



R七种武器之网络爬虫RCurl 第1周

DATAGURU专业数据分析社区

【声明】 本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

<http://edu.dataguru.cn>

关注炼数成金企业微信



- ◆ 提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！



- ◆ 数据抓取
- ◆ 数据加工
- ◆ 数据可视化



◆ 数据抓取

----- ?

◆ 数据加工

-----金融数据分析quantmod、数据加工厂plyr

◆ 数据可视化

-----数据可视化包ggplot2、交互化展示包shiny

巧妇难为无米之炊！

◆ 如何获取数据呢？

- 手动输入？
- 别人给予？
-

◆ 不若自己去获取！



◆ Duncan Temple Lang

现任加州大学 U.C. Davis分校副教授

致力于借助统计整合进行信息技术的探索

其omega团队已开发程序包达50余个

包括RCurl、XML、RSPython、Rmatlab等



作者个人主页：<http://anson.ucdavis.edu/~duncan/>

- ◆ The RCurl package is an R-interface to the [libcurl](#) library that provides HTTP facilities. This allows us to download files from Web servers, post forms, use HTTPS (the secure HTTP), use persistent connections, upload files, use binary content, handle redirects, password authentication, etc.
- ◆ RCurl这个程序包提供了由R到libcurl库的接口，从而实现HTTP的一些功能。例如，从服务器下载文件、保持连接、上传文件、采用二进制格式读取、句柄重定向、密码认证等等。

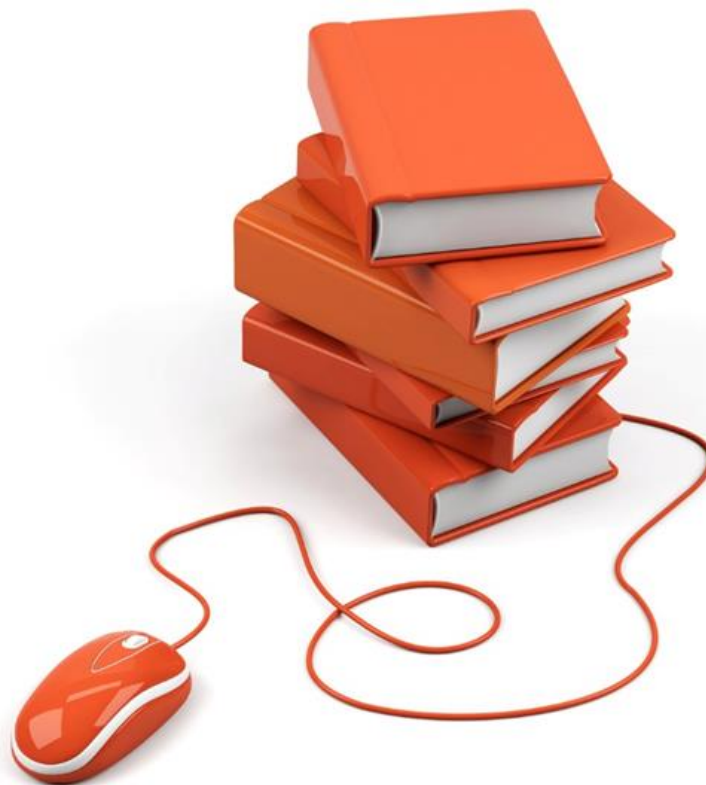
◆ 什么是curl&libcurl

- curl：利用URL语法在命令行方式下工作的开源文件传输工具
- curl背后的库就是libcurl

◆ 功能

- 获得页面
- 有关认证
- 上传下载
- 信息搜索
-

- ◆ HTTP 协议
- ◆ 正则表达式



- ◆ 协议是指计算机通信网络中两台计算机之间进行通信所必须共同遵守的规定或规则，超文本传输协议(HTTP)是一种通信协议，它允许将超文本标记语言(HTML)文档从Web服务器传送到客户端的浏览器
- ◆ 目前我们使用的是HTTP/1.1 版本



◆ 基本格式：schema://host[:port#]/path/.../[?query-string][#anchor]

scheme 指定低层使用的协议(例如：http, https, ftp)

