摘要

随着互联网信息技术的迅速发展，大约每天就有2000万的新网页诞生，且很多网页的信息熵不高，导致通用搜索引擎的信息覆盖率和检索精度都在迅速下降，因此发展面向专题的信息搜索与收集工具就成为趋势。

本文研究的面向专题的Web信息收集与过滤技术是这类工具的核心，围绕专题信息的特点，设计了一套个性化的专题信息查询表现、专题信息查询和专题信息过滤和收集方法，以适应专业用户信息收集的需要，提高信息收集的相关度和收集效率。文中重点对查询请求的提交和表现，以及信息的搜索与过滤策略做了研究。

本文以国家863高技术项目现实需求为背景，主要工作和研究内容如下：

（1） 个性化的查询表现技术，使用户不再为向系统提供简单有限的特征词而烦恼，可以帮助用户更好的表达个人的兴趣专题，定制专题知识库。同时，在信息过滤过程中，专题知识库也为文本过滤提供了匹配基向量。

（2） 提出基于专题的扩展查询技术，克服了有限个专题特征词的限制，使信息搜索过程中，查询关键词与文本特征不拘泥于形式上的“强吻合”，解决了搜索中存在的“漏搜”现象，最大限度的提高了搜索的查全率。

（3） 研究了检索中基于深度控制的链接过滤方法和下载控制方法，辅助新的相关信息资源的发现，使用户可以对下载的信息量进行间接的控制，搜集到相关度更大的专题信息。

（4） 改进了基于内容的文本过滤和基于链接结构相结合的过滤策略，有效的改善了检索中的主题“漂移”和搜索精度不高问题，提高了信息的下载精度。

关键词：专题信息；查询表现；查询扩展；基于内容；基于链接；过滤

ABSTRACT

With the rapid development of information on the Internet, around 20,000,000 web pages appeared, and lots of them have little entropy of intelligence. General search becomes weaker and weaker. The precision ratio and recall ratio have been declining that some special users have to look for a new method to fit their needs. It has become a tendency to develop specific Internet information retrieval tools.

The technique of topic-focused web information searching and filtering described in this paper. According to the personality of topic-focused information, we present a new Topic-focused method including information query representation, information retrieval and information filtering. The core technique is how to represent the query of information, how to search the relevant information and how to filter useful information.

Referring to national 863 high-tech Project, the main content in this paper was introduce as follows:

1. Representing personal information query, which make user no more fretful about useful keywords. Users can provide keywords more easily, customize specific topic and retrieval vector, which match the information user is looking for.
2. Introduced Topic-based query topology, which broken out the quantitative limitation of keywords and "word-mismatch". The method make the loss of the relevant information decrease and make the recall ratio go up.
3. Controlling of the Anchor-depth can make the system work more efficiently not search anchor greedily. It make the system work in high crawl efficiency
4. Improving on combining content-based and anchor-weight-based filter algorithm, which resolve the problem of "topic-drift". It works well on popular PC and achieves high topic-focused information recall ratio.

Keywords: Topic-oriented; query presentation; query expansion; content-based; anchor-based; filter

图表目录

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科学技术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谢意。

学位论文题目：面向专题的信息搜索与过滤技术研究

学位论文作者签名：[signature]　　日期：05年11月29日

学位论文版权使用授权书

本人完全了解国防科学技术大学有关保留、使用学位论文的规定。本人授权国防科学技术大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密学位论文在解密后适用本授权书。）

学位论文题目：面向专题的信息搜索与过滤技术研究

学位论文作者签名：[signature]　　日期：05年11月29日

作者指导教师签名：[signature]　　日期：05年11月30日

第一章 绪论

§1.1 选题的依据及课题来源

随着 Internet 的飞速发展，上网用户的数量呈指数型增长。根据 Google 主页发布的消息，截至 2004 年 2 月 27 日，全球网页约 43 亿，且以每天 2000 万的速度更新。在这样庞大的网页信息资源的使用过程中，大约有 85%的用户使用搜索系统去检索他们需要的信息[1]。但是，网络信息资源的指数增长，使搜索系统对网络信息的覆盖率在整体上呈急剧下降趋势，越来越难以应对用户的需要，就是号称功能最强大的搜索系统，也无法跟上网络信息的增长速度。尽管当前有些比较成熟的搜索系统，但是要准确、快速地查找所需信息却越来越困难。

（1） 由于自然语言的歧义性，一个关键字可以出现在多个不同的信息领域中，并有不同的意义，基于这个关键字的检索就会将所有涉及这个关键字的页面均返回给用户，检索精度低。一次搜索的结果可能有成千上万条，而在这过于庞大的信息群中，有用信息只是其中的小部分，检索结果中出现大量冗余信息。

（2） 目前的搜索系统都是服务器端软件，用户需要严格按照各搜索系统所要求的格式输入查询词，种种限制使用户不知道如何确切地表达自己的信息需求，也不知道如何更准确地寻找所需信息。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[13]。

（3） Web 信息资源的动态变化，搜索引擎无法保证对信息的及时更新；传统的搜索引擎不能满足人们对个性化信息检索服务的日益增长的需要。网络信息的急剧膨胀，使检索越来越难以控制，用户需求和服务间的巨大反差产生了强大的检索障碍。

（4） 搜索引擎一般不会遗漏较重要的网站，但由于对网站的描述较为简单，不能深入到网站的内部标引。对各个站点的标引仅是以宽度优先，而深度不足，常常导致相关的信息被“漏搜”，如著名的搜索工具 Google 只能对同一站点的三层以内的链接做出索引。

面对这些挑战，各类适应特定人群需要的“专题搜索工具”（Topic-Specific Search）应运而生并引起了研究者的重视。以何种策略访问 Web，以提高搜索效率，成为近年来专业搜索引擎研究的主要问题之一。垂直（专业）搜索引擎这种高度目标化、专业化的搜索引擎的优势在于，针对性强，对特定范围的网络信息的覆盖率相对较高，具有信息资源保障，有明确的检索目标定位，有效地弥补了综合性搜索引擎对专门领域及特定主题信息覆盖率过低的问题。同时，能够把具有相同兴趣点的人们集中在一个“主题社区”内，集中提供各种专业资源。

§1.2 国内外研究现状

在这样一个信息高速膨胀时代里，信息的价值是不言而喻的。因此越来越多的人

把目光转向了互联网这样一个信息资源丰富的宝库，自然地，互联网信息搜索工具就成为了研究的一个焦点。

当前对搜索工具的研究主要围绕以下三个方面进行：

一是在原有的体系上，加以改善，以更加适合用户的个性化需求；

二是检索的表现和提交技术，如何向用户提供一个友好的人机接口，最大限度的降低对用户的要求；

三是搜索工具所使用的搜索过滤技术上，使用更加完善高效的算法，力求高质量的完成用户的需求。

下面，本文就以这个三个方面简要的介绍一下国内外的研究现状。

1.2.1 搜索工具的体系结构研究现状

目前的互联网搜索工具基本上是客户端工具，采用浏览/服务器（B/S）结构。它的后端程序，如搜索机器人（Robot）采用不停的爬行的方法来收录并标引尽可能多的页面信息，对于本文中的情报人员来说是不必要的。搜索门户要建立自己庞大的索引数据库，或者使用元搜索引擎（Meta search engine）。其工作原理如图 1－1 所示。

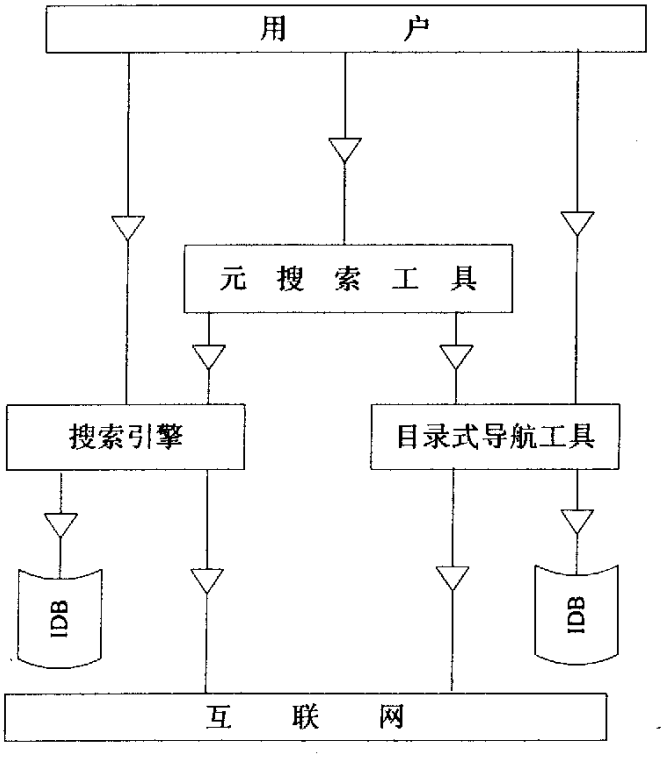

图 1-1　通用搜索工具的结构

国内著名的搜索引擎百度就是使用自己的爬行器，对互联网信息做标引，建立自主的标引数据库（IDB），当用户发出搜索请求时，搜索引擎在标引数据库的标引词中查找与用户的请求相关的信息，并将结果反馈给用户。刚刚由新浪开发的新一代搜索引擎爱问，虽然采用的仍然是这种结构，但是由于新浪可以利用自己作为一个信息

门户的优势，减小对信息源的依赖。

美国国家科学数字图书馆的 Collection Bilding Program(CBP)这个项目旨在为科学、数学、工程和技术创建大规模的在线数字图书馆，试图研究在某一主题上资源自动建设的可能性。CBP 具有自己的特点：第一，因为 CBP 是面向教育、面向教学，主题精确度(Precision)比覆盖度(Recall)更为重要;第二，CBP 不存储资源原文，而只是提供 url。第三，CBP 只需要用户最少量的输入，如关键词，系统就可以全自动的将有关该主题的最相关的有限数量 URL 返回给用户。

从检索性能来看，有些功能比较受限。对于目录式导航工具，如雅虎，如果分类不够科学合理，或者工具的分类方法与用户理解的分类有差别，那么，按分类目录浏览往往检索到很多无关的信息，易出现漏检和误检的情况。

无论采用搜索的机制，还是采用建立独立的标引数据库的机制，亦或是目录导航式，这三种方案都会对数据库服务器产生强大的依赖，没有索引数据库，搜索工具根本无法工作，无法完成搜索或者索引任务。通用搜索工具的良好搜索服务依赖于后台程序对网页的标引能力，它的目标是搜索机器人能够标引所有的网页，而并不关心网页的领域和相关度问题。因此，本文中针对情报人员的特殊需要，开发一个自主性更强的面向专题的单机版的搜索工具，搜索用户兴趣相关的专题信息并下载，该系统首先对链接及信息的相关度做出评价，优先提取相关度大的信息下载，而不是使用通用搜索工具的后台程序标引所有信息的机制。对专业用户来说，搜索工具运行的目的性更强，避免了盲目的提取链接，可以节约大量的系统资源。

1.2.2 搜索的人机交互技术研究现状

对于一个搜索工具而言，良好的服务界面和优秀的辅助用户搜索请求提交功能是重要的。当前，国内外在这方面做的工作不少，但是多数搜索工具都是客户端工具，因此，很难做到个性化。一些比较成熟的搜索工具所使用的方法也多半是局限在检索词的逻辑组合上，下面就对目前国内外在检索表现的研究状况作以简单介绍：

著名的搜索引擎 YAHOO 在它的主菜单中提供了 14 个主要的分类，在每个主分类中还有一些子分类。每个子类提供大量的可供检索的细小条目。用户可以通过关键词进行检索。关键词之间可用布尔运算符 AND 或 OR 连接，但是一旦选用了一个逻辑运算符，它就必须应用于所有的关键词。除了关键词的完全匹配之外，YAHOO 还提供关键词的部分匹配。

国外著名的分类搜索 INFOSEEK 将其搜索内容分成 12 个类，其中，每个类还有若干个子类。搜索在每个子类中进行。搜索内容为全文。用户通过关键词进行搜索。一次可输入若干个关键词。关键词之间用布尔运算符 AND 和 NOT 连接。输出结果按每个命中条目的得分排序。由于是在各个类中完成搜索，因此，可以有效地缩小查找的范围，减少查找的时间，提高搜索的效率。

搜索引擎 OPEN TEXT INDEX 虽然不提供分类，但提供二级关键词检索：简单检索和强化检索。简单检索中，用户可选择检索词或词组。强化检索可有 5 个关键词输入字段。对每个输入段，用户可选择搜索范围，如全文或标题等。用户也可用运算符列选择输入字段间的布尔运算符。可选的布尔运算符有 AND，OR，NOT，NEAR 或

FOLLOWED BY。运算符按出现次序起作用，而不是按运算符的优先级，检索结果按对应值排序。像 SQL 语言，那样复杂的查询语言在现有的搜索引擎中还不能应用。

北京优联克科技开发有限责任公司和香港优联克公司共同推出的智能型中文搜索引擎悠游(Go Yoyo)，可采用以下两种方法进行搜索：关键词搜索和分类方式在关键词窗口，输入检索关键词，单击“搜索(S)”框，可得到许多与关键词有关的相关网页。用户无需使用空格键把词分开，完全按照书写习惯输入检索要求即可。悠游(Go Yoyo)具有中文词自动切分功能，如输入关键词“氟污染”，将自动切分成有效关键词“氟”和“污染”进行搜索。除了用关键词进行搜索外，还可用分类方式进行搜索，对不习惯中文文字输入的用户尤为适用。在分类表中单击所需搜索类目，即可得到一系列相关网页。分类表中列有 12 个主项目：保健、财经、地区、电脑、教育、科技、社会、时事、文史、艺术、娱乐和政治，主项目下还列有各类分项。

现有 Web 搜索引擎在请求的提交上存在比较大的缺陷，大多数的关键词检索模式只配置了一个简单的检索框，或只提供关键词间最基本的布尔连接，逻辑运算符的组合比较呆板；要求与标引的内容间完成“强吻合”才能作为有用信息返回。提供的提问函数也是相当有限的，缺乏其他复杂高级的精确检索方式，不易于处理多词检索和限定词的检索；用户的个人需求表达比较受限，很难真实的表达用户的想法；限定载体类型、文档类型的检索比较困难，无法完全满足普通用户的要求。

1.2.3 搜索过滤技术的国内外研究现状

搜索过滤技术是一个搜索工具的核心技术，它的好坏，直接关系到一个搜索工具的性能，影响到查准率(precision ratio)、查到率(recall ratio)和响应的时间等重要评价指标。目前，国内外采用的搜索技术也比较多，比较好的技术主要有：

世界最大的搜索引擎 Google，使用的就是一种基于链接评价[1]的方法，来指导索引器的爬行，发现新的链接。具体实现的方法是：对当前的网页 A，假设指向它的网页有 T_1，T_2，…，T_n，则有：

$$PR(A)=(1-d)\frac{1}{T}+d(\frac{PR(T_1)}{C(A)}+\frac{PR(T_2)}{C(A)}+...+\frac{PR(T_n)}{C(A)}) \qquad (1-1)$$

其中，$C(A)$ 为从页面链出至其它页面的链接数目，$PR(A)$ 为页面 A 的 PageRank 值，T 为计算中的页面总量，d 为衰减因子（通常设为 0.85）。

从这个公式中可以清楚的看到，对链接的评价计算方法比较复杂，而且对收集的信息的地址要求也比较多，计算复杂度随访问页面和链接数量的增长而迅速增长，要求有强大的硬件资源支持，因此，不适应专业用户的使用需求。

Focus Project 系统[2]是由印度裔科学家 S.Charkrabarti 带头从事，他是最早从事这方面研究的人之一。该系统通过二个程序来指导爬行：一个是分类器 Classifier，用来计算下载文档与预定主题的相关度。另一个程序是净化器 Distiler，使用的是 HITS 算法。基于 HITS 方法的网络蜘蛛对每个已访问的页面计算其 Authority 权重和 Hub 权重，并以此决定页面中链接的访问顺序，设页面 P 的 Authority 权重和 Hub 权重分别为 $A[p]$ 和 $H[p]$，则按下列迭代公式计算：

$$A[p]=\sum_{q\in E}H[q] \qquad (1-2)$$

$$H[p]=\sum_{q\in F}A[q] \qquad (1-3)$$

其中，E 为所有指向页面 P 的页面集合，F 为被页面 P 中的链接指向的页面集合。

这类搜索策略的优点是考虑了链接的结构特征，但也存在一些缺陷，如忽略了页面与主题的相关性。不考虑语义的相关性，在某些情况下，会出现搜索偏离主题的“主题漂移”问题，影响检索的精度。

国内有关 Web 信息收集过滤的系统主要是四川联合大学计算机系开发的 WebMiner。WebMiner 系统进行挖掘的对象是 Web 上指定的某类事物，通过对该类事物领域知识和格式的学习，从中提取有用的数据放到数据仓库中，然后在数据仓库中进行挖掘，以期获得该事物的发展趋势和规律。该系统采用 agent 技术，把多维文本分析和文本挖掘这两种技术有机地结合起来，能够帮助用户快速、有效地对 WWW 上的 HTML 文档进行挖掘工作，实现了从语义的层次上，搜索相关的信息资源。

