上周听完，并没有急着去想这动态页面的爬虫该如何的去写好。当时只知道get与post最表面的意思，还有一些参数也不知道，http协议也不是很清楚，反正当时的第一想法就是想去了解这部分的知识点（偏离轨道有没有，我感觉还是有点小用）。

参考资料：

1）浅谈HTTP中GET和POST的区别：

<http://www.cnblogs.com/hyddd/archive/2009/03/31/1426026.html>

2）HTTP协议详解

<http://www.cnblogs.com/EricaMIN1987_IT/p/3837436.html>

3）关于HTTP协议这一篇就够了

<http://www.cnblogs.com/ranyonsue/p/5984001.html>

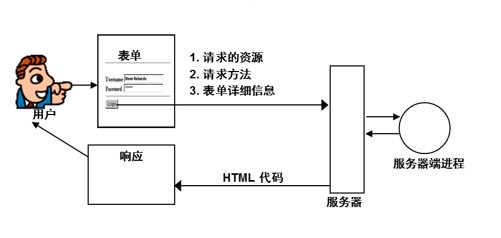
4）网页访问流程以及爬虫的基本原理

<http://www.hubwiz.com/class/570dce425fd193d76fcc723d>

**整理如下： （虽然都是copy，至少脑海里过了一下）**

**第一部分：网页的访问流程**

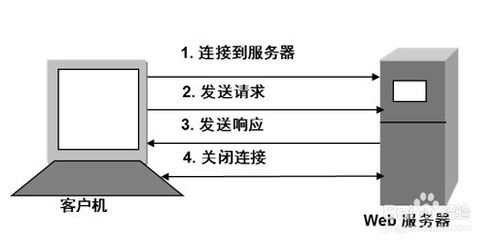
[**https://jingyan.baidu.com/article/4b52d702a863adfc5c774bf9.html**](https://jingyan.baidu.com/article/4b52d702a863adfc5c774bf9.html)

****

Web浏览器与Web服务器之间都经历了哪些过程？

1. ***建立TCP连接***

在HTTP工作开始之前，Web浏览器首先要通过网络与Web服务器建立连接，该连接是通过TCP来完成的，该协议与IP协议共同构建Internet，即著名的TCP/IP协议族，因此Internet又被称作是TCP/IP网络。HTTP是比TCP更高层次的应用层协议，根据规则，只有低层协议建立之后才能，才能进行更层协议的连接，因此，首先要建立TCP连接，一般TCP连接的端口号是80。

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=2)

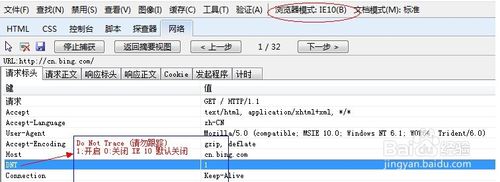
三次握手的过程

1. ***Web浏览器向Web服务器发送请求命令***

一旦建立了TCP连接，Web浏览器就会向Web服务器发送请求命令。例如：GET/sample/hello.jsp HTTP/1.1。

1. ***Web浏览器发送请求头信息***

浏览器发送其请求命令之后，还要以头信息的形式向Web服务器发送一些别的信息，之后浏览器发送了一空白行来通知服务器，它已经结束了该头信息的发送。

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=3)

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=4)

1. ***Web服务器应答***

客户机向服务器发出请求后，服务器会客户机回送应答， HTTP/1.1 200 OK ，应答的第一部分是协议的版本号和应答状态码。

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=5)

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=6)

1. ***Web服务器发送应答头信息***

正如客户端会随同请求发送关于自身的信息一样，服务器也会随同应答向用户发送关于它自己的数据及被请求的文档。

1. ***Web服务器向浏览器发送数据***

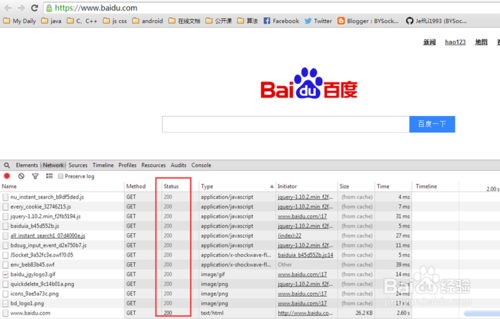
Web服务器向浏览器发送头信息后，它会发送一个空白行来表示头信息的发送到此为结束，接着，它就以Content-Type应答头信息所描述的格式发送用户所请求的实际数据。

1. ***Web服务器关闭TCP连接***

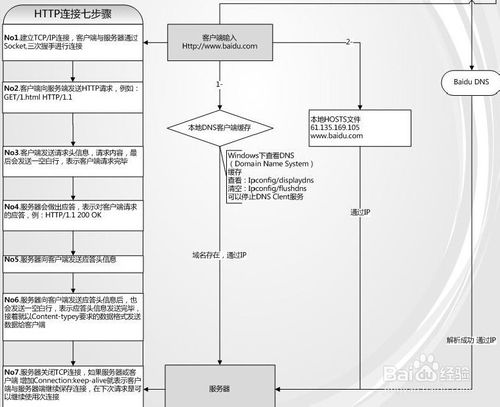
一般情况下，一旦Web服务器向浏览器发送了请求数据，它就要关闭TCP连接，然后如果浏览器或者服务器在其头信息加入了这行代码：

Connection:keep-alive

TCP连接在发送后将仍然保持打开状态，于是，浏览器可以继续通过相同的连接发送请求。保持连接节省了为每个请求建立新连接所需的时间，还节约了网络带宽。

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=7)

1. 完整的过程，如下图所示：

[](http://jingyan.baidu.com/album/4b52d702a863adfc5c774bf9.html?picindex=8)

**GET和POST请求的区别：**

**爬虫中的cookie：**

准备用fiddler抓学校教务系统的cookie，竟然没找到cookie，只在webform中找到学号与密码

浏览器把cookie通过HTTP Request 中的“Cookie: header”发送给Web服务器

Web服务器通过HTTP Response中的"Set-Cookie: header"把cookie发送给浏览器

当我们模拟登陆的时候就会用到cookie了：

cookie是很重要的，识别是否是登陆用户，就是通过cookie。  假如截获了别人的cookie是否可以冒充他人的身份登陆呢？  当然可以了！