

初识网络爬虫

在这个大数据的时代里,网络信息量变得越来越大、越来越多,此时如果通过人工的方式筛选自 己所感兴趣的信息是一件很麻烦的事情,爬虫技术便可以自动高效地获取互联网中的指定信息,因此 网络爬虫在互联网中的地位变得越来越重要。

本章将介绍什么是网络爬虫?网络爬虫都有哪些分类、网络爬虫的基本原理以及爬虫环境的搭建工作。

1.1 网络爬虫概述

网络爬虫(又被称为网络蜘蛛、网络机器人,在某社区中经常被称为网页追逐者),可以按照指定的规则(网络爬虫的算法)自动浏览或抓取网络中的信息,通过 Python 可以很轻松地编写爬虫程序或者是脚本。

在生活中网络爬虫经常出现,搜索引擎就离不开网络爬虫。例如,百度搜索引擎的爬虫名字叫作 百度蜘蛛(Baiduspider)。百度蜘蛛,是百度搜索引擎的一个自动程序。它每天都会在海量的互联网信 息中进行爬取,收集并整理互联网上的网页、图片视频等信息。然后当用户在百度搜索引擎中输入对 应的关键词时,百度将从收集的网络信息中找出相关的内容,按照一定的顺序将信息展现给用户。百 度蜘蛛在工作的过程中,搜索引擎会构建一个调度程序,来调度百度蜘蛛的工作,这些调度程序都是 需要使用一定算法来实现的,采用不同的算法,爬虫的工作效率也会有所不同,爬取的结果也会有所 差异。所以,在学习爬虫时不仅需要了解爬虫的实现过程,还需要了解一些常见的爬虫算法。在特定 的情况下,还需要开发者自己制定相应的算法。

1.2 网络爬虫的分类

网络爬虫按照实现的技术和结构可以分为通用网络爬虫、聚焦网络爬虫、增量式网络爬虫。在实际的网络爬虫中,通常是这几类爬虫的组合体,下面分别介绍。

1. 通用网络爬虫

通用网络爬虫又叫作全网爬虫(Scalable Web Crawler),通用网络爬虫的爬行范围和数量巨大,正

是由于其爬取的数据是海量数据,所以对于爬行速度和存储空间要求较高。通用网络爬虫在爬行页面 的顺序要求上相对较低,同时由于待刷新的页面太多,通常采用并行工作方式,所以需要较长时间才 可以刷新一次页面。所以存在着一定的缺陷,这种网络爬虫主要应用于大型搜索引擎中,有着非常高 的应用价值。通用网络爬虫主要由初始 URL 集合、URL 队列、页面爬行模块、页面分析模块、页面数 据库、链接过滤模块等构成。

2. 聚焦网络爬虫

聚焦网络爬虫(Focused Crawler)也叫主题网络爬虫(Topical Crawler),是指按照预先定义好的主题,有选择的进行相关网页爬取的一种爬虫。它和通用网络爬虫相比,不会将目标资源定位在整个互联网中,而是将爬取的目标网页定位在与主题相关的页面中。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快了,聚焦网络爬虫主要应用在对特定信息的爬取,为某一类特定的人群提供服务。

3. 增量式网络爬虫

增量式网络爬虫(Incremental Web Crawler),所谓增量式,对应着增量式更新。增量式更新指的 是在更新时只更新改变的地方,而未改变的地方则不更新。所以增量式网络爬虫,在爬取网页时,只 会在需要的时候爬行新产生或发生更新的页面,对于没有发生变化的页面,则不会爬取。这样可有效 减少数据下载量,减小时间和空间上的耗费,但是在爬行算法上增加了一些难度。

1.3 网络爬虫的基本原理

一个通用网络爬虫的基本工作流程,如图 1.1 所示。



图 1.1 通用网络爬虫的基本工作流程

网络爬虫的基本工作流程如下。

(1) 获取初始的 URL,该 URL 地址是用户自己制定的初始爬取的网页。

(2) 爬取对应 URL 地址的网页时,获取新的 URL 地址。

(3) 将新的 URL 地址放入 URL 队列。

(4)从 URL 队列中读取新的 URL,然后依据新的 URL 爬取网页,同时从新的网页中获取新的 URL 地址,重复上述的爬取过程。

(5)设置停止条件,如果没有设置停止条件,那么爬虫会一直爬取下去,直到无法获取新的URL 地址为止。设置了停止条件后,爬虫将会在满足停止条件时停止爬取。

1.4 搭建开发环境

1.4.1 安装 Anaconda

Anaconda 是一个完全免费的大规模数据处理、预测分析和科学计算工具。该工具中不仅集成了 Python 解析器,还有很多用于数据处理和科学计算的第三方模块,其中也包含许多网络爬虫所需要使 用的模块,如 requests 模块、Beautiful Soup 模块、lxml 模块等。

