基于Python的新浪微博爬虫研究

吴剑兰
(江苏警官学院,江苏 南京 210031)

摘 要: 对比新浪提供的API及传统的爬虫方式获取微博的优缺点,采用模拟登陆和网页解析技术,将获取的信息存入数据库中并进行分析。基于Python设计实现了新浪微博爬虫程序,可以根据指定的关键词获取相应的微博内容及用户信息。

关键词: 新浪微博; Python; 爬虫

0 引言

自2009年8月新浪推出微博业务以来,微博逐渐地进入人们的日常生活中。越来越多的人开始加入到社交网络中,与他人互动。继新浪之后,腾讯、网易等也相继推出微博业务,但新浪做为国内微博界的“元老”,仍是广泛受到人们的欢迎。如今,新浪微博用户已达5亿多人。[1]

随着使用人数的直线上升,带来的是信息量的急剧膨胀。每天都有数以万计的信息在奔流。微博通过点赞,转发,评论功能将个人的声音快速放大到社会空间,将个人的行为放大成为社会行为。作为网络新媒体的代表,微博用户产生的大量微博数据以及用户之间的互粉,转发等关系作为真实社会关系的一种写照,为社会网络研究提供了绝佳的研究数据。基于微博的数据研究已成为当今社会科学和计算机科学研究的重点。

1 新浪API

API接口使用较为方便,通过一个接口就可以很方便得获取所需的信息,而无须了解具体实现过程。但是新版的新浪API接口却有着很大的限制。最主要的一点,如果要想获得某人的微博个人信息和发表的微博内容,就必须得到对方的授权许可。

新浪API使用OAuth2.0授权机制。授权流程如图1所示。

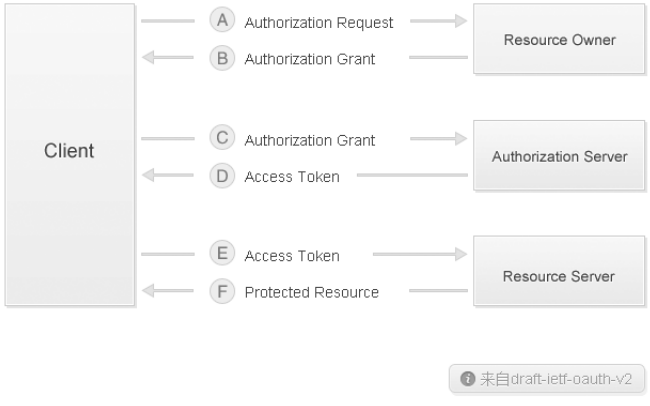

图1 OAuth2.0授权机制

其中Client指第三方应用,Resource Owner指用户,Authorization Server是我们的授权服务器,Resource Server是API服务器。

首先应用需要先引导用户到某个地址,用户授权后得到access_token,然后使用获取的access_token来调用API,以此来得到用户的信息和微博的内容。Access_token相当于是令牌,持有相应的令牌才能得到所需。除此以外,access_token还有授权有效期,对于测试应用来说只有一天的时间。

除了这些限制外,新浪API针对一个用户在使用一个应用的请求次数上还有限制。对于测试授权来说,单个用户每个应用每小时只能请求150次。这对于爬取微博信息来说是不够的。正因为有如此诸多的不便,尽管API的实现对开发者是透明的,但笔者仍决定采用传统的模拟登陆方法,然后通过分析网页源代码来获取信息。

2 模拟登陆

新浪微博的内容只有在登陆后才可以获取。通过firefox+httpfox分析网页版微博登陆方式可以发现主要分为三个步骤:(1)浏览器向新浪的服务器发送了一个GET请求报文,用于获取servertime,nounce字段,这两个字段是随机字段,每次登陆都不相同,用于加密用户名和密码;(2)用BASE64算法加密用户名,用RSA算法加密密码,向登陆URL发送包含加密后的用户名和密码的POST请求;(3)新浪服务器收到请求后与信息库进行比对,如果比对成功则发送一条含重定向的应答报文,浏览器解析得到最终跳转到的URL,打开该URL后,自动将该信息写入COOKIES,登陆成功。

3 网页分析

以新浪官方的搜索平台为搜索入口,输入关键词后,构造相应url。分析网页源代码,可以发现页面上的所有微博内容都在以<script>STK && STK.pageletM && STK.pageletM.view({"pid":"pl_weibo_direct",开头的行中。源代码中大多为反斜线(“\”),而中文则以UTF-8的格式表示,即“\uXXXX”(X为数字或字母),一眼望上去很凌乱。但是如果查看经过处理后的源代码,可以发现每条微博都有相似的格式,而且是以一种“树”的形式展现的。

