网络爬虫在搜索引擎应用中的问题及对策

刘沛鹏
广东省计算技术应用研究所 广东 广州 510000

【摘要】互联网和互联网技术的飞速发展为人们提供了巨大的机遇。但是，由于互联网上的信息是复杂且无序的，因此很难充分利用它们。快速，准确地从网上获取最有效的信息是用户需要快速的问题，搜索引擎已经很好地解决了这一问题。搜索引擎主要使用来自许多网络站点的信息来为用户提供他们所需的信息。在搜索引擎中，网络爬虫是所有数据源，起着非常重要的作用。爬虫设计的优缺点直接影响内容的丰富性和搜索引擎中更新的流畅性。因此本文从搜索引擎的概念、网络爬虫在搜索引擎中存在问题以及网络爬虫在搜索引擎应用对策等方面对本课题进行了分析。

【关键词】网络爬虫；搜索引擎应用；问题；对策

在讨论网络爬虫之前，让我们首先看一下计算机“机器人”，电脑机器人，本质是软件程序。该程序依赖计算机和网络以无限循环的方式执行网络操作。例如，搜索引擎，因此，基于用户提供的关键信息自动排除其他信息。网络拓扑由许多节点组成。Web 浏览器根据用户的关键字查找网页链接，并将其与网页的服务端关联，以便用户可以快速进行信息查找。

1 搜索引擎的概念

搜索引擎意味着使用某些计算机程序从互联网收集信息，组织和处理信息，将已处理信息提供给用户的方法。为用户提供搜索服务的系统。收集和整理互联网上的信息资源，并提供信息整理。它可以包含三个部分：信息检索，整理信息和用户查询。搜索引擎是提供信息服务的网站，使用特定程序对互联网上的所有信息进行分类将有助于用户在各互联网上找到所需的信息 [1]。

2 网络爬虫在搜索引擎中存在问题

2.1 在单机网络中的性能问题

在互联网时代，信息量不断增加，大量数据带来了性能问题。在极端的时间范围内从大量数据中提取用户所需信息的能力对搜索引擎网络爬虫来说是一项挑战，该程序的核心是搜索引擎，提高网络爬虫程序的性能是当前需要解决的问题 [2]。

2.2 页面资源下载缓慢

爬虫网络程序从多个 URL 获取信息，下载并定位相关链接。页面加载缓慢是需要解决的问题。另外，最新的爬虫网络搜索技术只能提取纯文本内容。但是对于网页，不能获取网页元素，并最终获得网页内容，这方面的研究还不是很多。

3 网络爬虫在搜索引擎应用对策

使用网络爬虫，需要收集与主题网页并减少不相关网页的下载。将主题网页与通用爬虫进行比较时，需要考虑如何描述和定义元素，如何确定网页内容与主题之间的关系以及确定网页的重要性（包括链接和资源），如何提高爬虫资源的覆盖率。以下四个方面描述了网络爬虫搜索的主要对策。

3.1 选择合适的主题集

在网络爬虫中，必须定义或描述一些元素，这些元素描述集合对科学分类和过滤的方向有很大帮助，主题集的利弊直接影响最终结果。主题可以是多个关键字或自然语言。用户可以根据自己的主题进行详细说明 [3]。

3.2 页面下载策略

3.2.1 采用 gzip/deflate 压缩编码传输

随着信息产业的快速发展，互联网资源的类型也有所不同，资源的类型，质量和运营环境条件也不同。对于网络资源，存在大小差异，在可靠性的运营环境中，可能会注意到更长的网络传输时间和更大的网络数据分组。因此，通过压缩网络传输的数据量可以加快数据传输过程的完成速度。

减少 Web 上可用数据量的最有效方法之一是压缩文件数据。gzip 是一个自由 GNU 文件压缩程序。此方法是使用最广泛的无损软件压缩算法。通过压缩以后，文件可以压缩到原来文件大小的 75% 左右，经过文件压缩以后，可以获得很多益处，比如可以提高网页运行速度，

除此之外，用户的体验也更加良好。超文本传输协议也可以与 Gzip 压缩编码一起使用。它的主要目标是改善和改进 Web 应用程序的性能。gzip 压缩使用更多流量的网站可以改善用户的在线体验。

