Stata 、cURL 交互与网络爬虫

小花经济学术

彭文威

MPhil in Social Science
Division of Social Science, HKUST

温州

2017 年 8 月 19-20 日

目录

1. 从爬虫说起
2. cURL 是什么?
3. cURL 爬虫
   - cURL 与微博 API
   - 实战之环保部空气质量日数据
   - 实战之中国土地网
4. cURL has limits

Table of Contents

网络爬虫

爬虫，是网络时代的产物，互联网上数据量呈现指数级增长。大量有用的数据可能是非结构化的，如果用人工收集会耗费大量的人力和时间。而程序工具的出现，为实现这种工作量巨大、内容繁琐重复的工作提供了解决办法。最初始的也是最常见的爬虫就是搜索引擎（google、baidu、bing...）
爬虫一般可以分为

- API 等规则化数据收集：这种爬虫非常简单，利用网站提供的 API 接口，可返回干净、结构化的数据。缺点在于，你只能获得他想让你获得的数据，并且出于保护自身利益和服务器成本的原因，API 往往有返回数据量限制。
- 非结构化数据收集（占比最大，难度高）。

Http 协议

超文本传输协议（英文：HyperText Transfer Protocol，缩写：HTTP）是一种用于分布式、协作式和超媒体信息系统的应用层协议。HTTP 是万维网的数据通信的基础。

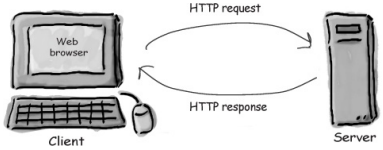


Figure: HTTP 协议

(Photo downloaded from: http://www.ruanyifeng.com/blog/2016/08/http.html)

HTTP 是一个客户端终端（用户）和服务器端（网站）请求和应答的标准（TCP）。通过使用网页浏览器、网络爬虫或者其它的工具，客户端发起一个 HTTP 请求到服务器上指定端口（默认端口为 80）。我们称这个客户端为用户代理程序（user agent）。应答的服务器上存储着一些资源，比如 HTML 文件和图像。我们称这个应答服务器为源服务器（origin server）。

HTTP 请求方法

HTTP/1.1 协议中共定义了八种方法（也叫“动作”）来以不同方式操作指定的资源：

- **GET**：向指定的资源发出“显示”请求。使用 **GET** 方法应该只用在读取数据，而不应当被用于产生“副作用”的操作中，例如在 **Web Application** 中。其中一个原因是 **GET** 可能会被网络爬虫等随意访问。
- **POST**：向指定资源提交数据，请求服务器进行处理（例如提交表单或者上传文件）。数据被包含在请求本文中。这个请求可能会创建新的资源或修改现有资源，或二者皆有。
- **HEAD**：与 **GET** 方法一样，都是向服务器发出指定资源的请求。只不过服务器将不传回资源的本文部分。它的好处在于，使用这个方法可以在不必传输全部内容的情况下，就可以获取其中“关于该资源的信息”（元信息或称元数据）
- **PUT**：向指定资源位置上传其最新内容。
- **DELETE**：请求服务器删除 **Request-URI** 所标识的资源
- **TRACE**：回显服务器收到的请求，主要用于测试或诊断。
- **.....**

Stata 爬虫

- GET
  - 当使用 GET 方法请求返回的数据是 Dcoument/TEXT 时，Stata 的 copy 命令可以轻松解决。
- POST
  - 当请求方法为 POST 时，copy 无能为力，一个较为简单的解决办法就是 Stata 与 cURL 交互。

Table of Contents

cURL

- cURL 是一个利用 URL 语法在命令行下工作的文件传输工具，1997 年首次发行。它支持文件上传和下载，所以是综合传输工具，但按传统，习惯称 cURL 为下载工具。cURL 还包含了用于程序开发的 libcurl。
- cURL 支持的通信协议有 FTP、FTPS、HTTP、HTTPS、TFTP、SFTP、Gopher、SCP、Telnet、DICT、FILE、LDAP、LDAPS、IMAP、POP3、SMTP 和 RTSP。

cURL 语法

- 下载单个文件，默认将输出打印到标准输出中 (STDOUT) 中：curl www.stata.com

- curl www.stata.com -o stata.txt

- 了解更多：curl –help

Table of Contents

1 从爬虫说起

2 cURL 是什么?

