基于网络爬虫技术的大数据采集系统设计

白天瑰
（甘肃钢铁职业技术学院　甘肃省嘉峪关市　735100）

摘　要：本文简单介绍了网络爬虫技术，论述了基于网络爬虫技术的大数据采集系统设计目标，探究了基于网络爬虫技术的大数据采集系统设计结构，并对基于网络爬虫技术的大数据采集系统设计实践进行了进一步探究，希望为基于网络爬虫技术的大数据采集系统设计提供一些参考。

关键词：网络爬虫技术；大数据采集系统；URL

在信息技术飞速发展进程中，网络成为汇聚信息的载体，网络平台中信息收集、提取效率在一定程度上影响着社会发展进程。与此同时，网络平台数据量、种类呈爆发式增加，数据多样化、分散性、关联性显著增强，传统数据收集方法无法满足数据分析需求。而网络爬虫技术是依据特定规律自动下载网络信息的计算机程序，可以按照用户需求提取网页信息，发掘信息资源价值。在考虑实际运用需求的情况下，采用网络爬虫技术设计大数据采集系统具有非常突出的意义。

1 网络爬虫技术

1.1 网络爬虫的内涵

网络爬虫又可称之为网络机器人，是一种根据一定规则自动遍历 Web（World Wide Web，全球广域网）的超链接结构，在遍历 Web 过程中，网络爬虫可以完成信息检索与定位 [1]。网站的某个网页是网络爬虫的入口，在网络爬虫读取网页内容过程汇总可自动寻找其他超链接，根据超链接寻找下一个网页，持续进行，直到抓取全部互联网网页。

1.2 网络爬虫的过程

网络爬虫的过程如图 1 所示。

图 1 中，DNS（Domain Name System，域名系统）是服务运营商的服务器地址，URL（Uniform Resource Locator）是统一资源定位符，也可称之为互联网标准资源地址，与互联网中每一文件一一对应，由协议（或服务方式）、资源主机 IP（Internet Protocol）地址（含端口号）、主机资源具体地址（目录、文件名等）组成 [2]。根据获取 IP 地址与访问内容可以封装 HTTP（Hyper text transfer Protocol，超文本传输协议资源），并打出超文本传输请求。进而由服务器接收信息寻找 Web 资源，在成功获得资源后创建超文本传输请求并封装，最终将超文本传输协议资源响应返回到爬虫，完成资源解析与保存。

2 基于网络爬虫技术的大数据采集系统设计目标

2.1 直观性

只有直观的数据才可以分析正确的结论。在直观性原则

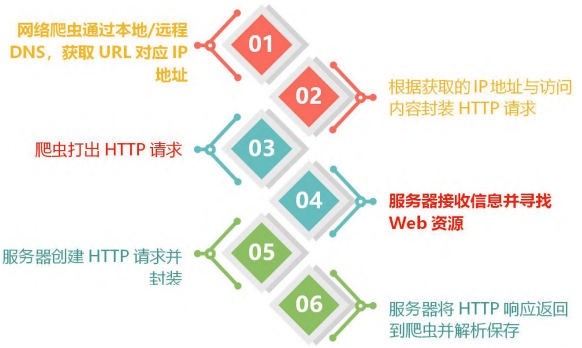

图 1：网络爬虫过程

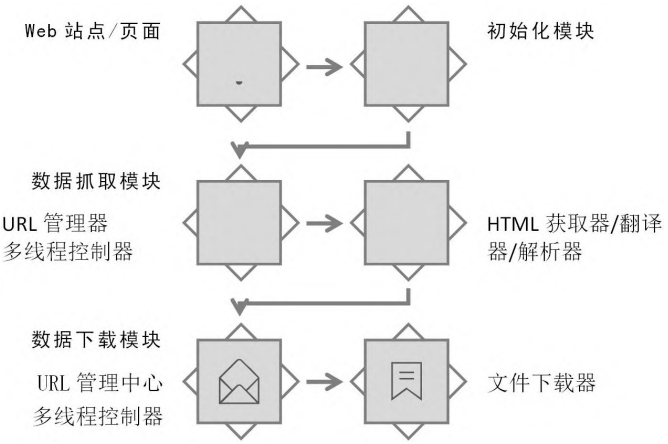

图 2：基于网络爬虫技术的大数据采集系统结构

引导下，设计人员应有意寻找基于网络爬虫的大数据采集系统拓展渠道，用一张图、一张表自然展现繁杂数据。借助图文并茂的方式，促使数据对比、趋势变化变得一目了然，有层次地展示数据，为数据采集结果审查提供依据，规避采集问题累计引发的系统运行成本增加问题。

