文章编号:1008-0570(2007)07-3-0119-03

基于 Web 的网络爬虫的设计与实现

Design and Implementation of Spider on Web-based Full-text Search Engine

(首都师范大学)徐远超 刘江华 刘丽珍 关 永
XU YUANCHAO LIU JIANGHUA LIU LIZHEN GUAN YONG

摘要:无论是站内信息检索还是特定的 Web 信息搜集,都离不开全文搜索引擎系统的核心模块——网络爬虫,本文详细介绍了一种设计及实现方案,包括页面搜集器和页面索引器的基本工作流程、数据存储结构、核心算法及主要的技术难点。该系统经实际运行,效果良好,最后给出了有待进一步改进的地方。

关键词:搜索引擎;网络爬虫;信息检索;页面索引
中图分类号:TP391 文献标识码:A

Abstract: Whether inside website information retrieval or special web information collecting, spider is the essential and most important module. One way of design and implementation of spider on web-based full-text search engine was introduced in detail, including the basic work principle、database structure、key arithmetic and technical difficulties about webpage collecting and webpage indexing. This blue print has been proved to be feasible. In the end it gives some aspects to be improved on.

Key words: search engine, spider, information retrieval, webpage indexing

技术创新

1 引言

搜索引擎 (Search Engine) 是随着 Web 信息的迅速增加,从 1995 年开始逐渐发展起来的技术。要在浩瀚的信息海洋里寻找信息,往往是“大海捞针”,无功而返,为了解决这个“迷航”,出现了搜索引擎技术。

搜索引擎以一定的策略在互联网中搜集、发现信息,对信息进行理解、提取、组织和处理,并为用户提供检索服务,从而起到信息导航的目的。搜索引擎提供的导航服务已经成为互联网上非常重要的网络服务,搜索引擎站点也被美誉为“网络门户”。尽管基于海量多媒体信息的语音、图形、视频搜索引擎技术成为搜索引擎领域的研究热点,但是基于 Web 的全文本搜索引擎仍然是使用最为广泛的,如信息量较大的专业门户网站的站内信息检索、基于互联网的特定信息搜集等等。

一般情况下,基于 Web 的全文搜索引擎均由页面搜集器、页面索引器、页面检索器等三个主要部分组成,如图 1 所示。

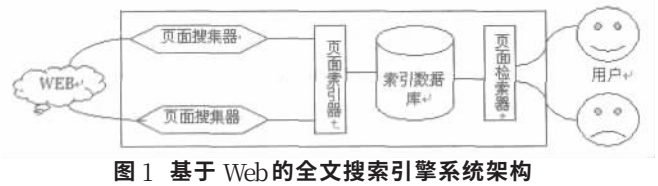

图 1 基于 Web 的全文搜索引擎系统架构

其中页面搜集器和页面索引器是搜索引擎最为核心的模块,主要由称为 spider 的爬虫程序来完成,以下给出一种详细的设计及实现方案。

2 页面搜集器的设计

2.1 基本工作流程

Spider 首先从 UnCheckedURL 数据库(表 1)中取出一条网址记录,检测当前网址链接的有效性。如果有效,则将当前网址对应的 HTML 页面保存到本地磁盘,然后将该 HTML 页面上的所有超链摘取出来,将此超链集合以追加的形式加入到 UnCheckedURL 数据库的尾部,并以广度优先搜索算法遍历 UnCheckedURL 数据库。

当然,在高性能计算机上,可以让 spider 程序并行工作。假设一台计算机是一个节点,每个节点上运行着 10 个 spider 程序,每个 spider 程序同时进行着页面的下载和超链的摘取工作,每个 spider 程序对应着自己的 UnCheckedURL 数据库,将下载的页面保存到本地磁盘的同一目录下,并将保存到目录下的页面文件以相同的命名规则命名,以供页面分析器使用。

2.2 UnCheckedURL 数据库

UnCheckedURL 数据库的作用在于存储从 HTML 页面上摘取的超链集合,搜集器就是通过该数据库取出网址,然后根据该网址去下载所指定的下一个 HTML 页面,并将下载下来的页面保存到本地磁盘,这些保存到本地磁盘的页面就成为后来的快照文件。

表 1 UnCheckedURL 数据库

字段名	描述	类型	PK	NULL
URL_id	记录标识	数字	N	否
URL	对应 HTML 页面的 URL	文本	Y	否
UpdateCycle	更新周期	日期时间	N	否
checked	当前记录的 URL 是否被检测过的标识	布尔型	N	否

