网络多媒体教育资源主题搜索算法研究

摘要

网络多媒体教育资源是指存在于 Internet 中的多媒体教学资源。随着网络与多媒体技术的发展，Web 中的多媒体教学资源，尤其是音频、视频和动画，也日益丰富，成为教育领域的重要组成部分。如何快速、准确地找到特定主题的多媒体教学资源，使其在信息化教育中充分发挥作用，是教育技术工作者亟待解决的问题，也使得传统的搜索引擎面临着巨大的挑战，各类多媒体搜索引擎也随之应运而生并受到广泛的关注。主题搜索器的搜索算法，是搜索引擎的核心，它决定了搜索引擎的搜索效率和质量。本文从主题网页和包含多媒体的网页在 Web 中分布的特点出发，围绕提高多媒体主题搜索效率的问题，提出一种基于 URL 链接规则的多媒体主题搜索算法。

本文首先介绍了在 HTML 文档中，与多媒体资源有关的文本信息和 HTML 标签，深入分析了主题页面和包含多媒体资源的网页在 Web 上的分布特征。在分析和比较现有的主题搜索算法的优缺点的基础上，归纳了提高搜索效率的几个关键因素。

本文对主题搜索领域中传统的主题搜索算法进行了详细的分析，在深入分析主题页面在 Web 上的分布特征与主题相关性判别算法的基础上，对 PageRank、Shark-Search 两种典型的主题搜索算法进行相关参数的改进，同时将网页内容相似度和网页链接相似度加入到算法的计算过程中，并将改进后的 Topic-PageRank 算法、Shark-Search 算法应用于多媒体主题搜索领域。

针对传统的主题搜索算法不能简单的应用于多媒体主题搜索领域的问题，本文基于包含多媒体的网页往往呈现出“资源相邻性”的特点，提出了一种基于URL链接规则的多媒体主题搜索算法，即从种子网站列表中自动学习出代表“多媒体资源区域”URL正则表达式，并用这些正则表达式来指导主题搜索器对网页的抓取。在详细介绍URL数据结构、URL距离的度量以及URL正则表达式的学习和指导过程的基础上，对PageRank算法进行改进，实现基于URL链接规则算法在链接方面的相似度计算。

为了验证基于 URL 链接规则多媒体主题搜索算法的高效性，本文采用统一的系统体系结构和软、硬件平台，对 Topic-PageRank 算法、改进的 Shark-Search 算法和基于 URL 链接规则的算法输入相同的种子页面集、限定同一搜索深度进行实验测试。文章从查全率和查准率两个角度，对三种算法进行比较。实验结果表明本文的工作是相当

有效的，尤其是提出的基于URL链接规则的多媒体主题搜索算法，具有相当的创新性和实际应用价值。

【关键字】 Web多媒体；主题搜索器；主题搜索算法

【分类号】 G434

Research on Network Multimedia Educational Resources Topic Search Algorithm

ABSTRACT

Network multimedia education resources refer to the multimedia teaching resources on the Internet. As the development of network and multimedia technology, Web multimedia teaching resources, especially audio, video as well as animation, are also increasingly rich and become an important part in the field of education. How to quickly and accurately find the specific topic multimedia teaching resources, and enable them to play the role in educational informationization sufficiently, is the urgent problem for educational technology workers, and also a great challenge to the traditional search engines. A variety of multimedia search engines emerged as the times required and got widespread concern. The algorithm is the core of the topic search engine and determines the search efficiency and quality. According to the characteristics of the distribution of topic page and multimedia page on the Web, the paper presents a multimedia topic search algorithm based on URL link rules that focuses on the problem of improving topic search efficiency.

The paper firstly introduce the text information and HTML tags related to multimedia resources in HTML documents, then make an deep analysis of the distribution characteristics of the topic page and multimedia page in the Web. Based on the analysis and comparison of the advantages and disadvantages of the existing topic search algorithms, the key factors of improving search efficiency are presented.

The traditional topic search algorithms in the topic search field get detailed analysis. Based on the deep analysis of the distribution characteristics of topic page in the Web and the topic relevance discrimination algorithm, the related parameters of the two typical topic search algorithms PageRank and Shark-Search are improved. Meanwhile, the web content similarity and the web links similarity are added to the calculation process of the algorithm. Then, the improved Topic-PageRank and Shark-Search algorithms are applied in multimedia topic search field.

Since traditional topic search algorithms can't be applied in multimedia topic search field, the paper presents a multimedia topic search algorithm based on URL link rules under consideration with the "resource adjacency" of the multimedia web. The new algorithm can learn automatically the URL regular expression on behalf of the "multimedia resource area" from the seed web list, and then use these regular expressions to instruct the topic search engine to capture pages. In the light of the detailed introduction on URL data structure, URL distance's measurement, and the study of URL regular expression, the PageRank algorithm is improved, eventually the link similarity is calculated based on the URL link rules algorithm.

To validate the high efficiency of multimedia topic search algorithm that based on URL link rules, this paper adopts a unified system structure and software and hardware platforms; inputs the same seed pages and limits the same search depth to conduct experiments for the Topic-PageRank, improved Shark-Search and URL-based algorithms. From the two aspects of recall and precision ratio, the three algorithms are compared. The experimental results show that our work is quite effective, especially the URL-based multimedia topic search algorithm which has considerable innovation and practical application value.

[Key Words] Web Multimedia; Topic Searcher; Topic Search Algorithm

[Category] G434

独 创 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得_____（注：如没有其他需要特别声明的，本栏可空）或其他教育机构的学位或证书使用过的材料。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。

学位论文作者签名：杨允广　　导师签字：

学位论文版权使用授权书

本学位论文作者完全了解 学校 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 学校 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。（保密的学位论文在解密后适用本授权书）

学位论文作者签名：杨允广　　导师签字：

签字日期：2009 年 4月 10日　　签字日期：2009 年 4月10日

第一章 绪论

1.1 引言

多媒体是综合性的信息资源，是文本（Text）、图形（Graphic）、声音（Sound）、动画（Animation）、视频（Video）等媒体元素的统称[1]。随着计算机技术的发展和Internet应用的普及，基础教育教学资源的信息化建设也得到了蓬勃的发展。随着各种类型教学资源专业网站的不断涌现，Web已经成为人们获取教学资源的主要途径。根据CNNIC发布的《第19次中国互联网络发展状况统计报告》，截至2006年底，中国网站的网页数量为44.7亿[2]，其中文本和图像仍然是网页最主要的内容形式，分别占据70.2%和29.5%的比例；视频网页占网页总数的0.3%。而按照多媒体格式分类：swf格式的网页占网页总数的1%，mp3格式的网页占网页总数的0.1%。同时根据CNNIC2008年8月发布的报告显示：目前中国网站数量已达191.9万个，年增长率达到46.3%，而且继续保持快速增长的势头。

美国教育传播技术协会（AECT）[3]90年代以后发布了2个不同的定义：AECT94定义认为："教育技术是为了促进学习，对有关的过程和资源进行设计、开发、利用、管理和评价的理论与实践"[4,5]。2004年，AECT又发布了对教育技术的新界定，对于该界定，彭邵东教授将其翻译为："教育技术是通过创造、使用、管理适当的技术过程和资源，促进学习和改善绩效的研究与符合道德规范的实践"[6]。从以上教育技术的定义中不难看出，对教育相关资源的利用和管理是教育技术的一个重要研究范畴，而多媒体资源是教育资源的重要组成部分。对于学习环境的构建和学习者的学习具有非常重要的作用。

但随着网站的增多，教学资源数量呈几何级数的增长，要想从Web上获取一条想要的多媒体资源的难度越来越大。目前网站教学资源都是以半结构化（Semi-structured）[7,8]的数据形式存在，大多数网页是通过HTML[9]语言来展现的，而HTML语言的一个显著特点是结构隐含、不规则或不完整。结果使得这种网站上的多媒体资源处于杂乱无序的状态，数据集成性非常差、应用程序无法直接解析、获取并利用Web多媒体资源困难，从而给教育教学资源信息化建设造成了极大的困难。

Internet具有海量、异构、动态变化等特性，使得用户试图通过浏览网页来发现所需要的多媒体资源已经变得越来越困难。如何能够实现从互联网自动的或者人尽量少

的参与下搜索到Web中的多媒体资源，并对搜索到的多媒体资源进行自动分类，然后对它们进行特征提取以确定多媒体资源的主题内容，这些工作都具有重要的现实意义。

基于此，论文立题研究网络多媒体教育资源主题搜索算法，根据包含多媒体的网页在Internet上分布的特征，通过一定的算法在尽可能少的耗费计算机资源的情况下快速地搜索到多媒体网页，从多媒体所在的网页中提取出多媒体相关文本，再从相关文本中提取出用于描述、标引Web多媒体的语义信息，从而确定多媒体的主题。将此方法应用到基础教育资源中，这对于学习环境的构建、教育资源信息化的建设具有重要的意义。

1.2 网络多媒体资源搜索的研究现状

目前，因特网上图形、图像、视频、音频、动画等多媒体信息日益丰富，各种基于网络的多媒体搜索引擎随之应运而生。根据它们的工作原理和方式的不同，主要分为两种类型：基于文本描述的多媒体搜索引擎和基于内容的多媒体搜索引擎。

基于文本描述的多媒体搜索引擎，这种方法主要是对含有多媒体资源的网站和网页进行分析，对多媒体信息的物理特征和内容特征进行著录和标引，把它们转换成文本信息或者添加文本说明，对这些文本信息建立数据库，索引时主要在此数据库中进行关键词的匹配。可以通过文件扩展名来确定多媒体的类型，如：图像文件常用.gif和.jpg作为扩展名，声音文件常用.mid、.wav、.mp3等作为扩展名，视频文件常用.avi、.mov、.mpeg、.rm、.rmvb作为扩展名。在多媒体搜索引擎中还可以利用超文本标识符来确定多媒体资源的类型，如：<Img src>和<Href>两个超文本标识符可以用来检测是否存在可显示的图像文件。嵌套在网页中的多媒体信息，在网页中往往带有与内容相关的标题或文本描述，这些信息在多媒体搜索过程中起着表征多媒体内容的作用。

目前，网上的大部分多媒体搜索引擎都属于此类。视频搜索引擎比较著名的有：优酷、六间房、百度视频、Google视频、土豆网等；音频搜索引擎比较著名的有：百度MP3、、爱听音乐、搜狗音乐、雅虎音乐等；动画搜索引擎比较著名的有：闪吧、小破孩、闪客帝国、TOM-Flash动画等。

基于内容的多媒体搜索引擎[10]一般包括两部分，数据库生成系统和查询子系统。具体而言，就是多媒体信息标引系统和检索系统。标引系统的任务是完成对多媒体的预处理和提取特征，然后建立起多媒体信息数据库系统，该数据库系统包括信息库、特征库和知识库。检索系统则负责对用户提交的多媒体信息进行特征提取，然后检索

多媒体特征库，并将用户要求最相似的信息输出出来。它和基于文本描述的搜索引擎最重要的区别是：基于内容的多媒体搜索引擎对多媒体的内容特征描述进行查询，这些多媒体信息内容特征的描述主要包括：图像的颜色、纹理、形状等；声音的音频、响度、频度和音色等；视频的视频特征、运动特征等。

基于内容的多媒体搜索引擎目前还不多见，常见的主要有以下几种：

（1）QBIC[11]（Query By Image Content）。QBIC（http://www.qbic.almaden.com）是20世纪90年代IBM公司研制的，它是标准的基于内容特征的检索系统。QBIC提供的检索途径有：利用系统提供的标准范围，用户自己输入图像、简图或者影像片段。

（2）上海交通大学的音乐数据库检索系统[12]，是为数不多的基于内容的声音检索系统。它不但提供了基于文本描述的声音检索系统的演奏者、作曲者、曲名、主题类别外，而且还提供了乐句和全曲作为检索途径。乐句作为乐曲的主题词，以简谱作为表现形式。检索时，将输入的字符序列和音乐数据库的字符序列相匹配。在乐句检索中，可以只输入乐句简谱的音高部分，而不用输入时值。在检索中，也可以采用前截或后截词检索，用“*”表示，如“*2323562*”表示前后截词。由于音乐的演奏形式会经常发生变化，如果检索者对旋律的记忆不很准确，这时就需要模糊检索功能，这是相当重要的。

（3）哥伦比亚大学研究的VideoQ是一种全自动面向对象基于内容的视频检索系统[13]。它对基于关键字或主题浏览的传统检索方式进行了拓展，同时提出了全新的基于视觉特征和时空关系的查询技术。此系统能自动切分并跟踪视频中任意形状的对象，并且提供包括颜色、纹理、形状和运动在内的丰富视觉特征库。目前，VideoQ视频库有超过3000段的视频，每段都被压缩成3层结构保持，可通过WWW互联网交互查询和浏览。

1.3 多媒体主题搜索相关技术介绍

1.3.1 主题搜索引擎

主题搜索引擎[14]以查询和检索某一专业领域或学科领域的因特网信息资源为目的，在互联网上智能的搜索符合特定查询主题的Internet信息资源。主题搜索引擎和通用搜索引擎在工作原理上是一样的，主题搜索器是其核心组成部分。不同的是，在主题搜索引擎中主题搜索器在进行网络信息采集时采用主题式搜索策略，按照管理员预先设定的主题去采集网上的相关信息，这样可以减少被采集的信息数量，提高索引数

据库中的信息质量。

主题搜索引擎一般由搜索器、索引器、检索器和用户接口四个基本部分组成。其基本结构如图1-1。

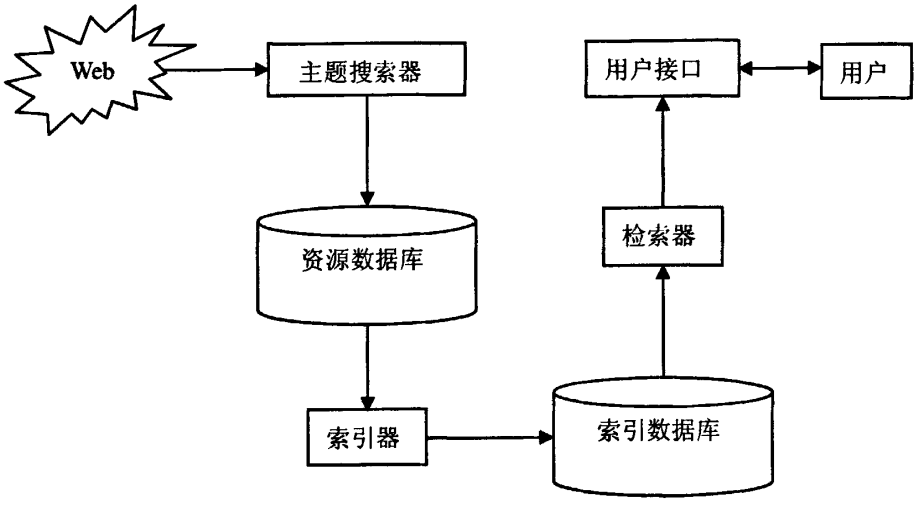

图 1-1 主题搜索引擎的基本结构

1. 主题搜索器

主题搜索器是一种网络资源发现与收集程序，通常从一个“种子集”（如用户查询、种子链接或种子页面）出发，通过HTTP等协议请求并下载网络资源，分析资源并提取链接，然后再以循环迭代的方式访问网络[15,16]。主要负责HTML页面的解析、爬行页面的选择和主题搜索算法的计算等工作。

主题搜索算法是主题搜索器的核心，它也是区别于通用搜索引擎的关键部分，决定了一个主题搜索引擎搜索的效率和返回结果的准确度。目前，在该领域很多专家、学者从理论和实践上做了很多研究工作，提出了许多主题搜索算法，包括：以Shark-Search[17]和Best-Fish[18]为代表的基于内容评价的搜索策略；以PageRank[19]和HITS[20]为代表的基于链接结构评价的搜索策略。文献[21]为了增强主题搜索的自适应能力，将巩固学习策略在预测远期回报的优势加入到搜索的学习过程中，用来预测待搜索链接未来回报价值。文献[22]在巩固学习的基础上，通过构建典型的“Web语境图”策略来估计目标页面的距离，加强了主题搜索器的自适应和增量反馈能力。

2. 索引器

索引器的作用是对主题搜索器采集来的网页进行处理，抽取出网页的索引项，然

后编制索引入库以备用户检索。网页处理的内容主要包括以下几个部分：文档特征提取、网页筛选、标引、相关度分析、归类和最后的入库、生成文档库的索引表。索引表一般使用某种形式的倒排序表（Inversion List）[23]，即由索引项查找相应的文档。索引表也可能要记录索引项在文档中出现的位置，以便检索器计算出索引项之间的相邻或接近关系（Proximity）。索引器可以使用集中索引算法或分布式索引算法进行索引，索引算法对索引器的性能有很大影响。

3. 检索器

检索器是根据用户的查询内容在索引库中快速检索出所需的资源，并对用户的查询主题与索引库中的相关资源进行相关度的评价，然后根据相关度的高低，通过用户接口返回给用户。其中检索算法、信息查询和组织方式都会在很大程度上影响检索器的性能。

4. 用户接口

用户接口的作用是输入用户查询内容、显示查询结果和提供用户交互反馈机制。主要目的是方便用户使用搜索引擎，能够高效率、多方式地从搜索引擎中得到有效、及时的信息。

1.3.2 多媒体资源的文件格式

主题搜索器在进行网页信息的采集过程中，通过分析嵌入到网页中的多媒体文件格式来判断多媒体的类型，然后获取多媒体资源的链接路径，并将其存入到多媒体数据库中。目前Internet上存在的网页大多以文本和图像形式存在，所以在搜索多媒体资源的时候图像暂时没有考虑。本论文搜集的多媒体主要包括三类：音频、视频、动画。它们的格式如下：