2.2 时效性

时效性是基于网络爬虫技术的大数据采集系统设计的重要原则，强调系统采集数据与当前项目需求相符，规避过时的低价值或无价值信息。同时，根据业务工作对数据的时效性要求（一个时期工作对时间的要求），及时更新网络爬虫程序，以年、季度、月份甚至星期、天、小时、分钟为单位进行系统实时变更，满足业务开展需求 [2]。

表 1：HTML 解析字典（局部）

中文名称	英文简写	标签内容
姓名	Xm	Table#tab>tbody>tr:eq(0)>table.InfoheaderTabel>xm
居住地址	Jzdz	Table#tab>tbody>tr:eq(1)>table.InfoheaderTabel>Jzdz
联系电话	Ixdh	Table#tab>tbody>tr:eq(0)>table.InfoheaderTabel>Ixdh

2.3 关联性

体量巨大、类别繁多是大数据的特点，而数据信息实用价值发挥的前提是收集相关的数据。因此，在基于网络爬虫技术的大数据采集系统设计时，应贯彻关联性原则，有针对性地摒除不相关的数据，在确保数据信息之间存在因果关系的同时，强化数据信息之间的印证作用，为大数据采集系统的现实应用提供依据。

2.4 全面性

数据采集全面性是做出准确判断的前提。因此，基于网络爬虫技术的大数据采集系统应贯彻全面性原则，覆盖全部与工作项目具有联系的数据、信息，确保内容全面性。同时，自动收集表格、图片等不同形式的业务数据，确保种类全面性。在这个基础上，依据相关制度，对标实际业务，收集足够的规范、案例，满足数据应用要求。

3 基于网络爬虫技术的大数据采集系统设计目标

在管理信息系统爆炸式发展进程中，互联网成为大数据应用的热点，互联网资源开发技术、应用框架日益成熟，基于 Spring 的开源系统、基于 DMDA（Data Model Driven Architecture，数据模型驱动方式）先后出现，为系统实现高效率、低成本开发提供了充足支持。与此同时，管理信息系统与大数据应用整合问题日益凸显，面临着专题信息数据后处理、业务应用数据化支撑双重要求 [3]。基于此，基于网络爬虫技术的大数据采集系统应立足系统业务面局限性，结合信息类型，自动获取意义明确的静态信息、动态时效性信息。其中静态信息是定义类、汇总类、表述类信息，具有权威性；动态信息源于新闻类站点、通用搜索引擎关键词检索栏，具有时效性。

本质上而言，基于网络爬虫技术的大数据采集系统设计目标是抓取网络数据，构建丰富的数据样本库。因网络数据量较大，拟将大数据采集系统划分为数据抓取部分、数据下载部分两个部分。同时，设计多线程控制器，满足硬件环境下程序运行效率最大限度提升要求；依据中文的语义表达方式，设计翻译模块，对提取信息进行翻译处理，解决外文网站信息阅读问题；设计 URL 管理器（统一资源定位符），在 HTML（Hypertext Markup Language，超文本标记语言）获取器下载页面管理 URL，并借助 HTML 获取器进行下载页面解析；设置存储模块，自动存储下载器下载的内容。

4 基于网络爬虫技术的大数据采集系统设计结构

根据基于网络爬虫技术的大数据采集系统设计需求，在直观性、全面性、关联性、时效性原则指导下，设计包含数据抓取、数据下载的系统结构 [4]。具体见图 2。

图 2 中，初始化模块用于主题设定、关键词设定、权威站点设定，内设爬取算法，可自动生成稳定的权威站点队列、关键词队列。数据抓取模块位于权威站点队列框架下，对应特定主题需求，受关键词队列的限制，可以生成爬取信息的基本索引，为索引信息中权威数据爬取提供依据。数据下载模块则针对抓取后信息，根据多线程控制器、URL 管理中心的分析下载所需文件。

