ISSN: 2661-3719(Print); 2661-3727(Online)

电影数据大批量获取研究

周　睿　通讯作者：王　争
电子科技大学成都学院　四川成都　611731

摘　要：随着大数据时代的不断发展，数据对于个人，企业乃至国家来说变得越来越重要。但当人为去翻页查询定位数据，下载数据，效率会很低。网络爬虫技术就能够在大量的数据中定位到有效数据，并把它获取出来，进行本地化保存或者数据库存储，方便用户更加直观对比和查阅。本文网络爬虫是基于Python语言，结合第三方xlwt库、lxml库和matplotlib库实现对网站电影数据的大批量获取、本地保存和数据可视化。

关键词：Python；xlwt；lxml；matplotlib；电影数据

1　引言

在大数据时代下，数据安全是每个人都关注的事，对个人来说，它包含了一个人的个人隐私；对于企业来说，掌握各类数据资料就如同掌握了更多的筹码。目前，电影消费已经成为生活娱乐的一部分，每一部电影有它的名字，导演，主演，评分，分类和简介等内容，但是当我们想对比查找时，就需要一页一页的切换，无疑是费时费力的，因此，我们可以通过代码来自动获取信息，并将数据保存在.csv或者.xls文件中，达到快速，便捷的效果，甚至做成饼状图或者直方图对比。

网络爬虫技术也被称网络机器人或者网络蜘蛛[1]，实质上就是用代码来模仿人类的行为，对网页发起访问，并对网站内容进行获取，结合第三方库实现本地保存，在运行时，为了避免被反爬机制检测到，通常需要添加UA伪装。目前为止大部分网站都是使用超文本标记语言（英语：HyperText Markup Language，简称：HTML），它是一种用于编写网页的标准标记语言[2]，这种语言编写的网页拥有标签，属性等，这种语言对网络爬虫提供了非常有利的条件。

本次电影数据的爬取全部使用Python语言爬取，Python语言具有可移植性强，易学，简单等特点；应用领域覆盖率科学运算、数据库编程、GUI编程、Web开发、系统运维等多方面。随着Python2.x和Python3.x的发展，实现的功能也在不断的丰富和强大。近年来，Python语言越来越受欢迎，在2021年10月，语言流行指数的编译器Tiobe将Python加冕于最受欢迎的编程语言，2020年，编程语言首次排名于Java、C语言和JavaScript之上。由此可见，Python语言的发展前景将会越来越好[3]。

2　Xpath解析应用

目前网络爬虫常用的解析方法有正则表达式，Beautiful Soup和Xpath解析；正则表达式又被称之为规则表达式，原理就是事先定义好一些特定的字符以及一些特定字符组合，组合成一种“规则字符串”，再通过一种过滤逻辑匹配到用户所需要的内容，Python语言的re模块提供了对正则表达式的支持，相对于其他两种解析方式，正则表达式解析局限性高；BeautifulSoup行业内也被称为美味汤，它是一个工具箱，通过解析文档为用户定位到指定位置，相对简洁简单，是一种常用的解析方法，使用BeatifulSoup解析不需要多少代码就可以写出一个完整的应用程序；Xpath（全称XML Path Language），即XML路径语言，Xpath依赖于Python中的lmxl库第三方库，lxml库主要功能就是解析和提取XML和HTML中的数据，功能十分强大。

Python中安装lmxl库的方法有很多，通常直接在cmd运行窗口中输入：pip install lxml；也可以使用Pycharm或者Anaconda进行安装。

在Xpath中，有七种类型的节点：分别为元素、处理指令、属性、文本、注释、文档节点以及命名空间[4]，节点是通过沿着step或者路径来选取的，使用这七种类型的节点，可以更加准确的从网页获取到用户需要的内容。

Xpath解析原理：①实例化一个etree对象，需要将被解析的页面数据加载在这个对象中；②调用etree对象中的xpath方法，并结合表达式进行数据获取。

使用方法：①实例化etree：from lxml import etree；②将网页数据加载到对象中：若解析本地html代码：使用etree.parse（FilePath）；若解析互联网获取的数据：使用etree.Html（response_text）；③路径表达式：包含属性

作者简介：周睿，男，汉族，2000年9月，籍贯：重庆市渝北区，学历：本科，毕业院校：电子科技大学成都学院，研究方向：智能科学与技术，邮箱：1191583334@qq.com。

