基于 Lucene.Net 校园网搜索引擎的设计与实现

蔡建超,郭一平,王 亮
(华中科技大学 控制科学与工程系,湖北 武汉 430074)

摘 要:在庞大的网络信息资源利用中,搜索引擎成为人们发现资源的有力工具。但是如果用常用的搜索引擎去搜索专门的门户网站,或者特定范围的网页,则显得力不从心。比如搜索一个学校内的网页,这些搜索引擎就很难进行全面高效的搜索。基于此,利用适应.Net环境的 Lucene.Net 作为基础,设计了自己的校园网搜索引擎,在特定范围的搜索表现出了自己独特的优势。

关键词:搜索引擎;爬虫;索引
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2006)11-0073-03

Design and Implementation of School Search Engine Based on Lucene.Net

CAI Jian-chao, GUO Yi-ping, WANG Liang
(Dept. of Control Sci. and Eng., Huazhong Univ. of Sci. and Techn., Wuhan 430074, China)

Abstract: In the vast sea of the information resources on Internet, search engine had become a powerful tool to look for what the people need. But if you want to search something in a portal site or in a limited domain, existing search engines can't do well with it. For example, these search engines can't cover all pages in the requested domain, so if you want to search pages in a school network, it will be very difficult. We have designed a search engine based on Lucene.Net, which can be used in .Net environment. It can show its particular advantage in domain search.

Key words: search engine; Web spider; index

0 引 言

搜索引擎在当今的网络资源应用中扮演着重要的角色,从 Baidu 和 Google 的业绩强势增长中也可以看到这点。有了搜索引擎,网络资源得以集中地管理和分类,从而使人们不用直接去网络上盲无目的地寻找自己需要的东西。搜索方法和结果较以前也便利、快捷并且更加有效。但是随着技术的发展,人们发现现在的搜索引擎并不能满足自己的特定要求,大众的搜索引擎也很难实现一个域范围内全面快速的搜索。比如说,在一个大学之内搜索,Baidu 和 Google 就不可能提供校园网所有网页这一特定范围的搜索功能,并做到快速地更新。

文中提出了一种基于 Lucene.Net 的校园网搜索引擎设计方法[1],通过该方法,可以较轻松地做出基于.Net 和 Windows 平台下适合自己的搜索引擎。具体的原理是先通过网络爬虫下载所有的网络资源,通过 Lucene.Net 索引,入库;然后定制网页,根据用户的查询要求排序输出。当然,还涉及到了更新、排序等一系列问题。通过此搜索引擎的设计,一些门户网站和一个域内资源的搜索可以做到方便、快捷、全面。

收稿日期:2006-02-19
作者简介:蔡建超(1980-),男,河南新密人,硕士研究生,研究方向为下一代网络技术和搜索引擎技术;郭一平,副教授,研究方向为数字图书馆体系结构、下一代网络技术。

1 搜索引擎及 Lucene.Net 概述

当今的搜索引擎大多采用集中式的搜索方式。所谓集中式就是通过很多服务器把网络资源全部下载到本地,然后做一些处理,为搜索做准备。搜索引擎结构大致分为搜索器、索引器和检索器等几部分组成[2],搜索器就是人们所说的网络蜘蛛(Web Spider)或者叫网络机器人。通过这种 Spider 程序,可以从一个网页出发,通过提取其中的 URL,在遵从 Robot Exclusion 协议的前提下,不断地提取得到的 URL,并且下载本 URL 的资源;而索引器的主要工作则是利用下载的网络资源,提取索引项,用于生成文档库的索引表;检索器主要是通过理解用户的查询需求,在文档库中检索出文档并且进行快速匹配,然后进行相关性排序,通过链接网页提供给用户检索结果。至此,完成搜索。

可以看出,一个优秀的搜索引擎只要把这几个部分做好,就可以满足用户的需求。为了在校园网内实现资源的有效检索,在 Lucene.Net 的基础上设计了这个搜索引擎。其实 Lucene.Net 并不是一个独立的开源项目,它是基于 APACHE 基金会 jakarta 的一个子项目 Lucene 的二次开发,目的是能够在.Net 环境下应用,现在通过对各个部分的改写,C# 下的 Lucene.Net 已经十分成熟。它秉承了

