基于 Python 的微博爬虫系统研究

陈政伊 袁云静 贺月锦 武瑞轩

（北华航天工业学院计算机与遥感信息技术学院，河北 廊坊 065000）

【摘 要】随着大数据时代到来，爬虫的需求呈爆炸式增长，以新浪微博为代表的一系列社交应用蕴含着巨大的数据资源。以新浪微博为研究对象，利用 Python 语言实现模拟登陆和网页解析技术，将获取的用户信息存为文档进行分析。文章分析了新浪微博模拟登陆时的加密方法，研究了验证码识别的实现方法，对挖掘的数据使用 TF-IDF 算法进行分析，提出了新的微博数据挖掘方向，论述了爬虫的国内外研究现状及开发难题。

【关键词】大数据；新浪微博；数据挖掘；Python 爬虫；模拟登陆

【中图分类号】TP393 【文献标识码】A 【文章编号】1008-1151(2017)08-0008-04

Python-based analysis of Microblog data mining

Abstract: With the coming of the age of Big Data,the need of Crawler growing explosively.The Social Network are influencing everyone's life,The Sina Microblog represented by a series of social application containing a large number of data resources.The Sina Microblog is the main study object,taking advantage of Python to realize the simulated landing and web page parsing,saving the downloaded data to analyze.The prode into the encode method during the simulated landing and how to realize the technology of recognizing the verified code,then making use of TF-IDF to further analyze the mined data.Putting forward the new direction of Sina Microblog data mining,discussing the research status of Crawler and the problems in the exploit process.

Key words: Big Data;Sina Microblog;data mining;Python crawler;simulated landing

1 引言

新浪微博自 2009 年 8 月进入公众视野，根据新浪发布的 2016 年财报，截止 2016 年底，微博月活跃人数已达 3 亿。微博已成为青年人生活的一部分，其中蕴含的巨大信息量的意义不言而喻。但是，与同类的国外社交网络社区如 Facebook，Twitter 等相比，新浪微博推出的供研究人员使用的数据接口尚不成熟，给数据分析工作带来了不小的压力。因此，许多技术成熟的科研团队自行开发爬虫系统来获取研究数据，同时，新浪出于安全考虑也在不断升级反爬技术。而爬虫技术难题之一就是反封锁，多数时候，有价值的信息一定采用了严格的反爬措施，比如验证码、防火墙、访问频率限制……。本文也将验证码作为一个重点探究对象，分析了新浪验证码识别的方法。

Python 作为一个语法简洁的程序设计语言，对于爬虫开发上有得天独厚的优势，在模拟浏览器行为登入网站时，Python 相比于 Java，C#，C++等拥有更简洁抓取接口，当模拟 session/cookie 的存储和设置时，Python 提供诸多优秀的第三方包譬如 Requests。在进行网页抓取后的处理工作时，Python 提供的 BeautifulSoup 库能用极简短的代码完成过滤 html 标签，提取文本的工作。

本文主要探究了爬虫的模拟登陆、识别验证码、内容解析、数据分析、数据整合等方面。

对于新浪微博最新的反爬虫设计进行了破解分析，同时探究了如何解析、分析获取的数据。

2 爬虫研究现状与难题

网络爬虫是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本，在机器学习和数据挖掘中，爬虫是最基础的一块内容，目前已经有各种完善的爬虫框架，例如 Crawler4j、WebMagic、WebCollector、Scrapy、Nutch 等。

爬虫面临的最重要的有三方面问题：

（1）法律道德风险：虽然爬虫抓取的各种网络信息都是公开的，但将这些信息商用甚至可能侵害源网站利益，既不合理也不合法。目前我国关于网络信息安全方面的法律法规几乎是空白。

（2）访问速度限制：爬虫的爬取速度主要依赖于服务器的出口宽带带宽和客户端的入口带宽，以及代码的质量，另外，大部分用户使用规模较大的网站都有反爬虫机制，其中最基本的就是频率限制，通常在频繁访问时间隔一段时间，甚至使用一套高效、可用的 IP 机制。

【收稿日期】2017-07-03

【作者简介】陈政伊，北华航天工业学院计算机与遥感信息技术学院学生，研究方向为数据挖掘，web 开发，机器学习。

（3）验证码：验证码的作用是用来区分机器人和人类。近年来验证码已经从简单的字母数字发展到拖动滑块甚至 Google 的 reCAPTCHA 这种以机器学习原理为基础开发的验证码模块，有时验证码的识别甚至比爬虫本身开发难度更大[4]。