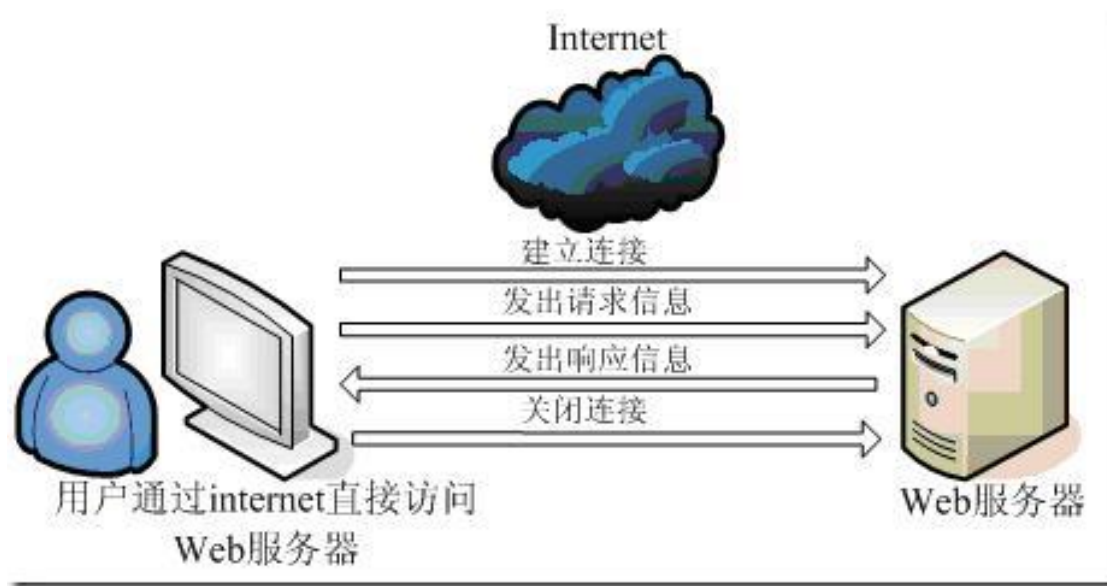
host HTTP服务器的IP地址或者域名

port# HTTP服务器的默认端口是80，这种情况下端口号可以省略。

path 访问资源的路径

query-string 发送给http服务器的数据

anchor- 锚



- ◆ 1. 提高访问速度，大多数的代理服务器都有缓存功能。
- ◆ 2. 突破限制，也就是翻墙了
- ◆ 3. 隐藏身份



◆ 请求行、请求报头、消息正文

METHOD /path - to - resource HTTP/Version-number
Header-Name-1: value
Header-Name-2: value
Optional request body

Method表示请求方法,比如 “GET” , “POST” , “HEAD” , “PUT” 等

Path-to-resource表示请求的资源

Http/version-number 表示HTTP协议的版本号

- ◆ Host
- ◆ Accept
- ◆ Accept-encoding
- ◆ Accept-language
- ◆ User-agent
- ◆ Cookie
- ◆ Referer
- ◆ Connection



◆ 状态行、消息报头、响应正文

Http/version-number	status code	message
Header-Name-1: value		
Header-Name-2: value		
Optional	Response body	