5 基于网络爬虫技术的大数据采集系统设计实践

5.1 初始化部分

Web 方式是庞大、繁杂互联网数据呈现的主要形式，确定主题是基于网络爬虫技术的数据采集的首要任务。根据主题关键词精准度对采集准确性、效率的影响，健全系统内部初始化部分，预先向用户提供关键词列表，并在关联信息主题方向明确的前提下设定主题方向语义，根据语义描述进行关键字、关键词的进一步分析 [5]。在初始关键词队列内，主题站点页面分布特性决定了权威站点链接主题相关性。基于此，可以贯彻准确性、全面性原则，在主题语义描述的基础上，将权威站点设立前移到用户设立之初，借助中文分词技术分析语义。进而从权威站点链接出发，进行关键词队列的健全，为主题相关性判定提供依据。同理，根据权威站点链接的站点主题的高度相关性，由初始权威站点出发，经链接爬虫，设置未来数据采集的关键数据源——权威站点队列，为信息获取提供依据。

5.2 数据抓取部分

数据抓取部分运行过程中，先在 URL 管理器输入最开始的 URL，再在 HTML 解析器内解析获取的网页，在解析期间提取新的 URL 并将其添加到新的 URL 列表，判断新的 URL 列表是否满足爬取条件。若无法满足爬取条件，则选择新的 URL 后开始新的抓取循环，若满足爬取条件，则将提取的信息保存到本地，借助翻译工具翻译提取信息，在数据库保存翻译结果，完成数据抓取。

5.2.1 URL 管理中心

URL 管理中心是解决 URL 指向循环（每一网页爬取信息均与其他网页的 URL 相互指向）、URL 重复问题的关键，负责管理待爬取 URL、已爬取 URL[6]。

较之传统的网络爬虫程序，URL 管理中心的重点是保证数据获取准确性、完整性。系统机构可以在网络爬虫阶段

表 2：Web 页面结构化下载数据结构（局部）

中文名称	变量	类型
网页标题	Title	String
网页 URL	urlRedirect	String
URL 长度	urlLength	Int

全部抓取主题 URL 指向的数据（均为有效目标数据）。同时，记录每一个 URL 的日志（含 URL 异常问题）。在抓取工作进入尾声后严格分析异常 URL 的原因，重新开启 1 次爬虫抓取循环。若异常 URL 原因不可控，则在记录 URL 日志的基础上将异常原因反馈给 HTML 解析器。

5.2.2 HTML 解析器

HTML 解析器是基于定向网络爬虫的应用，负责对 HTML 获取器下载页面进行解析后提取页面包含的 URL、定量数据。针对网络爬虫获取的半结构化、非结构化甚至无效数据混合情况，综合引入正则匹配、截取片段、文字替换、智能分段等数据整理挖掘工具。其中正则匹配用于逻辑化处理字符串操作，规则字符串由特定字符组合，特定字符前期已定义；截取片段用于截取描述性语句关键信息，对于一列字符串数据，需进行截取数据起始位置、结束位置的定义，将无效数据舍去；文字替换主要是将分类变量赋值为多种等级值，或者将解析变量数据内存在的无意义数据替换为空；智能分段主要将数值组合表示的变量切割为若干单一变量，根据“/”特殊字符自动分段。HTML 解析字典局部见表 1。

如表 1 所示，针对每一个数据页面，解析完成的数据表现为二维表的结构化形式，具有全面性、变量化、准确性。进而以对比解析数据、原始数据文件的形式，进行爬虫解析准确性验证，根据验证结果及时修改错误的解析字典，重新开始一次数据解析 [7]。

