南華大學
UNIVERSITY OF SOUTH CHINA

密　级：普通

文档编号：20170630-UOH　　版　次：2

分册　　第 1 册/共 1 册

分布式爬虫

（20170630）

作品说明书

编　　写　：　刘宝玉

校　　对　：　王浩茂、杨奎

审　　核　：　刘立

批　　准　：　刘立

生效日期　：　2017 年 6 月 30 日

南华大学计算机学院制

2017 年 6 月

文件修改控制

修改记录编号	修改状态	修改页码及条款	修改人	审核人	批准人	修改日期

目录

1. 作品简介

1.1.作品名称

分布式爬虫

1.2.作品类别

爬虫+Web 管理

1.3.中心内容

传统的爬虫系统，对所有的网页采用同样的办法处理，利用深度优先或广度优先的办法获取网页链接，下载网页，对网页中的所有的文本数据建立倒排索引。这种方式没有对网页数据的信息进行组织、归类。应大数据的需求，本小组开发了分布式爬虫系统。分布式爬虫，对同一个网站的同类数据，进行结构化。同时，能利用分布式的软件设计方法，实现爬虫的高效采集。本系统在爬虫时自动分析网页的组织形态获取新的链接，进行高速有效的下载；将爬取到的数据自动结构化，对于电商类网页，能对同一个网站的数据进行自动结构化，生成不同的表，例如商品表、店铺表、评价表等；对于新闻博客类网页，能进行网页正文的自动抽取，对正文进行自动摘要和关键词分析；使用 URL 去重算法对 URL 进行去重，对没有更新和已下载的的数据不再重复下载；对已经爬取到的数据与数据库中的同类信息进行对比分析，就电商网站这一类型，有助于用户对商品信息的真伪进行正确的识别；

采用分布式调度算法，将所有任务在多台机器上分布式执行，将不同网站的 URL 混合后，分配到多台机器上执行，并进行失败处理，在满足特定的条件下，实现最大的下载量。

为了方便后续的功能拓展以及使系统更加符合专业人员的需要，本小组为系统中用到的相关配置信息、相关功能介绍等信息提供了完整的后台配置界面，用户可随时对相关标准和指标做出调整，满足进一步的需要。

1.4.创新点与特色

- 客户端**跨平台**，用户只需使用浏览器即可完整使用系统（手机设备或 PC 设备）。
- 分布式+多进程进行爬虫任务，提高爬虫效率。
- 可对爬虫任务、进程、主机进行监控，保证爬虫的安全性和可靠性。
- 异常恢复

  由于发生异常中断的主机可恢复该主机上所有任务，使其从断点处继续执行。

2. 主要功能

该系统基于分布式调度算法、网页自动结构化、URL 去重算法的爬虫系统，自动的对相关信息建立表格、智能的对文章内容进行抽取、对关键词和摘要信息进行数据分析，将获取到的信息直观的提供给用

户。并对抽取到的数据进行测试、对所有任务进行统计、对爬虫任务进行监控、对参与爬虫任务的主从机和进程进行相关监控和操作、利用 graphite 监控 scrapy、对异常的恢复，从而大大地增加了爬虫的可信性和执行容错性。

主要功能有：

- **发布爬虫任务**

  发布爬虫任务要求用户设定的任务名称、输入任务的 URL、选择网站类型以及运行模式，系统将根据这些参数设置爬虫属性进行爬取。

- **查看爬虫任务**

  可以查看已经发布的爬虫任务，任务包括正在进行和历史任务，正在进行的任务两种状态：正在运行和暂停。历史任务全部是已经停止的爬虫。

- **爬取数据**

  从输入的起始 URL 进入，识别该页面中的需要的 URL 加入 redis 中的 URL 的队列中，分析页面结构抽取出有价值的内容插入 mongodb 数据库。

- **数据结构化**

  电商类：

  1. 从电商首页中获取导航栏中的商品分类。
  2. 从商品分类进入商品列表页，获取每页的商品信息并自动翻页。

3. 从商品信息进入商品详情页，获取更详细的商品信息。

新闻博客类：

1. 从新闻或博客网页中解析所有 URL,自动获取新闻或博客标题、正文、时间等。

- **查看数据**

可以在界面上查看爬取的已经结构化的数据。

- **进程管理**

查看正在执行爬虫任务的进程的状态。

- **正文测试**

正文测试分为批量测试和单例测试。批量测试是对大量抽取正文的文件与标准文件进行对比并计算其正确率以及显示各文件的测试成功与否；单例测试是对单个文件进行测试，将其内容展示出来，供用户查看对比。