3 cURL 爬虫

cURL 与微博 API

实战之环保部空气质量日数据

实战之中国土地网

4 cURL has limits

cURL 与微博 API

为何以微博为例？API 端口，参数尽知。而在实战的爬虫中，参数获取需要分析、经验与琢磨。

原理：

- 使用新浪微博接口 update. 授权方式:Basic Authentication
- 如果你真的想测试下这个效果，那么你需要去新浪开放平台（http://open.weibo.com/）创建一个应用. 并取得 APP Key（我们需要一个 APP key 作为传送的参数)。你创建的应用完全不必去审核 (未审核的应用允许最多 15 个测试账号访问接口)
- 使用的接口地址：https://api.weibo.com/2/statuses/update.json

微博 API 文档查看 request method

微博 API 文档查看 request method

微博 API 文档查看 POST 参数

open.weibo.com/wiki/2/statuses/update

微博·开放平台 | 微连接 | 微服务 | 文档 | 支持 | 推广 | 我的应用 | Wenwei_P...

- 微博登录
- 移动应用
- 网站接入
- 微博API
- 资源下载
- 联系我们
- 常见问题

访问授权限制

访问级别：**普通接口**

频次限制：**是**

关于频次限制，参见 接口访问权限说明

请求参数

	必选	类型及范围	说明
access_token	true	string	采用OAuth授权方式为必填参数，OAuth授权后获得。
status	true	string	要发布的微博文本内容，必须做URLencode，内容不超过140个汉字。
visible	false	int	微博的可见性，0：所有人能看，1：仅自己可见，2：密友可见，3：指定分组可见，默认为0。
list_id	false	string	微博的保护投递指定分组ID，只有当visible参数为3时生效且必选。
lat	false	float	纬度，有效范围：-90.0到+90.0，+表示北纬，默认为0.0。
long	false	float	经度，有效范围：-180.0到+180.0，+表示东经，默认为0.0。
annotations	false	string	元数据，主要是为了方便第三方应用记录一些适合于自己使用的信息，每条微博可以包含一个或者多个元数据，必须以json字串的形式提交，字串长度不超过512个字符，具体内容可以自定。
rip	false	string	开发者上报的操作用户真实IP，形如：211.156.0.1。

注意事项

评价此API

stata 透过 cURL 发微博

- 使用的接口参数：access_token ：授权登录获得的参数；status：发布的微博内容。（改版之前参数为: source = APP KEY;status= 发布的微博内容）
- 使用的 curl 的参数:-u 发送登陆请求格式是 -u username:password ;-d 发送 post 数据
- curl -u 微博账号: 微博密码 -d “source= 应用的 APP key&status=Test” https://api.weibo.com/2/statuses/update.json
- Stata : !curl -u 微博账号: 微博密码 -d “source= 应用的 APP key&status=Test ” https://api.weibo.com/2/statuses/update.json

