基于 Python 的网络爬虫技术在乡村空间规划中的应用研究

白红艳

（宁夏理工学院，宁夏 石嘴山 753000）

摘　要　在开展乡村空间规划过程中，需要对信息采集平台进行合理设计，这样才能够准确地掌握在空间规划过程中的各项信息内容，推动农村信息化的快速发展，这也是当前乡村振兴战略推进过程中的重要举措。在乡村空间规划系统设计过程中，需要从互联网发展技术出发对先进的 Python 网络爬虫技术进行充分应用。基于 Python 网络爬虫技术研发的数据采集系统，可以对主题数据进行自动采集，并且可以完成数据爬取、异常问题处理、robots 协议更新与管理、多线程爬取等不同任务。与传统人工数据采集方式相比，基于 Python 的网络爬虫技术在乡村空间规划中发挥的作用更加突出，能够大大提高乡村空间规划效率。

关键词　Python　网络爬虫　乡村空间规划　信息采集

中图分类号：TP31　　**文献标识码**：A　　**文章编号**：1007-0745(2022)11-0004-03

在互联网技术不断发展的背景下，大数据技术的应用越来越普遍，网络数据量不断增加，如何从海量数据快速获取有效信息，完成空间规划工作是当前乡村空间规划信息化以及现代化发展中面临的巨大挑战。网络爬虫技术的有效应用能够解决这些问题，还可以根据乡村空间规划的具体要求自行定制规则，选取特定的内容，能够精准地获取空间规划过程中的有效信息。此外，网络爬虫技术也可以根据乡村空间规划的实际需求进行自动化运行，进一步保证数据采集系统的现代化水平。

1 Python 与网络爬虫技术概述

1.1 网络爬虫技术

一般情况下，在网页中的信息包含超链接信息、文字信息，从功能上进行分析，爬虫处理主要完成数据采集、处理、存储等不同环节。在开展网络爬虫系统框架设计工作时，必须完成控制器、解析器、资源库等设计。

1.控制器为多线程中不同爬虫线程分配工作任务。

2.解析器完成网页下载，并处理页面进。在爬虫程序运行过程中，解析器主要完成基本工作。

3.资源库主要保存下载的网页资源。目前，不同的网络爬虫管理技术具有相似性。提取有效信息数据后可以对相关信息进行解析并保存数据[1]。

1.2 Python 概述

Python 作为解释型面向对象的动态数据类型高级程序设计语言，在应用过程中可以编译成 .pyc 跨平台自解码文件。该文件的主要优势是能够简单隐藏源码，并且对提高载入速度有积极作用，还可以实现跨平台应用，可以将 Python 与 Java、C++ 语言进行结合封装成 Python 可以调用的扩展库，程序员能够集中精力对程序逻辑进行设计和处理[2]。

2 乡村空间规划中对爬虫技术的应用要求

在对乡村空间规划数据信息采集系统设计过程中，需要以统一的标准和规范为基础整合乡村各类基础空间信息和农村土地流转信息。这样才能够形成县－镇－村不同级别共享的信息数据采集平台。需要以空间规划信息为主进行研发，保证平台采集功能的简单性和实用性。同时要根据不同地区乡村空间规划的实际需求灵活定制。在技术上要具有较强的可行性和超越性；在时间上可以分段实施，满足当前需求的同时，适应未来业务需求[3]。

目前，在乡村空间规划数据信息采集平台建设过程中，需要从以下原则出发，保证该系统平台的建设效果：

1.坚持高效性原则。该数据采集平台的不同数据必须得到有效组织和归类，保证信息查询更新工作顺利完成。同时要防止因为系统投入时间比较长对整体性能产生负面影响。

2.坚持集成性原则。在数据信息采集平台应用过程中，要确保不同业务流程能够顺利衔接，并利用数

据库关联业务、数据交换等不同技术实现数据信息共享和传播。

3. 标准化原则。在数据采集系统建设过程中，需要根据乡村空间规划的具体要求和规范对建设用地信息、空间坐标信息、行政区域编码信息以及元数据标准等进行全面掌握，保证数据采集的准确性和规范性。

