Python应用编程丛书

解析Python网络爬虫：

核心技术、Scrapy框架、分布式爬虫

JIEXI Python WANGLUO PACHONG
HEXIN JISHU Scrapy KUANGJIA FENBUSHI PACHONG

黑马程序员 编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

Python应用编程丛书

解析 Python 网络爬虫：核心技术、Scrapy 框架、分布式爬虫

黑马程序员　编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

网络爬虫是一种按照一定的规则，自动请求万维网网站并提取网络数据的程序或脚本，它可以代替人力进行信息采集，能够自动采集并高效地利用互联网中的数据，在市场的应用需求中占据着重要的位置。

本书以 Windows 为主要平台，系统全面地讲解了 Python 网络爬虫的相关知识。主要内容包括：初识爬虫、爬虫的实现原理和技术、网页请求原理、爬取网页数据、数据解析、并发下载、图像识别与文字处理、存储爬虫数据、初识爬虫框架 Scrapy、Scrapy 终端与核心组件、自动爬取网页的爬虫 CrawSpider、Scrapy-Redis 分布式爬虫。

本书适合作为高等院校计算机相关专业程序设计课程教材，也可作为 Python 网络爬虫的培训教材，以及广大编程开发者的爬虫入门级教材。

图书在版编目（CIP）数据

解析 Python 网络爬虫：核心技术、Scrapy 框架、分布式爬虫 / 黑马程序员编著 . —北京：中国铁道出版社，2018.8
（Python 应用编程丛书）
ISBN 978-7-113-24678-5

Ⅰ. ①解… Ⅱ. ①黑… Ⅲ. ①软件工具 - 程序设计
Ⅳ. ① TP311.561

中国版本图书馆 CIP 数据核字（2018）第 142754 号

书　　名：解析 Python 网络爬虫：核心技术、Scrapy 框架、分布式爬虫
作　　者：黑马程序员　编著

策　　划：秦绪好　翟玉峰　　**读者热线**：（010）63550836
责任编辑：翟玉峰　彭立辉
封面设计：王　哲
封面制作：刘　颖
责任校对：张玉华
责任印制：郭向伟

出版发行：中国铁道出版社（100054，北京市西城区右安门西街 8 号）
网　　址：http://www.tdpress.com/51eds/
印　　刷：中煤（北京）印务有限公司
版　　次：2018 年 8 月第 1 版　2018 年 8 月第 1 次印刷
开　　本：787 mm×1 092 mm　1/16　**印张**：17　**字数**：398 千
书　　号：1 ～ 2 000 册
书　　号：ISBN 978-7-113-24678-5
定　　价：52.00 元

PREFACE

前言

网络爬虫是一种按照一定的规则，自动请求万维网网站并提取网络数据的程序或脚本，它可以代替人力进行信息采集，能够自动采集并高效地利用互联网中的数据，市场的应用需求越来越大。

Python 语言的一个重要领域就是爬虫，通过 Python 编写爬虫简单易学，无须掌握太多底层的知识就可以快速上手，并且能快速地看到成果。对于要往爬虫方向发展的读者而言，学习 Python 爬虫是一项不错的选择。

为什么学习本书

随着大数据时代的到来，万维网成为了大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。基于这种需求，爬虫技术应运而生，并迅速发展成为一门成熟的技术。本书站在初学者的角度，循序渐进地讲解了学习网络爬虫必备的基础知识，以及一些爬虫框架的基本使用方法，以帮助读者掌握爬虫的相关技能，使其能够独立编写自己的 Python 网络爬虫项目，从而胜任 Python 网络爬虫工程师相关岗位的工作。

本书在讲解时，采用需求引入的方式介绍网络爬虫的相关技术，同时针对多种技术进行对比讲解，让读者深刻地理解这些技术的不同之处，以选择适合自己的开发技巧，提高读者的开发兴趣和开发能力。