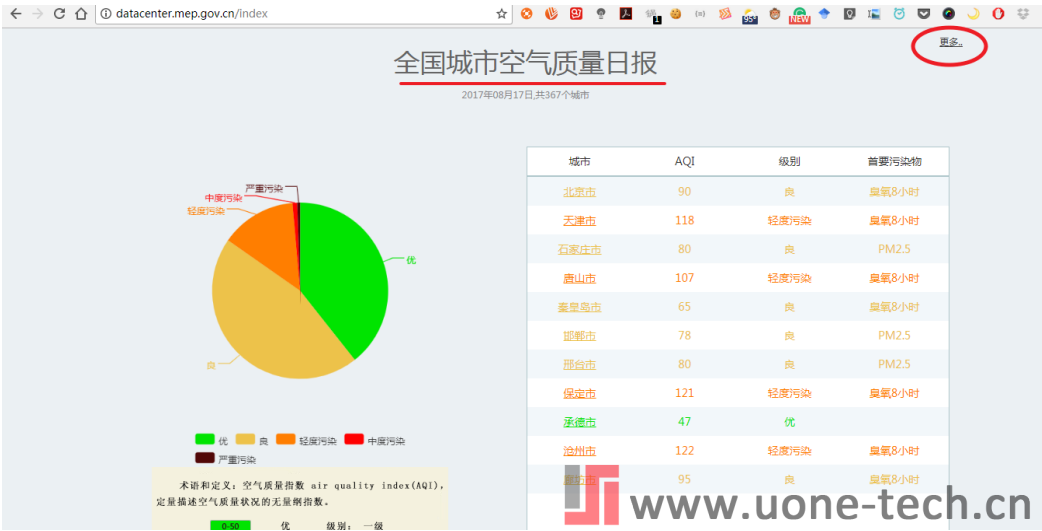
stata 透过 cURL 转发微博

- 参数: source:APP key; id：你想转发的微博 ID; status : 转发评论语。
- Stata code: !curl -u 18813795596:123456 -d "source=3186431957&id=3975371793325993&status= 期待各位优秀的大三学生申请我们的首届夏令营！" https://api.weibo.com/2/statuses/repost.json

实战之环保部空气质量日数据

实战之环保部空气质量日数据

空气质量日数据



城市	AQI	级别	首要污染物
北京市	90	良	臭氧8小时
天津市	118	轻度污染	臭氧8小时
石家庄市	80	良	PM2.5
唐山市	107	轻度污染	臭氧8小时
秦皇岛市	65	良	臭氧8小时
邯郸市	78	良	PM2.5
邢台市	80	良	PM2.5
保定市	121	轻度污染	臭氧8小时
承德市	47	优	
沧州市	122	轻度污染	臭氧8小时
廊坊市	95	良	臭氧8小时

寻找 POST 参数

城市 开始时间 2017-08-17 查询

截止时间 2017-08-17

序号	城市	AQI指数	首要污染物	日期	空气质量级别
1	北京市	90	臭氧8小时	2017-08-17	良
2	天津市	118	臭氧8小时	2017-08-17	轻度污染
3	石家庄市	80	PM2.5	2017-08-17	良
4	唐山市	107	臭氧8小时	2017-08-17	轻度污染
5	秦皇岛市	65	臭氧8小时	2017-08-17	良
6	邯郸市	78	PM2.5	2017-08-17	良
7	邢台市	80	PM2.5	2017-08-17	良
8	保定市	121	臭氧8小时	2017-08-17	轻度污染
9	承德市	47		2017-08-17	优
10	沧州市	122	臭氧8小时	2017-08-17	轻度污染
11	廊坊市	95	臭氧8小时	2017-08-17	良
12	衡水市	68	臭氧8小时	2017-08-17	良
13	张家口市	85	臭氧8小时	2017-08-17	良
14	太原市	117	臭氧8小时	2017-08-17	轻度污染
15	大同市	95	臭氧8小时	2017-08-17	良
16	阳泉市	87	PM2.5	2017-08-17	良
17	长治市	113	臭氧8小时	2017-08-17	轻度污染
18	晋城市	136	臭氧8小时	2017-08-17	轻度污染
19	朔州市	102	臭氧8小时	2017-08-17	轻度污染
20	晋中市	111	臭氧8小时	2017-08-17	轻度污染