音频：mp3、flac、wav、wma、midi、rm、au。其中以mp3、wav、midi、wma、这四种格式为主。

视频：avi、mpeg、rmvb、wmv、mov、asf、mpe、mpg、vob、dat、mlv、m2v、divx。其中以avi、mpeg、wmv、rmvb、asf、mov这六种格式为主。

动画：swf。

1.3.3 主题选择

在进行主题搜索的时候，首先要弄明白一个问题——什么是主题。针对不同主题的搜索系统，必须进行有效的主题选择，这样才能搜索到我们真正需要的主题页面。

一个主题就是一个“含义”，也可以叫做一个“概念”，它可以是一个词，也可以是一个短语，甚至是一个段落或者一篇文章。这个“概念”的范围可大可小，大的时候非常广泛，但同时意义也非常模糊；小的时候非常狭义，但这时它的意义却非常具体。因此系统在设计时需要考虑到主题变化的方便性。

搜索引擎返回结果的用户满意度在很大程度上决定于关键词的选择，现实中的主题范围太广泛，有些主题没有实际用途上的意义（“假如，并且”等）；而有些主题却又不能引起人们搜索的兴趣（“走”等）。为此，在设计系统时有必要对主题进行统一的分类，这有利于主题搜索系统从合适的主题范围和主题角度进行搜索。目前很多搜索系统主题的选择采用Yahoo主题分类目录，也有选择其它分类目录的，但所选择的分类目录必须是分类比较合理，同时具有一定的权威性。本论文主要研究教育领域的多媒体资源搜索，对教育主题的选择将在5.1.1节中进行详细介绍。

1.4 本文的主要工作和组织结构

主题搜索算法是目前主题搜索领域研究的热点，同样也是难点，但在多媒体主题搜索领域相关的研究还很少，简单地将主题搜索算法运用于多媒体资源的主题搜索过程中，并不能取得良好的效果。本文首先分析了多媒体资源在网页中分布的特点，在此基础上对传统的主题搜索算法进行相关参数的改进，使之应用于网络多媒体资源主题搜索领域中。同时，提出一种基于URL链接规则的多媒体主题搜索算法，通过实验结果验证基于URL链接规则的多媒体主题搜索算法，更能有效地提高多媒体资源主题搜索的查准率和查全率。

1.4.1 本文的工作

本论文以网络上包含教育多媒体资源的网页为研究对象，在已有主题搜索器的基础上，重点研究多媒体主题搜索算法，用于提高网络多媒体资源的主题搜索效率。主要工作包括以下几个方面：

1. 深入分析包含多媒体资源的主题页面在Web上的分布特征，通过采用通用搜索算法对种子网站进行搜索实验的结果分析，提出包含多媒体资源的网页在Web存在着“资源相邻性”的特点。

2. 对传统的主题搜索算法进行分析，同时针对包含多媒体资源的网页在Web中分布的特点，对Topic-PageRank算法、Shark-Search算法进行相关参数的改进，使它们应用于多媒体资源的主题搜索领域，并详细介绍了这两种算法的改进过程。

3. 根据包含多媒体资源的网页在Web中呈现出“资源相邻性”的特点，提出一种基于URL链接规则的多媒体资源主题搜索算法，并详细介绍了算法的实现细节和过程。

4. 在同一软、硬件平台上，在已有的主题搜索器系统上，对Topic-PageRank算法、改进Shark-Search算法、基于URL链接规则算法进行测试，通过实验结果证明基于URL链接规则的多媒体主题搜索算法，在多媒体资源主题搜索领域具有较高的查准率和查全率。

1.4.2 本文的组织结构

本文共分六章：

第一章 绪论。主要介绍了多媒体资源主题搜索的现状和相关技术，同时也对多媒体资源的文件格式、多媒体资源主题的确定做了简单的介绍。

第二章 网络多媒体Web页面主题信息提取。本章首先介绍了HTML文档中与多媒体资源有关的文本信息，然后详细介绍了HTML文档中能够表征多媒体资源信息的相关标签，最后对主题页面在Web中分布的特征进行了简单的阐述。

第三章 主题搜索算法研究。本章对目前主流的几种主题搜索算法进行了详细的介绍，结合多媒体资源在Web中分布的特点，选择PageRank算法和Shark-Search算法进行相关参数的改进，并详细介绍了算法改进的过程。

第四章 基于URL链接规则的多媒体主题搜索算法。本章详细介绍了HTML解析器的实现过程和文本信息处理、超链接分析和处理的关键技术，为多媒体主题搜索算法的设计打下了基础。通过实验验证了包含多媒体网页的“资源相邻性”特点，详细介绍了URL数据结构、URL距离的度量以及URL正则表达式的学习和指导过程，对PageRank算法进行改进，提出一种基于URL链接规则的多媒体主题搜索算法。

第五章 实验系统与测试分析。本章我们给出了系统的实现平台以及三种算法的测试结果，从查准率和查全率两个方面对三种算法进行比较，并对实验结果进行分析。

第二章 网络多媒体 Web 页面主题信息提取

目前大部分网页采用半结构化的HTML（HyperText Mark-up Language)语言编写，主题搜索器搜索页面时遇到的首要问题就是解析这些HTML页面，并且提取出HTML页面中包含的结构信息和内容信息。一般的，一篇HTML文档主要由文本、标签、注释这三部分组成。文本是指我们在HTML文档中看到的词句，除了脚本语言和注释，HTML文档中的所有数据，只要不是标签的组成部分，我们都可以将其认为是文本。文本是格式化的，并且由嵌套它的标签控制，如：<Title>基础教育资源网</Title>。处于不同标签内的文本传递的信息也是不同的，利用标签在网页中的重要程度，我们可以确定此文本对网页主题的贡献率。

2.1 HTML 简介

HTML是超文本标记语言的缩写，由Time Berbers-Lee在1900年提出，它是用于书写超文本文档的语言规范，目前它已经成为Web上的一种通用的描述语言。HTML用描述性的标签（或称为标记）来指明文档的不同内容，把HTML文档划分成不同的逻辑结构，如段落、标题和表格等。HTML标签有两个作用：一是定义文档结构，便于浏览器显示该文档；二是提供各种链接，把Web搜索器的搜索程序引导到该文档的关键区域。在其数年的演变过程中，HTML基于需求不断地更新，1997年W3C推出了HTML4.0作为推荐规范。

2.1.1 HTML 解析器

在主题搜索器中，HTML解析器的主要功能是从HTML源文件中分析出HTML文档中的链接结构，提取出其中包含的链接地址，使主题搜索器能够按照链接地址继续向前搜索，HTML4.01规范中可以看到HTML文档主要有如下四种链接结构：<A>（Anchor）tags、Img（Image）tags、<Map>and<Area>tags和<Frame>and<Iframe>tags。这四种链接结构都有很多属性与之相联系，具体属性描述略。

HTML解析的过程是将HTML文档的流式数据结构化的过程。根据HTML的语法定义，依次对输入的HTML文档词法和语法进行分析，其中词法分析是为了从字符流中识别出有意义的符号，这些符号是HTML语法的最小单位，包括标签、无标签文本单词、注释以及处理指令等。同时，HTML解析器依据HTML语法确定这些符号之间的关系，

如标签与标签之间的层次包含关系等。最后，HTML解析器将解析的结果以语法书的形式输出。

目前有公司和个人开发的HTML解析器非常多，如W3C提供的Ctidy软件以及IBM开发的Xerces等。然而，不同用途的主题搜索器对于HTML解析器在性能和功能上有不同的特殊要求。首先，Web文档书写的不规范造成了许多Web文档不合法，为保证解析过程的顺畅进行，必须特别考虑HTML解析器的容错性。其次，为了适应Web搜索器处理的需要，需要提供专用的访问接口（如访问页面链接的接口）。

2.1.2 HTML 文档中多媒体相关文本信息

HTML文档中包含很多信息，本文在对HTML页面进行解析时，只提取有助于揭示页面主题和多媒体主题的部分信息，包括：（1）元信息；（2）网页标题；（3）链接内容信息，包括URL地址信息和链接锚文本信息；（4）链接结构信息，包括父链接的信息和兄弟链接的信息；（5）在第三类、第四类信息基础上得到的信息，包括网页的物理层、逻辑层、网页的链接数量、网页是否包含多媒体、某网页链接兄弟链接的数量、某网页链接包含多媒体的兄弟链接的数量、网页链接的锚文本路径信息，网页链接的URL路径信息以及结合中英文分词词典和主题词典产生的URL翻译后的中文文本信息等；（6）其他与主题内容无关的，但在设计搜索算法时需要用到的信息，包括页面的访问时间、页面的错误信息等。以下对这几类信息进行简单的介绍。

1. 元信息

元信息即通常所说的Meta标签所包含的信息。本文只提取其中的Keywords（关键字）和Description（简介）这两种信息。Keywords为搜索引擎提供关键字列表，在HTML文档中的用法为<Meta name="Keywords" Content="关键词1,关键词2,关键词3,....." >；Description用来告诉搜索引擎网站的主要内容，在HTML文档中的用法为<Meta name="Description" Content="网页的简述" >。此类信息对网页主题有重要的指示作用。

2. 网页标题

网页标题是网页源代码中<Title>和</Title>标记之间的文字，也就是出现在浏览器界面的最左上方的文本内容。网页标题中的内容一般与网页的主题关系非常密切，起着概括全篇的作用。有人做过相关统计，如果标题中出现与某个主题相关的关键词，那么其主要内容与该主题也相关的网页数量占全部网页数量的97.8%。因此在判断网页

内容与查询主题相关度时，可以对出现在标题中的关键词赋较高的权值。

3. 链接内容信息

链接内容信息包括网页URL地址信息和网页锚文本信息。一般的网页制作者，都习惯于在自己制作的页面所对应的URL中，加入与该页面主题相关的信息来反映页面的主题，比如某个URL地址为http://www.cbe21.com/subject/physics/，那么该页面就很可能与“物理学科”这个主题有关。锚文本即超链接文本（Anchor），它也对该链接所指向的页面也起到了概括描述作用，这种概括在一定程度上可能会比该页面的作者所作的概括更为可观、准确。比如<A Href=" http://www.wsbedu.com/kejian/wulic3.asp">物理Flash课件，锚文本所链接的网页主题就很可能是关于“高一化学Flash课件”方面的。

4. 链接结构信息

在分析链接结构对网页主题贡献率的时候，我们提取父链接和兄弟链接两部分信息。如果网页P包含两个链接R和Q，我们称链接P是链接R和V的父链接，链接R和V是链接P的子链接，链接R和链接V互称为兄弟链接。根据主题页面的主题关联特征，可以得出下面的结论：

（1）如果父链接指向的页面内容与主题相关度较高，那么父链接所包含的子链接与主题的相关度可能也较高；

（2）如果兄弟链接指向的页面内容大部分与主题的相关度较高，那么链接的其他兄弟链接指向的页面也可能与主题相关。

因此把链接的父链接和兄弟链接所包含的信息提取出来，作为判断链接主题相关度的一个影响因素。

2.1.3 表征多媒体信息的相关标签介绍

Web 中的多媒体资源（视频、音频、动画、图片）分布在各类网站中，基本上按三种方式存在：第一种作为网页的组成成分嵌入在网页中，我们称之为嵌入式多媒体，如 Flash、在线音视频播放等；第二种通过网页的锚文本链接，可以自由下载，我们称之为超链接式多媒体；第三种存在于多媒体网络的数据库中，允许检索、浏览，但常常需要账号和密码。其中以第一种、第二种形式存在的多媒体，在网页的源文件中都有相应的链接地址和表征此多媒体类型的格式扩展名，依据多媒体的格式扩展名，网络多媒体主题搜索系统可以判断出此类多媒体的类型。第三种由于是动态生成的，没

有此类信息的提供，目前本文暂不对这样的网页进行分析。在 HTML 语言中用不同的标记来表征不同类型的多媒体信息。

1. <Img>标记信息

以<Img>标记主要是用来在网页中显示图片信息，其代码大致如下：

<Img src="image_url" width="xxx" height="yyy" alt="image-info">

其中 src 属性的值 image_url 该位置将提供图片的 URL，通过这个 URL，主题搜索器可以下载所对应的图片。alt 属性用于在图片无法正常显示时（通常是因为图片 URL 位置出错或者为死链接，或者是 Web 浏览器设置为不显示图片），用于在图片位置上替代显示的信息，该信息显示 image_info 的具体内容。当然，并不是每个<Img>标记都会有 alt 属性。但如果<Img>中包含了该属性，则其中的信息往往是对图片的内容一些具体描述。

Width 和 Height 属性是可选的，其本身不会带有关于图片内容相关的任何信息，只是对图片的位置和尺寸进行了描述，但是其提供的图片的位置在一定程度上也能够表明了此图片的重要性。

2. <Object>和<Embed>标记信息

<Object>标签只支持 IE 系列的浏览器或者其它支持 Activex 控件的浏览器，在此标签内可以插入的多媒体资源包括：视频、音频和动画。下面是<Object>标签中嵌入 swf 格式动画的例子：

```
<Object classid="clsid:D27CDB6E-AE6D-11cf-96B8-444553540000"
height="292"width="1000"codeBase="http://download.macromedia.com/pub/shockwave/cabs/flash/swflash.cab#version=5,0,0,0">
<PARAM NAME="_cx"VALUE="26458">
<PARAM NAME="_cy"VALUE="7726">
<PARAM NAME="FlashVars"VALUE="">
<PARAM NAME="Movie"VALUE="images/tongjivod.swf">
<PARAM NAME="Src"VALUE="images/tongjivod.swf">
<PARAM NAME="WMode"VALUE="Window">
<PARAM NAME="Play"VALUE="-1">
<PARAM NAME="Loop"VALUE="-1">
<PARAM NAME="Quality"VALUE="High">
```

```
<PARAM NAME="SAlign"VALUE="">
<PARAM NAME="Menu"VALUE="-1">
<PARAM NAME="Base"VALUE="">
<PARAM NAME="AllowScriptAccess"VALUE="always">
<PARAM NAME="Scale"VALUE="ShowAll">
<PARAM NAME="DeviceFont"VALUE="0">
<PARAM NAME="EmbedMovie"VALUE="0">
<PARAM NAME="BGColor"VALUE="">
<PARAM NAME="SWRemote"VALUE="">
<PARAM NAME="MovieData"VALUE="">
<PARAM NAME="SeamlessTabbing"VALUE="1">
<Embed src="images/tongjivod.swf"quality="high"
pluginspage="http://www.macromedia.com/shockwave/download/index.cgi?P1_Prod
_Version=ShockwaveFlash"type="application/x-shockwave-flash"width="1000" height="292">
</Embed>
</Object>
```

其中，<Embed>标签支持 Mozilla 系列的浏览器或其它支持 Netscape 插件的浏览器（Mozilla family of browsers）“pluginspage”属性告诉浏览器下载 flash player 的地址，如果还没有安装 flash player 的话，用户安装完后需要重启浏览器才能正常使用。为了确保大多数浏览器能正常显示 flash，通常把<Embed>标签嵌套放在<Object>标签内。支持 Activex 控件的浏览器忽略<Object>标签内的<Embed>标签。Netscape 和 Mozilla 系列的浏览器只读取<Embed>标签而不会识别<Object>标签。也就是说，如果你省略了<Embed>标签，那 Firefox 就不能识别你的 flash 了[24]。

3. <a>标记信息

以<a>标记来链接的多媒体与<Img>和<Object>、<Embed>的区别较大，此类标签表明多媒体是以超级链接的形式出现在网页当中，通常的代码大致如下：

```
<a href=”url_file.格式扩展名” >锚文本</a>
```

其中 href 属性的值 url_file 可以是任何文件的 URL 而不仅仅是多媒体，应该判断 url_file 实际指向文件的格式类型，对于静态的文件，通常包含扩展名，可以通过分析文件的扩展名来确定此文件的类型。<a>标记链接的多媒体可以是音频、视频、动画、

图像中的任何一种。对于动态的文件来说，判断就不是那么直观了，它通常是由带“？”的脚本生成的，而脚本通常都是一些 cgi 程序（扩展名为.cgi、.jsp、.asp 之类），显然通过判断扩展名已经没有用。一种办法是通过 HTTP 的 HEAD 请求来获取动态生成的内容的 MIME（Multipurpose Internet Mail Extension）类型，从而判断是否为多媒体数据，但是这样做需要进行 HTTP 请求，其代价是比较大的，而且这类信息通常是不稳定的，再加之目前互联网上的多媒体文件通常以静态的方式呈现个用户，所以本文采取最直接的办法，就是遇到动态的链接直接丢弃，根本不去处理，这也是目前大多数搜索引擎采用的处理办法。

通常，对于<a>标记表示的多媒体最准确的信息来自于链接文字，即锚文本。它是指当一个网页具有指向另外一个网页的链接时，与此链接相对应的描述性文字。一个网页的链接描述文档并不是该网页作者对自己网页的描述和说明，而是其它网页作者对该网页的简介描述或者页面内容的概述[25]。从这个角度讲，锚文本是对所链接网页客观的评价，它在一定程度上客观的表述了网页的主题。

锚文本是当前网页对所链接网页的说明，它一般比较精炼、简短，往往用一个词或者一句话来概括目标网页。用户在浏览网页时，可以根据锚文本的描述选择性的浏览自己感兴趣的页面。也就是说，锚文本在一定程度上决定了用户是否选择此链接的可能性。所以锚文本在计算查询主题与网页内容的相似度时起着重要的作用。目前不少商用搜索引擎开始利用锚文本信息来提高网页搜索的准确率。但是，由于锚文本的简短，这也限制了锚文本不可能表达出太多的内容，从而影响了网页检索时的查全率。在很多主题搜索算法中，为了提高搜索的查全率，往往也将锚文本的上下文相关文本提取出来，加入到主题相关度的计算过程中。另外，有时锚文本本身所含信息就是无效的，如网页中的“点击进入”、“下一页”，“取消”等，它们无法提供对目标页面的描述预测。

