爬到其它网页，重复这过程，并把爬过的所有网页收集回来。

②建立索引数据库

由分析索引系统程序对收集回来的网页进行分析，提取相关网页信息（包括网页所在 URL、编码类型、页面内容包含的所有关键词、关键词位置、生成时间、大小、与其它网页的链接关系等），根据一定的相关度算法进行大量复杂计算，得到每一个网页针对页面文字中及超链中每一个关键词的相关度（或重要性），然后用这些相关信息建立网页索引数据库。

③在索引数据库中搜索排序

当用户输入关键词搜索后，由搜索系统程序从网页索引数据库中找到符合该关键词的所有相关网页。因为所有相关网页针对该关键词的相关度早已算好，所以只需按照现成的相关度数值排序，相关度越高，排名越靠前。 最后，由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

搜索引擎的 Spider 一般要定期重新访问所有网页（各搜索引擎的周期不同，可能是几天、几周或几月，也可能对不同重要性的网页有不同的更新频率），更新网页索引数据库，以反映出网页文字的更新情况，增加新的网页信息，去除死链接，并根据网页文字和链接关系的变化重新排序。这样，网页的具体文字变化情况就会反映到用户查询的结果中。

互联网虽然只有一个，但各搜索引擎的能力和偏好不同，所以抓取的网页各不相同，排序算法也各不相同。大型搜索引擎的数据库储存了互联网上几千万至几十亿的网页索引，数据量达到几千 G 甚至几万 G。但即使最大的搜索引擎建立超过二十亿网页的索引数据库，也只能占到互联网上普通网页的不到 30%，不同搜索引擎之间的网页数据重叠率一般在 70%以下。我们使用不同搜索引擎的重要原因，就是因为它们能分别搜索到不同的网页。而互联网上有更大量的网页，是搜索引擎无法抓取索引的，也是我们无法用搜索引擎搜索到的。

（2）分类目录的基本原理

和全文搜索引擎一样，分类目录的整个工作过程也同样分为收集信息、分析信息和查询信息三部分，只不过分类目录的收集、分析信息两部分主要依靠人工完成。分类目

录一般都有专门的编辑人员，负责收集网站的信息。随着收录站点的增多，现在一般都是由站点管理者递交自己的网站信息给分类目录，然后由分类目录的编辑人员审核递交的网站，以决定是否收录该站点。如果该站点审核通过，分类目录的编辑人员还需要分析该站点的内容，并将该站点放在相应的类别和目录中。所有这些收录的站点同样被存放在一个“索引数据库”中。用户在查询信息时，可以选择按照关键词搜索，也可按分类目录逐层查找。如按分类目录逐层查找，用户不使用关键词也可进行查询，只要找到相关目录，就完全可以找到相关的网站。

## 1.3 搜索引擎的研究现状

（1）国外搜索引擎的研究现状

从 1969 年美国国防部的计算机网络 ARPANET 起步,INTERNET 不断发展壮大,至今已快有 40 年历史。随着 INTERNET 的飞速发展,网上的信息量越来越多,目前 INTERNET 已经成为世界上最大的信息宝库,它已成为全球范围内传播科研、教育、商业和社会信息的主要渠道。其中 WWW(World Wide Web)的发展速度更是惊人。据统计,自从 1991 年诞生以来,WWW 已经发展成为拥有约 1 亿用户,近千万个站点,600G 信息容量的巨大分布式信息空间,而且这个数字仍以每 4 到 6 个月翻一番的速度增加[1]。

WWW 是建立在客户机/服务器模式上, 以 HTML 语言和 HTTP 协议为基础, 能够提供面向各种 INTERNET 服务的、一致用户界面的信息浏览系统[2]。WWW 所具有的超文本和超媒体的特殊结构, 带来了信息出版和传播的一场革命。WWW 上存储着大量有价值的信息, 从电子期刊、电子工具书、商业信息、新闻报道、大学和专业机构介绍、软件数据库、图书馆资源、国际组织和政府出版物、统计资料、教学大纲、专家背景介绍, 到娱乐信息等等, 吸引了大量的用户去使用和开发它, WWW 已展现出良好的应用和发展前景。

WWW 是一个开放性的全球分布式网络, 资源分布在全球不同的地域, 而且网上的资源没有统一的管理和结构, 导致了信息搜寻的困难, 如何快速准确地从浩瀚的信息资源中寻找到所需要的信息已成为困扰网络用户的一大难题, 这就是所谓的“Rich Data Poor Information”问题[3]。美国 Lycos 公司最近的一项调查显示, 80%被调查者认为互联网非常有用, 但他们同时为查询所需信息花费了大量的时间精力而抱怨。为解决这个问题, 各种网络信息检索工具应运而生[4]。

在 1991 年, XWAIS 提供了一个有着友好界面的信息搜索系统, 这就是搜索引擎的早期雏形, 同一年还出现了另外一个信息搜索系统, 即我们所称之为 GOPHER 的搜索软件[5]。而最早的真正意义上的搜索引擎是 Lycos, 创建于 1994 年的春天, Lycos 是 Michael

Mauldin 将 John Leavitt 的 spider 程序接入到其索引程序中形成的[6-7]。著名的搜索站点 Yahoo 也是在当年成立的[8], Netscape 也出现在 1994 年[9]。如今，搜索引擎的核心是网络导航服务。搜索引擎已成为一个网络门户，它们提供新闻、在线图书馆、词典，以及其它网络资源，它们提供了不仅仅是网站搜索的服务，它们的涉及面越来越广，也越来越有用[10]。比如，Yahoo 注重地是网站分类汇总服务，而如 AltaVista、Excite 等则注重提供庞大的搜索数据库[11-12]。一些网络导航服务并不提供搜索功能，他们侧重的是其它服务，但不论如何，搜索引擎为我们的网络生活带来了极大的便利，而且是免费服务。目前，互联网上有名有姓的搜索引擎已达数百家，其检索的信息量也与从前不可同日而语。比如最近风头正劲的 Google，其数据库中存放的网页已达 30 亿之巨！

随着互联网规模的急剧膨胀，一家搜索引擎光靠自己单打独斗已无法适应目前的市场状况，因此现在搜索引擎之间开始出现了分工协作，并有了专业的搜索引擎技术和搜索数据库服务提供商。像国外的 Inktomi，它本身并不是直接面向用户的搜索引擎，但像包括 Overture（原 GoTo）、LookSmart、MSN、HotBot 等在内的其它搜索引擎则提供全文网页搜索服务。国内的百度也属于这一类，搜狐和新浪用的就是它的技术。因此从这个意义上说，它们是搜索引擎的搜索引擎即元搜索引擎（meta-search engine）[13-14]。

国外的搜索引擎主要有：

Yahoo（http://www.yahoo.com）。

WWW 上最流行的搜索工具之一，是一种典型的目录式搜索引擎。1994 年 4 月由 Stanford 大学的两位博士创建，将信息分为 12 大类，每一类又分为多个专题，只要单击这些链接点就可以逐级深入目录，最终达到所需网页，同时也提供关键字检索方式[15]。

Altavista（http://www.altavista.digital.com）。

由 Digital Equipment Cooperation 公司的 Alto Pola 研究实验室开发，其最大的特点是在检索语句上与传统的联机检索语言相似。它可以对返回结果的格式进行控制，分为标准、压缩和详细三种格式。它还能提供简单搜索和高级搜索。高级搜索包括了简单搜索的所有特性，还允许使用布尔运算符、接近操作符和括号等[16]。

最近 AltaVista 推出了新一代搜索引擎，它具有简单易懂、全菜单式的界面，还能够让用户通过日期、时间、国家和语言等参数进行搜索，同时用户也可以自定义搜索范围和结果。

Infoseek（http://www.infoseek.com）。

1995 年由 Infoseek 公司推出，特点是采用了词频统计方法来确定词语的重要性和相关性，可按次序检索。它的优点在于速度快和使用方便[17]。

WebCrawler（http://www.infoseek.com）。

现在由 America OnLine 公司赞助的 WebCrawler 是一个杰出的搜索引擎，它支持“自然语言搜索”。同时它还提供了一些特殊的服务，如“反向搜索网络”（可以看谁连到了你的网页上）和网络统计功能等[18]。

Lycos（http://www.lycos.com）。

卡耐基.梅隆大学的著名查询工具，是最早出现的搜索引擎之一。它最大的特点是采用了一种可以大大加速数据搜索速度的技术，称为 Centispeed，另一个特色是建立了一个 Lycos 数据库，含有最常用主页的主题目录。它的优点在于速度快、使用简便、索引很大[19]。

Excite（http://www.excite.com）。

由 Architext Software 公司开发，Excite 最大的特点是采用了一个称为“智能概念抽象”的专用查询软件，允许用户用自然语言提问，目前本服务只能处理简单的布尔逻辑检索，还不能处理高级查询服务，具有一定的按例查询功能[20]。

Google 的搜索引擎（http://www.google.com）。

由斯坦福大学学生创建的风险公司开发，具有先进的技术实力，多家著名网站纷纷升级到 Google 的搜索引擎。当使用 Google 搜索引擎搜索网站时，一般都在 0.5 秒以内完成搜索任务。Google 搜索引擎的特点在于使用了数据挖掘（Data Mining）的技术和网站评级方法。数据挖掘技术是寻找所要搜寻数据的技术，Google 的网站评级方法则是通过分析重要网站如何插入链接以及分析其结构来作为判断网站重要性的依据[21-22]。

此外，常用的国外搜索引擎还有：

HotBot（http://www.hotbot.com）

OpenText（http://www.opentext.com）

Highway61（http://www.highway61.com）

DigiSearch（http://www.digiway.com/digisearch）

World Wide Web Warm（http://www.wwww.com）等。

（2）国内搜索引擎的研究现状

随着互联网络在中国的迅猛发展和普及，互联网络上的中文信息资源和以中文为母语的网上用户也急剧增加，现有的外文搜索引擎不能适应中文双字节的特殊要求。于是许多中文搜索引擎应运而生，包括大陆、香港、台湾在内的许多以中文为母语的地区都开发出了各种各样的中文搜索引擎。

1996 年 2 月台湾的“番薯藤”中文搜索引擎正式启动，是较早的中文搜索引擎。97 年 5 月“悠游”公司在香港建立了“悠游”中文搜索引擎。97 年 5 月 4 日 Yahoo 发布了“雅虎”中文搜索引擎。国内的搜索引擎的建立基本是在 97 年底及 98 年初起步，

“网易”搜索引擎于97年5月开始建设，“北极星”中文站点信息检索系统于97年12月开通，98年5月“搜狐”搜索引擎建立。此后一大批中文搜索引擎相继建立，如“常青藤”、“华好”、“搜索客”等。虽然大陆中文搜索引擎发展的起步较晚，但发展的速度很快，许多信息公司或机构都先后开发出了各自的中文搜索引擎[23]。

但总体上来讲国内搜索引擎仍处于国外搜索引擎发展的“容量建设期”，大部分搜索引擎网页搜集不超过百万。而且不论大小都是综合型搜索引擎，没有很好的专业型搜索引擎，搜索引擎的查询范围较窄，匹配精度不高。但也应看到，国内在搜索引擎的发展上有许多超过国外搜索引擎的优势，比如在中文词语切分、自然语言处理、全文信息检索方面有很强的技术实力。

中文网站搜索引擎主要有：

搜狐 （http://www.sohu.com.cn）。

搜狐 (Sohu)是由爱特信 (ITC)公司于1998年2月25日在北京隆重推出的有“中文网路神探”之称的大型网上中文查找工具，其技术是由麻省理工学院支持。它是以提供分类目录为主的中文搜索引擎，其分类原则是以图书分类为基础，与日常应用习惯相结合，由编辑人员分类，因而分类质量较高。它的信息抓取范围较其它中文搜索引擎的范围要广，不仅有国内站点，还包含国外的中文站点，日访问率达上万人次。搜狐还提供新闻导读、娱乐天地、企业集锦和网猴等服务项目。进入新闻导读栏目可阅读由ITC整理的新华社环球新闻，包括业界动态、Internet、Intranet和电子商务四个栏目的新闻。企业集锦是将国内的企业分类集中提供给用户，为用户查询提供方便，更重要的是为企业宣传提供了一条有力的渠道[24]。

百度搜索引擎（http://www.baidu.com）。

由百度在线网络技术（北京）有限公司开发的商业化搜索引擎，也是全球更新最快的中文搜索引擎，搜索可在1秒钟内完成。百度搜索引擎在充分考虑了中华文化特点的基础上，采用具有国际先进水平的计算机技术和搜索技术，全面解决了当前中文搜索引擎存在的弊端。百度搜索引擎的核心技术主要是由以下六方面组成：①百度“东方之蛛”网页高速收集技术；②百度智能化中文语言处理技术；③百度智能化相关性算法及搜索结果排序技术；④百度高可配置性技术；⑤百度智能化分布式结构与容错设计技术；⑥百度高效的搜索查法和高反应速度的整体设计体系[25]。

新浪(http://www.sina.com.cn)

新浪(Sina)是最大的中文门户网站 ，收录了全球资讯逾万的中文网址 ，并分成娱乐休闲、商业经济、社会科学、教育就业、社会文化、参考资料、政法军事、体育健身、科学技术、新闻媒体、文学艺术、电脑网络、医疗健康、生活服务、参考资料、国家地

域等 15 大类 ，其下又细分多个小类，并提供了中文关键词的搜索功能。

网易(http://www.yeah.net)。

网易(Yeah)搜索工具由广州网易计算机系统有限公司开发研制。它提供了类目浏览和关键词检索两种方式，类目浏览中有商业、教育、电脑、运动、政治、科学、娱乐等 12 个大类，各大类下又细分为若干小类。关键词检索支持全文检索，反馈信息包括网址、提要、长度、最近修改时间和相关度等。该工具还设有热门站点、新到站点和登录站点等栏目，并提供了与江苏接入网、国讯网络、厦门新华信息网、瑞得在线、金华热线等网络站点的链接[26]。