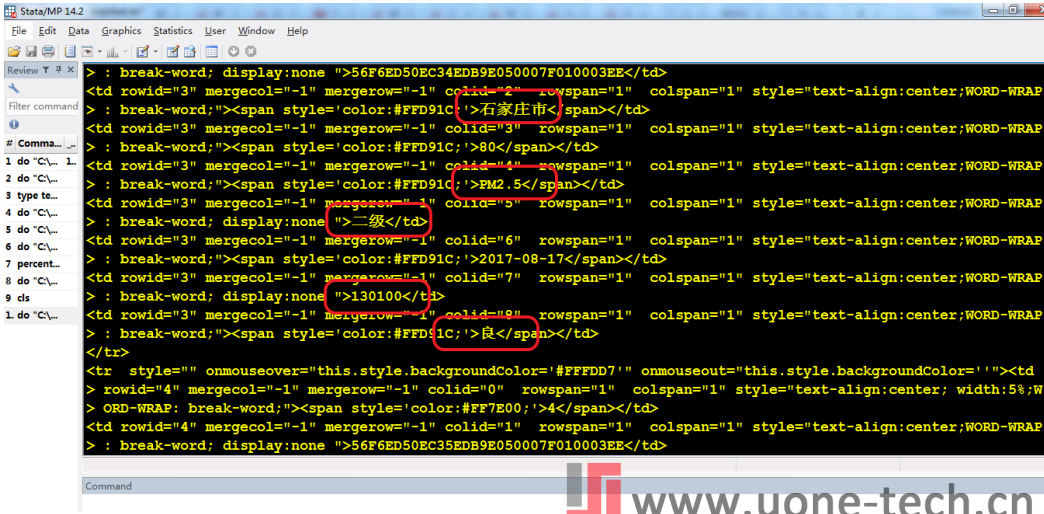
寻找 POST 参数

curl 模拟浏览器提交 POST 参数

```
global Begin = "2017-08-01"
global End = "2017-08-19"

# delimit ;
!curl "http://datacenter.mep.gov.cn:8099/ths-report/report%21list.action"
-H "User-Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, lik
Gecko) Chrome/59.0.3071.115 Safari/537.36"
-H "Content-Type: application/x-www-form-urlencoded" -H "Cache-Control: no-cache"
-d "page.pageNo=1&xmlname=1462259560614&queryflag=open&isdesignpatterns=
false&V_DATE=$Begin&_DATE=$End
--compressed -o temp.txt
;
#delimit cr
cls
type temp.txt
```