§1.3 本文的主要研究内容

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够使用数据挖掘技术，对其进行更高层次的分析，以便更好地利用这些数据，并将之应用在军事决策支持系统中，提高军事指挥、部队管理的质量和效能。

本文的研究内容就是在这种背景下，设计开发一个实际、可靠的原型系统，用于互联网上专题信息的搜集与过滤，为数据挖掘提供数据资源。

本文的主要研究内容包括：

（1） 研究了智能的查询请求表现技术，专题知识库的定制，如何简化专题信息查询表现，辅助用户请求的提交，使用户能够更好的表达个人的兴趣专题。

（2） 研究了检索向量的构成，强化检索向量的特征，提高检索中的排他性。如何在提高查全率和查准率的基础上，保证下载信息与专题的相关度。

（3） 研究了检索中基于深度控制的链接过滤方法和下载控制方法，使用户可以对下载的信息量进行间接的控制，提高收集信息的相关度。

（4） 重点研究了基于内容的文本过滤和基于链接结构相结合的过滤策略，改善了检索中的主题“漂移”和搜索精度不高问题，提高信息的下载精度和信息的收集效率。

（5） 设计并开发了一个原型系统用于信息的收集与过滤，该系统采用 C/S 结构，适应更多用户的个性化需求，具有更大的灵活性。用户可以按照个人的兴趣定制专题，也可以根据个人的需要对系统进行设置。

§1.4 论文的结构

本文共包括六章，后续章节基本上是按照上述研究内容逐一展开的。第二章介绍面向专题的信息搜索与过滤理论体系、系统体系结构和关键技术；第三章介绍检索请求的表现和提交技术，重点对基于专题(Topic)的查询扩展做了详细介绍；第四章介绍

了基于内容的文本过滤和基于链接分析相结合的过滤方法，同时研究了链接深度控制技术，从而实现了信息收集的精度和高效；第五章介绍了开发的信息收集与过滤系统的使用方法和工作流程，并对实验结果作了分析比较；最后一章总结和评价了论文的全部工作，并对未来需要研究的工作进行了展望。

第二章　信息搜索与过滤的关键技术及系统的体系结构研究

§2.1　信息搜索与过滤的相关问题

2.1.1　搜索过滤技术的分类

目前，国内外应用的搜索策略种类繁多，但是对各种搜索系统进行分析后，可以发现，其具体的搜索及过滤方式大致可分为以下四种类型：

(1)基于网页内容的搜索过滤技术

搜索“机器人”采集到 Web 空间中的网页后，系统根据网页内容经文本分析处理后建立索引库；当用户提出搜索条件时，网页的检索器会自动从索引库中查询有关的内容。基于网页内容的搜索技术直接使用网页的文本建立索引库，所以索引库的容量很大。而且从网页内容中判断是否符合查询条件比较困难，往往搜索出的网页很多，但真正有用的可能并没有多少。另外由于网页内容越来越丰富，不仅有文字内容还有各种多媒体数据，所以如何有效地对网页内容进行文本分析处理也是比较困难的。这类搜索技术是当前投入使用中的主流。

(2)基于目录的搜索过滤技术

在基于目录的搜索技术中，信息收集与索引主要依靠人工来完成，搜索系统的标引专家依靠手工来搜寻不断出现的新的网站，给每个网站一个标题和主题描述，并将其放入相应的类目体系中，在页面上表现为每个类目路径下排列着相关的网站。典型的基于目录的搜索系统有 Yahoo。但这类搜索系统有两大问题：①分类是按分类者或分类软件的分析而定，不一定与用户的意见一致；②如果你查找的信息没有对应的分类项，则可能无法继续进行搜索。

(3)元搜索过滤技术

元搜索技术将用户提交的搜索请求转换处理后，提交给多个预先选定的独立搜索系统，并将从各独立搜索系统返回的所有查询结果集中起来，过滤掉重复和相关度低的信息后，对来自多个引擎的结果进行排序后，再返回给用户。它主要考虑用不同的方法过滤从其他搜索系统接收到的相关文档，包括消除重复信息等。元搜索系统设计简单，但网络的负载太大且搜索效果始终不理想，所以没有哪个元搜索系统有过强势地位。典型的元搜索系统有 MetaCrawler 等。

(4)分布式搜索过滤技术

分布式搜索系统按区域、主题或其他标准创建分布式索引服务器。索引服务器相互之间可以交换中间信息，且查询可以被重新定向。如果一个检索服务器没有满足查询请求的信息，它可以将查询请求发送到具有相应信息的检索服务器。分布式搜索系统将索引数据库划分到几个分布的数据库中，每个数据库变得小一些，但所有搜索系统覆盖的范围变大，且很少有信息重复。作为分布式系统特性之一的可扩充性也是分布式搜索系统的优点之一。然而分布式搜索系统需要多个索引数据库协同工作，实现

较困难，目前尚未有真正的、实用的分布式搜索系统。

2.1.2 基于关键词的查询简介

现有的搜索引擎多采用全文检索技术，其核心是关键字符的机械式匹配。这种方式的固有缺点是参与匹配的只有字符的外在表现形式，而非它们所表达的概念。因此，经常出现所答非所问、检索不全的现象。

关键字匹配的方法是最基本也最常用的方法，几乎每一个搜索引擎都采用了这一方法。用户提交关键字，如果某网页中出现该关键字的频度较高，就将这一网页列进搜索结果。用户提交的关键字中还可以包括 and，or，not 等布尔检索来精确定位。

这种方法面临的问题比较突出的有：

（1） 查询结果完全依赖于用户所给出的关键字，系统和用户之间仅有单纯的请求一回答式交互，不能辅助、引导用户达到目的，造成检索效果比较差。

（2） 采用的索引的可能会与用户的检索词存在比较大的分歧，这时用户的检索要求就不可能会得到及时、充分的满足；

（3） 用户很难简单地用关键词及其及组合来真实地、准确地表达需求信息内容，导致检索困难；

（4） 搜索的结果不精确，常常输入一个或一组查询词，能返回数百篇结果，与用户的需求相差比较远；

（5） 不能根据户的请求来挖掘深结构的信息满足用户的最迫切的需求。

现阶段，在搜索用户所指定的关键词时，所采用的信息检索技术要么基于某种编码过程先对给定的术语进行预处理，要么先执行某种全文分析。这些方法一般只能反映用户所要检索内容的某一方面，无法保证内容的准确匹配。

扩展查询试图突破关键词匹配局限于表面形式的缺陷，从特征词所表达的专题领域的层次上来认识和处理用户的检索请求，从而优化了用户的检索请求，进而提高了查询的查全率。而用来对特定领域的概念及术语给予明确的形式化描述的专题(Topic)不仅为规范化资源描述及用户查询提供了基础，也为更准确地搜索信息提供了保证。

2.1.3 主题页面在 Web 上的分布特征

整个 Web 上的页面主题分布是混杂的，但同一个主题在 Web 上分布却有一些规律。我们将这些分布规律总结为四个特性：Hub 特性[17]、Sibling/Linkage Locality 特性、站点主题特性、Tunnel 特性。

(1)Hub 特性

美国康奈尔大学的教授 Jon M.Kleinberg 发现 Web 上存在大量的 Hub 页面，这种页面不但含有许多 outlink 链接(指出链接)，并且这些链接趋向于相关的主题。也就是说，Hub 页面是指向相关主题页面的一个中心。另外，他还定义了权威页面(authority)的概念，即许多其它页面都认同相关于这一主题有价值的页面。好的 Hub 页面一般指向多个 Authority 的页面，并且所指向的 Authority 页面越权威 Hub 页面的质量也越好；反过来，Hub 页面的质量越好，它所指向的每个页面也趋向于越权威。我们把主题在

Web 上的这一特性称为 Hub 特性。

(2)Sibling/Linkage Locality 特性

在 Hub 特性的基础上，人们又提出了 Sibling/Linkage Locality 特性。Linkage Locality，即页面趋向于拥有链接到它的页面主题的页面；Sibling Locality，对于链接到某主题页面的页面，它所链接到的其它页面也趋向于拥有这个主题。这实际上是 Hub 特性的变形，主要是从页面的设计者设计的角度考虑的。一个页面的设计者趋向于把本页面指向于与本页面相关的其他页面。我们把主题在 Web 上的这一特性称为 Sibling/Linkage Locality 特性。

(3)站点主题特性

我们发现，一个站点趋向于说明一个或几个主题，并且那些说明每个主题的页面较紧密地在此站点内部链接成块，而各个主题块之间链接较少。我们认为，这主要与网站的设计者的设计思路有关。每个网站在设计时都有目标，而这种目标往往就集中在一个或几个主题中。而网站的浏览者往往也有一定的目的性，这个目的性一般体现在用户趋向于浏览同一主题的页面。为了满足浏览者的这一需求，网站设计者需要将相关内容紧密地链接在一起。为了研究主题块的特性，我们设计了实验：首先将站点内的链接分为六类(下行链、上行链、水平链、交叉链、外向链、框架链)、站点内的页面分为四类(主页、索引页面、内容页面、参考页面)，并为每一类链接和页面赋予不同的权重，然后通过向量空间模型算法为每个页面分类，并在站点内部结构特征的基础上，对站点页面树按照自底向上进行主题聚类。试验结果证明了站点中存在着许多主题页面团。

(4)Tunnel 特性

在 Web 中还有一类现象，就是主题页面团之间往往需要经过较多的无关链接才能相互到达。这些无关链接就像一个长长的隧道，连接着两个主题团，因此我们把这种现象称为“隧道现象”(Tunnel)。在基于主题的页面采集过程中，Tunnel 的存在极大地影响着采集的质量。为了提高采集页面的准确率，我们需要提高过滤相关性判定阈值，而阈值的提高将过滤掉大量的 Tunnel，使得采集系统很可能丢失 Tunnel 另一端的主题团，进而影响了查全率(或者说资源发现率)。反过来，为了提高查全率，就得大量发现 Tunnel，降低过滤相关性判定阈值，但是阈值的降低使得混进了大量的无关页面，从而大大降低了页面的准确率。这是一个两难问题，但关键还是不能有效地区别 Tunnel 和其它大量无关页面。事实上，两个主题间的隧道数也较少。

§2.2 系统的关键技术

为了设计并实现一个优秀的信息搜索与过滤工具，不但要具有良好的人机交互接口，同时，还要在系统的搜索与过滤技术方面做深入的研究，因此，本文所研究的关键技术有以下几点：

1、查询(search)表现技术

在搜索工具中，用户需求的真实表达，是搜索工具正常工作的前提条件。而据统计，大约有 70%的用户不能真实表达个人的需求，或者是不能用简单有效的短语表达个人的需求。因此，检索请求的表现成为搜索工具的一个研究热点。当前的搜索工具大部分采用基于关键词的检索，常常导致用户提交的请求与搜索工具的标引不能够很

好的匹配，因此，本文对检索请示的表现做了一定的研究，采用基于向量的检索请求提交方法。由用户根据个人的兴趣定制专题关键词，并最终生成检索向量。向量的定制可以采用用户自定义方式和文档辅助定义两种方式。

2、基于专题(Topic-based)的概念扩展查询技术

现有的搜索引擎多采用全文检索技术，其核心是关键字符的机械式匹配。这种方式的固有缺点是参与匹配的只有字符的外在表现形式，而非它们所表达的概念。因此，经常出现答非所问、检索不全的结果。另一方面，查询结果完全依赖于用户所给出的关键字，系统和用户之间并无进一步的交互，也是造成检索效果比较差的原因之一。基于专题的概念扩展查询试图突破关键词匹配局限于表面形式的缺陷，从词所表达的专题领域的层次上来认识和处理检索用户请求，从而提高查全率。

3、基于内容的二重过滤技术

一次过滤是过滤掉页面中的标识符，提取出文本实体，同时，充分利用了页面的结构特征，对结构特征做加权，这是进行后续的文本信息提取的前提。这一任务主要是在文本解析器中完成。二次过滤主要是通过文档的知识提取功能，获得页面的信息特征，完成分类和主题相关度的计算，标记相关信息的来源地址的重要度。

4、基于内容的文本过滤与基于链接评价相结合的过滤技术

大多数的搜索工具单纯的采用基于宽度或深度的新链接发现机制，但是这种方法由于缺乏目的性，盲目地提取链接，因此标引工作量大，无论是对网络和硬件系统都会造成大量的资源浪费。因此，本文采用基于内容的文本过滤和基于链接结构过滤相结合的方法，基于内容的文本过滤作为链接评价的依据，既避免了单纯依靠链接过滤的盲目性，又避免了单纯依据内容过滤的“近视性”，提高了信息收集的精度和效率。

5、搜索链接深度的控制技术

通常情况下，链接之间的引用具有一定的相关性。同级链接的邻居间具有近似性，不同级链接间具有遗传特性，可以利用这个特性发现新的信息资源。但是这种相关关系，不可能会无限的、全部的继承下去，而是部分的继承，甚至在继承过程中会出现变异，衍生出新的主题。为防止链接的提取过程中，偏离定制的主题，产生“漂移”现象，系统采用深度控制技术。

§2.3 专题信息收集与过滤系统的体系结构

面向专题的信息搜索与过滤主要是通过面向专题的扩展知识库方法，采用基于相关专题信息内容和来源链接评价的过滤技术来实现。面向专题领域的知识库扩展可以强化专题信息的表述，提高查全率。链接的评价也是基于专题的相关性计算的，有利于提高链接提取的目的性和准确性。

专题信息收集与过滤系统是采用上述的技术，为数据挖掘提供数据准备的专业工具，其主要功能是从公共信息平台上，如互联网，收集用户根据个人需要定制的专题信息。文中提出了专题信息收集与过滤系统体系结构，如图 2-1 所示。

1、人机交互层

在本模块中，实现友好易用，更具有亲和力的交互式界面，最大限度地辅助用户提交检索请求。针对高级和初级用户，量身定做了两种机制辅助提交用户的检索请求。初级用户可以直接使用系统预置的 URLs 种子集，完成信息的采集和过滤；高级用户

则可以使用系统的知识提取功能生成检索向量，并可以使用基于专题(Topic)的扩展查询关键词的方式，建立知识库，向系统表达个人的兴趣专题。

2、预处理层

信息采集是把整个 Internet 看成是一个有向图，Internet 上的页面看成是顶点，页面上的超链接（Anchor）是有向边，则网页链接的提取原则如下：首先定位至当前系统读入的页面，由页面解析器对页面进行初步的分析后，提取出可以嵌入的链接。其中可能包括新页面的地址和媒体信息，如文本文件、图像、音频、视频等等。在本级过滤中，主要是对页面解析器解析的文档，做无用信息的过滤，主要是过滤网页中的标识符，如<body>、</body>、<color>、<font>等等页面本身的标记；提取净化后的文本，并能对结构加以识别。

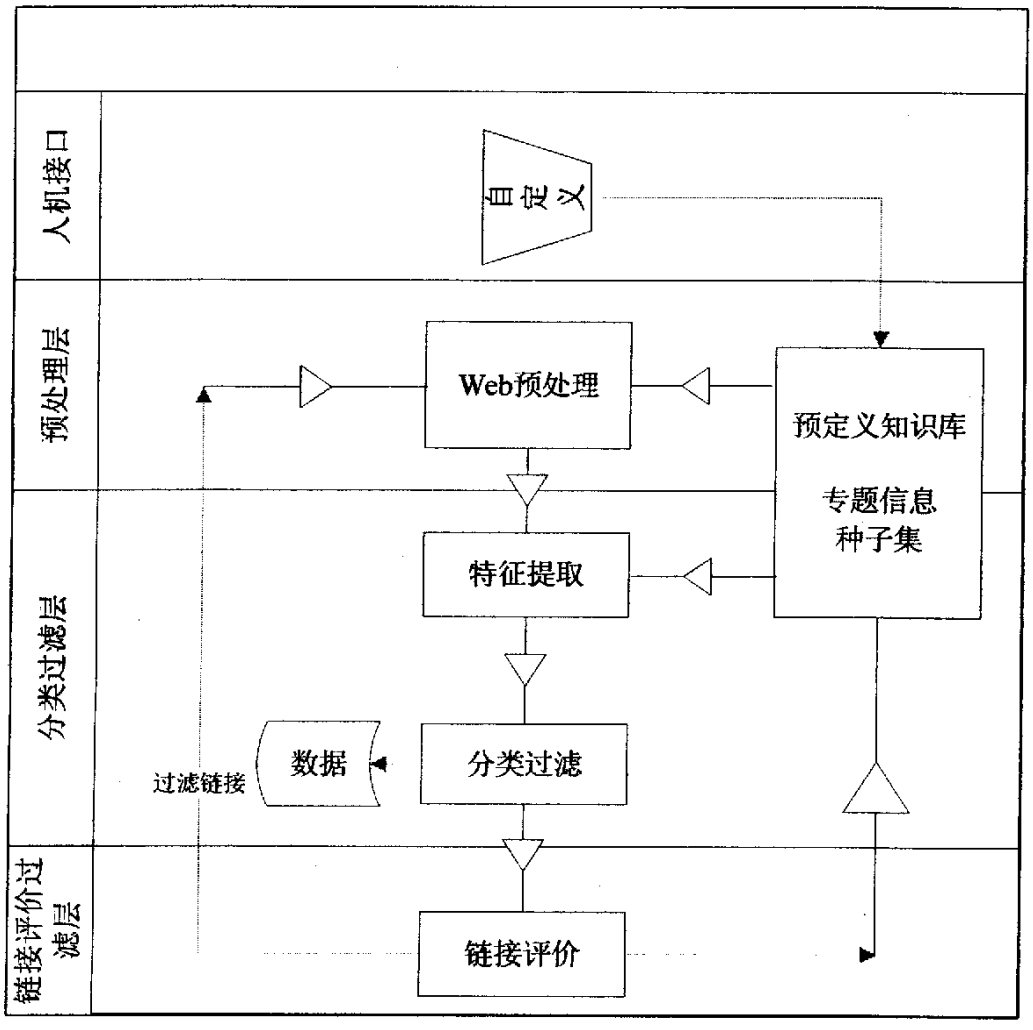

图 2-1　系统体系结构

3、分类处理与过滤层

对净化的有用信息，做进一步的过滤，分类，并以此作为链接评价的依据，如果

满足设定的专题相关性的阈值，则进入下一步。在文本的二次过滤中，主要是完成基于内容的文本过滤，滤出不相关的或与专题相关度小的信息，结束有用信息的提取分类，存入数据库，最后由开发的数据导入专用工具 Access2Sql 导入专业数据库 SQL server 中，为后续的信息挖掘和分析工作奠定了基础。