Java 下的 Lucene 的高效性,为许多的.Net 开发人员提供了进一步研究搜索引擎的平台。

现在介绍 Lucene 的文章[3]已经很多,简要的组成结构如表 1 所示。

表 1 Lucene.Net 结构功能表

Lucene.Net 结构功能表	
程序集名	功能
Lucene.Net.Analysis	语言分析器,主要用于切词,扩展后支持中文
Lucene.Net.Documents	索引存储时的文档结构管理,类似于关系型数据库的表结构
Lucene.Net.Index	索引管理,包括索引库的建立、删除等
Lucene.Net.QueryParsers	查询分析器,实现查询关键词间的运算,如与、或、非等
Lucene.Net.Search	检索管理,根据查询条件,检索得到结果
Lucene.Net.Store	数据存储管理,主要包括一些底层的 I/O 操作
Lucene.Net.Util	一些公用类

由此可以看出 Lucene.Net 提供了十分全面的索引、查询等一系列模块,基于此,通过扩展程序,做二次开发应该很容易做出适合自己使用的搜索引擎。

2 搜索引擎的实现

2.1 网络爬虫的编制

网络爬虫必须可以自动地搜索网页,提取网页中的链接,然后重复此工作,从而尽可能逐个网页地爬满整个 Internet。另外由于网络资源虽然在同个域范围内,但是为了保证效率,爬虫程序必须实现多线程才可以实现更新的快速性。搜寻网络资源时在遵循 ROBOT EXCLUTION 标准的前提,校园网内的网页实现全面下载还是比较容易的,因此,对网页的下载优先级不需要定义。另外,对网页的存储路径相对于网页的网址,需要作一个转换,从而可以更方便地在快照(即本地)中找到此网页。另外对于 IP 地址要做一个判断,以实现在特定范围内搜索。

主要程序有 start 函数实现,函数原型为: Start(Uri baseURI,int threads),前一个参数表示所要求下载的超级链接,后一个函数为线程数量。具体的编制参看有关 spider 的文献。

而对于路径的转换,只需要把字符转换一下即可实现,读者可以自行实现。具体的校园网搜索引擎中网络爬虫的实际工作流程如图 1 所示。

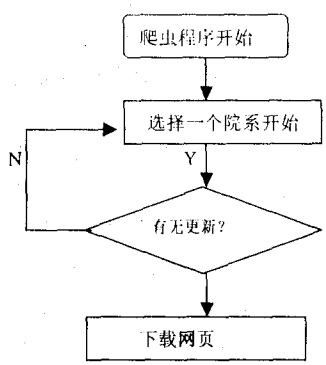

图 1 网络爬虫工作流程图

2.2 切词的研究

切词对于搜索结果来说是十分重要的环节,对于英文而言,只用关注空格即可实现切词,而对于中文,由于中文词汇非常丰富,词语的组成十分不固定,所以中文切词就变得比较困难。现在比较普遍的是中文单字切词、双字切词、字典切词,对于单字切词由于其一开始的不准确性就被很多人放弃,最准确的切词方法应该是字典切词,匹配方法大致有前向匹配、后向匹配和基于统计的匹配。对于各种方法,在此不再详细介绍,请参看相关文献。

简单举例来说,对于"中华人民"这句话,单字切词就是把这四个字一个作为一个词索引,搜索的时候逐字匹配,而双字索引是把这句话依次按两个字加以区分,分为"中华","华人","人民"这样切词。字典切词则是相对于词库来寻找哪个应该作为词,而哪个不是,由此也可以知道,字典切词是比较准确的,但是由于汉语词语的不断变化和扩充性,所以现在还要结合双字切词才可以更好地提高准确度。

由于字典切词和双字切词的配合使用,搜索引擎的匹配精度在很大程度上是可以得到保证的,期间参考了很多的切词算法[4,5],通过程序集 Lucene.Net.Analysis.CJK 来实现。

2.3 解析网页及索引入库

解析网页就是要把网页中的元素标记(Token)及其标记之后的内容提取出来,以利于入库,相对于每一个 Token 建一个相应的字段,而此 Token 的内容就是此字段的内容。实现方法:建立一个 Myparser 类,然后要实现读入文件流,接着解析成字符串格式输出,以备下一步处理,然后逐个提取 Token 及其内容。逐个提取 Token 的目的是为了以后搜索的时候可以给不同的 Token 加上不同的权值,从而可以很好地实现排序。提取 Token 之后开始入库:

```
Myparser parser=new Myparser(f);
Document doc=new Document();
doc.Add(Field.Text("title", parser.GetTitle()));
…用此方法逐个加入字段
return doc;
```