作为开发人员，要想真正掌握一门技术，离不开多动手练习，所以本书在讲解各知识点的同时，不断地增加案例，最大限度地帮助读者掌握 Python 网络爬虫的核心技术。

如何使用本书

本书基于 Python 3，系统全面地讲解了 Python 网络爬虫的基础知识，全书共分 13 章，具体介绍如下：

第 1、2 章主要带领大家认识网络爬虫，并且掌握爬虫的实现原理。希望读者能明白爬虫具体是怎样爬取网页的，并对爬取过程中产生的一些问题有所了解，后期会对这些问题提供一些合理的解决方案。

第 3~5 章从网页请求的原理入手，详细讲解了爬取和解析网页数据的相关技术，包括 urllib 库的使用、正则表达式、XPath、Beautiful Soup 和 JSONPath，以及封装了这些技术的 Python 模块或库。希望读者在解析网页数据时，可根据具体情况灵活选择合理的技术进行运用。

第 6~8 章主要讲解并发下载、动态网页爬取、图像识别和文字处理等内容。希望读者能够体会到在爬虫中运用多线程和协程的优势，掌握抓取动态网页的一些技巧，并且会处理一些字符格式规范的图像和简单的验证码。

第 9 章主要介绍存储爬虫数据，包括数据存储简介、MongoDB 数据库简介、使用 PyMongo 库存储到数据库等，并结合豆瓣电影的案例，讲解了如何一步步从该网站中爬取、解析、存储电影信息。通过本章的学习，读者将能够简单地操作 MongoDB 数据库，并在以后的工作中灵活运用。

第 10~12 章主要介绍爬虫框架 Scrapy 以及自动爬取网页的爬虫 CrawlSpider 的相关知识，通过对这几章知识的学习，读者可以对 Scrapy 框架有基本认识，为后面 Scrapy 框架的深入学习做好铺垫，同时，也可以掌握 CrawlSpider 类的使用技巧，在工作中具备独当一面的能力。

第 13 章围绕 Scrapy-Redis 分布式爬虫进行讲解，包括 Scrapy-Redis 的完整架构、运作流程、主要组件、基本使用，以及如何搭建 Scrapy-Redis 开发环境等，并结合百度百科的案例运用这些知识点。通过本章的学习，读者可在实际应用中利用分布式爬虫更高效地提取有用的数据。

在学习过程中，读者一定要亲自实践本书中的案例代码。另外，如果读者在理解知识点的过程中遇到困难，建议不要纠结于某个地方，可以先往后学习。通常来讲，通过逐渐深入的学习，前面不懂和疑惑的知识点也就能够理解了。在学习编程的过程中，一定要多动手实践，如果在实践过程中遇到问题，建议多思考，理清思路，认真分析问题发生的原因，并在问题解决后总结出经验。

致　谢

本书的编写和整理工作由传智播客教育科技股份有限公司完成，主要参与人员有吕春林、高美云、刘传梅、王晓娟、毛兆军等。全体人员在近一年的编写过程中付出了很多辛勤的汗水，在此表示衷心的感谢。

意见反馈

尽管我们付出了最大的努力，但书中仍难免会有不妥之处，欢迎各界专家和读者朋友来信提出宝贵意见，我们将不胜感激。在阅读本书时，发现任何问题或有不认同之处可以通过电子邮件与我们取得联系。

请发送电子邮件至：itcast_book@vip.sina.com。

黑马程序员

2018 年 3 月于北京

CONTENTS

目 录

第1章

初识爬虫

学习目标

- ◆了解爬虫产生的背景，能够体会到爬虫的顺势而为。
- ◆知道什么是爬虫。
- ◆了解爬虫的用途，进一步理解网络爬虫的便捷之处。
- ◆熟悉不同维度下网络爬虫的分类。

现阶段，互联网已成为人们搜寻信息的重要来源，人们习惯于利用搜索引擎根据关键字查找自己感兴趣的网站，那么搜索引擎是如何找到这些网站的呢？其实，搜索引擎使用了网络爬虫不停地从互联网爬取网站数据，并将网站镜像保存在本地，从而提供信息检索的功能。

网络爬虫技术经历了相当长时间的发展，用途也越来越广泛，除了各大搜索引擎都在使用爬虫之外，其他公司和个人也可以编写爬虫程序获取自己想要的数据。本章就对爬虫知识进行初步介绍，让大家对爬虫有基本的了解。