curl 下载网页源代码



清洗源代码

最后一步，利用 Stata 读入数据、清洗数据的强大功能对网页源代码进行整理。

```
. list ,clean
```

	Grade	Rownum	ID	AQI	CITY	STATUS	MAIN_PO~T	CITYCODE	OPER_DATE
1.	二级	1	56F6ED50EC32EDB9E050007F010003EE	90	北京市	良	臭氧8小时	110000	2017-08-17
2.	三级	2	56F6ED50EC33EDB9E050007F010003EE	118	天津市	轻度污染	臭氧8小时	120000	2017-08-17
3.	二级	3	56F6ED50EC34EDB9E050007F010003EE	80	石家庄市	良	PM2.5	130100	2017-08-17
4.	三级	4	56F6ED50EC35EDB9E050007F010003EE	107	唐山市	轻度污染	臭氧8小时	130200	2017-08-17
5.	二级	5	56F6ED50EC36EDB9E050007F010003EE	65	秦皇岛市	良	臭氧8小时	130300	2017-08-17
6.	二级	6	56F6ED50EC37EDB9E050007F010003EE	78	邯郸市	良	PM2.5	130400	2017-08-17
7.	二级	7	56F6ED50EC38EDB9E050007F010003EE	80	邢台市	良	PM2.5	130500	2017-08-17
8.	三级	8	56F6ED50EC39EDB9E050007F010003EE	121	保定市	轻度污染	臭氧8小时	130600	2017-08-17
9.	一级	9	56F6ED50EC3AEDB9E050007F010003EE	47	承德市	优		130800	2017-08-17
10.	三级	10	56F6ED50EC3BEDB9E050007F010003EE	122	沧州市	轻度污染	臭氧8小时	130900	2017-08-17
11.	二级	11	56F6ED50EC3CEDB9E050007F010003EE	95	廊坊市	良	臭氧8小时	131000	2017-08-17
12.	二级	12	56F6ED50EC3DEDB9E050007F010003EE	68	衡水市	良	臭氧8小时	131100	2017-08-17
13.	二级	13	56F6ED50EC3EEDB9E050007F010003EE	85	张家口市	良	臭氧8小时	131200	2017-08-17
14.	三级	14	56F6ED50EC3FEDB9E050007F010003EE	117	太原市	轻度污染	臭氧8小时	140100	2017-08-17
15.	二级	15	56F6ED50EC40EDB9E050007F010003EE	95	大同市	良	臭氧8小时	140200	2017-08-17
16.	二级	16	56F6ED50EC41EDB9E050007F010003EE	87	阳泉市	良	PM2.5	140300	2017-08-17
17.	三级	17	56F6ED50EC42EDB9E050007F010003EE	113	长治市	轻度污染	臭氧8小时	140400	2017-08-17
18.	三级	18	56F6ED50EC43EDB9E050007F010003EE	136	晋城市	轻度污染	臭氧8小时	140500	2017-08-17
19.	三级	19	56F6ED50EC44EDB9E050007F010003EE	102	朔州市	轻度污染	臭氧8小时	140600	2017-08-17
20.	三级	20	56F6ED50EC45EDB9E050007F010003EE	111	晋中市	轻度污染	臭氧8小时	140700	2017-08-17
21.	三级	21	56F6ED50EC46EDB9E050007F010003EE	139	运城市	轻度污染	臭氧8小时	140800	2017-08-17
22.	三级	22	56F6ED50EC47EDB9E050007F010003EE	110	忻州市	轻度污染	臭氧8小时	140900	2017-08-17
23.	三级	23	56F6ED50EC48EDB9E050007F010003EE	126	临汾市	轻度污染	臭氧8小时	141000	2017-08-17
24.	三级	24	56F6ED50EC49EDB9E050007F010003EE	122	吕梁市	轻度污染	臭氧8小时	141100	2017-08-17
25.	二级	25	56F6ED50EC4AEDB9E050007F010003EE	95	呼和浩特市	良	臭氧8小时	150100	2017-08-17
26.	二级	26	56F6ED50EC4BEDB9E050007F010003EE	75	包头市	良	臭氧8小时	150200	2017-08-17
27.	二级	27	56F6ED50EC4CEDB9E050007F010003EE	93	乌海市	良	臭氧8小时	150300	2017-08-17
28.	一级	28	56F6ED50EC4DEDB9E050007F010003EE	28	赤峰市	优		150400	2017-08-17
29.	一级	29	56F6ED50EC4EEDB9E050007F010003EE	37	通辽市	优		150500	2017-08-17

实战之中国土地网

中国土地网

序号	行政区	土地坐落	总面积	土地用途	供应方式	签定日期
1.	朝阳区	朝阳区大屯路36号	0.435872	科教用地	划拨	2017/8/10
2.	密云县	密云生态商务区A1地块	1.390900	街巷用地	划拨	2017/8/9
3.	房山区	房山区长阳镇	0.640000	科教用地	划拨	2017/8/8
4.	北京市本级	北京市丰台区长辛店镇辛庄村（一...	10.212050	其他普通商品住房用地	挂牌出让	2017/8/7

找到 POST 参数

<input id="TAB_QuerySubmitConditionData" name="TAB_QuerySubmitConditionData" type="hidden" value="9f2c3acd-0256-4da2-a659-6949c4671a2a:2017-8-1~2017-8-16:42ad98ae-c46a-40aa-aacc-c0884036eeaf:11▩~北京市" /><input id="TAB_QuerySubmitOrderData" name="TAB_QuerySubmitOrderData" type="hidden" value="282:False" /><script type="text/javascript">QueryAction.formId='mainForm';</script></TD>

