06 | 网络爬虫：爬取一个网站的流程是怎样的?

2023-04-21 黄鸿波 来自北京

《手把手带你搭建推荐系统》

你好，我是黄鸿波。

在前面的课程中，我们讲解了什么是推荐系统、推荐系统的数据构成与处理，还有数据库相关的知识，这些知识都是我们搭建推荐系统的基础。有了这些基础，下一步我们就可以尝试获取数据了。

那真实数据应该从何而来呢？前面我们讲过，推荐系统的数据来源是多样的，它可以由官方或者用户产生，也可以借助爬虫来获得。而我们这门课程的数据，主要是使用爬虫技术来获得。

这节课，我们就先来看看爬虫是什么，它的工作流程又是怎样的。

什么是爬虫?

爬虫的英文是 Spider，又称网络蜘蛛，它本质上是一种计算机程序。爬虫通过模拟人类的操作，并按照一定的规则自动浏览和检索网页的信息，将人们所需的数据抓取下来，然后对抓取到的数据进行处理，从而提取出有价值的信息。

网络爬虫按照系统结构和实现技术，大致可以分为下面四种类型。

1. 通用网络爬虫（General Purpose Web Crawler）。
2. 聚焦网络爬虫（Focused Web Crawler）。
3. 增量式网络爬虫（Incremental Web Crawler）。
4. 深层网络爬虫（Deep Web Crawler）。

实际的网络爬虫系统通常是由几种爬虫技术相结合来实现的。

最常见的爬虫系统就是各个搜索引擎了。像百度、谷歌等搜索引擎都有自己的爬虫程序，这些爬虫程序每天都在互联网中爬取各种各样的信息，然后将它们按照关键词和网站热度进行排序，最后将排序的结果呈现给用户，这就是我们最后通过引擎搜索出来的内容。

我们可以爬取哪些数据?

我们知道，互联网实际上是由一个个网页组成，借助超链接，网页之间可以相互跳转。理论上讲，我们可以通过超链接到达互联网中的任何一个网页，爬虫也可以抓取任意一个链接里的内容。

但事实上，我们并不能这么做，这不光是技术上的问题，还涉及法律层面的问题。

所有的爬虫程序要爬取互联网上的内容都需要遵循相关的协议，其中最重要的协议就是 robots 协议，它也叫做机器人协议或爬虫协议。这个协议其实很简单，它就是要告诉爬虫程序，在本网站下，哪些目录和文件是可以爬取的，哪些目录和文件不能被爬取。如果机器人不遵循 robots 协议，就有可能面临法律风险。

一般来讲，robots 协议是一个以 robots.txt 命名的文件，它被放在了网站的根目录里。也就是说，当爬虫程序看到这个文件，就应该遵循这个文件所定义的规范。当然，如果爬虫爬取到

的网站没有这个文件，从原则上讲，它就可以爬取这个网站下所有可以直接访问到的链接。

下面我列举几个关键的写法。

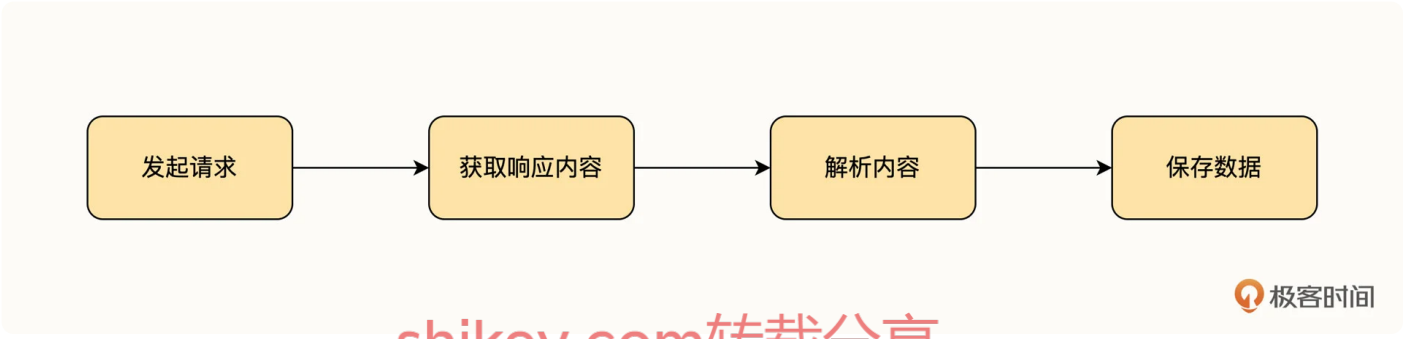
键值	附注
User-agent: *	这里的*是一个通配符，它代表所有的搜索引擎种类
Disallow: /admin/	禁止爬寻admin目录下面的目录
Disallow: /require/	禁止爬寻require目录下面的目录
Disallow: /cgi-bin/*.htm	禁止访问/cgi-bin/目录下的所有以“.htm”为后缀的URL（包含子目录）
Disallow: /*?*	禁止访问网站中所有包含问号“?”的网址
Allow: /cgi-bin/	允许爬寻cgi-bin目录下面的目录
Allow: /tmp	允许爬寻tmp的整个目录
Allow: .htm$	仅允许访问以“.htm”为后缀的URL

