<https://blog.csdn.net/maliao1123/article/details/80826974>

> 看看有哪些爬虫框架

#一、前言

目前市场上爬虫框架有很多，不同语言不同类型的爬虫框架都有，然而在开发预研的时候对于选择那种

框架对于很多开发者来说尤为头疼；

本篇主要总结一下市场上主流的开发语言中有哪些主流的爬虫框架，以及爬虫框架的优劣；希望在对你在选择合适爬虫框架中有所帮助。

#二、主流语言爬虫框架列表

##常用爬虫框架列表

JAVA | PYTHON | PHP | C# | C/C++

--- | --- | --- | --- | ----

Apache Nutch2 | scrapy | phpspider | DotnetSpider | open-source-search-engine

webmagic | Crawley | Beanbun |NWebCrawler | Cobweb

Heritrix | Portia | PHPCrawl | SmartSpider | upton

WebCollector | PySpider | php selenium | Abot | wombat

crawler4j | grab | | xNet | Spidr

Spiderman | cola | | AngleSharp | Larbin

SeimiCrawler | python selenium | |HtmlAgilityPack

jsoup | | |CSQuery |

java selenium |

htmlunit |

#三、主流爬虫框架简介

##1.Java爬虫框架

###Apache Nutch2

地址: http://nutch.apache.org/

Nutch 是一个开源Java 实现的搜索引擎。它提供了我们运行自己的搜索引擎所需的全部工具。包括全文搜索和Web爬虫。

Nutch 致力于让每个人能很容易, 同时花费很少就可以配置世界一流的Web搜索引擎. 为了完成这一宏伟的目标, Nutch必须能够做到:

\* 每个月取几十亿网页

\* 为这些网页维护一个索引

\* 对索引文件进行每秒上千次的搜索

\* 对索引文件进行每秒上千次的搜索

\* 提供高质量的搜索结果

简单来说Nutch支持分布式，可以通过配置网站地址、规则、以及采集的深度(通用爬虫或全网爬虫)对网站进行采集，并提供了全文检索功能，可以对采集下来的海量数据进行全文检索；假如您想完成对站点所有内容进行采集，且不在乎采集和解析精度(不对特定页面特定字段内容采集)的需求，建议你使用Apache Nutch,假如您想对站点的指定内容板块指定字段采集，建议您使用垂直爬虫较为灵活。

###webmgaic(推荐)

地址: http://webmagic.io/

WebMagic是一个简单灵活的Java爬虫框架。基于WebMagic，你可以快速开发出一个高效、易维护的爬虫。

特性：

\* 简单的API，可快速上手

\* 模块化的结构，可轻松扩展

\* 提供多线程和分布式支持

###Heritrix

地址: http://crawler.archive.org/

Heritrix 是一个由 java 开发的、开源的网络爬虫，用户可以使用它来从网上抓取想要的资源。其最出色之处在于它良好的可扩展性，方便用户实现自己的抓取逻辑。

###WebCollector

地址: https://github.com/CrawlScript/WebCollector

WebCollector是一个无须配置、便于二次开发的JAVA爬虫框架（内核），它提供精简的的API，只需少量代码即可实现一个功能强大的爬虫。WebCollector-Hadoop是WebCollector的Hadoop版本，支持分布式爬取。

###crawler4j

地址: https://github.com/yasserg/crawler4j

crawler4j是一款基于Java的轻量级单机开源爬虫框架，最大的一个特点就是简单。另外也支持多线程、支持代理、可以过滤重复URL

基本上从加载jar到工程里面 通过修改示例的代码就可以简单的实现一个爬虫的全部功能，而这一切动作加起来都不需要超过半个小时。

###Spiderman

地址: https://m.gitee.com/l-weiwei/spiderman

Spiderman 是一个Java开源Web数据抽取工具。它能够收集指定的Web页面并从这些页面中提取有用的数据。 Spiderman主要是运用了像XPath、正则、表达式引擎等这些技术来实现数据抽取。

