文章编号:1005－0523(2006)01－0109－04

网上求职机器人的软件设计

张红斌

(华东交通大学 信息工程学院,江西 南昌,330013)

摘要:给出了网上求职机器人的主要设计思想和设计流程,并对使用的关键技术进行了重点的介绍.

关 键 词:网上求职;机器人;Spider;Aggregator;多线程;JavaMail

中图分类号:TP393.4 **文献标识码**:A

1 网上求职概述

互联网的迅速发展,使得越来越多的求职者把网上求职作为一种新兴的求职方式.上网投递个人简历、求职信,甚至是网络面试都已成为时下流行的网上求职方法.众所周知,基于 Internet 的网上求职便捷、迅速、信息覆盖面广,能在较短时间花费较低的成本获取较好的效果,当然,如果求职盲目,或信息闭塞也会给求职者带来很多的不便,如何更好地挖掘网络招聘信息,必将有助于应聘者的成功择业.因此,笔者着手开发了一个网上求职机器人软件.开发工具选择 Java＋JDBC－ODBC 桥接器＋SQL Server 2000.

2 网上求职机器人的软件实现

2.1 网上求职机器人的设计思想

网上求职机器人的设计基于三个事实:1)求职信息覆盖面要广;招聘网站很多,如何在短时间内从大型招聘网站获取丰富的信息至关重要;2)求职信息内容要有筛选性;网上招聘信息纷繁复杂,如何有效地提取出有价值的信息十分必要.3)求职信息要能及时告之求职者.招聘职务转瞬即逝,如何在第一时间就把招聘信息通知求职者,十分关键.

所以,网上求职机器人的总体设计思想是:覆盖大型招聘网站→抓取招聘信息→提取有价值信息→告之招聘者.遵循这一思想,网上求职机器人的工作流程是:首先根据招聘者指定的网站域名或关键字进行网站信息下载,下载 HTML 网页到本地;然后,机器人自动对 HTML 中的信息进行解析,例如“招聘职务”、“工作地点”、“工作年限”等,并把解析的结果进行分类汇总;最后,机器人再从这些信息中选取出有价值信息,以邮件的形式通知招聘者.

2.2 网页下载模块的设计

网页下载程序也称 Spider 程序,它所做的工作就像蜘蛛一样在网状结构的 Internet 上不断地爬行.网页下载过程中对 URL 进行了队列的管理,共设置了 4 个 URL 队列,分别是等待队列、执行队列、错误队列和完成队列.

Spider 发现新的 URL 后,即将它加入等待队列,然后扫描等待队列,此时,等待队列若为空,Spider 工作完成,如果等待队列仍存在未被处理的 URL,则取出队首 URL,下载相应网页,并把该 URL 移入到运行队列中,若该 URL 还存在其他链接,则继续把该链接的 URL 加入到等待队列中,否则,该 URL 处理完毕,把它移入到完成队列中.在下载过程中出现了死链或出错则把该 URL 移入到错误队列中.

收稿日期:2005－10－20

作者简介:张红斌(1979－),男,江苏如皋人,讲师.

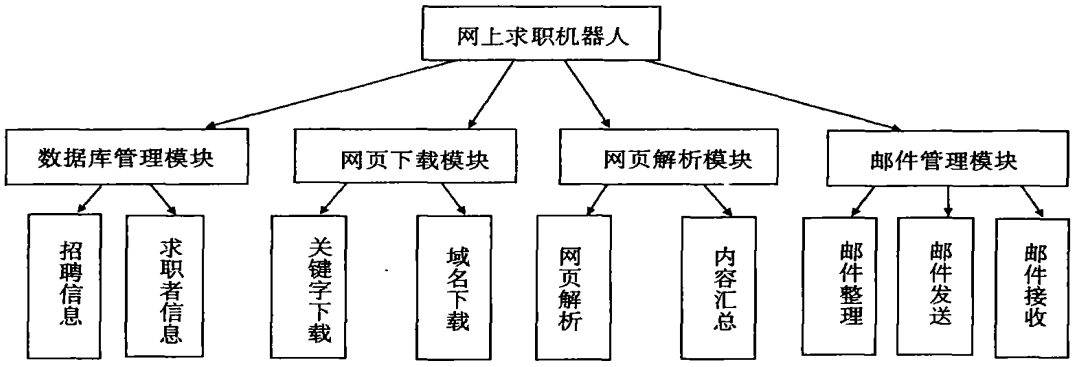

图 1 网上求职机器人软件模块构成

考虑到执行速度和效率，为了充分发挥 Java 的多线程优势，开辟多个 SpiderWorker 线程分别处理对应的等待队列中的 URL，即启动 Spider 对象后即从 URL 等待队列中取出队首 URL 并将其分配给一个 SpiderWorker 对象，具体网页的下载工作是由 SpiderWorker 完成的．同时，对 URL 状态队列进行基于 SQL Server 数据库的管理．