赛迪网推出垂直搜索引擎“IT 罗盘”（http://www.ccidnet.com）

“IT 罗盘”是国内第一个新一代基于 IT 行业的垂直搜索引擎，由国内领先的基于互联网的 IT 服务集成商——赛迪网（ccidnet.com）推出。垂直搜索引擎是面向某一领域、信息收录齐全、更新及时的垂直类搜索引擎。塞迪网的“IT 罗盘”，垂直定位于 IT，其中收录了大量经过严格过滤和人工加工的网站，结合了网站的分类检索、网址检索和网页精确检索等方面的优势，全面、精确地提供有关 IT 行业领域的信息资源和服务，更贴切地满足用户需求。“IT 罗盘”充分融入了个性化的设计，它为用户提供开放的接口，允许用户参与网站评价，同时允许用户定制自己的搜索需求[27]。

天网（http://bingle.pku.edu.cn）

由北大计算机系网络研究室设计开发，中国教育与科研计算机网示范工程应用课题之一，并被列入 CERNET 九五攻关项目。它提供一种检索 Web 资源及 FTP、NewsGroup 的手段。查询界面分为简单查询和复杂查询两种。由于该系统是基于分词的，因而人名和词库中没有的专业名词将查不到或查询效果较差[28]。

悠游（http://goyoyo.com）

由香港优联克公司和北京优联克科技开发有限责任公司共同开发。在北京和重庆设有镜像站点。悠游搜索引擎智能系统能对网上新网页和每日更新的信息进行自动搜索、识别，其中的关联性信息索引功能可自动在网页信息中搜索关键字，并将有关联性信息的网页一并找出。悠游能自动转化简繁字体、自动搜集英文、中文国际码和大五码的网页[29]。

（4）通用搜索引擎存在的问题和研究热点

搜索引擎的出现确实为人们在互联网上查找信息提供了有利的手段，然而现有的通用搜索引擎在搜索效率、信息维护、信息重复、专业化等方面还存在着一些问题和困难。

①大规模的分布式数据源

基于 Web 的自身特点，大量的数据分布在数以亿计的计算机互联网上，检索起来

困难重重。单个搜索引擎的索引数据库的覆盖率一般都低于 30%，很难索引所有 Web 资源[30]。

②网络信息的质量问题

互联网上的信息无论从数量和类型来看都呈指数增长，大量信息的存活期却缩短，索引数据库存储的文档和链接信息很有可能已经改变了位置或已经被删除。当用户沿着链接到远程站点访问这些信息时，便无法浏览到该网页。这个问题通常是通过使用一种称为链接机器人（Link Robot）的方法来解决的。目前，最常用的解决办法是该机器人定期对搜索树的一部分重新漫游，或重建整个搜索树，由于许多没修改的文档和站点也要重建，所以这并不是一个好的解决方案。

另外，网上大量的镜像站点和简单重复拷贝使得搜索引擎返回大量无用信息，搜索返回的结果成千上万，良莠不齐，造成“信息爆炸，资源库匮乏”。

③大量的动态网页无法检索

目前越来越多的 Web 网站使用了数据库和动态网页生成技术，而传统的搜索引擎无法检索到这些页面。

④异构数据源问题

网上检索要处理大量的多媒体信息，即使是文本信息也存在大量不同的文本格式。同时网上信息还存在多语种问题，亚洲语言字符的检索一直是信息检索界的一大难点。另外，Web 可访问多种格式的数据和多种类型的 Internet 站点，包括 FTP、HTTP、Gopher 以及 WAIS 等。网络搜索机器人和搜索引擎必须决定它将访问和检索哪些类型的 Internet 站点和哪些数据格式。

⑤忠实表达的问题

经典的信息检索界认为用户很难简单地用关键字来忠实表达他所真正需要检索的内容。表达的困难将导致检索结果的不理想，而且如何将结果表达成用户容易理解和使用的方式也是一个难题。

⑥搜索引擎的数据重复

常用的搜索引擎很少能够与其它的搜索引擎共享数据，其结果就是多个搜索引擎检索相同的资源和文档，多个机器人搜索访问同样的 Web 站点，这无疑带来了不必要的网络和服务器负载。

由于现有的搜索引擎有上述很多的缺陷，因此搜索引擎仍是网络和情报检索的研究热点，当前主要的研究热点有：

⑦能充分表达用户查询要求的查询语言

现有的搜索引擎的查询语言甚至比成熟的商业性的情报检索系统的查询语言还要

简单。当然这是由搜索引擎所处的网络环境所决定的。一套能充分表达用户要求但又不增加网络负载的查询语言是搜索引擎给用户的第一个良好的印象。

⑧索引数据库的组织和管理

与情报检索系统不同，搜索引擎的索引数据库是网络信息的一个轨迹。它要随网络信息的变化而变化，因此它除了数据增加以外还需要有数据的删除和修改功能。如何对大容量的、非结构化的信息进行增、删、改操作也是一个值得研究的问题。

⑨信息的自动加工

在传统的情报检索中，数据源基本上是人工加工且有标准的用词（词表），查全率和查准率都比较高。而搜索引擎对网上收集到的信息一般是采用自动加工，因此如何对信息进行准确的分类和标引是搜索引擎要研究的主要问题。

⑩提高检索的查准率

网上的信息相当丰富，现有搜索引擎的问题不再是能找到多少文献，而是找到了太多的文献，且许多文献不一定与用户要求非常相关。因此提高查准率是搜索引擎查找效率的主要体现。

⑪Web 信息的挖掘

信息挖掘是研究如何迅速发现和收集网上新加入的信息和被删除的信息，以及如何利用信息之间的各种关系等。网络搜索引擎对网络研究人员和情报检索研究人员都是一个值得研究的课题。

（5）专业搜索引擎的研究现状

通用搜索引擎的出现很大程度上解决了人们在互联网上查找信息的困难，但是目前通用搜索引擎在使用上面临上面提到的许多问题。随着信息社会的进一步发展，人们对信息的需求又有了新的趋势。近些年来，科学技术对于国民经济发展的带动作用越来越明显。高科技企业层出不穷，个个产业的科技成分也越来越高。如何为科技工作者搜集最新的科技信息，如何为商业决策者提供最新的业内新闻，对科技的发展和企业的经营都是至关重要的。

面对通用搜索引擎发展所遇到的困难和人们对信息的新需求，人们提出了对搜索引擎新的要求：

①运行在常规的软硬件设备之上。

②只搜索某一特定学科或特定专题的 Internet 信息资源。

③能够方便地对专题和学科进行配置。

为了满足这些新的要求，专业搜索引擎应运而生。

所谓专业搜索引擎就是以构筑某一专题或学科领域的 Internet 网络信息资源库为目

标，智能地在互联网上搜集符合这一专业或学科需要的信息资源，能够为包括学科信息门户、专业信息机构、特定行业领域、公司信息中心、行业专家等在内的信息用户，提供整套的网络信息资源开发方案。

专业搜索引擎与普通搜索引擎存在着很大的差别：

①服务目的不同

普通搜索引擎面向任何用户提供对任何信息的查询。而专业搜索引擎则面向专业用户向他们提供对其所在专业的信息检索。

②搜索方式不同

普通搜索引擎对网络进行逐页的爬行，试图遍历整个 Web。而专业搜索引擎则采用一定的策略预测相关网页的位置，动态的调整网页爬行方向，使系统尽可能的在与主题相关的网页集中的地方爬行，这节约了大量的网络资源。

③对硬件和网络的要求不同

普通搜索引擎需求过大，而专业搜索引擎由于没有遍历整个 Web 节约了大量的网络资源，并且没有自己的大型索引数据库，所以硬件要求也比较低。

目前，专业搜索引擎大都处于研究和试验阶段，利用它搜索的结果再经过专业人士的加工而形成的面向某一学科、领域的网络垂直门户网站也已经出现。本文所论述的搜索引擎系统就是属于这一类。

# 2 系统分析

## 2.1 系统背景

本文设计并实现的渔业信息搜索引擎系统来源于大连市科技计划项目《基于 WEB 的智能渔业信息检索系统》。

本人所在单位是一所水产特色院校，从事渔业相关技术的教学、科研以及与相关企业联合指导实际生产的工作。从网络获取渔业相关信息，已成为科研、教学、生产过程中获取信息的主要来源。鉴于目前的商业搜索引擎均为大型通用搜索引擎，缺乏专题或领域专用检索系统。通用搜索引擎，不能对专业的信息进行很好的解读，一种情况是对于一些专业术语不能识别，根本找不到相关信息；另一种情况是，所识别的信息面太大，参杂了许多与专业无关的垃圾信息，查找起来费时费力。再有，对一些专业的权威站点，搜索深度、精度不够，遗漏信息过多，而商业搜索引擎，由于考虑到商业目的，往往会将部分与专业关联不大的站点的信息，放在首位。当然，网络上也存在一些关于渔业信息的专业搜索引擎，但多数都是针对某一站点的站内搜索。因此研究一个具有个性化的、查准率较高的、具有定制服务功能或主动服务功能的中文渔业信息专用检索系统是十分必要的。该项目的成功实施将为渔业教学单位、生产企业和研究机构提供准确的信息、快速的技术服务和科学的知识，提高他们的工作效率，以适应现代社会的快速发展。用信息产业提升渔业这一传统产业，为建设“海上辽宁，海上大连”，建设东北老工业基地，建立和谐社会提供有价值的信息。

本系统的设计来源于该项目，但并不是该项目的全部，相当于整个项目的一个原型，其目的是研究搜索引擎的基本原理，包括搜索策略、分词技术、排序技术等方面的内容，为整个项目的实施，解决技术难题。

## 2.2 可行性分析

Internet 上资源是巨大的，可利用的信息是无组织的，并且信息每一天都在增加，更新。如何从 Internet 上获取对渔业教学单位、生产企业和研究机构最有价值的专业信息，是一个亟需解决的问题。因此，有必要设计并实现一个渔业信息搜索引擎。我们在设计该软件项目之前，应该进行可行性分析。国家《计算机软件开发规范》中指出，可行性分析的主要任务是“了解用户的需求及现实环境，从技术、经济和社会因素三个方面分析并论证软件项目的可行性，编写可行性研究报告，制定初步的项目开发方案”[31]。

（1）系统经济可行性

现今，多数水产行业工作者从网络搜索专业相关信息都是使用通用搜索引擎，但实

际效果并不好。例如：海水养殖工作者想知道最近关于海水盐度变化的信息，通过百度搜索引擎搜索“海水盐度变化”，搜索结果为找到相关网页约 40，200 篇，用时 0.0075 秒。然而，这些信息中，绝大多数都是一些关于地理知识的内容，以及一些商业站点提供的其它行业的付费技术文档，这些信息对于渔业生产没有任何意义。而工作人员从这些信息中，筛选有用的信息显然是低效的。另外，由于通用搜索引擎的网络蜘蛛，覆盖面比较大，不可能在短期内频繁爬行站点，用户获得的信息并不及时。因此，开发专门针对水产行业的搜索引擎，可以为渔业教学单位、研究机构提供准确的信息、科学的知识，提高他们的工作效率，为水产企业提供快速的技术服务，间接的提高了经济效益。另外，本系统开发成功后，可形成产品，直接为相关的院校网站、专业网站提供搜索引擎组件。可见，开发渔业信息搜索引擎在经济上是十分可行的。

（2）系统技术可行性

对于专业技术人员来说，他们所需要的信息是比较专业的，并不是杂乱，没有方向的信息。专业服务网站的搜索引擎就是从已知网站的网址出发，通过网站内部的超文本链接确定新的搜索点，然后用机器人周游这些新的搜索点，通过对网页的原文件分析，来查找有没有用户想要的信息。搜索算法可以采用深度优先搜索算法和宽度优先搜索算法，用户可以自主选择搜索算法。由于对指定网站进行搜索，降低了搜索范围和实现难度，使得该设计可以在现有的时间、能力上得以实现。

①开发的软件可行性

专业服务网站搜索引擎作为应用软件系统可以用于任何连网的计算机上。考虑到搜索信息的及时性及软件的运行速度等原因，选择 VC++作为开发工具。VC++为用户提供了 Windows 所一贯坚持的非常友好、操作简单的用户界面、完善而强大的网络和数据库操作功能（通过 VC++6.0 结合 SQL 语句实现）和简洁的数据库接口——开放数据库互连标准（ODBC）[32]。ODBC 由微软公司定制，它不但定义了 SQL 语法规则，而且还定义了 C 语言与 SQL 数据库之间的编程接口。这样，经过编译的单个 C 或 C++程序就可以对任何带有 ODBC 驱动程序的数据库管理系统（DBMS）进行访问了。ODBC 实质上使用户接口同具体的数据库管理过程分离，方便用户自己配置系统资源。

结合本系统的需求，使用 VC++6.0 工具进行开发完全能够满足要求。因此从软件技术角度看，开发实施本系统是可行的。

②开发的硬件可行性。

由于本项目是对大规模的网络信息加以分析、索引，所以对硬件的需求比较高，普通的个人计算机不能满足系统的要求，因此，在项目之初，已经购置一台 Dell 双双核心 Xeron 服务器，从硬件配置上，目前在没有大规模索引数据库的情况下，基本可以满足

要求。而本系统的前期开发阶段，主要是在原理上的研究、解决技术上的难题，因此，在普通的个人 PC 上即可完成。只要系统的配置可以满足 VC++6.0、SQL Server 2000 的运行即可，而当前的微机配置都远高于此要求，所以在硬件方面，本系统的开发也是可行的。

当然，在项目的正式实施阶段，根据实际的要求可能会需要构建一定规模的分布式系统。

③其他技术方面的可行性。

本项目的实施，计算机相关学科是重要的技术组成部分。除此之外，由于本项目是水产方面的专业搜索引擎，所以，与水产专业相关的技术知识也是本项目顺利实施的重要保证，在项目组中，包含了相关行业的专家，他们将在专业权威站点的搜集、专业词库的构建等方面提供更为有利的技术支持。

可见，本系统在无论在软件技术、硬件技术以及专业技术上均是可行的。

（3）系统应用可行性

本系统，所搜集到的内容，为各网站公开的内容，且系统并不向用户直接提供信息，主要向用户提供信息所在的 URL，用户要获得详尽的信息必须要链接到信息的原始出处，所以不涉及到违反相关法律法规的问题。