- **数据统计**

对当前时间下所有的爬虫任务的数据统计，包括历史任务统计和当前任务的统计，并使用柱形图和饼图直观地展示爬虫任务的数量、状态、类型的关系。

- **从机管理**

添加或删除参与爬虫任务的从机，显示从机的相关信息。

- **异常恢复**

当某台正在执行爬虫任务的从机因为某种因素而导致异常时，该从机将清除之前的进程信息，重新启动该从机上的所有爬

虫任务，爬虫任务将会从其断点处继续执行。

- **监控爬虫**

  对爬虫任务进行监控，并对其性能进行分析，将其可视化。即利用 graphite 监控 scrapy。

- **自动结构化**

  根据用户输入的网页 URL 将该网页上的数据结构化。分为正文抽取和批量抽取。正文抽取即根据用户输入的正文网页URL将网页上的正文自动结构化，从而将已经结构化的正文信息显示出来。批量抽取即根据用户输入的多个网页 URL 将网页上的数据自动结构化，从而将已经结构化的网页信息显示出来，并将显示结构化的信息链接，根据链接可查看详细内容。

3. 操作说明

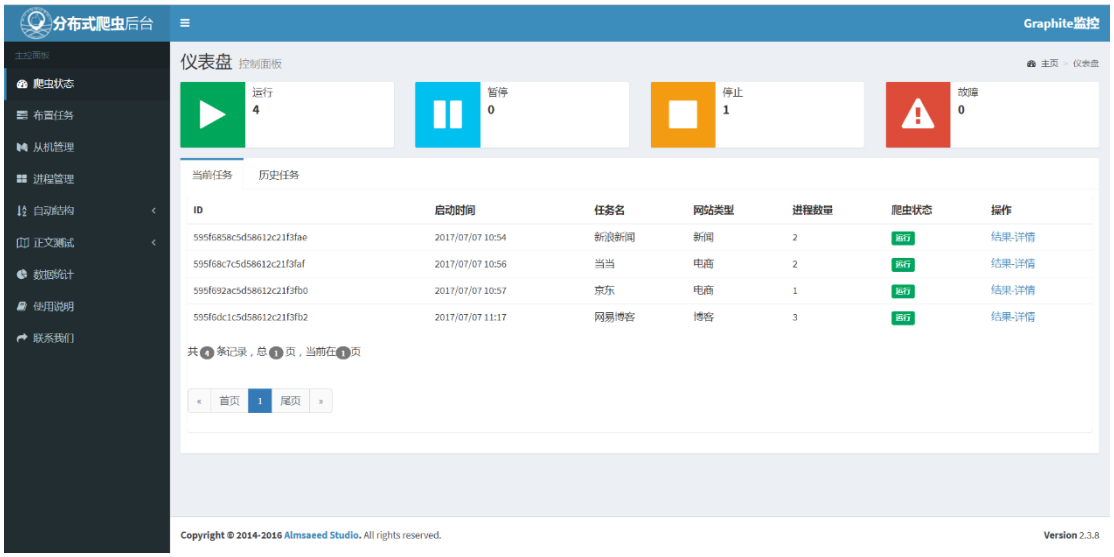
项目已部署，访问地址：http://123.207.230.48/crawlermanage/，登录名：admin，密码：a。

下面是针对用户的操作说明。

3.1. 界面说明

用户登录之后，就可以进入以下界面，进行相关操作。

3.1.1. 爬虫状态界面

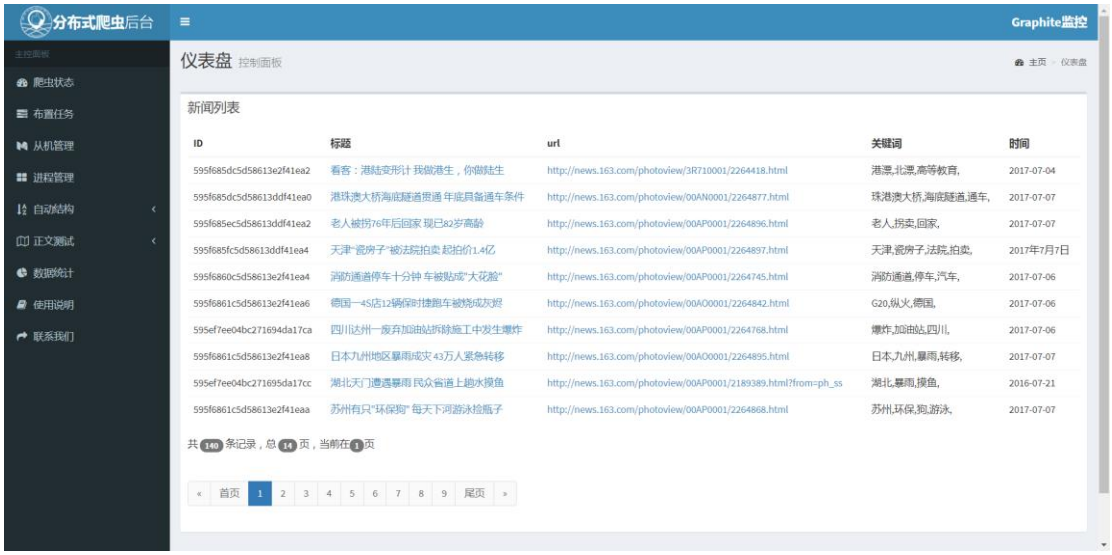


如上图为爬虫状态界面，在此界面上用户可以查看当前爬虫任务的信息和历史爬虫任务的信息，其中相关信息包括任务的 ID、启动时间、任务名、网站类型、爬虫状态、操作，用户可以通过操作一栏进而查看该任务的结果详情信息。该界面统计了任务的数量，用户可以点击翻页。