在 Windows 系统下的浏览器中打开 Anaconda 的官方地址(https://www.anaconda.com/distribution/) 下载对应的安装文件,如图 1.2 所示。



图 1.2 下载 Anaconda

这里笔者所选择的是 Windows (64-Bit Graphical Installer 为当时的最新版本),下载完成后直接双 击运行下载的文件,在 Welcome to Anaconda3 (自己下载的版本)窗口中直接单击 Next 按钮,如 图 1.3 所示。

在 License Agreement 窗口中直接单击 I Agree 按钮,如图 1.4 所示。

Anaconda3 2019.10 (64-bit) Setup — 🗆 🗙		Anaconda3 2019.10 (64-bit) Setup — 🗆 🗙
	Welcome to Anaconda3 2019.10 (64-bit) Setup	License Agreement Please review the license terms before installing Anaconda3 2019.10 (64-bit).
O ANACONDA.	Setup will guide you through the installation of Anaconda3 2019.10 (64-bit). It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.	Press Page Down to see the rest of the agreement. Anaconda End User License Agreement Copyright 2015, Anaconda, Inc.
Ĥ	Click Next to continue. 白击 Next 按钮	All rights reserved under the 3-clause BSD License: Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to instal 单击 I Agree 按钮
	Next > Cancel	< <u>Back</u> [<u>I</u> <u>Agree</u>] Cancel

图 1.3 Welcome to Anaconda3 窗口

图 1.4 License Agreement 窗口

在 Select Installation Type 窗口内选中 All Users(requires admin privileges)单选按钮, 然后单击 Next 按钮, 如图 1.5 所示。

在 Choose Install Location 窗口中选择自己的安装路径(建议不要使用中文路径),这里笔者选择一个自定义的安装路径,然后单击 Next 按钮,如图 1.6 所示。

Anaconda3 2019.10 (64-bit) Setup – 🗆 🗙	Anaconda3 2019.10 (64-bit) Setup — 🗆 🗙
Select Installation Type Please select the type of installation you would like to perform for Anaconda3 2019.10 (64-bit).	Choose Install Location Choose the folder in which to install Anaconda3 2019.10 (64-bit).
Install for: ① Just Me (recommended) ④ All Users (requires admin privileges) ④ All Users (requires admin privileges)	Setup will install Anaconda3 2019.10 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue. Destination Folder G:\Python\Anaconda\ Browse
● 单击 Next 按钮 Anaconda, Inc	Space required: 2.9GB Space available: 64.2GB ② 单击 Next 按钮 Anaconda, Inc.

图 1.5 选中 All Users(requires admin privileges)单选按钮

图 1.6 选择安装路径

在 Advanced Installation Options 窗口中,选中第一个复选框,将 Anaconda 加入环境变量,然后单击 Install 按钮进行安装,如图 1.7 所示。

由于 Anaconda 中包含的模块较多,所以在安装过程中需要等待的时间较长,安装进度如图 1.8 所示。

Anaconda3 2019.10 (64-bit) Setup — 🗆 🗙	Anaconda3 2019.10 (64-bit) Setup — 🗆 🗙
Advanced Installation Options Customize how Anaconda integrates with Windows	ANACONDA Installing Please wait while Anaconda3 2019.10 (64-bit) is being installed.
●选中复选框添加环境变量 ✓ Advanced Options ✓ Add Anaconda to the system <u>PATH environment variable</u> Not recommended. Instead, open Anaconda with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda. ✓ Register Anaconda as the system Python 3.7 This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.7 on the system.	Setting up the base environment
❷ 单击 Install 按钮	Anaconda, Inc

图 1.7 将 Anaconda 加入环境变量

图 1.8 安装进度

安装进度完成以后,将进入 Installation Complete 窗口中,在该窗口中直接单击 Next 按钮,如 图 1.9 所示。

由于 Anaconda 与 JetBrains 为合作关系,所以官方推荐使用 PyCharm 开发工具,在该窗口中直接 单击 Next 按钮,如图 1.10 所示。