1.1 爬虫产生背景

目前的互联网已经迈入大数据时代，通过对海量的数据进行分析，能够产生极大的商业价值。如果需要大量数据，有哪些获取数据的方式？常用的方式有以下几种：

1. 企业产生的数据

企业在生产运营中会产生与自身业务相关的大量数据，例如，百度搜索指数、腾讯公司业绩数据、阿里巴巴集团财务及运营数据、新浪微博微指数等。

大型互联网公司拥有海量用户，有天然的数据积累优势。一些有数据意识的中小型企业，也开始积累自己的数据。

2. 数据平台购买的数据

数据平台是以数据交易为主营业务的平台，例如，数据堂、国云数据市场、贵阳大数据交

易所等数据平台。

在各个数据交易平台上购买各行各业各种类型的数据，根据数据信息、获取难易程度的不同，价格也会有所不同。

3. 政府 / 机构公开的数据

政府和机构也会发布一些公开数据，成为业内权威信息的来源。例如，中华人民共和国国家统计局数据、中国人民银行调查统计、世界银行公开数据、联合国数据、纳斯达克数据、新浪财经美股实时行情等。这些数据通常都是各地政府统计上报，或者由行业内专业的网站、机构等提供。

4. 数据管理咨询公司的数据

数据管理咨询公司为了提供专业的咨询服务，会收集和提供与特定业务相关的数据作为支撑。这些管理咨询公司数量众多，如麦肯锡、埃森哲、尼尔森、艾瑞咨询等。通常，这样的公司都有很庞大的数据团队，一般通过市场调研、问卷调查、固定的样本检测、与各行各业的其他公司合作、专家对话来获取数据，并根据客户需求制定商业解决方案。

5. 爬取的网络数据

如果数据市场上没有需要的数据，或者价格太高不愿意购买，那么可以利用爬虫技术，爬取网站上的数据。

无论是搜索引擎，还是个人或单位获取目标数据，都需要从公开网站上爬取大量数据，在此需求下，爬虫技术应运而生，并迅速发展成为一门成熟的技术。

1.2 爬虫的概念

网络爬虫又称网页蜘蛛、网络机器人，是一种按照一定的规则、自动请求万维网网站并提取网络数据的程序或脚本。

如果说网络像一张网，那么爬虫就是网上的一只小虫子，在网上爬行的过程中遇到了数据，就把它爬取下来。

这里的数据是指互联网上公开的并且可以访问到的网页信息，而不是网站的后台信息（没有权限访问），更不是用户注册的信息（非公开的）。

1.3 爬虫的用途

认识了网络爬虫之后，会产生一个疑问，爬虫具体能做些什么？下面通过一张图来总结网络爬虫的常用功能，如图 1-1 所示。

（1）通过网络爬虫可以代替手工完成很多事情。例如，使用网络爬虫搜集金融领域的数据资源，将金融经济的发展与相关数据进行集中处理，能够为金融领域的各个方面（如经济发展趋势、金融投资、风险分析等）提供“数据平台”。