5.2.3 HTML 翻译器

HTML 翻译器是解决数据理解困难问题的关键，负责翻译并注解源于外文网站 HTML 数据，促使其语言、语义与中文语言表达方式相符。HTML 翻译器设计的前提是搜索目标网站数据信息语言、系统所需信息语言存在差异，设计依据是中文语义表达习惯，设计方法是直接调用百度翻译的 API（ApplicationProgrammingInterface，应用程序接口）翻译程序代码，满足实际需求。即：

```
Import requests
Url=’https://fanyi.baidu.com/sug’
Data={‘kw’:’hello’}# 根据需要改变 kw 对应数值
Res=requests.post(url,data=data).json()
Print(res[‘data’][φ][‘v’])
```

在“英译汉”时，样例输入 data={‘kw’:’hello’}，样例输出 :int. 打招呼，你好，表示问候。

5.2.4 多线程控制器

多线程控制器是解决程序运行速率低、运行时间长问题的关键，负责控制爬虫程序线程数，并根据数据要求控制程序，促使程序运行速率达到最高水平。根据程序高速运行要求，以串行网络爬虫为基础，将网页访问请求发送设置到爬虫程序开始环节，促使程序自动等待网页做出响应。根据程序等待网页响应时长，判定爬虫效率高低，程序等待网页响应时长与爬虫效率呈负相关。根据程序等待网页响应时长与爬虫效率之间关系，以进行多线程控制模式的优化。

5.2.5 HTML 获取器

HTML 获取器是整个网络爬虫数据抓取部分的核心，负责从网上查询 URL 对应网页，并在本地下载网页内对应信息，信息格式为 HTML。因多数数据网站要求具有时效性的登录权限，可每间隔 1000 条 HTTP 请求进行一次登录请求，为爬虫程序顺利进行提供依据 [8]。并对每次登录请求返回值进行判断，及时捕获、记录异常的数据并登录验证，规避 session 失效（用户会话与服务器连接的过程中长时间没有动作或连接超过有效时间后，用户会话的 session 数据被清空或回收）的情况。

5.2.6 数据库存储

数据库存储模块是存储 HTML 翻译后数据的载体，可以解决数据读取难题。数据库设计环境是 SQL Server 数据库，开发语言是 java，运行附属区与操作系统分别为 Apache Tomcat Web 服务器、Windows Server 2010 操作系统。

5. 3 数据下载部分

在数据下载流程中，先从数据库内提取数据，再借助 URL 管理器（或多线程控制器）判断是否满足下载条件。若不满足下载条件，则进行 URL（或下载程序）标注，重新进入 URL 管理器（多线程控制器）管理环节；若满足下载条件，则传递给文件下载器，下载数据，结束整个流程。

5.3.1 数据读取

数据读取模块负责读取数据库存储信息、URL。数据读取模块设计需要选择高效率、功能实现便捷、面向对象编程的 java 脚本语言，将源代码转换为字节码后翻译为计算机语言。根据网络爬虫程序复杂、Web 网站群结构层次多、数据量大、目录深度广的特点，在数据读取时应引入网页分析算法，自动滤除与主题联系较小的链接，确保读取链接与主题

表 3：网站数据爬取速度

爬取速度 \ 网站源	时间 /min	线程数 / 个	爬取页面个数 / 个	有效链接个数 / 个	文件个数 / 个	爬取链接个数 / 个
国内网站	24	12	1336	1215	15285	1331
国外网站	14	12	6352	277	3582	277

内容高度相关。进而根据读取策略，从相关队列内选择后续带读取内容，重复读取至达到检索要求。比如，利用 HTML 路径语言，将 HTML 路径语言导入 lxml 的 etree 模块，声明一段 HTML 文本后初始化，获得读取对象。在借助“//”开头的规则读取节点的过程中，文件类型位于读取元素后。此时，可以借助规则内置节点轴选择方法（子元素、兄弟元素等），先调用 ancestor 轴获得祖先节点（含 ul、himl、div、body 等），再选取 div 的祖先节点，逐步读取全部 HTML 信息。

