# Elise概要设计

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 作者 | 版本号 | 修改记录 |
| 2017-07-28 | 聂嘉文、夏湘东 | V1.0 | 初稿 |

## 1、总体架构

Elise主要包含四大组件：下载、抽取、调度、持久化。APP调用四个组件完成采集流程。



### 1.1、下载器

Downloader主要负责将种子任务下载下来然后将下载下来的页面交给Extractor进行抽取，并输出日志。



### 1.2、抽取器

Extractor主要根据下载下来的Page（网页内容），进行解析(html、json、jsonp)，得到新的url任务和页面属性抽取结果：



### 1.3、调度器

Scheduler主要管理待爬取的任务、新任务去重后存入Redis待抓取任务队列。



### 1.4、持久化

对抓取数据进行持久化存储。



### 1.5、APP

Elise程序入口，包含全局配置信息(缓存页面规则)，并调用四大组件，完成采集流程。

其中，规则更新流程：