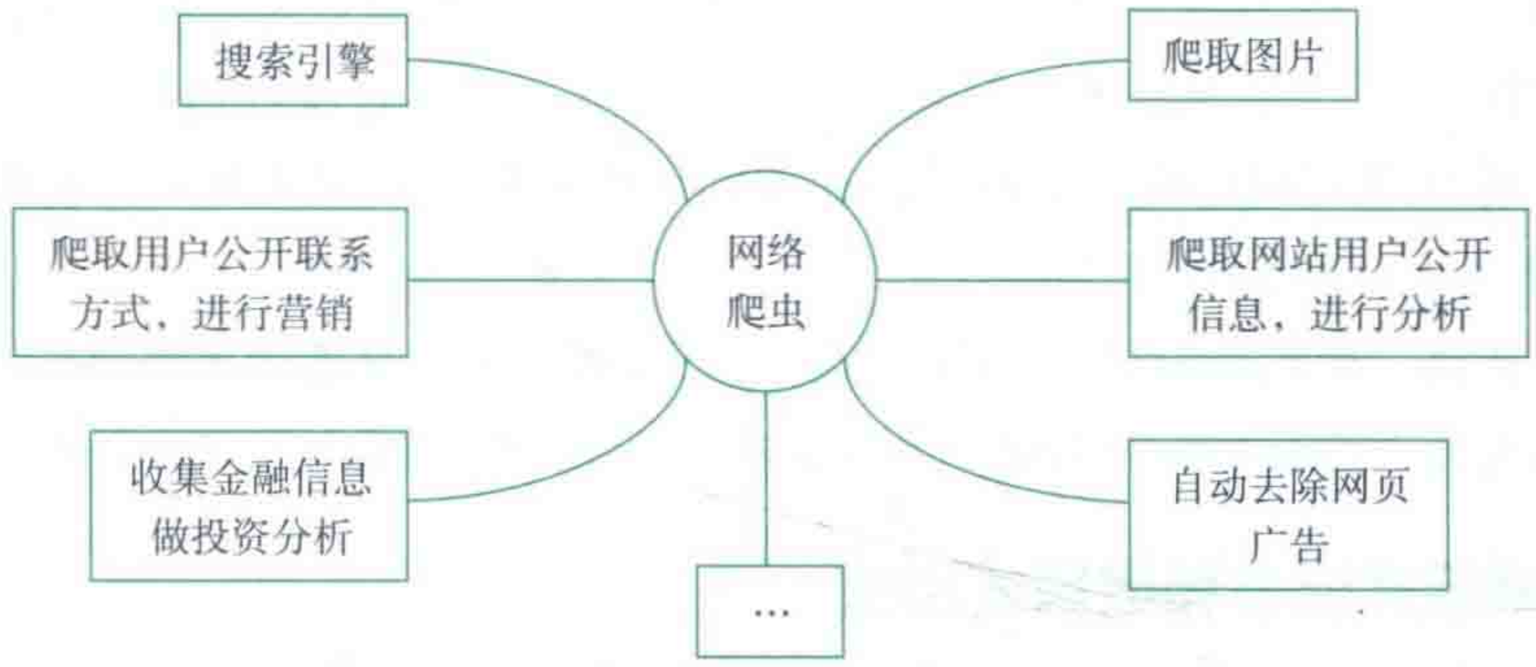

图 1-1 爬虫的常用功能

（2）浏览网页上的信息时，会看到上面有很多广告信息，十分扰人。这时，可以利用网络爬虫将网页上的信息全部爬取下来，自动过滤掉这些广告，便于对信息的阅读。

（3）想从某个网站中购买商品时，需要知道诸如畅销品牌、价格走势等信息。对于非网站管理员而言，手动统计是一个很大的工程。这时，可以利用网络爬虫轻松地采集到这些数据，以便做出进一步的分析。

（4）推销一些理财产品时，需要找到一些目标客户和他们的联系方式。这时，可以利用网络爬虫设置对应的规则，自动从互联网中采集目标用户的联系方式等，以进行营销使用。

总而言之，从互联网中采集信息是一项重要的工作，如果单纯地靠人力进行信息采集，不仅低效烦琐，而且消耗成本高。爬虫的出现在一定程度上代替了手工访问网页，实现自动化采集互联网的数据，从而更高效地利用互联网中的有效信息。

1.4 爬虫的分类

通常可以按照不同的维度对网络爬虫进行分类，例如，按照使用场景，可将爬虫分为通用爬虫和聚焦爬虫；按照爬取形式，可分为累积式爬虫和增量式爬虫；按照爬取数据的存在方式，可分为表层爬虫和深层爬虫。在实际应用中，网络爬虫系统通常是由几种爬虫技术相结合实现的。

1.4.1 通用爬虫和聚焦爬虫

通用爬虫是搜索引擎爬取系统（Baidu、Google、Yahoo 等）的重要组成部分，主要目的是将互联网上的网页下载到本地，形成一个互联网内容的镜像备份。聚焦爬虫，是“面向特定主题需求”的一种网络爬虫程序。