📄 Anaconda3 2019.10 (64-bit) Setup 🛛 🚽 🖂 🛛	Anaconda3 2019.10 (64-bit) Setup — 🗆 🗙
ANACONDA Installation Complete Setup was completed successfully.	Anaconda3 2019.10 (64-bit) Anaconda + JetBrains
Completed Show <u>d</u> etails Anaconda, Inc. etails Anaconda, Inc.	Anaconda and JetBrains are working together to bring you Anaconda-powered environments tightly integrated in the PyCharm IDE. PyCharm for Anaconda is available at: https://www.anaconda.com/pycharm

图 1.9 安装完成

图 1.10 PyCharm 开发工具提示

最后在"Thanks for installing Anaconda3!"窗口中根据个人需求,选中或取消选中(笔者选择取消选中)两个复选框,再单击 Finish 按钮,如图 1.11 所示。

将 Anaconda 安装完成以后并保证已经添加系统环境变量的情况下,打开"命令提示符"窗口,然 后输入"conda list"后按 Enter 键,即可查看当前 Anaconda 已经安装好的所有模块,如图 1.12 所示。





1.4.2 PyCharm 的下载与安装

PyCharm 是由 JetBrains 公司开发的 Python 集成开发环境,由于其具有智能代码编辑器,可实现自动代码格式化、代码完成、智能提示、重构、单元测试、自动导入和一键代码导航等功能,目前已成为 Python 专业开发人员和初学者使用的有力工具。

打开 PyCharm 官网的下载地址(https://www.jetbrains.com/pycharm/download/), 然后选择下载 PyCharm 的操作系统平台为 Windows, 单击开始下载社区版 PyCharm (Community), 如图 1.13 所示。



图 1.13 PyCharm 环境与版本下载选择页面

双击 PyCharm 安装包进行安装,在欢迎界面单击 Next 按钮进入软件安装路径设置界面,如图 1.14 所示。

图 1.11 安装结束

在 Choose Install Location 窗口中选择一个需要安装的路径,这里不建议将安装路径设置在默认的 C 盘中,笔者选择自定义安装路径,确认安装路径后单击 Next 按钮,如图 1.15 所示。

🖺 PyCharm Community Edition Setup 🛛 — 🗌 🗙		🛱 PyCharm Community Edition Setup 🛛 🚽 🗙
PC	Welcome to PyCharm Community Edition Setup	Choose Install Location Choose the folder in which to install PyCharm Community Edition.
	Setup will guide you through the installation of PyCharm Community Edition. It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer. Click Next to continue.	Setup will install PyCharm Community Edition in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue. ① 单击该按钮选择安装路径 ② 确认安装路径 G:\Python\PyCharm Community Edition 2019.3.3 Browse
Ē	单击 Next 按钮	Space required: 669.3 Ma Space available: 61.7 GB 单击 Next 按钮 < <u>Back</u> Next > Cancel

图 1.14 PyCharm 欢迎界面

图 1.15 设置 PyCharm 安装路径

在 Installation Options 窗口中首先在桌面快捷方式(Create Desktop Shortcut)中设置 PyCharm 程序的 快捷方式,笔者系统为 64 位,所以选中 64-bit launcher 复选框,然后设置关联文件(Create Associations), 选中".py"复选框,这样以后再打开.py(.py 文件是 Python 脚本文件,接下来编写的很多程序都是后 缀名为.py 的文件)文件时,会默认调用 PyCharm 打开,如图 1.16 所示。

在 Choose Start Menu Folder 窗口中直接单击 Install 按钮,如图 1.17 所示。

🖉 PyCharm Community Edition Setup 🦳 🗆 🗙	🖻 PyCharm Community Edition Setup 🛛 🚽 🗙
PC Installation Options Configure your PyCharm Community Edition installation	Choose Start Menu Folder Choose a Start Menu folder for the PyCharm Community Edition shortcuts.
● 选中 64-bit launcher 复选框 Create pesktop Shortcut Upuate PATH variable (restart needed) G64-bit launcher dir to the PATH	Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.
Update context menu Add "Open Folder as Project"	Accessibility Accessories Administrative Tools Anaconda3 (64-bit)
Create Associations ☑ ☑ ④ 选中".py"复选框 ③ 单击 Next 按钮	Java Development Kit Maintenance Microsoft Office 2013 Microsoft Office 2016 工具 MySQL PremiumSoft PyQt CPL v5.6 for Python 单击 Install 按钮
< gack Next > Cancel	< Back Install Cancel

图 1.16 设置快捷方式和关联

图 1.17 选择开始菜单文件夹窗口

安装进度完成以后,在 Completing PyCharm Community Edition Setup 窗口中,在不直接运行 PyCharm 开发工具的情况下,单击 Finish 按钮即可,如图 1.18 所示。

