大数据环境下基于 python 的网络爬虫技术

作者/谢克武，重庆工商大学派斯学院软件工程学院

摘要：随着互联网的发展壮大，网络数据呈爆炸式增长，传统搜索引擎已经不能满足人们对所需求数据的获取的需求，作为搜索引擎的抓取数据的重要组成部分，网络爬虫的作用十分重要，本文首先介绍了在大数据环境下网络爬虫的重要性，接着介绍了网络爬虫的概念，工作原理，工作流程，网页爬行策略，python在编写爬虫领域的优势，最后设计了一个通用网络爬虫的框架，介绍了框架中模块的相互协作完成数据抓取的过程。

关键词：网络爬虫；python；数据采集；大数据

DOI:10.16589/j.cnki.cn11-3571/tn.2017.09.017

引言

大数据背景下，各行各业都需要数据支持，如何在浩瀚的数据中获取自己感兴趣的数据，在数据搜索方面，现在的搜索引擎虽然比刚开始有了很大的进步，但对于一些特殊数据搜索或复杂搜索，还不能很好的完成，利用搜索引擎的数据不能满足需求，网络安全，产品调研，都需要数据支持，而网络上没有现成的数据，需要自己手动去搜索、分析、提炼，格式化为满足需求的数据，而利用网络爬虫能自动完成数据获取，汇总的工作，大大提升了工作效率。

1. 利用 python 实现网络爬虫相关技术

■ 1.1 什么是网络爬虫

网络爬虫（又被称为网页蜘蛛，网络机器人），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。它们被广泛用于互联网搜索引擎或其他类似网站，以获取或更新这些网站的内容和检索方式。它们可以自动采集所有其能够访问到的页面内容，以供搜索引擎做进一步处理（分检整理下载的页面），而使得用户能更快的检索到他们需要的信息。

■ 1.2 python 编写网络爬虫的优点

（1）语言简洁，简单易学，使用起来得心应手，编写一个良好的 Python 程序就感觉像是在用英语写文章一样，尽管这个英语的要求非常严格！ Python 的这种伪代码本质是它最大的优点之一。它使你能够专注于解决问题而不是去搞明白语言本身。

（2）使用方便，不需要笨重的 IDE，Python 只需要一个 sublime text 或者是一个文本编辑器，就可以进行大部分中小型应用的开发了。

（3）功能强大的爬虫框架 ScraPy，Scrapy 是一个为了爬取网站数据，提取结构性数据而编写的应用框架。可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。

（4）强大的网络支持库以及 html 解析器，利用网络支持库 requests，编写较少的代码，就可以下载网页。利用网页解析库 BeautifulSoup，可以方便的解析网页各个标签，再结合正则表达式，方便的抓取网页中的内容。

（5）十分擅长做文本处理字符串处理：python 包含了常用的文本处理函数，支持正则表达式，可以方便的处理文本内容。

■ 1.3 爬虫的工作原理

网络爬虫是一个自动获取网页的程序，它为搜索引擎从互联网上下载网页，是搜索引擎的重要组成。从功能上来讲，爬虫一般分为数据采集，处理，储存三个部分。

爬虫的工作原理，爬虫一般从一个或者多个初始 URL 开始，下载网页内容，然后通过搜索或是内容匹配手段（比如正则表达式），获取网页中感兴趣的内容，同时不断从当前页面提取新的 URL，根据网页抓取策略，按一定的顺序放入待抓取 URL 队列中，整个过程循环执行，一直到满足系统相应的停止条件，然后对这些被抓取的数据进行清洗，整理，并建立索引，存入数据库或文件中，最后根据查询需要，从数据库或文件中提取相应的数据，以文本或图表的方式显示出来。

■ 1.4 网页抓取策略

在网络爬虫系统中，待抓取 URL 队列是很重要的一部分，待抓取 URL 队列中的 URL 以什么样的顺序排列也是一个很重要的问题，因为这涉及到先抓取那个页面，后抓取哪个页面。而决定这些 URL 排列顺序的方法，叫做抓取策略。网页的抓取策略可以分为深度优先、广度优先和最佳优先三种：

（1）广度优先搜索策略，其主要思想是，由根节点开始，首先遍历当前层次的搜索，然后才进行下一层的搜索，依次类推逐层的搜索。这种策略多用在主题爬虫上，因为越是与初始 URL 距离近的网页，其具有的主题相关性越大。

（2）深度优先搜索策略，这种策略的主要思想是，从根节点出发找出叶子节点，以此类推。在一个网页中，选择一个超链接，被链接的网页将执行深度优先搜索，形成单独的一条搜索链，当没有其他超链接时，搜索结束。