至此,主要的工作原理基本完毕,已经可以实现对网页的双字索引入库。但是入库的时候要注意的是,提供了更新机制,现在采用的是简单的更新机制,即单双天机制,双天查找到更新的时候索引到目录一,单天的时候索引到目录二,既可以天天更新,又可以不影响使用,简单表示为图 2。

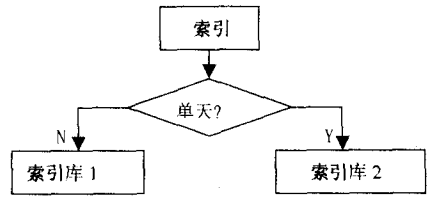

图 2 目录更新示意图

2.4 编制网页接口,实现检索输出

以上程序实现了索引,但是也还只是为搜索做好了准备工作,对于用户来讲是通过网页访问跟程序打交道的,所以检验结果一定要方便地通过网页显示出来,同时要兼顾更新、加亮、排序输出等功能。工作流程如图 3 所示。

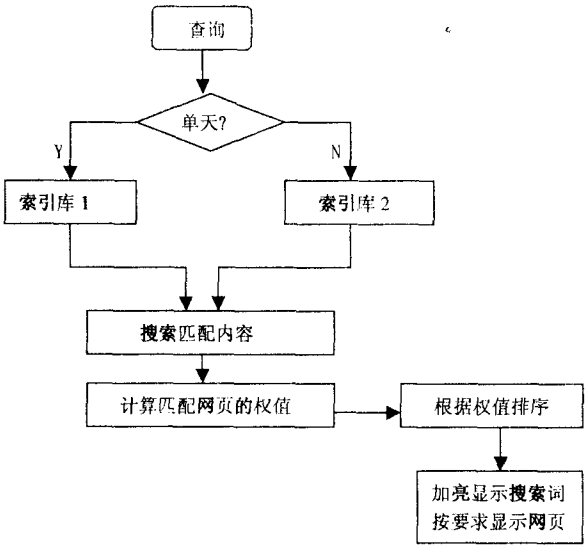

图 3　网页搜索排序程序流程图

主要的程序代码:

```
searcher = new IndexSearcher(IndexReader.Open(indexName));
queryString = Request.Params["word"];
page_num = System.Convert.ToInt16(Request.Params["num"].ToString());
Search(queryString);
```

搜索排序程序的主要代码:

```
Analyzer analyzer = new CJKAnalyzer(); //解析器
query = QueryParser.Parse(queryString, "contents", analyzer); //实现搜索
if (searcher != null)
{
thispage = maxpage;
hits = searcher.Search(query);
total_num = hits.Length();//结果的个数
}
Highlighter highlighter = new Highlighter(new QueryScorer(query));//加亮显示程序
highlighter.SetTextFragmenter(new SimpleFragmenter(100));
int maxNumFragmentsRequired = 3;//显示搜索到的字符串的最大个数,同时排序
for (int i=0;i<hits.Length(); i++)
{
System.String text = hits.Doc(i).Get("contents");
TokenStream tokenStream = analyzer.TokenStream("contents", new System.IO.StringReader(text));
highlightedText = highlighter.GetBestFragments(tokenStream, text, maxNumFragmentsRequired,"…");
}
```

排序的方法是通过对各个 Token 赋予不同的权值实现的,用公式 $W = \sum a_i * b_i$ 表示,W 表示总权值,a_i 表示各个 Token 权值,b_i 表示出现次数,W 越大,则排在最前边,最后则是通过网页页面显示出来。至此,编制一个校园网搜索引擎的全部工作已经完成。

图 4　搜索引擎搜索效果图

(下转第 80 页)

如果MAS系统只部署在一台机器上，上述方法是很有效的，如果需要部署在多台机器上，可以再设计一类本地协调主体，运行在每台机器上，各台机器上的主体都向本地协调主体注册，主服务器上运行全局唯一的系统管理主体，当主体之间需要合作时，先查找本地协调主体，看是否有中意的合作伙伴，若有则在本地通信，若没有则本地协调主体通过Socket通信向系统管理主体查询。由于本地协调主体承担了本地主体的管理，减轻了全局管理的压力，而且本地主体的通信不必查询系统管理主体，加快了本地主体通信的速度[9]。

