网络爬虫

陈华珊

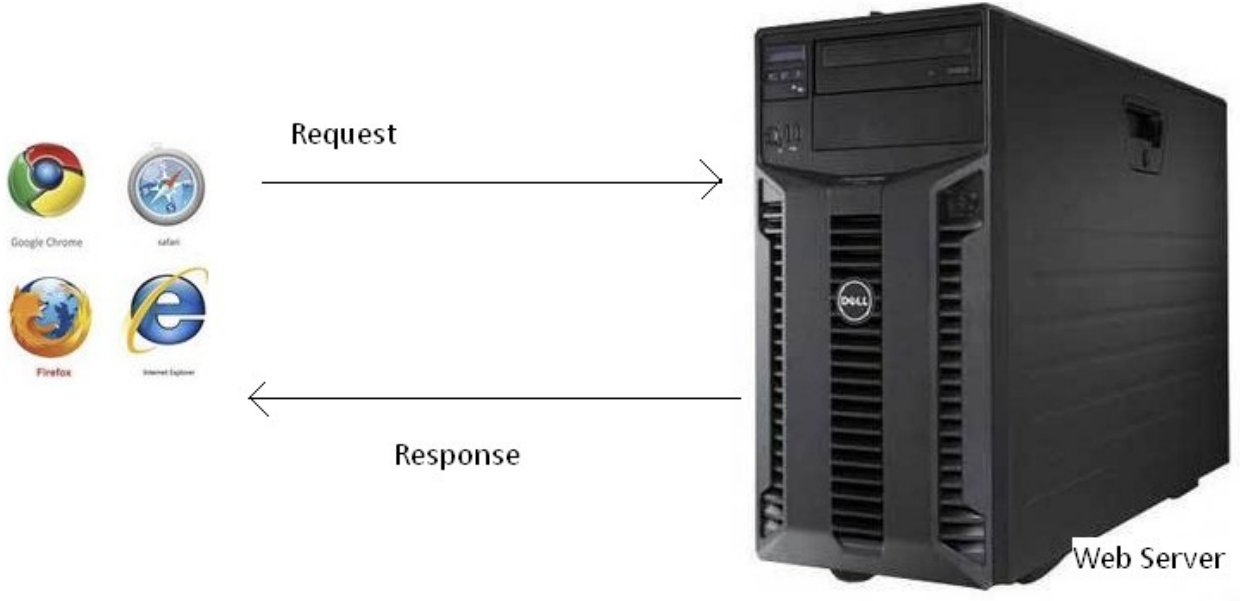
中国社科院社会发展战略研究院

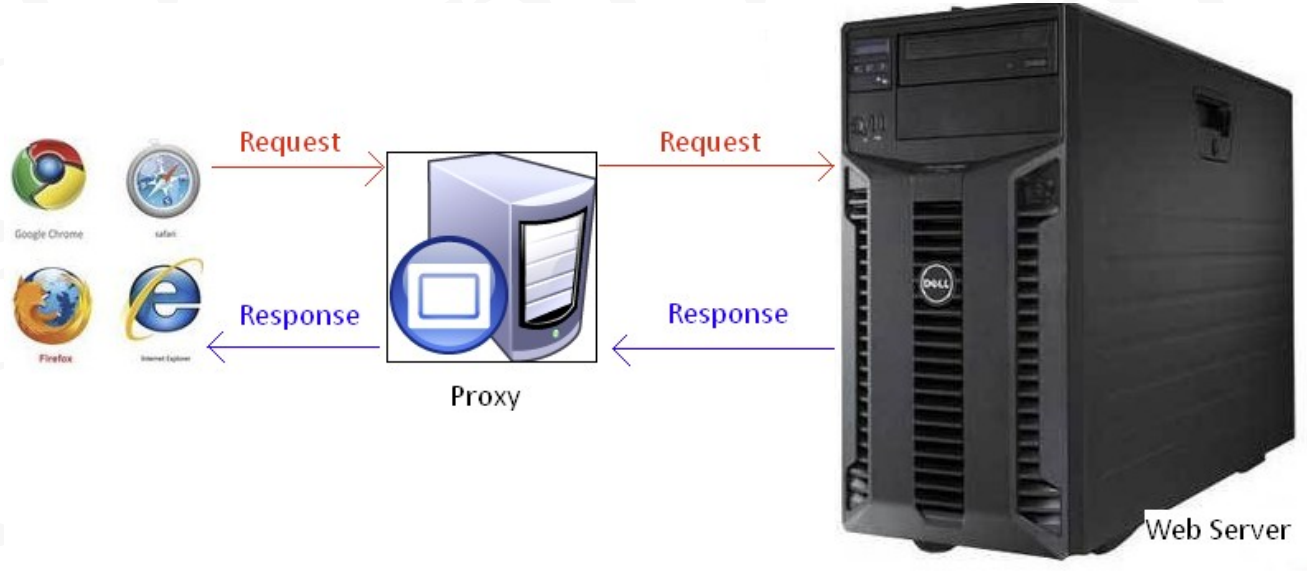
HTTP基础知识

◆ HTTP protocal (超文本传输协议)

- 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
- HTTP/2 is coming

◆ Web服务器，浏览器，代理服务器




Request
Request
Response
Response
Proxy
Web Server

◆ URL

URL(Uniform Resource Locator) 地址用于描述一个网络上的资源，基本格式如下