3 爬虫开发与研究内容

3.1 模拟登陆

本项目通过研究开发一款爬虫，探究了爬虫的模拟登陆、识别验证码、内容解析、数据分析、数据整合等方面。

新浪微博的通行证入口登陆较之新浪 PC 端登陆入口具有代码简洁易读等优点。在通行证入口登入时，需要经历三个过程，如图 1：

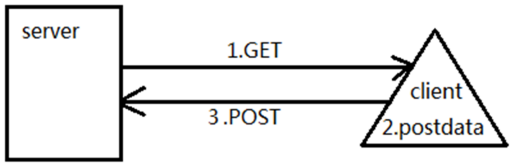

图 1 微博模拟登陆过程图

- 根据用户输入的用户名 username 经过 base64 算法加密得到参数 su
- 根据 su 得到一个 json 串，其中包含 rsakv，pubkey，servername，nonce，pcid 等参数
- 将密码进行 RSA 算法加密得到参数 sp 加入到 json 串中 POST 请求

3.1.1 用户名加密

观察 ssologin.js 页面，很容易找到关于用户名的编码方式。

“var username=config.username||"";

username=sinaSSOEncoder.base64.encode(urlencode(username));

delete config.username;

var arrQuery={

entry:me.entry,callback:me.name+".preloginCallBack",su:username,rsakt:"mod"};”

3.1.2 密码加密

密码加密的相关过程也在 ssologin.js 页面的源码，源码中可见当 loginType 为 rsa 时，客户端将密码经过 rsa2 算法进行编码，首先创建一个 rsa 公钥，公钥需要两个参数“pubkey”和“10001”，这里的“pubkey”是 GET 请求获得的，“10001”是 js 加密文件中的内容。绝大多数情况新浪登陆时使用此种加密方法。

“request.pwencode="rsa2";

request.rsakv=me.rsakv;

var RSAKey=new sinaSSOEncoder.RSAKey();

RSAKey.setPublic(me.rsaPubkey,"10001");

password=RSAKey.encrypt([me.servertime,me.nonce].join("\t")+"\n"+password)”

而当 loginType 为 wsse 时，客户端将密码经过 wsse 算法加密。

“request.pwencode="wsse";

password=sinaSSOEncoder.hex_sha1(""+sinaSSOEncoder.hex_sha1(sinaSSOEncoder.hex_sha1(password))+me.servertime+me.nonce);

request.sp=password;”

3.1.3 验证码识别

验证码识别是新浪爬虫开发的一个难点，也是新浪近两年新增的安全防御方式。在第一次客户端发送 GET 请求时，客户端同时发送一个用于请求验证码图片的 GET 请求，返回一个 pin.php 的 url 链接，是一个.png 或.image 格式的图片。使用 PyTesser 这个 python 的光学字符识别模块，这是一个结合 Tesseract OCR 引擎来使用的并且输出一个文本文件的验证码识别程序。

利用 split（）函数对 url 进行筛选得到验证码图片的 url，使用 requests 下载验证码图片，之后使用 pytesser 模块进行识别。获得 code 值连同 GET 获得的参数 pcid 加入 POSTDATA 数据字典。

图 2 模拟登陆成功图

3.2 网页下载器

模拟登陆成功后，找到需要爬取的微博博主用户名，根据分析博主的主页 url，如高了然的微博：“http://weibo.com/gaoliaoran?is_search=0&visible=0&is_tag=0&profile_ftype=1&page=1#feedtop”，马云的微博：“http://weibo.com/mayun?c=spr_qdhz_bd_360ss_weibo_mr&is_hot=1”等，可发现其中 userid 即为微博博主的用户名，提示用户输入用户名即 userid，爬虫即开始下载该网页源代码，最后将下载的源代码提供给网页解析器。

3.3 网页解析器

本项目中网页解析器采用的是 beautifulsoup 来提取数据，它可以使用 html 作为解析器也可以使用 lxml 作为解析器，所以功能相对强大。

首先根据已下载 html 文件定义一个 beautifulsoup 对象 soup，此时文件已由解析器解析成一个 html 节点树，原网页中审查元素了解所需段落在文件中的标签、属性等易于分辨的特征，然后通过 soup 对象即可很容易的获得这些节点标签，并得到其中的文字内容，本项目中该内容主要包括新浪微博用户的昵称，性别，以及他的微博原文等。