1. 通用爬虫

通用爬虫又称全网爬虫，它将爬取对象从一些种子 URL 扩充到整个网络，主要用途是为门户站点搜索引擎和大型 Web 服务提供商采集数据。

通用爬虫的爬行范围和数量巨大，对于爬行速度和存储空间要求较高，对于爬行页面的顺序要求相对较低。同时，由于待刷新的页面太多，通常采用并行工作方式，但需要较长时间才

能刷新一次页面。

2. 聚焦爬虫

聚焦爬虫又称主题网络爬虫，是指选择性地爬行那些与预先定义好的主题相关的页面的网络爬虫。

与通用爬虫相比，聚焦爬虫只需要爬行与主题相关的页面，从而极大地节省了硬件和网络资源；保存的页面也由于数量少而更新快，可以很好地满足一些特定人群对特定领域信息的需求。

1.4.2 累积式爬虫和增量式爬虫

1. 累积式爬虫

累积式爬虫是指从某一个时间点开始，通过遍历的方式爬取系统所允许存储和处理的所有网页。在理想的软硬件环境下，经过足够的运行时间，采用累积式爬取的策略可以保证爬取到相当规模的网页集合。但由于 Web 数据的动态特性，集合中网页的被爬取时间点是不同的，页面被更新的情况也不同，因此累积式爬取到的网页集合事实上并无法与真实环境中的网络数据保持一致。

2. 增量式爬虫

增量式爬虫是指在具有一定量规模的网络页面集合的基础上，采用更新数据的方式选取已有集合中的过时网页进行爬取，以保证所爬取到的数据与真实网络数据足够接近。进行增量式爬取的前提是，系统已经爬取了足够数量的网络页面，并具有这些页面被爬取的时间信息。

与周期性爬行和刷新页面的网络爬虫相比，增量式爬虫只会在需要时爬行新产生或发生更新的页面，并不重新下载没有发生变化的页面，可有效减少数据下载量，及时更新已爬行的网页，减小时间和空间上的耗费，但是增加了爬行算法的复杂度和实现难度。

面向实际应用环境的网络蜘蛛设计中，通常既包括累积式爬取，也包括增量式爬取。累积式爬取一般用于数据集合的整体建立或大规模更新阶段；而增量式爬取则主要针对数据集合的日常维护与即时更新。

1.4.3 表层爬虫和深层爬虫

Web 页面按存在方式可以分为表层网页和深层网页。针对这两种网页的爬虫分别叫作表层爬虫和深层爬虫。

1. 表层爬虫

爬取表层网页的爬虫叫作表层爬虫。表层网页是指传统搜索引擎可以索引的页面，以超链接可以到达的静态网页为主构成的 Web 页面。

2. 深层爬虫

爬取深层网页的爬虫就叫作深层爬虫。深层网页是那些大部分内容不能通过静态链接获取的、隐藏在搜索表单后的，只有用户提交一些关键词才能获得的 Web 页面。例如，用户注册后内容才可见的网页就属于深层网页。

与表层网页相比，深层网页上的数据爬取更加困难，要采用一定的附加策略才能够自动爬取。

深层爬虫爬行过程中最重要的部分就是表单填写，包含两种类型：

（1）基于领域知识的表单填写：此方法一般会维持一个本体库，通过语义分析来选取合适的关键词填写表单。

（2）基于网页结构分析的表单填写：此方法一般无领域知识或仅有有限的领域知识，将网页表单表示成 DOM 树，从中提取表单各字段的值。

小　结

本章引领大家进入了爬虫的世界，首先讲解了爬虫产生的背景，然后阐述了爬虫的概念，并针对爬虫的用途和分类进行了简要介绍。通过本章的学习，读者能够对爬虫建立初步的认识。

习　题

一、填空题

1. 网络爬虫又称网页蜘蛛或_____。
2. 网络爬虫能够按照一定的_____，自动请求万维网网站并提取网络数据。
3. 根据使用场景的不同，网络爬虫可分为_____和_____两种。
4. 爬虫可以爬取互联网上　　的且可以访问到的网页信息。