5.3.2 多线程控制器

多线程控制器负责控制读取的下载程序。因线程是独立运行的基本单元与进程执行路径，涵盖了独立堆栈、CPU 寄存器状态，由系统直接调度，一个进程内的若干线程共享地址空间、CPU、对象句柄等信息。多线程控制器内，每一线程具有独立于应用程序其他线程而运行的堆栈，满足一个进程内若干子任务并行执行要求。比如，2 个不同线程并发执行处理数据界面操作、实时数据读取任务。因基于 VB 的多线程不具备窗口类用户界面，可以利用动态链接库 DLL（Delay—locked Loop，延迟锁相环）代替 VB 串口通信控件 MSComm 完成串行通信。即在 VBStandard.EXE 程序内添加通用模块，在通用模块内，声明 Win32API 动态链接库（含创建线程函数），调用创建线程函数建立线程函数地址，根据线程函数地址进行线程体名定义。

5.3.3 文件下载器

文件下载器负责访问数据库内 URL 并下载要求的文件。基本的下载目录为存储路径 / 站点地址 / 结构化数据文件，原始页面结构化数据文化见表 2。

文件下载器下载量适中，可以满足基本管理、统计分析要求。文件下载器应用时，可以通过筛选直接导入下载目录并填充文件字段，对相关页面进行简单处理后下载。比如，基于权威站点关键词，将主题检索信息作为二次下载关键词集，下载信息作为时效性信息，并定义下次更新（或周期性更新）时间，满足未来分析要求。

6 基于网络爬虫技术的大数据采集系统设计效果测试

为确定基于网络爬虫的大数据采集系统应用效果，在 Window11 家庭版软件测试环境内，对爬取速度进行测试。测试 CPU 为 154700M，网络带宽为 100Mb，内存为 8G。得出结果见表 3。

由表 3 可知，基于网络爬虫的大数据采集系统分别爬取国内、国外网站数据的速度具有一定差异，但均满足具体应用需求，爬取大量数据耗费时间资源较少，多线程爬取、下载速率均处于较快的水平，且爬取国外网站数据可以达到中文语言表达规范要求。同时，在线程数一定的情况下，网络带宽对数据下载速度具有一定影响，在线程数达到 12 时，下载速度可以达到 8000kB/s，接近带宽最大速率。

7 结论

综上所述，大数据采集是大数据分析的入口，基于网络爬虫技术进行数据提取、处理、存储，可以根据获取的数据洞察内在知识，为用户决策提供参考。同时，采用网络爬虫技术设计的大数据采集系统可以满足互联网数据快捷获取需求，因此，应借助计算机编程语言编写代码，实现 Web 文本的自动收集，进一步丰富数据样本库，满足数据的高效率搜集、整理与查询应用要求。

参考文献

[1] 刘晓魁．网络爬虫技术与策略分析 [J]．网络安全技术与应用，2022(05)：17-19.
[2] 王辛浩，单艳．探究 Python 语言下网络爬虫的技术特点及应用 [J]．数字技术与应用，2022,40(10)：85-87.
[3] 龙辉．基于 Ajax 的聚焦网络爬虫技术在科研项目管理系统中的应用 [J]．电子元器件与信息技术，2022,6(07)：8-113.
[4] Jianghui WANG,Peng ZHOU,Yichen LIN, 等 .Spatio-temporal Variation of PM_(2.5) in China from 1998 to 2016[J].Meteorological and Environmental Research,2022,13(01)：21-25.
[5] 翟俊杰．社会工作者招聘的岗位特征与区域差异：基于网络爬虫的数据分析 [J]．南京工程学院学报（社会科学版），2022,22(02)：36-436.
[6] 刘业，吴建平．动态可配置网络爬虫系统的形式化研究 [J]．福建电脑，2022,38(08)：1-4.
[7] 张正阳，任保见，刘娜 .Python 网络爬虫在农业网络数据获取中的研究 [J]．现代化农业，2022(07)：50-53.
[8] 燕跃豪，宋建辉，鲍薇，孙晨光，原征．基于网络爬虫的电力营商环境大数据信息检索技术研究 [J]．电气时代，2022(10)：34-36.

作者简介

白天瑰（1987-），女，甘肃省白银市人。大学本科学历，中级职称。研究方向为计算机科学与技术。