文章编号：1007-757X(2010)7-0021-02

基于 Java 的多线程网络爬虫设计与实现

姜梦稚

摘 要：网络爬虫是目前比较流行的一种网页检索工具，其设计和实现也需要不断优化和改进。通过描述网络爬虫设计与实现中所碰到的问题，提供解决这些问题的方法，并给出实现这些目标的网络爬虫设计方法，提供该设计的 Java 语言版实现。

关键词：网络爬虫；链接检索；文字匹配；爬虫设计；多线程

中图分类号：TP316 文献标志码：A

0 引言

目前，对于全球大多数互联网用户来说，搜索引擎是其准确获得所需要信息或者知识的最有效的工具。但是对于所有的搜索引擎来说，最重要的性能指标有两个：查全率和查准率。查全率与搜索引擎搜集的网页数量和质量有关。

本文介绍的是用于搜集网页， 提高查全率的最重要的工具—网络爬虫(Web Crawler)的设计与实现。网络爬虫的主要作用是搜集互联网的网页，也可以用它来定期搜集某个网站的内容，跟踪判断网站的发展，或者做站内搜索引擎。从网络爬虫的工作原理来看，“网络爬虫”是一个比较形象的名字，它是在互联网内，通过网页链接，从当前网页爬到下一个网页来进行网页内容搜集的工具。它所需完成的工作[1]如下：(1)在一个网页上，获取网页的标题和网页中的摘要；（2）将搜集到的网页标题，链接，网页的摘要放入数据库中；(3)根据当前网页的内容，搜集网页中的链接信息，并根据链接顺序搜索相应链接网页的内容。

1 Web Crawler 中的若干问题

不同的 Web Crawler，在设计的时候侧重各有不同，本文所介绍的 Web Crawler 在设计的时候主要考虑解决以下几个问题：(1)Web Crawler 遍历网页中的所有链接，并且能对所搜索的网页进行搜索深度的限制；(2)Web Crawler 能够提取出网页中的摘要和标题信息，并且保存到数据库中；(3)要求能够对已有的搜索引擎的搜索结果再优化，提高所设计的 Crawler 的扩展能力；(4)要求能够采用多线程的方式，提高搜索的效率。

针对前述的四个目标，在设计 Crawler 的时候，具体考虑了如下一些问题。首先 Crawler 的搜索深度的问题，网页中的链接关系是相当复杂的，一组网页之间可以互相包含，网页 A 中的链接可以指向网页 B，同时网页 B 中的链接也可以指向网页 A。因此，网页的搜索深度必须作一定的限制，不能无限制的递归搜索；其次网页的搜索方式也有差别，有些爬虫采用广度优先策略 (BFS)，有些采用了深度优先策略(DFS)[2]，考虑到实现时采用多线程，且所设计的 Crawler 需在搜索的深度上作了限制，所以采用深度搜索的方式。对于搜集所有 url 链接，可以有不同的方式，本文采用的正则表达式。尽管具体的实现有多种方式，搜集链接也可以采用后面所提到的 HtmlParser 包，但是采用正则表达式，是一种方便，快捷的手段，Java 语言也提供对正则表达式的支持。

多线程程序是编程中比较复杂的问题，除了对线程的调试比较困难外，对线程所使用的资源的控制也同样复杂。在 Java 平台下对多线程的编程有充分支持，因此在设计 Crawler 的多线程实现时，采用了 Synchronous 关键字，原子数(AtomicInteger)和线程池[3]，通过使用线程池，在应用启动的时候，设置所需创建的线程数量，一旦线程池为空，则挂起当前请求，等待空闲的线程出现；原子数则可以保证程序中的计数以互斥的方式操作，保证了递增和递减操作的原子性；而 Synchronous 保证了不同线程在访问数据的时候不会出现两个线程同时访问一个数据。