4、链接评价与过滤层

在链接评价与过滤层中，判断是否满足用户定义的探测深度；计算该链接的相关度评价值 Correlation，Authority (page)，Hub(page)值，最后归一化，计算出该页面的重要度 Importance(page)。对链接按其重要度排序过、过滤，提取出与专题相关度大的链接，有选择的下载数据，这样下载到本地的数据量就会有大幅度的减少，减小了网络的传输压力和系统资源的浪费，提高了下载的效率。

第三章　搜索系统的人机交互设计

当前，随着信息技术的发展，网络搜索技术也取得了较大的发展，尤其是综合搜索工具目前已经发展的比较成熟了。但是，如何使搜索工具更加“人性化”、“智能化”，降低搜索工具对使用者的要求，是当前搜索工具研究的一个热点。因此，搜索请求的表现技术就成为检索研究工作中的重中之重。自然，面向专题的搜索工具也应朝着智能化、个性化方向发展。简单易用的“个性化”界面应是易于被用户理解和应用，容错能力强、灵活性和适应性强，能够高效检索到合适信息的智能型用户界面。在信息的检索系统中，信息问题(information problem)、检索问题(search question)代表不同的意义。使用者将信息需求语言化，称为信息问题或搜索请求问题，将信息问题带到检索系统称为检索问题或查询问题。

因此，设计一个人性化的人机接口是当前这类工具普遍关心的问题，提供一个有效的解决搜索请求问题和查询问题的方法是提高检索系统工作效能的前提。本章就围绕这两方面的技术展开研究，重点研究了基于专题的扩展查询技术。

§3.1　搜索的人机交互相关因素

3.1.1　个性化的人机接口

所谓个性化就是信息接受者存在个性差异，包括个人兴趣、个人习惯的差异。一般情况下，某类信息一旦生成就是固定不变地放在那里等待信息获取者来取，而不考虑信息的使用者是谁，从而也就不用考虑信息受众的个性。个性化的知识推送系统只有获取信息受众的个性化特征才能进行个性化知识推送。目前个性化的获取主要分为两类：一是用户手工输入用户个性化特征，二是智能跟踪用户行为，自动获取用户的个性化特征。用户自建个性化模型用户在使用时首先将个人兴趣、知识侧重进行手工输入，通过用户的手工输入信息，为用户建立初级个性化模型，即将用户输入的主题词、主题站点等信息加入到用户词典和收藏夹中，对用户个性化模式数据库进行初始化。这种方式能让用户首次使用系统就可获得个性化的知识服务。

个性化服务模块根据用户模型向用户提供相应的服务策略和服务内容，它负责提供具体的个性化服务，如个性化推荐、个性化信息检索等。个性化服务模块将所有内容与用户模型进行匹配，将匹配的内容推荐给用户。

在匹配过程中，首先是目标文档与用户个性相关度的计算，目的是将文档按照相关度的大小排序后再推荐给用户。系统的文档信息过滤算法和相关度计算的算法是基于向量空间模型来完成的，向量空间模型法主要采用下述的模式进行过滤：用户提交自己的个性化描述请求，该请求能够最大程度地反映用户的兴趣爱好。其它需要过滤的文档与该请求进行比较，与该请求越相似的文档其获得推荐的机会就越大。这里的相关度指的是文档与用户个性的相关度。其次，对检索系统返回的结果进行个性化过

滤，就是把检索主题和它对应的相关主题一起提交给后台的信息检索系统。这个过程包括利用回溯对比算法去掉重复的检索结果，并按满足用户兴趣度的程度从大到小排序。当信息检索系统把检索结果送给用户浏览之前，要经过用户个性化模式的信息过滤，这样不仅过滤掉不相关的文档和重复文档，还可以把文档分成已经浏览过的文档和新文档，并按满足用户兴趣度的大小排序。经过查询过滤后的搜索结果按照相关度和用户兴趣度递减的顺序排列。如果经过过滤查询之后输出的网页数目大于用户要求输出的网页数目，那么用户可以重新调整相关性评估的匹配关键词以缩小范围、减少最终相关网页的输出。

用户是个性化服务的享用者，同时用户对个性化服务的反馈也可以用于调整个性化服务系统。如用户可以直接修改定制的用户模型，定制个人感兴趣的专题；可以定制收集信息的类型、搜索深度和信息的过滤选项，以调整个性化服务系统的性能；系统可以根据用户对个性化内容的选择改进用户建模模块和个性化服务模块的性能。每当用户需要此类个性化知识推送服务时，就可以得到自己个性化的知识推送服务，即使用户无法寻求一个确切描述个人兴趣爱好的关键词，仍然可以通过高级用户接口享受个性化服务，同时还可以享受的主动知识服务。因此，如何表现用户的检索请求是在信息收集与过滤中的一个至关重要的内容，构建具有文档分析与管理能力、信息代理和信息推送能力用户查询界面是研究的重点。

3.1.2 主题的相关性

一般而言，最常使用的主题相关定义是 Cuadra 和 Katter 在 1967 年所提出的定义，即为：相关是信息条件叙述(即输入系统之检索问题)和文章内容间之一致性(correspondence)。即文章所涵盖的内容对信息条件叙述的适合(appropriate)程度。

Saracevic 将 Cuadra 和 Katter 所提出的影响相关的变量重组为五类，分别为：文件及文件表征、检索问题、判断情境、判断尺度、以及判断者。其中影响最大的变量是判断者，而对判断者影响最大的变量则为其主题知识。一般而言，主题知识越丰富的人，所判断出的相关文献越少，而情境变量(包括：问题参与程度、文件预期用途等)则始终发挥其一定的影响力。

Saracevic 根据文献分析，提出相关概念的三大假设[23]：

(1) 只有信息需求者有资格作相关判断，因为相关是极为主观的判断；
(2) 对同一位判断者，其相关判断的结果会随着时间变化，因此个人认知的动态变化在相关判断中扮演相当重要的角色；
(3) 不同的判断情境会导致不同的相关判断结果，如信息需求者所处的环境及信息预期的使用目的等。

在本文中，我们也是基于以上的假设收集与专题相关的信息，为数据挖掘提供信息准备，控制收集信息的相关度。

3.1.3 搜索请求问题

在搜索工具的表现中，主要包括两个方面的内容：搜索请求的表现和检索结果的表现。前者，是检索正确高效执行的前提条件，直接影响到检索结果的质量和数量；

后者是对检索结果的表现，即以何种形式表现给用户，由于本系统是信息的搜索与过滤一体化的智能系统，所以搜索结果并没有以直观的形式表现给用户，但是搜索结果的表现，即如何从众多的链接中提取相关度更大的链接，仍然是一个极其重要的内容，而且在无人参与评判的状态下，检索结果的表现就更加重要了。因此，信息的表现技术在检索中是一个重要内容。

信息检索系统中，检索表现问题就转化为：将用户的兴趣转化为计算机语言，同时将文章内容以描述文章的索引词汇表示。完成这一任务采用的主要方法主要有两种：一是借助于计算机；二是使用人工提取。

利用计算机处理或者专业人员，将文献的内涵用各种方法表示，例如：题名、关键词、摘要、目次、书后索引、分类、编码等足以代表原文献的内容或主题；不论是文献本身、文献的结构、文献的其它形式、主题索引、自然词汇、主题内容、时间特征等，只要能将文献的特征表现出来的任何形式，都可以称作是文本表示法。信息组织依据信息结构，例如：作者、题名、出版者、时间属性等加以分析著录之外，也要分析信息的内容，称为主题索引(subject indexing)，两者均属于文本表示法。

文本表示法按其形成大致可以分为以下三种：

（1） 传统的分类编目、索引摘要等编制工作，都属于人脑和手工处理阶段。

（2） 计算机辅助文本表示法出现，例如：利用计算机编制关键词索引。

（3） 自动化摘要和自动化分类研究渐有进展，两者与自然语言处理研究关系密切。

§3.2 人机接口

为了更有效的表达用户的检索请求，简化用户操作，提供两种请求表现方法：自定义方法和知识提取方法。前者由用户直接定义专题特征词，用户具有更大的自由度。后者，用户只需要提供一篇参考文档，系统使用知识提取功能即可从中得到专题特征。

1、初级用户接口

在下载链接的初始化方面，系统已经为用户预置了 URL 种子集，用户无需提供新的链接地址就可以进入工作状态，解决了许多用户无从下手的烦恼；在专题的定制方面，用户可以提交文档或者是 Web 页面，使用知识提取来获得文档摘要，删除不能表达专题特征的词和虚词，写入知识库，操作简单，极大的方便了初级用户的使用。

2、高级用户接口

在下载链接的初始化方面，系统可以自定义添加一个或者多个新的链接作为初始种子集；在专题的定制方面，用户可以直接在专题知识库的定义模块中，添加专题和关键词，从用户提交的文档提取的关键词，赋予不同的权重，以突出相应主题。自定义功能使用户具有更大的自由度，可以根据个人兴趣的需要提供更加专业的信息收集站点，节省搜索时间，提高下载效率。用户也可以自由定义专题的表述，可以随意的修正关键词，最终生成匹配向量，实现与用户兴趣的无缝连接。

3、辅助功能设置

用户可以通过对系统的设置来决定是否下载动画、声音、视频和图片等多媒体信息。通过对网页源代码的解析处理，提取相关的媒体类型。同时，根据相关结构描述标记对多媒体信息分类。例如标记<IMG>的“ALT”属性说明。<IMG>定义一个内嵌

行内图像，<IMG>元素的“ALT"属性定义一个字符串，这个字符串能够反映图片的内容，对索引该页面提供了重要的文本信息，因此识别这个标记属性对信息的内容的表述有重要的辅助作用。我们可以根据这个属性和链接的后缀名来提取多媒体信息。如图 3-1 所示。

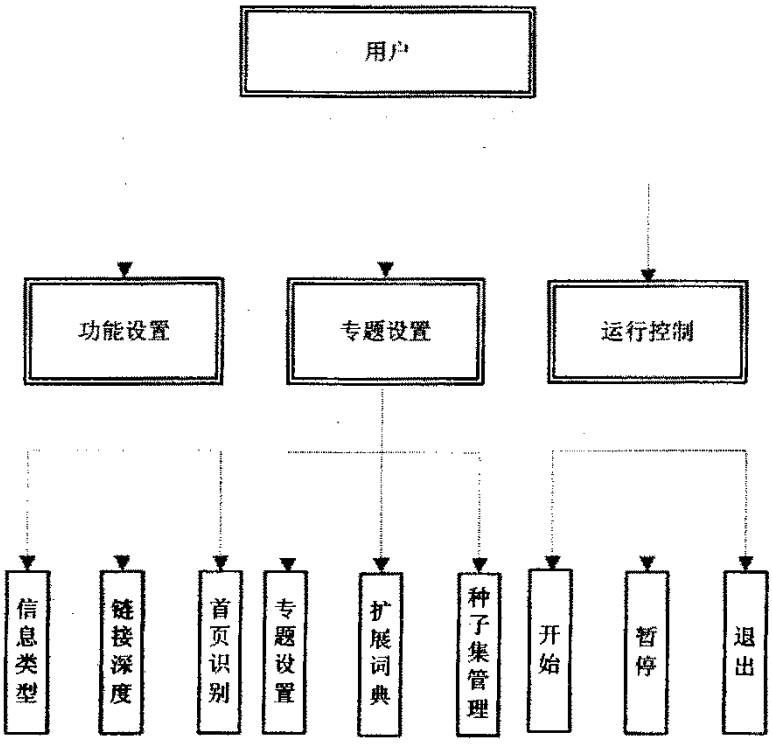

图 3-1　系统功能与设置

§3.3　搜索请求问题的实现技术

信息需求者的搜索问题如何转化成检索词汇，是许多信息检索国际会议和网络信息检索研究者重视的问题。Spink 和 Saracevic 研究检索词汇[21]来源，发现 61%的检索词汇来自检索者(38%来自于检索的书面，23%为来自于检索过程的反馈)；检索者提供的词汇可获得 68%的相关文献，是检索到相关文献的最高记录。

检索词汇与索引词汇间的关系，从 70 年代检索系统开始运作，研究者就相当关心其使用情况。到 1972 年 Bates 进行检索系统使用研究，结果发现使用主题标目检索完全相配情形仅为 20%-35%；五年后再次进行研究，发现相同结果，检索失败高达 65%-80%；80 年代几个研究结果也并无明显差异；90 年代，无论中、外的检索系统使用研究都和 70 年代研究结果相差不多。由此可见，20 余年来，信息检索技术改进有限。

搜索系统作为一个大众使用的工具，必须首先要被用户所接受，简单而实用且便于用户的操作是基本的要求，同时，还要能够满足任务的需要。因此，用户如何通过一个人性化的界面表达个人的检索请求是一个重要的、不容忽视的研究问题。

对于通用的搜索工具而言，如使用百度，Google 等搜索工具，主要是如何提交有

效的关键词。关键词的分析必须保证两点：

一是专业性，即这个词要求是很精练的，这样有可能极大地简化你的搜索结果，因而词甚至可以怪一点，尽量避免大众化的词，如“装备”，“精神”等。即使需要大众化的词，如“装备”和“精神”，但若加入一两个限定特征词，如加入“战斗”二字加以限制，辅助信息特征的表达，则结果将大大简化，提高查准率。

二是具有代表性，也即这个词具有代表意义，用户所需的网页内必须包含该关键词。由于现在的通用工具多半是以全文索引或者是结构性的标引，所以如果用户提供的关键词并不在标引词之内，就会造成部分有用信息的遗漏。

而对于本文研究的专题搜索及下载工具而言，则考虑因素要更复杂一些。由于该系统不是一个普通的搜索引擎，为适用专业的需要，并不使用先标引，然后提交检索词汇做匹配计算的方案。本文采用的搜索与下载一体化的方案，系统可在无人值守的状态下完成工作。用户只要在工作前明确任务要求，也就是要对用户的查询进行表现，并提交给搜索及下载工具，主要包括专题的定制，知识库的建立。如何能够使用有限的特征词，更好的表现用户需求信息的特征，降低对用户的工作能力要求，保证用户易于操作，是检索表现的一个重点研究内容，因此本文提出基于专题（Topic-based）的扩展查询技术。

3.3.1 基于专题(Topic-based)的扩展查询

由于在定制主题过程中，一般用户通常很难用简单的关键词来表达出感兴趣的专题，因此，在检索中词的不匹配问题比较严重。本文采用基于专题(Topic)的查询扩展方法，允许用户对每一个专题进行关键词的扩展，进一步修改知识库，完成交互式的辅助查询提交功能。

由于基于专题（Topic）扩展能够“给出构成相关领域词汇的基本术语，以及利用这些基本术语，构成的规定一些词汇外延的信息”。因此，使用基于专题（Topic-based）的扩展查询，有助于提高信息的表达，提高查全率，减少“漏搜”的有用信息。基于专题的扩展查询主要是完成对用户检索请求的智能优化，同义扩展、词义蕴含扩展、外延扩展、相关扩展等功能。