###SeimiCrawler

地址: http://seimi.wanghaomiao.cn/

一个敏捷的，独立部署的，支持分布式的Java爬虫框架

SeimiCrawler是一个强大的，高效敏捷的，支持分布式的爬虫开发框架，希望能在最大程度上降低新手开发一个可用性高且性能不差的爬虫系统的门槛，以及提升开发爬虫系统的开发效率。在SeimiCrawler的世界里，绝大多数人只需关心去写抓取的业务逻辑就够了，其余的Seimi帮你搞定。设计思想上SeimiCrawler受Python的爬虫框架Scrapy启发很大，同时融合了Java语言本身特点与Spring的特性，并希望在国内更方便且普遍的使用更有效率的XPath解析HTML，所以SeimiCrawler默认的HTML解析器是JsoupXpath,默认解析提取HTML数据工作均使用XPath来完成（当然，数据处理亦可以自行选择其他解析器）。

###jsoup

地址: https://jsoup.org/

jsoup 是一款Java 的HTML解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据。

##2.Python爬虫框架

###scrapy (推荐)

地址: https://scrapy.org/

Scrapy，Python开发的一个快速、高层次的屏幕抓取和web抓取框架，用于抓取web站点并从页面中提取结构化的数据。Scrapy用途广泛，可以用于数据挖掘、监测和自动化测试。

Scrapy吸引人的地方在于它是一个框架，任何人都可以根据需求方便的修改。它也提供了多种类型爬虫的基类，如BaseSpider、sitemap爬虫等，最新版本又提供了web2.0爬虫的支持。

Scrap，是碎片的意思，这个Python的爬虫框架叫Scrapy。

###Crawley

地址:http://project.crawley-cloud.com/

高速爬取对应网站的内容，支持关系和非关系数据库，数据可以导出为JSON、XML等

###Portia

地址:https://scrapinghub.com/portia

Portia 是一个用 Python 编写无需任何编程知识，就能可视爬取网站数据的开源工具。无需下载或安装任何东西，因为，Portia 是运行在您的 Web 浏览器中。

Portia 是 scrapyhub 开源的一款可视化爬虫规则编写工具。Portia 提供了可视化的 Web 页面，只需通过简单点击，标注页面上需提取的相应数据，无需任何编程知识即可完成爬取规则的开发。这些规则还可在 Scrapy 中使用，用于抓取页面。

###PySpider

地址:http://www.pyspider.cn/

PySpider：一个国人编写的强大的网络爬虫系统并带有强大的WebUI。采用Python语言编写，分布式架构，支持多种数据库后端，强大的WebUI支持脚本编辑器，任务监视器，项目管理器以及结果查看器。

###grab

地址:http://www.imscraping.ninja/posts/introducing-grab-framework-python-webscraping/

网络爬虫框架(基于pycurl/multicur)。

###cola

地址:https://github.com/chineking/cola

一个分布式爬虫框架。项目整体设计有点糟，模块间耦合度较高，不过值得借鉴。

##3.PHP爬虫框架

###phpspider

地址:https://github.com/owner888/phpspider

phpspider是一个爬虫开发框架。使用本框架，你不用了解爬虫的底层技术实现，爬虫被网站屏蔽、有些网站需要登录或验证码识别才能爬取等问题。简单几行PHP代码，就可以创建自己的爬虫，利用框架封装的多进程Worker类库，代码更简洁，执行效率更高速度更快。