在万维网络中 gzip 压缩的过程如下：（1）HTTP 将用户请求转发到 Web 服务器。如果用户的要求包括 Accept-Encoding 这样的字符，则意味着它包含 gzip 压缩数据，因此应事先检查服务器配置是否包括 gzip 压缩配置。（2）如果服务器具有 gzip 压缩装置。压缩后直接在浏览器中显示。（3）使用静态文件（HTML，CSS 等）搜索时，它将自动检查服务器文件夹中是否包含压缩文件。（4）如果压缩的文件在缓冲目录中没有进行存储，那么网络服务器就会将这个现象进行反馈，并请求将压缩文件存储在缓冲目录中。（5）缓冲目录是进行文件压缩的主要场所。（6）如果用户需要获取动态文件类型，那么 Web 服务器会根据用户的需求将动态内容提供给用户。这时，缓冲目录中将不会存储压缩文件。

3.2.2 异步非阻塞下载，提升 CPU 利用率

网络爬虫效率实际上会影响网页上的数据。网络爬虫根据 URL 要求发送数据请求，当页面接收到数据并返回时这需要一定时间的间隔，如果网络爬虫在这段时间内未参与其他任务，那就会浪费资源并且网络爬虫工作效率相对较小。相反，如果网络爬虫可以充分利用这个间隔执行任务，它将等待接收数据返回后，就可以大大提高 CPU 资源利用率，这种机制称为非阻塞异步请求。

3.3 提取所需 Web 信息

网络爬虫在采集的起始点开始，对站点的相关信息源，并使用适当的互联网协议自动在每个站点上搜索相关信息。为了更好获取相关信息，搜索引擎主要使用多线程获取 Web 信息。

3.4 URL 搜索策略

有两种查找和选择页面的方法。其中一种是根据遍历表查找的方法，搜索顺序通常是广度优先或深度优先，第二个是跟随“最好”优先的原则对主题进行搜索，广度优先选择是使用最广泛的，其实现的理论也基于互联网的存在。换句话说，选择一个网页并下载以加载相关内容，从而获取到网页，比如 HTML 文件，该文件包含 3 个超链接。网络爬虫会选择以下选项之一在操作过程中进行下载和处理，并处理其他连接，接着深层次处理 URL，相对于深度搜索，广度优先的优势显而易见。（1）由于广度优先首先通过浅层次进行处理，因此无论结构如何复制的分支，都可以返回文档。（2）广度优先原则是可以快速，广泛地搜索高质量页面，而不是浅层次的相关页面；（3）广度优先原则可以是多个网络爬虫共同在一起获取数据。它可以从内部链接扩展到外部链接，从而提供广泛的覆盖范围。深度优先是网络爬虫的另外一种方法，它的性能原理是选择一个浅层链接，并探索该链接下的详细数据，到达链接的末尾时，它将返回到数据并选择一个新的链接起始，直到所有链接都被操作后，搜索结束。

3.5 对链接进行过滤

为了提高检索主题信息的速度和准确性，系统必须确定主题和用户界面的完整 URL 之间的关系，最常见的过滤算法是 EPR。需要增加链接主题的权重，然后输入链接网页主题的权重，这是 EPR 算法的结构 [4]。

3.6 URL 提取策略

网络爬虫请求的页面数据但返回 HTML 代码，用户在浏览器看到的内容就是 HTML 执行和处理后的网页化呈现，网络爬虫只有通过从网页文件获取 URL 超链接并从中提取出来，才算完成整个爬行过程。

4 结束语

总的来说，网络爬虫搜索技术的使用为搜索引擎的发展奠定了良好的基础。但是，随着网络爬虫技术的飞速发展，现代人对搜索引擎的需求日益增长。信息服务一直在朝着个性化和规范化的标准发展。当然，这也对网络爬虫的设计提出了非常严格的新要求。考虑到如何实现页面动态变化和原始搜索统计信息的结合来提高爬取功能的效率，这值得研究。

【参考文献】

[1] 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.

[2] 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.

[3]Kegzipin. 网络爬虫技术原理 [J]. 计算机与网络 ,2018,44(10):38-40.

[4] 饶军 , 华申峰 , 吴晓璐 . 关于互联网视听节目监测中网络爬虫的应用研究 [J]. 江西通信科技 ,2015(03):34-36.

项目：基于网络爬虫的信息系统价格库 项目编号：（LX2019004）