当然，除了上面这些最常用的规则之外，还有很多种写法，如果后面有需要，我们再进行讲解。

爬虫的工作流程

现在，我们已经知道了爬取网站时应该遵循的规则，那爬取网站的流程是怎样的呢？

实际上，爬虫爬取网站的流程和我们浏览网页时浏览器的操作是一样的，它主要经过了下面这几步。



首先，爬虫程序要向链接发起请求，然后等待网页响应请求。服务器响应之后，会将响应之后的内容返回给浏览器。此时，爬虫需要对内容进行解析。对于普通的用户来说，这一步对应的就是浏览器内核解析数据的过程。得到数据之后，用户会浏览数据，但是对于爬虫项目来讲，我们的目的是获取数据，所以，爬虫程序在解析完数据之后会将自己需要的那一部分内容保存下来，从而形成我们最原始版本的数据集。

接下来，我们详细拆解一下这几个部分。

发起请求

发起请求是指爬虫程序向指定的网页发送请求的过程。一般来讲，在 HTTP 协议中，请求的方式有两种，一种是使用 GET 请求，另外一种是使用 POST 请求。而请求的链接我们称之为 URL，也就是统一资源定位符。

无论是什么请求，我们都要在请求的内容中加上请求头，只有加了请求头，网站所在的服务器才会把你当作一个正常的用户，否则就是一个非法用户。

对于网站来说，**不同类型的内容会接收不同类型的请求。**一般来说，如果请求只包含数据内容，在这一过程中不涉及内容保密和文件传输，我们会使用 GET 请求。比如我们要获取一个类别的列表或者获取一个链接的详情页等，都会使用 GET 请求。

对于那种需要提交文件或者图片，以及对整体安全性要求比较高的内容，我们一般使用 POST 请求。比如在论坛中回复帖子、注册和登录网站、上传文件信息等操作。

为什么说对安全性的要求会影响请求的方式呢？

如果使用 GET 请求，实际上就是在我们原有的请求字符串后面加一个问号，然后加上一系列的 key=value，并且每个 key-value 对都会用“&”隔开，有几个参数就有多少个 key-value 对。例如，下面是我使用百度搜索 Python 这个关键词后，地址栏所产生的链接如下。

复制代码

```
1 https://www.baidu.com/s?wd=python&rsv_spt=1&rsv_iqid=0x9f44c577000d575e&issp=1&f=
```


得到的内容如图。

我们可以看到，在这串链接中，我们的基础链接是https://www.baidu.com/s，后面跟着若干个参数，每个参数都有独特的含义。每个参数的形式都是 key=value，每个参数的后面都有一个 & 符号，后面再跟下一组参数，这就是典型的 GET 请求方式。但要注意的是，GET 请求参数的总长度是有限制的，对于不同的浏览器来说，限制的长度也不一样，具体的长度限制你可以去浏览器中搜索一下。

也就是说，使用 GET 请求的时候，我们能够清楚地从地址栏看到我们都使用了什么参数，这种请求方式对于列表类的信息，或者说不需要加密的信息来说非常适用，但是对于一些需要保密的信息，比如用户名和密码就不是很适用了。因为这样很容易将隐私数据暴露出来，降低安全性。

因此，对于在安全性方面有要求，或者参数比较大（比如说一个文件、一个视频或一个图片等）的情况，GET 请求就不太适合了，这个时候我更建议使用 POST 请求。

与 GET 请求相比，POST 请求的最大优点是**所有的请求参数都不会暴露在地址栏中，也就是说，它不会作为 URL 的一部分展示出来。另一方面，相比较 GET 请求的字符限制特性，POST 请求能发送的数据量更大。**再回看 GET 请求，视频和较大的图像很难通过，即使请求的数据在字符限制之内，由于网络传输问题，图片或视频在上传时也容易中断或丢包，导致上传的内容不完整，无法观看。总之，在对安全性和数据量有要求时，POST 请求是更好的选择。

获取响应内容

发起请求之后，服务端就开始处理我们的请求。一旦服务端处理好我们的请求，就会将处理结果返回给用户。这一步服务器端的操作就是响应请求并处理，而对于客户端或者爬虫程序来说，就是获取响应内容。