本系统，网络蜘蛛在搜集信息的过程中，对每个站点不会采用过多的线程，并且会控制一定的频率，另外，网络蜘蛛搜集信息，是周期性的，并不频繁，而用户的检索操作是针对本系统的信息库，而不是直接检索相应的站点。因此，不存在对网站的恶意访问的问题。

本系统，信息检索功能与当前网络上一般的搜索引擎的使用方式基本一致。而信息搜集功能有良好的用户界面，与普通的 Windows 程序的使用基本一致。因此，本系统具有操作可行性。

综上所述，本系统在经济、技术、应用上是可行的。

## 2.3 系统设计要求

本论文研究的重点是设计并实现出一个基于 WEB 的智能渔业信息检索系统。这个系统主要面对想要在网络上获得渔业领域信息的人员。针对渔业信息相关网站，该搜索引擎系统主要需求如下：

（1）为用户提供只与渔业相关的搜索信息。

（2）可根据用户的实际需要，提供单关键词搜索；多关键词搜索。

（3）为用户提供准确的、及时的搜索结果。

（4）用户定制搜索。

①用户可以锁定目标搜索。如锁定目标为某个网站，或者为某些网站。

②用户可以设定搜索深度，如设定搜索为一级搜索，或者为二级搜索，这样可以降低搜索范围，提高搜索效率。

（5）对于搜索到的结果集，应尽量将重要的信息放在前面。

（6）对于搜索到的结果集，应尽量去除重复信息。

# 3 系统设计

## 3.1 系统总体设计

渔业信息搜索引擎总体模块划分，如图 3.1 所示：

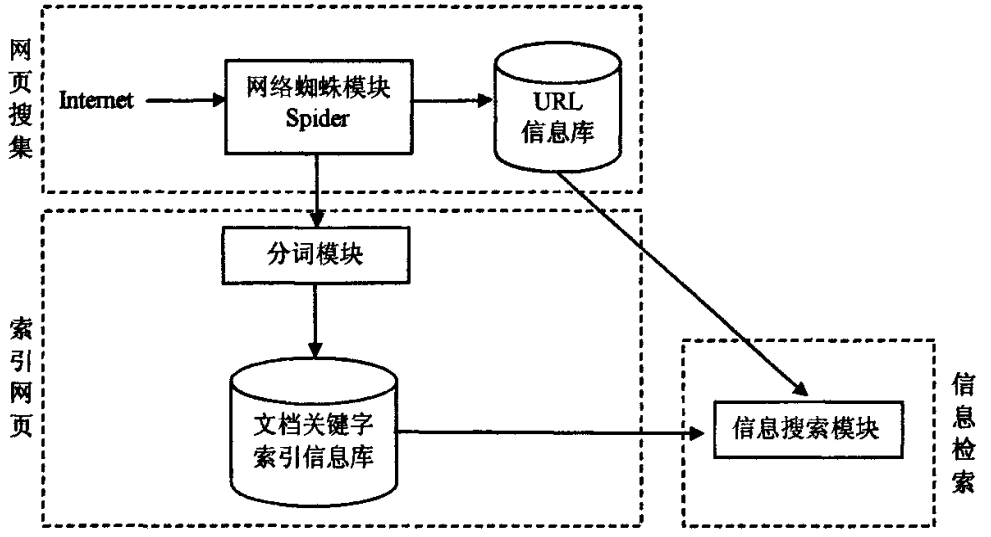


图 3.1 系统总体结构图
Fig. 3.1 System Structure Fig

（1）网络蜘蛛模块（Spider）

Spider（也称为 Spider 或 Crawler）实际上是一个基于 Web 的程序，它从一个初始网页出发遍历互联网，自动地采集网上信息。当 Spider 进入某个超文本页面时，它利用 HTML 语言的标记结构来搜索信息和获取指向其它超文本的 URL 链接，通过一定的算法选择下一个要访问的站点继而转向另一个站点继续搜索信息。理论上，如果为 Spider 建立一个适当的初始文档集，它就可以遍历整个网络。但实际上，这样的目标是不可能实现的，再有，对于专业搜索引擎，这也是不现实的。本文设计的搜索引擎根据具体的要求，可设置网络蜘蛛的初始站点集；设置爬行层次；设置站点范围约束。

（2）URL 信息库

URL 信息库主要存储 Spider 搜集来的网页，主要存储网页的 URL 信息，用来记录本搜索引擎目前所能涉及到的网页文档的集合。在这些信息库中，由于信量庞大，要充分考虑对信息的快速检索。

（3）分词模块

搜集到的网页，以一篇文档的形式给出，如果所有的检索请求都从这些文档中逐字直接匹配来获得，是不可能的。一般做法是，将文档分割成词，对词项加以匹配。对于英文这是很简单的过程，因为英语是以词为单位的，词与词之间用空格分隔。但是，对于中文，是以字为单位，如何将字组成词，是个难题。分词成功后，还应对网页信息库建立以词项为关键字的索引。

（4）信息检索模块

针对用户具体的搜索请求，检索网页信息库，寻求与用户请求相符的网页的URL。还涉及到对搜集到的网页排序，去除重复网页。

## 3.2 网络蜘蛛设计

### 3.2.1 网络蜘蛛的基本原理

网络蜘蛛即 Web Spider，把互联网比喻成一个蜘蛛网，那么 Spider 就是在网上爬来爬去的蜘蛛。网络蜘蛛是通过网页的链接地址来寻找网页，从某一个或某一组初始页面开始，读取网页的内容，找到在网页中的其它链接地址，然后通过这些链接地址寻找下一个网页，这样一直循环下去。对于搜索引擎来说，要抓取互联网上所有的网页几乎是不可能的，许多搜索引擎的网络蜘蛛只是抓取一定的链接深度，或是抓取一些特定站点。对于渔业信息搜索引擎来说，属于专业搜索引擎的范畴，大多数有价值的信息都集中在少数的权威的专业站点内。本系统中，Spider 的初始站点集，可由用户添加；搜索层次，搜索范围，均由用户指定。

在抓取网页的时候，网络蜘蛛一般有两种策略：深度优先遍历和广度（宽度）优先遍历。深度优先遍历是指网络蜘蛛会从某一层次的某一链接开始，一个链接一个链接跟踪下去，处理完这条线路所有链接之后再转入该层次的其它链接，继续跟踪。而广度优先搜索策略是指在抓取过程中，在完成当前层次的搜索后，才进行下一层次的搜索。由于多数网站，都采用扁平化的网站结构设计，重要的信息存放在较浅的层次上，且最深层次一般在 3～4 层之间，因此广度优先遍历更适合，这也是最常用的方式。广度优先遍历，如图 3.2 所示。

### 3.2.2 网络蜘蛛的基本流程

网络蜘蛛按照广度优先遍历的方法，从网络上不断获取新的URL，对于每一个URL又要获取其网页文档。网络蜘蛛的基本流程如图 3.3 所示。

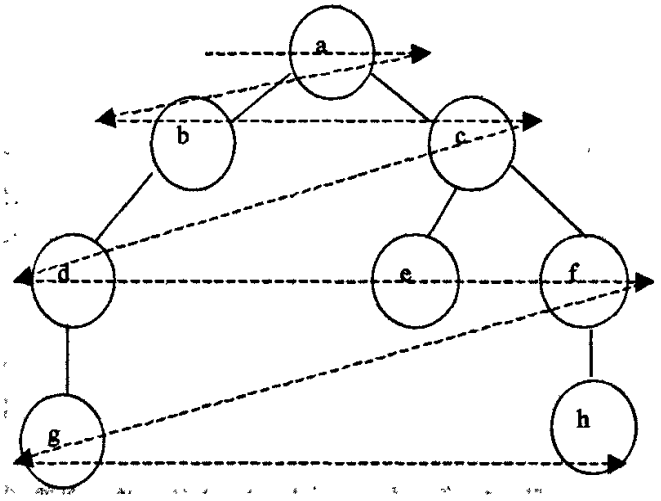


图 3.2 广度优先遍历方法

Fig. 3.2 Breadth-First Search

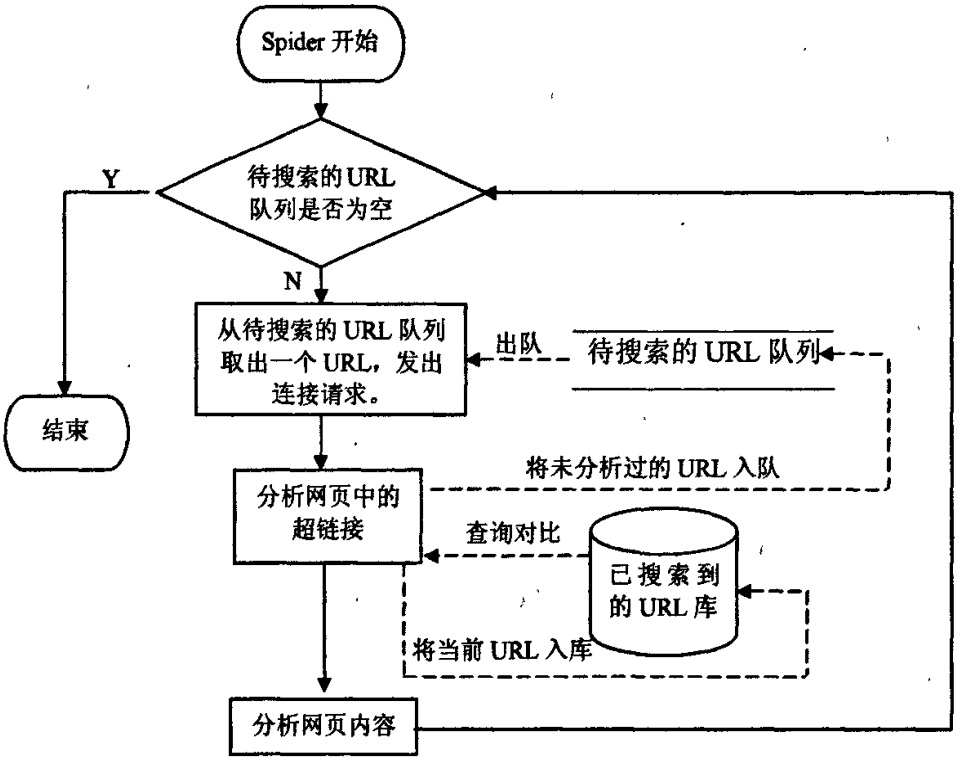


图 3.3 网络蜘蛛流程图

Fig. 3.3 Flow Chart of Web Spider

首先，先将起始 URL 存入到待搜索的 URL 队列中，Spider 程序每次从队列中取出一个 URL 发出链接请求，获取网页源文档，然后，加以分析，并且将已分析过的 URL 保存到数据库之中。对网页的分析主要包含两个方面：一方面，提取网页中的超链接信息，以获取新的 URL，为了避免 URL 的重复搜索，每个获取到的 URL，都应查询已搜索到的 URL 库进行对比，确认其并未搜索过，再将它存入到待搜索队列的尾部。另一方面，应分析网页的内容，进行分词，提取关键词等操作。应该说网络蜘蛛在整个信息搜集系统中处于核心地位，分词模块、对网页的索引等功能均由它来调用。

### 3.2.3 网页的初步分析

Internet 上，Web 网页主要使用的是 HTML（超文本标记）语言，因此对于网页的分析也主要是对 HTML 语言的分析。HTML 文档实际上是一种纯文本文档，在文档中以<>括起的部分称为标签，标签主要是用来标注文档的结构、外观、链接等信息。

（1）对网页内容的初步分析

当向某一 URL 发出链接请求后，便可获得该 URL 对应的文档，当然所获得的并不一定是 Html 文档，有可能是一些其它类型的文件，在本系统中，目前，仅考虑 Html 文档的情况，所以，当获取某一 URL 的响应后，首先，先判断文件类型、长度，对于非 Html 文档予以舍弃。

对于获取的 Html 文档，要进行初步的分析，包括：获取文档标题、获取文档关键字、获取超链接、剔除无用标签、获取内容文本。

网页的标题，是对整篇网页内容的统领。对于文档标题的获取，只需提取文档中的<title></title>标签中间的内容即可。但可能有一些文档的标题是如下内容：“无标题”、“欢迎访问”、“Untitled Document”等，这类标题，对整个网页文档的内容没有任何统领作用，这样的标题，可以保存下来，但在以后的分析过程中，不做任何考虑。

现在许多网页的设计者，特别是专业网页的设计者，一般会在网页的设计中直接指出文档中的关键字。关键字在 Html 文档中是在 Meta 标签中给出的，<Meta Name="KEYWORD" CONTENT="具体的关键字">，因此只需要分析 META 标签，并且 Name 属性为 KEYWORD 时，提取 CONTENT 属性的内容即可。

对网页超链接的提取。HTML 语言中超级链接的语法表示为：<a href="http://www.域名">或者<a href=" http://fishery.aweb.com.cn/technic.html">。这样可以通过提取<a>标签的 href 属性来提取超级链接。但是对于一些网页中使用相对 URL，例如：<a href="technic.html">或者<a href="/technic.html">或者<a href="../ technic.html">等情况，需结合当前页面的 URL 和当前站点的信息，将相对的 URL 转换成绝对的 URL。Web 网页中，

除了静态的 HTML 网页外，还存在大量服务器端解析的动态网页，如 asp、jsp、php 等，其 URL 形如："http://fishery.aweb.com.cn/technic.jsp" 或者 "http://fishery.aweb.com.cn/DisplayNews.jsp?NewsID=63572" 的形式。此类网页，其主要特点是用户的请求会在服务器端做特定的处理后，然后转换成相应的 HTML 文档，再发送给用户，即服务器端解析，而对于用户来说，获得的仍然是一篇 HTML 文档，所以处理方式与 HTML 网页基本相同。