Figure: 中国土地网源代码

(Photo cropped from: http://www.landchina.com/default.aspx?tabid=263&ComName=default)

Download

```
# delimit ;
!curl "http://www.landchina.com/default.aspx?tabid=263&ComName=default"
-H "User-Agent: Opera/9.80 (Android 2.3.4; Linux; Opera Mobi/build-1107180945; U; en-GB) Presto/2.8.149 Ver
-H "Content-Type: application/x-www-form-urlencoded"
-H "Cache-Control: no-cache"
-d "__EVENTVALIDATION=/wEWAgKg+ojgCwLN3cj/BEjpHh2IVeSnH7YAAyO/TP/eDYTr/Sg8kIkYrlhhTyNZ
&hidComName=default
&TAB_QueryConditionItem=9f2c3acd-0256-4da2-a659-6949c4671a2a
&TAB_QueryConditionItem=42ad98ae-c46a-40aa-aacc-c0884036eeaf
&TAB_QuerySortItemList=282:False
&TAB_QuerySubmitConditionData=9f2c3acd-0256-4da2-a659-6949c4671a2a:

2017-8-1~2017-8-16|42ad98ae-c46a-40aa-aacc-c0884036eeaf 11~ 北京市

&TAB_QuerySubmitOrderData=282:False

&TAB_QuerySubmitPagerData=1"

-x 123.103.93.38:80  -o temp.txt
;
#delimit cr
```

爬虫最后一步：数据清洗

Stata 出色的数据管理能力使我们在使用 copy 或 cURL 下载网络源代码以后，能轻松的整理成格式化的数据。

```
cap unicode analyze "temp.txt"
cap unicode encoding set gb18030
cap unicode translate "temp.txt", transutf8
cap unicode erasebackups, badidea
qui set obs 1
qui gen v = fileread("temp.txt")
qui split v,p(`"gridTdNumber"')
qui drop v v1
qui sxpose ,clear
qui split _var1,p(`"</td><td class="queryCellBordy">"')
qui gen num = ustrregexs(1) if ustrregexm(_var11,`"^">([0-9]+)\."')
qui gen city = ustrregexs(1) if ustrregexm(_var12,`"<span title="(.+)">"')
qui replace _var12 = city if city~=""
qui drop city
qui ren _var12 city
qui gen url = ustrregexs(1) if ustrregexm(_var13,`"href="(.+)""')
qui gen address = ustrregexs(1) if ustrregexm(_var13,`"target="_blank">(.+)</a>$"')
qui gen date = ustrregexs(1) if ustrregexm(_var17,`"^([0-9]+/[0-9]+/[0-9]+)</td>"')
```

爬虫最后一步：数据清洗

```
. list num city area type way address date url in 1

  1.  | num      city       area       type       way       address            date
      |   1    朝阳区   0.435872   科教用地     划拨    朝阳区大屯路36号     2017/8/10
      |
      | url
      | default.aspx?tabid=386&comname=default&wmguid=75c72564-ffd9-426a-954b-8ac2df0903b7&recorderguid=350bd..

. list num city area type way address date url in 2

  2.  | num      city       area       type       way       address                 date
      |   2    密云县   1.390900   街巷用地     划拨    密云生态商务区A1地块      2017/8/9
      |
      | url
      | default.aspx?tabid=386&comname=default&wmguid=75c72564-ffd9-426a-954b-8ac2df0903b7&recorderguid=14961..
```

Figure: 北京市土地结果公告

cURL 的本领

- POST
- Headers
- Cookies
- 上传、下载、重定向....

Table of Contents

cURL has its' limits

cURL 不知道何为 Javascript

cURL 无法解决验证码

Thank you!