定位、索引定位和文本获取等。

3 数据获取

首先，在爬取网页前，应该检查该网页是否有robots协议（统一小写：robots.txt），该协议是君子协议，它是一种存放在网站根目录下的文本文件，它通常提示了该网站中的有哪些数据可以被网络爬虫获取，哪些不可以被获取，robots协议并不是一个规范，而是行业内的规则，所以不能完全保护网站的隐私，随着大数据时代的发展，以及大数据公司逐渐变多，君子协议也逐渐被忽视，因此，在爬取网站时，合理选择网站，切勿强行爬取数据，从而防止因为兴趣学习网络爬虫导致从入门到入狱的后果。robots协议查看方法一般是直接在网页地址后面添加‘/robots.txt’查看。（例如：https：//www.xxx.com/robots.txt）。

其次，通过import导入本次爬虫所需要的requests库、xlwt库、lxml库和matplotlib库。然后定义一个URL（Uniform Resource Locator，统一资源定位器，简称网络地址），该网址就是需要爬取的地址，电影网页需要换页操作，观察两张及以上的电影网页变化点在何处，电影网站爬取是数字从1、2、3……的变化，使用.format（）传入字典的形式传送数字，再加上for循环语句实现输入一个数字，实现从起始页开始爬取到该数字的页面数结束。为了实现代码完整性，爬取时添加UA伪装。

对爬取网页进行检查，观察获取形式为get请求，使用response=requests.get（url）对网页进行获取，防止后续出现编码错误，用response.encoding=response.apparent_encoding对编码格式进行修改，再使用response.text获取页面文本信息。

使用Xpath对获取到的互联网文本信息进行解析，再使用.xpath（）对需要的属性定位，可能出现一个父标签下拥有多个子标签，这时就需要索引定位，索引时Xpath与Python语法不同，Python是从0开始索引，而Xpath是从1开始，获取到需要的数据后返回是一个列表。

对获取到的列表进行循环从而访问到父标签下的子标签，发现获取到的电影链接是不完整的，需要使用字符串拼接对详情页面链接进行补全，然后再次对电影详细界面进行解析、索引定位、属性定位、文本获取，提取出需要的数据。将获取到的电影数据以.append（）形式循环插入一个空列表中。用户使用代码创建空表并写入表头，循环放入列表中的数据，保存为.xls格式，再对获取到的数据进行简单处理，使用matplotlib库绘制成饼状图形式，实现数据的可视化。电影数据整体爬取如下图1所示。

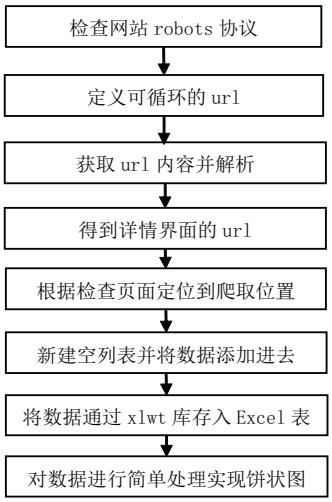

图1 电影数据获取流程图

4 数据保存及可视化

电影数据保存，可视化利用到了Python语言的xlwt库和matplotlib库。xlwt第三方库是一个用于生成与Microsoft Excel 95到2003年版本兼容的表文件的库，完全使用Python语言编写，与其他包或者模块没有依赖关系。matplotlib库是一个Python的2D绘图库，目的是实现数据可视化，更方便用户得到数据分析的结论。

在数据保存中，使用xlwt.Workbook（）函数创建工作簿，利用函数worksheet=workbook.add_sheet（）创建一个工作表，再通过worksheet.write（行，列，名称）为工作表写入表头，最后通过for循环将列表中的数据依此写入表中，使用workbook.save（）函数对其进行保存。具体Excel表表头内容如图2所示。

电影名	电影评分	主演	导演	类别	地区	上映年份	网站内更新时间	简介
我的姐姐	7.3	张子枫、肖央、朱媛媛	殷若昕	电影	中国大陆	2021	11.09	影片医