本文中实现的 Crawler 中第三个考虑的问题就是获取网页的内容、提取网页摘要信息和标题信息。网页内容的获取方式有多种，比较常用的就是想网页发出一个 Http 请求，并获取返回的字符流。考虑到实现这种请求/响应方式的复杂性，本文采用了 HtmlParser[4]包来具体实现网页内容的获取。对于标签的获取采用两种方式，一种是采用 HtmlParser 包来获取，另外一种提取文本的方式也可以使用正则表达式[5]，在构造合适的正则表达式时，需要考虑到标签的特殊结构，为了提高文字的抽取效率，可以对一段 html 源码首先过滤掉一些不需要的标签。采用 HtmlParser 包除了能够获得网页的内容外，该包还能提供一系列的获取网页内容的工具类，获取特定标签，并通过标签筛选规则的运用，获取包含有文字的标签，提取出其中的文字作为摘要；对于标题通过获取<title>标签，获得网页的标题。

2 Web Cralwer 的设计

本节介绍 Web Crawler 的设计，包括类的设计和时序图的设计。本文所实现的 Web Crawler 采用 MVC 的设计方式，前台设计成 Web 页面，发送 Web 请求；后台 Servlet 接受前台发送过来的请求。在后台，有如下几部分的内容：描述数据实现的 ICrawlerModel 接口及其实现；表示多线程搜索的 ICrawler 接口、AbstractCrawler 和 MultipleThreadCrawler 类；工具类 IParser 接口和 HtmlParser 类；表示链接的数据结构 Link 和 LinkDepth 类；以及存储结果的 DbAccess 类。上述

作者简介：姜梦稚（1980-），男，上海电机学院电子信息学院，讲师，主要研究方向：形式化技术及其应用、软件开发技术及其应用，上海 200240

类的实现都是采用 Java，数据库使用 MySQL，除此以外还要用 Tomcat Web 服务器，如图 1 所示。

前台的设计使用 jsp+jQuery 的方式，jQuery 是一种 javascript 工具包，提供对 ajax 的支持，能够实现无刷新的页面布局。Ajax 是目前一种流行的设计网页的方式。后台采用 Servlet 运行方式，获取前台通过 ajax 方式提交的参数，根据不同的选择（可能是关键词的搜索，也可能是某一个网址的搜索）进行处理。

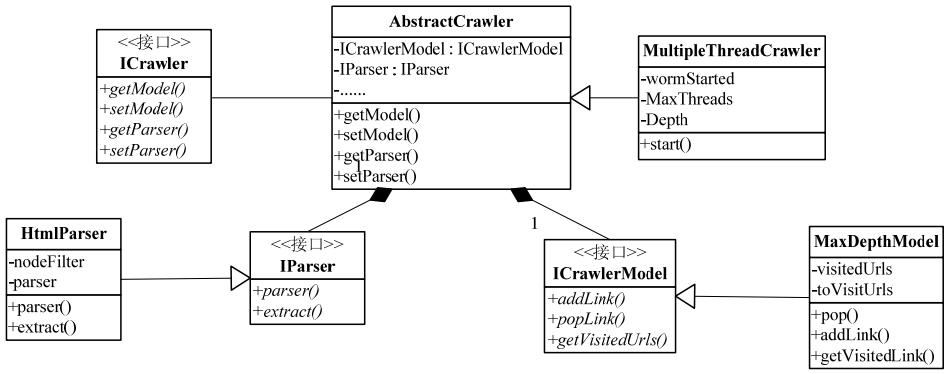

图 1　Web Crawler 的类图[6]

MultipleThreadCrawler 实现了抽象类 AbstractCrawler，它是 Crawler 实现的核心内容，主要对每一次的搜索数据的处理，多线程的协调等。在该类中实现两个私有类 Worm 和 TextExtractor，前者实现对网页链接的搜索，并填充入 Model 中的 toVisitUrls 数据结构中，后者则是对当前搜索的网页提取标题和摘要信息。

MaxDepthModel 是一个存储数据的类，其包括已经访问过的 Url 和未访问过的 Url，并且提供向数据结构中填充未访问的 Url。采用接口的方式，也是为了能够在今后扩展具体实现。