###Beanbun

地址:http://www.beanbun.org/#/

Beanbun 是用 PHP 编写的多进程网络爬虫框架，具有良好的开放性、高可扩展性。

\* 支持守护进程与普通两种模式（守护进程模式只支持 Linux 服务器）

\* 默认使用 Guzzle 进行爬取

\* 支持分布式

\* 支持内存、Redis 等多种队列方式

\* 支持自定义URI过滤

\* 支持广度优先和深度优先两种爬取方式

\* 遵循 PSR-4 标准

\* 爬取网页分为多步，每步均支持自定义动作（如添加代理、修改 user-agent 等）

\* 灵活的扩展机制，可方便的为框架制作插件：自定义队列、自定义爬取方式…

###PHPCrawl

地址:http://phpcrawl.cuab.de/

PHPCrawl是一个PHP开源的Web检索蜘蛛(爬虫)类库。PHPCrawl抓取工具“ Spider ”的网站，并提供一切有关网页，链接，文件等信息。

PHPCrawl povides可以选择性的指定的爬虫的行为，比如喜欢网址、内容类型，过滤器、cookie的处理等方式。

##4.c#爬虫框架

###DotnetSpider

地址:http://www.dotnetspider.com/

DotnetSpider这是国人开源的一个跨平台、高性能、轻量级的爬虫软件，采用 C# 开发。目前是.Net开源爬虫最为优秀的爬虫之一。

###NWebCrawler

地址:http://nwebcrawler.codeplex.com/

NWebCrawler是一款开源的C#网络爬虫程序更多NWebCrawler

###SmartSpider

地址:http://www.softpedia.com/get/Internet/Download-Managers/SmartSpider.shtml

SmartSpider爬虫引擎内核版，全新的设计理念，真正的极简版本。

###Abot

地址:https://github.com/sjdirect/abot

Abot是一个开源的.net爬虫，速度快，易于使用和扩展。

###xNet

地址:https://github.com/X-rus/xNet

这个一个俄国牛人写的开源工具，为啥说他强悍了，因为他将所有Http协议的底层都实现了一遍，这有啥好处？只要你是写爬虫的，都会遇到一个让人抓狂的问题，就是明明知道自己Http请求头跟浏览器一模一样了，为啥还会获取不到自己想要的数据。这时你如果使用HttpWebReaquest，你只能调试到GetRespone，底层的字节流是调试不到了。所以必须得有个更深入的底层组件，方便自己调试。

###AngleSharp

地址:http://anglesharp.github.io/

解析HTML利器AngleSharp介绍解析HTML利器AngleSharp介绍AngleSharp是基于.NET（C#）开发的专门为解析xHTML源码的DLL组件。

###HtmlAgilityPack

地址:http://htmlagilitypack.codeplex.com/

HtmlAgilityPack是 .NET 下的一个 HTML 解析类库。支持用 XPath 来解析 HTML 。命名空间： HtmlAgilityPack

###CSQuery

地址:https://github.com/jamietre/CsQuery

CsQuery 犀利的html代码分析库，像jq一样用c#处理html

##5.C/C++爬虫框架

###open-source-search-engine

地址:https://github.com/gigablast/open-source-search-engine

基于C/C++开发的网络爬虫和搜索引擎.

###Cobweb

地址:https://github.com/stewartmckee/cobweb

非常灵活，易于扩展的网络爬虫，可以单点部署使用.

###upton

地址:https://github.com/propublica/upton

一个易于上手的爬虫框架集合，支持CSS选择器.

###wombat

地址:https://github.com/felipecsl/wombat

基于Ruby天然的支持DSL的网络爬虫，易于提取网页正文数据.

###Spidr

地址:https://github.com/postmodern/spidr

全站数据采集，支持无限的网站链接地址采集.

###Larbin

地址:http://larbin.sourceforge.net/download.html

larbin是一种开源的网络爬虫/网络蜘蛛，由法国的年轻人Sébastien Ailleret独立开发，用c++语言实现。larbin目的是能够跟踪页面的url进行扩展的抓取，最后为搜索引擎提供广泛的数据来源。 Larbin只是一个爬虫，也就是说larbin只抓取网页，至于如何parse的事情则由用户自己完成。另外，如何存储到数据库以及建立索引的事情 larbin也不提供。

larbin最初的设计也是依据设计简单但是高度可配置性的原则，因此我们可以看到，一个简单的larbin的爬虫可以每天获取500万的网页，实在是非常高效。

利用larbin，我们可以轻易的获取/确定单个网站的所有联结，甚至可以镜像一个网站；也可以用它建立url 列表群，例如针对所有的网页进行 url retrive后，进行xml的联结的获取。或者是 mp3，或者定制larbin，可以作为搜索引擎的信息的来源。

#四.总结

爬虫框架很多，如果有兴趣我们自己也可以写一些。我们不需要掌握每种主流的爬虫框架，只需要根据自己擅长编写语言，深入的掌握一种爬虫框架即可。大多数爬虫框架实现起来都大同小异。

如果你是python开发者我推荐你学习目前流行的scrapy,如果你是java开发者我推荐你学习webmagic。

---------------------

作者：Lateautunmxxxxx

来源：CSDN

原文：https://blog.csdn.net/maliao1123/article/details/80826974

版权声明：本文为博主原创文章，转载请附上博文链接！