🖻 PyCharm Community Edition Setup 🛛 🚽 🕹				
Completing PyCharm Community Edition Setup				
PyCharm Community Edition has been installed on your computer.				
Click Finish to close Setup.				
Run PyCharm Community Edition				
取消选中复选框不运行				
单击 Finish 按钮				
< <u>B</u> ack <u>Einish</u> Cancel				

图 1.18 完成安装

1.4.3 配置 PyCharm

双击 PyCharm 桌面快捷方式,启动 PyCharm 程序。选择是否导入开发环境配置文件,这里选择不导入,单击 OK 按钮,进入阅读协议页,如图 1.19 所示。

🖺 Import PyCharm Se	如果之前使用过 PyCharm,可 配置文件,快速设置 PyCharm	以导入之前的 开发环境 ×
O Previous version	C:\Users\Administrator\.PyCharmCE2	019.2\config ~
	不导入环境配置文件	单击 OK 按钮
Do not import set	tings	ОК

图 1.19 环境配置文件窗口

在 Set UI theme 窗口中可以根据个人需求选择开发工具的主题样式,笔者这里选中 Light,使用白色的主题颜色,然后单击 Next:Featured plugins 按钮,如图 1.20 所示。

在 Download featured plugins 窗口中,直接单击 Start using PyCharm 按钮,如图 1.21 所示,此时程 序将进入欢迎界面。

进入 PyCharm 欢迎页,单击 Create New Project,创建一个新工程文件,如图 1.22 所示。

在 New Project 窗口中,首先选择工程文件保存的路径,然后单击 Create 按钮,如图 1.23 所示。



Python 网络爬虫从入门到精通

图 1.21 下载特色插件





1: Proj

Import Settings. Export Settings... Settings Repository.. Save All

G Reload All from Disk

Add to Favorites

Associate with File Typ

File Encoding Remove BOM

Print...

-2

Invalidate Caches / Restart.. Export to HTML ...

Ctrl+S

Ctrl+Alt+Y

第1章 初识网络爬虫

在 Settings 窗口中依次选择 Project:demo (demo 为自己编写的工程名称) → Project Interpreter, 然 后在右侧的下拉列表中选择 Show All...,将打开 Project Interpreters 窗口,如图 1.25 所示。

图 1.24 打开设置窗口

Show tips on startup

×

Previous Tip

Close

C Event Log

- Sectings	6 单击下打	☆列表		^
Q•	Project:	T J J L Pr	🖻 For current project	
Appearance & Behavior	Project Interpreter:	Python 3.7 (demo)	G:\Pvthon\demo\venv\Scripts\pvthon.exe	/ 🌣
Appearance	<n(< td=""><td>o interpreter></td><td></td><td></td></n(<>	o interpreter>		
Menus and Toolbars	Package 👫 F	ython 3.7 (demo)	G:\Python\demo\venv\Scripts\python.exe	+
> System Settings	pip	• 11		1
File Colors	setuptools	40.8.0		
Scopes 💿				0
Notifications	4 选	择 Show All.		
Quick Lists				
Path Variables				
Keymap Editor ① 选择 Projec Plugins	t:demo	_		
▼ Version Control 2 选择	Project Interpreter			
Project Structure				
Build Execution Deployment				
Languages & Frameworks				

Python 网络爬虫从入门到精通

图 1.25 进入设置窗口

在 Project Interpreters 窗口中,单击右侧的"+"按钮,如图 1.26 所示。

Project Interpreters	×
Python 3.7 (demo) G:\Python\demo\venv\Scripts\python.exe	+
单击该按钮 ——	Ţ
	>>
ОК Са	ncel

图 1.26 单击按钮

在 Add Python Interpreter 窗口中,首先单击左侧的 System Interpreter 选项,然后在右侧的下拉列表中选择 Anaconda 中的 python.exe,最后单击 OK 按钮,如图 1.27 所示。

Add Python Interpreter	×
🖶 Virtualenv Environment	Interpreter: G:\Python\Anaconda\python.exe ···
 Conda Environment System Interpreter 	❷ 选择 Anaconda 中的 python.exe
Pipenv Environment	T System Interpreter 选项
	● 单击 OK 按钮 OK Cancel

图 1.27 添加 Python 编译器

12

返回 Project Interpreters 窗口后,选择新添加的 Anaconda 中的 python.exe 编译器, 然后单击 OK 按钮, 如图 1.28 所示。



