# 网络爬虫技术的研究与实现

陈进才
广东省计算技术应用研究所 广东 广州 510000

【摘要】自互联网出现以来，互联网数据经历了爆炸性的增长。我们已经进入了一个数据爆炸的时代。在互联网上搜索信息时，许多人面临不同的问题。在这方面，已经出现了搜索引擎，可以帮助人们找到大量有用的信息，并从有用的信息中找到更可靠的信息。但是，随着在线数据的爆炸式增长，传统搜索引擎已很难满足人们的实际需求。因此，网络爬虫作为搜索引擎主体的作用逐渐凸显了出来。

【关键词】网络爬虫技术；实现；策略

基本上，搜索引擎通常是大型计算机程序。网络爬虫技术是搜索引擎的重要组成部分。搜索引擎从互联网收集了数千个信息丰富的网页，并对网页上的每个单词进行了索引。收集知识渊博的网页是人们构建搜索引擎过程中非常重要的一部分。Web 浏览器是用于收集网页的程序。网络爬虫是搜索引擎中收集信息的部分。搜索引擎对所有索引网页的质量，数量和刷新周期都会影响网络爬虫技术的性能。因此，搜索网络爬虫的研究有着深远的意义。

## 1 网络爬虫的概念及其分类

### 1.1 网络爬虫的概念

网络爬虫是搜索引擎的最基本部分，主要是用于加载网页（也称为网络蜘蛛）的计算机程序或脚本。通常网络爬虫每周都会移至 URL 行，先等待 URL 顺序，然后以特定的 URL 顺序状态加载此页面。浏览网页后，获取新的 URL 并确保输入队列中的输入队列为空并重复直到满足停止爬行的条件，从而遍历整个网络。这是网络爬虫所做的一个过程，称为互联网爬行 [1]。

### 1.2 网络爬虫的分类

网络爬虫有很多分类。通过设计和技术实施，它可以分为4种类型。在实际应用中，一般分为通用网落爬虫，聚焦网络爬虫，增量式网络爬虫以及深层网络爬虫。

1.2.1 通用 网络爬虫

在正常情况下，通用网络爬虫对爬行页面的顺序要求不是很高。他们中的大多数使用并行方法，刷新页面需要花费很长时间。因此，网络爬虫时代正在尝试使用流动技术来减少流动时间。最常用的两个是第一深度和第一宽度。深度是指从低浓度到高参考深度浓度的第一次对高浓度的访问，宽度是指对网页的第一目录的浅层内容进行爬网并向深层目录进行爬网。网络爬虫通常包含投影，Google Crawler 是分布式网络的其余部分，它支持使用并行 I/O 进行并行化，而 URL 是用于存储队列的单独过程。Google Crawler 使用 Page Rank 和其他算法来提高系统性能。

1.2.2 聚焦 网络爬虫

聚焦网络爬虫可以显示与所选项目相关的页面。由于扫描仪的扫描距离短，因此页面将刷新得更快，从而有效地节省了网络资源。聚焦 网络爬虫的爬网策略主要是通过评估页面内容和链接结构来确定的。基于内容分析的策略，更完善的算法主要用于评估与请求页面之间的关系，并基于链接结构分析（主要的 Page Rank 算法）使用爬网策略。之后通过网页解析模块将链接进行分析，有序进行排列 [2]。

1.2.3 增量式 网络爬虫

扩展浏览器基本上会刷新加载的网页并检查刷新的网页，以确保新网页是最新的。增量网络爬虫可以大大减少浪费的网络资源，因为它通常在需要页面时才开始爬网，并且不会多次加载没有变化的页面，但是该算法更复杂且更难以实现。

## 2 采用的关键技术

### 2.1 选取种子集合

如果在总种子集合的选取过程中发生错误，则会导致用户收集的材料中只在所需总材料的极少部分。在这方面使用的方法通常是第一个在数据收集过程中使用手动干预来执行大量数据收集的方法。然后，根据该集合，搜索将完成或委托给整个集合。在此处选择有关性的决定，并且在下一个研究中，研究过程将相应发展。它还可以主动将网站管理员 URL 提交给搜索引擎。如果它在后台运行，则必须在允许的时间内路由系统分发网络交换机。

### 2.2 分布式策略技术

从地域角度来看，这种区别可以分为基于局域网和宽带网络的数据收集方法。顾名思义，局域网是组装在一起的。所有歧管通过高速连接相互连接。它们应通过互联网大范围分布在不同区域中，并且最终组装应在远程完成。在现实生活中，常见的 LAN 方法更为普遍，这主要是根据当事人的搜索习惯造成的。在研究过程中，目标存档要比整个站点受到更多关注。WAN 信息收集过程中使用的生产和运营成本相对较高，并且更适合具有特殊需求的大型企业。根据不同节点的工作方式，网络爬虫分配可以分为三种类型：动态分配，静态分配和独立模式。