（3）最佳优先搜索策略，该策略通过计算 URL 描述文本与目标网页的相似度，或者与主题的相关性，根据所设定的阈值选出有效 URL 进行抓取。

■ 1.5 网络爬虫模块

根据网络爬虫的工作原理，设计了一个通用的爬虫框架结构，其结构图如图 1 所示。

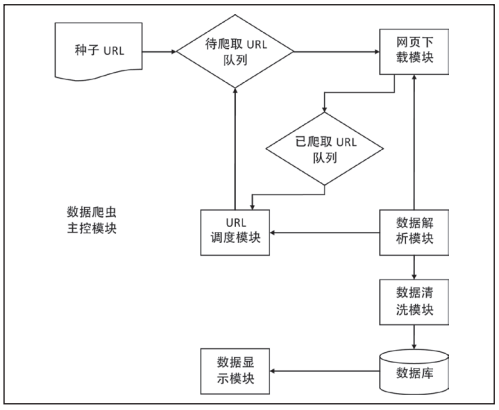

图 1

网络爬虫的基本工作流程如下：

（1）首先选取一部分精心挑选的种子 URL；

（2）将这些 URL 放入待抓取 URL 队列；

（3）从待抓取 URL 队列中取出待抓取在 URL，将 URL 对应的网页下载下来，将下载下来的网页传给数据解析模块，再将这些 URL 放进已抓取 URL 队列。

（4）分析下载模块传过来的网页数据，通过正则表达，提取出感兴趣的数据，将数据传送给数据清洗模块，然后再解析其中的其他 URL，并且将 URL 传给 URL 调度模块。

（5）URL 调度模块接收到数据解析模块传递过来的 URL 数据，首先将这些 URL 数据和已抓取 URL 队列比较，如果是已经抓取的 URL，就丢弃掉，如果是未抓取的 URL，就根据系统的搜索策略，将 URL 放入待抓取 URL 队列。

（6）整个系统在 3-5 步中循环，直到待抓取 URL 队列里所有的 URL 已经完全抓取，或者系统主动停止爬取，循环结束。

（7）整理清洗数据，将数据以规范的格式存入数据库。

（8）根据使用者偏好，将爬取结果从数据库中读出，以文字，图形的方式展示给使用者。

2. 系统模块

整个系统主要有六个模块，爬虫主控模块，网页下载模块，网页解析模块，URL 调度模块，数据清洗模块，数据显示模块。这几个模块之间相互协作，共同完成网络数据抓取的功能。

（1）主控模块，主要是完成一些初始化工作，生成种子 URL，并将这些 URL 放入待爬取 URL 队列，启动网页下载器下载网页，然后解析网页，提取需要的数据和 URL 地址，进入工作循环，控制各个模块工作流程，协调各个模块之间的工作

（2）网页下载模块，主要功能就是下载网页，但其中有几种情况，对于可以匿名访问的网页，可以直接下载，对于需要身份验证的，就需要模拟用户登陆后再进行下载，对于需要数字签名或数字证书才能访问的网站，就需要获取相应证书，加载到程序中，通过验证之后才能下载网页。网络上数据丰富，对于不同的数据，需要不同的下载方式。数据下载完成后，将下载的网页数据传递给网页解析模块，将 URL 地址放入已爬取 URL 队列。

（3）网页解析模块，它的主要功能是从网页中提取满足要求的信息传递给数据清洗模块，提取 URL 地址传递给 URL 调度模块，另外，它还通过正则表达式匹配的方式或直接搜索的方式，来提取满足特定要求的数据，将这些数据传递给数据清洗模块。

（4）URL 调度模块，接收网页解析模块传递来的 URL 地址，然后将这些 URL 地址和已爬取 URL 队列中的 URL 地址比较，如果 URL 存在于已爬取 URL 队列中，就丢弃这些 URL 地址，如果不存在于已爬取 URL 队列中，就按系统采取的网页抓取策略，将 URL 放入待爬取 URL 地址相应的位置。

（5）数据清洗模块，接收网页解析模块传送来的数据，网页解析模块提取的数据，一般是比较杂乱或样式不规范的数据，这就需要对这些数据进行清洗，整理，将这些数据整理为满足一定格式的数据，然后将这些数据存入数据库中。

（6）数据显示模块，根据用户需求，统计数据库中的数据，将统计结果以文本或者图文的方式显示出来，也可以将统计结果存入不同的格式的文件将中（如 word 文档，pdf 文档，或者 excel 文档），永久保存。

3. 结束语

现在已经进入大数据时代，社会各行各业都对数据有需求，对于一些现成的数据，可以通过网络免费获取或者购买，对于一下非现成的数据，就要求编写特定的网络爬虫，自己在网络上去搜索，分析，转换为自己需要的数据，网络爬虫就满足了这个需求，而 python 简单易学，拥有现成的爬虫框架，强大的网络支持库，文本处理库，可以快速的实现满足特定功能的网络爬虫。

参考文献

＊ [1] 于成龙，于洪波．网络爬虫技术研究 [J]．东莞理工学院学报，2011，18(3):25-29.

＊ [2] 李俊丽．基于 Linux 的 python 多线程爬虫程序设计 [J]．计算机与数字工程，2015，43(5):861-863.

＊ [3] 周中华，张惠然，谢江．基于 Python 的新浪微博数据爬虫 [J]．计算机应用，2014，34(11):3131-3134.