数据提取简单流程图描述如图 3 所示：

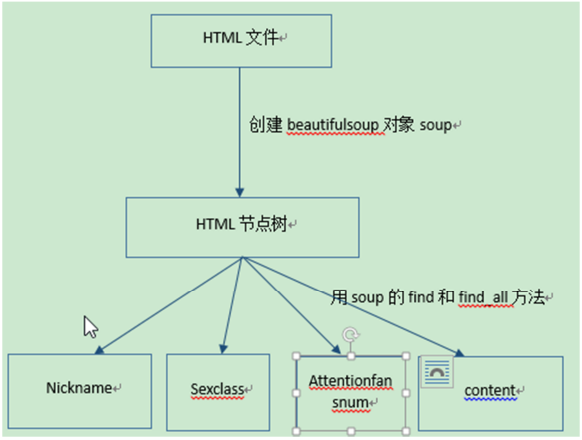

图 3　数据提取流程图

部分源码如下：

```
def _get_new_data(self,soup):
res_data = {}
 #获取昵称
nickname=soup.find('h1',class_='username').string
 #获取性别
sexclass=soup.find('span',class_='icon_bed').a.i['class']
if (sexclass==[u'W_icon', u'icon_pf_male']):
sex='M'
else :
sex='F'
#获取关注数
attentionsnum=soup.
        find('div',id='Pl_Core_T8CustomTriColumn__3').
        table.tr.contents[1].strong.string
#获取粉丝数
fansnum=soup.
        find('div',id='Pl_Core_T8CustomTriColumn__3').
        table.tr.contents[3].strong.string
#获取微博数
textsnum=soup.
        find('div',id='Pl_Core_T8CustomTriColumn__3').
        table.tr.contents[5].strong.string
#逐条获取微博内容
i=0
content=[]
whilei<int(textsnum) :
content.append(soup.find('div',id='plc_frame').
            find_all('div',class_="WB_text
W_f14")[i].get_text())
i=i+1
#将获得的内容放到字典中
res_data['nickname']=nickname
res_data['sex']=sex
res_data['attentionsnum']=attentionsnum
res_data['fansnum']=fansnum
res_data['textsnum']=textsnum
res_data['weibotext']=content
returnjson.dumps(res_data).decode("unicode-escape")
```

该部分运行结果如图 4：

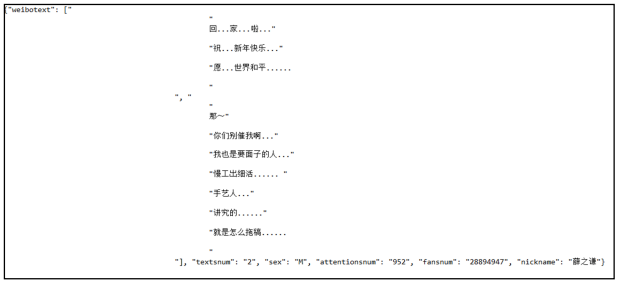

图 4　数据提取结果

存入文档后的形式如图5，数据提取成功后保存在磁盘并交由数据分析模块进行重要内容的分析和提炼。

图 5　数据提取文档图

3.4　数据分析

由于获取的数据是文本文件，可以使用 TF-IDF 算法提取关键词，此算法是基于“字词的重要性与其在文章中出现次数成正比，又会随着字词在语料库中使用次数成反比”这一理论的算法。以下为此算法的实现步骤：

（1）打开微博预处理文档，并读取一行字符串，赋给变量 text，同时用 text_num 表示字符串（文档）的数量；

（2）如果 text 不为空，将 text 拆分，并记录此文档为 text_name，创建列表 text_list1，存储拆分之后的词语，text_name_num 记录该文档的总词数，此时 text_num 加一；

（3）继续读取字符串，拆分出的词语 word（预先称为关键字），先判断此时 text_name 代表的文档的所有词语中是否有此词语 word，若有，判断 word 是否在其他文档里出现，若没有，在字典 word_all 增添此词，并赋键值（key）为一；否则，将此关键字的键值（key）加一；

（4）如果字典 word_all 没有此时 text_name 变量指向的文档名的元素，即当前文档包含已经遍历过了的关键字，则新建一个文件名为 text_name 变量指向的文档名，键值为一空字典；

（5）空字典 word_all2 里新建一列表，记录当前的关键字，及其数量与文档总词数量；

（6）重复以上 i-v 直到读完整个预处理文档 text,最后应得到两个字典 word_all 和 word_all2。在 word_all2 中记录了单篇文档中的关键字个数，以及总词数。 在 word_all 中，记录每个关键字在多少篇文档中出现。最终求出 TF-IDF 值，找到关键字。

3.5 数据应用

本文的爬虫有两个创新的应用方向：

3.5.1 青年人租房信息平台

近年来国内关于房产的社会问题日趋严重，国家努力采取一系列措施控制楼市，鼓励大学毕业生等年轻人以租房代替买房，但目前尚无较为独立的提供租房参考信息的平台，本文的爬虫可以以微博巨大的信息量为基础，爬取网民的日常微博言论，大数据整合出各城市各区域的治安，环境，人员复杂度等信息，在一个独立的平台的展现。