```
schema://host[:port#]/path/.../[;url-params][?query-string][#anchor]
```

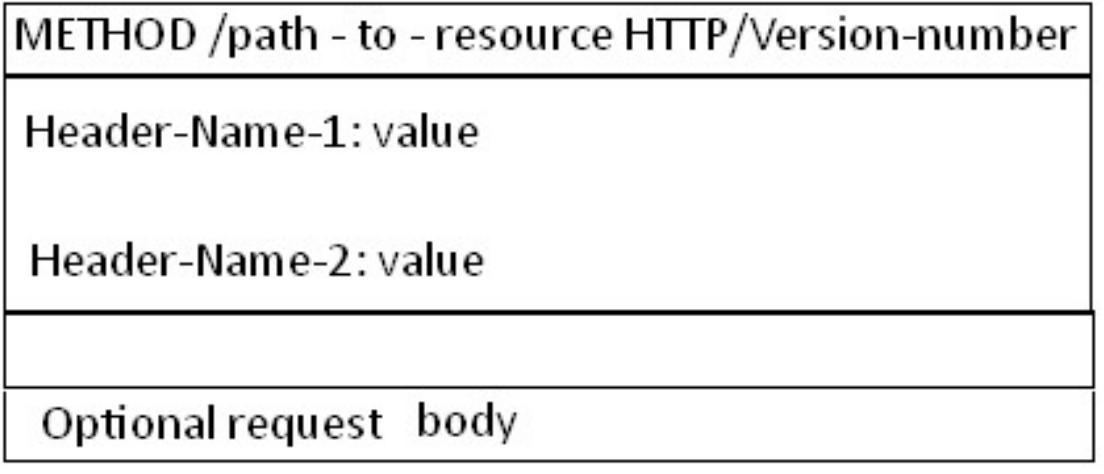
scheme	指定低层使用的协议(例如：http, https, ftp)
host	HTTP服务器的IP地址或者域名
port#	HTTP服务器的默认端口是80，这种情况下端口号可以省略。如果使用了别的端口，必须指明，例如 http://www.cnblogs.com:8080/
path	访问资源的路径
url-params	
query-string	发送给http服务器的数据
anchor	锚