由此可见,存储在服务器上的源文件是“原生”的,而用户在浏览时,后台通过Javascript程序处理这部分代码,将其生成xml格式代码,进而交给浏览器去解析。

Python的第三方库BeautifulSoup/lxml是Python的html/xml解析器,可以很好地处理不规范的标记并生成剖析树。Lxml库支持XPath规范,XPath是一种在xml文档中查找信

作者简介: 吴剑兰(1993-),男,江苏苏州人,本科。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

息的语言，用于在xml文档中通过元素的属性进行导航，利用XPath可以方便在html文档中定位感兴趣的节点。[2]

仔细观察可以发现，每条微博都以<div class='WB_cardwrap …>作为起始，而其中的<a class="W_texta W-fb"…>节点含有昵称，<p class="comment_txt">节点含有微博内容，以此类推可以得到时间，转发数，评论数等信息。

对于获取微博用户个人的信息，也是使用与此相类似的方法。通过分析用户个人主页的源代码，可以得到UID，和Page-id。Page-id用于构建指向用户个人信息的URL地址，其格式为：'http://weibo.com/p/'+page-id+'/info'。此即为要进行分析的URL地址。

对于获取用户发表的微博这块，有一个难点。在使用浏览器浏览用户发表的微博时，一开始不会将一页上的所有微博都显示出来，而是当滚动到底部时自动加载，如此滚动加载两次才能把一页上的微博都显示出来。获取得到的网页源代码同样也是不完整的，只含有每页的前十条左右，必须进行手动滚动才能显示完整。因此，可以采用发送HTTP请求的GET方法，构建相应的URL来模拟这一滚动过程。

4 关键词的提取

这个爬虫程序还有一个可以对爬取到的微博内容进行分析，提取关键词的功能。使用TF-IDF算法来实现。TF-IDF算法的思想如下：为了提取关键词，一个容易想到的思路就是找到出现次数最多的词。如果某个词很重要，它应该在其中多次出现，于是，进行“词频”（TF）统计。但是，出现次数最多的词是“的”“是”“在”这一类词，这些词叫做“停用词”，对结果没有帮助，需要过滤掉。

过滤掉之后，可能会有多个词出现的次数一样多，但这并不意味着这些词的关键性是一样的。因此，还需要一个重要性调整系数来衡量一个词是不是常见词。如果某个词比较少见，但是它在其中多次出现，那么它就可能就是我们需要的关键词。用统计学语言表达，就是在词频的基础上，要对每个词分配一个"重要性"权重。最常见的词给予最小的权重，较常见的词给予较小的权重，较少见的词给予较大的权重。这个权重叫做"逆文档频率"（IDF），它的大小与一个词的常见程度成反比。知道了"词频"（TF）和"逆文档频率"（IDF）以后，将这两个值相乘，就得到了一个词的TF-IDF值。某个词对文章的重要性越高，它的TF-IDF值就越大。所以，排在最前面的几个词，就是关键词。

根据这一算法思想，爬虫程序可以根据爬取的一系列微博条目，获得这些条目的关键词。[3]

5 结语

文章分析了新浪API的一些认证限制，新版的API需要被搜索用户提供相应的授权，因此采用传统爬虫的方式。然后模拟登陆、网页分析、关键词提取等三个方面介绍了如何爬取新浪微博信息，研究用户登陆微博的过程，从网页源代码中构造利于分析的DOM树并提取所需信息，运用TF-IDF算法获取微博集中的关键词，最终实现了一个基于Python的新浪微博爬虫程序。

[参考文献]

[1]郭晓云.基于Python和Selenium的新浪微博数据访问[J].电脑编程技巧与维护，2012.
[2]齐鹏，李隐峰，宋玉伟.基于Python的Web数据采集技术[J].电子科技，2012.
[3]阮一峰.TF-IDF与余弦相似性的应用[EB/OL].(2013-03-15).http://www.ruanyifeng.com/blog/2013/03/tf-idf.html.

Sina Micro-blog Crawler Based on Python

WU Jianlan
（Jiangsu Police Institute，Nanjing 210031，China）

Abstract：The advantages and disadvantages of obtaining micro-blog contrast Sina provides API and traditional crawler style，using simulated landing and Webpage analysis technology，the information stored in the database and analysis. The design and implementation of Python based on the Sina micro-blog crawler，can obtain micro-blog content and user information corresponding to the specified keyword.

Key words：Sina micro-blog；Python；Crawler

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net