二、判断题

1. 爬虫是手动请求万维网网站且提取网页数据的程序。（　　）
2. 爬虫爬取的是网站后台的数据。（　　）
3. 通用爬虫用于将互联网上的网页下载到本地，形成一个互联网内容的镜像备份。（　　）
4. 聚焦爬虫是“面向特定主题需求”的一种网络爬虫程序。（　　）
5. 通用爬虫可以选择性地爬取与预先定义好的主题相关的页面。（　　）

三、简答题

1. 什么是网络爬虫？
2. 简述通用爬虫和聚焦爬虫的区别。
3. 简述使用网络爬虫的优点。

第2章

爬虫的实现原理和技术

学习目标

- 掌握通用爬虫和聚焦爬虫的工作原理，能够理解两者存在的不同。
- 熟悉爬虫爬取网页的流程，为后续框架开发埋下伏笔。
- 了解通用爬虫的网页分类，明确动态爬虫与互联网网页间的关系。
- 了解爬虫要遵守的协议及智能爬取更新网页的文件。
- 熟悉防爬虫的一些应对策略，可以根据实际情况灵活地运用。
- 了解使用 Python 语言做爬虫的优势。

在上一章，我们已经初步认识了网络爬虫，并了解了网络爬虫的应用。本章将分别对通用爬虫和聚焦爬虫的实现原理和相关技术进行介绍，让大家对这两种爬虫有更深入的了解。然后，使用带界面的八爪鱼采集器工具带领大家实现一个简单的爬虫，以加深对聚焦爬虫工作流程的认识。

2.1 爬虫实现原理

不同类型的爬虫，具体的实现原理也不尽相同，但是这些原理之间会存在很多共性。下面就以通用爬虫和聚焦爬虫为例，讲解这两种爬虫是如何工作的。

2.1.1 通用爬虫工作原理

通用爬虫是一个自动提取网页的程序，它为搜索引擎从 Internet 上下载网页，是搜索引擎的重要组成部分。

通用爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在爬取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的停止条件。图 2-1 所示为通用爬虫爬取网页的流程。

通用爬虫从互联网中搜集网页、采集信息，这些网页信息用于为搜索引擎建立索引提供支持，它决定着整个引擎系统的内容是否丰富，信息是否及时，因此其性能的优劣直接影响着搜索引擎的效果。

但是，用于搜索引擎的通用爬虫其爬行行为需要符合一定的规则，遵循一些命令或文件的内容，如标注为 nofollow 的链接，或者 Robots 协议（关于 Robots 协议的详细内容，参见 2.4 节）。

图 2-1　通用爬虫爬取网页流程

多学一招：搜索引擎的工作流程

搜索引擎是通用爬虫的最重要应用领域，也是人们使用网络功能的最强助手。下面介绍搜索引擎的工作流程，其主要包含以下几个步骤。

第一步：爬取网页

搜索引擎使用通用爬虫来爬取网页，其基本工作流程与其他爬虫类似，大致步骤如下：

（1）选取一部分种子 URL，将这些 URL 放入待爬取的 URL 队列。

（2）取出待爬取的 URL，解析 DNS 得到主机的 IP，并将 URL 对应的网页下载下来，存储至已下载的网页库中，并将这些 URL 放进已爬取的 URL 队列。

