基于Python的网络爬虫技术研究

王碧瑶
(山东农业大学信息学院,山东泰安 271000)

摘要:专用型的网络爬虫能够得到想要的返回结果,本文就以拉勾网作为例子,对基于Python的网络爬虫技术进行研究和分析。
关键词:Python;网络爬虫技术;搜索引擎
中图分类号:TP393 **文献标识码**:A **文章编号**:1007-9416(2017)05-0076-01

1 爬虫系统需求的分析和设计

利用模块化的设计来对网络爬虫系统进行开发,一个通用的爬虫架构需要有爬虫调度端,URL管理器,网页下载器,网页解析器这4个模块。爬虫调度端去启动、停止或者监视爬虫运行情况,URL管理器去对将要爬取的URL和已经爬取过的URL进行管理,网页下载器将URL管理器指定的URL网页下载下来存储为字符串,字符串传送给网页解析器进行解析,并将其中未被抓取过的URL送入URL管理器中。

公司名、地址以及薪水等都需要被抓取然后保存到文件中。拉勾网加载职位信息采用异步加载方式,所以对一系列网络请求分析之后,发现是positionAjax.json请求去响应职位信息,网页存储职位信息采用的是json格式,并且json的层级结构为content-positionResult-result,所以采用json格式读取这种层级结构下的数据。其次就是分页的设计,在json格式content-positionResult-totalCount下存储着该种搜索下职位信息的总个数,发现每页的职位个数为15,只要totalCount/15就可以得出爬取的页数。

2 基于Python的网络爬虫技术的实现

下面就利用拉勾网来进行基于Python的网络爬虫的实现。爬虫首先要获取的是所有待抓取的链接地址,再就是对链接地址中的职位信息进行有效解析,随后保存在文件当中。

2.1 抓取与解析的实现

首先确定URL为'https://www.lagou.com/jobs/positionAjax.json',浏览器提交的FormData有三个参数first、pn、kd,后两个分别代表当前页数和搜索的关键词,页数每次增加1,并需要模仿浏览器发送post请求,并将headers进行伪装,如下所示:

```
headers={'User-Agent':'Mozilla/5.0(Windows;U;Windows NT 6.0) Gecko/20070309 Firefox/2.0.0.3'}
```

得到页面信息之后再对我们想要的信息进行抓取,对于这次要爬取的信息,json是非常适合的方法,仅列出薪水部分实现如下:

```
for posi in self.json['content']['positionResult']['result']
i['positionSalary']=position['salary']
```

2.2 爬虫的抓取策略概述

策略一是深度的优先遍历策略,起始的URL会被网络爬虫最先开始,随后对每个URL进行追踪直到处理结束该处的URL,随后就转向下一处的URL继续追踪;策略二是反向的链接数,其是其他的URL指向一个网页的数量并以此对网页重要程度进行评价,优点是能够根据重要性的网页进行优先爬取;策略三就是宽度优点的遍历策略,将新下载网页包含的链接直接追加到待抓取URL队列末尾;策略四是最佳的优先搜索策略,通过分析目标网页主体和目前的URL的关联性,并对评价好的URL进行择优抓取;总而言之,采取不同爬取策略是殊途同归的,达到重要网页的优先爬取。

2.3 反爬虫的策略采取概述

如果认为单线程爬虫设计方法速度较慢,不会发生被封的情况,其实是不对的,必须利用其他有效策略来操作才行,下面简要分析相关的策略。

策略一是禁止cookies,若有些网站发现某个访问频繁发生,可能会将其视作爬虫,并且通过cookies去识别访问者的身份而拒绝该用户的访问,所以在scrapy爬虫框架中设置COOKIES_ENABLES=FALSE即可禁止网站访问cookie,对于通过cookie识别来禁止爬取数据的网站非常的有效。策略二,HTTP协议中的一个字段是User agent,伪装其在浏览器是个很好的策略。策略三,可以对访问频率进行降低,让对方无法从访问量中看得出来。有众多的办法去实现:对每个页面间歇抓取、每天抓取的页面数量进行限制等,其实scrapy爬虫框架中,就可以设置下载等待时间download_delay来减少被禁止爬取的风险。

3 结语

网页的内容以及网站的信息可以利用Python自带的框架进行获得,然后利用正则的表达式等对需要的信息进行提取以及分析。互联网站点无数,其开发过程遵循的原则、代码的风格也不一样,所以处理不同风格的代码确实存在不小的挑战。

参考文献

[1]钱程,阳小兰,朱福喜.基于Python的网络爬虫技术[J].黑龙江科技信息,2016,36:273.
[2]姜杉彪,黄凯林,卢昱江,张俊杰,曾志高,刘强.基于Python的专业网络爬虫的设计与实现[J].企业科技与发展,2016,08:17-19.

收稿日期:2017-04-26
作者简介:王碧瑶(1996—),男,汉族,山东济宁人,本科在读,研究方向:计算机。