基于Python网络爬虫技术的数据采集系统研究

田春荣
（辽河油田信息工程公司集团业务技术中心，辽宁 盘锦 124010）

摘 要：针对数据采集系统的构建问题，本次研究结合我国数据采集系统的发展现状，提出了一种技术Python网络爬虫的数据采集技术，首先对Python网络爬虫技术进行全面的介绍，在此基础上，提出了基于该种方法的数据采集系统构建方法，为推动我国数据采集技术的进一步发展奠定基础。研究表明：Python属于一种面向对象的高级语言，其在计算机科学领域已经开始得到大面积的推广及应用，由于其具有非常强大的第三方库，因此，利用该种技术进行数据采集工作具有非常明显的优势，因此，未来需要加快基于该种技术的数据采集系统构建，进而可以推动数据采集领域的进一步发展。

关键词：Python；网络爬虫技术；数据采集系统；技术介绍；系统构建
【DOI】10.12231/j.issn.1000-8772.2021.01.195

1 前言

随着社会的快速发展，网络中的数据量逐渐的提升，如何从广泛的数据资源中寻找有用的数据资料属于一项重大问题，而网络爬虫技术正是寻找有效数据资料的关键技术，其可以对网络中的数据进行有效的采集，以便工作人员可以对数据资料进行合理的利用，这对于推动我国社会的进一步发展十分有利[1]。针对数据资料的采集问题，本次研究提出了一种基于Python语言的数据爬虫技术，为推动我国数据采集领域的进一步发展奠定基础。

2 Python网络爬虫技术介绍

在人工智能及大数据处理等技术快速发展的前提下，Python迎来了巨大的发展机会，这属于一种面向对象的高级程序，其在计算机科学领域的应用相对较广，同时，还可以对各种类型的数据资料进行全面的处理。在使用该种技术进行网络爬虫的过程中，其主要是提供了多种类型的基本库，主要的第三方库主要包括三种类型，分别是Urlib库、Beautiful Soup库以及Threading库。在Urlib库方面，这属于Python语言中已经内置的基本库，其主要可以用来进行网络请求，通过使用URL的方式，对网络中的数据资源进行全面的爬取，该种类型的基本库又可以分为四个模块，分别是urllib.request模块、urllib.error模块、urllib.parse模块以及urllib.robotparser模块，urllib.request模块的主要作用是通过向浏览器中发送相关的请求，最终得到网站中的数据信息；urllib.error模块的主要功能是对网站的处理异常问题进行全面的爬取，对部分异常问题进行合理的处理；urllib.parse模块的主要功能是对网站中的数据资料进行全面的解析，对部分有用的数据资料进行提取[2]；urllib.robotparser模块的主要功能是对网站上的部分协议进行解析处理，并根据协议的要求，明确数据爬取的目录。在Beautiful Soup库方面，这属于一种相对较为先进的第三方库，其主要是对网站中的数据资料进行全面的解析，使得数据资料呈现出结构化的特征，同时，其还可以根据用户的基本要求，对文档进行合理的修改及查找，该种类型的第三方库可以将网站文档转化为树形结构，与其它类型的库相比，其相对较为简单，应用相对较为高效，其内部含有大量的解析器，可以解析的数据类型相对较多，同时，还可以对部分解析的数据进行处理，最终实现数据输出的功能。在Threading库方面，这属于程序语言中已经内置的库，可以采用多线程的方式对数据资料进行合理的处理，最终实现多线程爬虫的基本功能，其内部含有的模块数量相对较多，例如含有各种类型的函数以及处理对象等，采用多线程的方式对数据进行处理，可以使得网络爬虫的效率得到全面的提升，数据采集的时间缩短，由此可见，采用该种类型的程序语言构建数据采集系统具有非常明显的优势[3]。

3 基于Python网络爬虫技术的数据采集系统研究

（1）设计目标。在进行数据采集的过程中，其主要的目标就是通过构建某种类型的爬虫程序，对网站中的部分数据资料进行有效的采集及处理，需要根据数据结构的不同，将其储存到不同类型的数据库中，对于非结构化的数据而言，其主要是储存在本地的硬盘之中。在使用Python语言进行数据爬取的过程中，首先需要使其鲁棒性得到提升，需要对爬虫程序的各种优势进行充分的考虑。构建的爬虫程序需要具备三方面的要求：①可行性，网络中的数据资源数量相对较多，通用性的爬虫无法对目标数据进行有效的采集，因此，需要对爬虫的目标进行合理的规划，对爬虫的范围进行精确的把控，对部分数据资源进行筛选，最终实现精确爬取的目的；②健壮性，目前爬虫方面的技术发展速度相对较快，反爬虫领域也取得了较大的发展，因此，在进行数据采集的过程中可能会出现众多的异常问题，需要全面提高爬虫的健壮性，防止程序在运行的过程中出现死循环问题；③提高数据采集效率，即使爬虫软件设定了目标，但是由于网络中的数据量过于庞大，单线程的程序仍然无法满足要求，因此，必须引进多线程的技术，进而使得数据采集效率可以得到全面的提升。

（2）采集系统设计。根据数据采集的相关要求，需要将基于Python程序的爬虫软件分为六个模块：①总调度，该模块主要是对整个程序进行全面的管理，这属于数据采集系统的入口，可以对其它模块的运行情况进行调度管理，只有在一个命令完成以后，程序才能进入到下一个命令之中，直到所有的程序都完成工作位置；②管理器，对所有的URL进行合理的管理，其包含的信息相对较多，也属于整个数据采集系统的关键组成部分；③下载器，在网络中找到相关的数据资源以后，需要获取数据资源所在的位置，然后通过调用数据下载功能，对需要的数据资料进行下载，一般情况下，数据资源的类型不同，所使用的下载功能也会存在一定的差别；④解析器，在数据资料下载以后，需要使用该功能对所有的数据进行合理的处理，对噪音进行全面去除；⑤存储模块，其主要的功能就是对处理以后的数据资料进行合理的储存；⑥线程管理，其主要是对线程的数量进行合理的设定，通过线程管理的方式，可以使得数据的采集效率得到全面的提升。

4 结束语

综上所述，在目前网络数据日益庞大的前提下，采用爬虫技术对网络中的有效数据资源进行合理的采集十分关键，由于Python程序内置了大量的先进库资源，使用这些库资源必然可以实现数据采集的基本功能，因此，构建基于该种程序的爬虫技术十分关键，在构建此方面技术的过程中，需要满足可行性、健壮性以及高效率的基本要求，进而使得数据采集效果和采集效率都可以得到全面的提升。

参考文献

[1]严家馨.基于Python对资讯信息的网络爬虫设计[J].黑龙江科技信息，2020(05)：57-58.
[2]魏冬梅，何忠秀，唐建梅.基于Python的Web信息获取方法研究[J].软件导刊，2018(01)：41-43.
[3]楼姗姗.大数据环境下基于python的网络爬虫技术探讨[J].决策探索(中)，2019，33(11)：94.

作者简介：田春荣(1984-)，女，辽宁开原人，中级工程师，信息工程岗(室内)。