（3）分析已爬取 URL 队列中的 URL，分析其中的其他 URL，并且将 URL 放入待爬取的 URL 队列，从而进入下一个循环。

那么，搜索引擎如何获取一个新网站的 URL？

（1）新网站向搜索引擎主动提交网址（如百度 http://zhanzhang.baidu.com/linksubmit/url）。

（2）在其他网站上设置新网站外链（尽可能处于搜索引擎爬虫爬取范围）。

（3）搜索引擎和 DNS 解析服务商（如 DNSPod 等）合作，新网站域名将被迅速爬取。

第二步：数据存储

搜索引擎通过爬虫爬取到网页后，将数据存入原始页面数据库。其中的页面数据与用户浏览器得到的 HTML 是完全一样的。

搜索引擎蜘蛛在爬取页面时，也做一定的重复内容检测，一旦遇到访问权重很低的网站上有大量抄袭、采集或者复制的内容，很可能就不再爬行。

第三步：预处理

搜索引擎将爬虫爬取回来的页面，进行各种预处理，包括：提取文字、中文分词、消除噪声（如版权声明文字、导航条、广告等）、索引处理、链接关系计算、特殊文件处理……

除了 HTML 文件外，搜索引擎通常还能爬取和索引以文字为基础的多种文件类型，如 PDF、Word、WPS、XLS、PPT、TXT 文件等。在搜索结果中经常会看到这些文件类型。

但搜索引擎还不能处理图片、视频、Flash 这类非文字内容，也不能执行程序。

第四步：提供检索服务，网站排名

搜索引擎在对信息进行组织和处理后，为用户提供关键字检索服务，将用户检索的相关信

息展示给用户。

同时会根据页面的 PageRank 值（链接的访问量排名）来进行网站排名，这样 PagePank 值高的网站在搜索结果中排名会靠前。当然，也可以直接付费购买搜索引擎网站排名，付费购买排名是搜索引擎公司的盈利手段之一。

图 2-2 所示为搜索引擎的工作原理和主要组成部分。

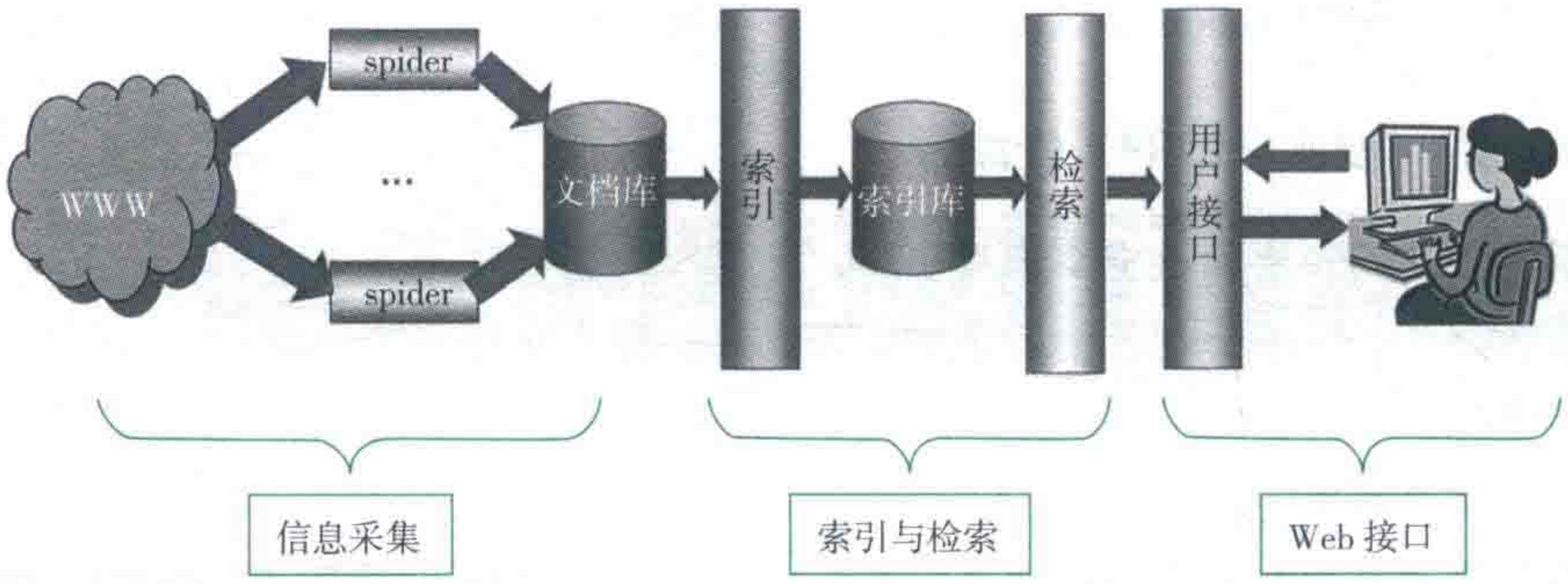

图 2-2 搜索引擎的工作原理和主要组成部分

2.1.2 聚焦爬虫工作原理

与通用爬虫相比，聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接，并将其放入等待爬取的 URL 队列。然后，它将根据一定的搜索策略从队列中选择下一步要爬取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止，如图 2-3 所示。