Topic 能够辅助一个系统识别处理需求，定义各种规范。根据应用的领域规模和目的的不同，对 Topic 研究的侧重也有所不同；涉及特定的领域，被称为领域 Topic；涉及通用的世界知识，被称为高层模型或通用的 Topic；涉及知识表示语言，则称为表示 Topic 或 MetaTopic。从这个意义上可以看到，任何一个给出的专题，都可以在相应的领域中确定其应有的位置，并和其他的相关概念建立语义联系。也就是说，给出的任何一个领域专题(Topic)都应该包含该领域中的所有可能术语描述的信息，同时还应该包含同义词和相关词汇描述的信息。一般来说，形式化程度越高，越有利于计算机的自动化处理。

定义的任何领域的形式化专题(Topic)至少有以下两方面的作用：

（1） 自动分析文档的领域属性。一般来讲，被检索的大多数文档都列出了关键词和内容摘要，将这些信息结合文档主要内容，在专题(Topic)知识的协助下，可以判断该文档属于哪个领域，并以此对文档进行分类。经过这一步处理，事实上可以过

滤掉不相关领域，得到所有可能与该文档相关的领域。而且，还可以根据近似语义相关匹配，使某一领域的相关文档按其相关程度进行排序。这样不仅可以使文档自动分类，信息的查寻不再是遍历扫描，而且可以使检索结果排序输出，从而大大节省检索时间和加快用户对检索的语用相关性判断。

（2） 智能化规范和显性化的查寻需求。由于用户对信息需求认识的模糊性，因此应利用一定的手段使模糊的信息需求在检索的初始阶段得以显性化和清晰化。对于用户给出的查寻关键词，在专题(Topic)知识的协助下，可以有效地判断其所属的专题，然后分别将该专题及其属下的相关概念与定义罗列给用户，用户据此进行相应的选择，一方面通过这一选择过程帮助用户明确其信息需求，把未意识到的、未清晰表达的客观情报需求进一步显性化；另一方面让系统了解用户所关心的专题，为检索过程提供更为精确的信息，在客观上有利于使相关性的判断向语用相关靠近和转移。

在完成用户检索请求的智能优化，同义扩展、词义蕴含扩展、外延扩展、相关扩展等功能的同时，扩展查询功能还允许用户指定主查询关键词和辅助查询关键词，前者是信息必然包含的词汇，而后者则允许不出现在信息文档中，起着辅助解释功能的作用。最后，将检索词汇按照权重分配组成专题特征向量，进行搜索。这样可以降低搜索中“遗漏搜索”和“冗余搜索”现象的出现。

用户可以自定义任何一个感兴趣的主题，而且可以界定并丰富该主题可以涵盖的内容。解决思路就是采用一定的策略扩展用户提交的初始查询.例如在收集中国航天计划的信息中，如果用户的原有查询词是“航天”一词，则会收集到大量中国航天信息，当然，同时也会收集到大量的世界航天方面的知识，但是如果在前面加上“航天”一词，则可以滤出其他国家的相关信息，解决无关网页的问题。如果只有“航天”一词，那么对我国的“神舟”宇宙飞船的相关信息就有可能漏掉，因此，可以通过基于专题(Topic-based)的概念扩展，追加“宇宙飞船”一词，就可以收集到我国的“神舟”方面的相关信息，则网页漏检的情况也可以解决。

3.3.2 概念之间的关系判定

在基于专题（Topic-based）的概念扩展查询中，最重要的是解决词库的扩展问题。由于词库不可能无限制的扩展，因此如何就要对是否要追加新的词条，做出判断。概念是通过概念关系联结起来的，因此在提取出与查询相关的概念后，就需要判定出这些概念之间的关系。概念之间存在很多种的关系：等价概念关系、近似概念关系、概念包含关系、否定概念关系等。在本文构造的知识库中，我们主要考虑两种类型的概念关系：一是包含关系；二是相似关系。

(l)包含关系　如果概念 A 表达的是概念 B 的一个方面或是一个部分，我们就称 B 包含 A，A 为细化词，B 为概述词，两者形成包含关系对，记为 S(B，A)。例如，S(“软件”，“操作系统”)。

在本文中我们采用源于 sanders 和 croft 在 1999 年提出的包含方法，判定概念之间的包含关系。这个方法已被证实是一种可靠也可行的技术。X 和 Y 是两个术语，如果有 Y 出现的文档集是有 X 出现的文档集的子集则可认为 X 包含 Y，后来 Wuyi_Fang

提出了包含方法的修正规则(revised subsumption approach)。本文中，我们所用的 X 包含 Y 规则如下：

$$P(X|Y) \geq N \geq P(Y|X), \qquad 0 < N < 1 \qquad (3-1)$$

其中：$P(X|Y) = \frac{N_{xy}}{N_x}$；$P(Y|X) = \frac{N_{xy}}{N_y}$；$Nx$ 是有 X 出现的文档的数目；N_y 是有 Y 出现的文档的数目；N_{xy} 是有 X 和 Y 出现的文档的数目；N 为阈值。

阈值越大，则从文档中提取的术语对就会相对较少，反之亦然。这个阈值根据需要适当地选择。我们设定 N 为 0.8，因为在搜索过滤系统中，根据这个值所提取的包含关系对建立的知识库结构上较为合理。

(2)相似关系　所谓相似关系实际也就是同义关系。Grefenst 于 1994 年提出上下文相似法用于判定概念间的同义关系。上下文相似法认为如果两个词的上下文相同的话，则这两个词是同义的。在构造知识库时，我们首先根据包含方法分析出了概念之间的包含关系，因此每个概念分别具有一个相关的概述词集和细化词集。在构造知识库过程中，我们将 Grefenst 的上下文分析法进行了演变：如果两个术语的相同概述词和细化词的个数都超过某个阈值的话，则认为它们的概述词集和细化词集是相似的，从而认为这两个术语是相似词。在系统测试过程中设定这个阈值为 50。通过以上方法决定这两个概念是否为相似关系。

当不存在上述两个关系时，系统将从知识提取器中，将提取的主题特征，写入知识库，以动态地丰富、完善知识库，提高知识库与信息检索和过滤过程中的匹配能力。

3.3.3 扩展查询的实现

扩展查询的实现主要依赖于词库的增量化管理。信息时代的最大特色就是新信息产生速度快，尤其 Internet 网的信息扩展是非常迅速的，知识库必须实现增量化管理，才能同步地为用户提供服务。

检索系统中实现查询扩展功能主要采用以下步骤完成：

（1） 首先，利用静态知识库进行查询扩展。这种类型的检索系统中由专家或者用户，按照一定的组织结构预先建立一个静态知识库存储专题领域知识。检索过程中，从相关的信息文档中查找与静态知识库中的术语匹配的部分。按照相关性关系，对其中的信息特征扩展知识库。

（2） 然后，利用检索结果进行查询扩展。这种类型的检索系统中，系统执行两次文档查询以实现用户的一次检索，第一次的查询结果用于扩展用户的初始查询，对反馈的文档进行知识提取，用户获得重要的反馈信息后，或者从相关的信息中获得相关的启示后，可以进一步修正用户的原始定义。

在(1)中，知识库是由专家或者用户按照对本领域知识的理解建立的，因此根据知识库进行的查询扩展符合人的认知规律。但也正因为知识库的这种构造特点，整个知识库的组织和管理与用户对领域或者相关知识的认知相关。

在(2)中，则是利用文档集反馈信息做扩展查询，因此实现的查询扩展更贴近文档集的内容，表达的主题也会更加有效。

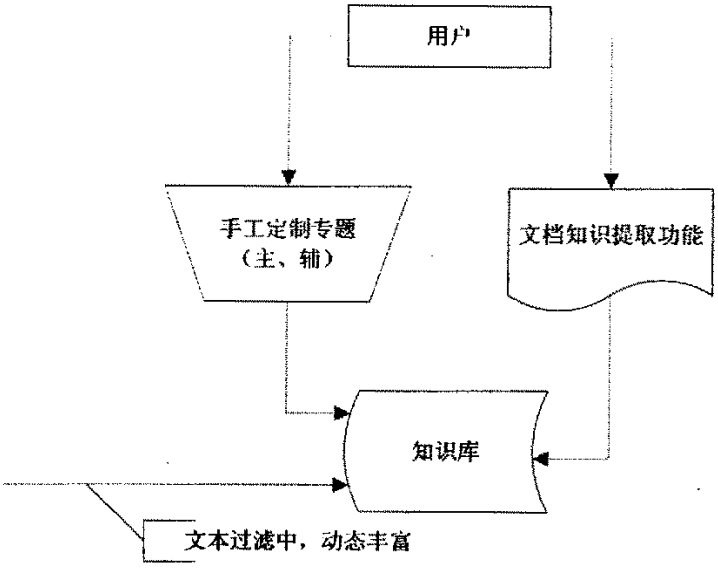

图 3-2　扩展查询的实施方案

采用基于专题的扩展查询具有以下四个特点：

（1） 实现了专题层次的检索，突破了关键词检索局限于形式的固有缺陷。

（2） 实现了对用户检索请求的合理化联想，整个搜索引擎像领域专家一样，提高了信息的查全率。

（3） 知识库的各个环节，包括同义描述元素集合、关系表等都实现了增量化管理，具有良好的可扩展性。

该系统也有一些缺点，我们可以使用一些统计的或基于规则的方法来自动获取领域概念，但整个知识库的组织和管理还是需要人的干预，需要手工对知识库进行修改、添加、删除等操作。

§3.4　查询问题的实现技术

3.4.1　HTML 网页的解析

HTML 文档实际上是一种纯文本文档，可以在写字板中打开、编辑，但与常规的正文文本相比有着很大的差别。HTML 中包含了正文文本和大量的 HTML 标记，因此要对网页作解析处理[4]。这些 HTML 标记主要用于定义文档的标题、字符集等属性信息，控制文本的显示格式和表现效果，以及引入超链或各种媒体类型等。这些 HTML 标记相当于插在书本中的一个个书签，为提取文档中的有用信息提供了很大帮助。所以，在对 HTML 文档进行扫描处理前，首先需要对 HTML 标记进行正确的识别和处理，并根据 HTML 标记对网页不同部分的文本进行加权处理。利用 HTML 标记对索引词加权的思想来源于传统信息检索中对文本结构信息的挖掘。如：对段首句和段尾句中的词给予更多的权重；对文章首段中出现的词赋以更高的权值。本文发掘 Web 页面的“半

结构化”的特点，充分利用 Web 文档结构的结构特征，对索引词加权提供了更多的信息，所以在本系统中首先根据标记的修饰作用对索引词加权。处理过程如图 3-3 所示。

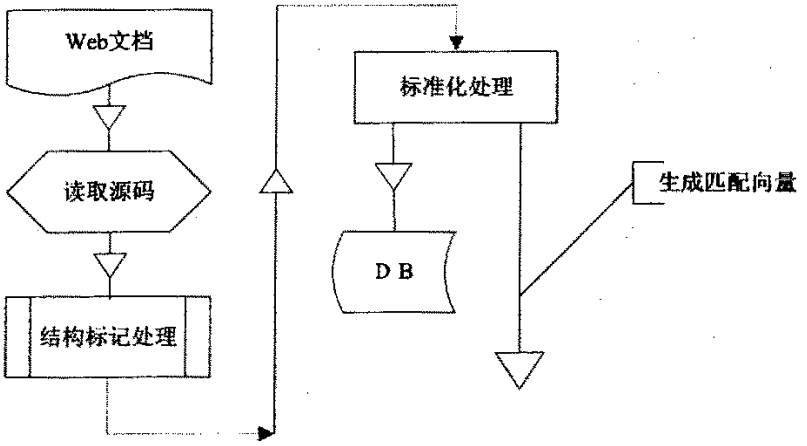

图 3-3　页面解析处理过程

3.4.1.1　HTML 标记的处理

在 HTML 网页的标记中，根据需要我们将之分为两类：

一是对我们的信息提取有辅助作用的。虽然有的标记本身及其所修饰的内容均不在浏览器中显示，如包括<Meta>(元数据)中、<TITLE> </TITLE>(标题)和<!->(注释)，但是对 HTML 网页文本的分析，却起着非常重要的作用。因此，可以在解析器中加以利用，辅助文档特征的提取。但是，<Meta>中的内容多为网页作者对网页中关键内容的描述，许多搜索引擎也以此作为索引网页的依据，但是正因为这种原因，现在许多网页在<Meta>中大量加入与本网页并不相关的内容来“欺骗”搜索工具，使<Meta>中的数据失去了原有的作用，所以系统中也不能完全依靠页面结构，来提取信息特征。

二是我们不准备采用的标识符。虽然有的标记修饰的内容在浏览器上显示，绝大多数的标记属于这一类，它们又可以分为三个子类：改变文本的显示效果，如<B>(粗体显示)、<I>(斜体显示)；改变文本的内容样式，通过改变文本的显示效果来实现，但这些标记对反映其修饰内容的属性作用并不大，以上标记的特征是：标记所修饰内容的显示效果还可以用别的标记代替，具有不确定性，所以为提高系统的执行效率，我们暂不考虑这些特性。

基于上面的分析，我们采用了表 3-1 所示的 HTML 标记加权方案。

表 3-1　HTML 标记加权方案

<TITLE>...</TITLE>	1
<H1>...</H1>	0.7
<Content>...</Content>	0.5
<Meta>...</Meta>	0.3

3.4.1.2 HTML 文档的预处理

HTML 文档的预处理主要完成的工作是：扫描 HTML 文档源代码，对于需要加权的 HTML 标记所修饰的内容进行加权处理，不需要加权的按照默认权值 0.5 处理，将被修饰的正文内容交由下面的语句切分模块进行处理。

经过上面的分析后，HTML 标记对于处理程序来说可以分为两类：需要加权的和不需要加权的 HTML 标记。对于不需加权的标记，只作为语句分隔符进行处理，这样就简化了预处理算法的工作。预处理的结果保存到队列中，作为语句切分的源数据。

由于网页作者的操作规范不统一，造成 HTML 网页的非标准化，给网页分析造成困难，例如：

（1） 空白字符(tabs)字符，回车符以及换行符等可能会以不定数量出现在有用元素之前或之后，也可能出现在元素的中间；
（2） 标一记大小写不统一；
（3） 标记之后可能没有，也可能会有一个或多个参数；
（4） 参数值中引号等分界符使用比较混乱，如对于相对链接中有的网页使用单引号，有的网页会使用双引号；
（5） 标记参数的次序没有统一规则。

所以在预处理中，要对网页中以上这类非标准信息，做标准化处理，转化为程序可识别的、可处理的格式。

3.4.1.3 网页扫描算法

因为HTML语法的不规范性，可能出现HTML标记嵌套的现象，所以在算法实现时设计了标记属性值和加权属性值进行处理，对于被两个或两个以上需要加权的标记修饰的内容，按最大加权值进行处理。算法的基本处理过程描述如下：

（1） 按步骤（2）—（3）对 HTML 文档进行扫描，直到文档结束；
（2） 初始化标记属性值、加权属性的初始值，设当前加权值为 0.5；
（3） 对 Web 文档进行扫描，如果文档内容为：
  a) 正常文本，继续扫描直到遇到HTML标记或文档结束，将扫描过的文本保存到结果队列中；
  b) 需要加权的HTML开始标记，将该标记入属性值，比较该标记的加权值和当前加权值，将小的加权值写入属性值，大的加权值作为当前加权值；
  c) 不需要加权的HTML标记的开始或结束标记，将前面扫描过的文本末尾加入分隔符，跳过该标记继续扫描后面的文档；
  d) 需要加权的HTML标记的结束标记，将扫描过的文本保存到结果队列中，标记属性值更新，权值更新值作为当前加权值。若属性值为空，则以0.5为当前加权值。

为了提高程序的执行效率，在算法实现时对于需要加权处理的 HTML 标记，从数据库中读入，保存在平衡的检索二叉树的结点中，程序对 HTML 文档扫描时从检索二叉树中查找特定的标记。

3.4.2 查询向量的建立

为了更好的收集到用户感兴趣的专题信息，既要注意信息的质，又要关心信息的量，使收集的信息在保证一定的相关度的同时，能够收集到更全面的信息，以为数据的挖掘提供最充足的信息资源。但是，单纯的依靠用户提供的关键词完成匹配，进而检索相关信息是远远不够的。单纯的依靠关键词会导致大量的“漏搜”，不能最大限度的收集到用户所需的相关主题的信息。主要原因如下：