通过对大量包含多媒体的网页分析我们发现：包含多媒体的网页链接在其父网页中通常以链接列表的形式出现，这些链接列表我们称之为“主题团”，将“主题团”中包含的锚文本称之为“主题团”标题，它们对这些链接的主题起着指示性作用。在网页的源文件中，我们对锚文本做如下处理：首先将网页进行网页分块，分块原则是按照网页中的<Table>标签分为不同的链接序列，然后按照 4 个启发式规则提取出“主题团”内的标题[26]。规则如下：

（1）该文本的字号比周围文本的大；

（2）该文本与周围文本的颜色不同；

（3）该文本字数很少（一般少于 10 个）；

（4）该文本独立成段。

如果满足其中任意 2 个，则将其认定为“主题团”标题。以此作为表征多媒体主题的重要信息。在第 3.4.2 节中将此类信息加入到算法的计算过程中。

2.2 主题页面在 Web 上的分布特征

网络资源在 Internet 中的分布呈现出更新快、规模大、增长快、动态网页多、半结构化或者无结构化和异构性的特点。这些特点使得网页看起来似乎杂乱无章，无任何规律可循。但是通过查看大量的多媒体资源在网上的分布，并结合前人的研究成果，发现这些网页在网上的分布是有一定规律可循的，可以将这些主题页面的分布规律总结如下几个基本特征：中心页面（Hub）特征、主题关联（Linkage/Sibling Locality）特征、主题聚集（Cluster）特征和隧道（Tunnel）特征、资源相邻性特征，主题搜索器可以根据这些特征发现包含特定主题的多媒体网页。

1. 中心页面（Hub）特征

美国康奈尔大学 Kleinberg 教授发现 Web 上存在大量具有如下特征的页面[27]：这种页面不但含有许多出向链接并且这些链接趋向于同一个主题。Kleinberg 教授将这类页面称之为 Hub 页面。换而言之，Hub 页面是指向相关主题页面的一个中心页面。另外，Kleinberg 教授还给出了权威页面（Authority）的概念，即权威页面是那些关于某一个主题有价值的页面。好的 Hub 页面指向多个 Authority 的页面，并且所指向的 Authortiy 页面权威性越高，Hub 页面的质量越好；反之，Hub 页面的质量越好，它所指向的页面也越权威。根据这个思想，Kleinberg 教授提出了 HITS 算法[27]，该算法在计算广泛主题的应用中取得良好的效果。本文在第 3.2.2 小节对 HITS 算法进行了详细的阐述。

2. 主题关联（Linkage/Sibling Locality）特征

Aggarwal 等人[28]提出了主题页面的 Linkage/Sibling Locality 特性，也有学者把它翻译为主题关联特性。Linkage Locality 是指一个页面的主题倾向于和它的父页面的主题相同，也就是说如果一个页面与主题相关，那么它的子页面也可能与主题相关。在预测子页面的主题相关度时，可以给它乘上一个较高的权重，以说明父页面的相关度对子页面相关度的影响。Sibling Locality 是指对于一个链接到某个主题页面的页面而言，

它所链接指向的其他页面也倾向于和这个主题相关。该特征其实是 Hub 特性的另一个表达形式，只不过它是从页面编辑者的角度来考虑的：即一个页面的编辑者总是倾向于在本页面中添加指向与本页面相关的其他页面的超级链接。该特征在 4.3 小节，本文设计的多媒体主题搜索算法中有具体的体现。

3. 主题聚集特征

研究人员还发现，大部分非门户的普通站点都趋向于说明一个或者几个主题，而且那些相同的主题页面较紧密地在此站点的内部链接成团，但是不同的主题之间却很少有相互的链接。这种特征体现了人们使用分类分层的思维对网站的总体布局的设计思想，每一个网站都有一个明确的设计目标，而这种目标往往就集中在一个或者几个主题上。而 Web 上的浏览者在浏览网页的同时也往往带有一定的目的性，也趋向于浏览同一个主题的网页。

4. 隧道（Tunnel）特征

在 Web 中还有一种现象，就是尽管在 Web 上存在很多主题页面团，但是在这些页面团之间，往往是通过若干个主题无关的链接连接在一起，这些主题无关的链接在页面主题团之间，就好像一条长长的隧道，这就是 Tunnel 特征。

在主题搜索器搜索过程中，隧道特征极大的影响页面搜集的质量和搜索器最终资源的发现率。为了提高页面搜索的准确率，需要提高主题相关性判别算法中的过滤阈值，但是过滤阈值的提高往往会过滤掉大量的 Tunnel，使得采集系统很可能丢失 Tunnel 另一端的主题页面团，从而影响了查全率。反之，为了提高查全率，就必须保存 Tunnel，需要降低主题相关性判别算法中的过滤阈值，从而造成页面的查准率的降低。这是一个两难的问题，传统的主题搜索算法很难兼顾两者。本文在第 4.3.2 小节中提出了一种解决的方法。

2.3 HTML 文档信息的提取方法

目前关于网页信息提取方法的研究主要分为两大类：（1）基于一个或多个网站中的页面集进行页面的模板检测。（2）基于单一页面的处理，根据所处理页面的DOM结构，可视信息等应用一些启发式规则把页面内的噪音去除，抽取出主题内容[29]。

论文基于本实验室其他工作人员开发的HTML解析器（详细见参考文献[30]），对HTML文档信息的提取采用第2种情形，把HTML网页表示成DOM树，然后找到HTML文档树中包含的<Table>节点，如果<Table>节点之间包含<a>标记、<Object>标记、

<Embed>标记等链接序列，则将链接序列提取出来，否则不提取。

需要注意的是，大多数标签，解析它们并不困难，因为HTML解析器可以轻易的判断他们的起始标签和结束标签。然而对于某些特殊的HTML标签，解析起来却非常困难，目前最常见的就是超链接的链接值是一个JavaScript函数。由于链接的Href属性的值并不是一个具体的URL，而是一个JavaScript函数，点击此链接锚文本时，可以触发JavaScript函数，进而调用此函数，因此要想彻底解决此问题，就必须理解JavaScript，该方法的本质是使用Rhino扩展解析器，使其能够识别JavaScript。因为其复杂性，本文并不采取这种办法（该方法可在http://www.mozilla.org/rhino中找到更多的相关信息）。本文从实用角度出发，当遇到链接是以“JavaScript:”开始时，就将其忽略。

2.4 本章小结

本章首先对 HTML 文档做了简单的介绍，包括 HTML 解析器的实现原理、HTML 文档中能描述多媒体内容的文本信息以及 HTML 文档中能够表征多媒体信息的相关标签的介绍。同时介绍了在 Web 中主题页面的分布特征，包括中心页面（Hub）特征、主题关联（Linkage/Sibling Locality）特征、主题聚集特征和隧道（Tunnel）特征，这些特征是主题搜索算法实现的依据。最后介绍了如何通过 HTML 解析器来提取 HTML 文档中包含的多媒体信息。

第三章 主题搜索算法研究

主题搜索算法是整个主题搜索系统的核心，它也是区别主题搜索器和通用搜索器的关键部分。传统的通用搜索器采用广度或深度优先的策略搜索Web，以求较高的Web覆盖率。主题搜索器服务于特定人群，其索引的内容只限于特定的主题和专门领域，在主题搜索器中网络蜘蛛[14]通常采用“最好优先”策略访问Web，即为快速、有效的获得更多的主题相关页面，每次选择最有价值的链接进行访问。因此，如何评价链接的价值是决定主题搜索算法优劣的关键因素，本文基于这个因素对现有的主题搜索算法进行了分类，系统分析、比较了它们的优缺点，并分析现有的主题搜索算法运用于多媒体主题搜索系统中的可能性。

3.1 基于内容的搜索算法

基于内容的搜索算法，主要是根据主题（关键词、主题相关文本）与文本内容（包括网页内容、链接文本）“语义”的相似度来评价链接价值的高低，以此决定链接访问的顺序。链接文本是指链接周围的说明文字和链接URL上的文字信息。该类算法的代表是Fish-Search算法和Shark-Search算法，Fish-Search算法是De Bra等[31]早期提出的一个的主题网页动态爬行算法。Fish算法将主题蜘蛛在Web中爬取网页的过程模拟为鱼群在大海中觅食的过程，当鱼找到食物时（相似性大于一定值），鱼的繁殖能力就增强，反之鱼则逐渐消亡。Shark-Search算法在Fish-Search算法的基础上做了2种主要的改进。首先，用一个连续的值函数来表示相关性，取值在0-1之间，而不是Fish-Search的二值判断；另外，待爬行链接的主题相关性受锚文本、锚文本上下文和父链接相似性继承的影响。在Fish-Search算法和Shark-Search算法中，经常使用向量空间模型来计算主题内容的相似度。

3.1.1 向量空间模型（VSM）

向量空间模型基于这样一个关键假设，即组成文章的词条所出现的顺序是无关紧要的，它们对于文章的主题所起的作用是相互独立的，因此可以把文档看作一系列无序词条的集合[32,33]。

VSM模型以特征项作为文档表示的坐标，以向量的形式把文档表示成多维空间中的一个点，特征项可以选择字、词和词组等（根据实验结果，普遍认为选取词作为特

征项要优于字和词组），表示向量中的各个分量。它的基本思想是这样的，把文档d_i看成是由一组词条（$T_1,T_2,...,T_n$）构成的，对于每一个词条T_i，都可以根据它在文档中的重要程度赋予一定的权值w_i。可以将T_1，T_2，...，T_n看作一个n维坐标系，w_1，w_2，...，w_n为对应的坐标值，因此每一篇文档都可以被看作向量空间中由一组词条矢量构成的一个点，如图3-1。

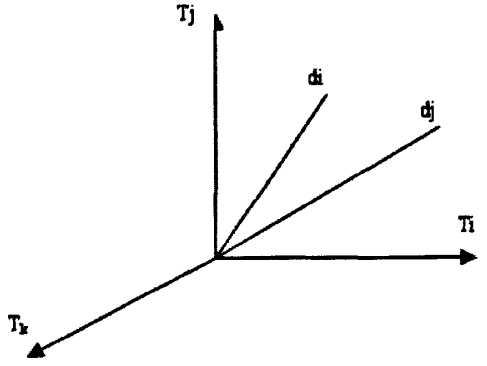

图 3-1 VSM模型示意图

向量空间模型中文档的特征表示方法如下：

设D是一个包含m篇文档的文档集合，$D=\{d_1，\ ...d_i，\ ...d_m\}$，$i=1，2，...,m$集合中的每篇文档$d_i$都被表示成向量$d_i=(w_{i1},...,w_{ij},...w_{in})$，$i=1,2,...,m$; $j=1,2,...,n$。其中，w_{ij}表示第j个特征项T_j在文档d_i中的权值。

权值的计算方法有以下三种：

（1）$w_{ij}=\begin{cases}1 & \text{第 j 个特征项属于文档 } d_i\\ 0 & \text{第 j 个特征项不属于 } d_i\end{cases}$

（2）$w_{ij}=\begin{cases}t_{ij} & \text{第 j 个特征项属于文档 } d_i \text{ 出现的次数}\\ 0 & \text{第 j 个特征项不属于 } d_i\end{cases}$

（3）TF-IDF词频统计方法[34]。该方法基于这样一个假设：在真实语料中，出现词频越高的词条（特征）带有较少的信息，而出现频率越少的词条带有较多的信息。TF-IDF

的值表示权重，词条T_j在文档d_i中的TF-IDF值有以下公式定义：

$$w_{ij} = tf_i \times \log(N/df_i) \tag{3.1}$$

其中，tf_i是词条T_j在文档中出现的次数；df_i表示整个文档集D中包含词条T_j的文档树，称为文档频率，IDF为df_i的倒数，成为逆文档频率；N表示统计语料中的文档总数。因此，文档d_i可以表示成一个特征向量：

$$d_i = (w_{i1},...,w_{ij},...w_{in}) \tag{3.2}$$

3.1.2 页面主题相关度

页面主题相关度的计算有多种方法，例如 Naïve Bayes、神经网络、实例映射模型、向量空间模型等[35]。其中向量空间模型对训练文档的要求较低，主要的目标特征可以从少量的训练文档中提取出来，而且计算比较简单、正确率比较高，适用于网络信息的发现。本论文在后面计算网页的内容相似度时，运用的算法基础就是基于向量空间模型的简单向量距离算法。该算法的基本思想是计算图 3-1 中两个向量之间的夹角余弦值，下面对该算法进行简单介绍。

简单向量距离算法的实现步骤如下：

（1）将训练文档集中的文档表示成特征向量。

（2）计算特征项词条在该类中的权值。将特征项词条在该类所有文档特征向量中权值的算术平均值，作为该词条在类别特征向量中的权值。

（3）将待分类的新文本表示成特征向量。

（4）用公式（3.3）计算新文本特征向量和类别特征向量之间的相似度。

（5）比较待分类文本与每类类别特征向量之间的相似度，将文本分到相似度最大的那个类别中。

$$sim(d_i,q_j) = \cos(\theta) = \frac{M\sum_{k=1}^{n} w_{ij} * w_{ik}}{\sqrt{\sum_{k=1}^{M} w_{ik}^2 * \sum_{k=1}^{M} w_{jk}^2}} \tag{3.3}$$

其中，d_i为待分类的文本特征向量，d_j为第 j 类的中心向量，M 为特征向量的维数，w_{ik}和w_{jk}分别为向量的第 k 维在文本d_i和d_j中对于的权值。w_k采用公式 3.1 来计

算。

基于内容的主题搜索算法，根据语义相似度的高低来决定链接的访问顺序。这种方法起源于文本检索中对文本相似度的评价，优点是计算简单。但是由于Web页面是一种半结构化的文档，包含很多结构信息，同时页面与页面之间通过超级链接相互联系在一起，基于内容的主题搜索算法忽略了网页链接方面对链接价值的影响，忽视网页的“主题聚集特征”和“主题关联特征”。

3.2 基于链接结构的搜索算法

在搜索引擎中人们都是希望能够找到最有权威性的网页，权威性（Authority）的网页一般都隐藏在Web页面链接中。Web由页面和从一个页面指向另一个页面的超链接组成，超链接包含了大量潜在的注释，这些注释有助于网页权威性的判断。当一个Web页面的作者建立指向另一个页面的链接时，可以看作是作者对另一个页面的认可。通常认为具有较多入链或出链的页面具有较高的价值。基于链接结构的搜索算法，就是通过对Web页面之间相互引用关系的分析来确定链接的重要性，进而决定链接访问的顺序。Page-Rank算法和HITS算法是其中具有代表性的方法。

3.2.1 PageRank 算法

PageRank超链接分析算法由斯坦福大学博识研究生Sergey Bring和Lawre Page提出并实现的[36]，是Googe搜索引擎的核心技术之一，主要用于对搜索结果的排序上，Google通过PageRank元算法计算出网页的PageRank值，从而决定网页在结果集中的出现位置，PageRank值越高的网页，在结果中出现的位置越靠前。简单地说，PageRank值代表网络上某个页面重要性的一个数值。

PageRank算法主要基于以下假设：一个网页被其它网页链接的次数越多，则它可能是越重要的；一个网页虽然没有被多次引用，但是被重要的网页链接引用，则它也可能是很重要的；一个网页的重要性平均地传递到它所链接的网页。

基于以上思想，我们可以得出页面 p 的PageRank值 $PR(p)$ 的迭代公式：

$$PR(p)=\frac{1-d}{N}+d*\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{3.4}$$

其中：$PR(p)$ 代表网页 p 的PageRank值；$PR(T_i)$ 代表网页 T_i 的PageRank值，网页 T_i 指向网页 p；d 为阻尼系数；$C(T_i)$ 网页链入网页 p 的网页数量。

PageRank算法是一种随机漫游模型[37]，算法在迭代计算过程中，权值是按当前网页的出度（out-degree）平均分配的，没有考虑到网页的相对重要性。但用户在实际的访问过程中，用户会根据链接与主题的相似度，选择性的访问网页。例如一个页面X有4个链接，分别指向页面A、B、C、D，四个页面与主题的相似度分别为：0.2、0.4、0.6、0.8。则用户在选择网页连接时，选择页面D的机率要远远大于页面A。而在PageRank算法中网页的PageRank值被平均的分配到它所链接的网页中去，同时这类方法只考虑了链接结构和页面之间的引用关系，但忽略了页面与主题的相关性，在有些情况下，会出现搜索偏离主题的“主题漂移”问题[38]。

3.2.2 HITS 算法

同PageRank算法一样，HITS算法也是根据链接结构和页面之间的引用关系决定页面重要性的算法。HITS算法中定义了两个重要概念：Authority和Hub[39]。

Authority表示一个权威页面被其它页面引用的数量，指向该权威页面的数量越大，则该网页的Authority值越大。但是在Web链接结构中，并不是每一个超链接都代表作者寻找认可，如一个网页导航或付费广告。同时，由于存在着商业竞争，很少有Web页面会指向其竞争领域的权威页面。这些Web链接结构存在的局限性，会导致页面Authority值很大，但却并不是人们所需要的页面。由此人们提出另外一种重要的Web页面：Hub。Hub是指一个或多个Web页面，它提供了指向权威页面的链接集合，网页指向其他页面集合越大，此网页的Hub值越大。Hub页面本身可能并不突出，或者说可能没有几个链接指向他们。但是，Hub页面却提供了指向就某个公共话题而言最为突出的站点链接，比如（http://www.hao123.com）。一般来说，好的Hub网页指向许多好的权威网页；好的权威网页是有许多好的Hub网页指向的Web网页，这种Hub与Authority网页之间的相互加强关系，可用于权威网页的发现和Web结构和资源的自动发现，这就是Hub/Authority方法的基本思想。