3.5.2 高校转型建设建议平台

国内高校数量众多，多数都在努力转型，从新浪网发布的 2016 年新浪微博财报数据来看，目前，新浪微博用户拥有大学及以上高等学历的用户占 77.8%，30 岁以下用户占 80%以上，显然，本文探究的爬虫爬取的多是与年轻人相关的数据，利用这些数据进行筛选整合搭建一个反映大学生真实诉求的平台，显然这些数据的真实性将远远超过一些调查问卷获得的数据，将这些内容反映给各大高校用于高校的转型建设。

4 结语

文章分析了新浪爬虫模拟登陆时的细节实现，由于新浪微博入口页的代码较为复杂，选择了新浪通行证入口进行登陆，然后分析了新浪用户名密码的加密方式，以及验证码识别的方法，使用 requests 下载网页的 html 代码，然后利用 BeautifulSoup 进行网页解析，研究了数据的分析方法 TF-IDF 算法，进行关键信息的提取，分析与利用。对于国内外爬虫开发现状以及面临的问题进行了论述。此外，提出了基于广大网民的原创微博的关于“租房”“高校建设”等应用的数据挖掘方向，将是笔者的研究方向之一。

【参考文献】

[1] 林晓丽,胡可可,胡青.基于 Python 的微博用户关系挖掘研究[J].情报杂志,2014,33(6):144-148.

[2] 张玉芳,彭时名,吕佳.基于文本分类 TFIDF 方法的改进与应用[J].计算机工程,2006,32(19):76-78.

[3] 赵世杰,陈秋.基于语义和 TF-IDF 的项目相似度计算方法[J].计算机时代,2015(5):1-5.

[4] Shih-Yu Huang,Yeuan-Kuen Lee,Graeme Bell,Zhan-he Ou,et al. An efficient segmentation algorithm for CAPTCHAs with line cluttering and character warping[J].Multimedia Tools and Applications,2009,48(2):267-289.

（上接第 5 页）

【参考文献】

[1] 程甫.大数据时代高校教育管理模式的创新研究[J].中国培训,2016(5):35-37.

[2] 姜楠,许维胜.基于数据挖掘技术的学生校园消费行为分析[J].大众科技,2015(1):22-23.

[3] 廖珣.基于 K-means 和 CBR 方法的高校就业预测模型应用研究[J].人力资源管理,2010,3(3):79-80.

[4] 叶炼.电信客户行为分析系统数据仓库的设计与实现[D].西安:西安电子科技大学,2009.

[5] 廖崇良.数据安全运维管理平台的建设[J].电子技术与软件工程,2017(2):213.

[6] 刘卫萍,王宁,周晓磊,等.数据融合技术在环境监测领域的应用[J].计算机系统应用,2016(6):88-93.

[7] 陈为,沈则潜.数据可视化[M].北京:电子工业出版社,2015:104,89,116,88.

[8] 阿不都克里木,高永强,迟忠先.数据库质量及其应用[J].计算机工程,2002,28(4):28-30.

[9] 郭崇慧,田凤占,靳晓明,译.Dunham M H.数据挖掘教程[M].北京:清华大学出版社,2005:79-85.

[10] 赵宝华,阮文惠.高校财务数据仓库的设计与实现[J].计算机工程.2008,34(17):266-268,273.

[11] 沈学利,钟华.决策树与数据挖掘结合的研究与应用[J].计算机工程,2011,37(11):89-91.

[12] 孙国强,韩强飞,陈俊.Kettle 在企业仓库建设中的应用与研究[J].信息系统工程,2017(2):28.

[13] 刘静萍.数字化校园建设中基于 ODI 的数据集成平台研究[J].青海师范大学学报(自然科学版),2016(2):16-20.

[14] FAYYAD U M,DJORGOVSKI S G,WEIR N.From Digitized Image to Online Catalogs Data Mining a Sky Survey[J].AI Magazine,1996,17(2):51-66.

[15] 雷芸,涂庆华,宋俊飞,等.大数据时代高校智慧校园服务平台建设与研究[J].通讯世界,2017(1):275-276.

[16] 黄成兵.大数据环境下高校智慧校园建设应用探讨[J].智能计算机与应用,2017,7(1):131-133.

[17] 汶向东.大数据时代“智慧校园”的应用研究[J].中山大学学报(社会科学版),2015,31(5):62-65.