（1） 由于用户的提供的关键词数量有限，一般用户不会提供三个以上的关键词，而根据统计，一般要在 3 至 5 个关键词才能基本上表达清楚一篇文档的主题；

（2） 由于用户提供的关键词的质量问题，由于一般用户难以用有限的关键词来准确表达个人的要求，总是与实际的主题表述有一定的偏差；

（3） 相关度的评价很大程度上存在着主观因素，与信息的需求者之间存在不确定性，同一信息对不同用户的可能度可能并不相同。

因此，我们认为检索问题和收集的目标文章所承载信息之间不是一一对应的映射关系，而是基于专题(Topic)的一对多的概念上的映射，如图 3-4 所示。

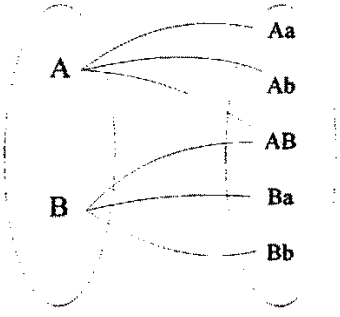

图 3-4 基于专题的一对多映射

为提高搜索工具的效能，采用检索向量构成信息需求。信息需求不是简单的关键词的罗列，而是由一系列的关键词按照一定的权重分配构成的。例如：半结构化的 Web 页面，对<content>、<title>、<meta>中的结构词可以赋予较高的权重。但是，对于网页的标题权重又不能太大，因为一篇文档的内容具有一定的范围，如果对网页的标题赋予的权值太大，则会导致主题过于集中，不能覆盖全部文档信息；对用户自定义的兴趣专题，用户可以根据主观的要求，对相应的关键词赋以相应的权重，构成检索矢量。

查询矢量的生成算法如下：

文档用 D(Document)表示；特征项用 T(Term)表示；K（Keyword）表示出现在文档 D 中且能够代表该文档内容的基本语言单位，主要是由词或者短语构成，则文档文本可以用特征项集表示为：

$$D(t_1,t_2,...,t_n) \qquad (3-2)$$

其中 t_k 是特征项，$1 \le k \le n$。

例如：一篇文档中有 a、b、c、d 四个特征项，那么这篇文档就可以表示为 D（a、

b、c、d）。对含有 n 个特征项的文本而言，为区别各个特征项的效用值，突出主题，我们赋予每个特征项一定的权重表示其重要程度。则文档 D 可以表示为：

$$D=(t_1,w_1;t_2,w_2;...t_n,w_n) \qquad (3-3)$$

简记为：

$$D=(w_1,w_2,...w_n) \qquad (3-4)$$

我们称之为文档 D 的向量表示。其中 w_k 是特征项 t_k 的权重，标示出特征项 t_k 在文档 D 中对文档 D 的主题的贡献度。假设特征项 a、b、c、d 的权重分别为 0.4, 0.3, 0.2, 0.1，那么该文档的特征向量则可以表示为 D（0.4, 0.3, 0.2, 0.1）。

3.4.3 文档的相关度计算

相关性是考虑文章和文章之间的关系，相关不可能纯粹由主题决定，许多非主题因素也会影响读者的相关性判断，因此，相关绝不能被简单定义为主题相关。Goffman 理论[18]的主要研究成果就是在测量检索问题和文章所承载信息间之相关关系，Goffman 认为相关绝不止于传统检索词汇和描述文章词汇间的简单吻合关系，应该加入其它考虑才可能完全定义相关。因此我们采用查询向量进行搜索，而不是简单的依靠关键词的一一对应来完成检索。

在得到具有权重分配的文档的检索向量之后，就可以对文档的相似度进行计算。按照文档特征向量的相似度分类存储，其具体计算方法为：对定制的专题 T_1 和下载文档 T_2 之间的内容相关度 $sim(T_1,T_2)$ 用两个向量之间的夹角的余弦值来表示：

$$sim(T_1,T_2)=\cos\theta=\frac{\sum_{k=1}^{n} w_{1k}\times w_{2k}}{\sqrt{\left(\sum_{k=1}^{n} w_{1k}^{\ 2}\right)\left(\sum_{k=1}^{n} w_{2k}^{\ 2}\right)}} \qquad (1\le k\le n) \qquad (3\text{-}5)$$

其中，w_{1k} 和 w_{2k} 分别表示定制标准主题 T_1 和网页 T_2 第 K 个特征项的权值。

当 $\cos\theta$ 小于指定的阈值 V 时，则认为是不相关的，不能属于参与比较的定制专题，要与下一个定制专题进行匹配，如此循环，完成信息的收集。

第四章　搜索过滤技术研究

搜索过滤技术是信息收集与过滤系统的核心技术，它直接决定整个系统的效能。一个良好的过滤技术可以最大限度地降低网络的传输负载，节省系统的资源，进而提高搜索工具的执行效能。

本文研究的过滤技术主要包括两个方面：基于内容的文本过滤和基于链接的过滤。前者是对下载的文本做基于内容的精细过滤，保证下载后分类存储的主题不产生漂移，同时，对信息的来源地址做出评价，并为基于链接的过滤提供依据。后者对提取的链接进行分析、筛选，然后写入地址表，以供系统的进程调用。这种有选择的链接提取方法，可以提高搜索和采集的效率，而传统的使用搜索引擎的方法则是贪婪的遍历策略，需要访问所有的链接，这种方法使得系统搜索效率下降。

§4.1　基于内容的文本过滤及分类技术

虽然网页具有一定的组织结构，可以供我们分析文档的内容，但是这只能起到一个辅助的作用。由于现在 Web 文档的结构上还没有统一的标准，如在<meta keyword>节中，出现的内容并不一定是本部分信息的表述（当前由于受到一些商业动机的影响，可能会采用这一特性来欺骗通用搜索引擎）；而且不同的人对相同的文档的理解是有差异的，所以 Web 工作者提取的中心内容可能也不一定是我们所感兴趣的内容。因此，如果单纯的依靠这种结构性特征对网页的内容进行评判，然后就收集下载，势必会有大量的信息被误判，从而导致大量无用信息的错误收集，大量有用信息却被遗漏。为了更加准确真实地表达某一个页面或者是板块的主题，收集到相关度更高的信息。

根据主题（如关键词、主题相关文档）与链接文本内容的相似度来评价链接价值的高低，以此决定其搜索策略，本文称之为基于内容的搜索过滤策略。基于内容的过滤策略可以有效的避免以下问题的发生的概率：

（1）　由于 Web 文档的结构上没有统一公认的标准，基本上是由网页发布人员主观决定，易产生不公正的评价；

（2）　受竞价商业模式影响，投机某些搜索门户的搜索策略，标记结构中使用不恰当的关键词，容易影响搜索结果的客观性，常使搜索结果与用户需求之间产生错位；

（3）　网络信息质量控制存在欠缺，任何人只要具备相应的条件就可以把任何信息送到网上,而这些信息不经任何质量控制就被搜索引擎标引，未经质量控制的信息必然会影响搜索结果的质量。

通过信息过滤，与用户个性化模式不相关的文档或用户不感兴趣文档被过滤掉了，对用户感兴趣的信息文档做基于向量空间模型的聚类处理，对相关度大于预定阈值的文档，写入数据库。

采用基于内容的过滤策略还有一个重要的作用：对文本与知识库比较，记录比较结果，标记当前页面的主题与定制主题相关度，从而对该信息的来源链接做出评价，以作为链接过滤的参考。对已收录的信息文档的来源做反馈标记，强化有效链接的权

威度，衰减不相关页面或者相关度低的页面的权威度，引导新的链接资源的提取和过滤。

基于向量空间模型的聚类算法一个很大的缺陷在它没有考虑文本的上下文间的语义关系和潜在的概念结构（如词汇间的共现关系、语义关系等）。同时，向量空间模型基于分量无关的假设在实际中也是不可能的，作为分量的词条往往有很大的相关性。针对这个问题，本文改进了特征向量的选取算法，结合语义权重组成特征向量。对下载信息与定制的专题进行相似度计算。

4.1.1 基于内容的过滤和聚类预处理

要实现检索矢量的作用，就要正确提取出网页的特征，否则信息表现做的工作只能是徒劳。网页的特征表示工作中最重要的一步就是特征选择，特征选择是选择那些最具有区分性和排他性的特征，也就是最能把类别区分开来的特征，而不是大多数对象都具有的特征。

4.1.1.1 问题描述

文档的自动过滤与分类可以描述为如下过程：抽取文档特征，将文档表示为统一的标准方式；使用分类器(Classifier)判断文档所属类别，分类器是分类系统的核心，可以通过学习不断改进和完善，增加、更新类别和知识(增加专有词汇等)，对类别的描述如下：

C：类别集合；T：所有文档集合；D：训练文档集合，$D \in T$；

S：文档特征向量空间；d：文档；

R：$T \rightarrow S$，映射，将文档转换为特征向量；

U：$D \rightarrow C$，已知的映射，训练集合中的文档已经分好了类。

分类的工作是给出映射 R，并以此为基础，使用 U,R, D, C 构造映射 H： $T \rightarrow C$，使得 H 与 U 尽可能接近。

本文中，文本分类的研究所采用的分类模型，如图 4-1 所示。在模块中，首先将文本集向量化，得到特征的集合；类别特征向量空间生成器从特征的全集中抽取经过权重计算的特征子集构成文本分类的特征向量空间。然后，将测试文本用特征向量表示，再经过分类器分类，得到所属的类别。

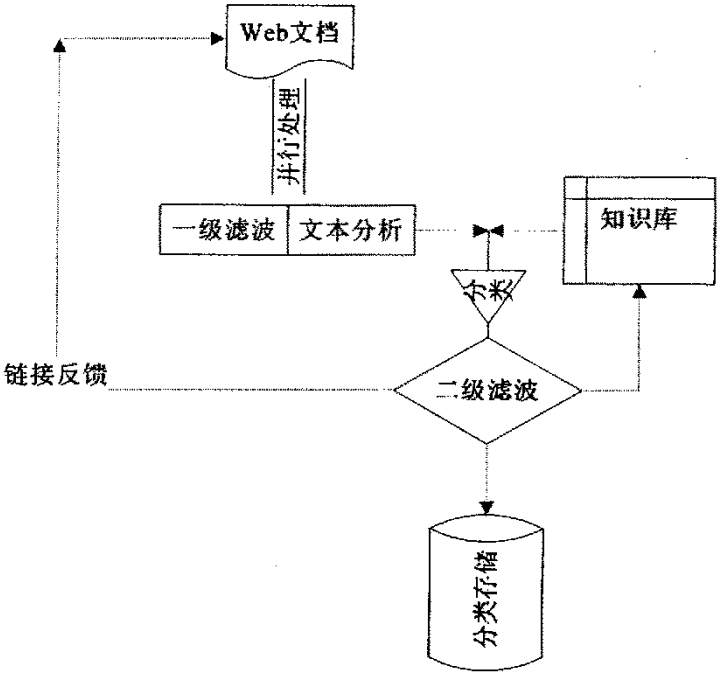

图 4-1 文本分类模型

4.1.1.2 分类及过滤方法中的关键技术

在文本的分类及过滤部分中，文本的处理采用的是基于向量空间模型（VSM）的文本处理方法。基于 VSM 的特征提取方法都是统计的方法，首先利用不同的方法对特征项进行评分，然后选出分值较高的作为特征构成特征向量空间。常用的特征提取方法有文档频率、信息增益、互信息等，本系统采用文档频率作为特征提取标准，该方法计算复杂度低，适合在线分类。但是由于通常的基于向量空间模型的方法不考虑上下文和语义关系，因此本文引出针对网页结构的权重特征，并提出对低频的专有名词的特殊处理方法。

1、特征项的抽取

特征提取是文本分类系统中十分关键的问题，它可降低向量空间的维数，提高系统的速度，提高系统的精度和防止过拟合[22]。

通过大量的文本分析，可以发现，能够反映文本类属性的特征词的词性具有一定的规律，特征词大多是名词，少数动词和少数形容词。一般来说，介词、副词、感叹词、冠词、限定词等不能作为特征词。另一个因素，特征词主要是区分各类的特征，虽然有的词在一类中出现很多，但在别的类中也可能出现很多，这一类词(或称为普通词)一般也不能作为特征词。去掉普通词的一般方法是建立一个禁用词表，效果不是很好。与一般纯文本文件不同，WWW 网页是 HTML 格式的超文本，页面中含有<title>、<meta>、<head>等标记，以及描述此页面的标题(title)、页面描述(description)、关键词（keyword）、超链接(URL)等。这些标记可能包含重要的分类信息。基于以上的考虑，本系统采用利用专业词典，从专业词典中为各类抽取特征词，这样会大大提高抽取特征项的准确度，有效地保证最后提取的特征项数目较小且相互独立，同时也减少了训

练量。

在分类系统中，使用基于文档关系的加权相似度的特征选取算法，这是一种类似于过滤器的方法，特征选择独立于分类器的学习算法，去除不相关的或者相关度不大的特征项。

设 d_i，　d_j 为两个文档变量，它们之间的相似度为：

$$s(dij)=\begin{cases}0, & if(dij>D)\\ 1-d_{ij}^{\ 2}/D^2, & otherwise\end{cases} \qquad (4-1)$$

dij 是矢量 d_i 和 d_j 的欧几里得距离，类的形状是超球面；D 是类之间的最小距离。如果两个文档的相似度越大，那么它们属于同一类的概率也越大。

但在这实际应用中的效果不很好，因此引用加权相似度。改进 dij 采用如下公式计算：

$$dw_{ij}=\sqrt{\sum_{f=1}^{p} w_f^{\ 2}(d_{if}-d_{jf})^2} \qquad (4-2)$$

其中：w_f 为文档的特征 f 针对结构上的加权相似度。

采用上述公式的加权距离来提高性能，使类的形状成为超椭球面。这样，在计算相似度时，w_f 使得特征 f 的相关程度更加明显。w_f 值越大说明特征 f 的重要程度也越大。

$$D=\sum_{i=1}^{N}\max_{dj\in T} dEij/n \qquad (4-3)$$

用来替代公式(4-1)中 D，该算法不仅考虑了文档矢量空间相近性，而且引入了 Web 文档的结构特征。

2、类特征的权重计算

设带有类标识的训练样本库 T，T={t_1，t_2，…t_n}。训练样本库 T 共有 N 个类样本。假设对第 n 类进行特征向量的抽取，对每个样本 t∈T，做以下处理：首先除去停用词，然后对文本 t，进行切词，设词典为 Dict，提取词典中出现的词构成关键词集合。关键词集合表示如下：

$$CWord=\{w|w\in Dict, w\in t_n\} \qquad (4-4)$$

关键词集合(即特征项集合)确定后，需要确定每一特征项在这一文本类中的权重。一般认为：词条的重要性正比于词条在文档内出现的频数，反比于训练文本内出现该词条的文档频数。因此有些系统选用可构造词条权值的评价函数：

$$w=tf*\log(N/n) \qquad (4-5)$$

其中，tf 表示词条 W 在文档 T 中的出现频数，N 表示全部训练文本中的文档数，n 表示词条 W 文档频数。

本系统中考虑到网页的特点，它与普通的文本不同之处在于网页中除了含有纯文本信息以外，还有其他描述信息，这些描述信息中出现的关键词包含网页的重要信息，对分类有较大的作用。因此，对从网页中提取的这些信息进行了加权处理。

设样本 t 的关键词 W 在标题、页面描述、关键词、超链接中的词频分别为 tf_{title}，tf_{des}，tf_{key} 和 tf_{url}。总词频 tf 为：

$$tf = tf + a * tf_{title} + b * tf_{des} + c * tf_{key} + d * tf_{url} \qquad (4-6)$$

其中：a, b, c, d 为大于 0 的权重参数。

3、专有词汇的处理方法

分词[24]是大规模中文文本处理的最基础的步骤，但是 Dict 未包含词的分词处理问题，尤其是专有名词的处理问题是一个难点，由于在文档中专有名词一般来说，是使用频次很低的，如一个地名，但是对于使用文档频次来计算权重的算法，分词的准确率和分类系统的准确率就会影响很大。因此，本文提出对专有名词的特殊处理方法。据统计，真实文本的分词精度一般达不到 80%。分词精度过低的最主要原因不是分词中的歧义现象，而是分词词典词汇量的限制。汉语中有几十万个词汇，诸如人名、地名、机构名、术语等这些专有名词往往和特定的文档有关，具有较高的区分度值，对确定文档的类别有一定的作用，但绝大多数专有名词却没有被通用词典收录。如果一味地扩大词典规模也是不现实的，除少量常用词外，大多数词汇出现频率极低，且在特定环境下出现。因此，本文利用统计方法来自动地从文档集中抽取专有词汇。在使用通用词典进行第一遍切分之后，搜索所有可能的专有词汇，称为候选词，再运用计算 λ^2 统计量的方法，计算每个候选词的相关度，最后根据相关度大小决定该候选词是否为专有词汇。

4.1.2 匹配阈值的选取

系统采取“排列分类[23]”（ranking classification），即计算待分类文档与所有类别的相似度，然后进行排列。一般情况下，文档属于相似度最高的类别。但有时候，一篇文档可能属于多个类别。如在本应用中，一篇有关以色列问题的文档既可以归入美国类，也可以归入以色列问题类，又因为每个类别与对应文档的相似度分布不尽相同，因此本文为每一个类别均设置了一个匹配阈值。相似度高于此匹配阈值的文档即属于该类别。至于阈值的选取，目前理论上还没有很好的解决办法。在本文中，对于阈值的设置提出了一种“平均值”法。基本思想是：对于每个类别 C，计算在 V 集中属于类 C 的所有文档与 C 相似度的算术平均值，作为此类别的初始阈值。然后在训练过程中，可以根据需要进行调整。调整的幅度是各个相似度与当前阈值的差的平均值。举例说明如下：假设 V 集中文档 d1,...dn,属于类别 C，初始阈值 $K_0 = \sum_{i=1}^{n} sim(di,dj) / n$，阈值调整公式为：

$$k = k \pm \Delta k \qquad (4-7)$$

其中 k 为当前阈值，$\Delta k = \sum_{i=1}^{n} \frac{|k_0 - k_i|}{n}$，$k_i = sim(di,dj)$，经验证明，这种方法可以使分类性能快速达到最优，在效率上高于凭经验调整阈值的传统方法。