3 基于 Python 网络爬虫技术在乡村空间规划中的应用

3.1 基于乡村空间规划信息的采集结构设计

在基于 Python 网络爬虫技术的数据采集系统设计过程中，需要完成分布式爬虫架构，设计网络爬虫，主要是对网络中特定信息进行采取，从而为乡村空间规划提供可靠的数据信息。因此，在具体的设计中需要利用垂直搜索引擎获取数据。为了保证网络爬虫任务能够顺利完成任务，收集更多符合需求的信息。在应用过程中可以从分布式架构方式出发对乡村空间规划过程中的信息数据进行爬取。分布式架构包含分节点的工作状态和监控、URL 的分发工作等不同内容。分节点能够获取主节点发放的工作任务，并且能够根据任务要求开展爬取工作，并将爬取结果及时反馈到主节点。爬虫网络的组成结构如图 1 所示。

在分布式爬虫架构应用过程中，计算机并行处理具有至关重要的作用。利用 Nutch 框架可以完成分布式网络爬虫架构。在乡村空间规划数据爬取过程中，主节点可以将所有带 URL 下载任务的内容分配到不同的分节点完成工作。设计的分布式下载任务在调度过程中能够及时将 URL 映射到服务器进行下载。计算公式如下：

$$Node=hashFun(MD5(URL))\%n$$

在该公式应用过程中：

Node 代表即将分配任务的节点。

hashFun（）表示构造的哈希函数。

n 表示节点的数量。

在分布式网络应用过程中，其具有较强的伸缩性，一旦节点数量出现变化，可以完成数据二次爬取。URL 对集中式分割方法进行应用，可以构建哈希函数完成工作任务。在具体的使用中需要先构造哈希函数，第一次计算获取新的 URL，并将其映射到工作任务表中之后，对工作任务表进行哈希映射，将映射获取的结果划分到分节点 [4]。

3.2 确定设计目标以及要求

在对乡村空间规划数据采集系统设计时，其主要目标是利用编写爬虫程序对乡村空间规划的主题数据进行爬取。需要对爬虫程序的可行性、合规性、效率以及健壮性等问题进行充分考虑。

1. 可行性问题。在海量的互联网信息中获取有效的乡村空间规划数据信息难度比较高。普通用户无法利用通用性爬虫获取相关信息。因此，需要对爬取目标以及主题进行科学规划，定义爬虫范围，并制定网址过滤规则和内容筛选规则，才能够实现乡村空间规划信息的精准爬取。

2. 合规性问题。Robots 协议是国际互联网界的通用道德规范，在协议应用过程中可以告知爬虫程序的具体爬取范围，明确哪些页面能够被抓取，在爬虫程序运行过程中可以自觉遵守 Robots 协议。

3. 效率问题。虽然在乡村空间规划中对基于 Python 的网络爬虫技术进行应用，可以明确空间规划的主题信息，但是与主题相关的 URL 数据量比较大，如果利用单线程结构完成数据采集无法满足数据应用需求，在爬虫程序编写时，需要利用多线程爬虫技术。

4. 健壮性问题。随着反爬取技术的不断发展，有一些网站在应用过程中会设置反爬取策略，这就意味着在爬取过程中会出现不同问题。例如 URL 无法顺利连接、URL 显示不存在、网络不畅通等。为了保证爬虫程序有效应用，并对异常情况科学处理，需要保证爬虫程序的健壮性，防止爬虫程序陷入死循环。

3.3 采集系统模型设计

在采集系统模型设计过程中，主要从以下模块出发，保证系统设计的全面性。

1. 总调度模块。这是乡村空间规划数据采集系统的大脑，对各个模块进行总调度。爬取后返回数据信息，触发下一个任务，一直到完成所有爬虫任务。

2.URL 管理器模块。该模块的主要功能是对所有 URL 进行科学管理，主要包含待爬取 URL、新 URL、URL 的有效判别以及已经爬取的 URL 编码转换等各项工作。

3. 页面下载器模块。该模块在应用过程中需要从 URL 管理器获取 URL 之后调用页面管理器的下载功能，获取相关数据。获取的数据类型主要包含 XML 数据以及 HTML 数据和 JSON 数据。

4. 页面解析器模块。解析器主要是对下载器获取的数据进行解析处理，去除噪声，获取目标数据。

5. 数据存储模块。完成数据解析处理后，获取目标数据，需要调用数据存储模块，将结构化数据直接存储在数据库内，非结构化数据需要存储在本地硬盘，并与数据库建立索引关系。

6. 线程管理模块。在基于 Python 的网络爬虫技术应用过程中，为了保障乡村空间规划信息获取的精准性，可以设定爬取作业的线程数量，利用多线程爬虫提高数据爬取效率。