- **详情界面**

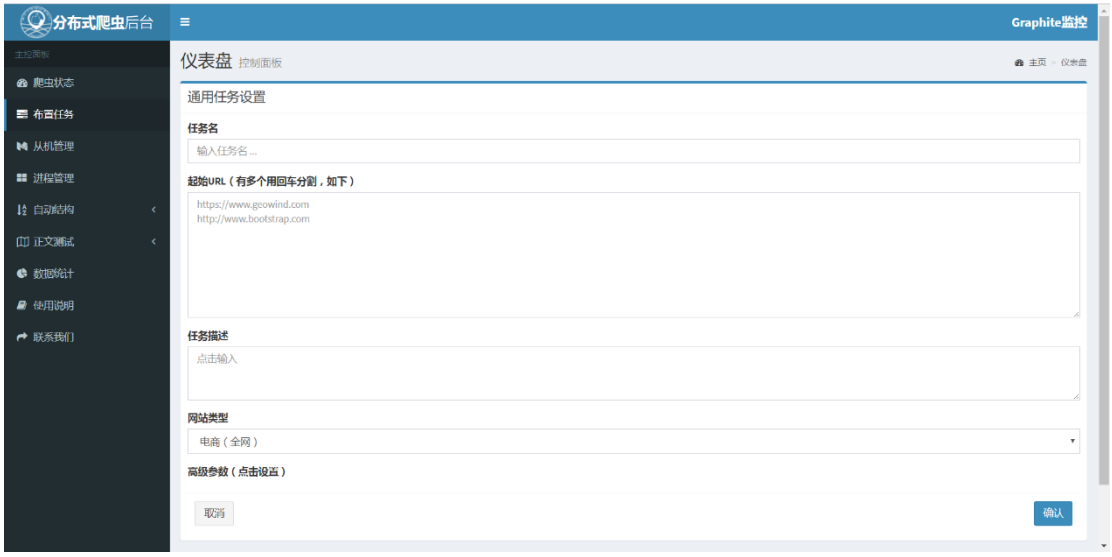
如上图为详情界面，在该界面上可以查看到爬虫任务的基本信息项。

➢ 结果界面



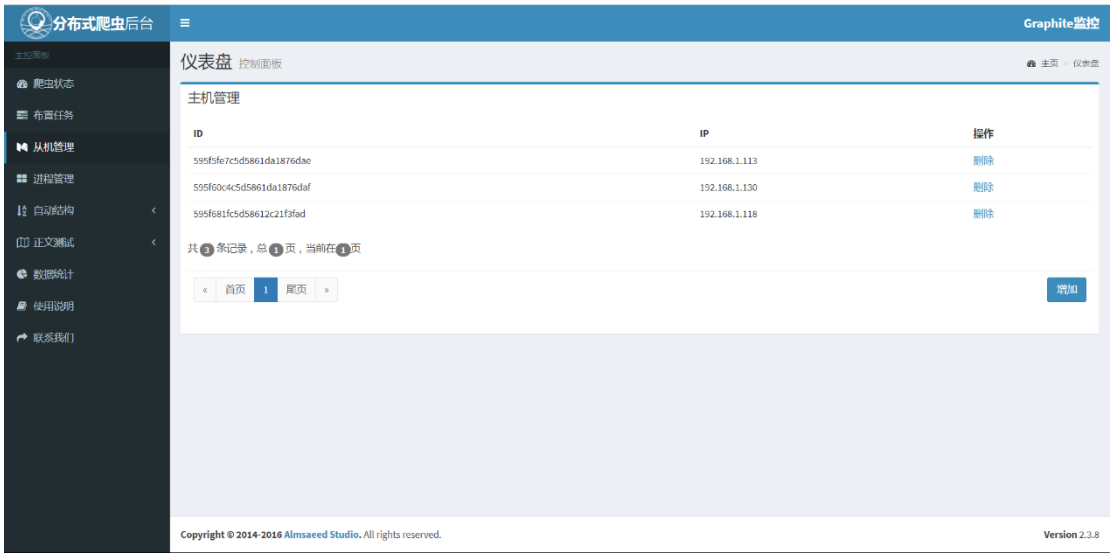
如上图为结果界面，在该界面上显示了爬虫任务结果的列表。

3.1.2. 布置任务界面

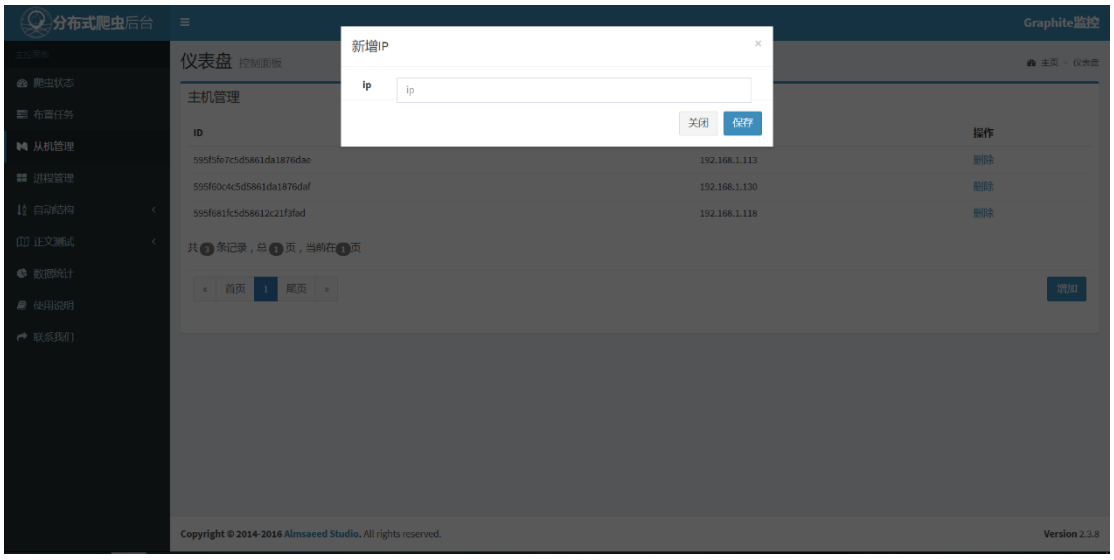