基于以上思想，我们可以利用Hub页面去找到权威页。Kleinberg在90年代末提出的一种链接分析算法（HITS算法），与PageRank算法等实用性算法不同，HITS算法更大程度上是一种实验性质的尝试。HITS算法对每个访问页面计算其Authority权重和Hub权重。设页面P的Authority权重和Hub权重分别为A[p]和H[p]，它们分别按下列迭代公式计算：

$$A[p]=\sum_{q:(q,p)\in E} H[q]$$

$$H[p]=\sum_{q:(q,p)\in F}A[q] \qquad (3.5)$$

其中，E为所有指向页面p的页面集合，F为被页面p中链接指向的页面集合。

PageRank算法和HITS算法均为基于链接分析的搜索算法，在评价链接价值时采用“在线”评价方式，其评价的依据主要是在线获得的信息，因此这两种方法也被称为基于立即回报价值评价的搜索算法。

3.3 基于先验知识的搜索算法

近年来对Web信息资源分布特点的研究表明，同一类型网站的构建方式、同一主题网页的组织方式存在某种程度的“相似性”。有的学者根据这种相似性，首先对主题爬虫进行训练，使其具备一些“先验知识”，然后利用这些“先验知识”来预测较远的未来回报价值，文献[40]将这种链接价值评价方式的搜索策略称为基于未来回报价值的搜索策略。代表性的方法有基于 “巩固学习”[40]的搜索算法和基于“语境图”[41]的搜索算法。

巩固学习是指代理体在一个顺序的任务中，通过接受来自环境的奖励和惩罚来提高自身性能的学习方式，它的优势在于能够预测远期的回报价值。MaCallum将其引入主题搜索领域，以解决基于内容的搜索算法存在的“近视”缺点，并将主题搜索器的搜索过程表述为马尔科夫决策过程[42]（Markov Decision Processes,MDPs）。主题搜索器被看作代理体，Web页面的动态结构代表状态，主题搜索器对链接的访问代表行动，搜索过程被划分为训练和搜索两个阶段。由于在巩固学习模型中，未来回报价值用Q价值表示，因而这种算法的核心是学习如何计算链接的Q价值。在训练阶段，利用巩固学习算法计算每个链接的Q价值，并用Q价值的大小将链接分类，然后用每一个类中的链接文本训练一个贝叶斯分类器；在搜索阶段，对每个未知的链接根据其链接文本，用训练好的贝叶斯分类器计算链接落在每一类中的概率，并以此为权值计算链接的Q价值。Q价值反映的是远期价值的预测值，对于当前与搜索页面主题不相关的页面，这类搜索算法也可以根据预测的未来价值确定正确的搜索方向。

基于巩固学习的搜索算法虽然能够确定主题搜索器的搜索方向，但无法估计距离目标页面的距离，为此，Diligenti等人提出通过构建典型Web页面的“语境图”（Context Graph）来解决这个问题，该算法同样分为训练和搜索两个阶段。在训练阶段，首先选

择典型的主题相关页面作为种子集，并从这些种子集出发，通过一些提供引链查询服务的通用搜索引擎检索出所有指向它们的页面，并以这些页面作为第一层次集（表示到目标页面的距离为1），并用这一层次集中的页面文本训练一个贝叶斯分类器*C1*；然后再从第一层次集中的页面出发，按同样的方法得到第二层次集（表示到目标页面的距离为2）和分类器*C2*；如此重复，直到某个预先指定的层次。由此便可得到一个种子页面集和周围页面之间层次关系的“语境图”。在搜索阶段，面对一个新的页面，用训练好的贝叶斯分类器确定该页面属于哪个层次集，并将其中的超链接也放入相应的层次集中，链接的爬行优先级按照链接距离的大小确定，链接距离越小优先级越高。基于先验知识的搜索算法本质上是通过训练挖掘出链接文本中隐含的结构信息，从而判断出搜索页面与目标的距离远近。基于先验知识的搜索算法对于预测未来搜索方向有一定优势。

3.4 对两种典型主题搜索算法的改进

如前面所述，多媒体资源在网页中基本上按三种方式存在：嵌入式多媒体、超链接式多媒体、动态生成多媒体。其中以第一种、第二种形式存在的多媒体，在网页的源文件中都有相应的链接地址和表征此多媒体类型的格式扩展名，依据多媒体的格式扩展名，网络多媒体主题搜索系统可以判断出此多媒体类型。第三种由于是动态生成的，没有此类信息的提供，目前很难对这样的网页进行分析。

多媒体主题搜索和主题搜索有相似之处，即它们都是搜索网络中特定的主题。而多媒体主题搜索还要求搜索到的网页中包含特定主题的多媒体资源。我们基于多媒体在网页中分布的特点，对已有的主题搜索策略进行分析发现：基于“巩固学习”的主题搜索策略和基于“语境图”的主题搜索策略很难应用于多媒体主题搜索中，两种搜索策略都将搜索过程分为训练和搜索两个阶段，根据训练构建网页分类器，再用于指导搜索阶段网页主题的判断，而由于多媒体在网页中分布的复杂性和多样性，通过训练很难构建具有一定普适性的分类器。本文从基于内容评价搜索策略和基于链接评价搜索策略中，选择两种典型的主题搜索算法进行改进，下面详细介绍一下两种算法的改进过程。

3.4.1 基于 Topic-PageRank 主题搜索算法

PageRank算法是一种随机漫游模型，算法在迭代的过程中权值是按当前网页的出度平均分配的，即在权值分配中平等对待链出的每个网页，没有考虑网页的相对重要

性。但实际上，从链接结构上看，网页按入度和出度的不同是具有相对重要性的，入度和出度较大的网页比较重要，因此应该分得较高的权重。基于此方面的考虑Xing[43]提出加权PageRank（Weighted PageRank，简称WPageRank）算法，其中网页的重要性和网页的入度和出度成正比。同时考虑到：网页间的链接反映的是一种认可关系，网页A中有链接指向网页B，说明网页B的内容与A相关或具有一定的价值，同一网页中不同链接指向的网页内容与当前网页内容的相关程度是有差别的。基于此思想，Ingongngam等人[44]提出了以主题为中心的PageRank（Topic Centric，简称TCPageRank）算法，算法指出网页权值的分配应和网页内容的相似度成正比，被链接的网页内容与当前网页的内容越相似分配到的权值比重就越大。

对于PageRank算法的改进在一定程度上提高了知名网站在迭代过程中获得的权重，但相对于某个特定的查询，即使它们与查询有着微弱相关也会排在结果的前列，这种现象称为主题漂移（Topic Drift）。主题漂移现象使得查询的相关性遭到极大破坏，WPageRank算法在分配权值时以网页重要性为比例，因此知名网站会获得更高的权重，在一定程度上也加剧了主题漂移现象的发生。在TCPageRank算法中根据网页的相关性来分配权值可以有效解决主题漂移现象，但却忽略了排序中对权威性的需求。

为了提高主题搜索的准确性，我们从网页内容和链接分析两个角度对 PageRank 算法进行改进，从网页内容相关性分析角度确定网页与主题的相关性，从链接分析角度确定网页的权威性和主题资源搜索的覆盖率。我们在利用 PageRank 算法对待爬行队列中的页面进行排序时，把网页的链接信息相似度加入到 PageRank 计算公式中。该算法认为：用户在查资料时，用户点击网页内链接的机率不是相等的，而是和两个因素有关：链接锚文本的主题相关度和链接所指向网页的实际主题相关度。链接被点击的概率，同这两个因素成正比，这里我们将改进后的 PageRank 算法称为 Topic-PageRank 算法，具体公式如公式 3.6。

改进后的 Topic-PageRank 算法为：

$$TPR(p)=\frac{1-d}{N}+d*\sum_{i=1}^{n}TPR(T_i)*P(T_i,p) \qquad (3.6)$$

在公式中：$TPR(p)$代表网页 p 的 PageRank 值；$TPR(T_i)$代表网页T_i的 PageRank 值，其中网页T_i指向网页 p；d 为阻尼系数；$P(T_i,p)$为从页面T_i到达页面 p 的概率；N为已下载到待爬行队列中与主题相关的网页数量；n 为链接到网页 p 的网页数量。

前面提到网页T_i到达页面p的概率$P(T_i,p)$受两个因素的影响：链接锚文本的主题相关度和链接所指向网页的实际主题相关度。$P(T_i,p)$的值与这两个因素成正比，锚文本的主题相关度越高，则$P(T_i,p)$越大。$P(T_i,p)$的计算公式如公式 3.7。

$$P(T_i,p)=\omega*\frac{Sim_{content}(p)}{\sum_{i=1}^{n}Sim_{content}(i)}+(1-\omega)*\frac{Sim_{link}(p)}{\sum_{i=1}^{n}Sim_{link}(i)} \tag{3.7}$$

$\sum_{i=1}^{n}Sim_{content}(i)$表示从网页$T_i$中链出的所有网页内容相似度的集合；$\sum_{i=1}^{n}Sim_{link}(i)$表示从网页$T_i$中链出的所有网页的链接信息相似度的集合，其计算公式为 3.10；ω是一个影响因子，取值范围为 0~1。

在 PageRank 算法迭代的过程中，由于待爬行链接的网页内容相似度是通过继承其父网页的内容相似度得到的，所以在计算其内容相似度的时候，只能获取网页的锚文本和 URL 信息，通过布尔模型来计算网页的内容相似度。具体公式为：

$$Sim_{content}(i)=\alpha*Sim_{author}(i)+\beta*Sim_{URLFanyi}(i) \tag{3.8}$$

$$Sim_{anchor}(i)=\frac{|D_{anchor}(i)\cap T|}{|T|} \tag{3.9}$$

$$Sim_{URLFanyi}(i)=\frac{|D_{URLFanyi}(i)\cap T|}{|T|} \tag{3.10}$$

其中，$Sim_{content}$代表网页的内容相似度；Sim_I代表每部分文本内容与主题词集的内容相似度。D_I为文档集合，T为主题关键词集合。URLFanyi 是利用中英文翻译词典将网页的 URL 地址翻译成中文文本内容。通过统计，不同位置的文本内容对网页的标识程度不同，为 Anchor 、URLFanyi 两部分文本内容分配不同的权重α：β=6：4。最后计算将得到的$Sim_{content}$作为一个字段和网页的其它信息存入数据库中，在计算链接分析相似度时还要用到。

根据 PageRank 算法的思想：我们把 Web 看成一个有向图 G=（V，E），可以用图这种数据结构来表示链接关系。其中定点 V 为所有网页集合，边 E 为网页间的链接集合。对于有向边（p，q）∈E 代表网页p链接指向网页q，有向边（p，q）上的权重代表网页与主题的相关度，节点的出度(out-degree)指的是节点链出的网页数量，而节点的

入度(in-degree)则指的是链接指向节点的网页数量。

因此根据网页的这种结构特点，我们可以做出以下假设：具有相同主题的网页更倾向于聚集在一起；网页 X 的入度越高，它与主题相关度就越高；如果网页 X 的出度越高，则它的主题相关度也高。

$$Sim_{link}(d_j) = (1-\lambda) * \frac{\sum_{j=1}^{n} Sim_{content}(p_i)}{n} + \lambda * \frac{\sum_{i=1}^{N} Sim_{content}(b_i)}{N} \qquad (3.11)$$

式中：$Sim_{link}(d_j)$——页面 d_j 的链接信息主题相关度预测值；

$Sim_{content}(p_i)$——第 j 个链入到网页 d_j 的页面主题相关度；

$Sim_{content}(b_j)$——第 i 个网页 d_j 链出的网页的主题相关度；

λ——是影响因子，$\lambda = 0 \sim 1$ ；

在公式 3.11 中，网页 d_j 的主题相关度与网页的链接入度的平均主题相关度和网页 d_j 的链接出度的平均主题相关度有关。影响因子 λ 用于调节链接入度和链接出度两个因素之间权重分配。λ 越大，则公式更倾向于网页的链接入度，λ 越小，则公式更倾向于网页的链接出度。在搜索过程中，对于一些底层网页，它的链接出度几乎为零，而对于一些网页的初始页，它的链接入度远远小于它的链接出度的数量。所以要及时调整 λ 的大小，以确保搜索结果的准确性。

由于网页 p_i 和 b_i 在爬行过程中已经被主题蜘蛛下载到本地，并且通过基于内容相似度计算出来。$Sim_{content}(p_i)$ 和 $Sim_{content}(b_j)$ 的值可以直接利用，无需重新计算。因此页面 d_j 的主题相关度可以在较少的时间内计算出来，从而保证主题蜘蛛的爬行效率。

本文对 PageRank 改进后的算法和 Lawrence Page 和 Sergey Bring 提出的 PageRank 计算公式最大的区别在于：后者认为，用户点击页面中 URL 的概率都是相等的，而实际上，当用户在搜索某个领域的资料时，与主题相关的页面总是获得较高的点击概率，在改进的算法中，每个利用公式 3.7 进行迭代计算时，TPR 值都将乘以此链接点击的概率再被传递到下个页面。改进后的计算公式受链接关系和主题相关度两个参数的影响。Topic-PageRank 算法改进的具体细节为：

（1）链接主题相关度越高，对此页面的 TPR 值贡献越大。

（2）网页入度页面 TPR 值越高，则此页面的 TPR 值也越高。

（3）TPR 值不再以均等的概率被传递到下一个页面。

3.4.2 基于改进 Shark-search 的主题搜索算法

1．Fish-Search 算法与 Shark-Search 算法

De Bra 等[31]提出的 Fish-Search 算法是一个早期的主题网页动态爬行算法。Fish 算法将主题蜘蛛在 Web 中爬取网页的过程模拟为鱼群在大海中觅食的过程，算法中的每条鱼代表一个 URL。当鱼找到食物（发现相关网页）时，它的繁殖能力增强（搜索宽度增加），并且它繁殖的后代寿命与它自身相同（搜索深度不变）；当没有发现食物（没有发现相关网页）时，它的繁殖能力保持不变（搜索宽度不变），并且它的后代寿命缩短（搜索深度-1）；当进入污染区（网页不存在或者读取时间太长），这条鱼死去（放弃对该链接的爬行）。此算法是基于页面内容与主题的相关性以及链接选取的速度来确定待爬行 URL 的优先级。然而它对相关性的判断是离散的二值判断，即：相关/不相关。在 Fish-Search 算法的基础上，Hersovici 提出了 Shark-Search 算法。

Shark-Search[45]算法在 Fish-Search 算法的基础上提出了两种主要的改进。首先，用一个连续的值函数来表示相关性，取值在 0-1 之间，而不是 Fish-Search 的二值判断；另外，待爬行链接的主题相关性受锚文本、锚文本上下文和父链接相关性继承的影响。文献[46]对网页中不同区块的链接进行聚类，然后将相同类的所有链接锚文本作为该类的描述文本，以此来计算该类与主题的相关性。用来替代 Shark-Search 算法中锚文本上下文对链接相关性的影响。文献[47]从网页页面、链接块，以及链接本身三个粒度上对网页的相似性分别进行计算，然后将三者按照不同权重结合，进而确定整个网页的相似性。

2．改进 Shark-Search 算法

（1）网页链接分块

在本文第 2.1.3 节中，我们通过分析得到：网页链接在其父网页中通常以链接列表的形式出现，这些链接列表称之为“主题团”。用来表述“主题团”的文本称为“主题团”的标题，它们对这些链接的主题相关性具有指示作用，我们把一个“主题团”内的链接划分成一个网页链接块。

（2）网页内容相似性计算

网页链接按照 table 标签分为不同的“主题团”，将“主题团”标题与主题的相关度

作为待爬行链接中锚文本和URL地址的权重。这样待爬行链接的主题内容相关度的计算公式如公式3.12。

$$Content\_score(u_i) = Score(block\_title)[\beta * Score(anchor) + (1-\beta) * Score(url)] \tag{3.12}$$

其中，$Score(block_title)$是链接u_i所在“主题团”标题与主题的相关度，计算时采用向量空间模型VSM，在向量空间模型中，所有的检索关键词t形成关键词集合$T=(t_1,t_2,...,t_n)$，“主题团”标题文档D中的每一个文档d都被表示成一个范式的矢量$V_i(d)=(t_1,w_1(d);...;t_i,w_i(d);...t_n,w_n(d))$，其中$w_i(d)$为$t_i$在文档$d$中的权重，权重的计算采用TF-IDF词频统计方法计算，如公式3.13。采用向量空间模型计算“主题团”标题与主题的相关度，如公式3.14。$Score(anchor)$和$Score(url)$分别表示链接u_i的锚文本和URL地址与主题的内容相关度，采用布尔模型进行计算，具体计算过程如公式3.9，3.10。β为相关子，用以调节链接的锚文本和URL地址所占的比重。

$$w_i(d) = \frac{tf_i \lg(\frac{N}{nt_i} + 0.01)}{\sqrt{\sum_{i=1}^{n}(tf_i \lg(\frac{N}{nt_i} + 0.01))}} \tag{3.13}$$

其中，tf_i表示关键词t_i在文档d中出现的频率；N表示用于特征提取的全部训练文本的文档总数；nt_i表示出现关键词t_i的文档频率。

$$Score(block\_title) = sim(d,q) = \cos(\theta) = \frac{\sum_{i=1}^{n}(w_i(d) * w_i(q))}{\sqrt{\sum_{j=1}^{n} w_j^2(d) * \sum_{j=1}^{n} w_j^2(q)}} \tag{3.14}$$

其中，$w_i(q)$为关键词t_i在查询q中的权重，通常当查询中包含就为1；否则就为0。“主题团”标题与查询主题的相关度就表示为两个范化矢量之间夹角的余弦。

（3）网页链接相关度计算

在本文第2.2中介绍到：在链接结构方面，主题网页表现出“主题关联特征”和“主题聚集特征”。通过对大量网页的观察，对于同一种子网站中，包含多媒体的网页不仅符合上述两个特征，往往还呈现出“资源相邻性”的特点。所谓“资源相邻性”是指

