Python3 编程实现网络图片爬虫

涂辉，王锋，商庆伟
（徐州工业职业技术学院现代教育技术中心，江苏 徐州 221140）

摘　要：在大数据时代，网络数据的获取需要通过计算机自动实现，网络爬虫可实现对网页上的图片的抓取。Python 语言的易读、易学、可移植等优点使其成为人工智能大潮下最炙手可热的语言之一。通过 Python3 实现网络爬虫，并将获取到的图片自动存储到本地，为后续的机器学习、人工智能奠定了数据基础。

关键词：Python3 语言；网络爬虫；图片抓取
DOI:10.16184/j.cnki.comprg.2017.23.006

1　概述

近年来，随着信息技术的进步，互联网发展突飞猛进，中国已有接近 7 亿互联网用户，互联网也已深入到各地区各行业，爆炸式的数据增长使传统的依靠信息管理人员人工获取数据变得不可能。网络爬虫[1]是一种按照特定的规则，对网络信息自动抓取的程序或者脚本，也被称为网络机器人或网页蜘蛛。网络爬虫通过模仿浏览器对网页的 URL 地址访问的方式进行，用户不需要人工操纵即可自动地获取所需要的数据[2][3]。

Python 语言自 1991 年诞生，现已发展到 3.6.3 版本，因其功能强大、开源、语法简洁清晰，几乎在目前所有的操作系统上都能够运行[4]，具有丰富和强大的库被逐渐广泛应用于系统管理任务的处理和 Web 编程中，2017 年 7 月 20 日 IEEE 发布 2017 年编程语言排行榜[5]，Python 高居首位，超过 C 语言与 Java。

使用最新版本 Python 编写爬虫模拟浏览器访问目标页面获取目标图片数据，并将这些图片保存到本地文件夹，为进一步的图像数据挖掘与数据分析提供基础。使用爬虫程序能够让数据分析人员将更多的精力放在数据分析上面，而不是在程序开发的细节上消耗大量时间，同时爬虫还能够对海量数据起到过滤作用。

2　Python 爬虫的设计

爬虫是抓取糗事百科上的 JPG 及 GIF 格式趣图，方便离线观看。爬虫用的是 Python3.X 版本开发，主要用到了 urllib 的 request 和 re、os 模块，模块是一个包含变量、函数或类的定义的程序文件，正是 Python 大量的第三方库支持使得 Python 开发简单易学，使用模块前只需要通过 import 导入模块即可。

urllib 模块提供了从万维网中获取数据的高层接口[6]，当用 urlopen() 打开一个 URL 时，就相当于用 Python 内建的 open() 打开一个文件。但不同的是，前者接收一个 URL 作为参数，并且没有办法对打开的文件流进行 seek 操作，而后者接收的是一个本地文件名。

抓取到网页含有包括动画、图片、文档等各种格式元素。这些文件被爬虫抓取下来后，需要将其中的目标信息提取出来。正则表达式是一种在文本中寻找特定字符串的方法，能够准确地提取文档的特定信息。re 模块（regular expression）是 Python 中支持正则表达式的库。Pattern 实例是 re 处理文本并获得匹配结果的必须步骤，Re 库必须先将用户给定的正则表达式字符串编译为 Pattern 实例，Pattern 实例也被称为 Match 实例，它是程序获得信息并作其他操作的基础。

OS 模块是一个 Python 的系统编程的操作模块，提供了丰富的适用于 Mac、NT、或 Posix 的操作系统函数，这个模块允许程序独立地与操作系统环境、文件系统、用户数据库以及权限进行交互。

2.1　爬虫准备

```
    # 导入所需的 urllib、re、os 库
    import urllib.request, re, os
    # 定义抓取到的文件保存路径
My_targetPath = "F:\\python3.5 学习 \\01_spider\\qiubai"
    # 目标网址
    My_url = "http://www.qiushibaike.com/"
```