7.robots 管理器模块。该模块主要完成爬取网站的 robots 协议下载或者更新，并根据具体的协议内容对 URL 管理器进行调用，确定爬取地址的目录结构。

8. 异常处理模块。该模块在应用过程中需要将数据爬取中的所有模块接入异常处理模块。一旦出现异常情况，需要及时触发异常处理进程，并将其写入日志信息库。

3.4 数据采集系统的运行步骤

完成信息采集系统各模块设计工作后，需要在总调度程序的统一调度下开展数据采集工作。具体运行步骤如下：

1. 先确定乡村空间规划的爬取目标并完成数据库创建，之后对总调度模块的不同参数进行初始化配置，输入待爬取网站的入口，明确解析器的参数配置、并发线程数量，之后将其导入网站 robots 启动数据采集工作流程。

2. 总调度模块会先从 URL 管理器中提取 URL 并调用 robots 管理器，对 URL 的目录是否合规进行检查。如果合规可以直接调用页面下载完成数据下载；如果不合规需要调用 URL 管理器请求新的待爬取 URL[5]。

3. 页面下载器完成数据下载后，需要调用页面解析器对数据进行处理。如果存在问题会触发异常处理模块。

4. 完成数据解析处理后，将结果进行科学分类，主要包含主题数据和 URL 数据。如果为 URL 数据，需要利用 URL 管理器方法将其写入待爬取 URL 库；如果为目标数据也就是主题数据，需要调用数据存储模块；如果不为以上两种数据会触发异常处理模块。

5. 在数据存储模块中会写入数据，之后开始下一轮爬取工作，如果不能顺利进行下一步爬取会触发异常处理模块。

3.5 实际应用环节

在本次研究过程中主要利用基于 Python 的网络爬虫技术完成乡村空间规划过程中的数据资源的采集工作，主要对数据资源层进行深入分析。在数据资源层应用过程中，其作为整个平台架构的重要基础，包含数据库、社会经济数据、附件文档等不同信息内容。这些数据都会按照统一的技术规范进行整合处理，并利用分布式存储和管理模式提高数据的应用价值。在数据采集系统运行过程中，基础数据层需要利用标准数据交换格式与服务层完成数据交换，在基于 Python 的网络爬虫数据采集系统应用中，需要对空间基础数据库进行科学设计，并且要明确在乡村空间规划过程中，不同类别基础数据的坐标数据、平台符号库、规范标准等差异，完成异构数据统一，为数据采集奠定有利基础。在异构数据融合统一时，需要按照统一的技术标准进行操作。在平台设计中可以对乡村空间规划现有数据进行科学整理，形成规范范围、比例尺、坐标体系、规划用地分类体系等空间基础数据库。在空间基础数据库应用过程中，可以利用网络爬虫技术对基础地理数据库、现状数据库、空间规划数据库以及社会经济数据库等进行精准的数据采集和提取。还要通过云技术以及虚拟技术建立元数据库，对空间基础数据库利用统一的编码进行数据入库处理。

4 结语

综上所述，在乡村振兴战略的影响下，我国乡村的发展水平在不断提升。乡村振兴战略本身是利国利民的重大工程，也是系统、长期的复杂任务。在大数据时代，为了保证乡村振兴战略顺利实施，需要加强乡村空间规划工作。为了准确提取在乡村空间规划中的各类主题信息数据，需要对基于 Python 的网络爬虫技术进行应用，才能获取更加符合条件的网页信息。在分布式网络爬虫技术应用中，可以大大提高数据采集效率。在乡村空间规划信息采集过程中，可以根据具体的空间规划目标制定主体信息，利用网络爬虫技术精准采集数据信息，提高数据采集效率，保证数据的应用价值。

参考文献：

[1] 张宝刚 . 基于 Python 的网络爬虫与反爬虫技术的研究 [J]. 电子世界 ,2021(04):86-87.
[2] 王碧瑶 . 基于 Python 的网络爬虫技术研究 [J]. 数字技术与应用 ,2017(05):76.
[3] 李彦 . 基于 Python 的网络爬虫技术的研究 [J].2021(03):39-40.
[4] 李培 . 基于 Python 的网络爬虫与反爬虫技术研究 [J]. 计算机与数字工程 ,2019,47(06):1415-1420,1496.
[5] 朱思柱 , 张萌 . 区块链技术在农业农村中的应用与对策研究 [J]. 中国农机化学报 ,2021,42(07):170-176.