**参考：**[**http://cuiqingcai.com/942.html**](http://cuiqingcai.com/942.html)

[Python爬虫入门二之爬虫基础了解](http://cuiqingcai.com/942.html)

[Python](http://cuiqingcai.com/category/technique/python)  [崔庆才](http://cuiqingcai.com/author/cqcre)  3年前 (2015-02-12)  217726浏览  [10评论](http://cuiqingcai.com/942.html#comments)

**1.什么是爬虫**

爬虫，即网络爬虫，大家可以理解为在网络上爬行的一直蜘蛛，互联网就比作一张大网，而爬虫便是在这张网上爬来爬去的蜘蛛咯，如果它遇到资源，那么它就会抓取下来。想抓取什么？这个由你来控制它咯。

比如它在抓取一个网页，在这个网中他发现了一条道路，其实就是指向网页的超链接，那么它就可以爬到另一张网上来获取数据。这样，整个连在一起的大网对这蜘蛛来说触手可及，分分钟爬下来不是事儿。

**2.浏览网页的过程**

在用户浏览网页的过程中，我们可能会看到许多好看的图片，比如 <http://image.baidu.com/> ，我们会看到几张的图片以及百度搜索框，这个过程其实就是用户输入网址之后，经过DNS服务器，找到服务器主机，向服务器发出一个请求，服务器经过解析之后，发送给用户的浏览器 HTML、JS、CSS 等文件，浏览器解析出来，用户便可以看到形形色色的图片了。

因此，用户看到的网页实质是由 HTML 代码构成的，爬虫爬来的便是这些内容，通过分析和过滤这些 HTML 代码，实现对图片、文字等资源的获取。

**3.URL的含义**

URL，即统一资源定位符，也就是我们说的网址，统一资源定位符是对可以从互联网上得到的资源的位置和访问方法的一种简洁的表示，是互联网上标准资源的地址。互联网上的每个文件都有一个唯一的URL，它包含的信息指出文件的位置以及浏览器应该怎么处理它。

URL的格式由三部分组成：  
①第一部分是协议(或称为服务方式)。  
②第二部分是存有该资源的主机IP地址(有时也包括端口号)。  
③第三部分是主机资源的具体地址，如目录和文件名等。

爬虫爬取数据时必须要有一个目标的URL才可以获取数据，因此，它是爬虫获取数据的基本依据，准确理解它的含义对爬虫学习有很大帮助。

**4. 环境的配置**

学习Python，当然少不了环境的配置，最初我用的是Notepad++，不过发现它的提示功能实在是太弱了，于是，在Windows下我用了PyCharm，在Linux下我用了Eclipse for Python，另外还有几款比较优秀的IDE，大家可以参考这篇文章 [学习Python推荐的IDE](http://www.oschina.net/news/57468/best-python-ide-for-developers) 。好的开发工具是前进的推进器，希望大家可以找到适合自己的IDE

下一节，我们就正式步入 Python 爬虫学习的殿堂了，小伙伴准备好了嘛？

转载请注明：[静觅](http://cuiqingcai.com/) » [Python爬虫入门二之爬虫基础了解](http://cuiqingcai.com/942.html)