2.3 网页解析汇总模块的设计

网页解析汇总程序也称 Aggregator 程序．它把 Spider 下载的网页进行信息提取，对于网上求职机器人来说，“公司名称”、“联系方式”、“公司网站”、“招聘职务”、“工作年限”等均是感兴趣信息．在对多个网页进行信息提取之后，再把这些信息汇总到某个信息载体中，如 HTML 网页、数据库或 XML 中，以便对信息的再检索．其中 HTML 网页可以直接发布到 WEB 上，供求职者查询．

3 网上求职机器人设计的关键技术

3.1 多线程技术

由图 2 可知 Spider 是多线程的，必须有一种方法在不同的线程间分配任务．这个工作由 SpiderWorker 对象来完成．通过这个对象来维护线程池，同时承担 Spider 创建和销毁线程对象的任务．SpiderWorker 对象子类化 Thread 对象，在构造 Spider 对象的同时，生成了 SpiderWorker 对象以进行线程的管理．SpiderWorker 对象的生成在 Spider 类中，根据线程池大小创建 SpiderWorker 对象．SpiderWorker 对象的开始执行在 Spider 对象的 run()方法中，通过 pool[i]．start ()方法启动线程．Spiderworker 对象的 start ()方法被调用后，执行 SpiderWorker 的 run()方法，该方法唤醒已经初始化过的 Spider 线程，为进一步处理网页做好准备．

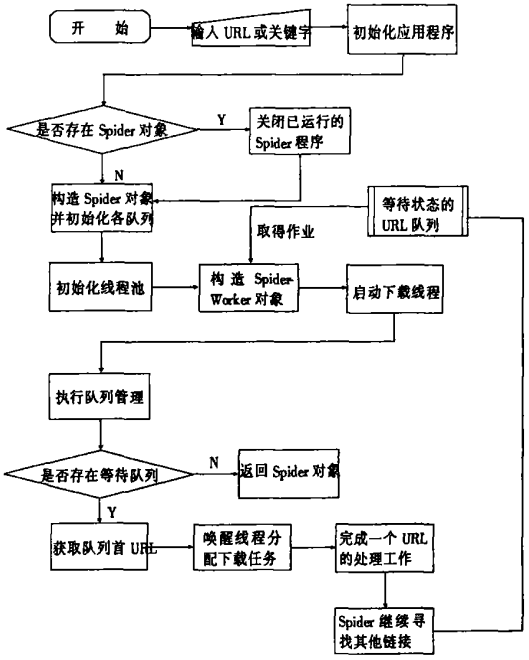

图 2 网页下载模块的实现流程图

3.2 套接字技术

Java 中的套接字编程方法定义了两个类 Socket 和 ServerSocket，其中 ServerSocket 类用于设计服务器程序，而 Socket 类用于设计客户端程序．一旦服务器和客户端连接成功，就使用输入/输出流技术来完成客户端和服务器之间的通信，任一端都可以往套接字读写数据．构件主要基于 Socket 套接字基础之上的 Http、SMTP 以及 FTP 协议的编程．在 Spider 程序中使用套接字技术完成与对应 URL 的连接．

3.3 输入/输出流技术

Spider 程序和 Aggregator 程序都使用到了输入/输出流技术．在 Spider 程序中当创建了套接字连接后，需要使用文件输出流把 HTML 网页保存到本地

硬盘中,而在 Aggregator 中需要把提取的网页信息写入 HTML 和 XML 中.

3.4 JavaMail 技术

JavaMail 规范封装了所有常规消息系统的公共特性和方法,并且它可以实现所有的邮件协议,包括 SMTP,POP3,IMAP 等,所以,在本软件中选择 JavaMail 技术进行邮件编程.JavaMail 定义了五个重要的类:Session,Message,Transport,Store 和 Folder.

本软件中在邮件发送过程中,首先从数据库中取出求职人的 Mail 地址,接着从指定的 SMTP 服务器获取一个 Session 对象,然后以这个 Session 对象为参数,创建一个 Message 对象,指定好 Message 对象的各个属性后,直接调用 Transport 对象的静态方法 Send 就可以完成邮件发送了.