图2 电影表头图

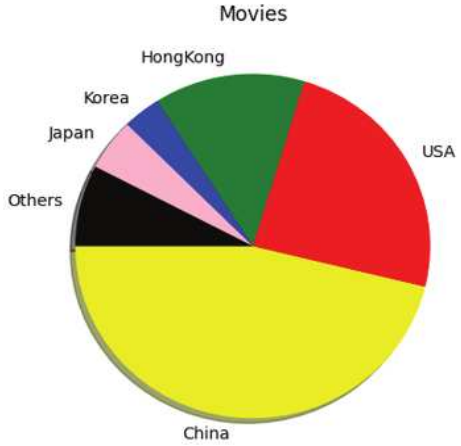

图3 电影数据可视化

为了体现数据的准确性和丰富性，爬取电影数据

ISSN: 2661-3719(Print); 2661-3727(Online)

209页，合计10032部电影，对电影出版国家进行了数据分析，提前定义全局变量，使用判断语句，当判断一致时，对应变量加一处理，将变量的百分比放入一个新的列表中，将名称也对应写入一个新的列表中，定义图片标题，字体和各个数据代表的颜色，使用plt.savefig（）函数对显示的图片进行本地化保存，最后plt.show（）将饼状图进行显示，如图3所示。

5 反爬虫机制

整个互联网上，有非常高额比例的流量其实是爬虫，例如，某公司某个页面的接口每分钟访问量是1.2万左右，但这里面有多少是正常用户呢？正确答案是500以下，也就是说，一个单独的网页界面，12000的访问量里，只有大概500是正常用户，其余是爬虫，注意在统计爬虫的时候，考虑到不可能识别出所有的爬虫，因此这500个用户里面，其实还隐藏着一些爬虫，爬虫率高达百分之九十以上[4]。过高的爬取，频繁的访问网页，会导致网站服务器崩毁，因此，反爬虫机制出现在了用户的视野中。

目前有三种常见的反爬机制和应对策略：①从用户请求的headers反爬虫，这是最常见的反爬虫机制，在访问某些网站的时候，网站通常会用判断访问是否带有头文件来鉴别该访问是否为爬虫，用来作为反爬取的一种策略，如果遇到了这类反爬虫机制，可以直接在爬虫中添加headers，也就是前文所说的UA伪装，将浏览器检查中的user-agent复制到代码headers中或者将referer值直接修改为爬取目标网站的域名。②基于用户行为反爬虫，被爬取的网站通过检测用户的行为，例如同一IP地址在短时间内多次访问同一个页面或者同一个账户短时间内多次进行类似操作，这种防爬，需要有做够多的IP来应对，有了大量代理IP后可以每请求几次更换一个IP就行或者可以在每次请求后随机间隔几秒再进行下一次请求，如果有多个账户切换使用，效果将会更好[5]。③动态页面的反爬虫，这种网站需要爬取的数据是通过Ajax请求得到或者通过Java生成的，比如获取淘宝的个人详情地址，解决方法就是使用selenium和phantomjs。

6 结束语

总的来说，在科技的不断发展下，数据显得尤为重要，网络爬虫获取数据占据了主要地位，但是由于反爬机制的不断完善，信息量，信息维度越来越多，爬取信息的难度也在逐渐提升。Xpath解析功能强大，使用方法简单、灵活，可以在知道数据爬取的HTML位置或者位置相对固定的地方使用，与xlwt库和matplotlib库搭配使用可以将获取到的数据进行本地化保存以及数据可视化。最后，在获取数据时，首先得遵循行业基本规则，获取到的数据切勿进行商业化。技术是无罪的，学习是可以的，实际操作就要适可而止，不要触碰到了红线。

参考文献：

[1]王康，史雅婷，梁洪炎，吉卓嘎，强巴卓玛.基于XPath的天气数据的爬取研究[J].江苏通信，2021，37（05）：83-84.

[2]卢江，刘文正.基于爬虫技术的图书购买推荐与比价策略研究[J].科技资讯，2021，19（01）：214-219.DOI：10.16661/j.cnki.1672-3791.2010-5042-8461.

[3]胡正雨.基于Python的网络爬虫技术研究[J].科技 风，2020（20）：102.DOI：10.19392/j.cnki.1671-7341.202020080.

[4]孙亚红.基于Python的招聘信息爬虫系统设计[J].软件，2020，41（10）：213-214+235.

[5]曾燕清，陈志德，李翔宇.应用树结构的Xpath自动提取算法[J].福建电脑，2020，36（07）：34-38.DOI：10.16707/j.cnki.fjpc.2020.07.008.