基于 Python 的网络爬虫程序设计

文/郭丽蓉

摘要

网络信息量的迅猛增长，对如何从海量的信息中准确的搜索到用户需要的信息提出了极大的挑战。网络爬虫具有能够自动提取网页信息的能力。本文根据某信息网的特点，提出了一种基于 Python 的聚焦爬虫程序设计。实验结果表明：本程序具有针对性强，数据采集速度快、简单等优点，有利于对其它的数据进行后续的挖掘研究。

【关键词】网络爬虫 Python

1 爬虫技术

网络爬虫，又称网页蜘蛛（web spider），是一个功能强大的能够自动提取网页信息的程序，它模仿浏览器访问网络资源，从而获取用户需要的信息，它可以为搜索引擎从万维网上下载网页信息，因此也是搜索引擎的重要组成部分。

根据爬取的对象、使用的结构及技术，爬虫可分为：

1.1 通用型爬虫

该爬虫又称为全网爬虫，主要用在搜索引擎，从初始的 URL 到全网页面，但需要的存储容量大，速度要求快，工作性能强大。

1.2 聚焦型爬虫

该爬虫专注某一方面，只搜索事先定义的关键信息。

1.3 增量型爬虫

每隔一段时间更新，重新爬取，更新数据库。

1.4 深层爬虫

该爬虫一般需要登录提交数据，才能进入页面提取信息。

利用网络爬虫，能够帮助用户解决上网浏览过程中的一些信息的快速抓取及保存。比如日常上网浏览网页过程中，经常会看到一些喜欢的图片，希望保存下来作为素材使用，一般的方法就是通过单击鼠标右键选择另存为来保存图片，如果批量保存图片工作量会比较大，而利用设计的网络爬虫来爬取图片，自动化处理，快速高效。同时，利用爬虫可以获取大量的感性认识中得不到有价值数据，为一些决策提供依据。

表 1：各个数据定位的 Class

数据	元素	Class
公司名称	div	comp_name
职位	span	name
薪水	'p	job_salary

2 Python概述

Python 语言是一种功能强大面向对象的解释型计算机程序设计语言，能有效而且简单地实现面向对象编程。Python 语言属于语法简洁清晰的开源编程语言，特色之一是强制用空白符 (white space) 作为语句缩进。

Python 具有丰富的标准库和强大的第三方库。它常被昵称为胶水语言，能够和其他语言制作的各种模块（尤其是 C/C++）很轻松地联结在一起，易于扩展。常见的一种应用情形是，使用 Python 快速生成程序的原型（有时甚至是程序的最终界面），然后可以用更合适的语言改写其中有特别要求的部分，比如对于性能要求特别高的 3D 游戏中的图形渲染模块，完全可以用 C/C++ 重写封装为 Python 可以调用的扩展类库。

在使用之前，必须搭建好使用环境。到 Python 官网下载针对用户所使用的操作系统 Python 版本来安装，安装完成后需要设置环境变量便于启动 Python。同时可选择一款合适的编辑工具来完成爬虫的编写。

目前 Python 的版本有 2.X 和 3.X。两者主要在语法、编码、性能、模块上有些不同。

使用 Python 开发爬虫的优点：

（1）语言简洁，使用方便。

（2）提供功能强大的爬虫框架。

（3）丰富的网络支持库及网页解析器。

本文中的爬虫是在 Python 3.6 环境下调试完成的。

3 爬虫案例

本文通过 Python 语言来实现一个简单的聚焦爬虫程序，把需要的招聘信息爬取保存到本地。该爬虫的功能是爬取某信息网上关于互联网职位的信息，并将其发布的招聘信息保存在 Excel 文档中。

3.1 解决Where、What、How的问题

（1）Where：爬哪里，确定要抓取的页面。解决这个问题由用户对数据的需求来决定。

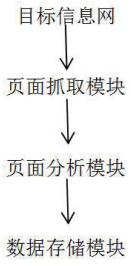

图 1：爬虫流程图

（2）What：爬什么，分析上述页面，确定从页面中爬取的数据。

（3）How：怎么爬，可使用 Python 强大丰富的标准库及第三方库来完成。这是爬虫的核心部分。尤其是对网页的解析，可使用正则表达式、BeautifulSoup、lxml 来解析网页，三种方法各有千秋，使用时可根据用户的熟练程度和需要选择一种适合的解析方法。