### 2.3 网络爬虫技术

2.3.1 基本原理

在实施网络爬虫技术时，系统通常会选择一个包含许多链接的网页。目的是通过使用许多网页上包含的链接从那里获得更多信息。然后将所得的URL放入URL库，在 URL 网络爬虫上读取，并将 URL 读取发送到 DNS 模块。可加载 URL 和可移植 URL 可以在其他 URL 库中使用，以避免重复抓取。如果在搜索过程中发现未扫描的 URL 仍然存在于已扫描的网页上，则可以通过下载模块获得它。如果在库中找到了在搜索过程结束时获得的 URL，则将其过滤掉，如果找不到，则将其恢复。如果无法获得 URL，则将其放入库中也将无法获得 URL，并且所有网络爬虫的获取均已正式进行 [3]。

2.3.2 基本机构

(1) 配置模块：网络爬虫系统功能的执行不能与先前的配置分开。通常，取决于网络爬虫的配置，最大线程数和最大调查深度网络爬虫包括在配置过程中，以便可以有效地执行该功能。

(2)URL 识别模块：在流动过程中，应避免重复使用与 URL 识别模块相同侧的流动，并且应通过连续检查来过滤重复的 URL 传递。

(3)Robots 协议模块：扫描页面时，网络爬虫标识在机器人协议下操作的页面，这些页面无法在访问机器人安全模块和指令的页面上进行扫描。

(4) 网页抓取、解析和存储模块：数据网络爬虫是一个复杂的过程。第一步是扫描网页，加载网页以进行爬网，并使用网络分析引擎分析和清理链接。Web 爬网必须支持和维护网页存储引擎的数据结构。

## 3 网络爬虫技术的实现

### 3.1 插入 URL 列表功能

在实现外接程序URL功能时，应重点注意以下几点：运行 URL 并将精确的拆分转换为运行 URL 格式该功能的实现基本上可以分为两个阶段：显示和分段切割。

### 3.2 生成 URL 列表模块功能

在开发 URL-list 模块期间，需要执行以下操作：根据文件系统的操作和客户端初始化确定 Map 的个数，并进行设置。确定 URL 是否在可接受的范围内，如果不符合规格，则根据间隔范围和权职范围等计算出URL分数。在 Reduce 阶段，板上获得的数据应存储在此文件夹中。这个过程仍然是一个相对独立的部分，使其成为相对独立的 Segement。

### 3.3 网页抓取、解析以及存储

网页抓取模块的实现采取的步骤应保持在最低限度。在抓取阶段，用户可以定义外链接的深度 Reduce 阶段是指不断为用户的网页选择有效的高质量数据。最常用的搜索工具或正则表达式来执行网页分析器的功能。正则表达式可以有效地检索，并且将所需的信息存储在网页中。

### 3.4 URL 去重模块

目前，去重策略可以大致分为 3 种类型。（1）根据启动时使用的数据库，可能会存储网络爬虫可以访问的相关数据；（2）将数据集 URL 放入 hash set 将有助于搜索过程更快，更有效地找到兼容的参数；（3）在计算哈希函数 URL 时，必须使用哈希函数启动引擎 [4]。

## 4 结束语

互联网技术的飞速发展使得搜索引擎越来越重要。网络爬虫是搜索引擎的重要组成部分。相关讨论主要是关于技术相关性和功能网络爬虫的联合实现。现如今，依托于机器的网络爬虫已不能满足用户收集信息以方便实施分布式技术网络爬虫的需求。一旦创建了分布式系统，就可以使用大数据来提高网络爬虫数据收集和存储性能的速度。

## 【参考文献】

[1] 庄礼金 , 戴泽鑫 . 网络爬虫的设计与实现 [J]. 信息技术与信息化 ,2020(12):47-49.

[2] 吴宇鹏 . 分布式网络爬虫技术的研究与实现 [J]. 电脑编程技巧与维护 ,2020(11):9-10+19.

[3] 杨国军 . 基于 Python 的数据爬虫的设计与实现 [J]. 数字技术与应用 ,2020,38(10):153-154+158.

[4] 高文超 , 李浩源 , 徐永康 . 基于网络爬虫的搜索引擎的设计与实现 [J]. 电脑知识与技术 ,2020,16(30):6-9+12.

项目：基于网络爬虫的信息系统价格库 项目编号：（LX2019004）