获取网页文本内容。HTML 网页中，一部分是标签；另一部分是文本内容。文本内容是一篇网页中真正的信息体，因此，必须提取文本内容。并且，后续对网页的进一步分析（分词），主要是对文本内容的分析。一般情况下，剔除所有 HTML 标签后，即是文本内容。也有例外，如：<Script>标签标注的文本，是一段脚本程序，不是真正的信息。还有，<Style>标签标注的文本，在定义、使用样式表，也不是真正的信息。对于这些标签中修饰的文本，应剔除掉。除了这些外，在本系统中，也没有单纯的提取文本内容，还保留了一部份标签，因为，一些标签对内容是有辅助作用的，如：<B>标签标注的内容代表其文字是粗体显示，显然，是因为里面的内容比较重要。对于这类标签应保留，以备后面细致分析网页时使用。本文考虑到的这类标签主要有：<title>、<Hn>类标签、<B>、<Big>、<UL>、<OL>等等，不一一列举。

本系统中对网页初步分析的设计如下：先定义一个"读取标签的模块"，该模块的作用是，能够读取文档中的一个标签及属性。网页初步分析采用一遍扫描法，网页初步分析模块的流程如图 3.4 所示。

（2）获取网页源代码的方法

网页 Html 源代码的下载是使用 WinInet API 函数实现。WinInet 是以 WinSock 为基础，比 WinSock 更高层的 API，它在编写高性能的客户程序方面远远超过了 WinSock。WinInet API 提供给程序员一些用于编写 Internet 程序的高层借口，这些接口隐藏了编程过程中的烦琐的细节。WinInet API 具有降低编程难度、安全性好、兼容性好、支持多线程、支持数据缓存等优点。WinInet API HTTP 编程的步骤为图 3.5 所示。具体如下：

① 调用 InternetOpen()完成对 Internet DLL 初始化工作，并返回一个会话句柄。

② 使用返回来的会话句柄调用 InternetConnect()，同时设置 INTERNET_SERVICE_HTTP 标志，开始建立与 HTTP 服务器的连结，并返回一个连结句柄。

③ 利用连结句柄调用 HttpOpenRequest()或 HttpOpenRequestEx()函数打开一个 Http 请求，并返回 Internet 句柄，该句柄可以用于其它 HTTP 函数。但 HttpOpenRequest()并不会把 HTTP 请求直接发给 HTTP 服务器，发送请求由 HttpSendRequest()函数完成。

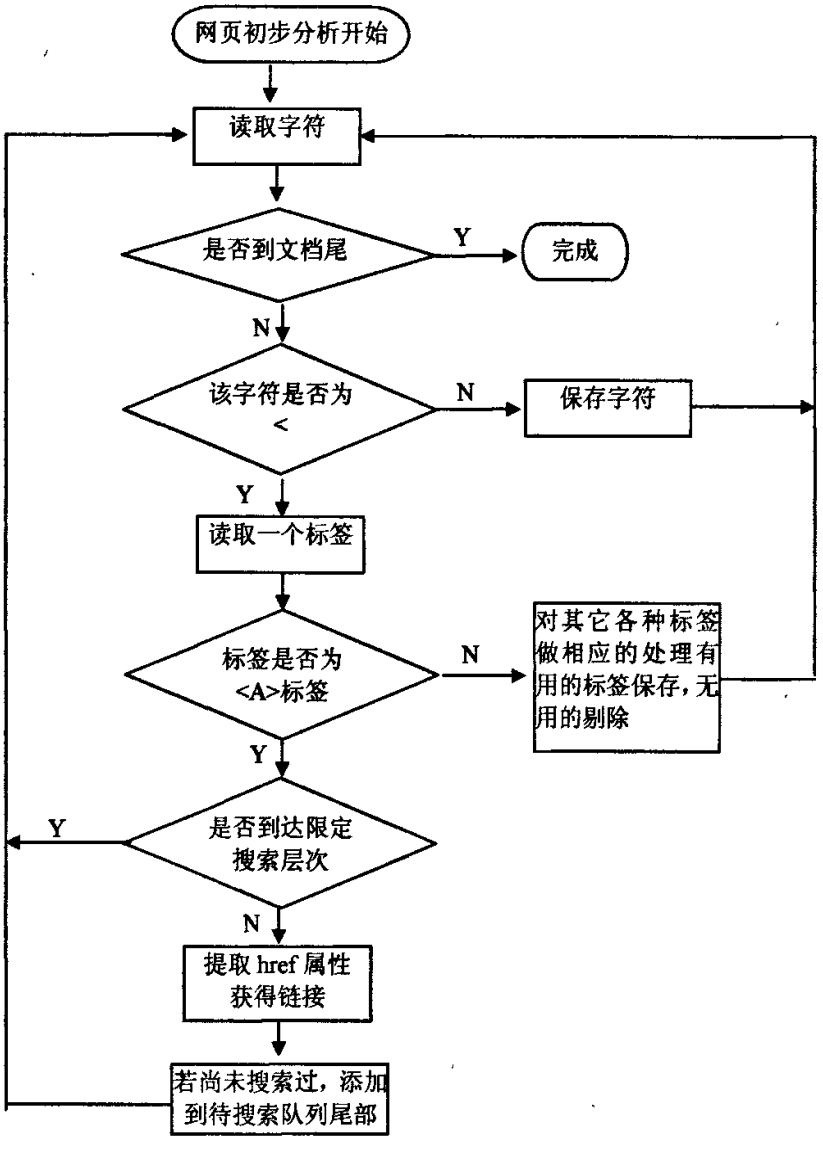


图 3.4 网页初步分析流程图

Fig 3.4 Flow Chart of Web Page Initial Analysis

④ 利用 Internet 句柄调用 HttpSendRequestEx()返回的句柄有效，那么就可以利用由 HttpOpenRequest()或 HttpOpenRequestEx()返回的 Internet 句柄进行诸如上传数据、查询信息或下载数据操作。

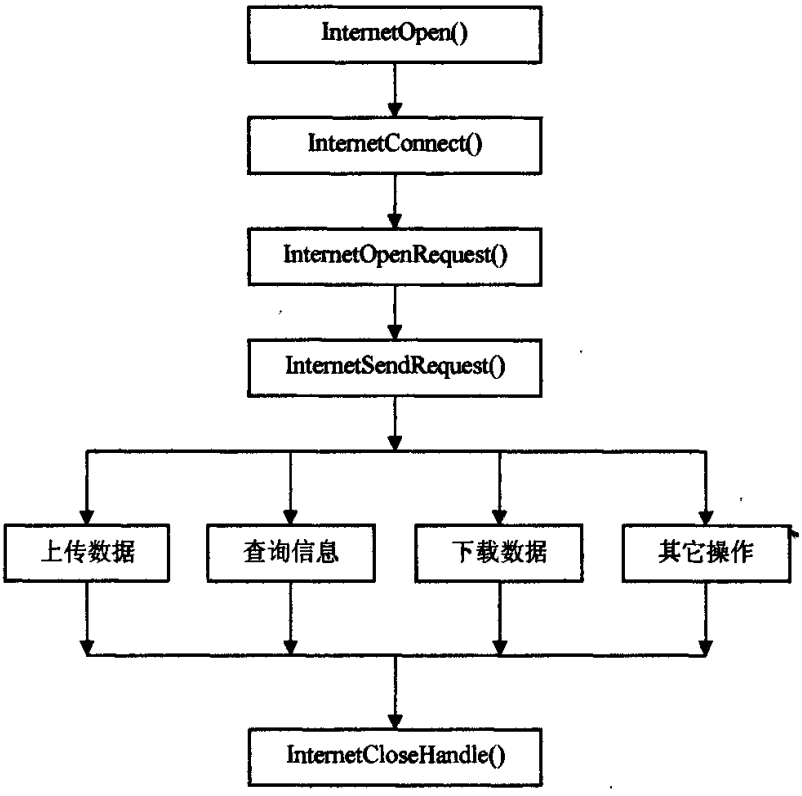


图 3.5 HTTP 编程的步骤
Fig. 3.5 Process of HTTP Programme

### 3.2.4 多线程提高效率

由于网络蜘蛛所要爬行的 URL 是大量的，只靠单线程去爬行，显然是低效的。为了有效地提高爬行效率，充分的利用系统资源，提高系统的吞吐量，对于网络蜘蛛采用多线程的设计方案，这样相当于多个蜘蛛并行爬行。

线程在计算机中是独立运行的基本单位。一个运行着的程序，称为进程，它是独立分配资源的基本单位。而在一个进程内可以创建多个线程。线程可以并发执行，有效的提高系统效率。在线程的并发执行中，要充分的考虑到对共享资源的访问，如果处理不当，会引起执行结果错误，严重的会引起死锁。除此之外，多个线程之间并不是独立的，线程间还要考虑通信的问题。

Visual C++中对多线程提供了很好的支持。Visual C++提供了两种线程：用户界面线程和工作者线程。用户界面线程派生于 CWinTread 类，较复杂。而工作者线程容易实现，也最常用。创建辅助线程，首先，定义线程处理函数，把需要并发执行的代码编写在其

中，线程处理函数的形式如下：

```
UINT 线程处理函数名（LPVOID pParam）
{
    并发执行的代码；
}
```

若该函数要么是全局函数，要么是静态（static）成员函数。其中，pParam 是一个指向处理参数的通用指针。定义好该函数后，就可以用 AfxBeginThread（线程处理函数名，参数指针）启动线程。当该线程处理完函数返回后，线程被撤销。

Visual C++中还对线程的同步提供了几种方案：

（1）临界区（CriticalSection）。可针对某段代码作 Lock 操作，被锁定后的代码，不允许其它线程访问，直到执行 UnLock 操作。

（2）互斥量（Mutex）。该对象只允许一个线程访问，一般用来管理对互斥资源的访问。

（3）信号量（Semaphore）。该对象包含一个计数器，一般用来代表系统中某项临界资源的可用数目。

（4）事件对象（Event）。一般用于两个线程间的通信。它有两种状态，有信号态和无信号态。一个线程可用 WaitForSingleObject 函数等待某一事件对象。

（5）消息机制。

本系统中，网络蜘蛛的多线程的设计，采用了 VC 的工作者线程方式，即设计了一个线程处理函数。该处理函数，与单线程的处理基本一致，但还需考虑以下问题。

对于待搜索的 URL 队列、URL 信息库均属于多个线程的共享资源，即每个线程都需要从待搜索的 URL 队列取出 URL，分析 URL 所对应的网页超链接，核对 URL 信息库，将没有分析过的 URL 添加到待搜索的 URL 队列；还要将已分析好的 URL 添加到 URL 信息库中。因此，要实现对待搜索的 URL 队列、URL 信息的互斥访问。

线程数目的问题。系统内的线程数目并不是越多越好。一方面，对系统线程的总数目要加以限制，系统线程当到达一定数量后，由于线程间的同步问题，以及系统为管理各个线程所付出的代价，反而会使系统性能下降。因此，应将系统内的总线程的数目小于一个固定值，这一点比较容易实现。另一方面，对访问某一网站的线程数目，也要加以控制，因为许多网站，为防止用户恶意攻击，对同一用户访问网站的连接数加以控制。另外，由于太多的线程同时访问一个站点，所获得的重复 URL 也会增多。对于这一点，系统中，定义了一个数组，用来存放当前系统中，正在访问站点的域名，及正在访问该站点的线程数目，当某一线程从待搜索的 URL 队列中读取（并不出队）一个 URL，其

恰好属于正在访问的某一站点，且已达到该站点的连接限制，则放弃分析，从待搜索的URL 队列中向后再取一个单元。在本系统中，最大的线程总数目为 10，某一站点最大线程数目为 3。