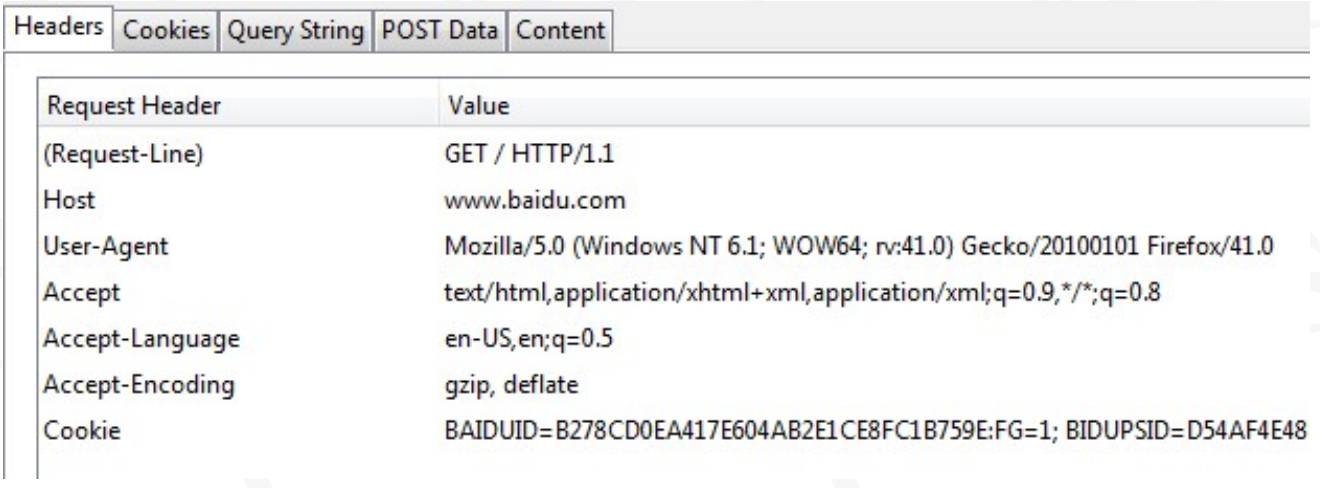
URL 的一个例子

```
1  http://www.mywebsite.com/sj/test;id=8079?name=sviergn&x=true#stuff
2
3  Schema: http
4
5  host: www.mywebsite.com
6
7  path: /sj/test
8
9  URL params: id=8079
10
11 Query String: name=sviergn&x=true
12
13 Anchor: stuff
```

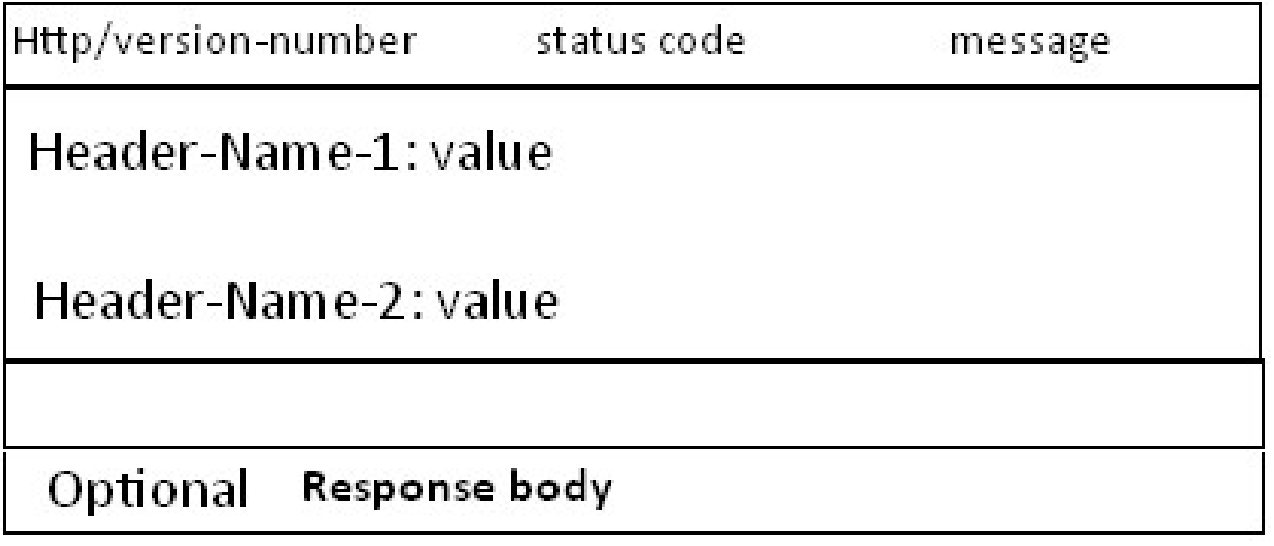
◆ HTTP消息的结构

- Request
  - 请求行
  - http header
  - body




Headers
Cookies
Query String
POST Data
Content
Request Header
Value
(Request-Line)
GET / HTTP/1.1
Host
www.baidu.com
User-Agent
Mozilla/5.0 (Windows NT 6.1; WOW64; rv:41.0) Gecko/20100101 Firefox/41.0
Accept
text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language
en-US,en;q=0.5
Accept-Encoding
gzip, deflate
Cookie
BAIDUID=B278CD0EA417E604AB2E1CE8FC1B759E:FG=1; BIDUPSID=D54AF4E48