§4.2 深度控制技术

在互联网中近 20 亿的网页中，所有网页具有一个引用权威和相关主题的倾向。据统计，近九成网页是通过互相引用和链接提供给用户的，因此这些网页之间存在一个链接深度的问题，有的可能是在主页提供的，而有的可能是在其子页面中提供的。虽然网页是一个松散的半结构化文档，但是正规的信息站点，在信息的组织上仍然存在共同的约定，按相关的内容分成多种板块（或称为团），在每个板块进一步会展开相关的内容。如果将其中的板块看成一个节点，链接看成一个边，那么整个结构就构成了一个树结构，如图 4-2 所示。

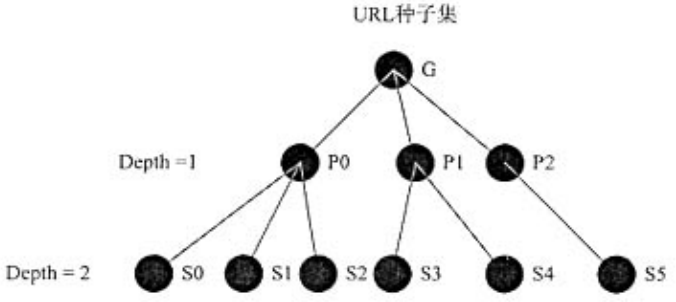

图 4-2 链接结构示意图

这棵树根据信息的相关密度和信息量，可以不断长大。因此，搜索工具可以跟踪子树，发现新的信息源。但是如果搜索工具无限制的跟踪链接也是不可行的：

一是由于在树的成长过程中，除了继承父亲的特性，外还会发生变异，产生新的个性，而这些新的特征不是我们所需要的特性，如果不适用后文的基于内容的过滤机制，将会出现主题内容的漂移（drift）；但是如果采用基于内容的过滤，虽然能够保证收集到的专题信息的相关性，但是却会产生新的问题，即遍历所有的链接信息，会浪费大量处理器的资源，也会增加网络的传输压力。

二是由于各种站点的更新机制不同。如搜索主题站点百度的更新就较 Google 快，而百度对链接分析能力较 Google 差，Google 能分析到下四层链接，而百度可能只能至第三层，Google 的中文数据库一月一更新（英文的一天一更新），但 Google 在全球有数千台服务器，它们对各种语种更新的速度也是不同的，并且更新方式采用梯度更新，即先更新某些服务器，再进行全面更新。类似地，其他的信息站点也是一样的，一些重要的信息总是放在主页上，能让用户在第一时间接触到，而在深层次的板块投入的精力明显不够，导致更新周期变长。因此，我们有必要对链接的深度加以控制。

为了对链接的提取数量做有效的控制，对链接的深度加以控制的同时，本文还采用宽度优先与深度控制相结合的链接提取方法，既保证了收集信息的扩展范围，又控制了系统对链接提取的贪婪度。

4.2.1 深度控制的实现算法

搜索深度控制算法主要是依据链接的继承特性，防止无限制的深度跟踪会损失信

息采集的相关性和广泛性，造成有用信息树枝的丢失，其主要的实现过程如下：

(1) 首先，从 HTML 解析器的输出中，提取出超链接（Anchor）；

(2) 分割出超链接的最后部分（最后一个斜线后面部分），判断是嵌入媒体还是指向新的链接；

例：对链接 http://news.163.com/military/cn/123.rm 提取出“123.rm”，则算法跳出。否则进入深度累计程序；

(3) 通过类型识别系统，判断其链接类型；对深度的控制算法如下所示：

```
Length = 0 ;
if （文件名＝*.rm or *.rmvb or *.avi or (等视频类)）
{
   Anchor 标准化；
  Anchor 写入  table of video；
}
   if （文件名＝*.jpg or *.gif or *.jpeg or (等图像类)）
{
   Anchor 标准化；
  Anchor 写入  table of image；
}
   if （文件名＝*.html or *.shtml or (等静态网页类)）
            and (Depth < D0 )
{
    标记深度 Depth = Depth + 1；
    计算信息的相关度；
    评价信息的来源地址；
    预测相邻链接的信息相关度值；
    预测子链接的信息相关度值；
    计算链接的重要度(Importance)；
    提取满足要求的链接地址(Anchor)；
    Anchor 标准化；
    Anchor 写入  table of TempUrl ；
}
回溯法访问链接树；
```

(4) 当 Inet 控件读入新的链接后，重复上述执行流程。

Anchor 标准化是将页面信息能够正确下载的重要保证，由于相当的页面在引用链接时，并不使用完整地址，而是使用相对地址。这时，如果通过使用 HTML 解析器提取出的链接，直接保存，那么下载信息时从链接表（TempUrl）中读入的链接地址，系统并不能正确识别。因此，要在使用标准化过程中，记录当前主页的地址，识别并判定当前的地址类型，对将相对地址标准化。例如，系统的实验过程中，选取的种子集为网易新闻链接(http://news.163.com)，页面解析器提取的链接为../world/eu/051107.html，如果不实施这一链接地址的标准化，直接写入地址表，则

当系统再次读入这一信息时，就无法识别服务器地址。所以要对链接标准化，最终记录的链接地址为(http://news.163.com/world/eu/051107.html)，这一信息就可以支持系统在任何时候，由任何线程调用。

4.2.2 宽度优先搜索可达搜索判断法

设在链接关系模式 G(U,D)中，有属性集 V 和 W，采用宽度优先搜索法判断从 V 到 W 是否可达，就是首先扩展较浅层次的节点，同一层次(深度相同)的节点按任意顺序排列。由于与或图的有向弧之间存在“与关系”，所以当把一个节点的所有子节点扩展完后，还要考察这一节点和已经扩展出的所有节点(包括深度比它浅的节点)的组合(即与关系)能否扩展新的节点。针对本系统收集信息的使用目的，收集更多的相关专题信息，本文对原始的链接种子集采用宽度优先算法提取链接。

宽度优先搜索算法实现步骤如下：

（1） 把属性集 V 中的所有节点依次放入 TempUrl 表，节点放置顺序任意，深度为 0。

（2） 将指针 p 置于 TempUrl 的第一个节点处，如果 p 为空，则无解失败退出；否则继续。

（3） 扩展节点 p，若 p 无后继节点，则转向（5）；有后继节点，则转向（4）。

（4） 扩展节点 p 的所有后继节点，依次放入 TempUrl 表，节点放置顺序任意。

（5） for(J=1；J<p；J++)

{对 TempUrl 表中的第一个节点到第 p-1 个节点作长度为 J 的组合，然后将每一组合与 p 节点相并后再做扩展，若有后继节点就将所有后继节点依次放入 TempUrl 表，节点放置顺序任意。}

若这一轮扩展的所有节点都和 TempUrl 表里的节点重复，则失败退出。

（6） 查看扩展的 TempUrl 表中是否包含属性集 W 的所有节点，若包含，则说明从 V 到 W 可达，成功退出程序。

（7） p++，若 p 不为空，则转向（3）；否则继续（8）。

（8） 从 V 到 W 不可达，退出程序。

对同一站点采用宽度优先搜索的方法需要扩展较多的节点，不会在最初的链接提取中，出现“木桶效应”。

宽度搜索方法虽然保证了信息的收集的广泛性，但是这种方法只是盲目地从属性集 V 出发进行扩展，而目标节点集 W 的信息在整个算法过程中并没有用到。系统必须将大量资源站点的内容传送至搜索站点本地，然后进行分析，这样大规模的资源文件的传送和处理无疑会增加网络传输的负担，使网络变得更加拥塞，此外也大量占用了被搜索站点和搜索站点本身的 CPU 资源，致使用户的访问不能得到系统及时的响应。因此，为提高信息收集的效率，减少系统资源的占用，降低网络的传输压力，采用链接的过滤技术。

§4.3 基于链接分析的搜索过滤技术

最近的研究表明，单纯依靠基于内容的网络搜索与过滤策略在距离相关页面集较

近的地方搜索时会表现出良好的性能。但由于页面中的文本信息缺乏“全局性”，很难反映 Web 的整体情况。而单纯依靠链接的访问数量来评价链接的重要程序则存在计算复杂，运算量随着链接数量的增加呈指数增长的趋势，不适合一般的用户需要。因此在保证收集的专题信息的相关度的前提下，进一步提高信息的收集效率，采用基于内容和链接评价相结合的搜索过滤策略。

对于新信息源的发现，我们采用的方法有：①从用户提供的链接种子集中获取；②在有确定目标的主动搜索中，通过系统的推送功能来获取；③从相关页面提供的链接中获取。对于方法①比较容易实现。对于方法②采用基于案例的统计学习算法实现。而对于方法③，由于传统的搜索引擎主要的问题就是搜索的盲目性，即不论什么问题都是采用同一种方式或策略，搜索效率比较低。因此，本文中采用启发式的搜索方法，发掘 Web 的结构特征和语义特征，对链接提取采用基于内容的与基于链接相结合的过滤策略。

采用基于链接评价的过滤技术，可以增加有用链接的发现距离，更加精确的提取出扩展链接，滤出无效的链接和噪声。把内容相关度作为链接评价的依据，对链接进行过滤，既解决了单纯依靠内容检索的“近视”问题，又保证了信息的收集的相关度，同时，也避免了传统方法的贪婪搜索带来的资源浪费。

4.3.1 基于链接过滤的指标算法

一个好的网页总是少不了大量的链接（Anchor），但是这些引用和链接不是随意组成的，而是有着特殊的意图的。通过研究我们发现链接间的引用具有相关性，统计表明，同级页面具有引用相同主题页面的特点；子页面具有继承主页面主题的特点。基于链接分析的过滤主要是发掘页面间的相关性，继承性。充分利用链接间的这种特性，发现新的链接，并有选择的提取，这样将大大提高信息收集的效率，降低网络的传输压力和系统资源的消耗。

为了使得整个页面收集过程尽量持续下载相关页面，在从下载库中取一个链接爬行时，必须依据一定的标准来选取一个最有可能是相关页面的链接去爬行。这个标准在本系统中，就是每个链接 URL 的预测相关度，它用来表征链接指向的文档与特定主题的可能的相关程度。

对于链接的预测相关度的计算，本文借鉴了 HITS 搜索算法。考虑适用于一般的专业用户，没有商业化的搜索引擎服务器强大硬件支持，PR 算法只有在收录比较多的链接时，才能发挥出链接评价的准确性。因此，本文采用改进的 HITS 链接评价[39]算法，不但可以保证良好的链接过滤效果，而且可以有效的减少计算量，降低对硬件的要求。

在 Authority－Hub 链接评价算法，灵活采用网络检索深度控制、检索节点个数控制和检索时间控制技术。系统可由管理员发布命令结束收集过程，也可以在下载库为空时收集过程自行终止。

该算法将 Web 看成是一个具有大量节点和节点间连线的图，页面是节点，超链接是节点间的连接。算法从一个网络节点开始，给定代表专题的信息和链接深度限制条件，以该原始节点为中心，查看周边节点，查找与主题最可能相关的页面节点。它用

矢量模型来表示文档，用矢量模型的相关度计算模型来计算页面文档间的相关度 Correlation (p , q)。该算法发掘了页面间的特性，对页面的相关度采取了继承机制，对于相关度高的节点页面文档，其子页面文档的预测相关度就要比那些相关度低的页面文档的子页面文档的预测相关度要高，它的子链接被选中的概率也更大。在计算子链接文档的预测相关度过程中，还综合利用了文档链接的上下文信息，包括链接中的 Anchor 文本，以及链接周围的文字，都参加了该链接指向文档的预测相关度的计算。这样计算出的链接评价值具有局部特性，但是却缺乏全局的观点，因此本文中将链接相关度的评价值与 HITS 算法相结合，实现了全局特性与局部特性的统一，不但能够发现最有效的链接，同时也能防止提取链接的不妥“隧道”性。下面来介绍链接相关度的计算方法：

第 1 步，取得输入参数：包括专题的信息的特征（RV）和链接级限制条件，页面内容相关度阈值 Cthreshold，anchor 文字相关度阈值 Athreshold、相关度遗传衰减参数 d，以及两个用于计算链接上下文相关度和链接预测相关度的比例调节参数 α 、 β 。

第 2 步，当待下载链接库非空，并且未接收到停止下载的命令时，从下载库中取得链接重要度 Importance(page)最高的链接 URL 下载(初始时为经过系统预置的种子链接或者是系统推送的链接)。

第 3 步，提取链接并对 html 页面计算其与定制的专题特征生成的检索矢量（RV）的相关度：

Correlation (current node) = Similarity (RV, current node)　　(4-9)

如果该页面的相关度低于预先设定的阈值，则抛弃该 HTML 页面，回到算法第一步，并且抽取该页面中所有的子节点（例如图 4-2 中的 P0,P1,P2）。计算每个子节点的遗传相关度 inherited score (child node)：

If Correlation (current node) > Cthreshold

Then inherited score (child node) = d* correlation (current node)

Else inherited score (child node) = d* potential score (current node)

计算每个子节点的 Anchor 文字(anchor text)的相关度：

Anchor score = similarity(RV，anchor text)；　　(4-10)

计算每个子节点的 Anchor 文字周围指定范围内的上下文文本(anchor text context)的相关度 Score of anchor text context：

If anchor score > Athreshold

Then Score of anchor text context = 1

Else anchor context score = similarity (RV, anchor text context)

计算每个子节点的 Neighborhood 的相关度 Neighborhood score：

Neighborhood score= α * anchor score + (1- α) * anchor context score　(4-11)

计算每个子节点的预测相关度 potential score(child node)：

potentials core(child node)= β * inherited score(child node)+(1- β)* neighborhood score (child node)

计算结束每个链接节点的相关度值或者预测值之后，记入链接评价表中，供链接评价模块调用，进入链接的重要度 Importance(page)计算。

4.3.2 链接评价计算模型

由于收集到的信息用于数据挖掘，要求信息是非易失性，即数据保持不变，按计划添加新数据。为了能够不断捕获并追加的新的信息资源，在本系统中是根据网站的重要程度，来决定下一个链接的提取。

在计算网站的重要程度(用于赋予相应的页面优先级，优先提取重要度大的页面)时，改进了的 Hub Page 和 Authority Page 的计算方法，引入了相关度的评价值加权的方法，来区别链接的效用值，实现了权威站点的推送功能，辅助新资源的发现。

改进的 HITS 方法的对每个已访问的页面中的子节点计算其关于预测相关度的 Authority 权重和 Hub 权重，并以此决定页面中链接的访问顺序。设页面 P 的 Authority 权重和 Hub 权重分别为它们分别为 Authority (page)和 Hub (page)，按下述公式计算：

$$\text{Authority (page)} = \Sigma \text{Hub(link page)} * \text{potential score(link page)} \quad (4\text{-}10)$$

$$\text{Hub (page)} = \Sigma \text{Authority(link page)} * \text{potential score(link page)} \quad (4\text{-}11)$$

重要度计算公式：

$$\text{Importance (page)} = D * \text{Hub(site)} + (1-D) * \text{Authority(site)} \quad (4\text{-}12)$$

其中：D 为 Hub 特征和 Authority 特征的权重分配系数。

在本文中，将 Importance(site)作为度量网站重要性的标准，以便在更新数据库以及下载链接的选择过程中，能够提取相关度更大，使用价值更高的站点下载。同时，根据站点的重要程度，实现了权威网站的推送功能，启发新资源的发现。

第五章　实验过程与结果分析

在本章中，将简述本文中的技术验证系统工作流程，同时对实验结果进行分析。本文中主要用到的工具有两个：专题信息收集与过滤系统和 Access2Sql 数据转换工具。

系统开发运行硬件环境与软件环境：

硬件环境：CPU1.4GHz，256M 内存，80G 硬盘，2Mb/s ADSL

软件环境：Windows2000 professional SP4 系统平台，Visual Studio 6.0，DirectX 8.1 SDK，SQL Server2000，锐捷网络管理 Rejie 4.0。

§5.1　专题信息搜索与过滤系统实现与工作过程

本系统是一个专门为数据挖掘提供数据准备的工具，可以安装在互联网或者局域网等公共信息平台，在无人值守的状态下工作。由于考虑降低前台系统工作的硬件要求，前台使用了 Microsoft Access 数据库，最后由 Access2Sql 数据转换工具导入专业数据库 SQL Server2000。