综合，考虑以上因素，本系统的主线程，如图 3.6 所示。

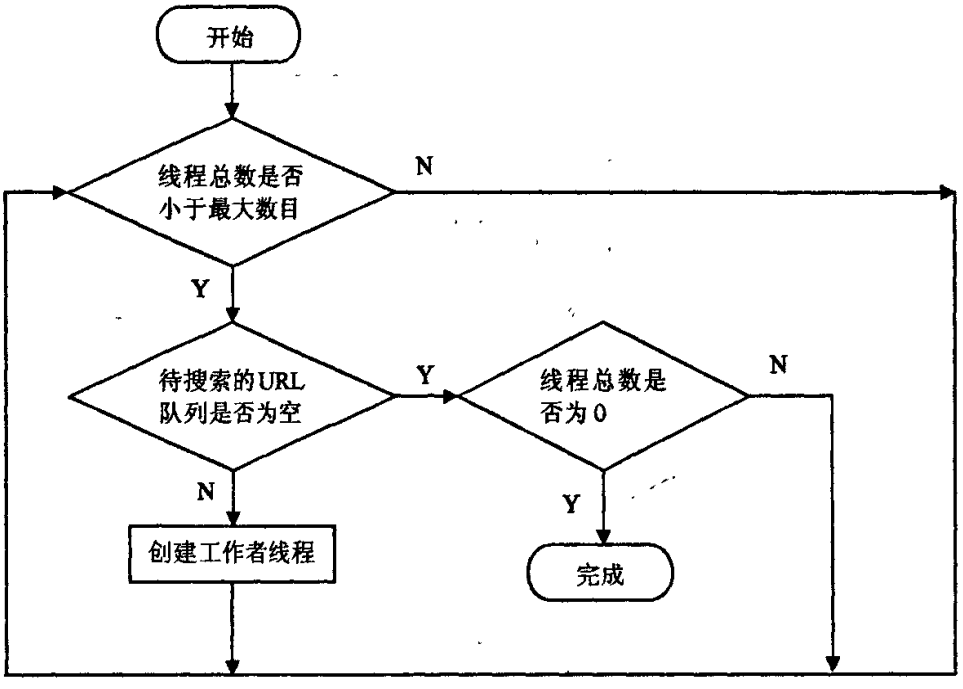


图 3.6 主线程流程
Fig. 3.6 Flow Chart of Main Thread

每个工作者线程的流程，如图 3.7 所示。

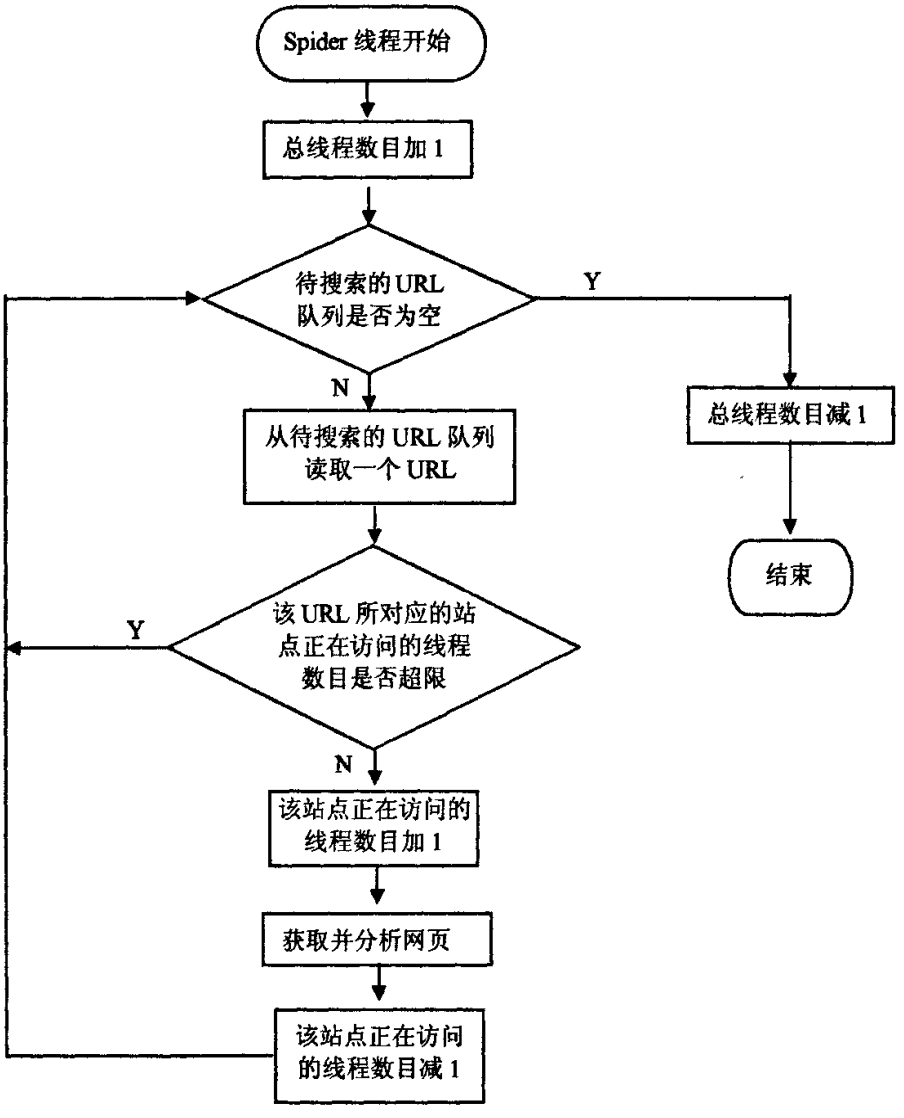


图 3.7 工作者线程流程
Fig. 3.7 Flow Chart of Worker Thread

### 3.2.5 相关类的设计

本系统，采用了面向对象的程序设计方式，将一些主要数据、过程封装在类之中。与网络蜘蛛相关的类主要有以下几个。

（1）CURL 类。该类中封装了有关于 URL 的属性，及对 URL 的操作。其主要成员的类图如图 3.8。

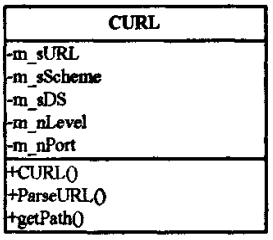


图 3.8 CURL 类图
Fig. 3.8 CURL Class Figure

m_sURL：存储 URL 字符串。为 CString 类型。

m_sScheme：存储 URL 的协议名。

m_sDS：存储该 URL 的域名。为 LPCSTR 类型。

m_nLevel：存储该 URL 所在的层次。为 int 型。

m_nPort：用此 URL 连接时，所用的端口。为 int 型。

ParseURL（）：从 m_sURL 成员中分析出，该 URL 所对应的协议名、域名、所用端口，分别赋给 m_sScheme、m_sDS、m_nPort 成员。

getPath(int)：从 m_sURL 成员分析出，相对路径所对应的字符串，其中 int 型的参数表示相对路径的级别，0 为当前路径，1 为上一级路径，-1 为根路径。返回值为 LPSTR 型。

（2）CWebPage 类。表示网页类，它封装了网页的基本属性、以及获取网页、分析网页。其主要成员的类图如图 3.9 所示。

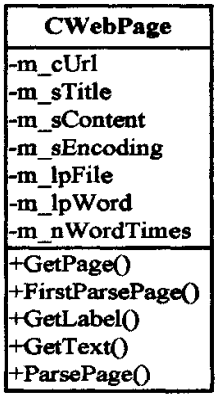


图 3.9 CWebPage 类图
Fig. 3.9 CWebPage Class Figure

m_cURL：表示该页面的 URL。为 CURL 型。

m_Title：表示网页的标题。为 LPSTR 型。

m_sContent：表示网页的关键字。为 LPSTR 型。

m_lpFile：存储网页内容的文件的指针。

m_lpWord：该网页分词后所得到的词汇列表的指针，每个词项是一个结构类型，包含该词的词名，出现的次数。

m_nWordTimes：记录网页中的总的词的使用次数，为后面统计词频使用。

GetPage（）：该方法用来发出连接请求，并获取网页。

FirstParsePage（）：该方法初步分析页面。

ParsePage（）：该方法是对页面全面的分析，之中会先调用 FirstParsePage（）方法，然后，完成分词操作。

GetLabel（）：从当前位置上向后读取的一个标签，包括它的标签名、属性值。

GetText（）：从当前位置起向后到第一个标签止，读取一段文本。

### 3.2.6 网页信息库的设计

对于网络蜘蛛所搜集到的 URL、以及获取的有用网页文档，应将其保存起来，以备信息检索模块查询。由于网络蜘蛛搜集到的信息量庞大，因此，如何建立一个高效的索引系统，以便从庞大的数据中完成快速检索，是搜索引擎研究的一个课题。

在本系统中，对于网页信息的存储、索引，正在研究中，目前并未定义单独的索引文件系统，而是基于数据库实现的。数据库管理系统，采用的是微软的 SQL Server 2000。数据库名为 SearchEngineDb，其中，与网络蜘蛛模块相关的表，主要有下面两个。

（1）表 URLInfo。该表主要是用来存储蜘蛛所搜索到的 URL。其结构如表 3.1 所示。

蜘蛛模块每次搜集到的 URL 都须与已经搜集的 URL 表项对比，为了在大量的 URL 信息中，快速检索，该表须对 url 字段作索引。

（2）表 WebDocInfo。该表主要用来存储所搜集到的 URL 对应的有用的页面内容。其结构如表 3.2 所示。

表 3.1 URLInfo 表
Tab. 3.1 URLInfo table

| 字段名 | 类型 | 索引 | 备注 |
|---|---|---|---|
| url | varchar | 有，主键 | 网页的绝对 URL 字符串 |
| dn | char | 无 | 主机域名 |
| scheme | char | 无 | 协议名，一般情况为 HTTP |
| port | int | 无 | 端口号 |
| level | int | 无 | 该 URL 所处层次 |
| haveAnalize | int | 无 | 为 1 时，表示该网页已分析完毕，且该网页有效；为 2 时，表示该网页已分析完毕，且网页无效；为-1 时，表示该网页尚未分析；为 0 时表示网页已分析完毕 |
| DocId | long | 有 | 表示该 URL 分析后的结果保存在 WebDocInfo 表中的编号，在允许为空 |

表 3.2 WebDocInfo 表
Tab. 3.2 WebDocInfo table

| 字段名 | 类型 | 索引 | 备注 |
|---|---|---|---|
| DocId | varchar | 有，主键 | 文档的编号 |
| DocType | varchar | 无 | 文档的类型 |
| Title | varchar | 无 | 文档标题 |
| Encoding | varchar | 无 | 文档编码方式 |
| Content | varchar | 无 | 文档的关键词 |
| WordList | text | 无 | 文档分词后词的列表，包括每个词的词频 |
| WordTimes | int | 无 | 总的词的使用次数，统计词频用 |

以上两个表之间具有一对一的联系，联系的字段为 DocId。但两个表不能合为一个表，因为在蜘蛛在爬行期间要频繁使用 URL 信息，要靠通过不断的查询以往搜集到的 URL 信息来避免重复爬行。对于蜘蛛模块，一旦将某个网页分析成功后，将其信息保存在 WebDocInfo 表之中，而该信息以后对蜘蛛模块就没有什么作用了，它主要是为信息检索模块服务的。因此，将 URL 信息与文档信息分开存储。再有，并不是每个 URL 所

对应的网页都有意义，对于一些无用网页，只需存储它的URL，以表示该URL已搜集过了，而没有保存网页文档的必要。

## 3.3 中文分词

对于搜集到的连续的网页文档，要将其分割成以词为单位的信息体。英文本身就是以词为单位的，词和词之间是靠空格隔开，分词不存在难度。而中文是以字为单位，句子中所有的字连起来才能描述一个意思。例如，英文句子I am a student，用中文则为："我是一个学生"。计算机可以很简单通过空格知道student是一个单词，但是不能很容易明白"学"、"生"两个字合起来才表示一个词。把中文的汉字序列切分成有意义的词，就是中文分词，也称为切词。我是一个学生，分词的结果是：我/是/一个/学生。

### 3.3.1 分词在搜索引擎中的作用

网络蜘蛛搜集到的大量的网页，如果用户的每个实际的搜索请求都从搜集到的网页的内容中直接寻求匹配，则搜索效率会很低，且会搜索到大量的无用网页。对于一篇网页，其内容中会存在大量的无实际意义的词汇，如："的"、"得"、"呢"等，而代表其关键信息点的词汇相对来说数量较小，且这样的关键词汇一般会重复出现多次。因此，利用分词将网页内容分隔成一个一个词语，分析这些词语，从中去除大量无实际意义的，提取少量的关键词汇，一篇网页的内容就可以用这些关键词来替代，而用户的实际搜索请求也不必从每篇网页的内容中进行搜索，只需与关键词匹配即可。显然，分词的准确程度对于搜索引擎的准确度影响很大。

### 3.3.2 中文分词技术

现有的分词算法可分为三大类：基于词典匹配的分词方法、基于统计的分词方法和基于理解的分词方法。

（1）基于词典匹配的分词方法

这种方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个"充分大的"机器词典中的词条进行配，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，又可以分为单纯分词方法和分词与标注相结合的一体化方法。常用的几种机械分词方法，有以下几种。

①正向最大匹配法（由左到右的方向）

该算法通常称为MM法，其基本思想为：设D为切分参考字典，Max表示为D中

的最大词长，Str 为待分的句子或字串，MM 法是每次从 Str 中取出长度为 Max 的一个子串（假设 Str 长度大于 Max，当小于 Max 时，则取出整个子串）。把该子串与 D 中的词进行匹配，若成功，则该子串为词，指针后移 Max 个汉字后继续匹配。若不成功，则把该子串最后一个字去掉，再与 D 中的词匹配，如此匹配下去，直至匹配成功或至该串只剩一个字为止（表示该字可当作词，可在该字后面开始切分）。该切分算法优点是执行起来简单，不需要任何的词法、句法、语义知识。没有很复杂的数据结构，唯一要求就是需要一个词典 D。缺点是不能很好的解决歧义问题，不能认识新词。根据资料统计，匹配的错误率为 1/169[33]。正向最大匹配法流程如图 3.10 所示。

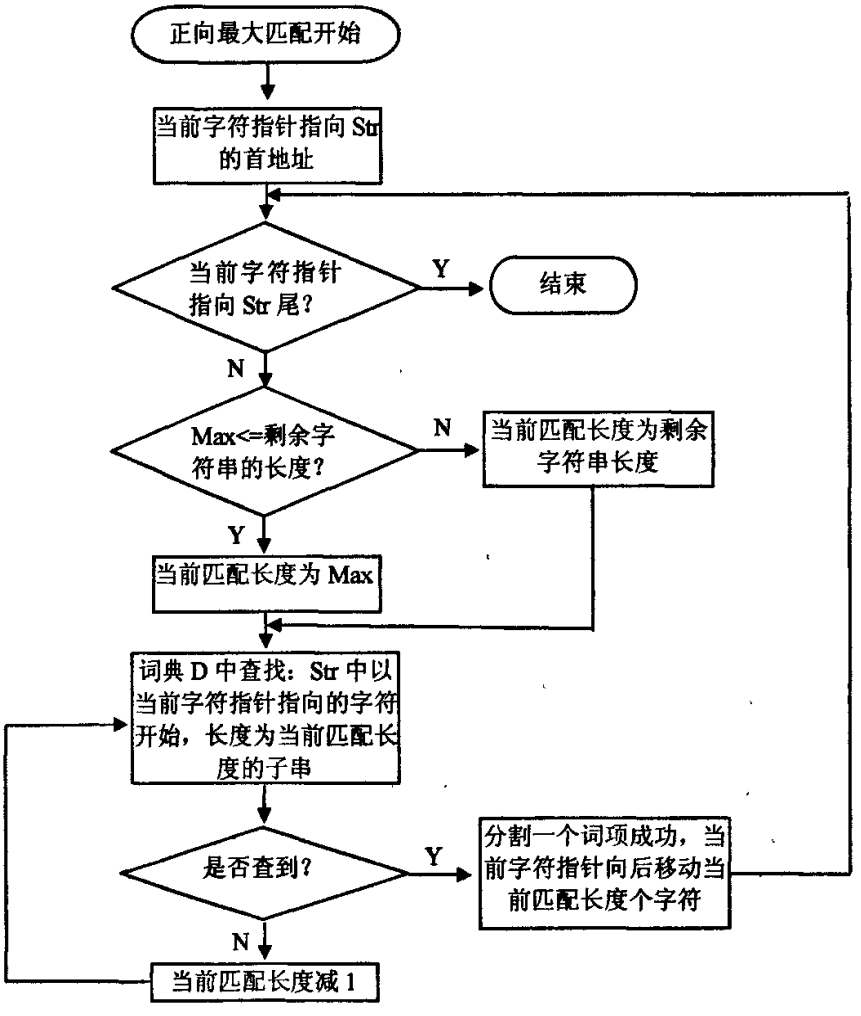


图 3.10 正向最大匹配法流程图
Fig. 3.10 Flow Chart of Maximum Matching Method

②逆向最大匹配法（由右到左的方向）

该算法通常称为 RMM 算法，该算法的基本原理与 MM 算法一样。不同的是分词的扫描方向，它从右向左取子串进行匹配。对于一个字串或句子 Str，从右边往左取 Max 个汉字，与 D 中进行匹配，若匹配，指针前移，再取 Max 个字串，直至第一个字为止，若不匹配，则去除最左边的字，再进行匹配，直至右边的词（n 个字，n>=1）被匹配出，然后，指针前移 n 个字，再取 Max 个字串，如此往复，直至该 Str 串全部被切分出。该切分算法与 MM 算法一样，实现简单，不需要任何的词法、句法、语义知识，需要一个词典 D。缺点也是不能很好的解决歧义问题，不能切分新词。RMM 切分算法比 MM 算法有更高的切分正确率，切分错误率减小到 1/245。