- Response
  - 请求行
  - request header
  - body



◆ Method

与服务器交互的方法，最基本的有4种，分别是GET,POST,PUT,DELETE;

- GET vs. POST
  - GET 提交的数据会放在URL之后，以? 分割URL和传输数据，参数之间以 & 相连，如

    EditPosts.aspx?name=test1&id=123456
  - POST方法是把提交的数据放在HTTP包的Body中;
  - GET 提交的数据大小有限制（因为浏览器对URL的长度有限制）；POST 方法提交的数据没有限制;

◆ 状态码（Status）

Response 消息中的第一行叫做状态行，由HTTP协议版本号，状态码，状态消息三部分组成。

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

1 + 1XX 提示信息 - 表示请求已被成功接收，继续处理

2

3 + 2XX 成功 - 表示请求已被成功接收，理解，接受

4

5 + 3XX 重定向 - 要完成请求必须进行更进一步的处理

6

7 + 4XX 客户端错误 - 请求有语法错误或请求无法实现

8

9 + 5XX 服务器端错误 - 服务器未能实现合法的请求

404 Not Found

---

nginx/1.8.0

很抱歉，您要访问的页面不存在！

温馨提示：

1. 请检查您访问的网址是否正确
2. 如果您不能确认访问的网址，请浏览百度更多页面查看更多网址。
3. 回到顶部重新发起搜索
4. 如有任何意见或建议，请及时反馈给我们。

R调用HTTP解析

包的加载及调用

```
1  library(httr)
2
3  r <- GET("http://www.baidu.com")
```

查看 response 对象，有用的信息包括：真实的URL（URL跳转）、HTTP状态码、内容类型、大小、etc.

```
r
## Response [http://www.baidu.com/]
##   Date: 2015-10-25 15:23
##   Status: 200
##   Content-Type: text/html; charset=utf-8
##   Size: 96.5 kB
## Error in gregexpr("\n", text, fixed = TRUE) : invalid multibyte
string at '<b5>"<20>title="鎵 扮 櫨 搴┖椤<b5>"></a><form id="form"
name="f" action="/s" class="fm"><input type="hidden" name="ie"
value="utf-8"><input type="hidden" name="f" value="8"><input type="hidden
name="rsv_bp" value="1"><input type="hidden" name="rsv_idx" value="1"><in
type=hidden name=ch value=""><input type=hidden name=tn value="baidu"><in
type=hidden name=bar value=""><span class="bg s_ipt_wr"><input
id="kw" name="wd" class="s_ipt" value="" maxlength="255" autocomplete="of
class="bg s_btn_wr"><input type="submit" id="su" value="鐗      惧
害    涓€涓<8b>" class="bg s_btn"></span><span class="tools"><span
```

```
8    id="mHolder"><div id="mCon"><span>杈 撳 磩 娉<95></span></div><ul
     id="mMenu"><li><a href="javascript:    ;" name="ime_hw">鎵       嬪
     啓</a></li><li><a href="javascript:    ;" name="ime_py">鎷       奸
     煡</a></li><li class="ln"></li><li><a href="javascript:;" name="ime_cl">鍏
     抽    榔</a></li></ul></span></span><input type="hidden" name="rn"
     value=""><input type="hidden" name=
```