在该数据库中,URL 字段设为主键,这样可以防止重复链接的出现。设置 checked 字段的目的主要是标识爬虫程序的起始点,在数据库中将检测过的 URL,无论是否有效,都将 checked 字段置为 0,将刚从 HTML 页面摘取出来的 URL 所对应的 checked 字段置为 1,这样,在 Spider 程序重启的时候,都能在中断的那条记录继续往下走,保证了 Spider 搜集器的运行效率。

徐远超:讲师 硕士
基金项目:北京市自然科学基金资助项目(4062009);北京市教育委员会科技发展计划面上项目(KM200610028014)

2.3 核心算法与难点

在搜集器的设计中，难点之一是页面超链的摘取，包括 HTTP 协议的超链、其他协议的超链以及相对路径的超链。作者所用的开发语言是 C#，摘取超链的核心算法如下：

[1]创建一个 C# 自带的 REGEX 类的对象，该类是用来进行正则表达式的匹配文本类。

[2] 创建一个 C# 里读取文件的 StreamReader 类的对象，将当前的 HTML 页面逐行读入，如果当前一行句子匹配了<a href = (.*)>，那么就将小括号里的内容取出，进入 3，否则返回到 2。

[3]如果小括号里的内容与 http 协议头匹配，那么进入 4，否则返回到 2。

[4]如果小括号里的内容不含有 ".gif"，".jpg"，".swf"，那么进入 5，否则返回到 2。

[5]认为当前内容是一条合法的 URL，以追加的形式写入未检测网址数据库当中，返回到 2。

为了防止超链搜集重复而进入死循环，一方面定义 URL 为主键，另一方面设置更新周期，即只要更新周期还没到的已搜集页面均不再重新搜集，提高了页面搜集器的运行效率。搜索引擎的更新周期对搜索引擎搜索的查全率查准率有很大影响，周期太短增加系统运行负担，周期太长死超链新网页不能及时发现，因此如何定义更新周期也是一个值得探讨的问题。

在检测链接有效性的问题上，根据 HttpWebResponse 类提供的 StatusCode 属性来检测一个 URL 是否有效，如果 StatusCode 的值是 200，则证明当前链接是有效的，否则是无效的。在请求页面资源的时候，设置了等待响应时间为 5000 毫秒，当对方 5 秒后不响应，则自动放弃当前资源的请求，从而保证了搜集器能够快速稳定地运行。

关于搜集子程序的并行问题，采取的办法是每个 Spider 程序都对应着自己的 UnCheckedURL 数据库，都从自己的 UnCheckedURL 数据库中取出链接，以广度优先算法进行遍历，每个 Spider 程序都将下载下来的页面存放到同一目录下，从而提升 Spider 搜集页面的效率。

目前，Web 信息很大一部分通过程序或脚本访问数据库动态生成，由于信息的不确定性，这些信息是很难搜集的。因此，该系统只搜集静态 Web 页面。

3 页面索引器的设计

3.1 基本工作流程

将保存在本地磁盘的 HTML 页面取出，首先进行标题的摘取，将信息写入到索引数据库中的 Title 字段，接着进行内容摘要的提取，将信息写入到索引数据库的 Body 字段，然后提取 10 个关键字，将信息写入到索引数据库的 key1 到 key10 十个字段中，最后是网页的优先级评判，以便检索时进行页面排序。

3.2 PageIndex 数据库

首先设计 PageIndex 数据库，PageIndex 数据库是后台索引器与前台检索器共用的，其作用是将已经保存到本地的 HTML 页面索引到索引数据库里，以供检索器查询结果并且返回结果给客户端。

3.3 核心算法与难点

首先是标题的摘取。创建正则表达式对象，匹配页面内容，如果匹配了含有<title>(.*)</title>标签，就将该标签对内的小括号里的内容取出来，写入索引数据库里的 Title 字段。

接着是内容摘要的提取。所谓内容摘要，就是将当前 HTML 页面上的所有的标签全都去掉之后剩下的正文部分。只需用到两个正则表达式语句就可以解决此问题，要想写一个效率很高的正则表达式是相当困难的，经反复推敲，最终写出了以下的两个正则表达式<script.*?>[\s\S]+?<\/script>和<style.*?>[\s\S]+?<\/style>|\s|)。

表 2 PageIndex 数据库

字段名	描述	类型	PK	NULL
Index_id	记录标识	数字	Y	否
URL	对应 HTML 页面的 URL	文本		否
content	对应的页面被保存的路径，供检索器返回快照文件	文本		否
Title	页面的标题	文本		否
Key1-Key10	描述页面内容的 10 个关键字	文本		否
Body	页面摘要	文本		否
Priority	优先级，前端根据此字段的值按照大小排列顺序返回结果	文本		否