如上图为布置任务的界面，任务即爬虫任务。在此界面上，用户可以设定相关参数来发布爬虫任务。

3.1.3. 从机管理界面

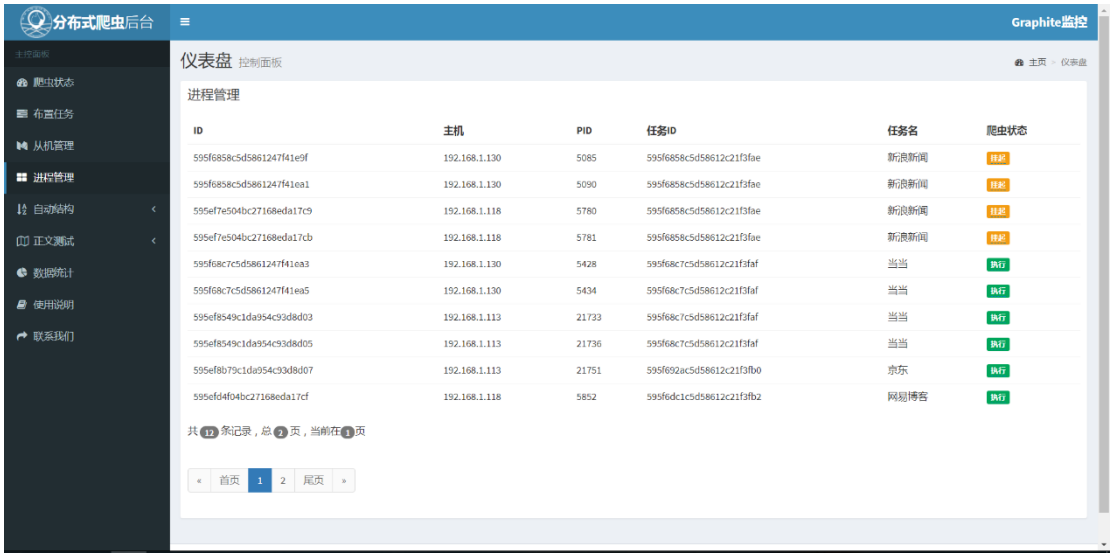


如上图为从机管理的界面，用户可以在此界面上查看到相关主机的信息（主机 ID、主机 IP、操作），此外，用户可以通过操作一栏对主机进行删除操作，通过下方的添加按钮添加主机，添加时要输入主机的 IP。该界面也对所有主机进行了数量统计，用户可通过下方的页码或方向标进行翻页。