4 网络机器人的软件实现

4.1 网页下载模块的实现

构造一个 Spider 程序必须明确几个类和接口.首先,必须实现一个名为 ISpiderReportable 的接口,Spider 程序将把所有找到的网页返回给实现这个接口的类.其次要实现一个名为 SpiderWorker 的类来管理每个 Spider 线程.ISpiderReportable 定义如下:

```
Public interface ISpiderReportable
{
Public boolean foundInternalLink(String url); //发现内部链接
Public boolean foundExternalLink(String url); //发现外部链接
Public boolean foundOtherLink(String url); //发现其它链接
Public void processPage(HTTP page); //处理 Web 页面
Public void completePage(HTTP page,boolean error); //判断页面处理是否完成
Public boolean getRemoveQuery(); //是否清空查询字符串
Public void spiderComplete(); //工作是否完成}
```

SpiderWorker 泛化之 Thread 类,它的主要函数如下:

```
{
Public boolean isBusy()//判断该 SpiderWorker 当前状态是忙还是空闲
Public void SpiderWorker()//构造 SpiderWorker 对象
Public void run()//唤醒已经初始[illegible]过的 Spider 线程,准备下载工作
Public void processWorkload()//给它所管理的 Spider 对象分配下载任务
Public boolean getHTTP(String http)//获取 HTTP 链接状态}
```

4.2 网页解析模块的实现

根据招聘网页上的信息描述,共设计了 14 个解析函数分别解析相关信息.它们是 getZipcode()、getCompanyinfo()、getHR()、getPhone()、getFax()、getInfo()、getSalary()、getResumelang()、getLang()、getYear()、getPlace()、getDate()、getMail()、getNum()和 getSite().现以 getInfo()为例说明其实现方法,其他均同理.

```
public static String getInfo (String p)//指定待解析网页的本地地址
{
String info= "";
try {
String url;
url = "http://127.0.0.1/job/"+p; //因为已经把网页下载到了本地,所以直接
//向本地 WEB 服务器发送套接字连接要求
HTTPSocket http = new HTTPSocket(); //创建套接字连接
http.send (url, null);//发送连接要求
int i = http.getBody().indexOf("职位描述");//解析职位描述信息
int j = http.getBody().indexOf("申请这个职位");
info = http.getBody().substring(i,j);
}catch (Exception e) { }
return info;//返回解析的结果 }
```

4.3 网页汇总模块的实现

网页汇总模块就是对信息进行分类汇总.当然 Spider 程序会定期的对这些保存在磁盘中的 Web 页面进行更新,每周进行一次.下面以发布方式选择 HTML 文件为例说明其实现方法,随着 Web 页面的更新,这些存储在 HTML 文件中的信息也会被定期更新.

```
ps.println ("<table width= '75%' border= '1' align= 'center' >");
ps.println("<tr><td align= 'center'>"); //表头开始
```

```
ps.println("<td align='center'>"); ps.println("电邮</td>");
ps.println("<td align='center'>"); ps.println("发布日期</td>");
ps.println("<td align='center'>"); ps.println("工作地点</td>");
ps.println("<td align='center'>"); ps.println("招聘人数</td>");
ps.println("<td align='center'>"); ps.println("工作年限</td>");
ps.println("<td align='center'>"); ps.println("薪水范围</td>");
ps.println("<td align='center'>"); ps.println("外语要求</td>");
ps.println("<td align='center'>"); ps.println("简历语言</td>");
```

其他数据类似,在此省略.

5 总 结

本文讨论了网上求职机器人的设计思想和设计流程,并给出了实现它的关键技术,笔者使用该软件对 51job 进行了测试,测试环境是 PⅣ 2.8G,内存 512 M 的主机,网络环境是 100 M 内网,100 M 交换机,以"高级程序员"关键字进行测试.测试结果表明软件能在规定时间内正确地下载指定页面,并在较短的时间里将下载结果汇总发布.考虑到软件的通用性和智能性,笔者还将就机器人的自主学习、相似信息的过滤以及软件的分布性进行进一步的研究和探讨.

参考文献:

[1] Jeff Heaton,童兆丰,等.网络机器人 Java 编程指南[M].北京:电子工业出版社,2002.

[2] 谭淑英,刘丽华.Web Robot 技术及其 Java 实现[J].中南工业大学学报,2001(3).

[3] 印 昊,王行言.Java 语言与面向对象程序设计[M].北京:清华大学出版社,2003.

[4] 于 磊,潘 郁.智能学习型网络机器人[J].计算机工程,2004,30(13).

[5] 侯晓强,徐春荣,勾海波.Java 服务器编程实例[M].北京:清华大学出版社,2003.

Software Design of the Bot Seeking-job on Internet

ZHANG Hong-bin

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: It mainly introduces the design idea and design flows of the bot on Internet; it focuses on the key technologies used in the software.

Key words: seeking-job on internet; bot; spider; aggregator; multi-thread; javaMail