(2)任务协调。

主体内部的任务触发机制是源于一个目标(Goal)的生成。一个Goal生成之后，主体自动检查自身是否能够完成该任务以及是否占有完成该任务所需的全部资源，若否，把所缺少的资源，以子目标(SubGoal)的形式向相关主体发出询问，寻求合作。模型中任意两个主体的A和B之间的交互行为执行算法为：①A向B发送“请求服务”的消息；②B收到之后向A发送accept消息，并在自己的任务集中寻找与之相吻合的操作；③如果没有相吻合的计划，则向A发送一个reject消息，否则执行其操作。完成之后，用result操作将结果信息发送给A。因此，在确定Agent内部的任务处理之后，要描述整合以本体为信息流的任务之间的相互关系，明确子目标，从而反映出代理之间的任务协调关系。

图9是监视主体接到一个MT的连接请求时所作的任务协调过程。

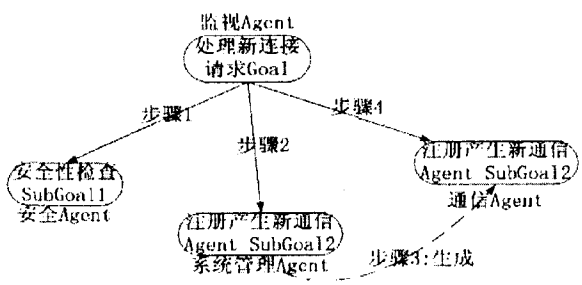

图9 新通信连接的任务协调过程

3 结束语

为了印证基于MAS的集成通信服务器的可行性与有效性，文中使用Zeus开发平台，构建了一个系统原型，把以上系统中的几类主体，部署在多台服务器与工作站上，用来模拟通信服务器的运行，经过五矿货运广东公司的反复测试，在GPS终端数量达到100台的运行条件下，和原有软件对比，丢包率由原来的15～20%减为3～5%，处理一条消息的平均延时由原来的386ms缩短为平均117ms，特别是对报警类紧急信息，系统响应时间大幅度缩短，取得了良好的运行效果。

参考文献：

[1] 原仓周，柳重堪，张其善.大规模车辆监控通信服务器的设计与实现[J]. 北京航空航天大学学报，2004，30：232－235.

[2] 陈 斌，李德华，姚 迅.一种基于GPRS技术的可扩展车辆监控系统的设计与实现[J]. 计算机应用研究，2005(6)：175－178.

[3] 原仓周，柳重堪，张其善.大规模车辆监控SMSGPRS通信服务器参数分析[J]. 小型微型计算机系统，2005，26(5)：775－778.

[4] 刘小明，王飞跃.基于Agent的单路口交通流控制的研究[J]. 系统仿真学报，2004，16：853－855.

[5] Wooldridge M. 多Agent系统引论[M]. 石纯一，等译. 北京：电子工业出版社，2003.

[6] Manvi S S，Venkataram P. Application of agent technology in communications：a review[J]. Computer Communications，2004，27：1493－1508.

[7] 高国军，段永强，张申生.基于CORBA和多代理技术的可重构企业信息系统[J]. 计算机集成制造系统－CIMS，2000(3)：26－30.

[8] 郭 中，王惠芳，黄永忠，等.软件Agent的通信模型[J]. 计算机工程与设计，2002，23(11)：9－11.

[9] Gaspari M. Concurrency and Knowledge－level communication in agent language[J]. Artificial Intelligence，1998，105：1－45.

(上接第75页)

3 结果及结论

通过检验，编制的搜索引擎基本可以实现使用要求。在使用的过程中，更新问题及搜索效率都可以很好地保证，当然这也体现了Lucene.Net的高效性，搜索结果如图4所示，为了方便观看总体效果，图4是采用一个搜索词并采样最后一页得出的效果图。

搜索引擎的设计已经全部结束了，不过也有不足之处需要进一步的改进和加强，比如说搜索引擎的切词问题，如何做到最优，还有排序问题，如何才能做到最合理，都需要进一步的研究。

参考文献：

[1] 赵 汀，孟祥武.基于Lucene API的中文全文数据库的设计与实现[J]. 计算机工程与应用，2003(20)：179－183.

[2] 李晓明，刘建国. 搜索引擎技术及趋势[EB/OL]. 2003－04. http://www.se－express.com/se/se07.htm.

[3] 张校乾，金玉玲，侯玉波.一种基于Lucene检索系统的一种全文数据库的设计与实现[J]. 现代图书情报技术，2005(2)：40－44.

[4] 车 东. 在应用中加入全文检索功能——基于Java的全文索引引擎Lucene简介[EB/OL]. 2002－08. http://www.chedong.com/tech/lucene.html.

[5] 郭 辉，苏中义，王 文，等.一种改进的MM分词算法[J]. 微型电脑应用，2002，18(1)：13－15.