HtmlParser 提供了对文本解析的方法，图 2 给出的用户提交待搜索的网站的时序图：

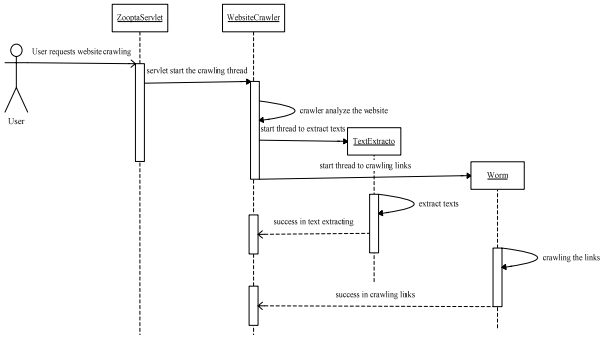

图 2 时序图[6]

3　Web Crawler 的实现

本文设计的 Web Crawler 采用 Java 语言和 MySql 来实现，作为开源工具的组合，被应用在很多重要的领域。除了前述提到的实现细节外，在获取 html 文本中的 url 链接时使用了正则表达式，可以提高文字的匹配速率和程序的运行效率（采用 Java 包匹配往往比较复杂）。同时在抽取 html 文本的文字信息时，也使用了这种技术，由于正则表达式表述的灵活性，并没有一种能够适合所有情况的匹配表达式，只有在实践过程中不断改进和优化。

4　结束语

本文总结了 Web Crawler 在设计和实现过程中遇到的问题，并结合软件设计模式，给出一种 Web Crawler 可行的软件结构，并实现检索，存储，显示等一系列问题，所给出的解决方法有一定的通用性，软件的框架能够根据实际的需要进行改写，可以在对当前已有得搜索引擎的搜索结果进行优化。Web Crawler 是一个需要不断优化和改进的工具，其设计和实现可以采用多种方式，也可以根据 Crawler 的实际需要来设计。

参考文献

[1] 宋晖,张岭,叶允明.基于标记树对象抽取技术的 Hidden Web 获取研究[J].计算机工程与应用,2002(23).

[2] 赫枫龄.用有向图法解决网页爬行中循环链接问题[J].吉林大学学报,2004.3.

[3] 陈昊鹏,饶若楠.Java 编程思想,第 3 版[M].机械工业出版社,2005.5.

[4] Derrick Oswald 等 HtmlParser 参考文档, http://htmlparser.sourceforge.net. [OL].

[5] 史寿伟.正则表达式参考文档,http://www.regexlab.com/zh/regref.htm [OL].

[6] Martin Fowler UML 精粹:标准对象建模语言简明指南[M].2006.3.

(收稿日期：2010-03-19)

ISSN1007-757X

Microcomputer Applications

Monthly (Since 1985)

Zhu Zhongying

Editor-in-Chief

Vol.26, No.7 (General No.207)
July 2010

CONTENTS

Abstract: Web crawler has been a popular searching tool nowadays in people's usage with internet, its design and implementation has always been a ever-improving and optimizing process. In this paper, we present several topics that we design and implement a web crawler, and give some solution to these problems; finally, we give a design of this web crawler and provide java version implementation.

Key words: Web Crawler; Link Searching; Word Match; Crawler Design; Multiple-threaded

Jiang Yilian (Technology Department, Shanxi Radio and TV University, Xi'an 710068, China)

Abstract: Based on the definition, characteristics and architecture of grid, the architecture and resource management pattern of the grid system are provided to implement the distance education resource shares. The application and character of the distance education resource shares are also outlined.

Key words: Grid Technology; Resource Shares; Distance Education; Grid Architecture

Liao Jin, Wang Lin, Xie Jianying (Automation Department, Shanghai Jiaotong University, Shanghai 200030,China)

Abstract: Compared with other semi-dry flue gas desulphurization technologies, the circulating fluidized bed flue gas desulphurization(CFB-FGD) has higher efficiencies and economics. In this paper, A CFB-FGD network control system is designed and implemented. Consisting of Fast Ethernet ,Industrial Ethernet and Profibus ,this system is a 3-level control network .Also it's rational redundant structure improves the dependability of the system .Actual operation indicates the design is rational and stable ,realize the automatic control of flue gas desulphurization ,and improve the desulphurization efficiency .