3.2 具体实施

该爬虫系统主要由三个模块：页面抓取模块、页面分析模块、数据存储模块，三个模块之间相互协作，共同完成网页数据的抓取。

（1）爬虫实现流程如图 1 所示。

（2）打开某信息网招聘信息，该网站 URL 是爬虫主要的处理对象，打开互联网职位页面并分析网页源代码结构，代码如图 2 所示。

分析代码过程中，可利用开发者工具确定每个数据对应的元素及 Class 名称。例如本页面的公司名称、职位、薪水对应的元素及 Class 如表 1 所示。

（3）确定爬虫方法，导入程序中所用到的库。对服务器发出请求，打开网页，需要使用 requests 库，本爬虫中分析网页使用的是 BeautifulSoup 方式，需要用到 BeautifulSoup 库，存储数据用到库 xlwt，整个爬虫程序中用到的库都需要导入。部分代码及注释（# 开始为注释行）为如下：

```
# 导入程序中所用到的库
import requests
from bs4 import BeautifulSoup
importxlwt
# 打开网页
```

图 2：部分网页源代码

```
r=requests.get(link,headers=headers)
# 通过 BeautifulSoup 的 find_all 方法解析页面
soup=BeautifulSoup(r.text,"lxml")
job_list=soup.find_all('li',class_="job_itemclearfix")
# 查找公司名称、职位、薪水等信息
for job in job_list:
comp=job.find('div',class_='comp_name').text.strip()
name=job.find('span',class_='name').text.strip()
salary=job.find('p',class_='job_salary').text.strip()
```

（4）爬取结果，本地保存。爬虫从网页中提取数据后，需要将数据保存下来，Python 提供存储数据的方法：可保存到文件中（包括 TXT 文件和 XLS 文件），也可存储在数据库中（包括 MySQL 关系数据库和 MongoDB 数据库）。该爬虫对招聘信息进行本地保存，存储格式为 .XLS 文件。部分代码如下：

```
# 创建 Excel 文档保存将要爬取的数据
book=xlwt.Workbook(encoding='utf-8')
sheet1=book.add_sheet(u'sheet1',cell_overwrite_ok=True)
# 在创建好的 Excel 文档中写入信息标题
heads=[u' 公司名称 ',u' 职位 ',u' 薪水 ']
ii=0
for head in heads:
sheet1.write(0,ii,head)
ii+=1
# 调用 write 方法将查找到的信息写入 Excel 文档中
for data in (comp,name,salary):
sheet1.write(i+1,0,comp)
sheet1.write(i+1,1,name)
sheet1.write(i+1,2,salary)
i+=1
# 保存 Excel 文档
book.save('joblist.xls')
```

保存结果如图 3 所示。

可以看到和网页中提供的招聘信息是一致的。

3.3 改进和完善

上述爬虫只能爬取网页上的第一页招聘信息，如果需要爬取所有页面信息，可根据分析网页 URL 地址的规律，使用 for 循环爬取。分析比较：

0d303655-002e-4964-f90b-fa2a1cb1f628&ClickID=3

0d202408-002e-46bc-d2a9-36b62fb6175b&ClickID=2

到下一页 ClickID 的值增加 1，需重新构造 URL。

部分代码如下：

```
fori in range(0,10):
link1=link+str(i)
```

4 结束语

总之，在大数据时代的今天，用户对各类数据的需求越来越多，对数据进行有效的分析可对相关决策提供依据，爬虫作为一种自动收集数据的手段，有广阔的应用。同时，结合学院实际情况，可以将爬虫技术应用在网络舆情分析及网络犯罪线索发现中。

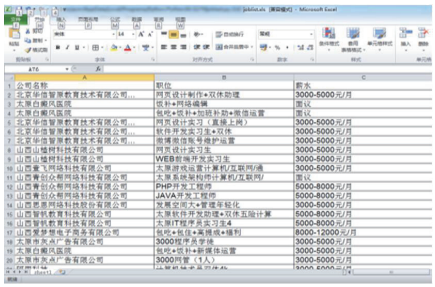

图 3：存储的 Excel 文档

参考文献

[1] https://baike.baidu.com/item/Python/407313?fr=aladdin.

[2] 谢克武 . 大数据环境下基于 Python 的网络爬虫技术 [J]. 软件开发 , 2017, 5 (44).

[3] 唐松，陈智力铨 . Python 网络爬虫 [M]. 北京：机械工业出版社 , 2017.

[4] 陈琳，任芳 . 基于 Python 的新浪微博数据爬虫程序设计 [J]. 信息系统工程 , 2016, 9: (97-99).

[5] 于成龙，于洪波 . 网络爬虫技术研究 [J]. 东莞理工学院学报 , 2011, 6: (25-27).

[6] 韩菲，金磊等 . 基于 Python 的实时数据库设计 [J]. 仪表仪器用户 , 2017, 6 (28).

作者简介

郭丽蓉 (1979-)，女，山西省文水县人。讲师。

作者单位

山西警察学院网络安全保卫系　陕西省太原市 030021