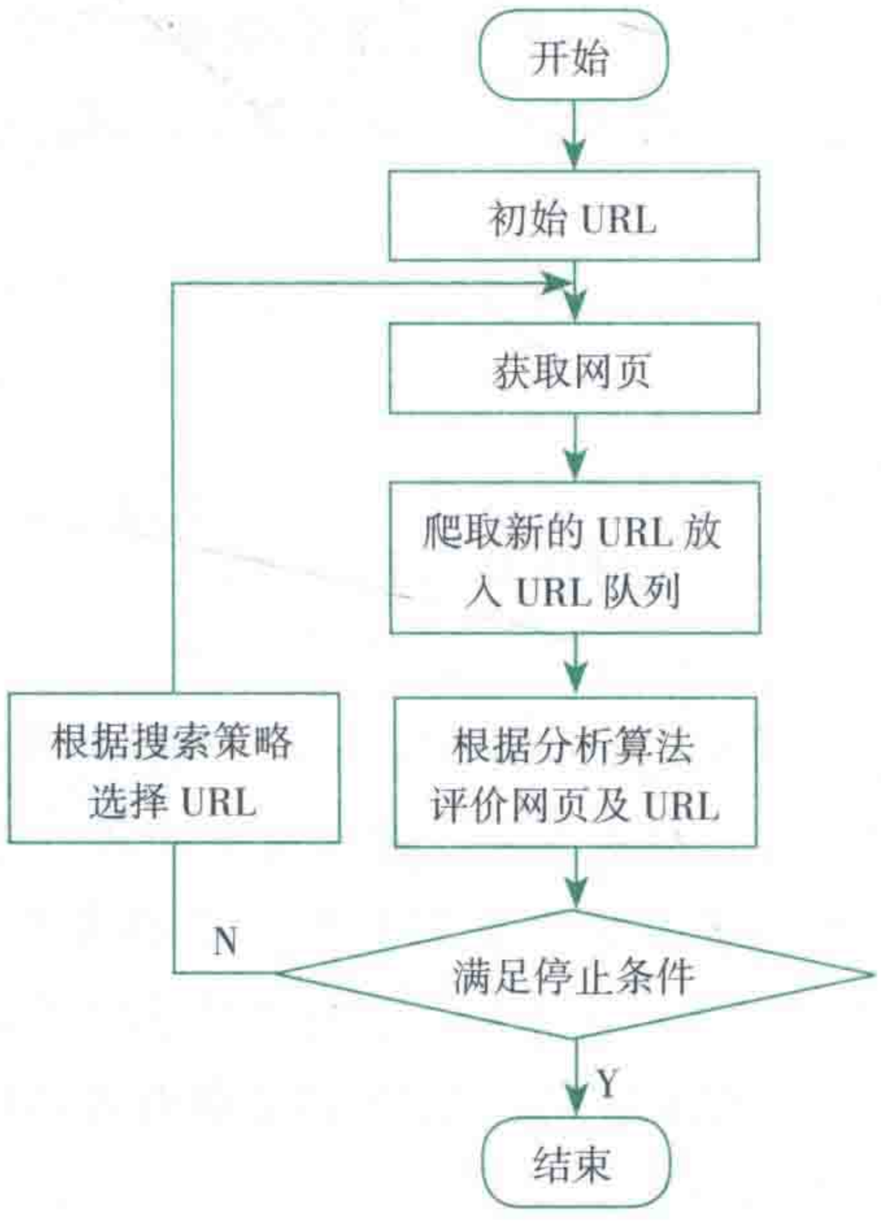

图 2-3 聚焦爬虫工作原理