如上图为添加从机时的界面，用户输入合法 IP 便可成功添加从机。

3.1.4. 进程管理界面

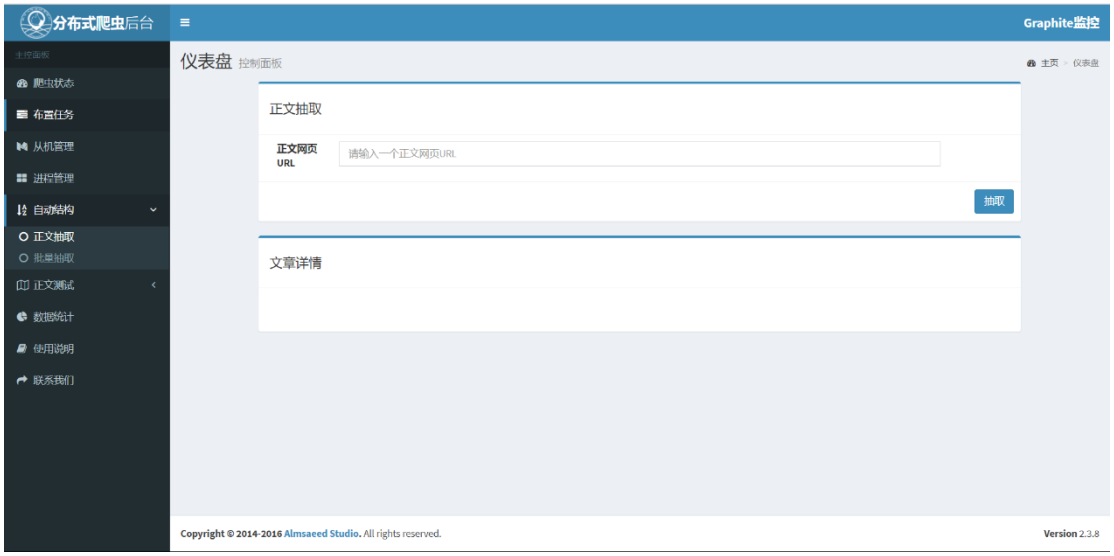


如上图为进程管理界面，用户可以在此界面上查看正在执行爬虫任务的进程的状态以及任务的部分信息（ID、主机 IP、PID、任务 ID、任务名、爬虫状态），在状态栏可以点击改变相应进程的状态。用户可通过下方的页码或方向标进行翻页。

3.1.5. 自动化结构界面

自动化结构界面分文正文抽取和批量抽取两个界面。

- **正文抽取**



如上图为正文抽取界面，用户输入合法正文网页 URL，点击抽取按钮，就可进行正文抽取。在文章详情处可看到结构化后的文章详情。

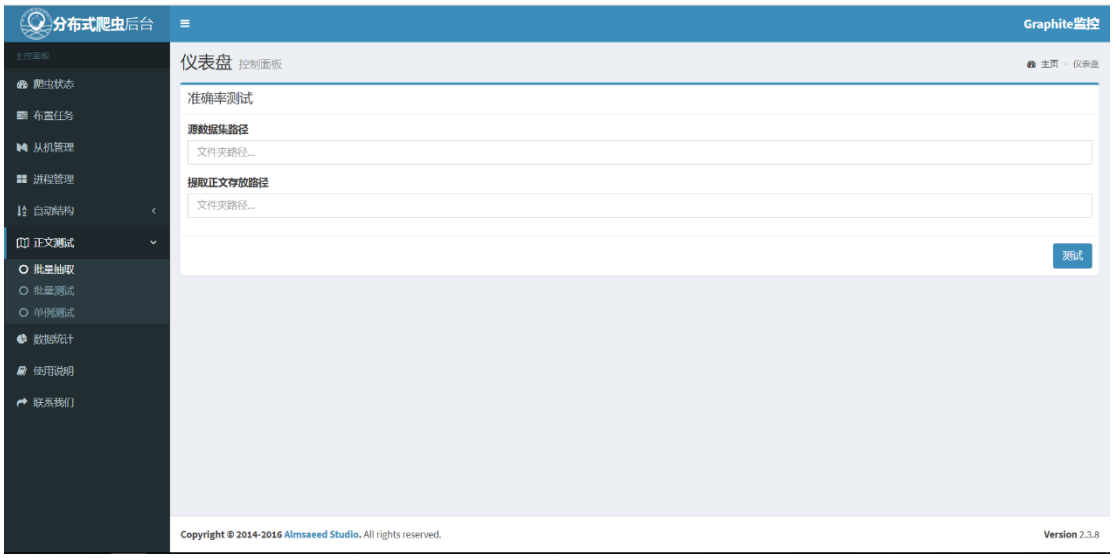
- **批量抽取**

如上图为批量抽取的界面，用户可输入多个具有相同模块的合法 URL，选择网页类型，点击确认按钮便可进行批量抽取。在测试结果会出现相关链接、标题等，点击标题可看到详细信息。