HTTP/version-number表示HTTP协议的版本号

status-code 和message表示状态码以及状态信息

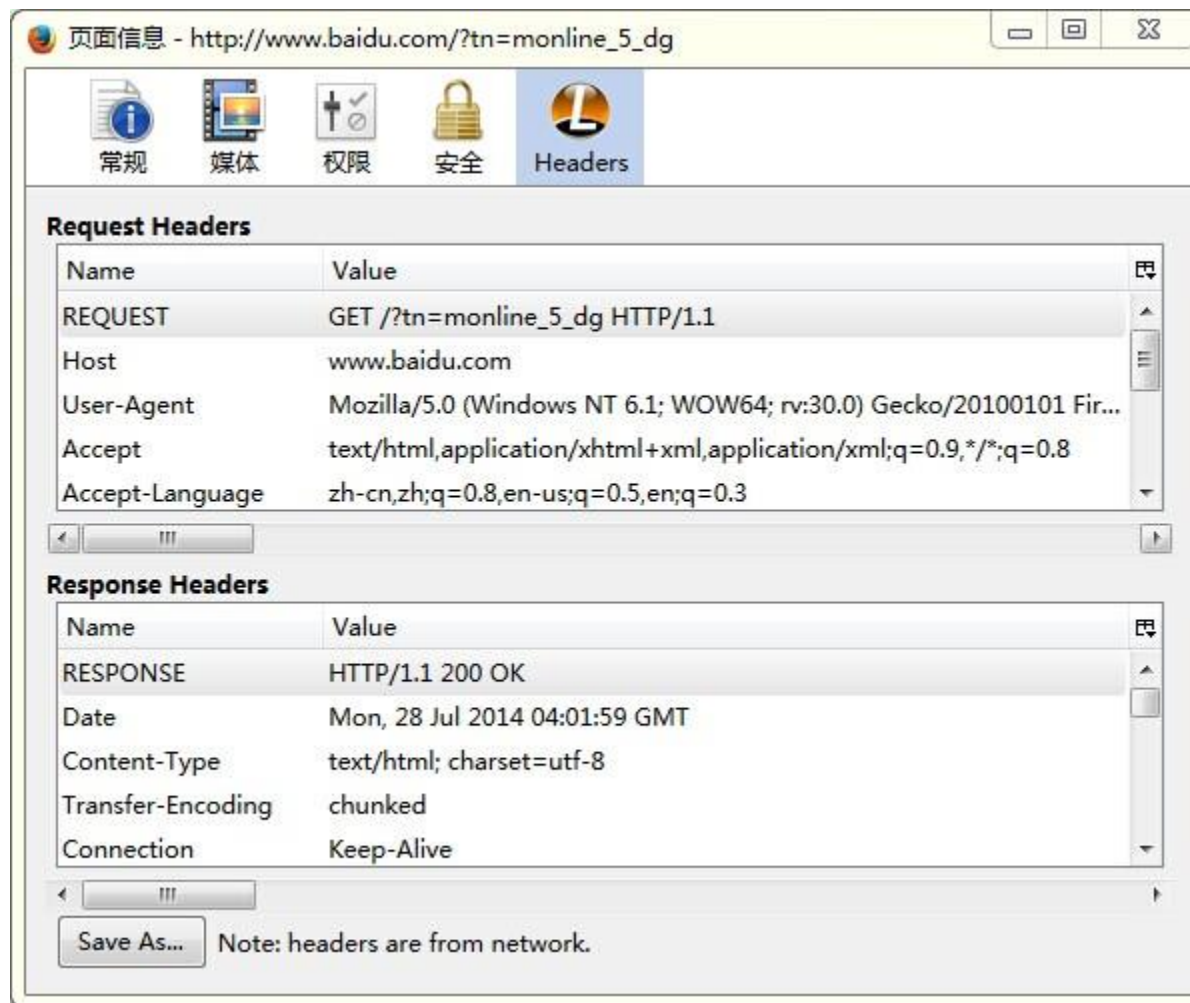
- ◆ 状态码用来告诉HTTP客户端,HTTP服务器是否产生了预期的Response.
- ◆ HTTP/1.1中定义了5类状态码，状态码由三位数字组成，第一个数字定义了响应的类别
 - 1XX 提示信息 - 表示请求已被成功接收，继续处理
 - 2XX 成功 - 表示请求已被成功接收，理解，接受
 - 3XX 重定向 - 要完成请求必须进行更进一步的处理
 - 4XX 客户端错误 - 请求有语法错误或请求无法实现
 - 5XX 服务器端错误 - 服务器未能实现合法的请求

- ◆ Server
- ◆ Date
- ◆ Last-Modified
- ◆ Content-type
- ◆ Connection
- ◆ X-Powered-By
- ◆ Content-Length
- ◆ Set-Cookie

```
> k=getURL("http://www.dataguru.cn/",headerfunction = h$update)
> h$value()

      Server      Date
      "nginx" "Tue, 05 Aug 2014 03:40:08 GMT"
Content-Type Content-Length
      "text/html"      "225561"
Last-Modified      Connection
      "Tue, 05 Aug 2014 03:10:01 GMT"      "keep-alive"
      Vary      ETag
      "Accept-Encoding"      "\"53e04b09-37119\""
Accept-Ranges      status
      "bytes"      "200"
statusMessage
      "OK"
```

利用Firefox的插件来具体查看



- ◆ getURL()
- ◆ getForm()
- ◆ postForm()



利用getURL()查看相关信息

◆ url.exists()

◆ d = debugGatherer()

```
temp <- getURL("http://www.dataguru.cn/", debugfunction=d$update, verbose = TRUE)
```

```
cat(d$value()[3])#提交给服务器的头信息
```

```
cat(d$value()[1])#服务器地址以及端口号
```

```
cat(d$value()[2])#服务器端返回的头信息
```

利用getURL()查看相关信息

- ◆ 查看服务器端返回的头信息
- ◆ ##字符串形式
- ◆ headers = basicTextGatherer()

txt=getURL("http://www.dataguru.cn/",headerfunction = headers\$update)

names(headers\$value())#说明是字符串形式

headers\$value()

```
> ##字符串形式
> headers = basicTextGatherer()
> txt=getURL("http://www.dataguru.cn/",headerfunction = headers$update)
> names(headers$value())#说明是字符串形式
NULL
> headers$value()
[1] "HTTP/1.1 200 OK\r\nServer: nginx\r\n"
```