在一个网站中，相同主题的网页往往存在于这个网站的某一部分或某几部分区域中。并且处于同一区域的网页主题也往往是相同的。根据此特点，在计算网页链接相关度时，我们用父网页和兄弟网页的链接相关度来揭示链接结构对一个 URL 链接相关度的影响。为了把这种影响实时地反馈给每个子链接，引入一个动态因子。用公式 3.15 来表示链接结构对一个 URL 链接相关度的贡献：

$$Structure\_score(u_i) = \sum_{j=1,\ u_i \in d_j}^{t} \lambda(d_j) P(d_j) \Big/ t \qquad (3.15)$$

其中，u_i 是正在爬行的链接，t 是父链接的总数，$\lambda(d_j)$ 是动态因子，用公式 3.16 进行计算；$P(d_j)$ 表示从父链接继承来的链接相关度和已爬行过兄弟链接的平均链接相关度，它来衡量通过父链接能爬行到多少主题相关页面的能力，用公式 3.17 进行计算；

$$\lambda(d_j = (n' + \theta) \ / \ (n + \theta) \qquad (3.16)$$

其中，n' 是父链接 d_j 的已爬行子链接中主题相关页面的个数，n 表示父链接 d_j 已爬行子链接的总个数，θ 是归一化因子，通常取 0.5。在爬行的过程中，$\lambda(d_j)$ 会不断的动态调整。

$$P(d_j) = (1-\sigma) R(d_j) + \sigma \sum_{k=1, d_k \in d_j}^{N} P(d_k)/N \qquad (3.17)$$

其中，σ 是偏置因子，$R(d_j)$ 为父链接 d_j 的主题相关度，d_k 为 d_j 的一个已爬行子链接，N 为 d_j 已爬行子链接的总数，$\sum_{k=1, d_k \in d_j}^{N} P(d_k)/N$ 是父链接 d_j 中已爬行子链接的平均链接得分。

3.4.3 内容相似度和链接相似度的归一化

两种算法都是从网页内容和网页链接的角度来计算多媒体网页的主题相似度的，为了提高整个网页的主题相关性和权威性，我们采用内容相似度和链接相似度按不同权值相加所得结果来标识。在判断网页的主题相似度时将二者归一化，将计算得到的值作为主题搜索器未来爬行链接的依据，具体实现过程将在第 4.6 节中进行详细介绍。

3.5 本章小结

本章首先介绍了目前主流的几种主题搜索算法：基于内容的搜索算法、基于链接

结构的搜索算法和基于先验知识的搜索算法等，并对这几种算法的特点进行分析，提出可以通过对基于内容的搜索算法和基于链接结构的搜索算法进行相关参数的改变，从而应用到多媒体主题搜索领域中。

选择 PageRank 算法和 Shark-Search 算法进行改进，详细介绍了两种算法改进的过程。为了提高算法的查全率和查准率，两种改进后的算法都将网页内容相似度和网页链接相似度加入到算法的计算过程中。

第四章 基于URL链接规则的多媒体主题搜索算法

通过对大量包含多媒体网页的分析，我们发现在同一网站中，包含多媒体的网页往往呈现出“资源相邻性”的特点。即在同一个网站中，包含多媒体资源的网页往往存在于这个网站的某一部分或某几部分区域中，并且处于同一区域的多媒体资源的主题也是相同的。依据此特点，我们对PageRank算法在网页链接方面进行相应的改进，提出一种改进的PageRank算法——基于URL链接规则的多媒体主题搜索算法。本章从理论上介绍了此算法在多媒体资源主题搜索领域的适用性，并详细介绍了算法的实现过程。

4.1 文本信息的处理

在主题搜索过程中，对于HTML解析器提取出的文本信息，要首先进行一系列处理以后才能将其加入到算法的计算过程中，对于锚文本和网页标题等信息需要进行中文分词，网页链接URL地址字符需要进行英文和汉语拼音翻译。以下对相关的技术进行具体介绍。

4.1.1 中文分词算法

目前，中文分词算法大体可以分为三类[48]：基于字典匹配的分词算法、基于词频统计的分词算法和基于知识理解的分词算法。本系统采用基于教育主题字典的正向最大匹配算法（MM算法）。该算法的思想是：一个正确的分词结果应该由合法的词组成，这些词在当前待切分的句子中，并且属于词典中的一个分词。在分词过程中，按照正向的扫描方向，采用最大长度优先匹配的方式进行分词。具体描述为：假设自动分词词典中的最长词条所含汉字个数为MaxLen个，则取被处理材料当前字符串中的MaxLen个字作为匹配字段，查找分词词典。若分词词典中有这样的一个MaxLen个字的词，则匹配成功，匹配字段作为一个词被切分出来；若在分词词典中找不到，则匹配失败，匹配字段去掉最后一个字，剩下的作为新的匹配字段，进行新的匹配，如此进行下去，直至切分成功为止。然后再按上面的步骤进行下去，直到切分出所有的词为止。在系统中主要针对爬行网页的中文锚文本、网页<Title>标签中的中文文本进行中文分词。

4.1.2 英文和拼音翻译技术

网页的URL一般都是由字符串组成的，有的是英文单词，有的是汉语的拼音、有

的是汉语拼音的缩写等等。比如，http://www.chuzhong2wuli/kejian/index.axp，可以表示成初中、物理、课件方面的网站。在判断网页内容和主题相似度的时候，网页URL也是表征网页内容的重要组成部分。所以，在分析此部分的时候，需要首先把网页的URL转化成中文字词，然后和教育主题词典进行匹配，从而判断出网页的内容相似度。将网页URL字符串分成一个个的单词，然后再把每个单词通过中英文词典翻译成中文词语。技术的具体过程为：中英文翻译词典载入、根据定义好的分隔符进行分词、根据大写首字母进行分词、进行中英文翻译和无关词过滤。由于网页URL字符串的不确定性，我们在系统运行的过程中，尽量多地囊括我们搜集到的字符串。同时在计算网页内容相似度的时候，也降低了此部分在整体相似度中所占的比重。

4.1.3 布尔模型

在网页内容主题判断时，由于网页锚文本、URL翻译后的中文文本一般都是由几个关键词组成，所以它们的主题内容相似度采用布尔模型来计算。在布尔模型中，一个文档通过一个关键词集合来表示。同时，网页主题内容表征也以关键词集合的形式表示，在判断文档与某主题的相关度的过程中，相当于是计算两个关键词集合的交集。对基于布尔模型的主题判别模型来说，交集中含有的元素越多，则认为与主题的相关度就越高。可以用文档D与主题关键词集合T之间的交集元素的个数占集合T的比例来代表文档D的主题相关度Sim(D)，公式表示如下：

$$Sim(D)=\frac{|D\cap T|}{|T|} \tag{4.1}$$

布尔模型的主要缺陷在于每个关键词的权重都是一样的，它不支持设定关键词的相对重要性。优点是易于实现，计算代价较小。

4.2 超链接的分析与处理

主题搜索器在主题搜索的过程中，首先从一个初始的超链接集合出发，将这些超链接全部放入到一个有序的待提取超链接队列里，然后在从这个队列里按顺序取出，通过 Web 上的协议，获取超链接 URL 所指向的页面，再从这些已获取的页面中分析提取出新 URL，并将它们继续放入到待提取 URL 队列里，然后重复上述过程，直到 Web 信息提取器根据自己的搜索策略停止采集为止，可以看出，超链接是主题搜索器中最为关键的信息。

在 HTML 页面中，主要有四种标签用于超链接：（1）A 和 Base 标签；（2）Img

标签；（3）Map 和 Area 标签；（4）Frame 和 Iframe 标签。在这四种标签中，A 和 Base 标签最常用，常用属性有 Href,Title,Alt 和链接文本；Img 标签和超链接相关的属性有 Src 和 Alt；对于 Map 和 Area 标签，常用属性和 A 标签基本相同；Frame 和 Iframe 一般与 Frameset 一起使用，用于网页进行分割，常用属性有 Src 和 Name 等。由于我们设计的这个主题搜索器的最终目标是提取出包含教育多媒体的页面，这些多媒体数据可以以链接的形式出现，也可以嵌入到 Web 页面中，以链接形式出现的多媒体数据大多包含在标签<A>中，嵌入到 Web 页面中的多媒体数据包含在标签<Object>和<Embed>中，所以除了上面提到的四类标签外，还需要提取包含嵌入式多媒体链接的标签 Object 和 Embed；由于<Embed>标签未被 w3c 收录，所以在 Swing 包中并不支持这个标签，因此只通过<Object>标签来提取（多媒体文件的路径包含在这个标签的<Param>标签中）。

URL 全称是 Uniform Resources Locator，翻译成中文是统一资源定位符，它用于完整地描述 Internet 上网页和其他资源地址的一种标识方法。Internet 上的每一个网页都具有一个唯一的名称标识，也就是我们通常所说的 URL 地址。超链接 URL 的格式分为两类：（1）Protocol://hostname:port/path/?Query.（2）Protocol://hostname:port/path/#anchor，这种格式被称之为绝对 URL，它指向一个确定、无歧义的 Internet 资源位置。但是由于从 HTML 解析器中提取出来的 URL 大部分不是上述格式的绝对 URL，而是相对 URL，相对 URL 总是与它所在的网页结合在一起，一旦离开它所在的网页，该相对 URL 便没有了意义。因此为了使每个链接具有独立性，必须把相对 URL 解析成绝对 URL。解析相对 URL，可以分为两种类型：（1）相对 URL 以斜杠（/）开始，它表示“直接来源于主机”，这意味着把主机名直接加在该相对 URL 路径的前面。（2）不以斜杠（/）开始，这种类型的相对 URL 在解析时，将把该相对 URL 直接连接到正在查看页面所在的目录；在这种类型中，有两种特殊情况，一种是以“..”开始，这表示父目录；另一种情况是以“.”开始，这表示当前目录。

4.2.1 Swing 包中使用的核心类

本文使用本实验室其他人员开发的 HTML 解析器系统，系统的实现是基于 JAVA 语言中 Swing 包中的几个核心类，它们分别是 HTMLEditorKit.Parser、HTMLEditorKit.ParserCallback、HTML.Tag、HTML.Attribute。HTMLEditorKit.Parser 是 HTMLEditorKit 类的内部类，使用它的关键是将其实例化。但 HTMLEditorKit 类并没有

提供获得 Parser 类的公共接口，因此必须重载 HTMLEditorKit 类的 get Parser 方法来实例化 HTMLEditorKit.Parser 对象。然后调用该对象的 parse 方法，该方法只是简单的从头开始读取 HTML 代码，并将在读取文档的过程中遇到标签、文本或注释发送到回调类 HTMLEditorKit.ParserCallback，然后调用相应的回调方法进行实际的解析，但是由于这些回调方法原来只是空实现，因此需要解析人员自己对回调类中的回调方法进行重写来处理自己感兴趣的数据。HTML 解析器具体的结构请参照参考文献[30]。

4.2.2 页面解析器的实现

在本节对参考文献[30]的 HTML 解析器进行介绍，系统设计的 HTML 解析器大体分五部分：协议转换器、网页读取器、信息提取器、URL 路径转化器、信息存储器。实现过程如下：

（1）设置一个时间变量，用来保存访问该页面的时间，并将这个时间变量值存入数据库；设置一个布尔型变量 isHTTPS 来判断该站点是否使用 HTTPS（超文本传输安全协议）来与浏览器交换数据；

（2）若 isHTTPS 值为 true，将 URL 中 protocol 的值由 HTTPS 变为 HTTP，将端口号设为 443，并用 SSL（加密套接字协议层）建立与服务器端的连接，否则，只需简单的创建一套接字连接即可；

（3）向指定 URL 的网页服务器发送 HTTP 请求；

（4）读入服务器端的响应标题，并判断标题域 Location 的值是否为空，若不是，则将该标题域 Location 的值作为当前要访问网页的 URL 即进行 URL 的重定向，转 2；若标题域 Location 的值为空，则转 5；

（5）通过 URLConnection 建立与 URL 指定的数据源的动态链接；

（6）若连接成功，将网页 HTML 代码通过输入流读入阅读器 Reader 中，转（7）；否则，则生成错误信息，将这个错误信息存入数据库；

（7）实例化解析器类 HTMLEditorKit.Parser 对象，然后调用其 parse()方法，该方法将从 Reader 中读取 HTML 代码，直到读完整个文档；

（8）实例化回调类 HTMLEditorKit.ParserCallback 对象，并将其作为参数传递到解析器 HTMLEditorKit.Parser 对象的 parse()方法中；该类将进行实际的 HTML 的解析工作。

（9）若遇到标签<Meta>，则调用回调函数 handleSimpleTag，提取出该标签属性

name 值，若值是 Keywords 或者 Description ，则提取出 content 的值存入数据库；若遇到标签<Title>，则调用回调函数 handleText,提取出网页的标题存入数据库；

（10）若遇到标签<A>、<Area>、<Map>，则调用回调函数 handleStartTag 和 handleEndTag,提取出该标签的属性 HREF 值；若遇到标签<Iframe>，则调用回调函数 handleStartTag 和 handleEndTag,提取出该标签的属性 Src 值；若遇到标签<Base>，则调用回调函数 handleSimpleTag，提取出该标签属性 Href 值；若遇到标签<Frame>，则调用回调函数 handleSimpleTag，提取出该标签属性 Src 值；若遇到标签<Img>，则调用回调函数 handleSimpleTag，提取出该标签属性 Src 值；若遇到标签<Param>，则调用回调函数 handleSimpleTag，提取出该标签的属性 VALUE 值，而每个<Object>标签一般会嵌套多个<Param>标签，在此只提取 VALUE 值是链接地址的那个属性；

（11）若提取出的 Href 值、Src 值或者 Value 值的扩展名不是.html 或者.htm 或者.shtml 或者.asp 或者.jsp，则直接放弃；否则，如果是上述格式中的一种且是相对 URL 地址，需要通过该网页的链接地址利用 URL 类提供的构造函数对该相对 URL 地址进行解析，将其转化成绝对 URL 地址，以便离开该网页，该 URL 地址仍然可以被正确访问；

（12）为了对该链接指向的网页内容有一个大体的预测，提取出锚文本（锚文本是超链接的载体，用户点击它就可以链接到一个新的页面或同一页面的另一个位置）作为该链接的内容描述；由于只有锚标签才可能有锚文本（因为若以图片作为超链接的载体，则该锚标签就不会有锚文本，在此称该图片为链接图片），所以只对锚标签的锚文本进行提取，其他的标签或以链接图片作为链接载体的锚标签，则将其链接文本设为 null；提取锚文本的方法如下：设置一个布尔型变量 isLink，当遇到锚标签时，就将该变量的值置为 true，然后 HTMLEditorKit.Parser 对象处理到链接文本时，调用回调类的 handleText 方法，只有当 isLink 的值为 true 时，该文本才被读出，处理完此文本后，变量 isLink 自动复位为 false,从而实现链接 URL 值与锚文本之间的一一对应关系；

（13）构造一个链接的容器类 Link，用来存放链接 URL 值与锚文本，一方面在存储上保持了链接 URL 值与锚文本之间的一一对应关系，另一方面也为以后的访问带来方便，将链接 URL 值与锚文本一起存入数据库。

4.2.3 页面数据的组织

已提取 Web 页面的相关信息和网页的超链接一起存入链接数据库中。所提取的 Web

页面相关信息包括：网页URL经中英文翻译后的词语、链接的锚文本、网页标题、网页中所包含的所有链接URL、网页的物理层和逻辑层数、提取时间等，这些信息在计算网页内容相似度和链接相似度时都会用到。每个页面对应记录中的一条记录，在后续的处理过程中可以从中获得必要信息。链接数据库的具体逻辑结构将在5.2.2节中进行详细介绍。

4.3 基于URL规则的链接相似度计算

4.3.1 多媒体资源主题相邻性

在同一网站中，包含多媒体的主题网页表现出“资源相邻性”的特点。也就是说，包含多媒体资源的网页往往存在于这个网站的某一部分或某几部分区域中。并且处于同一区域的多媒体资源的主题也是相同的。根据此特点我们可以做出如下假设：

（1）如果一个网页是与主题相关的包含多媒体的网页，那么此网页的子链接很可能是与主题相关的包含多媒体的网页；

（2）如果一个网页是与主题相关的包含多媒体的网页，那么此网页在父网页中的兄弟链接很可能是与主题相关的包含多媒体的网页。

具体而言：包含多媒体的网页一般都存在于网站的某一个或几个目录下，而且目录与目录之间的差别明显。同一个网站中一个目录下可能包含大量的多媒体，而另一个目录下很少、甚至不包含多媒体。

1. 静态网站

（1）网站http://www.fxzx.net/中所有的多媒体内容都存放在http://www.fxzx.net/bxzy/目录下。

（2）网站http://teacher.cqbxzx.com包含多媒体的网页目录为http://teacher.cqbxzx.com/xy/jc/、http://teacher.cqbxzx.com/xy/wxs/和http://teacher.cqbxzx.com/xy/wz/，其余的目录下就很少包含多媒体。

2. 动态网站

通过观察，网页http://www.cbe21.com/subject/physics/courseware.php?photo_id=100003中包含多媒体，那么以此模板产生的http://www.cbe21.com/subject/physics/courseware.php?photo_id=100002和http://www.cbe21.com/subject/physics/courseware.php?photo_id=100001也包含相似内容的多媒体资源。

以上分析可以得出：同一个网站中属于同一主题的页面的URL有着很强的相似性，

也就是说这些网页的URL可以用一个(或者若干个)正则表达式来概括。如：

http://teacher.cqbxzx.com/xy/* ；

http://www.cbe21.com/subject/physics/courseware.php?photo_id=* 。其中星号代表全部网页的不同部分。有了这个正则表达式，我们可以根据正则表达式来指导主题蜘蛛的爬行方向，判断一个未知的URL包含多媒体的可能性。