3.1.6. 正文测试界面

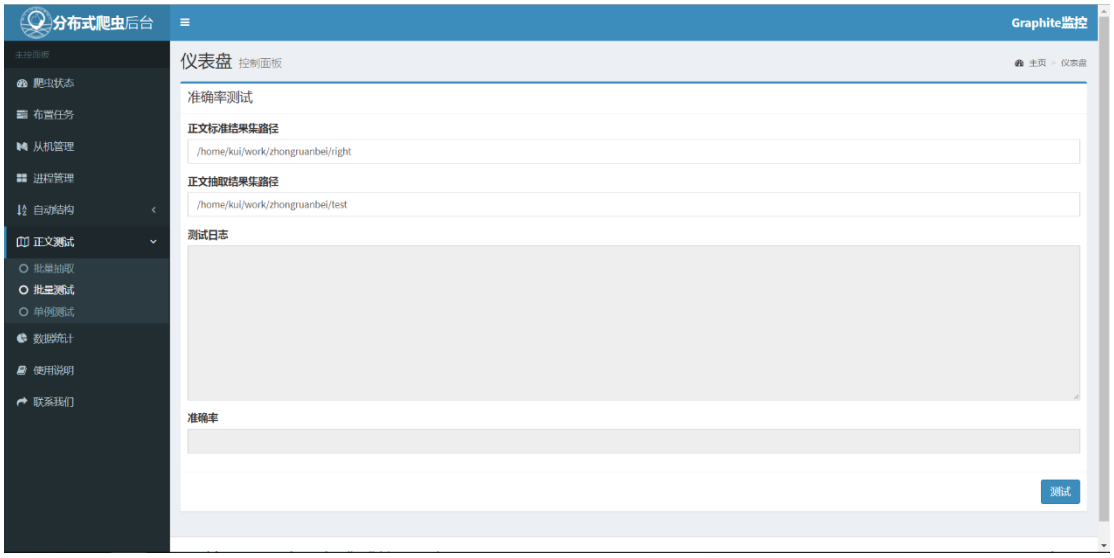
正文测试包括了批量抽取、批量测试、单例测试。

- **批量抽取界面**



如上图为批量抽取的界面，用户输入源数据路径即存放文件的文件夹路径、提取正文存放路径即对源文件夹中的文件进行了正文抽取后要存放的文件的文件夹路径。

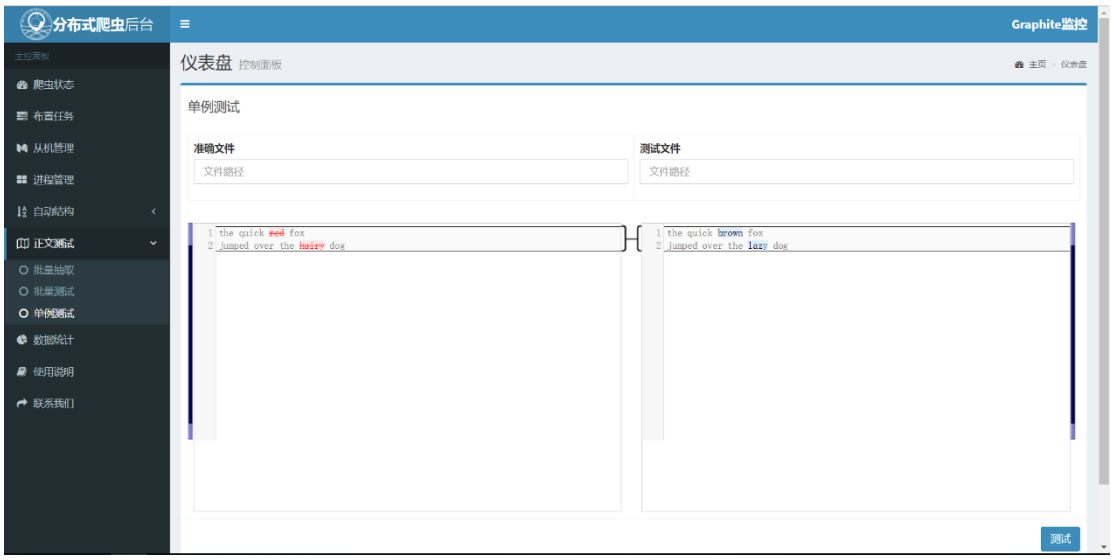
- **批量测试**



如上图为批量测试的界面，用户可以输入正文标准结果路径和通过批

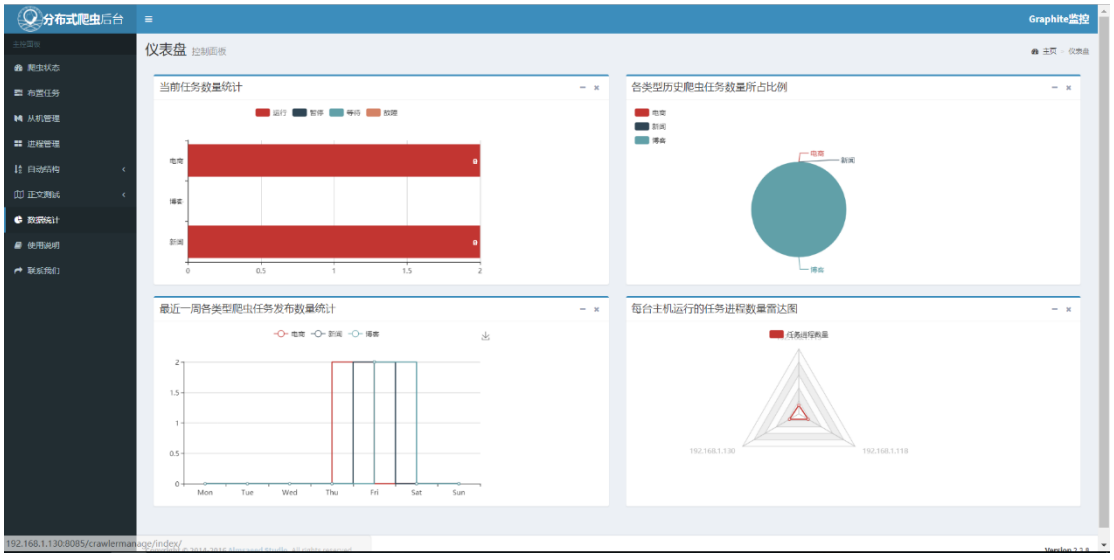
量抽取的文件结果的路径，点击测试按钮测试其准确率。

- **单例测试**



如上图为单例测试的界面，用户输入准确文件的路径、测试文件的路径，点击右下方的测试按钮，系统将会对两个文件进行比对，并将比对的内容展示在界面上，让用户可以直观的得出结果。