2.2　伪装成浏览器

对于一些网站，如果不是从浏览器发出的请求，则得不到响应。所以，需要将爬虫程序发出的请求伪装成浏览器。User Agent 是 Http 协议中的一部分，中文名叫

作者简介：涂辉（1987-），男，硕士，研究方向：数据挖掘。
收稿日期：2017-09-13

用户代理，属于头域的一部分。在访问网站时通过用户代理向服务器提供用户使用的操作系统及版本、浏览器版本及类型、浏览器的内核等信息标识。通过改写 User-Agent 将 Python 爬虫伪装成浏览器。

```
My_headers = {
        'User -Agent': 'Mozilla/5.0  (Windows NT 10.0; Trident/7.0; rv:11.0; JuziBrowser) like Gecko'
                'Chrome/51.0.2704.63 Safari/537.36'
    }
```

2.3 存储图片函数

通过之前导入的 os 模块操作图片的存储路径，文中主要用到 os.path.isdir 来判断指定对象目录是否存在，如果不是指定文件目录使用 os.mkdir 来创建目录，最后使用 os.path.join 将目录与图片的存储名称连接。

```
def saveFile(path):
    # 对保存路径有效性的识别
    if not os.path.isdir(My_targetPath):
        os.mkdir(My_targetPath)
    # 设置抓取到的图片的存储路径
    My_pos = path.rindex('/')
    t = os.path.join(My_targetPath,path[My_pos+1:])
    return t
```

2.4 爬虫主体函数

首先通过 urllib 的 request 和 urlopen 方法模拟浏览器访问目标页面获取网页数据，然后对数据进行适当的编码操作，其次通过空白符切割网页数据，使用正则表达式获取 jpg 和 gif 格式的文件，最后将获取的图片文件存储到本地指定的文件夹。

```
def spider_qiubai(url):
    My_req = urllib.request.Request (url= My_url, headers=My_headers)
    My_res = urllib.request.urlopen(My_req)
    My_data = My_res.read()
    data = My_data.decode('GBK')
# 本网页不适合 utf-8 编码,只能用 GBK 编码
    # 对母网页内容处理,
    k = re.split(r'\s+', data)
    s = []
    si = []
    for i in k:
 if (re.match(r'.*?jpg"', i) or re.match(r'.*?gif"', i)):
            s.append(i)
    # 获取这些图片链接的内容,并保存成本地图片
    for it in s:
        re_m1 = re.search(r'src="(.*?)"', it)
        My_iturl = re_m1.group(1)
        print(My_iturl)
        try:
urllib.request.urlretrieve(My_iturl, saveFile(My_iturl))
        except:
            print('失败')
```

2.5 多页面抓取

分析目标页面发现，多个连续的页面只是 url 的某个值不同，通过设置抓取网页的起始页与终止页的页码范围与 url 相同部分进行拼接，模拟浏览器的分页操作，实现对相应的多个页码的所有数据的抓取。

```
#main
if _name_ == '_main_':
    # 来判断本 py 程序是直接运行还是被引用
    sta_page=1# 网站的起始页
    end_page=500# 网站的终止页
    while sta_page<end_page:
        print("现在爬取的是第"+str(sta_page)+"页")
        My_url = url+str(sta_page)+'.html'
        spider_qiubai(My_url)
        sta_page += 1
```

3 结语

通过 Python3 编程实现了网络爬虫对指定网页的 jpg 和 gif 格式图片的抓取，通过实验对 914 个页面的 7312 张图片共计 3.41G 进行抓取，在 20 分钟之内完成所有图片保存到本地。实验表明本程序能有效实现相应爬虫功能。而且这种爬虫编程简洁明了，对于初学者有很好的指导作用，对于专注数据分析的研究者能够节省编码时间，将更多精力投入数据挖掘中。

参考文献

[1] 刘金红，陆余良. 主题网络爬虫研究综述 [J]. 计算机应用研究，2007，(10)：26-29+47.

[2] 李琳. 基于 Python 的网络爬虫系统的设计与实现 [J]. 信息通信，2017，(09)：26-27.

[3] 周德懋，李舟军. 高性能网络爬虫：研究综述 [J]. 计算机科学，2009，36 (08)：26-29+53.

[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，(09)：97-99.

[6] 王弘博，孙传庆. Python3 程序开发指南. 2 版. 北京：人民邮电出版社，2011.