为了更好的验证算法的可行性，我们人工选取10个高中物理网站作为种子链接，开辟了10个线程，以盲搜索的方式搜集包含多媒体Flash的网页，实验结果如表4-1。

表4-1 盲搜索实验结果

Flash			
种子个数		10	
内部网页总数		38820	
错误网页数		2045	
有效网页（包含 Flash 的网页）数		1316	
效率		3.390%	
共有 Flash 文件个数：5089			
物理层		没有限制	
物理层数	共有网页	有效网页	有效网页/总有效网页
1	621	3	0.2280%
2	1010	72	5.471%
3	5250	178	13.53%
4	1824	103	7.827%
5	18851	775	58.89%
6	730	20	1.520%
7	2734	84	6.383%
8	2129	64	4.860%
9	3498	5	0.3800%
10	1325	8	0.6080%
11	838	0	0%

通过分析实验结果可以得出：Flash 文件主要集中在网站物理层的 2—9 层且往往集中在某一物理层上。同时，我们爬取了 http://www.wsbedu.com/（为您服务教育网）种子网站，具体实验结果如表 4-2。

表 4-2 搜索“为您服务教育网”实验结果

Flash			
种子网站	http://www.wsbedu.com/		
共有内部网页	9817		
错误网页	1141		
有效网页（包含 Flas 的网页）	435		
效率	5.01%		
共有 Flash 文件个数: 1121			
物理层		没有限制	
物理层数	共有网页	有效网页	有效网页/总有效网页
1	1	0	0%
2	393	9	2.06%
3	1390	69	15.86%
4	3805	124	28.5%
5	4046	229	52.64%
6	1110	4	0.91%
7	72	0	0%
8	0	0	0%

根据表 4-2 的实验结果，我们可以看出包含 Flash 多媒体的网页绝大多数都分布在表 4-3 中的 4 个目录下。主题搜索器在搜索过程中，如果能抽取出这 4 个目录的 URL 正则表达式，然后通过分析计算出搜索网页链接 URL 与抽取出的 URL 正则表达式之间的距离。通过距离的大小来判断待搜索网页包含多媒体网页的可能性，就能有效地提高多媒体资源主题搜索的效率。

表 4-3 多媒体资源分布目录

物理路径	有效网页	有效网页/总有效网页
http://www.wsbedu.com/wu51/*	69	15.86%
http://www.wsbedu.com/lin/hs/*	209	48.04%
http://www.wsbedu.com/kejian/shengk/*	52	11.95%
http://www.wsbedu.com/mid/soft/*	34	7.81%

4.3.2 解决隧道（Tunnel）问题

上面提到，只要找出代表包含多媒体主题页面的URL的正则表达式似乎就可以很好的指导主题搜索器抓取包含多媒体的主题相关页面。但是现实情况不是这样的，仅仅只有包含多媒体主题相关网页的正则表达式是不够的。在第一章中提到，互联网存

在隧道问题，这对于同一个网站内的网页也是适用的。通往包含多媒体主题相关页面往往要先经过主题无关页面，而且一个网页的首页和首页链接出去的页面往往不是包含多媒体主题相关页面，如果直接用代表包含多媒体主题相关页面的正则表达式预测这些页面的URL，那么毫无疑问这些页面都不是包含多媒体主题相关页面，他们的URL与包含多媒体主题相关页面的URL正则表达式是不匹配的，也就不会去抓取这些页面，这样就不能通过这些页面抓取到更深层次的包含多媒体主题相关页面。因为包含多媒体主题相关页面只有通过先抓取这些页面才能抓取到。

为了解决这个问题，主题搜索器在搜索过程中，先采用宽度优先策略对种子网页所链接的每个内部链接都进行抓取。在URL正则表达式学习阶段，主题搜索器主要目的是探测多媒体网页所在的物理目录和物理层数，形成URL正则表达式。由于起初阶段，并不清楚多媒体在整个网站中的位置，在搜索过程中，必须保证搜索的宽度和深度。为了解决这个问题，主题搜索器在搜索的过程中，对同一物理层的网页URL只取*NUM*条记录进行网页分析。由于考虑到包含多媒体的网页一般分布在某一个目录的最底层，而且当网页的物理层数n>6时，网页的数量骤减。所以在公式中没有将这方面的因素考虑在内，直接取n的正比例函数。我们以第2个物理层抓取的网页数量为基准，*NUM*的数值如公式4.2。实验证明该公式能够发现更多的多媒体网页。

$$NUM_{Pn} = NUM_{(P2)} * (n-1) \quad (n \leqslant 10) \tag{4.2}$$

NUM_{Pn}表示物理层n应该抓取的网页，$NUM_{(P2)}$表示第2个物理层抓取的网页数。n表示物理层数，当n=10 时停止盲搜索。

4.3.3 URL 数据结构

在介绍URL正则表达式学习算法之前，我们先介绍一下URL的数据结构。这是比较URL相似度的关键，我们把一个URL分成三个部分（去掉http协议部分）：host，path，query。其中path由一系列directory组成，query由一系列键值对组成。比如URL http://www.cbe21.com/subject/physics/courseware.php?photo_id=100002，其中host为www.cbe21.com；path为/subject/physics/courseware.php，组成该path的directory为subject，physics，courseware.php；query为photo id=100002，组成该query的键值对为（photo_id，100002）。用Java表示URL数据结构如图4-1。

```
public class UrlStruct{
    private String host;
    private String[] path;
    private ArratyList <Pair<String ,String>> query;
}
```

图 4-1 URL数据结构

4.3.5 URL 距离的度量

把一个URL解剖成上面的URL数据结构后我们就可以基于这个URL数据结构各个部分的距离来计算URL的距离[49]。用i，j表示两个网页的URL，它们之间的距离用$dURL_{ij}$表示，如公式4.3。

$$dURL_{ij}=(dHost_{ij})*(dPath_{ij}+1)*(dQuery_{ij}+1)-1 \quad (4.3)$$

其中$dHost_{ij}$表示网页i，j的host部分的距离，$dPath_{ij}$表示网页 i，j的path部分的距离，$dQuery_{ij}$表示网页i，j的query部分的距离。这三部分的距离的计算步骤如下：

(1) 如果两个URL的host部分不相同，则不进行距离判断。

（2）如果网页i，j的path部分分别有m个directory和n个directory，设$m \leq n$，记前m个相应位置上的不等的directory的数量为k，则

$$dPath_{ij}=k\times 4+(n-m)\times 2 \quad (4.4)$$

（3）对$dQuery$进行简单的设置，若相等，则$dQuery_{ij}$为0，否则$dQuery_{ij}$为1。因为对于大多数网站，如果其他都部分都相等，只有query部分不等，则这两个URL所指的页面基本都是同一个模板产生的，他们一般也是属于同一类页面。

4.3.6 URL 正则表达式的抽取

对于URL距离满足一定条件的URL簇，进行抽取以获得URL正则表达式[49]。抽取的具体过程为：首先把同一站点的URL分解为host，path和query三部分，并把path分解成一系列directory，把query分解成一系列键值对。由于host部分肯定是相同的，所以就把host按照原样记在正则表达式中。把path部分的各个directory对齐，若相应位置上的directory相同则把这部分值加进正则表达式中，否则用*号代替加入到正则表达式中。对于query部分也采用和path部分类似的方法。最后我们就可以得到一个正则表达式。其流程图如图4-2。

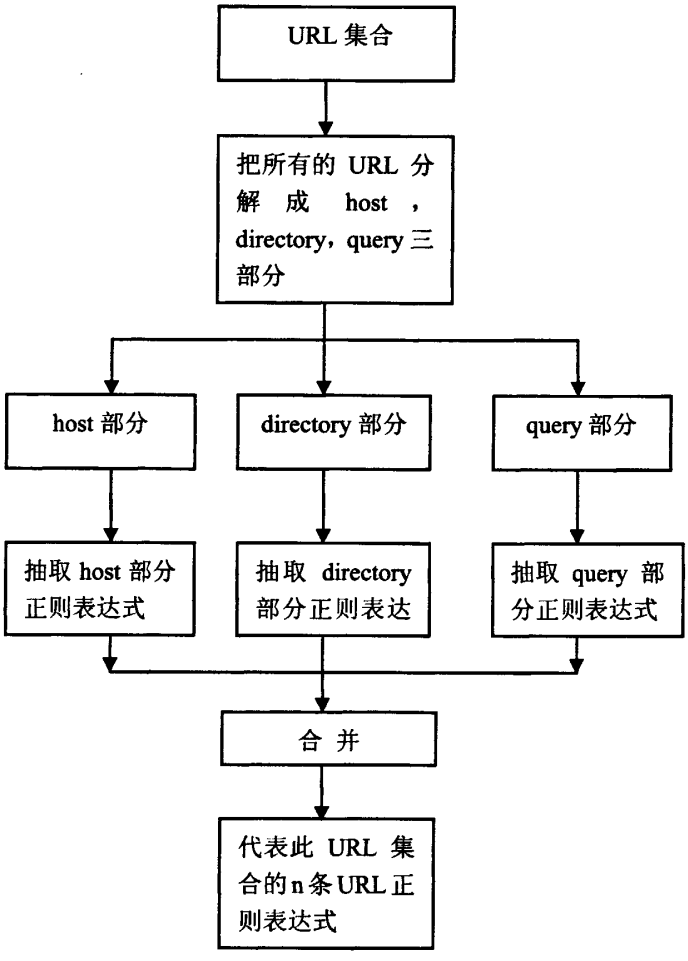

图 4-2 URL 正则表达式的抽取过程

4.3.7 基于 URL 规则主题搜索的实现

基于上述事实，我们提出一个基于URL规则的主题搜索的解决方案，该方案大致有二步组成：

一是实验搜索阶段，对于每个种子网站，首先进行盲搜索，从搜索到的包含多媒体的网页中学习出一些URL正则表达式。

二是指导搜索阶段，该阶段利用从实验爬虫阶段学习出的URL正则表达式来指导爬虫进行实际的网页抓取。

实验搜索阶段流程如图4-3。

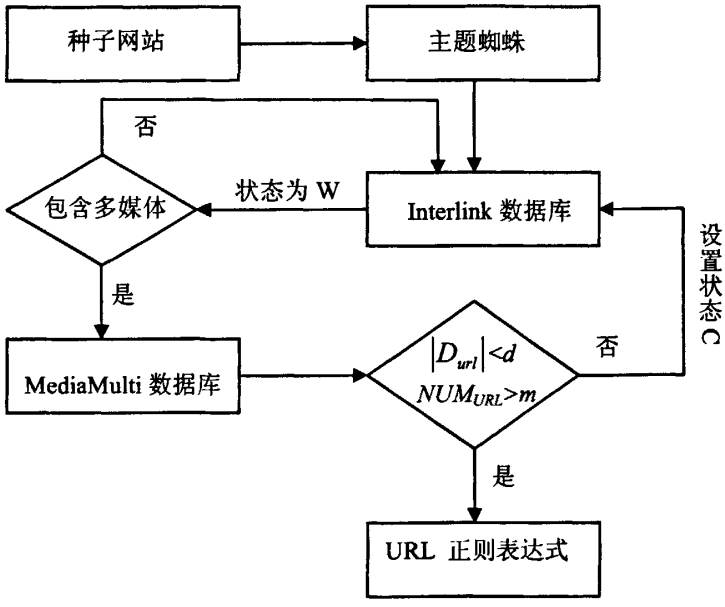

图 4-3 实验搜索阶段

种子网站列表作为搜索的起始站点，主题搜索器在网页爬取过程中，不断地获取网页链接并将其存入 InterLink 数据库中，然后对于状态为 W 的链接，首先判断网页主题的相关性和网页中是否包含多媒体资源，将满足这两个条件的链接存入 MultiMedia 数据库中。当 MultiMedia 数据库中的网页数量 $NUM_{URL}>m$，且网页 URL 之间的距离 $|D_{url}|<d$ 时进行 URL 正则表达式的学习。具体学习的过程见 4.3.6 节中的 URL 正则表达式的抽取。

由于在实际的实现中，由于主题判断方面的误差，进行正则表达式学习的网页 URL 可能是少量与主题无关页面。这样最终从一个网站中学习出来的代表主题相关的 URL 正则表达式中，可能就存在一些实际上代表主题无关页面的 URL 正则表达式。比如，某一网站导航条中，任何一个网页中都插入一个 swf 格式的动画多媒体，这样主题搜索器在搜索时就会将所有的网页都加入到 MultiMedia 数据库中，从而造成 URL 正则表达式的学习的不准确。所以程序在实际的运行过程中，我们还要加入一定的限制条件，对于多次出现在 MultiMedia 数据库中的网页 URL 进行剔除。

指导搜索阶段流程如图4-3。

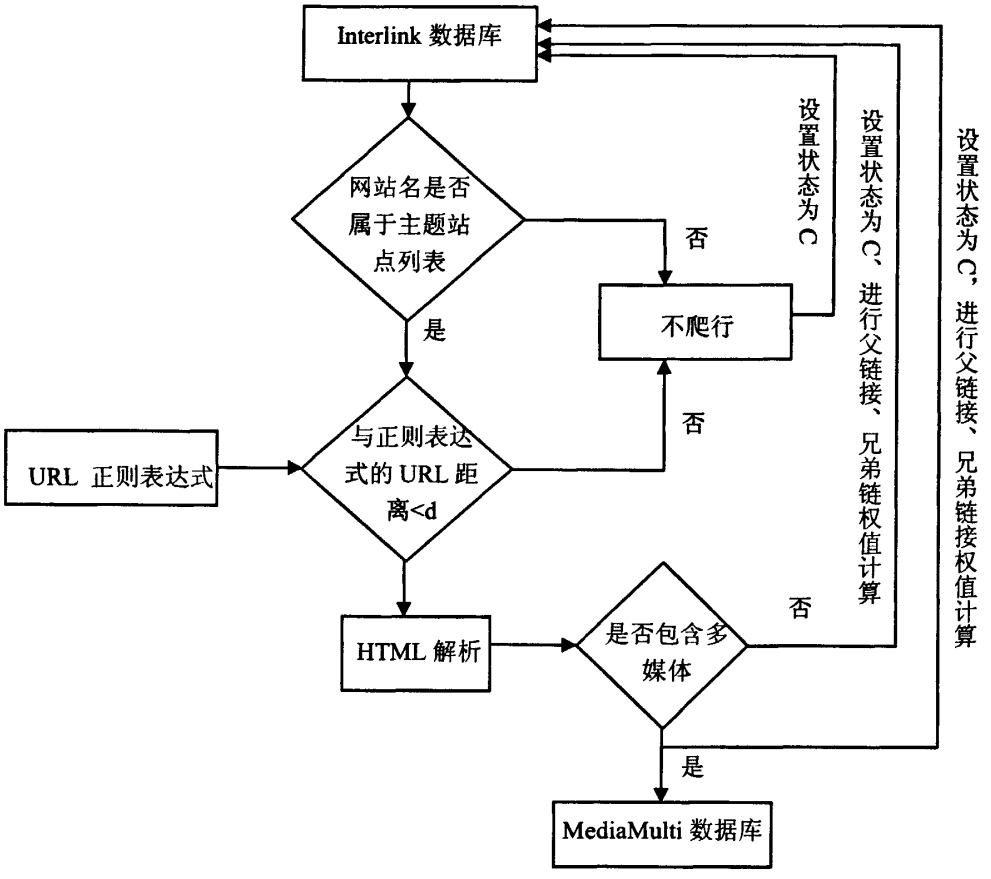

图 4-4 指导搜索阶段

从上一步我们可以得到每个网站中的URL正则表达式，有了这些URL正则表达式，我们就可以用它们来指导下一步网页链接的爬行方向。具体实现过程为：

（1）从InterLink数据库中取出状态为W的网页链接，判断新的URL的domain是不是属于主题站点列表，如果不是则不抓取，结束流程；

（2）从URL正则表达式列表中寻找一条与之匹配的正则表达式，如果不存在则不抓取，将其在InterLink数据库中的状态设置为C，结束流程；如果存在多条与之匹配的URL正则表达式，则选择那条最优的，即归纳URL数量最多的。

（3）对匹配的网页URL链接进行HTML解析，判断此网页是否包含多媒体，如果包含则将此网页存入MultiMedia数据库中，同时将其在InterLinks数据库中的状态设置为C，并进行父链接、兄弟链接权值计算，以指导下一步的搜索方向；如果不包含，则对此网页进行父链接和兄弟链接权值计算后加入到InterLink数据库中，并将其状态设为C。

4.3.8 基于 URL 链接规则相似度的计算

以上介绍了URL正则表达式的实现和计算URL距离时使用的数据结构，通过这些我们能够判断出待爬行队列中的URL和正则表达式之间的距离，从而计算出待爬行队列中每个链接的链接信息得分，确定每个链接爬行的优先级。在链接结构方面，主题网页表现出“资源相邻性”的特点，包含多媒体的网页往往存在于这个网站的某一部分或某几部分区域中。并且处于同一区域的多媒体网页主题也往往是相同的。所以在计算网页链接相关度时，我们用父网页和兄弟网页的链接相关度来揭示链接结构对一个URL链接相关度的影响。

1．PageRank算法的缺陷分析

首先看一下 PageRank 算法的公式：

$$PR(p)=\frac{1-d}{N}+d*\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \qquad (4.5)$$

我们发现在公式 4.5 中每一个指向页面 p 的页面 T_i，它的重要度平均地传给了此页面中的每一个链接指向的页面，也就是说每个子链接都平均地接受父链接的 $1/C(T_i)$ 的页面重要度。我们认为，对基于主题的重要性来说，这是不合理的，而应该是跟链接连接到的页面主题相关度高低成比例的。

2．对 PageRank 算法的启发式步骤进行改进

在公式 4.5 中计算页面的 PageRank 值时，每个页面的初始化 PageRank 值都相同，该方法关注的仅仅是页面之间的链接关系，不关注语义方面的含义。因此，初始时每个页面的权重均是相同的。而且，它对链接也没有进行主题语义相关的加权处理，页面重要度基于链接的传递都是平均的。而对于基于 URL 链接规则相似度的计算算法来说，我们认为如果一个网页 URL 与正则表达式的距离越小，那么它越接近包含多媒体的区域，在网页链接权值传递的过程中，应该赋予较高的权重。同时将兄弟链接对此网页的链接权值传递也加入到算法中来。