<script.*?>[\s\S]+?<\/script>是匹配所有的<script>…</script>脚本语言，即凡是碰到有脚本语言的地方，都被去除掉，取而代之的是空格，用正则表达式类的 Replace 方法，将匹配到的部分内容用指定的字符来代替。<style.*?>[\s\S]+?<\/style>|\s|)部分是用来剔除一些样式声明的，在 html 语言中，有很大地一部分是页面样式声明，这一部分也应该剔除掉。接下来，就将标签的左右尖括号剔除掉，这样，就剩下了正文部分，也就是摘要了。这两个正则表达式经实践证明，可以适用于所有的 html 文本，而且效率也是相当的高。正文摘要写入索引数据库的 Body 字段中。

然后是 10 个关键字的摘取。当前分析器对关键字的摘取主要从两个地方入手：Meta 值和 Title 值。如果当前页面的头部提供了<meta name=keywords content=(.*)>字样的话，那么我们就将小括号里的内容提取出来，通过中文分词组件分词，将前 10 个词组写入索引数据库里的 key1 到 key10 字段中；如果当前页面没有提供 meta 值，那就将其标题提取出来，并经过中文分词写入索引数据库的 key1 到 key10 字段中。

中文分词在全文搜索引擎中极其重要，不仅要准，还要快，两者同时满足才有应用价值。歧义识别和新词识别一直是中文分词的两大难题。开发一个中文分词组件是容易的，但要最好很难，为了把更多的精力和时间放在整个系统的实现上，作者暂且借用互联网上公开的一些中文分词组件，如 C# 代码编写的 RainSoft (http://www.rainsts.net)，C++代码编写的 ICTCLAS(http://www.nlp.org.cn)。前者分词效果不佳，后者分词效果较好，不过.net 版需要自己改写，java 版(http://www.sourceforge.org)网上已发布。

最后是网页的优先级评判，以确定网页排序。在优先级的评定问题上，提出了以链接数量来衡量网页重要程度的方法。即，如果当前页面上的超链数越多，就认为该网页越重要。各大搜索引擎公司采用的方法基本类似，都是基于超链的分析，如 Google 的 PageRank 算法和 HillTop 算法，百度也是用的超链分析法，具体算法没有公开。除上述排序算法外，很多搜索引擎都不约而同的使用了搜索引擎优化技术，即收费排名，实质上，这是一种盈利模式，而并非一种排序技术。

4 结语

在调试过程中，遇到了诸多问题，尤其是 CPU、内存资源占用以及 Spider 的搜索效率问题。现在的 Spider 程序在占用 CPU 资源较少的情况下能够良好稳定地运行，在带宽资源能够保证的情况下，页面下载的速度基本保持在 1 秒钟 10 个页面左右。

技术创新

受硬件条件所限，无法进行企业级的集群系统测试，所以在页面搜集和索引效率上还有较大的提升和优化空间，小型应用已没有问题。搜索引擎技术是海量信息存储、集群并行运算和自然语言处理等众多高端技术的集成，对中文搜索引擎来说，中文分词也尤为重要。因此，该系统还有很多需要提升和完善的地方，包括：系统的可伸缩性；中文分词系统的自主性；网页排序的公平性；数据存取及运行效率的高效性等等。

本论文的主要创新点：要想在 4 页纸的版面内把一个全文搜索引擎 2/3 的内容叙述清楚，的确很难，尽管如此，本文也尽可能较为全面而详细地阐述了基于 Web 的全文搜索引擎中的核心模块——网络爬虫的设计和实现过程，对于大量存在的专业门户网站，如中国自控网，完全可以从中获取一些有益的信息来开发自己的真正意义上的站内全文检索系统。对于其他搜索引擎系统、如智能问答系统，本文也有很强的借鉴意义。

参考文献

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

[2]印鉴，陈忆群，张钢. 搜索引擎技术研究与发展[J].2005，31(14):54-56 .

[3]陈刚，卢炎生.教育网 BBS 搜索引擎设计与实现[J]微计算机信息，2006，6-3:34-36

[4]Winter.中文搜索引擎技术揭密：网络蜘蛛[EB/OL]. http://article.bwtech.net/artshow_33.htm

[5]Winter.中文搜索引擎技术揭密：中文分词[EB/OL]. http://article.bwtech.net/artshow_30.htm

[6]Winter.中文搜索引擎技术揭密：排序技术[EB/OL]. http://article.bwtech.net/artshow_31.htm

作者简介：徐远超(1975-)，男，湖北武汉人，汉族，讲师，硕士，现主要从事 Internet 网络技术和嵌入式系统研究；刘江华(1982-)，男，广西柳州人，助理工程师，研究方向：搜索引擎技术、自然语言处理；刘丽珍(1966-)，女，副教授，博士，主要研究领域为网络环境下的知识发现和知识管理。关永(1966-)，男，内蒙古包头人，教授，博士，研究方向：智能信息处理。

Biography:Xu Yuan-chao, male, born in 1975, Master, lecturer, His research interests includes Internet technology and embedded system.