继续挖掘

```
status_code(r)
## [1] 200
headers(r)
## $date
## [1] "Sun, 25 Oct 2015 15:23:20 GMT"
##
## $`content-type`
## [1] "text/html; charset=utf-8"
##
## $`transfer-encoding`
## [1] "chunked"
##
## $connection
## [1] "Keep-Alive"
##
## $vary
## [1] "Accept-Encoding"
```

```
18  ##
19  ## $`set-cookie`
20  ## [1] "BAIDUID=680D02880F58BA09472B3FBE7D3722D4:FG=1; expires=Thu,
    31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; domain=.baidu.com"
21  ##
22  ## $`set-cookie`
23  ## [1] "BIDUPSID=680D02880F58BA09472B3FBE7D3722D4; expires=Thu,
    31-Dec-37 23:55:55 GMT; max-age=2147483647; path=/; domain=.baidu.com"
24  ##
25  ## $`set-cookie`
26  ## [1] "PSTM=1445786600; expires=Thu, 31-Dec-37 23:55:55 GMT;
    max-age=2147483647; path=/; domain=.baidu.com"
27  ##
28  ## $`set-cookie`
29  ## [1] "BDSVRTM=0; path=/"
30  ##
31  ## $`set-cookie`
32  ## [1] "BD_HOME=0; path=/"
```

```
33  ##
34  ## $`set-cookie`
35  ## [1] "H_PS_PSSID=17519_17386_13550_1452_17636_17620_17640_12825_14429_1
    path=/; domain=.baidu.com"
36  ##
37  ## $p3p
38  ## [1] "CP=\" OTI DSP COR IVA OUR IND COM \""
39  ##
40  ## $`cache-control`
41  ## [1] "private"
42  ##
43  ## $cxy_all
44  ## [1] "baidu+d1cb143c7b192070cdeb480150617f40"
45  ##
46  ## $expires
47  ## [1] "Sun, 25 Oct 2015 15:22:48 GMT"
48  ##
49  ## $`x-powered-by`
```

```
50  ## [1] "HPHP"
51  ##
52  ## $server
53  ## [1] "BWS/1.1"
54  ##
55  ## $`x-ua-compatible`
56  ## [1] "IE=Edge,chrome=1"
57  ##
58  ## $bdpagetype
59  ## [1] "1"
60  ##
61  ## $bdqid
62  ## [1] "0xafc6a8ef0003da4e"
63  ##
64  ## $bduserid
65  ## [1] "0"
66  ##
67  ## $`content-encoding`
```

```
68  ## [1] "gzip"
69  ##
70  ## attr(,"class")
71  ## [1] "insensitive" "list"
72  str(content(r))
73  ## Classes 'HTMLInternalDocument', 'HTMLInternalDocument', 'XMLInternalDo
    'XMLAbstractDocument' <externalptr>
```

◆ Body

- 文本内容

```
content(r, 'text')
content(r, "text", encoding = "ISO-8859-1")
```

- 非文本内容

```
content(r, "raw")
writeBin(r, 'aaa.jpg')
```

- 数据

JSON automatically parsed into named list

```
content(r, "parsed")
```

◆ Cookie

```
r$cookies
##          domain  flag path secure          expiration       name
## 1    .baidu.com  TRUE    /  FALSE 2083-11-13 02:37:25    BAIDUID
## 2    .baidu.com  TRUE    /  FALSE 2083-11-13 02:37:25   BIDUPSID
## 3    .baidu.com  TRUE    /  FALSE 2083-11-13 02:37:25       PSTM
## 4 www.baidu.com FALSE    /  FALSE                <NA>    BDSVRTM
## 5 www.baidu.com FALSE    /  FALSE                <NA>    BD_HOME
## 6    .baidu.com  TRUE    /  FALSE                <NA> H_PS_PSSID
##
## 1                                                      680D02880F5
## 2                                                          680D02
## 3
## 4
## 5
## 6 17519_17386_13550_1452_17636_17620_17640_12825_14429_17447_17001_174
```

传输 cookie

```
1    r <- GET("http://httpbin.org/cookies/set", query = list(a = 1))
2    cookies(r)
```

◆ 传递参数

```
1  r <- GET("http://httpbin.org/get",
2    query = list(key1 = "value1", key2 = "value2")
3  )
4  content(r)$args
5
6  r <- POST("http://httpbin.org/post", body = list(a = 1, b = 2, c
   = 3))
7
8  # 上传文件
9  POST(url, body = upload_file("mypath.txt"))
```

爬虫实践