专题信息搜索与过滤系统程序流程如下：

```
SearchandFilter()
{
    当前页面;
    页面解析器()
    {  净化信息;
       基于内容的文本过滤;
       N＝链接数;
       For (I =N;  I>0;  I--)
           {
              if (Url <控制深度)
               {  提取文本信息;
                  文本分类;
                  计算 Importance( site)值;
                  筛选链接进入写入地址表;
               }
           }
    }
}
```

本系统通过使用智能化的人机接口，既可以使高级用户自由的表达个人的兴趣专题，又可以辅助普通用户快捷的定制自己的专题，如图 5-1 所示。

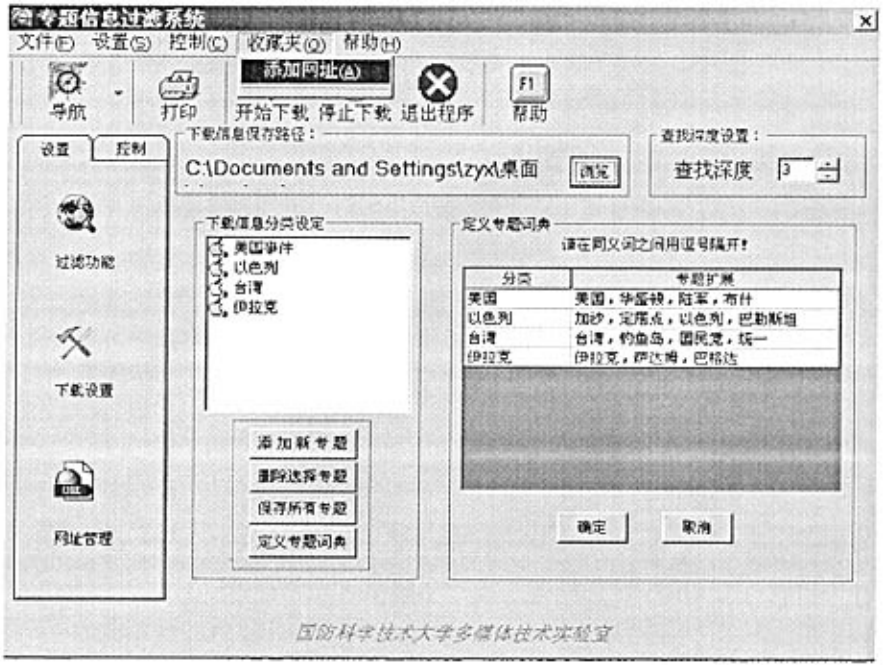

图 5-1　系统人机接口界面

系统在运行过程中，采用基于内容和基于链接评价相结合的过滤策略，运行过程如图所示。

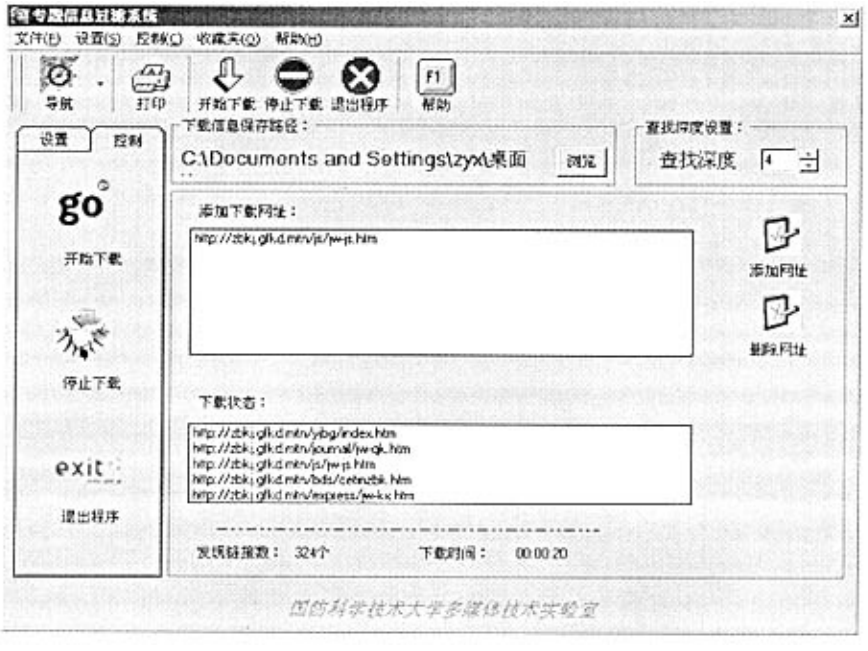

图 5-2　系统运行时界面

系统首先使用基于内容的搜索，可以保证初始信息与主题的相关度，防止搜索进入

迷盲状态；对新定位的页面使用基于评价的过滤策略，克服了传统的贪婪搜索策略的效率低下问题，既提高了搜索和下载的效率，又提高了收集的专题信息的相关度。

最后，由数据导入工具将信息搜索与过滤系统中的数据导入专业数据库中，完成数据的全部准备工作。数据导入工具的工作界面如图 5-3 所示。

图 5-3　数据导入工具的工作界面

§5.2　实验评价指标

检索评估可谓系统评估的基础，检索评估的目的在于了解检索结果满足信息需求的程度。通常信息需求的满足可以从很多不同的角度进行探讨，其中较常被提及的观点[52]包括：检索的质量、检索的效率、检索系统本身及检索者的检索技巧等。本文将从以下几个方面对信息收集与过滤系统的性能作评价分析：

一、查准率(precision ratio)

查准率的公式如下：

$$p = a / e \qquad (5\text{-}1)$$

其中，p 表示查准率，a 表示检出相关文章的篇数，e 表示所有检出文章的篇数。

大部分的学者对查准率都还算满意，但他们也发展，两次检索的查准率即使相同，所得到的相关数据笔数却可能差距很大。

为弥补上述现象所造成的问题，查到率成为第二个用来评估检索效益的测量值。一般而言，查到率是指相关文献被检出的比例，因此，系统评估不仅考虑到拒绝不相关文献的能力(查准率)，同时也测量系统找到所有相关文献的能力(查到率)。

二、查到率(recall ratio)

查到率的公式如下：

$$r = a / f \qquad (5\text{-}2)$$

其中，r 为查到率，a 为检出相关文章的篇数，f 表示所有相关文献的篇数。

查到率与查准率之间存在一种反比的关系，因此，如果 A 系统查准率高但查到率低，而 B 系统查准率低但查到率高，则很难判断系统性能的优劣。

三、噪声比

噪声比(noise ratio)也是评价检索效能较重要的测量值之一，又称为原子尘(fallout)或废弃物(discard)，其代表检索者不希望见到的现象，因此其比值自然越低越好。

噪声比的定义公式如下：

$$f = b / m \qquad (5\text{-}3)$$

其中，f 表示噪声比，m 表示页面中所有不相关文章的总数，b 表示检出不相关文章的笔数。

§5.3 实验与结果分析

5.3.1 实验专题定制

考虑信息的及时性和权威性，我们的实验数据主要来自于选择网易主页(www.163.com) 新闻板块（http://news.163.com），包括国际新闻、国内新闻、体育新闻以及部分广告内容。接下来的所有实验如无特殊声明，都是在这个数据集上进行的。选择的主题为 4 个，如（美国，台湾，伊拉克，以色列），新闻专题资料的收集的定制情况如表 5-1 所示。网页新闻（http://news.163.com）是一个综合性的新闻网站，信息内容丰富，资料齐全，所以选择该网站作为权威网址，进行信息的收集与过滤试验。

表 5-1 新闻专题及其关键字词列表

专题编号	专题名称	关键字词列表（以逗号分隔）
1	美国	美国，布什，华盛顿，陆军，白宫
2	伊拉克	伊拉克，萨达姆，巴格达，
3	以色列	以色列，加沙，定居点，巴勒斯坦
4	台湾	中国，台湾，陈水扁，连战

从开始下载到系统自动停止，即该网站的网页全部分析完毕，共运行 2543 秒，共分析网页 2647 个，分析出网址（包括图片与视频链接）37649 个，共收集资料 182314K。

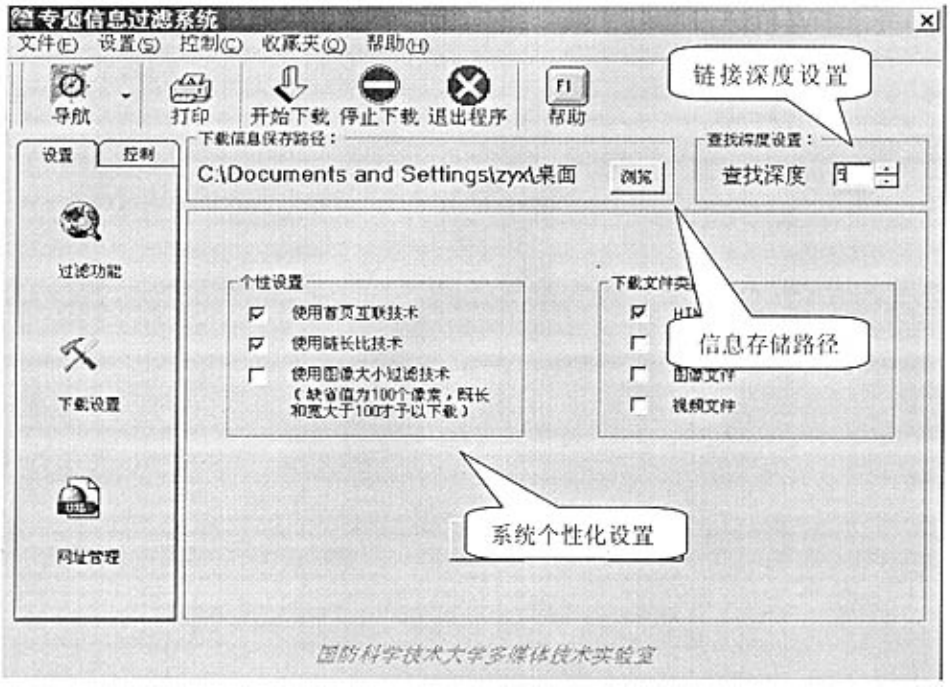

图 5-4　系统的相关功能设置

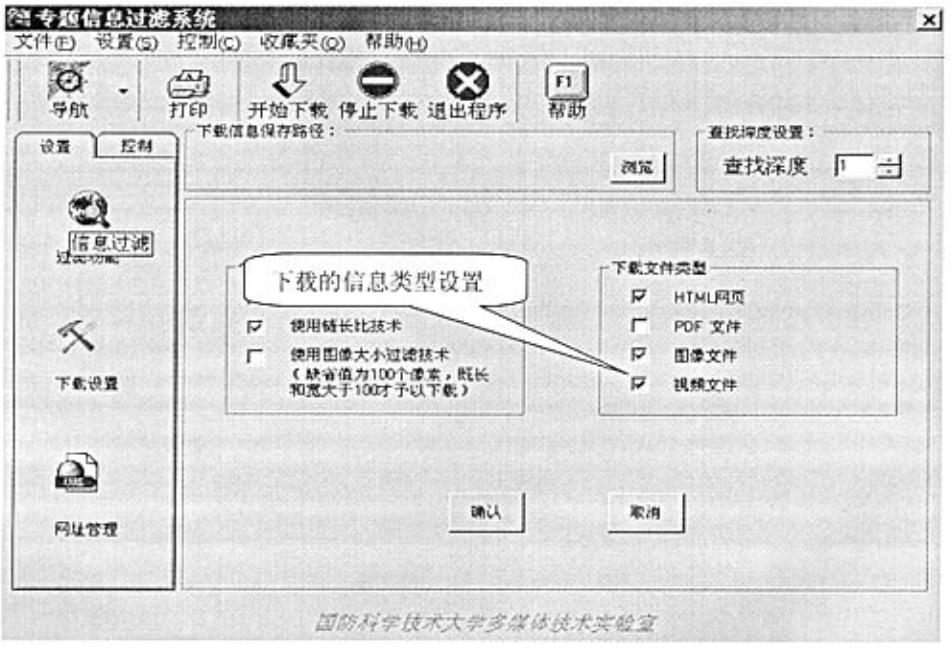

图 5-5　下载信息类型设置

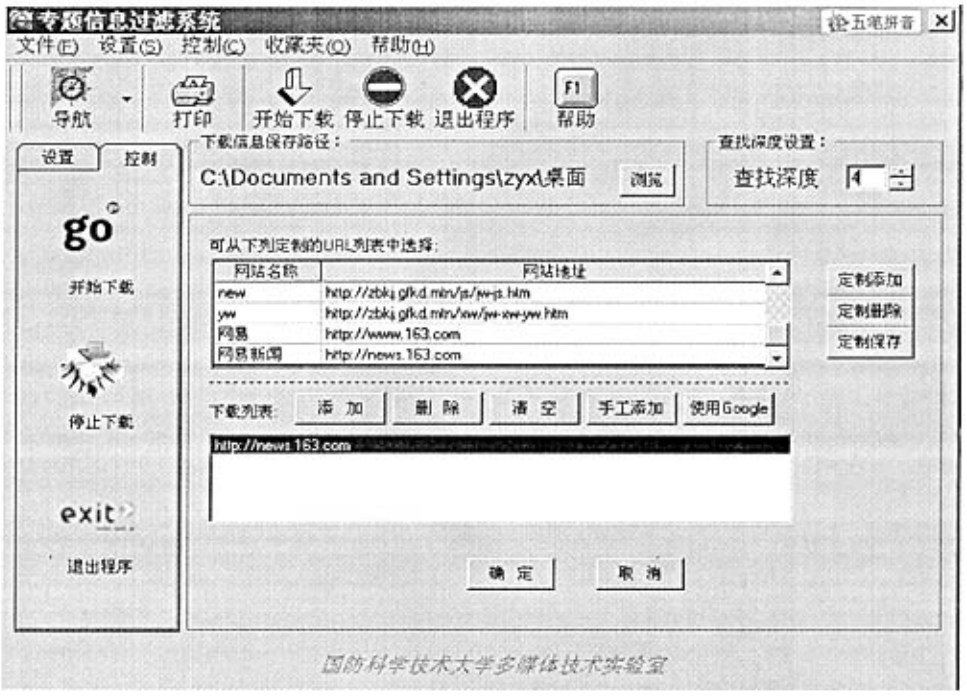

图 5-6 种子集设置

5.3.2 实验结果与分析

当链接种子集选择网易新闻 (http://news.163.com) 时共发现链接 711 个，由于本系统为专题信息收集工具，并未使用通用搜索引擎的“爬行器——标引器”机制，因此，为方便计算相关的评价指标，本文只对链接深度小于 2 时作了分析和跟踪。

当链接深度 depth = 1 时，该页面的链接共有 711 个，其中不相关广告链接 41 个。当链接深度 depth = 2 时，根据以上用户定制的专题，系统能够根据链接的评价最先取得网易新闻板块 (http://news.163.com)中的国际板块（http://news.163.com/world），于是本文也对这一链接的做了实验。对这两个链接的实验结果如下所示：

表 5-2 Depth=1、2 时实验结果

项目	数目（个）		大小（K 字节）		查准率（%）		查到率（%）		噪声比（%）	
链接深度	1	2	1	2	1	2	1	2	1	2
HTML	23	45	101	210	97	96	94	89	—	0.4
图片	11	21	510	1010	98	97	96	97	—	—
视频	1	2	13000	34000	100	100	100	100	—	—

从实验结果可以看出，当链接深度 depth =1 时，在查准率和查到率是较高的，噪声比极低；当 depth = 2 时，在同样的查准率的状态下，查到率有所下降，也出现了噪

声，对其结果分析发现：

在 depth = 2 时，收集相关文档的查到率下降的主要原因，我们发现在其次级链接中存在三个板块，16 个链接到国际实事论坛，其中有 2 个链接是我们定制的相关专题的信息。由于该论坛是动态网页，因此，本原型系统不能很好的识别，导致查到率下降。同时，由于互联网信息的发布没有统一严格的规范，尤其是各种论坛中的言论都比较自由，其信息可靠性和信息熵具有不确定性，这一部分信息的收录正在研究中。

当链接深度 depth = 4 时，收集专题的定制同上时，专题信息收集与过滤系统经过 267 秒完成收集下载任务。各专题的收集信息量如下所示：

表 5-3 Depth = 4 实验结果

专题类型	专题 1	专题 2	专题 3	专题 4
Html 文档	42	23	37	13
图片	25	19	21	5

以上数据是 2005 年 11 月 7 日的实验结果。由于实验前没有选择在校园网（互联网）下载数据，因此上述数据记录完全是在一次实验上收集的信息。当多次选择在同一种子集上进行实验时，系统会根据信息的更新时间和发布位置，追加新信息，而不是更新，以适应本系统收集的使用目的。

本专题搜索工具具有以下几个优点：

（1） 基于知识库和使用关键字词向量的专题搜索与过滤系统能保证收集资料的快速、准确、全面。无需训练的用户兴趣表达方式更加直接与方便，系统的设置灵活方便，也更加符合用户习惯；