具体公式为：

$$PR(p)=\frac{1-d}{N}+d*(\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}+\sum_{j=1}^{m}\frac{PR(T_j)}{C(T_j)})*\frac{1}{|D_{url}|} \qquad (4.6)$$

其中 $\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$ 表示父链接对子链接权值的传递；$\sum_{j=1}^{m}\frac{PR(T_j)}{C(T_j)}$) 表示兄弟链接对网页

p 权值的传递。$|D_{url}|$ 表示网页 p 与 URL 正则表达式的距离。

4.4 网页内容相似度的计算

在这里我们使用改进 Shark-Search 算法中网页内容相似度的计算公式，网页链接按照 table 标签分为不同的“主题团”，将“主题团”标题与主题的相关度作为待爬行链接中锚文本和 URL 地址的权重。这样待爬行链接的主题相关度为：

$$Content\_score(u_i) = Score(block\_title)[\beta * Score(anchor) + (1-\beta) * Score(url)] \quad (4.7)$$

算法具体的实现过程见第 3.4.2 节。

4.5 基于 URL 链接规则的多媒体主题搜索算法的具体描述

基于 URL 链接规则的多媒体主题搜索算法的具体工作过程，首先为主题搜索器输入种子集 U_B 并赋予较大的链接初始值Q，搜索深度 D，主题集 T，然后进行以下步骤。

（1）建立两个数据库：链接数据库 InternalURLs、多媒体数据库 MultiMedia；将种子集 U_B 放入 InternakURLs 中，并设其状态为 W（准备）；

（2）从链接数据库 InternalURLs 中提取状态为 W 的链接 J_u；

（3）对此链接进行源文件下载和 HTML 解析，提取出 J_u 链接中的子链接列表 L_U；

If J_u 的物理层数 n>10

转到步骤 2；

Else If （网页 J_u 包含多媒体&&网页 J_u 与主题相关）

将网页 J_u 加入 MultiMedia 数据库中，根据公式 4.3 进行 URL 正则表达式抽取；

Else

转到步骤 2；

（4）根据公式 4.3、4.6 计算 L_U 中各链接的链接结构相似度 S_U，将 L_U 中的链接和相应的 S_U 按优先级存入 InternalURLs 数据库中，将链接的状态设为 W；

（5）按优先级提取 InternalURLs 数据库中状态为 W 的链接 I_U；

While（I_U 不为空&&爬取层数<D && $I_U \in$ URL 正则表达式）

（6）对链接 I_U 进行源文件下载和 HTML 解析；

If I_U 中不包含多媒体

提取 I_U 中子链接列表 L_U；

转到步骤 4；

Else

I_U 包含多媒体；

（7）分析链接 I_U，提取 I_U 中的“主题团”标题 B_u，根据公式 4.7 计算“主题团”标题 B_u 与主题 T 的内容相似度 S_B；提取“主题团”中的链接 URL，分别根据向量空间模型公式 3.13 计算每个链接的链接内容相似度 C_U；

If C_U>某一特定值

将 I_U 存入 MultiMedia 数据库中；

Else

转到步骤 2；

（8）End While；

4.6 内容相似度和 URL 链接规则相似度的归一化

为了提高算法的查准率和查全率，算法在计算过程中将网页内容和网页链接都加入到计算过程中，采用一定的动态因子调节网页内容相似度和网页链接相似度在网页相似度中所占的比重。在这里将二者归一化，计算得到的值作为主题搜索器即将爬行链接的依据。由于第三章的Topic-PageRank算法、改进的Shark-Search算法也要进行内容相似度和链接相似度的归一化。因此三种采用统一的计算公式，如公式4.8。

$$S_{(i)} = \partial * Content\_score(u_i) + (1-\partial) * PR(p) \qquad (4.8)$$

4.7 本章小结

本章介绍了在网页中存在的多媒体资源往往呈现出“资源相邻性”的特点，依据此特点提出一种基于URL链接规则的多媒体主题搜索算法，详细介绍了算法实现过程

中的文本信息处理和链接结构的分析和处理；然后详细地阐述了算法中URL正则表达式的实现过程以及如何使用URL正则表达式来指导主题搜索器获取网页链接，并给出了基于URL链接规则的多媒体主题搜索算法的具体描述。最后，为了提高多媒体主题搜索的准确率，算法将进行内容相似度和URL规则链接相似度的归一化计算。

第五章 实验系统与测试分析

5.1 简述

本文的研究工作是山东省教育厅科技计划项目——基于语义的Web中多媒体信息个性化搜索系统研究（网络多媒体搜索模块开发与系统集成）的一部分。根据前文提出的算法，在WinXP操作系统下，在已有的多媒体主题搜索器基础上对Topic-PageRank算法、改进Shark-Search算法、基于URL链接规则的多媒体主题搜索算法进行了实现，通过实验结果，对算法的搜索效率作出评价和分析。

5.1.1 主题词集的建立

为了确定在Web中搜索到的多媒体资源的主题，我们从人教版中小学课本中提取了与多媒体资源可能有关的主题词，如表5-1。如果表征多媒体内容主题词和多媒体主题词集有相交的词语，那么搜索到的多媒体就属于相互交叉的学科。不同的学科之间有一些主题词有相互交叉的现象，如表5-2。同时为了满足更多教育工作者的需要，我们设置(http://www.cbxy.sdnu.edu.cn/cbxy/wordup/index.asp)搜集页面，用户可以根据学段、学科加入主题词，我们收集后会加入到相应的主题词集中，用于指导多媒体主题搜索器的爬行。

表 5-1 基础教育网络多媒体教学资源主题词集记录数

	小学				初中					高中				
类型	语文	数学	科学	社会	语文	数学	物理	化学	生物	语文	数学	物理	化学	生物
视频	165	141	72	49	138	49	83	40	153	123	27	76	116	60
音频	90	8	5	5	180	19	22	25	19	152	30	12	18	8
动画	308	285	63	56	149	304	94	40	120	167	72	98	154	41

表 5-2 网络多媒体教学资源主题词集中词语交叉示例

绿色食品 分类/类型	小学				初中					高中				
	语文	数学	科学	社会	语文	数学	物理	化学	生物	语文	数学	物理	化学	生物
视频	√		√	√	√			√	√	√		√	√	√
音频					√		√		√					√
动画			√	√	√			√	√	√		√		√

系统在运行时用到的词集还有：中英文翻译词典（共有词17323条）负责对搜索到包含多媒体的网页URL进行中英文翻译，为了提高翻译的准确性和扩充面，通过分析一般的多媒体网页，我们还搜集了一些常用的英文标识，比如：wuli — 物理；中文分

词词典（共有词85406条）对提取的网页 <Title>标签的主题、网页锚文本等进行中文分词，然后再和主题词集进行主题词的匹配，匹配方法采用布尔模型和向量空间模型。

5.1.2 主题搜索器的体系结构

整个搜索器分为在线训练和离线训练两部分，其中在线训练主要负责网页信息提取和待爬行队列中 URL 的选择，离线训练主要负责 URL 规则训练和网页相关度的计算。其处理流程按照如图 5-1 所示的主题搜索器的体系结构进行。

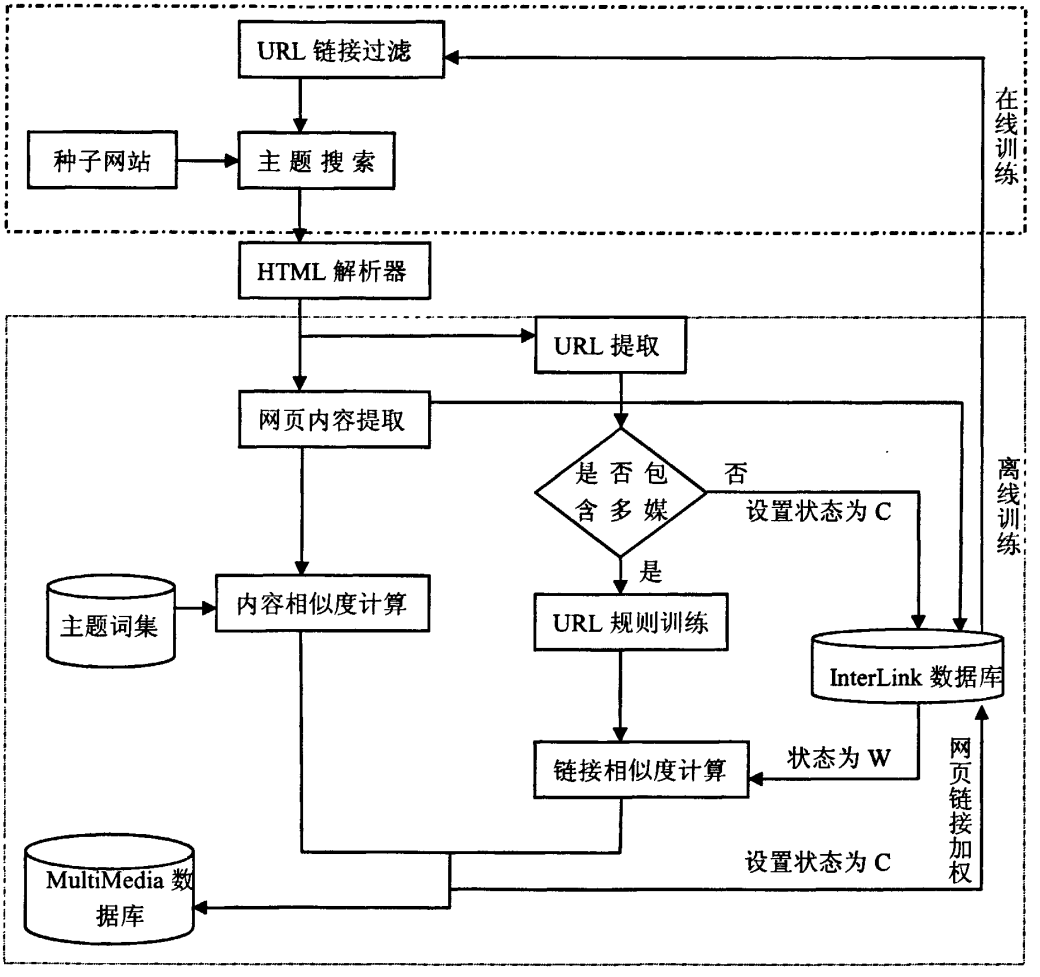

图 5-1 主题搜索器体系结构

“主题搜索器”负责从互联网上抓取网页，提取网页的两部分信息：（1）获取网页内容，以此来确定此网页与查询多媒体主题的相关度；（2）提取网页链接，确定主题搜索器的即将爬行页面。“主题搜索器”通过“HTML解析器”获取此网页的文本信息，

并将其和网页链接存入“InterLink数据库”中。“网页内容相似度计算”用于判断此网页和查询多媒体主题的相关度；“HTML解析器”同时提取出网页的URL，然后判断此网页是否包含多媒体，如果包含则进行“URL规则训练”，将其用于“链接相似度计算”，经过内容和链接相似度计算后的网页链接与“InterLink数据库”中的网页链接进行“网页链接加权”，从而确定下一步要搜索的网页，经过“URL链接过滤”确定“主题搜索器”的爬行方向；如果不包含多媒体，则直接将提取的网页链接存入“InterLink数据库”中，等待“网页链接加权”对其进行权值的分配。最终搜索得到的多媒体链接存入“MultiMedia数据库”中，同时也要将其存入“InterLink数据库”中用于指示“主题搜索器”下一步搜索的方向。

5.2 主题搜索算法性能的评价

5.2.1 测试页面集的选择

种子页面是主题爬行的起点，往往种子选择的好坏也关系到爬行的效果，好的种子集可以使爬行尽快的进入与搜索主题相关的领域，避免了爬行起始阶段的盲目性。本文选择初始种子页面的过程主要分两步：首先通过Google官网的搜索页面进行某一学科多媒体资源网站的通用搜索；然后在返回的结果集中加入用户的参与，进行初始种子的选择。在人工选择种子页面的时候，选择那些标题和主题有关的并且包含丰富多媒体资源的网页。本文选择30个与物理有关的网页（10个教育综合性的网站、10个中学门户网站、10物理多媒体专题网站），然后随机选择10网站作为种子链接，如表5-3。

5.2.2 数据库设计

本系统的运行需要三个数据库：用户注册信息数据库UserInfo、提取的网页链接数据库InterLink数据库、搜索到的多媒体网页数据库MultiMedia数据库。其中InterLink数据库、MultiMedia数据库都分别包含三个数据表：Audio、Video、Flash，分别用来存放从音频、视频和动画的网页信息。MultiMedia数据库主要存放搜索到的多媒体信息，其中的字段包括：网页的URL地址、网页的HTML代码、多媒体的数量、访问时间等，在此就不一一介绍了。InterLink数据库详细的字段如表5-4。

表5-3 实验种子链接

网站名称	网站URL地址
新课标资源网	http://www.k868.cn/Index.html
初中物理在线	http://www.czwlzx.com/
国家基础教育资源网	http://www.cbern.gov.cn/
物理资源网	http://physweb.51.net/
中学物理网	http://wuli.jiaoyudaohang.com/kejian/Index.asp
南海一中	http://www.nhyz.org/
肥西中学	http://www.fxzx.net
谢恩在线物理教学网	http://xieenonline.51.net
中国基础教育网	http://www.cbe21.com/
八九物理资源网	http://www.szxcedu.com:8001/physource/

表5-4 InterLink数据库

字段名称	数据类型	字段意义	备注
ID	自动编号	数据库中记录的个数	主键
URL	备注	链接的URL地址	
Status	文本	链接的访问状态	W：没有访问 R：正在访问 C：访问完成 E：访问出错
URLFanyi	备注	URL经中英文翻译之后的词语	
Anchor	文本	链接的锚文本	
Title	文本	网页标题	
Keywords	备注	网页元标记中的关键词	
Description	备注	网页元标记中的网页描述	
FileContentIntroduce	备注	URLFanyi+Anchor+Title +Keywords+Description	
AnchorPath	备注	从网页入口到该网页所经过的所有锚文本	
URLPath	备注	从网页入口到该网页所经过的所有链接URL	
InternalURLs	备注	网页中包含的所有URL	
PathScore	数字	经过该网页的所有网页中包含多媒体网页的个数	
TotalPathScore	数字	经过该网页的所有网页个数	
Contain_or_not	文本	是否包含多媒体	

ContentScore	数字	FileContentIntroduce 的主题相似度得分	
linkStructureScore	数字	链接结构得分	
LinkRatio	数字	pathScore 与 totalPathScor 的比值	
Hyper_Links	数字	网页中链接的总数量	
L_Layer	数字	网页的逻辑层	根据从网站首页到达该网页的步数进行计算
P_Layer	数字	网页的物理层	根据 URL 地址中的“/”个数进行计算
IsChecked	文本	该网页是否已经爬行	根据网页的状态判断
VisitedTime	日期/时间	访问网页的时间	
ExMessage	文本	访问网页时的出错信息	

5.2.3 测试指标

系统采用在搜索引擎领域中两个非常重要的测试指标：信息提取的准确率（Precision），和信息提取的资源召回率（Recall）。

提取准确率*Pr*定义为：

Pr=已提取页面中主题相关的页面数目/所有提取的页面数目

资源召回率*Re*定义为：

Re=已提取页面中主题相关的页面数目/测试集中主题相关的页面数目

5.2.4 测试方法

为了对算法性能进行定量比较，从相同的初始URL集合出发，分别使用通用搜索算法、基于Topic-PageRank算法、基于改进Shark-Search算法、基于URL链接规则算法对信息进行提取。为了保证主题搜索的效率，减少算法的计算复杂度，在测试时，判断主题内容相似度时，首先对URLFanyi、Anchor的文本内容进行中文分词，统计词语的词频，根据布尔模型计算内容相似度。并根据词语在HTML不同位置赋予相应的权重因子。中文分词采用MM（正向最大匹配）算法。

5.2.5 相关参数的设定

（1）基于Topic-PageRank主题搜索算法相关参数的设定

此算法在计算网页内容相似度时，对于公式 3.8，不同位置的文本内容对网页的标识程度不同，为 Anchor 、URLFanyi 两部分文本内容分配不同的权重α：β=6：4。在

公式 3.6 中，参数d取值 0.8，在公式 3.11 中，由于实验网页集中链出网页比链入网页更有可能与主题相关，因此取λ的值为 0.85。

（2）基于改进Shark-Search的多媒体主题搜索算法

经反复测试，在网页内容相似度计算时，公式 3.12 选择参数$\beta=0.8$。链接相似度计算时，公式 3.15、3.17 选择参数$\sigma=0.6$，$\lambda=0.5$时，实验效果最好。

（3）基于URL链接规则的多媒体主题搜索算法的测试

在公式3.12中，经实验测试，在网页内容相似度计算时，采用和基于改进Shark-Search算法内容相似度的计算公式，选择参数$\beta=0.8$；在计算链接相似度时，对于改进的PageRank算法中，选择参数$d=0.85$。

为了比较三种算法的搜索效率，在计算网页内容相似度和链接相似度的归一化时，网页内容相似度和链接相似度占同样的比重，公式4.8中参数∂统一设定为0.4。

5.2.6 仿真试验环境

为了验证以上三种不同的搜索算法的效率，采用相同的仿真试验环境，使用同一多媒体主题搜索器，算法都采用 Java 语言进行实现。为提高搜索效率，采用多线程技术同时搜索不同的站点，系统开启了 50 个线程。实验环境为：WindowsXP 操作系统，PIII CPU，512M 内存。

5.3 实验结果

先用通用搜索算法运行，运行结果如表 5-5 所示。然后再分别用 Topic-PageRank 算法、改进 Shark-Search 算法和基于 URL 链接规则算法运行，运行结果如表 5-6 所示。