(100037 北京 首都师范大学 信息工程学院) 徐远超 刘江华 刘丽珍 关永

通讯地址：(100037 北京 北京西三环北路 56 号北京 801 信箱信息工程学院)徐远超

(收稿日期：2007.5.03)(修稿日期：2007.6.05)

(上接第 110 页)

而且通过仿真实验分析了基于 D-STTD 的 MIMO-OFDM 系统的性能及其在相关信道情况下的性能，仿真结果表明该系统在相关信道的情况下仍有良好的性能。因此在下行链路中采用 D-STTD 的方式可以达到最大的信道利用率并且系统的复杂度可以降低，是一种适宜的选择方式。

参考文献

[1]李静，候思祖. "OFDM 误码率性能分析与研究"，[J]微计算机信息，2006,22:261~264.

[2] "Double-STTD scheme for HSDPA systems with four transmit antennas: Link level simulation results," Temporary document 21 (01)-0701, 3GPP TSG RAN WG1, June 2001, release 5.

[3]A. Forenza, A. Pandharipande, H. Kim, and R. W. Heath Jr., "Adaptive MIMO transmission scheme: Exploiting the spatial selectivity of wireless channels," accepted to IEEE Veh. Technol. Conf., May 2005.

[4]Klein A, Kaleh G K, Baier P W. "Zero forcing and minimum mean-square-error equalization for multiuser detection in code-division multiple-access channels[J]," IEEE Trans. Veh. Technol , 1996.45(5):276~287.

[5]邵怀宗，彭启琮，李玉柏. "一种空时块编码 OFDM 系统中的自适应均衡算法，"电子与信息学报，2004 年 2 月.

[6]Shiu D. "Fading correlation and its effect on the capacity of multi-element antenna systems [J]. " IEEE Trans. Commun., 2000, 48(3): 502~513.

作者简介：胡学斌，男，1980 年出生，山西临汾人，电子科技大学在读硕士研究生，主要研究方向为现代通信中的信号处理；张忠培，男，1967 年出生，重庆万州人，电子科技大学副教授，硕士生导师，主要研究方向为移动通信；黄炜，男，1952 出生，北京市人，电子科技大学副教授，硕士生导师，主要研究方向为现代通信中的信号处理。

Biography:Hu Xuebin (1980-), male, master candidate, research fields: signal processing in modern communication.

(610054 成都 电子科技大学 通信与信息工程学院) 胡学斌 张忠培 黄炜

通讯地址：(610054 成都 电子科技大学 通信与信息工程学院) 胡学斌

(收稿日期：2007.5.03)(修稿日期：2007.6.05)

(上接第 115 页)

同时，模型在形式上也较为统一。因此，有效地解决了传统的 RBAC 模型中存在的问题，成为一种有效而实用的安全模型。

本文作者的创新点：依托信息管理系统(MIS),在其上面定义角色，根据角色的定义来定义不同的用户群，这样就非常方便的通过定义角色把用户和权限的关系直接分开管理，大大的减少了管理上的工作量。

参考文献

[1]洪帆、何绪斌、徐智勇.基于角色的访问控制.小型微型计算机系统.2000.2

[2]季永志、阎保平、续岩、吴开超、沈志宏. 基于角色的访问控制框架的研究和实现.2005.8

[3]叶锡君、许勇、吴国新. 基于角色的访问控制在 Web 中的实现技术.2002.1

[4]肖剑锋. 基于 Struts 与 Hibernate 的 MIS 开发。[J]微计算机信息。2006.7

作者简介：丁振国，1959 年生，，男，陕西省三原人，博士，教授，主要从事计算机网络与信息处理技术方面的研究与教学。陈敏，1981 年生，男，湖北黄冈人，硕士研究生，主要研究方向为计算机网络与信息处理。

Biography:Ding ZhenGuo,male,1959, Born in SanYuan,ShanXi Province Doctor, professor of xidian university,Main research direction is computer network and information process.。ChenMin, male,1981s,Master, Born in HuangGang,HuBei Province, Main research direction is computer network and information process.

(710071 西安 西安电子科技大学)丁振国 陈敏

通讯地址：(710071 西安电子科技大学 396 信箱)陈敏

(收稿日期：2007.5.03)(修稿日期：2007.6.05)

技术创新