（2） 支持多专题同时下载与处理，节省网络与计算资源，与通用搜索工具相比，搜索过滤下载实现了一体化，处理速度更快；

（3） 支持收集多种数据格式的资料和文件。通过结构挖掘与内容发现后，可以收集与主题内容相关的 HTML 网页、PDF 等其他类型的文档和图片、视频等多媒体信息；

（4） 由于实验数据的限制，在链接深度较小时，查准率和查到率都较高；当链接深度增加后，通过主观信息相关度评判，信息的相关密度仍然很高，可以满足使用的目的。

主要不足：

（1） 对非 HTML 类型的文档，分类的精确度不够。因为在分类中用到了 HTML 类型文档的结构特征，因此不能做到内容和权重分配的正确性；

（2） 用户提交或者知识库中的种子集的依赖程度比较大，种子集的选择会影响到所收集信息的数量和及时性。当用户新指定的种子集链接上的资料信息不够及时、不够全面或者该网站的主题与收集主题相距甚远，则重新发现、挖掘新的权威网站将影响信息的收集的效率；

（3） 在信息的收集过程中，虽然在不追求效率的情况下，对运行系统的硬件要求不高，但是资源独占性比较严重。

5.3.3 系统的整体操作性评价

为评价本文提出的基于专题的扩展查询功能的使用特性，也对系统的操作功能评

判结果如表 5-4 所示。

表 5-4 系统的操作功能

操作特性	深度控制	推送功能	导航设置	多信息类型支持	基于内容的多媒体分类
状况	有	有	有	有	无

第六章　结论及展望

§6.1　本文的主要研究工作

本文结合当前的国内外研究现状和当前的使用任务需要，对公共信息平台的专题信息收集和过滤技术做了一定的研究，主要工作和研究成果如下：

（1）根据互联网公开情报资料收集与检索的特点，研制开发出一个具有基于内容过滤和链接评价相结合的搜索与过滤工具。在实际的系统试验与检测中发现，该系统可以非常有效地在权威网站或者指定网站中自动收集相关内容信息。这对于在那些不提供站内搜索，或者站内搜索涵盖不全面的网站进行信息检索来说，其效果是显而易见的。

（2）为克服在搜索中存在的“漏搜”现象，最大限度的提高了搜索的查全率，提出了基于专题的查询扩展技术，用户可以在对该领域内的认知基础上，扩展查询关键词，组成检索向量。降低了单纯依靠有限个专题特征词的搜索过程中，大量有用信息由于不能和用户提供的关键词拟合而造成的信息丢失。

（3）提出基于深度控制的链接过滤方法和下载控制方法，使用户可以对下载的信息量进行间接的控制，搜集到相关度更大的专题信息，保证收集信息的全面性和广泛性，适应本专题的使用需要。

（4）改进了基于内容的文本过滤和基于链接结构相结合的过滤策略，有效的改善了检索中的主题“漂移”和搜索精度不高问题，提高了信息的下载精度。另外，它可以建立在综合搜索引擎之上，充分利用已经收集的大量的数据，在其上进行检索、挖掘等应用处理。

§6.2　今后的工作

由于时间和能力有限，本文只作了以上方面的研究，在研究中也感觉到还有以下等方面的内容有待于完善：

在系统的功能上还有多方面的问题待于完善：

（1）对网页类型的支持，目前由于受网页的生成机制的影响，本文研究的内容还不能支持动态生成的网页，虽然这部分动态的页面只占了整个互联网信息中的极少的一部分，但是这个动态页面发展的趋势，应该予以关注。比如：应该支持动态页面(*.aspx)的支持。

（2）由于本文研究的信息搜集主要集中在一些报道，具有新闻的部分特点，因此新闻组也是一个比较不错的信息来源。在下一步的研究中，应该把新闻组也引入，作为一个数据获取渠道。

（3）本文虽然能够根据站点的重要度，对用户有推送服务，但是在某一专题的智能化方面还不够，应该在这一方面作进一步的探讨，优化用户查询界面。新一代垂直

搜索引擎的用户界面应朝着智能化、个人化方向发展，向容错能力更强、灵活性和适应性更强，能够高效检索到合适信息的智能型用户界面发展。其目标是使用户界面具有文档分析与管理能力和更强的信息推送能力。

（4） 对部分类型的信息的分类工作做的研究还不够，如对多媒体数据，声音和图像，只能依靠链接中的结构信息分类，并不能对多媒体信息作基于内容的分类。

在技术方面，有以下几个方面需要做进一步深入研究：

（1） 完善专题词库的更新方法，加强扩展查询的智能化方法的研究。引入支持自然语言扩展查询，提供能充分表达用户查询要求的检索功能，使信息查询变得更加方便、快捷和准确。

（2） 在用户的兴趣发现方面，做一定的研究，能够从用户的使用习惯中捕捉有用信息，提供更全面的智能服务。

致谢

在我攻读硕士学位期间，自始至终得到了许多老师、同学的帮助，在此向他们表示深深的谢意。

首先要感谢我的导师吴玲达教授。她为我指明了研究的方向和前进的目标，使我在课题的研究过程中，个人的综合素质和独立科研能力得到了极大的提升。她的耐心、细致与自信，为我树立了很好的榜样，她的远见卓识和敏锐的意识也让我折服。当前随着网络的蓬勃发展，虽然有大量的搜索引擎存在，但是在本文所研究的专题自动收集下载的研究方面并不是很深入，开展相关研究的基础薄弱，难度很大，困难也很多。导师总能在百忙之中抽出时间来与我进行细致的交流，每次交流都能够使我有所收获，及时调整研究方法。同时，导师为我的课题研究提供了许多便利条件，帮助我排忧解难，使我的研究工作得以顺利进行。我的课题研究思路和成果，许多都直接源自于与导师的悉心指导。

在这里也要感谢谢毓湘博士和其他几位博士。自从我进入课题研究以来，他们为我提供了许多帮助。谢毓湘博士通过定期讨论对课题的系统理论研究和资料收集方面提供了非常多的帮助。栾悉道博士，韩智广博士和文军博士在分别在算法研究和编程等问题上为我提供了大量的帮助。正是有了他们的无私帮助，我才能在课题研究中克服种种困难，达到课题研究的预期目标。

感谢在实验室一起工作和学习的孙文广，孟国明和其他硕士同学，在这样一个同甘共苦的集体中学习，我感到非常温暖和愉快。最后要感谢所有关心我、帮助我的人，也要感谢教研室所有老师和同学提供的帮助，感谢父母及亲人对我的支持与鼓励。

再次谢谢大家！

参考文献表

[1] Josluis.Ambite, Craig A Knoblock, and Maria Muslea, "Conditional Constraint Networks for Interleaved Planning," IEEE Intelligence Systems, Vol 7,2005

[2] William Hersh and Jeffery Pentecost, "A Task-Oriented Approach to Information Retrieval Evaluation," Jan. 1996: 50-56.

[3] Anthony K.H. Tong, Raymond T. Ng, Laks V.S. Lakshmanan, Jiawei Han, Geo-Spatial Clustering with User-Specified Constraints, In Proc. Of the 1st International Workshop on Multimedia Data mining (MDM/KDD'2000), August 20, 2000,Boston, MA, USA. pp: 1~7

[4] BrinS. Extracting patterns and relations from the World Wide Web .In: Proc of Web DB Workshop at EDBT'98. Valencia, 1998

[5] Gudivada V N, Information retrieval on the World Wide Web, IEEE Internet Computing, 1997, 1(5)

[6] Duminda Wijesekera, Daniel Babara, Mining cinematic knowledge: Work in progress-An extended abstract, In Proc. Of the 1st International Workshop on Multimedia Data mining (MDM/KDD'2000), August 20, 2000, Boston, MA, USA. pp: 98~103

[7] Schapire.R , Singer,Y. BoosTexter: A boosting-based system for text categorization, Machine Learning, 2000,39(2/3):135~168

[8] Siirtola H. Direct Manipulation of Parallel Coordinates, Proc. of the IEEE International Conference on Information Visualization (IV-2000), London, 2000-07:373

[9] Martin Mcchalowshi, Josluis.Ambite, Snehal Thakark,and Rattapoon Tuchinda, "Retrieving and Semantically Integrating Heterogeneous Data from the Web," IEEE Intelligence Systems,Vol 19,No 4,2005

[10] Evern Sirin, Bijian Parsia, and James Hendler, "Filtering and Selecting Semantic Web Services with Interactive Compositon Technique," IEEE Intelligence Systems, Vol 19,No 3,2005

[11] Florescu D, Levy A, Mendelzon A, Database techniques for the World Wide Web: A survey. ACM SIGMOD Record 27:3,1998(9)

[12] Dreilinger D, Howe A E, Experiences with selecting search engine using metasearch, ACM Trans on Inf Sys,1997,15(3):195~222

[13] Fredirik Espinnoza, Kristina Hook, An interactive WWW interface to an adaptive information system, In: Proceedings of the Reality of Intelligent Interface Technology Workshop, Massachusetts: User Modeling Inc. 1997

[14] Marko Balabanovic, Agents'97 Marina del Rey CA USA, an Adaptive Web/Page Recommendation Service, New York: ACM Press, 1997

[15] Pokorny, J., Web searching and information retrieval, Computing in Science & Engineering, July-Aug. 2004, 43~48

[16] Lertnattee, V.; Theeramunkong, T., Multidimensional text classification for drug information, Information Technology in Biomedicine, IEEE Transactions on, Sept.

2004, 306~312

[17] Murata, T., Visualizing the structure of Web communities based on data acquired from a search engine, Industrial Electronics, IEEE Transactions on, Oct. 2003, 860~866

[18] Buntine, W.; Lofstrom, J.; Perkio, J.; Perttu, S.; Poroshin, V.; Silander, T.; Tirri, H.; Tuominen, A.; Tuulos, V., A Scalable Topic-Based Open Source Search Engine, Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on, 228~234

[19] Wang Liang; Guo Yiping; Fang Ming, Internet search engine evolution: the DRIS system, Communications Magazine, IEEE, Nov. 2004, 30

[20] Sciascio, E.D.; Donini, F.M.; Mongiello, M.; Piscitelli, G., Design and implementation of a Web-search engine based on computation tree logic, Electrotechnical Conference, 2004. MELECON 2004. Proceedings of the 12th IEEE Mediterranean, 2004, Vol.2, 705~708

[21] Chesnevar, C.I.; Maguitman, A.G., ArgueNet: an argument-based recommender system for solving Web search queries, Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference, 2004, Vol.1, 282~287

[22] Yuen, L.; Chang, M.; Lai, Y.K.; Chung Keung Poon, Excalibur: a personalized meta search engine, Computer Software and Applications Conference, 2004. COMPSAC 2004. Proceedings of the 28th Annual International, 2004, vol.2 , 49~50

[23] Nikravesh, M., Web intelligence: conceptual search engine and navigation, Industrial Informatics, 2003. INDIN 2003. Proceedings. IEEE International Conference on, Aug. 2003, 390~395

[24] Miyakawa, A.; Sugita, K.; Ishida, T.; Shibata, Y., Implementation and evaluation of a tradition search engine using sensitivity searching method, Advanced Information Networking and Applications, 2004. AINA 2004. 18th International Conference, 2004, Vol.1, 630~635

[25] Sato, N.; Udagawa, M.; Uehara, M.; Sakai, Y, Searching restricted documents in a cooperative search engine, Distributed Computing Systems Workshops, 2004. Proceedings. 24th International Conference, March 2004, 44~49

[26] Risvik, K.M.; Aasheim, Y.; Lidal, M., Multi-tier architecture for Web search engines, Web Congress, 2003. Proceedings. First Latin American, Nov. 2003, 132~143

[27] Maleki-Dizaji, S.; Othman, Z.A.; Nyongesa, H.O.; Siddiqi, J, Evolutionary reinforcement of user models in an adaptive search engine, Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference, Oct. 2003, 706~709

[28] Yan Li; Xin-Zhong Chen; Bing-Ru Yang, Research on Web mining-based intelligent search engine, Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference, Nov. 2002, 386~389

[29] NikRavesh, M., Fuzzy conceptual-based search engine using conceptual semantic indexing, Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American, June 2002,146~151

[30] Takagi, T.; Tajima, M., Query expansion using conceptual fuzzy sets for search engine, Fuzzy Systems, 2001. The 10th IEEE International Conference, Dec. 2001, 1303~1308

[31] Zhang, W.; Xu, B.; Yang, H., Development of a self-adaptive Web search engine, Web Site Evolution, 2001. Proceedings. 3rd International Workshop, Nov. 2001, 86~93

[32] Baowen Xu; Weifeng Zhang; Hongji Yang; Chu, W.C., A rough set based self-adaptive

Web search engine, Computer Software and Applications Conference, 2001. COMPSAC 2001. 25th Annual International, Oct. 2001, 377~382

[33] Kyung-Joong Kim; Sung-Bae Cho, A personalized Web search engine using fuzzy concept network with link structure, IFSA World Congress and 20th NAFIPS International Conference, 2001, vol.1, 81~86

[34] 欧阳柳波，李学勇，李国徽，王鑫，专业搜索引擎搜索策略综述，计算机工程，2004,7（13），32～46

[35] 王灏，黄厚宽，田盛丰，文本分类实现技术，广西师范大学学报（自然科学版），2003，3（21）

[36] 曾楠，俞宁，张艳伟，张连发，利用智能搜索引擎构建网上知识平台，计算机应用研究，2004，2

[37] 王会进，陈超华，李清，基于动态知识库搜索引擎的技术，暨南大学学报（自然科学版），2004，2，25

[38] 陶兰，杨睿，李四明，WMS，面向领域的 Web 的信息挖掘系统，小型微型计算机系统，2004，4

[39] 李名智，中文搜索引擎发展的现状问题及对策，中国信息导报，No.2，1999（30～32）

[40] 智能搜索引擎关键技术研究与实现，哈尔滨工程大学硕士学位论文，中国知识基础设施工程

[41] 杜阿宁，方滨兴，胡铭曾，云晓春，中文交互式网络搜索引擎及其自学习能力，计算机工程与应用，2003，10

[42] 孙建涛，沈科，陆玉昌，石纯一，网页分类技术，清华大学学报（自然科学版），2004，Vol. 44，No.1

[43] 张茂元，卢正鼎，基于特征选取及模糊学习的网页分类方法研究，小型微型计算机系统，2004，7，Vol.25, No.7

[44] 潘春华，武港山，面向主题的 Web 信息收集系统的设计与实现，小型微型计算机系统，2003，12，Vol.24 No.12

[45] 张礼东，汪东升，郑纬民，基于 VSM 的中文文本分类系统的设计与实现，清华大学学报（自然科学版），2003.Vol.43，No.9

[46] 李雪蕾，张冬茉，一种基于向量空间模型的文本分类方法，计算机工程，2003，Vol.29，No.1

[47] 武旭，须德，基于向量空间模型的文本自动分类系统的研究与实现，北方交通大学学报，2003，Vol.27，No.2

[48] 成奋华，吴家强，数字图书馆基于向量空间模型的文档分类系统，情报技术，2004，7

[49] 徐德智，吴敏，陆文彦，基于 Agent 的专业搜索引擎的研究和构造，计算机工程，2002，28(10)，99~101

[50] 汪晓岩，胡庆生，李斌，庄镇泉，面向 Internet 的个性化智能信息检索，计算机研究与发展，1999，36(9)，1039~1046

[51] 栾悉道，互联网公开情报收集与处理技术研究，国防科学技术大学研究生院硕士学位论文，2001

[52] 徐德智，吴敏，陆文彦，基于 Agent 的专业搜索引擎的研究和构造，计算机工程，2002，28(10)，99~101

[53] 董建设，基于 HTML 标记分析及中文切词的网页索引研究与实现，兰州理工大学，2003

[54] 陈福集，杨善林，一种基于 SOM 的中文 Web 文档层次聚类方法，情报学报，2002，4（2），174~178

[55] 王建会，王洪伟，申展，胡运发，一种实用高效的文本分类算法，计算机研究与发展，2005，42（1），85～93

[56] 张健滈，刘洋，杨静，代坤，搜索引擎结果聚类算法研究，计算机工程，2004，3（5），95～97

[57] 陈彤兵，汪保友，胡金化，施伯乐，一个实时搜索引擎的设计，小型微型计算机系统，2003，5（5），856～858

[58] 凤元杰，刘正春，王坚毅，搜索引擎主要性能评价指标体系研究，情报学报，2004，2（1），64～69

附录 A　作者发表的论文

[1] 张玉新，吴玲达，谢毓湘，栾悉道，一种基于小波变换与分形编码的新闻图片检索方法，计算机应用研究，2005 年第 2 期

[2] 张玉新，吴玲达，谢毓湘，孙文广，面向专题的信息搜索过滤系统设计与实现，国防科学技术大学第 5 届研究生学术活动节论文集

[3] 孙文广，吴玲达，宋汉辰，张玉新，基于红外卫星云图的云的三维表现，微计算机信息，已录用，2006，7

附录 B　作者参加的科研项目

[1]　国家 863 高技术项目，战略多媒体情报收集、处理与态势表现构件技术研究
[2]　“十五”国防预研项目，军用多媒体数据内容处理与管理技术研究