网络爬虫（一）： 网络爬虫科普与URL含义

2017-11-08 1332

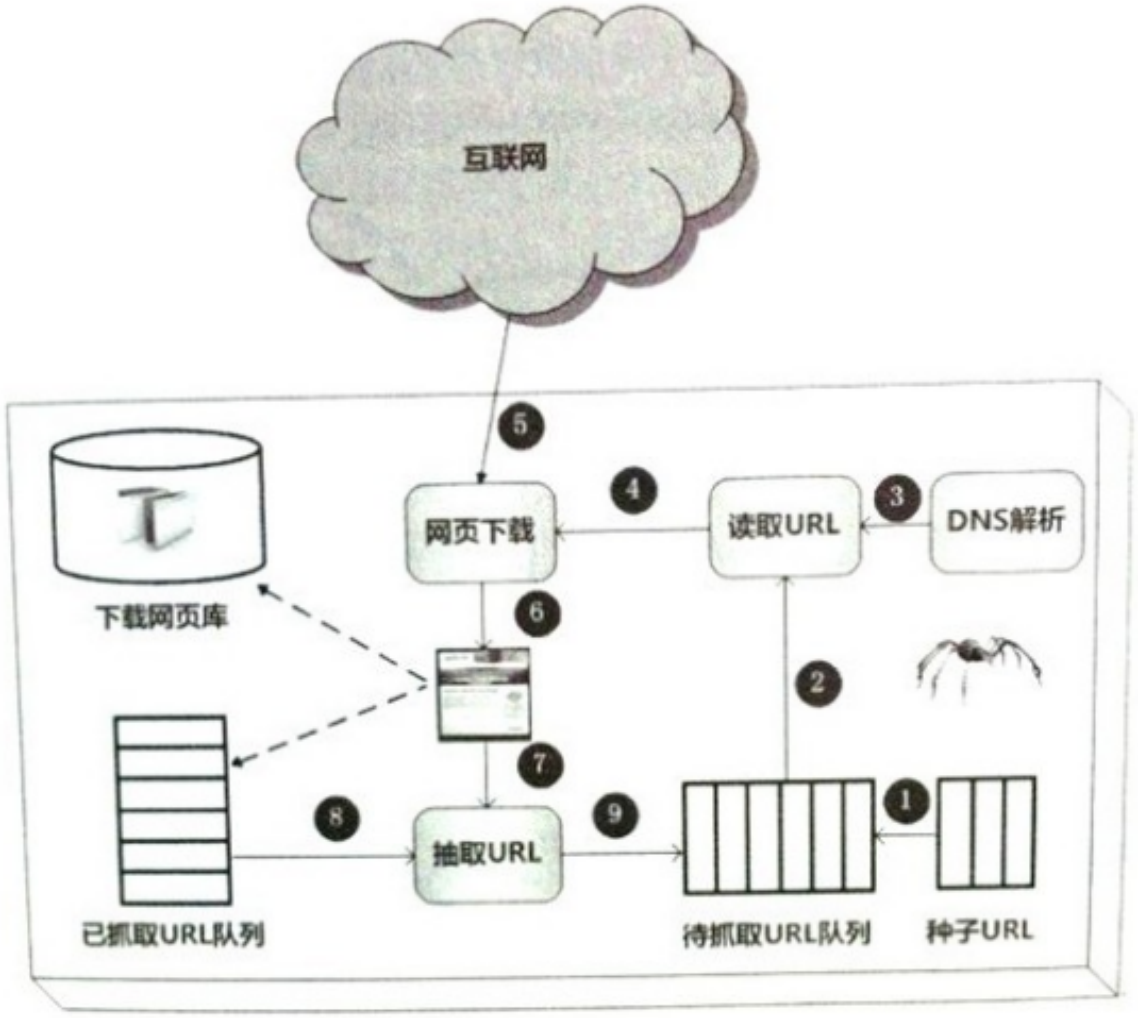
1. 科普

通用搜索引擎处理的对象是互联网的网页，目前网页的数量数以亿计，所以搜索引擎面临的第一个问题是如何设计出高效的下载系统，已将海量的网页下载到本地，在本地形成互联网网页的镜像。网络爬虫就是担当此大任的。

抓取网页的过程其实和读者平时使用IE浏览器浏览网页的道理是一样的。比如说你在浏览器的地址栏中输入 www.baidu.com 这个地址。打开网页的过程其实就是浏览器作为一个浏览的“客户端”，向服务器端发送了一次请求，把服务器端的文件“抓”到本地，再进行解释、展现。浏览器的功能是将获取的HTML代码进行解析，然后将原始的网页转化为我们看到的网站页面。

网络爬虫最基本的思路就是：从一个页面开始，分析其中的url，提取出来，然后通过这些链接寻求下一个页面。如此往复。

2. 通用爬虫框架



首先，从互联网上精心选择一部分网页，以这些网页的链接地址最为种子URL，将这些种子URL存入待抓取的URL队列(1)，从待抓取的URL队列开始读取一个url(2)。其中的链接地址经过DNS解析(3)转化为网站服务器对应的IP地址。网页下载器根据IP地址向服务器发送请求，获得网页(5)，下载好网页后一方面作为原始数据保存到页面库中，等待建立索引等后处理。另一方面将该网页的地址存到已经抓取的队列(8)（避免重复爬取）。对于刚才爬取的网页进行解析(6)，抽取其中的url(7)。对于不再已抓取URL队列中的URL和存于待抓取URL队列(9)，重复刚才的操作。

3. URL

爬虫最主要的处理对象是URL。简单说url就是输入的网址（例如：http://www.cnblogs.com/kaituorensheng/）。理解URL之前首先理解URI.

Web上每种可用的资源，如 HTML文档、图像、视频片段、程序等都由一个通用资源标志符(Universal Resource Identifier， URI)进行定位。
URI通常由三部分组成：

1. 访问资源的命名机制
2. 存放资源的主机名
3. 资源自身 的名称，由路径表示

如URI：http://www.why.com.cn/myhtml/html1223/

我们可以这样解释它：

- 这是一个可以通过HTTP协议访问的资源
- 位于主机 www.why.com.cn上
- 通过路径“/myhtml/html1223/”访问

URL是URI的一个子集。它是Uniform Resource Locator的缩写，译为“统一资源定位符”。
通俗地说，URL是Internet上描述信息资源的字符串，主要用在各种WWW客户程序和服务器程序上。
采用URL可以用一种统一的格式来描述各种信息资源，包括文件、服务器的地址和目录等。
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

第一部分和第二部分用“://”符号隔开
第二部分和第三部分用“/”符号隔开
第一部分和第二部分是不可缺少的，第三部分有时可以省略

3. 1 HTTP协议的URL示例
使用超级文本传输协议HTTP，提供超级文本信息服务的资源。
例：
其计算机域名为www.peopledaily.com.cn。
超级文本文件(文件类型为.html)是在目录 /channel下的welcome.htm。

3.2 文件的URL
用URL表示文件时，服务器方式用file表示，后面要有主机IP地址、文件的存取路径(即目录)和文件名等信息。
有时可以省略目录和文件名，但“/”符号不能省略。
例：file://ftp.yoyodyne.com/pub/files/foobar.txt

上面这个URL代表存放在主机ftp.yoyodyne.com上的pub/files/目录下的一个文件，文件名是foobar.txt。