**主要开源网络爬虫简介**

**张劲暾 PB16111485**

**1, Python语言开源爬虫：**

1. scrapy, 快速,高级封装的爬虫框架, 方便修改, 提供了多种类型爬虫的基类, 便于开发, 从这个角度看, scrapy类似于爬虫框架的内核, 基于这个内核开发的爬虫及爬虫框架还有：diango-dynamic-scraper, scrapy-redis, scrapy-cluster, distribute\_crawler等.
2. Pyspider,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. 可以用任何HTML解析包, WEB 界面编写调试脚本, 起停脚本, 监控执行状态, 查看活动历史,获取结果产出, 数据存储支持MySQL, MongoDB, Redis, SQLite, Elasticsearch; PostgreSQL及SQLAlchemy,队列服务支持RabbitMQ, Beanstalk, Redis和Kombu,支持抓取 JavaScript 的页面,组件可替换,支持单机/分布式部署, 支持Docker部署, 强大的调度控制, 支持超时重爬及优先级设置
3. cola,分布式的爬虫框架, 用户只需编写几个特定的函数, 而无需关注分布式运行的细节. 任务会自动分配到多台机器上, 整个过程对用户是透明的. Cola集群需要一个master和若干个worker, 对于每台机器, 只能启动一个worker. 但是集群不是必须的, 在单机模式下亦可以运行. Cola还不够稳定, 目前会处于持续改进的状态。
4. Scrapely, 一个可以从HTML网页中提取结构数据的库. 给定一些网页例子和需要提取的数据, scrapely会对类似网站构造一个通用解析器.
5. Grab, 具有自动cookie支持和HTTP和SOCKS代理,支持IDN, 可以灵活定制HTTP请求, 具有自动的字符集检测功能和强大的API, 可通过XPATH查询从HTML文档的DOM树中提取数据,异步API可以实现数千个同步查询
6. mechanicalsoup,自动存储和发送cookie,遵循重定向,并可以跟踪链接和提交表单, 不处理JavaScript.
7. portia,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 Web 浏览器中,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.
8. crawley,基于Eventlet构建的高速网络爬虫, 可以将数据储存在Mysql,Sqlite等关系数据库中,拥有cookie处理程序可以抓取受登录保护的页面.
9. brownant,轻量级web数据爬取框架,主要由豆瓣东西使用

**2, C语言开源爬虫：**

1. httrack: 一种离线浏览器, 可以将网络站点爬取到本地目录并保存相关网页文件, 即建立镜像站点.

**3, Java语言开源爬虫：**

1. Apache Nutch,非常成熟的产品化网络爬虫,支持细粒度配置,以Apache Hadoop数据结构为依托,提供了良好的批处理支持.Nutch不仅具备了插件式和模块化优点,还提供了可扩展的功能接口,比如解析,索引和自定义ScoringFilter实现,像是用Apache Tika进行解析.除此之外,Nutch还为Apache Solr,Elastic Search,SolrCloud等提供了插件式索引.Nutch可以自动发现网页超链接,减少很多维护工作,比如检查坏链接,为所有访问过的页面建立拷贝进行搜索.
2. Crawler4j, Java实现的开源网络爬虫.提供了简单易用的接口,可以在几分钟内创建一个多线程网络爬虫.
3. Jsoup,HTML解析器,可直接解析某个URL地址,HTML文本内容,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.
4. Gecco,一款用java语言开发的轻量化的易用的网络爬虫.Gecco整合了jsoup,httpclient,fastjson,spring,htmlunit,redission等优秀框架,只需要配置一些jquery风格的选择器就能很快的写出一个爬虫.Gecco框架有优秀的可扩展性,支持爬取规则的动态配置和加载,支持页面中的异步ajax请求,支持页面中的javascript变量抽取,支持下载代理服务器随机选取.
5. Spiderman,基于微内核+插件式架构的网络蜘蛛,它的目标是通过简单的方法就能将复杂的目标网页信息抓取并解析为自己所需要的业务数据.
6. Heritrix, 良好的可扩展性，方便用户实现自己的抓取逻辑,

关键特性：

1.用单个爬虫在多个独立的站点一直不断的进行递归的爬.

2.从一个提供的种子进行爬,收集站点内的精确URI,和精确主机.

3.主要是用广度优先算法进行处理.

4.主要部件都是高效的可扩展的

5.良好的配置.

局限：

1.单实例的爬虫,之间不能进行合作.

2.在有限的机器资源的情况下,却要复杂的操作.

3.只有官方支持,仅仅在Linux上进行了测试.

4.每个爬虫是单独进行工作的,没有对更新进行修订.

5.在硬件和系统失败时,恢复能力很差

**4, JavaScript语言开源爬虫：**

1. simplecrawler, 用于抓取网站的库.爬网程序从爬网页面提供全面的数据,可用于页面分析,索引,可访问性检查等.可以指定限制来限制二进制文件的爬网.

**5, C#语言开源爬虫：**

1. DotnetSpider, 至力于打造一个轻量化,高效率,易开发,可管理的一体化爬虫框架.它的核心要素是URL调度,基本的网页内容下载器,基于爬虫实体类的爬虫实现模式,常用数据库的支持,多线程的支持,状态的监控,运行日志,网页端管理. 具有相当高的可定制性,支持包括MySQL,PostgreSQL,MongoDB 在内的多种数据库.
2. Abot, 专为提高速度和灵活性而打造,它负责低级管道（多线程,http请求,调度,链接解析等）,只需注册事件来处理页面数据,还可以插入核心接口实现,以完全控制爬网过程,轻松定制（可插拔架构允许您决定爬行的内容和方式）,经过严格的单元测试（高代码覆盖率）,非常轻便（不是过度设计）,没有进程外的依赖项（数据库，安装的服务等......）

**6, PHP语言开源爬虫：**

1. Goutte, 屏幕抓取和Web爬行库, 提供了一个很好的API来抓取网站并从HTML/XML响应中提取数据.
2. Php-spider,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。
3. 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…