表 5-5 通用搜索算法实验结果

种子个数	网页总数	有效网页个数	有效网页占有率	运行时间	平均爬行速度
10	73952	3059	4.464%	30.91 小时	40 个/分钟

表 5-6 三种算法实验结果比较

算法	种子个数	网页总数	有效网页个数	运行时间	平均爬行速度	查准率	查全率
Topic-PageRank 算法	10	26872	984	18.44 小时	24 个/分钟	2.67%	32.17%
改进 Shark-Search 算法	10	21691	1643	24.51 小时	14 个/分钟	7.58%	53.71%
基于 URL 链接规则算法	10	15892	2048	15.27 小时	17 个/分钟	12.89%	66.95%

实验还从算法的查准率和爬行时间的关系出发，对三种算法进行测试，每隔 1 小时统计一次，结果如表 5-7 所示。

表 5-7 三种算法查准率随时间变化实验结果比较

查准率/（%）			时间/小时
Topic-PageRank 算法	改进 Shark-Search 算法	基于 URL 链接规则算法	
2.256%	7.485%	6.527%	1
3.246%	12.267%	16.463%	2
3.568%	19.619%	21.493%	3
3.896%	21.572%	25.691%	4

5.4 结果分析

表 5-6 显示了各种方法的优劣。在算法的查准率方面，最低的为 Topic-PageRank 算法，最高的为基于 URL 链接规则算法；在查全率方面，由于通用搜索算法中，所有的有效网页并不一定是主题相关网页，所以查全率的高低只能反映算法的相对比较情况。最低的为 Topic-PageRank 算法，最高的为基于 URL 链接规则算法。在测试结果中，Topic-PageRank 算法的查准率和查全率都很差，这是因为它优先提取的都是基于某个主题重要度的页面，并没有对这个主题下包含多媒体链接网页进行特殊的加权考虑。改进 Shark-Search 算法，在搜索的查准率和查全率方面都介于另外两种算法中间，主要是因为 Shark-Search 算法，在搜索到“多媒体资源区域”后，能够及时调整爬行方向，较好的预测了主题多媒体资源分布的位置。但是对于干扰网页过多的情况下，此算法就很难找到“多媒体资源区域”，从而使得算法的查准率和查全率就大大降低。基于 URL 链接规则的算法，由于网页呈现多媒体“资源相邻性”特点，在搜索的初始阶段，就以搜索“多媒体资源区域”为目标，在搜索到相关区域后，通过对链接进行主题相关性的判别，优先提取此区域内的主题链接，所以运行结果的查准率和查全率都是最高的。

表 5-7 比较了三种算法的查准率随时间的变化情况，Topic-PageRank 算法由于关注的是搜索网页 PR 值的大小，即搜索网页在整个网络中的重要性，所以其在单位时间内的查准率是最低的。在主题搜索的初始阶段，由于基于 URL 链接规则算法，首先要进行盲搜索以发现“多媒体资源区域”，然后再根据 URL 链接规则进行多媒体资源的搜索，改进的 Shark-Search 算法能够搜集到更多的多媒体主题网页。随着时间的推移，基于

URL 链接规则算法的查准率逐渐提高上来。

综合比较上述三表，从另外的角度反应了各个算法的特性，由于三种主题搜索算法计算的时间复杂度都很大，所以它们的平均爬行速度都远远小于通用搜索算法的平均爬行速度。三种主题搜索算法，在搜索过程中，采用一定的搜索策略，对于判断与主题无关的网页不进行任何分析，三种主题搜索算法的爬行网页总数也远远少于通用搜索算法。在搜索的有效网页个数方面，由于算法的局限性，不可能将全部与主题相关的多媒体网页搜索到，相比较而言，基于 URL 链接规则的多媒体主题搜索算法的查全率是最高的。影响三种主题搜索算法运行时间的主要因素是算法的时间复杂度。三种算法在网页内容相似度计算方面采用同样的算法，不同的是网页链接相似度的计算。算法的时间复杂度是改进的 Shark-Search 算法最复杂，Topic-PageRank 算法次之，基于 URL 链接规则算法最简单。

结束语

网络教育多媒体主题搜索算法以网络多媒体教育资源为研究对象，在已有主题搜索器的基础上，重点研究多媒体主题搜索算法，从而提高网络多媒体资源的主题搜索效率。针对传统的主题搜索算法，不能简单地应用于多媒体主题搜索领域的问题，本文对传统的 PageRank 主题搜索算法、Shark-Search 主题搜索算法进行深入分析，结合多媒体资源在网页中分布的特点，对这两种典型的算法进行相关参数的改进，使之应用于多媒体主题搜索领域。本文还在分析传统主题搜索算法的基础上，根据包含多媒体资源的主题页面在 Web 上的分布特征，提出一种基于 URL 链接规则的多媒体主题搜索算法，算法从种子网站列表中自动学习出代表“多媒体资源区域”的 URL 正则表达式，并用这些正则表达式来指导主题搜索器对网页的抓取。同时详细介绍了 URL 数据结构、URL 距离的度量以及 URL 正则表达式的学习和指导阶段。

为了验证基于 URL 链接规则多媒体主题搜索算法的有效性，本文采用统一的系统体系结构和软、硬件平台，对 Topic-PageRank 算法、改进的 Shark-Search 算法和基于 URL 链接规则的算法输入相同的种子页面集、限定同一搜索深度进行实验测试。文章从查全率和查准率两个角度，对三种算法进行比较。实验结果表明本文的工作是相当有效的，尤其是提出的基于 URL 链接规则的多媒体主题搜索算法，具有相当的创新性和实际应用价值。

由于时间关系和本人的技术水平有限，论文只是对传统经典的 PageRank 算法、Shark-Search 算法，进行了详细的分析和改进。对于目前流行的基于巩固学习的搜索算法、基于概念语义的搜索算法、基于动态价值的搜索算法等，没有进行详细的分析，并将它们应用于多媒体主题搜索领域。在今后的工作中还要围绕以下几个方面做深入的研究和分析：

1．对主题页面在网络中的分布规律，进行更深入的研究，尤其是包含多媒体的网页在 Web 上的分布特征，为提高算法的查准率和查全率提供更多的选择参数。

2．包含多媒体资源的网页往往分布在网站的最底层，对链接价值的评价深入到更微观层面，找出蕴涵其中更具有链接结构意义链接分析方法。

3．目前算法只学习种子网站列表中的 URL 正则表达式，如何获取种子页面上有价值的外部链接，并对其进行 URL 正则表达式的学习，有待于进一步研究。

4．算法中参数的设置直接决定了算法的优劣，参数的选择需要大量的实验数据支持，目前这方面做得还不够。

5．现有主题搜索系统的性能还有待于进一步优化，包括：主题搜索器下载网页的速率、下载资源存储策略以及对搜索资源种子的选取方案等。

6．本文实验和分析的网页全部是静态网页，目前动态网页的数量已经超过静态网页的数量，如何对动态网页进行分析仍然是一个重点和难点，有待于进一步深入研究。

参考文献

[1] 邢志宇．多媒体搜索引擎指南［EB/OL］. http://www.sowang.com/zhuanjia/xzhy/20031118-1.htm，2003．11.

[2] 中国互联网信息中心．中国互联网发展状况统计调查[EB/OL]. http://www.cnnic.cn/index/OE/00/11.

[3] Association for Educational Communications and Technology[EB/OL]. http://www.aect.org/default.asp.

[4] 尹俊华，庄榕霞，戴正南．教育技术学导论[M]．北京：高等教育出版社，2002.

[5] 何克抗，郑永柏，谢幼如著. 教学系统设计[M]．北京：北京师范大学出版社，2002，10.

[6] 彭绍东. 解读教育技术领域的新界定[J]. 电化教育研究，2004，（10）：8-17.

[7] D. Floresu, A.Levy, A. Mendelzon. Database Techniques for the Word-Wide Web: A Survey. ACM SIGMOD Record,27(3),1998.

[8] P.Buneman. SemistruedData. In Proceedings of the ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems. Tucson, Arizona, PP. 117-121, 1997.

[9] HTML 4.01 Specification. http://www.w3.org/TR/html4011.

[10] 章毓晋[J]．基于内容的视觉信息检索[M]．北京：科学技术出版社，2003.

[11] M.M.Gorkani, R. W. Picad. Texture Orentation ForSorting Photos “at a glance” . In 12th Conference on Pattern Recognition. Pp. 459-464, October 1994.

[12] 音乐数据库检索系统[EB/OL]. http://www.1ib.sjtu.edu.cn/music.htm.

[13] S. -F . Chang, W. Chen, H.Meng, H.Sundaram, and D. Zhong, “VideoQ: an automated content based video search system using visual cues” , in Proc. ACM Multimedia, November 1997.

[14] Modha D. .Spangler W. Clustering Hypertext with Application to Web Searching . In: Proceedings of the 11th ACM Conference on Hypertext and Hypermedia.San Antonio: ACM Press,2000:143-152.

[15] Aggarwal C, AI_Garawi F, Yu S P. Intelligent carwling on the World Wide Web with

arbitrary Predicates[C]. In:Proc of the 10th International World Wide Web Conference, 2001.

[16] Menczer F. Complementing search engines with online Web mining agents[J]. Decision Support Systems, 2003, 35(2): 195-212.

[17] Bra D P, Houben G, Kornatzky et al. Information Retrieval in Distributed Hypertexts[C]//Proceeding ofthe 4th RIAO Conference 1994.

[18] Cho J, Garcia-MolinaH, Page L.Efficient Crawling Through URL Ordering[J]. Computer Networks, 1998, 30(1-7):161-172.

[19] Page L, Brin S, Motwani R, et al.The pagerank citation ranking:Bringing order to the web[R]. Standfrd Digital Libraries SIDL-W P-1999-0120, 1999.

[20] Bharat K, Henzngер. Improved Algorithms for Topic Distillation in A Hyperlinked Environment[C]. In:Proc. of SIGIR Conference on Research and Development in Information Retrieval, 1998.

[21] Rennie J,McCallum A.Using reinforcement learning to spider the Web efficiently[C]. In: Proc of the International Conference on Machine Learning (CML 99),1999.

[22] Diligenti M,Coetzee F M,Lawrence Set al,Focused carwling using context graphs[C].In:Proc of the International Conference on Very Large Database (VLDB '00), 2000: 527-534.

[23] 李晓明 闫宏飞 王继民著. 搜索引擎—原理、技术与系统[M]. 科学出版社, 2005.

[24] 席一凡, 刘培奇编著. 动态网页设计教程[M]. 西安电子科技大学出版社, 2005. 02.

[25] S.Chakrabarti,B.Dom,P.Raghavan,etc.Automatic Resource Compilation by Analyzing Hyperlinkage Structure and Associated Text Computer Networks and ISDN Systems.1998,(30):65-74.

[26] 宋宇, 孟祥增. 基于改进 Fish-search 算法的多媒体检索[J]. 计算机工程, 2008, 34(11):189-193.

[27] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, 1998: 668–67.

[28] C. Aggarwal, F. Al-Garawi , P. Yu. Intelligent Crawling on the World Wide Web with Arbitrary Predicates. In Proceedings of the 10th International WWW Conference, Hong

Kong, May 2001.

[29] 许文, 都运程, 李渝勤, 施水才. 《一种通用HTML网页主题信息提取方法》[J]. 现代图书情报技术 2007, 1:40-43.

[30] 宋宇. 面向教育的多媒体主题搜索器设计与实现[D] . 济南:山东师范大学硕士学位论文, 2008.

[31] L Zhang,L Chen, M Li, et al. Efficient propagation for face annotation in family albums[A]. Proceedings of the Eleventh ACM International Conference on Multimedia[C]. New York, USA:ACM Press, 2004. 716-723.

[32] 王灏, 黄厚宽, 田盛丰.文本分类实现技术[J]. 广西师范大学学报（自然科学版), 2003, (01).

[33] 王伟强, 高文, 段立娟.Internet上的文本数据挖掘[J] . 计算机科学, 2000, (04.

[34] T J. A Probabilistic Analysic of the Rocchio Algorithm with TFIDF for Text Categorization[J] . In Proceedings of the 14th International Conference on Machine Learning, 1997:143-151.

[35] 刘玮玮. 搜索引擎中主题爬虫的研究与实现[硕士]. 南京理工大学, 2006.

[36] Arind Arasu, Jasmine Novak, Andrew Tomkins, John Tomlin. PageRank Computation and the Structure of the WEB:Experiments and Altorithms.In Proceedings of 11th International World Wide Web Conference, 2002.

[37] 钱功伟，倪林. 基于网页链接和内容分析的改进 PageRank 算法[J]. 计算机工程与应用, 2007, 43 (21) 160-164.

[38] Menczer F, Pant G,Srinivasan P. Topic Web crawlers:Evaluation adaptive alogorithms. ACM Trans. on Internet Technologies, 2003, 26(1):89-113.

[39] Bharat K, Henznger. Improved algorithms for topic distillation in a hyperlinked environment. In:Proc of SIGIR Conference on Research and Development in Information Retrieval. ACM press, 1998.

[40] Rennie J, McCallum A. Using Reinforcement Learning to Spider the Web Efficiently[C]. In: Proc. of the International Conference on Machine Learning (ICML 99), Bled, Slovenia, 1999, 335-343.

[41] Diligenti M, Coetzee F M, Lawrence S, et al. Focused Crawling Using Context graphs[C]. In:Proc. of the International Conference on Very Large Database (VLDB00),

Cairo, Egypt, 2000, 527-534.

[42] Sutton R S, Barto A G. Reinforcement learning: an introduction. MA: MIT Press, 1998.

[43] Page L, Brin S. The PageRank Citation Ranking: Bringing Order to the Web[EB/OL] . http://www.db.stanford.edu/~backub/PageRank. 1998-2001.

[44] Haveliwala T H. Topic-sensitive PageRank[C] . Proceedings of the Eleventh International World Wide Web Conference, Hoho Lulu Hawaii, 2002.

[45] D M Bikel, R L Schwartz, R M Weischedel. An algorithm that learns what's in a name[J] . Machine Learning, 1999, 34(1-3):211-231.

[46] 苏祺, 项锟, 孙斌. 基于链接聚类的 Shark-Search 算法[J] . 山东大学学报:理学版, 2006, 41(3).

[47] 陈骏, 陈竹敏. 基于网页分块的 Shark-Search 算法[J] . 山东大学学报:理学版, 2007, 42(9):62-66.

[48] 邱正国. 主题蜘蛛的研究与实现[D]. 南京:南京师范大学硕士学位论文, 2007.

[49] 叶勤勇. 基于 URL 规则的聚焦爬虫及其应用[D]. 杭州:浙江大学硕士学位论文, 2007.

[50] 李超峰, 卢炎生.基于 URL 结构和访问时间的 Web 页面访问相似度度量[J] . 计算机科学, 34（4）:207-209.

致　谢

值此论文完成之际，谨向在论文写作过程中给予指导和帮助的各位老师、同学和我的亲人表示感谢！

首先要向我的导师孟祥增教授致以崇高的敬意和深切的谢意。衷心感谢孟老师多年来的支持、鼓励、关怀和信任，本文是在孟老师的精心指导下完成的，整篇论文都凝聚着孟老师的心血和汗水，孟老师始终以他宽阔的视野、渊博的学术知识和严谨的治学态度对我言传身教，并将使我受益终生。孟老师为我开启了了解和研究多媒体资源的大门，学高为师、身正为范是学生对孟老师最好的敬意。同时感谢和蔼可亲的刘瑞敏师母，使我们身在异处却能感受到家的温暖。

特别感谢王玲老师和谭金波老师，感谢你们的关心、信任和帮助，让我不仅在学术上获益良多，开阔了视野，同时也进一步提高了自己的能力和团队合作的优良品质。同时，感谢马池珠老师、陆宏老师、宫淑红老师、白成杰老师对我开题报告的指导和学习上的帮助。

感谢我的师兄师姐，他们对我尽快进入科研状态起到了极大地促进作用，尤其感谢宋宇师兄，在我研究之初，给了我很多指导和帮助。感谢我的同门原佳丽、井艾斌，为我的论文提供了很多参考数据。感谢我的师弟师妹们，他们对我的启发让我开拓了研究的思路。

感谢 2006 级全体研究生在这三年来给我的帮助，我会永远记得这个和睦的大家庭。

最后深深的感谢给予我生命并培养我成人的父母，他们无怨无悔，无私的奉献着他们的一切。感谢他们多年来对我精神和物质上的关心和照顾，使我能够顺利完成学业。

需要感谢的人还很多，在此难以穷举。我将心怀感激，永远铭记。

攻读硕士学位期间发表论文及参加的项目

一、攻读学位期间发表的学术论文

1．网络多媒体教学资源主题搜索研究，电化教育研究（中文核心期刊、CSSCI 检索源期刊），2009（6），第一作者

2．一种基于网页内容和链接分析的主题搜索算法，情报杂志（中文核心期刊、CSSCI 检索源期刊），2008（6），第一作者

3．一种改进 Shark-Search 的多媒体主题搜索算法，计算机工程与应用（中文核心期刊、中国科技核心期刊），2010（1），第一作者

4．网络多媒体主题搜索策略比较研究，图书情报工作（中文核心期刊、CSSCI 检索源期刊），第一作者（录用）

5．基于 Web 多媒体基础教育资源检索系统的设计，中国教育信息化（CSSCI 扩展版来源期刊），2008（1），第一作者

6．网络多媒体主题搜索研究及其教学应用，现代远距离教育（CSSCI 扩展版来源期刊），2009 年（1），第一作者

7．网络多媒体主题搜索策略研究，中国科技资源导刊，2009（2），第一作者

8．多媒体教育资源主题搜索研究，第 7 届教育技术国际论坛论文集，2008（9），论文获得三等奖，第一作者

9．Java 调用 VC_的动态链接库，电脑知识与技术（学术交流版），2007（6），第二作者

10．快速双向中文分词算法，山东师范大学学报（自然科学版）（中国科技核心期刊），2009（1），第二作者