还可以将上述两种方法相互组合起来构成双向匹配法。由于汉语单字成词的特点。一般说来，机械分词的精度还远远不能满足实际的需要。实际使用的分词系统，都是把机械分词作为一种初分手段，还需通过利用各种其它的语言信息来进一步提高切分的准确率。

（2）基于统计的分词方法

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。定义两个字的互现信息，计算两个汉字 X、Y 的相邻共现概率。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，因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，并且对常用词的识别精度差，时空开销大。实际应用的统计分词系统都要使用一部基本的分词词典（常用词词典）进行串匹配分词，同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

（3）基于理解的分词方法

这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想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。由于汉语语

言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段[34]。

对于任何一个成熟的分词系统来说，不可能单独依靠某一种算法来实现，都需要综合不同的算法，需要多种算法来处理不同的问题。

在中文分词过程中，除了分词算法，还应考虑两大难题：歧义识别和新词识别。

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，可切分为"表面/的"和"表/面的"两种情况。歧义主要有以下两大类。

第一类，交叉歧义。在字段 AJB 中，AJ∈D，且 JB∈D，则称 AJB 为交叉歧义。其中 A、J、B 为字串，D 为字典。例如上面的"表面的"。

第二类，组合歧义。在字段 AB 中，AB∈D，且 A∈D，B∈D，则称为组合歧义。例如："门把手"中"把手"是一个词，而"把手拿开"中"把"和"手"分别为词

新词识别，是指对那些在字典中都没有收录过，新出现的词的识别。比如一些新出现的专业名词。这是分词需要考虑的另一大难题。

### 3.3.3 渔业信息搜索引擎分词的实现

由于本系统的搜索引擎属于专业搜索引擎，用户的搜索请求，大多为专业词汇，因此在分词过程中，专业词汇优先，基本方法选择逆向最大匹配法。

对于词典，主要建立 3 个词典。

（1）专业词典。主要存储关于渔业信息的专业词汇。如："饵料"、"海参"。

（2）通用词典。主要存储除了专业词汇，但有实际意义的普通词汇，主要以名词、动词为主。

（3）停用词词典。主要存储汉语中无实际意义的词汇。如："的"、"得"、"是"。

分词过程中，将文本划分成逐级细化的组织结构，如图 3.11 所示。

首先根据文本中的非中文字符（包括英文、标点、数字）作为分隔点，将文本划分成短句（两个分隔点之间的字串）。每个短句内再进行分词。在分词过程中，要尽量减少因歧义，而对专业词汇的错误分析。一般说来，在文档中，专业词汇之间不容易发生歧义，而专业词汇与普通词汇之间、普通词汇与普通词汇之间发生歧义的可能性较大。专业文档中，主要信息点一般是由专业词汇表达的，普通词汇间的少量歧义，对文档的分析结果影响较小。为了能减少专业词汇与普通词汇间的歧义，系统对专业词汇优先分析。对已划分好每个短句内，以专业词典为依据，应用逆向最大匹配法，划分出专业词汇。若有专业词汇，以每个专业词汇为分割点，将短句继续划分成更短的短句，对每个更短的短句内，以通用词典和停用词词典为依据，应用逆向最大匹配法，划分出其它词

汇。最后，用划分后的有意义的词建立网页索引。其主体流程如图 3.12 所示。

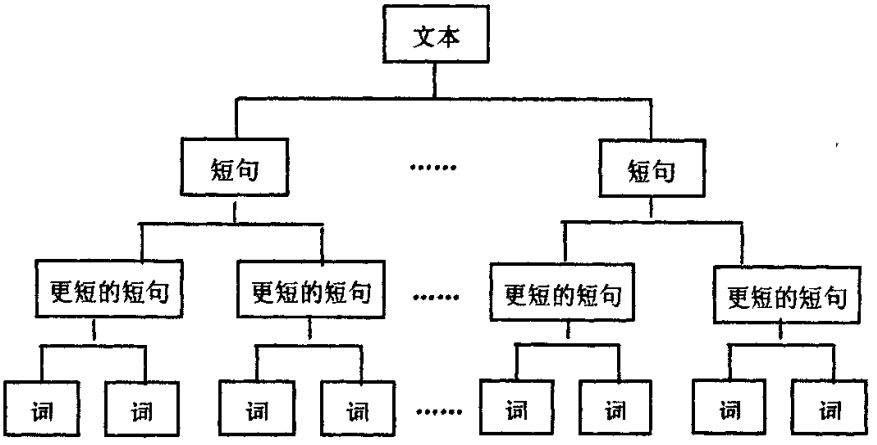


图 3.11 文本切分层次
Fig. 3.11 The Layers of Text Segmentation

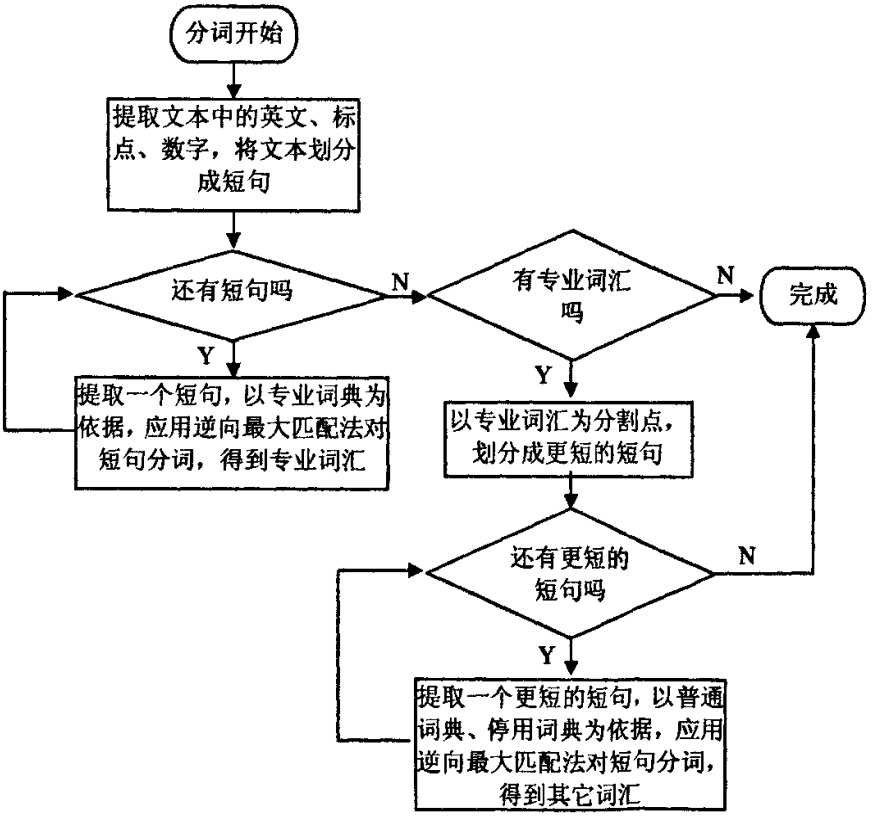


图 3.12 分词流程
Fig. 3.12 Flow Chart of Word Segmentation

### 3.3.3 词典的组织

词典是分词过程中的依据，它记录了大多数的汉语词汇，词典收录词条的多少直接影响分词的质量，因此，词典的数据量是较大的。另外，由于汉语的特点，词的长度也不尽相同，从1个至7个汉字不等。分词过程中，每个词的分析都依赖于查询词典以获得其信息，对词典的访问是频繁的。如何从大量的、不等长的记录中，快速的查询到所需的信息是非常重要的。

为了有效的提高词典的查询速度，在词典的组织上，用散列法为词典建立索引，并且采用了2级索引。

（1）词典包括三部分：首字Hash表、词索引表、词典正文，其组织形式如图3.13所示。

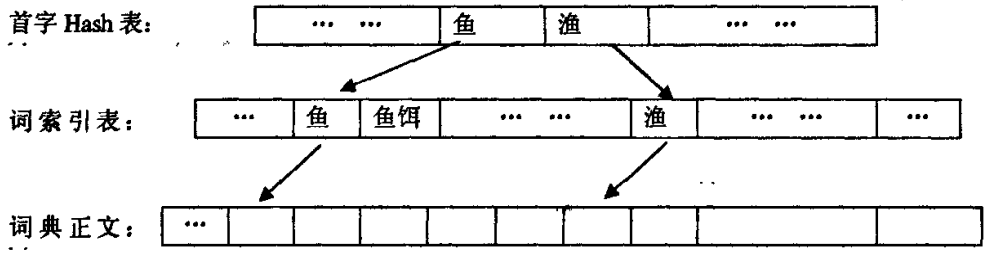


图 3.13 词典结构
Fig. 3.13 Structure of Dictionary

（2）首字Hash表的结构。对所有的前缀汉字在Hash表中都有唯一的一项与它相对应。Hash表中每一项的结构如图3.14所示。

| C | $P_0$ | $P_1$ | flag |
|---|---|---|---|

图 3.14 首字哈希表结构
Fig. 3.14 Structure of First Word Hash Table

其中，C：该前缀汉字；$P_0$：以该汉字作为前缀的首词在词索引表中的位置；$P_1$：以该字作为前缀的最后一个词在词索引表中的位置；flag：以该字为前缀的词条是否只有单字词的标志。

（3）词索引表的结构。对于词典正文中的每一个词在词索引表中都有唯一的一项与其对应，此索引表中每一项的结构如图3.15所示。

| W | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | $P_7$ | L |
|---|---|---|---|---|---|---|---|

图 3.15 词索引表结构
Fig. 3.15 Structure of Word Index

其中，W：为该词；Pi：以该词前 i 个汉字组成的单词在词典正文中第一次出现的位置，其中 i 从 2 到 7；L：以该词为前缀的词条数目。该索引表是有序表，在查找时，可采用二分查找的方法。

（4）词典正文结构。

字典正文主要有词名、词性等属性。

### 3.3.4 倒排索引网页

对某一网页文档分词后，得到一组词项的集合，这组词项作为网页文档的主要特征，自然要将它们添加到网页文档信息库之中，于是得到了以文档为索引，与词项之间的对应关系，如图 3.16 所示。

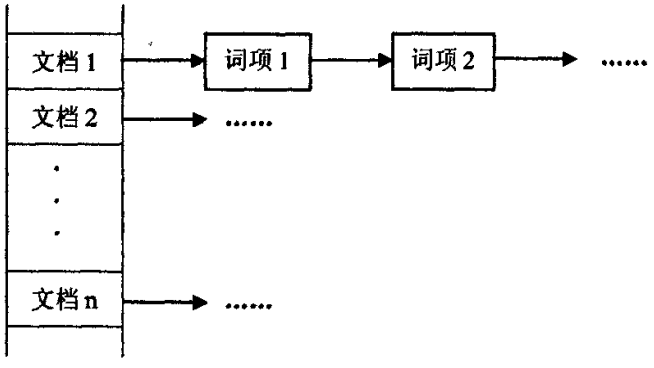


图 3.16 以文档为索引的结构
Fig. 3.16 Structure of Index as Document

这种文件可称为正排文件。显然，通过正排文件，很容易以文档为出发点，得到这篇文档包含哪些词项。但是，这样的索引形式，不利于用户的搜索请求，因为用户的搜索请求是以一组关键词为出发点的。当用户给出某一词的搜索请求，为了确定哪些网页中含有该词，则需要查询网页信息库中收录的所有网页文档后才能得到结果，这是非常低效的。因此，应该建立从词项到文档的索引文件，这种文件称为倒排文件。如图 3.17 所示。

这样，对于用户提出的某一词的搜索请求，只需查询倒排文件中相应的词项，即可

得到包含该词的所有文档，而对于倒排文件中词项可采用散列方法组织，进一步提高查询速度。

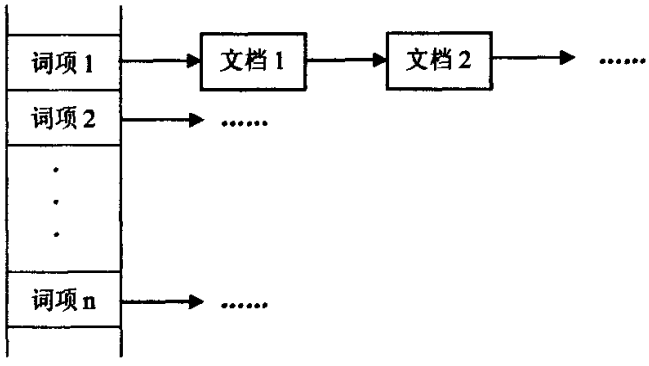


图 3.17 倒排文件结构
Fig. 3.17 Structure of Inverted File

实际建立倒排索引，除了要指明词项和文档的对应关系外，还应指明该词项在文档中出现的频数和位置，所以每个倒排表项为：词项（(文档编号 1，频数（位置 1，位置 2，……位置 n)),（文档编号 2，频数（位置 1，位置 2，……位置 n)) , …… )。例如：海参（(1，3（11，26，51)，(7，2（6，21))。这说明“海参”这个词在 1 号文档中出现过 3 次，位置分别为 11、26、51，在 7 号文档中出现过 2 次，位置分别为 6、21。