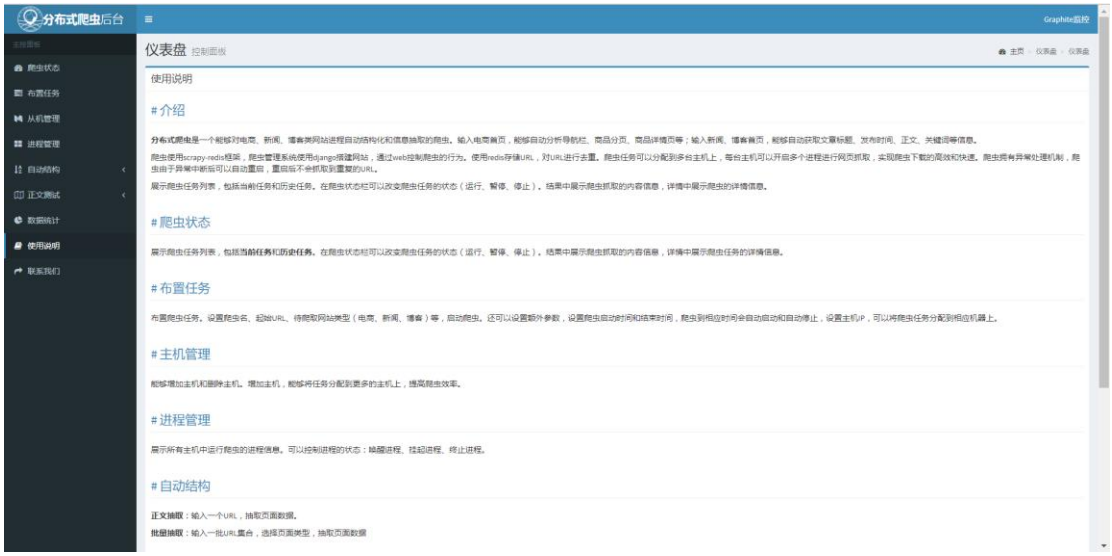
3.1.7. 数据统计界面



如上图为数据统计的界面，用户可以看到当前尚未完成的爬虫任务的

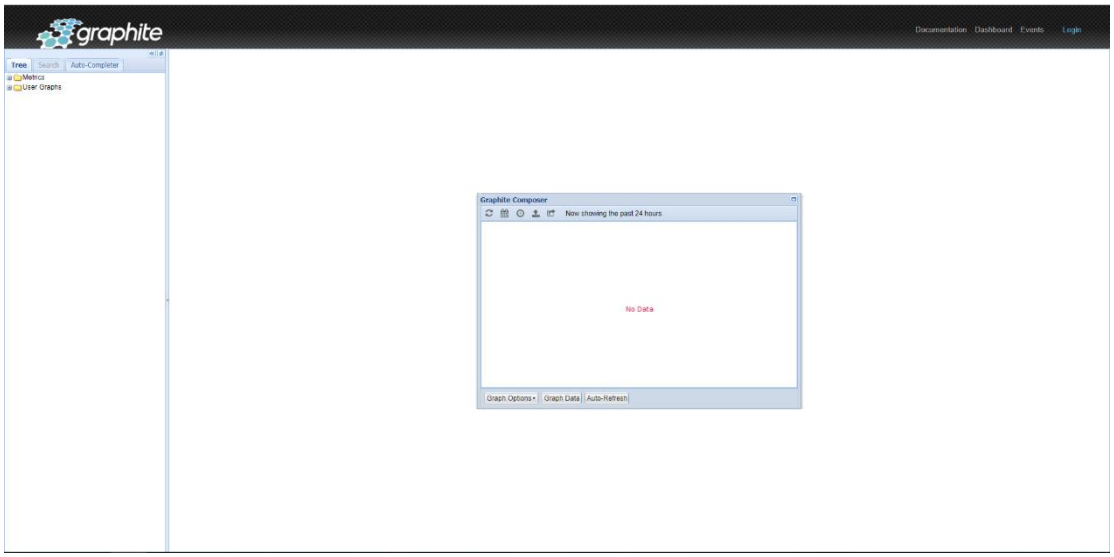
状态、种类、数量的柱形图，历史爬虫任务种类所占比例的饼图，最近一周内爬虫任务的数量统计的柱形图，每台主机的任务进程数量的雷达图。通过此界面，用户可以完全了解当前系统所有爬虫任务的数量种类以及状态的信息。