Key words: Network Control System; Industrial Ethernet; CEB-FGD; Redundancy

Xu Yijun, Zhang Weidong, Yang Yeqing (Department of Automation, Shanghai Jiaotong University, Shanghai200240, China)

Abstract: For the use of combined control with direct lateral thrust and aerodynamics force interceptor missiles within the atmosphere, the paper proposed an optimal sliding-mode mobile terminal guidance law which is suitable to intercept high-speed, large incoming missiles. first, the model of the relative motion will be simplified as the form of time-varying linear equations. Then, the paper proposes an improved optimal sliding-mode guidance law so that the guidance command can be determined for Normal Acceleration directly. In the end of this paper, the results of numerical simulation verify the validity of the design.

Key words: Direct Lateral Jet; Optimal Sliding-mode;High-speed

DEVELOPMENT AND APPLICATION

Wang Chen, Zhang Xianmin (Multimedia System Laboratory, Shanghai Jiaotong University, Shanghai200240,China)

Abstract: This paper focuses on the internal structure and syntax standards of PMS multiplexed stream which is based on the definition of China Mobile Multimedia Broadcasting standard GYZ 234.As a theoretical foundation, using program means to achieve data de-multiplexing of CMMB. We mainly use C language to de-multiplexing data content of TOD, MFS description message and MFS data corresponding to the different types of payload. By analysis and process, we stored the results in an array of structures. And use VS2008 as a design platform for the preparation of man-machine interface, through the user's choice, display the result to users.

Key words: CMMB; Standard; Multiplexed Stream; C Language

Yang Litao, Lu Linji, Fan Zhengyu (Department of Automation ,Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Super-resolution reconstruction refers to restoring a high-resolution image from a series of low-resolution images degraded by warping, blurring and aliasing. A SR reconstruction method for video images based on POCS algorithm has been proposed. POCS algorithm is a widely used method for image reconstruction. This article has modified this method in two aspects. First, the original high-resolution image estimation has been instead of bicubic interpolation image. Second, the problem of edge preservation in reconstruction has been solved by edge detection and PSF modification. Experimental results show that the high-resolution image reconstructed not only increases the resolution but also preserves the details.

Key words: Super-resolution; Projection onto Convex Set (POCS); PSF; Edge Preservation

Wu Liang, Li Shuguang (Intelligent System Laboratory, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: With the development of computer science and database subject, digital information is becoming the first choice of data forms. Nowadays, with the help of large scale search engine, users could find valuable information rapidly. The improvement of search engine with high efficiency which used in enterprise is now a hot subject. This paper describes a search engine based on Keyword Selection (KWS) which aimed to enterprise data structure. By using dictionary based on Hash Structure and measures of Coupling Degree of Double Characters, keyword strings would be splitted into pieces and results would be cached as well. Meanwhile, Sphinx and MySQL database ensure high accuracy and quick response.

Key words: Enterprise Search Engine; Hash Structure; Coupling Degree of Double Characters; Cache

Cui Xuerong[1,2],Zhang Hao[2](1.College of Computer and Communication Engineering,China University of Petroleum (East China), Qingdao 257061, China; 2. College of Information Science and Engineering,Ocean University of China, Qingdao 266100,China)

Abstract: In order to meet users' flexible Web application requirements and avoid the re-development and re-design of software, the template-based Web applications using OWC (Office Web Component) control was proposed. Complete flows to create templates, edit data and view data were designed and the key technologies and realization methods were introduced. In this way, can greatly reduce the workload of software developers, enhance system flexibility, meet the users to change the display, input methods and improve man-machine interface friendly.

Key words: Office Active X Control; OWC; Template; Web Application

TECHNICAL COMMUNICATION

Situ Fang[1], Zhang Haolong[2], Cao Jian[1] (1.Department of Computer Science and Technology, Shanghai Jiaotong University, Shanghai