本系统的实现过程中，并未单独定义索引文件系统，而是基于数据库实现的，其倒排文件的实现方法为：在数据库 SearchEngineDb 中，建立 InvertedFile 表，该表的结构如表 3.3 所示

表 3.3 InvertedFile 表
Tab. 3.3 InvertedFile table

| 字段名 | 类型 | 索引 | 备注 |
|---|---|---|---|
| WordName | char（20） | 有，主键 | 词项名 |
| DocCount | long | 无 | 记录出现该词的文档总数 |
| DocList | varchar | 无 | 记录出现该词的每个文档编号，出现次数 |

其中，DocList 的格式是：文档编号，频数 / 文档编号，频数 / ……，每次只需读取该字段，获得一个字符串，对字符串分析，即可得到相应的文档编号及频数信息。

## 3.4 信息检索模块

该模块的主要作用是根据用户的搜索请求，检索到符合条件的网页的集合，并对集合作相应的处理，以得到对用户最有意义的结果集。

### 3.4.1 基本检索功能

用户的搜索请求最简单的情况就是单关键词检索，对于这种情况，只需直接对倒排文件检索，即可得到包含该词的网页文档编号的集合，然后，根据网页文档编号分别检索 URL 信息表、网页内容信息表，便可得到相应网页的 URL 和内容信息

若用户的搜索请求是多关键词检索，应对每一个关键词分别检索倒排文件，得到一组分别包含相应关键词的集合，然后，再对这组集合取交集，即可得到包含所有关键词的网页文档编号的集合。再检索 URL 信息表、网页内容信息表，便可 URL 和内容信息。

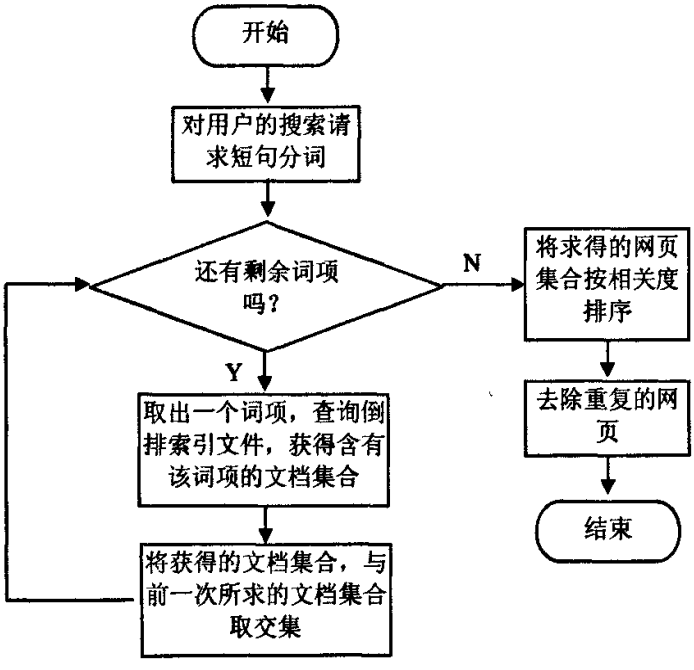


图 3.18 信息检索流程
Fig. 3.18 Flow Chart of Information Search

用户的搜索请求也可能不是词汇级的检索，而是给出一个短句，这种情况下，应先对用户的短句分词，将短句变成词的集合，然后，与多关键词检索的处理相同。

信息检索的流程如图 3.18 所示。

### 3.4.2 相关度排序

通过基本的检索功能，可以得到包含用户所有关键字的网页的集合，这样可能会有很多网页，但是，在这些网页中，仅有少部分是对用户有意义的，根据查询结果与用户的查询请求之间的相关程度，对网页排序，将那些与搜索请求相关程度大的排在前面显示。

（1） 相关度排序技术。

①文档的向量表示

对于一个文档 D，经过分词后，将得到由词项 t1，t2，…，tn 组成的集合。每个词在文档 D 中起的作用不同，把每个词在文档中的起的作用称为该词的权重。tk 词对应的权重记为 Wk。文档 D 的特征，可以看作由两个方面来构成。一方面，各个词权重之间的比例；另一方面，所有词共同作用的结果，文档越长共同作用的值就越大。将所有词项的权重 W 组成一个向量 $(W_1, W_2 \cdots, W_n)$。如图 3.19 所示。

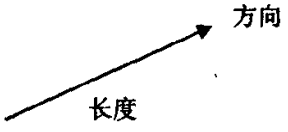


图 3.19 文档向量
Fig. 3.19 Document Vector

显然，该向量的方向可以代表各权重分量间的比例，而该向量的长度可以代表所有分量的共同作用。因此，文档 D 可以看成是由 n 个词项 t1，t2，…，tn 所对应的权重 $W_1, W_2 \cdots, W_n$ 组成的向量，这就是文档 D 的向量表示。

②余弦相似度

当用户给出一组关键词或是一个短句的搜索请求，便可得到一个查询向量 q。一个文档向量 D 与 q 的相似程度，可以作为相关度排序的依据。为了让短文档和长文档具有同样的机会，并不考虑向量长度的因素，因此，只要向量 D 与向量 q 方向相近，即可认为 D 与 q 有较高的相似程度。如图 3.20 所示。

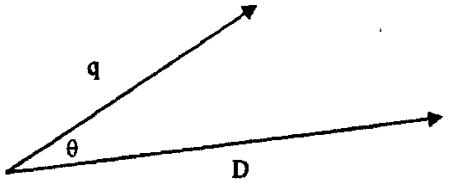


图 3.20 文档向量相似度
Fig. 3.20 Similitude of Document Vector

q 与 D 方向相近，就是 q 与 D 之间的夹角 θ 要尽量的小。由向量的内积公式，如公式（3.1）

$$\vec{D} \bullet \vec{q} = |\vec{D}||\vec{q}|\cos\theta \tag{3.1}$$

可得 D 与 q 之间夹角 θ 的余弦值，如公式（3.2）

$$Sim(\vec{D},\vec{q}) = \cos\theta = \frac{(\vec{D} \bullet \vec{q})}{|\vec{D}||q|} \tag{3.2}$$

其中，$Sim(\vec{D},\vec{q})$表示 D 与 q 之间的相似度。

这种相似度称为余弦相似度。由于，向量的各分量的值均为正数，所以向量 D 与 q 之间的夹角最大为 90°，$Sim(\vec{D},\vec{q})$的值为 0，此时，可以认为 D 与 q 之间没有任何相似。向量 D 与 q 之间的夹角最小为 0°，$Sim(\vec{D},\vec{q})$的值为 1，此时，可以认为 D 与 q 最相似，但不是相同。

由上面的结论，可以看出，在查询结果集中，应对每个文档与查询向量之间求余弦相似度，然后，按照余弦相似度的降序排列。

③权重的衡量

一个文档中，将所有词项的权重 W 组成的向量，便是该文档的向量表示。最著名的的权重衡量方式，就是 TF*IDF 公式。其中 TF 称为“单文本词汇频率”（Term Frequency ），即某一词在该文档中出现的频率，也就是用词的次数除以网页的总字数。显然，TF*IDF 公式考虑到在 IDF 一定的情况下，一个词在一篇文档中出现的频率越高，权重越大。IDF 称为“逆文本频率指数”（Inverse document frequency ），它考虑的是，对于所搜集到的所有文档而言，若某一个词在大量的文档中都出现过，则该词预测主题的能力较弱，在该文档中起的作用就小，如“的”、“得”、“是”这样的词汇，几乎在所有的文档中都出现过，但它们对预测一篇文档的主题没有任何作用。相反，若某一个词在少量的文档中出现过，则该词预测主题的能力较强，在该文档中起的作用就大。IDF

的公式为 Ln(D/Dt)，其中 D 是全部网页数， Dt 表示含有词汇 t 的网页数目。比如，我们假定总网页数是 D=10000，停用词“的”在所有的网页中都出现，即 Dt=10000，那么它的 IDF=Ln(10000/10000)= Ln (1) =0。假如专业词“烂鳃病”在 20 个网页中出现，即 Dt=20，则它的权重 IDF=Ln(500)=6.2。又假定普通词汇“预防”，出现在 5000 个网页中，它的权重 IDF=Ln(2)=0.7。由此，可见 TF*IDF 公式充分考虑到了，一个词在一篇文档中出现的频率及该词预测主题的能力两个因素。

（2）本系统的相关度计算。

本系统中的相关度计算基于公式 3.2 的余弦相似度，其中，权重计算公式采用了 TF*IDF 公式。但是在对词频 TF 的计算中，还考虑到了 HTML 标签的因素。

HTML 中的一些标签，对词项的权重是有影响的，比如说，<title>标签是用来标记网页的标题，而在标题中出现的词，对主题的预见能力要比普通网页文本的预见能力强的多，因此，用<title>标记的词应该获得较高的权重。HTML 标签中，类似<title>这样，能够影响词项权重的标签有很多，像<h1>、<b>、等，但并不是所有的 HTML 标签都对词项的权重有影响，像<img>、<frame>、<form>等，对词的权重无任何作用。对于能够影响权重的这类标签所标记的词汇，为了提高相应权重，系统中在词频上予以加倍计算。例如，可以认为在<titile>标签中出现 1 次的词，相当于该词在普通网页文本中出现 40 次，利用这种方法，可使该词的权重得到显著的提高。影响权值的标签及它们的加倍系数，如表 3.4 所示。

表 3.4 HTML 标签加倍系数表
Tab. 3.2 Times coefficient of HTML Label table

| 标签名 | 加倍系数 |
|---|---|
| <TITLE> | 40 |
| <CITE> | 8 |
| <EM> | 2 |
| <STRONG> | 4 |
| <B> | 4 |
| <I> | 2 |
| <BIG> | 4 |

续表

| | |
|---|---|
| <H1> | 12 |
| <H2> | 8 |
| <H3> | 4 |
| <H4> | 2 |
| <H5> | 1 |
| <LI> | 4 |
| <UI> | 4 |
| <A> | 4 |

### 3.4.3 网页的消重

对网页进行相关度排序后，可以保证用户最先看到是最重要的页面。但是，在查询结果中，还会有很多重复页面，因为在网络中，信息的转载是很平常的事情，特别是一些作用重要的页面，被转载的机会和次数可能就越大。如果，搜索结果中重复信息越多，则垃圾信息的含量就越高。为了让用户搜索到更为纯净的信息，系统在完成相关度排序后，为页面再加以消重的操作。

（1）消重的技术。最简单的判断两个网页的内容是否重复的方法就是两个网页的内容直接对比，但这样做显然是低效的。一种改进的作法是，对网页的内容采用 MD5 算法，然后再行比较。MD5 算法针对某一字符串可产生一个 128 位的唯一标识，只要字符串不同，产生的标识也不同，因此，两个网页的内容相同，则产生的标识也相同。这种方法可以严格的判断两个页面内容是否重复。

（2）本系统中相似度的计算

对于大多数网页在转载的过程中，不一定能保证内容与原文完全一致，多多少少会添加一些无用内容，因此上面的方法，过于严格。可采用前面所述的余弦相似度的作法，求两个文档向量 D1、D2 夹角的余弦值，来确定网页是否重复。但这种做法是有缺陷的，因为，余弦相似度只考虑到了两个文档向量的方向相近的问题。例如：文档 D1 的向量为（a，b，c），而文档 D2 的向量恰为（xa，xb，xc），即向量 D2 的每个分量恰好是 D1 每个分量的 x 倍，显然两者的夹角为 0°，余弦相似度为 1，但两个文档并不相同。因

为，余弦相似度求的只是“相似度”，而非“相同度”，而页面的重复，应该取决于两个网页的相同程度。因此，在判断两个文档是否重复时，所计算的相似度，不但应该考虑文档向量的方向相似，还应考虑长度相似。

本系统网页消重的计算基于这样的原理：

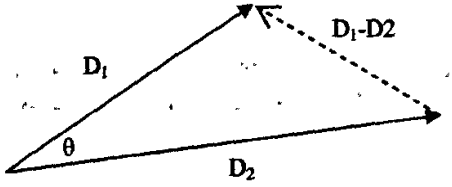


图 3.21 文档相同程度

Fig. 3.21 Degree of The Sameness of Document

如图 3.21,图中的虚线边的长度为|D1-D2|，这个长度可兼顾 D1、D2 之间的角度和长度因素。综合考虑长短文档的问题，应对|D1-D2|的值作归一化处理，让|D1-D2|/max（|D1-D2|)。显然，max（|D1-D2|）$=\sqrt{|\vec{D}_1|^2+|\vec{D}_2|^2}$。有公式（3.3)。

$$Sim(\vec{D}_1,\vec{D}_2)=\frac{\sqrt{|\vec{D}_1-\vec{D}_2|^2}}{\sqrt{|\vec{D}_1|^2+|\vec{D}_2|^2}} \tag{3.3}$$

本系统就是应用公式 3.3 来衡量相似度的，该相似度综合考虑到夹角和长度因素。但是这个相似度与余弦相似度正好相反，当两个向量的夹角为 90° 时，相似度为 1，此时两个向量差距最大；而当两个向量的夹角为 0° 且长度相等时，此时相似度为 0，而两个向量的差距最小。

## 3.5 系统应用

本文所设计并实现的是一个应用于水产渔业领域的中文专业搜索引擎，系统运行的环境为 Windows98 以上的操作系统，开发工具为 Visual C++6.0，后台数据库为 SQL Server 2000。本系统的界面主要包括：网址搜集界面、分词结果界面、信息检索界面等。

专业服务网站搜索引擎系统的网址搜集界面如图 3.22 所示。

在该界面上用户要选择搜索起始网址，确定搜索级别和搜索范围等项目。其中起始网址列表框，显示网络蜘蛛爬行的起始网址，单击添加按钮可添加新的起始网址，单击删除按钮可删除选中的起始网址。深度文本框可设置网络蜘蛛爬行的深度，若设置为 0，则表示不限定深度，此时，网络蜘蛛爬行的结束，需要通过单击“结束”按钮来完成。

