Java实现利用搜索引擎收集网址的程序 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/220/2021_2022_Java_E5_AE_9E_E7_8E_B0_c104_220841.htm 我这里讲的不是怎么使用搜索引擎，而是怎么让程序利用搜索引擎来搜集网址，这有什么用？很有用！网上动辄有人叫卖网址数据库，如发布软件网址、邮件地址、论坛网址、行业网址，这些网址是怎 么来的呢？不可能是人手工收集而来的，都是让程序利用搜索引擎取到的，如果您需要某类网址信息数据，就跟我来一起研究一下，非常简单。本文采用Java语言写成，以google和百度搜索引擎为对象。我们要利用google、百度搜索引擎的搜索规则中的两条，关键字搜索和inurl搜索。什么是inurl搜索，就是你所要搜索的网址中本身带有的关键字，比如http://www.xxx.com/post.asp ,这个网址就含有post.asp这样的关键字，在搜索引擎中填写规则是 inurl:post.asp,这是收集网址的关键，因为很多网址本身会带有特定的信息，比如软件发布的网页网址信息中多含有 publish、submit、tuijian这样的信息，如http://www.xxx.com/publish.asp,这样的网址多是发布信息的网页，在结合网页中本身可能含有的关键字，就可以用搜索引擎搜索出结果，然后我们利用程序将结果取回，对HTML页面进行分析，去除没有用的信息，将有用的网址信息写入文件或者数据库，就可以给其它应用程序或者人来使用了。第一步，用程序将搜索结果取回，先以百度为例，比如我们要搜索软件发布的网页，关键字采用“ 软件发布 版本 inurl:publish.asp",先登录百度看看，将关键字写入，然后提交，在地址栏就会看到

http://www.baidu.com/s?ie=gb2312amp.sr=amp.cl=3amp.wd=软件发布 版本 inurl:publish.aspamp.si=amp.ie=gb2312amp. wd=软件发布 版本 inurl:publish.aspamp.cl=0，其中rn表示一页显示多少个结果，wd=表示你要搜索的关键字，pn表示从第几条开始显示，这个pn将是我们程序循环取结果的变量，每20条循环一次。我们用Java写的程序来模拟这个搜索的过程，用到的关键类为 java.net.HttpURLConnection,java.net.URL，先写一个提交搜索的class,关键代码如下：以下是引用片段： class Search { public URL url. public HttpURLConnection http. public java.io.InputStream urlstream. ..... for(int i=0.i .i { ..... try { url = new URL("www.baidu.com/s?lm=0amp.rn=20amp.ct=0amp.pn=" beginrecord "amp.hl=zh-CNamp.newwindow=1amp.sa=N&amp.ie=UTF-8,其中编码要用ie=UTF-8,start表示从第几条记录显示，需要注意的是google对浏览器还要检查，如果浏览器不符合它的要求，将返回错误代码，所以在模拟浏览器提交中，我们要多加一行代码，修改关键部分要将http属性中的User-Agent设置为常用的浏览器，比如Mozilla/4.0,代码如下：以下是引用片段： try { http = (HttpURLConnection) url.openConnection(). http.setRequestProperty("User-Agent", "Mozilla/4.0"). http.connect(). urlstream = http.getInputStream(). }catch(Exception ef){}. 100Test 下载频道开通，各类考试题目直接下载。详细请访问 www.100test.com