一般来讲，我们获取响应内容要注意以下几个点：**响应状态码、响应头、响应体。**

响应状态码也叫做 HTTP 状态码（HTTP Status Code），是用以表示网页服务器超文本传输协议响应状态的 3 位数字代码。它由 RFC 2616 规范定义，并经过了 RFC 2518、RFC 2817、RFC 2295、RFC 2774 与 RFC 4918 等规范的扩展，所有状态码的第一个数字代表了响应的五种状态之一，我们比较常见的状态码有下面这几个。

响应状态码	含义
200	OK，请求已成功，请求所希望的响应头或数据体将随此响应返回。出现此状态码表示状态正常。
301	Moved Permanently，被请求的资源已永久移动到新位置，并且将来任何对此资源的引用都应该使用本响应返回的若干个 URI 之一。如果可能，拥有链接编辑功能的客户端应当自动把请求的地址修改为从服务器反馈回来的地址。除非额外指定，否则这个响应也是可缓存的。
400	Bad Request，返回这个状态码可能有下面两种原因。 • 语义有误，当前请求无法被服务器理解，除非进行修改，否则客户端不应该重复提交这个请求。 • 请求参数有误。
403	Forbidden，服务器已经理解请求，但是拒绝执行它。与401响应不同的是，身份验证并不能提供任何帮助，而且这个请求也不应该被重复提交。如果这不是一个 Head 请求，而且服务器希望能够讲清楚为何请求不能被执行，那么就应该在实体内描述拒绝的原因。当然，假如它不希望让客户端获得任何信息，服务器也可以返回一个404响应。
404	Not Found，请求失败，未在服务器上发现请求所希望得到的资源。
500	Internal Server Error，服务器遇到了一个未曾预料的状况，导致它无法完成对请求的处理。 一般来说，这个问题都会在服务器端的源代码出现错误时出现。
501	Not Implemented，服务器不支持当前请求所需要的某个功能。当前服务器无法识别请求的方法，并且不支持其对任何资源的请求。
502	Bad Gateway，作为网关或者代理工作的服务器尝试执行请求时，从上游服务器接收到无效的响应。
503	Service Unavailable，由于临时的服务器维护或者过载，服务器当前无法处理请求。这个状况是临时的，并且将在一段时间以后恢复。如果能够预计延迟时间，那么响应中可以包含一个 Retry-After 头用以标明这个延迟时间。如果没有给出这个 Retry-After 信息，那么客户端应当以处理500响应的方式处理它。

一般来讲，我们在爬取数据的时候，最重要的一个状态码就是 200。因为只有在接收到 200 这个状态码时，才说明我们后面的内容有可能是正常的。**也就是说，只有接收到 200 状态码，才有返回数据。**

解析和保存数据

接收到正常数据之后，下一步就是要解析和保存数据了。我们在上一个步骤中得到的数据一般来讲都是 HTML 数据，我们要做的就是从这些 HTML 数据中提取出我们想要的内容，然后再将这些内容存储到我们指定的数据库或者文本中，供后续加以利用。

一般来讲，在爬虫实践中，我们可以将解析的数据分成 3 大类。

第一类就是 HTML 数据，这也是最基础的数据。我们要将网页的结构解析出来，并提取我们想要的内容。

第二类数据是 JSON 数据。因为我们爬取出来的内容常常是被 JSON 格式包裹着，所以我们要做的就是将这一部分数据解析出来，获取到里面的内容加以利用。

第三类数据是二进制数据。因为有些时候开发人员会以二进制的形式写入数据，所以我们有时还要解析二进制格式的数据。

具体的操作方法我们后续还会详细去实操，一步步完成网站的解析。

总结

最后我来给你总结一下这节课的内容。学完这节课，你应该了解下面这些知识点。

1. 爬虫的英文是 Spider，又称网络蜘蛛，它的主要目的是让程序像人一样对网站进行浏览并抓取数据。
2. 爬虫主要分为通用网络爬虫、聚焦网络爬虫、增量式网络爬虫和深层网络爬虫四种类型，而实际的网络爬虫系统通常是由多种爬虫技术相结合来实现的。
3. 爬虫的工作流程共分为四步，即发起请求、获取相应内容、解析内容和保存数据。我们应该清楚地了解每个步骤所做的事情。

课后题

学完这节课，给你留两道课后题。

1. 我们说，网络爬虫系统通常是由几种爬虫技术结合起来实现的，那么这几类爬虫技术分别有什么特点呢？
2. 请你预习一下 Scrapy 框架，这也是我们下节课的重点。

欢迎你在留言区与我交流讨论，我们下节课见！

精选留言 (1)

peter

2023-04-21 来自北京

请问：爬虫有开源的吗？类似于工具软件那种，拿来就能用，这样就不需要自己开发了。

作者回复: 同学你好，后面我会把我们的代码放到github上，到时候大家可以去下载。

1