3.1.8. 使用说明界面



如上图为使用说明界面，用户通过此界面可以了解到该爬虫系统以及爬虫系统相关功能介绍，以方便用户快速熟悉该系统的使用。

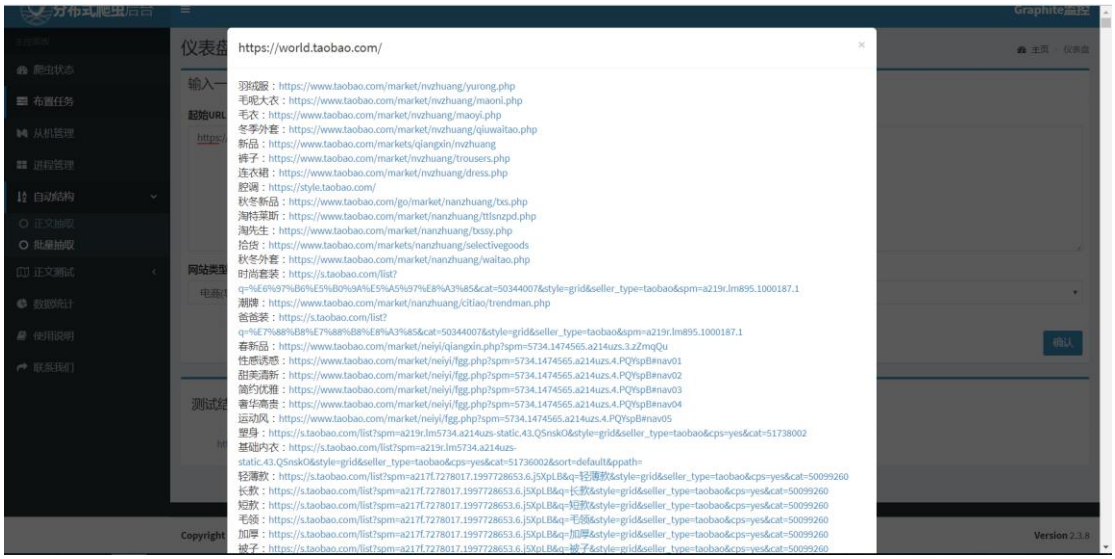
3.1.9. Graphite 界面



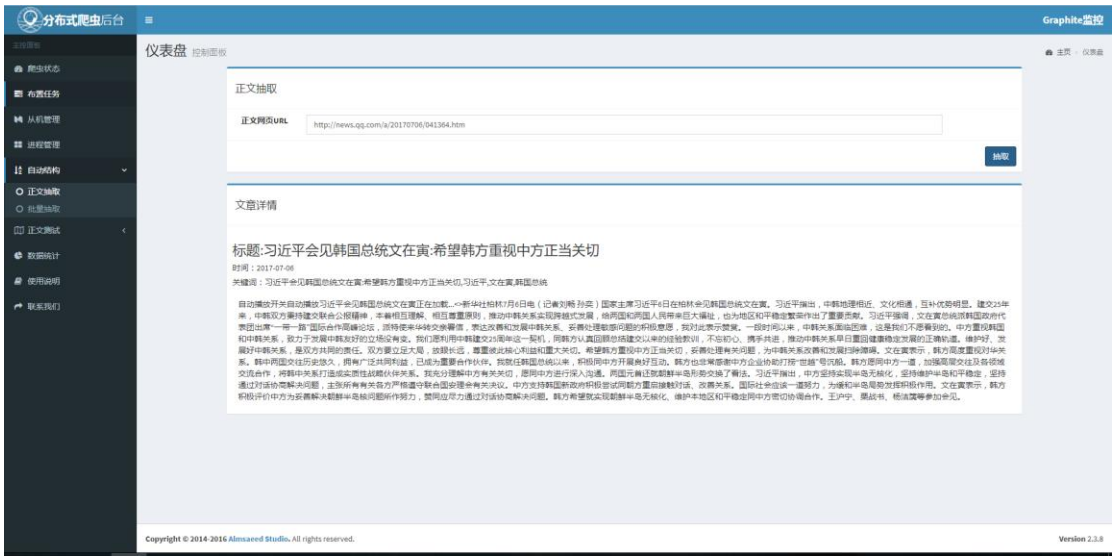
如上图为 Graphite 监控爬虫界面，用户可通过选择点击左侧的菜单栏从而观测到爬虫任务的监控情况。

3.2. 运行实例

淘宝网页导航栏的数据结构化结构：



正文抽取的一个运行实例：



3.3. 参数说明

参数含义	说明	实例取值
任务名	输入一个用户名，可为中文英文	爬虫 1….
URL 网址	可输入多个有效网址，采用，进行分割	https://www.geowind.com
任务描述	可输入中文英文	
网站类型	声明输入的网址类型	电商网址
运行时间区间	设置爬虫进行的开始时间和结束时间	2017/05/21/10:10-2017/05/21/10:30
分布式主机	爬虫时可使用的主机	
IP	合法 IP 即可	192.168.1.1
文件夹路径	合法路径，即路径存在且路径无错	E:\xx\xx

关键字		上衣