筛选条件，主要是限定蜘蛛爬行的网站范围，如图 3.22 所示的该文本框的值“aweb.com.cn”，说明网络蜘蛛在爬行的过程中，只对域名以“aweb.com.cn”结尾的网址，进行搜集，还可设置多个筛选条件，中间用分号相隔，如“aweb.com.cn; yahoo.com”。搜索时间主要用来显示搜集网址所经过的时间。“定时”按钮可设置定时搜集的时间，当单击该按钮时，显示如图 3.23 所示的界面。


图 3.22 URL 搜集界面
Fig. 3.22 URL Search Interface

图 3.22 整个界面的下方的列表控件，主要用来显示目前所搜集到的网址，当在该控件内选中某一网址，单击“显示内容”按钮，便可显示所对应网址的标题及分词结果，如图 3.24 所示。

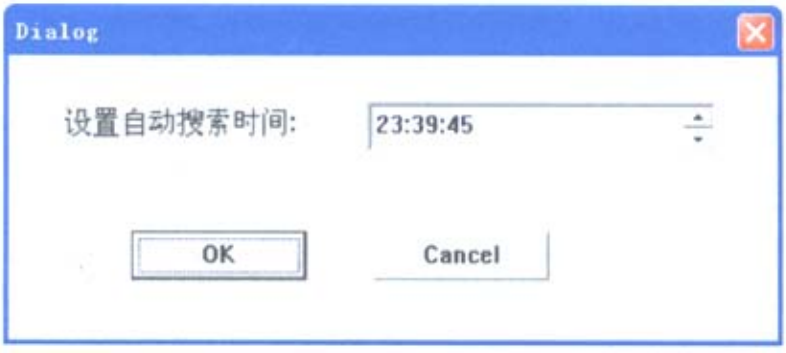


图 3.23 定时设置界面
Fig. 3.23 Time Setting Interface

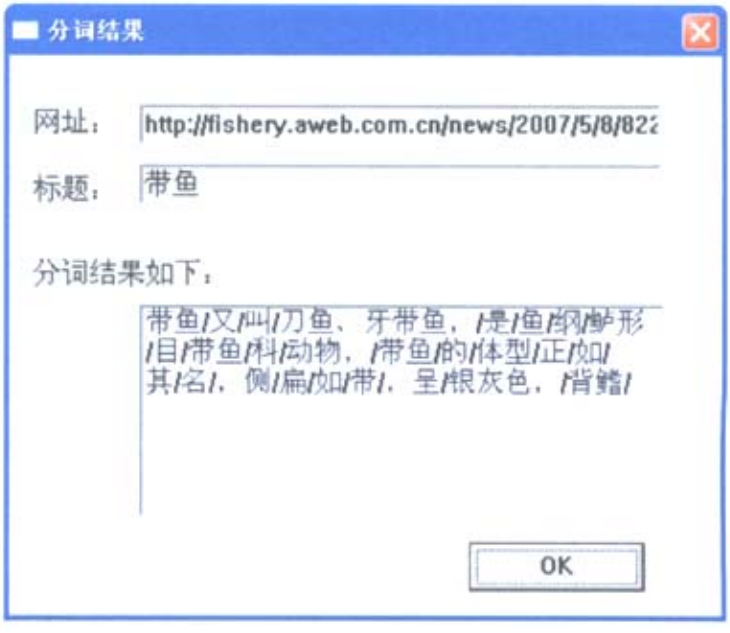


图 3.24 分词结果界面
Fig. 3.24 Word Segmentation Result Interface

本系统是为水产渔业领域的专业人员设计的专用搜索系统，主要用来搜索国内水产渔业领域专业网站的信息，如：水产渔业网，中国渔业网等。下面我们用本搜索引擎对“水产渔业网”的内容进行搜索。水产渔业网网址为“http://fishery.aweb.com.cn/”。

首先，将水产渔业网网址添加到“起始网址”列表。设置深度为 4，仅在该站站内搜索，所以筛选条件设置为“aweb.com.cn”。现在，可以进行网站内容的实时搜索，搜索主界面如图 4.1 所示。一级实时搜索速度较快，搜索时间在 10 秒以内，但是二级以上

网址搜索所需时间较长。总共搜集了从起始网址出发的4级页面，网页总数为512个，用时约2.7298分钟，速度还是可以令人满意的。

水产渔业网主要为JSP技术构成的动态网站，其中有相当一部分以jsp为扩展名的动态页面，经过测试，本搜索引擎能够准确的获得这类页面的内容，由此可见，本搜索引擎无论对动态页面还是静态页面，都有很好的适应能力。

通过上述试验，我们看出该搜索引擎可以很好的完成实时搜索功能，能够对动态网页进行搜索，具有用户定制功能，很好的满足了水产专业的工作者获得信息的要求，提高了工作效率。

由于本系统的词库尚在建设过程中，所以，分词结果及信息检索结果的准确度还不是很高，但随着词库的建立越来越完整，分词及信息检索的准确度将得到有效的提高。

# 4 系统展望

本系统所实现的渔业信息搜索引擎，属专业搜索引擎。为了增大系统的搜索规模、增加系统的搜索精度，还需要做很多方面的研究，扩展系统。下面分别从系统的主要模块出发，说明系统还需做的研究工作。

（1）网络蜘蛛模块

①随着搜索网站规模的扩大，数目的增多，为了提高搜集 URL 的速度，应尽量避免重复搜索。然而，在网络上，同一个资源的 URL，有可能有多种表达形式，比如：某些网站存在镜像站点，存在 URL 不同，但网站内容相同的现象。对于这种情况，可从一些 URL 的规律上加以解决，许多镜像站点的 URL 与主站点的 URL 存在一些相似之处。另外，也可搜集一些主要站点的镜像站点信息，建立镜像站点的信息库。还有一种情况，某些网站可以用域名加虚路径的形式访问，也可以用 IP 地址加虚路径的形式访问。对于这种情况，有些网站可直接用 IP 地址代替其域名。但是，有些却不行，有些网站采用了虚拟主机技术，同一个 IP 地址，可能对应许多不同的站点，因此，不能用 IP 地址替代其域名。对于，以上的情况还需加以研究，以寻求好的解决方案。

②本搜索引擎，目前仅考虑了搜索 HTML 网页的情况。在许多专业站点中，一些专业文档还可能不是通过 HTML 网页的形式给出，而是通过象 OFFICE 文档、PDF 文档等非 HTML 文档形式给出的，因此，还需研究常用的非 HTML 文档格式的编程接口，以使专业搜索引擎有更强的搜集信息的能力。

（2）中文分词模块

中文分词的准确程度，直接影响了搜索引擎的精度。分词的研究重点还是在解决歧义、识别新词方面。

①在解决歧义方面，除了应用正、逆向匹配法，或两者相结合的方法之外，还应与基于概率的分词方法结合。基于概率的方法，主要是考虑利用词语在文档中的搭配概率来解决歧义问题。

②在新词识别方面，主要还需研究新词的自动识别。由于一些专业知识发展速度比较快，一些新的专业词汇产生，原有词典中没有相应的条目，可是，这些专业词汇又对文章的主题起着很重要的作用。一种想法是，可对用户的实际搜索请求建立日志，当大量的用户都有针对某一词汇的搜索请求，而词典中又没有该词的情况下，可将此词汇作为新词添加到词典中。

（3）信息检索模块

信息检索模块主要还是要从提高检索速度方面加以研究。

①现有系统，由于搜索的网站规模不大，搜索的信息量相对较少，所以，其信息库、索引文件都是基于数据库实现的。但是，随着规模的扩大，还应考虑建立专门针对搜索引擎的索引文件系统，研究相应的检索算法。另外，还可考虑建立分布式系统，提高并行处理能力。

②系统中还应考虑利用自动分类技术提高检索速度。自动分类技术，主要根据网页文档内容的主要特征，将网页自动划分到各个类别中去。这样，当用户提出某一关键词的搜索请求后，系统首先可根据关键词的类别，检索属于该类别的网页的集合，有效地提高检索速度。

# 结　论

本文设计并实现的渔业信息搜索引擎系统来源于大连市科技计划项目《基于 WEB 的智能渔业信息检索系统》。该搜索引擎主要面对水产渔业领域的专业技术人员。

本文讨论了水产专业技术人员使用传统搜索引擎搜索信息面临的问题：不能对专业信息很好的解读，查不到，查不准，垃圾信息过多。本文系统的完整的设计了一个应用于水产渔业领域的中文专业搜索引擎的设计与实现过程。研究探索了适用于专业搜索引擎的网络爬行技术、中文分词技术、信息检索技术、网页的排序及消重等技术。该引擎的开发，采用了面向对象的设计方法，开发语言选用 Microsoft 的 Visual C++ 6.0，数据库管理系统采用 SQL Server 2000。本文所述的搜索引擎，是专门针对水产行业的搜索引擎，其搜索结果中，不含有与水产渔业无关的信息，相对于通用搜索引擎要精确的多。而相对于一些站内的专业搜索引擎，本系统的搜索引擎可由用户自由添加起始网址，因此搜索的范围要大的多。

本文的渔业信息搜索引擎属专业搜索引擎。本文给出了专业搜索引擎的通用设计模式，本文的搜索引擎具有广泛的适用范围。

# 参 考 文 献

[1] 张士青，童卫东. Internet 基础. 现代通信，2002，6：15.
[2] 杜元清. 网络信息检索系统 WWW. 情报理论与实践，1995，6：43-45.
[3] 张利，邵世煌，吴晓琼等. Current application of search engines and their developing trend. Journal of Dong Hua University，2002,19(2)：126-130.
[4] Duan Lian, Zhang Haoliang, Wu Weiling. The user-level security of mobile communication systems. The Journal of China Universities of Posts and Telecommunications, 2002, 9(3): 62-68.
[5] Porter Michael. Gopher brings order to the amorphous mass that is the internet. Personal Engineering and Instrumentation News, 1995, 12(3): 66-68.
[6] Kotlowitz R W, Taylor L R. Compliance metrics for the inclined gull-wing, spider J-bend, and spider gull-wing lead designs for surface mount components. Transactions on Components, Hybrids and Manufacturing Technology, 1991, 14(4): 771-779.
[7] Chen Hsinchun, Chung Yiming, Ramsey Marshall, Yang C C. Smart itsy-bitsy spider for the Web. Journal of the American Society for Information Science, 1998, 49(7): 604-618.
[8] DeNero K A. Netscape quick tour for Macintosh. Electronic Library, 1995, 13(6): 582.
[9] Byrne, J A. The virtual corporation. Business Week, 1993, 2:98-102.
[10] 郭飞跃. 搜索引擎是如何工作的. 现代电子技术，1999（4）：33-35.
[11] Martin Philippe, Eklund P W. Knowledge retrieval and the www. Intelligent Systems and Their Applications, 2000, 15(3): 18-25.
[12] Ozmutlu Seda, Spink Amaanda, Ozmutlu H C. Multimedia Web searching trends: 1997-2001, 2003, 39(4): 611-621.
[13] Aone C, Bennett S W, Gorlinsky J. Multi-media fusion through application of machine learning and NLP. Proceedings of the AAAI Symposium on Machine Learning in Information Access, (USA), 1996.
[14] Chen H, Lynch K J. Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man and Cybernetics, 1992, 22(2): 885-902.
[15] 周树威. 中文搜索引擎分类体系研究与实施:(硕士学位论文). 北京：中国科学技术信息研究所，2002.
[16] 康桂英. 搜索引擎 Altavista 研究. 情报杂志，2000，19(3):34-35.
[17] 阮延生. Infoseek 搜索引擎的研究. 情报探索，2000，4：30-31.
[18] 朱蓓. 简评 WebCrwaler 与 MetaCrawler. 情报杂志，1999，18（5）：27-28.
[19] 邵莹. Lycos:可怕的狼蜘蛛. 电子商务，2000（9）：38-39.
[20] 陆建平. Excite 搜索引擎. 电脑技术，1998（8）：20-21.
[21] 朱俊卿. 搜索引擎 Google 研究. 广州大学学报（社会科学版），2001，15（11）：7-8.

[22] 马静. Google 的搜索机理和搜索技巧. 图书情报工作，2001，9：69-70.
[23] 张元馨. 基于用户档案的个性化搜索引擎的研究与实现：(硕士学位论文). 西安：西安交通大学硕士学位论文，2001.
[24] 孙丽，陈通宝，乔晓东. 网上中文检索工具的比较研究. 情报学报，1999，6：225-227.
[25] 徐颢. 浅谈百度中文搜索引擎的应用. 中华医学图书情报杂志，2003，12（5）：54-55.
[26] 芦新华. 解读网易搜索引擎的分类体系. 河南图书馆学刊，2001，9：62.
[27] 郝力. 中国最大的中文 IT 专业网站“赛迪网”开通. 中国信息导报，2000，4：14.
[28] 雷鸣，王建勇等. 第三代搜索引擎与天网二期. 北京大学学报，2001，5：734-736.
[29] 王红松，唐芬，马朝智. 网络信息检索的途径及方法. 江汉石油学院学报，2002，3：69-70.
[30] Steve Lawrence，Giles C L. Searching the World Wide Web. Science，1998，280：98-100.
[31] 韩义. 企业管理信息系统的可行性分析研究. 天津师大学报(自然科学版)，2000，6（20）：29-32.
[32] 林勇，宋征等. Visual C++6.0 应用指南. 北京：人民邮电出版社，1999.
[33] 李晓明，闫宏飞，王继民. 搜索引擎——原理、技术与系统. 北京：科学出版社，2005.
[34] 瞿锋，陈纪元. 汉语自动分词算法综述. 福建电脑，2006，4：23-25.

# 致　　谢

本论文是在导师马洪连副教授的悉心指导下完成的。在此，谨对导师的辛勤培养和关怀表示衷心的感谢。

我同时还要衷心感谢本项目的负责人于红教授对论文进行了评阅，并提出了许多宝贵的意见。

感谢同项目组的合作伙伴史鹏辉老师，在整个系统设计阶段，给了我很大的帮助。