利用getURL()查看相关信息

- ◆ 查看服务器端返回的头信息
- ◆ ###列表形式
- ◆ h = basicHeaderGatherer()

```
txtt=getURL("http://www.dataguru.cn/",headerfunction = h$update)
```

```
names(h$value())
```

```
h$value()
```

```
> k=getURL("http://www.dataguru.cn/",headerfunction = h$update)
> h$value()

      Server      Date
"nginx" "Tue, 05 Aug 2014 03:40:08 GMT"
Content-Type      Content-Length
"text/html"      "225561"
Last-Modified      Connection
"Tue, 05 Aug 2014 03:10:01 GMT"      "keep-alive"
Vary      ETag
"Accept-Encoding"      "\"53e04b09-37119\""
Accept-Ranges      status
"bytes"      "200"
statusMessage
"OK"
```


利用getURL()查看相关信息

◆ 查看curl请求的访问信息

◆ curl = getCurlHandle()

d=getURL("http://www.dataguru.cn/", curl = curl)

getCurlInfo(curl)\$response.code

getCurlInfo(curl)

```
> getCurlInfo(curl)
$effective.url
[1] "http://www.dataguru.cn/"

$response.code
[1] 200

$total.time
[1] 0.39

$namelookup.time
[1] 0.016

$connect.time
[1] 0.063

$pretransfer.time
[1] 0.063

$size.upload
[1] 0
```

```
◆ myheader <- c(  
  "User-Agent"="Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN; rv:1.9.1.6) ",  
  "Accept"="text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8",  
  "Accept-Language"="en-us",  
  "Connection"="keep-alive",  
  "Accept-Charset"="GB2312,utf-8;q=0.7,*;q=0.7"  
)
```

◆ 如何确认自己的header提交上去了呢？

除了header还有什么

- ◆ 可设参数listCurlOptions()
- ◆ 可设参数170余个！！
- ◆ 下节课会讲解几个常用参数的设置。



The screenshot shows an R console window with the following output from the listCurlOptions() function:

```
[101] "postfieldsize"      "postfieldsize.large"
[103] "postquote"          "postredir"
[105] "prequote"           "private"
[107] "progressdata"       "progressfunction"
[109] "protocols"          "proxy"
[111] "proxy.transfer.mode" "proxyauth"
[113] "proxypassword"      "proxyport"
[115] "proxytype"          "proxyusername"
[117] "proxyuserpwd"       "put"
[119] "quote"              "random.file"
[121] "range"              "readdata"
[123] "readfunction"       "redir.protocols"
[125] "referer"            "resume.from"
[127] "resume.from.large"  "seekdata"
[129] "seekfunction"       "share"
[131] "sockoptdata"        "sockoptfunction"
[133] "socks5.gssapi.nec"  "socks5.gssapi.service"
[135] "ssh.auth.types"     "ssh.host.public.key.md5"
[137] "ssh.private.keyfile" "ssh.public.keyfile"
[139] "ssl.cipher.list"    "ssl.ctx.data"
[141] "ssl.ctx.function"   "ssl.sessionid.cache"
[143] "ssl.verifyhost"     "ssl.verifypeer"
[145] "sslcert"            "sslcertpasswd"
[147] "sslcerttype"        "sslengine"
[149] "sslengine.default"  "sslkey"
[151] "sslkeypasswd"       "sslkeytype"
[153] "sslversion"         "stderr"
[155] "tcp.nodelay"        "telnetoptions"
[157] "tftp.blksize"       "timecondition"
[159] "timeout"            "timeout.ms"
[161] "timevalue"          "transfertext"
[163] "unrestricted.auth"  "upload"
[165] "url"                "use.ssl"
[167] "useragent"          "username"
[169] "userpwd"            "verbose"
[171] "writedata"          "writefunction"
```

◆ 探寻搜索原理

- 以百度搜索“RCurl”为例

◆ 提取关键字符

◆ 替换成新的搜索



◆ 提交表单

- 新浪登录的尝试
- 如何确定信息提交上去了？

◆ <http://www.dataguru.cn/article-873-1.html>



The image shows the Sina Weibo login interface. At the top left is the Sina logo and '新浪通行证 - 登录'. At the top right are links for '新浪首页' and '帮助'. The main area contains a large empty text box for a message. Below it are input fields for '登录名' (Username) and '密码' (Password). The username field has a placeholder '微博帐号/邮箱/博客' and a red error message '注册新浪通行证 请输入登录名'. The password field has a placeholder '密码' and a link '找回密码'. Below the password field are two checkboxes: '下次自动登录' and '安全登录'. At the bottom is a blue '登录' (Login) button.

- ◆ **Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。**
- ◆ **关于逆向收费式网络的详情，请看我们的培训网站 <http://edu.dataguru.cn>**



Thanks

FAQ时间