图 1.28 选择 Anaconda 中的 Python 编译器

返回 Settings 窗口,此时窗口中将自动显示出 Anaconda 内已经安装的所有 Python 模块,然后单击 OK 按钮,如图 1.29 所示。

Settings	● 确认 Anac	onda 中的 python e	×
Q*	Project	conda (1.11) pytholi.e.	rent project Rese
✓ Appearance & Behavior	Project Interpreter: 0	Python 3.7 G:\Python\Anaco	onda\python.exe 🗸 🗘
Appearance			
Menus and Toolbars	Package	Version	Latest version +
> System Settings	_ipyw_jlab_nb_ext_conf	0.1.0	0.1.0 –
File Colors 🛛 🖻	alabaster	0.7.12	0.7.12
Scopes 💿	anaconda	2019.10	2019.10
Notifications	anaconda-client	1.7.2	1.7.2
Quick Lists	anaconda-navigator	1.9.7	1.9.7
Path Variab Anaconda 内的	anaconda-project	0.8.3	▲ 0.8.4
Keymap python 模块	asn1crypto	1.0.1	▲ 1.3.0
> Editor	astroid	2.3.1	▲ 2.3.3
Plusing	astropy	3.2.1	4 .0
Numine Control	atomicwrites	1.3.0	1.3.0
	attrs	19.2.0	▲ 19.3.0
V Project: demo	babel	2.7.0	▲ 2.8.0
Project Interpreter 🛛 🖻	backcall	0.1.0	0.1.0
Project Structure	backports	1.0	1.0
> Build, Execution, Deployment	backports.functools_lr	1.5	▲ 1.6.1
> Languages & Frameworks backports.os		0.1.1	0.1.1
② 単击OK按钮 OK Cancel Apply			

图 1.29 显示 Anaconda 内已经安装的 Python 模块

1.4.4 测试 PyCharm

右击新建好的 demo 项目,在弹出的快捷菜单中选择 New→Python File 命令(一定要选择 Python File 项,这个至关重要,否则无法后续学习),如图 1.30 所示。

172			D () D	T 100 110	III dama () dama 1	
	<u>File</u> Edit	<u>v</u> iew <u>N</u> avigate <u>C</u> ode	Refactor Run	<u>loois VCS</u> indow	Help aemo [\aemo]	- U X
	demo				Add Configuration	▶ ∯ 15, ■ Q
rites 📑 📑 🛨 Troject	Project demo demo Comparison Scratch	▼ ③ ÷ ♥ New a S Cut h ⊆ Cor Cop D Paste Find Usages Find Usages	Ctrl+X W菜单+C Ctrl+V Alt+F7 Ctrl+Shift+F	Add Configuration File New Scratch File Directory Python Package Python File HTML File EditorConfig File Resource Bundle Bar Alt+Home	File	
🕈 2: Favo		• 石古 demo 坝目	Ctri+Shirt+K			
arre		Clean Python Comp	iled Files	iere to open		
ructi		Add to Favorites	>			
: Str		<u>R</u> eformat Code	Ctrl+Alt+L			
		Optimi <u>z</u> e Imports	Ctrl+Alt+O			
	Terminal Creates a Py	Show in Explorer Directory <u>P</u> ath	Ctrl+Alt+F12			C Event Log

Python 网络爬虫从入门到精通

图 1.30 新建 Python 文件

在新建文件对话框输入要建立的 Python 文件名 hello world,如图 1.31 所示。随后按 Enter 键,即 可完成新建 Python 文件工作。

New Python file						
🝰 hello world — 输入 hello world 文件名						
🐍 Python file						
🛃 Python unit test						
🛃 Python stub						

图 1.31 输入新建的 Python 文件名称

在新建文件的代码编辑区输入代码 "print ("hello world!")", 如图 1.32 所示。

14





在编写代码的区域右击,在弹出的快捷菜单中选择 Run 'hello world'命令,运行测试代码,如图 1.33

所示。



图 1.33 运行 Python 测试代码

如果程序代码没有错误,那么将显示运行结果,如图 1.34 所示。



图 1.34 显示程序运行结果

1.5 小 结

本章首先介绍了什么是爬虫,然后介绍了爬虫都有哪些分类(通用爬虫、聚焦爬虫以及增量式爬虫)、爬虫的基本原理,接着学习了如何搭建爬虫的开发环境,这里推荐读者安装 Anaconda,这样可以避免频繁地安装很多第三方模块。为了提高开发效率,推荐读者使用 PyCharm 开发工具来编写爬虫程序。

15