# 爬虫系统使用说明

# 1. 设置

平台部署完成之后，需要对整个服务进行一些全局参数的设置。只有正确设置这些参数，才能保证服务的稳定运行。

## 1.1 数据库配置

点击顶栏的设置按钮，进入数据库配置页面。如下：


在数据库链接一栏中填入该服务所使用的 MySQL 数据库链接。JDBC URL 的语法规则可自行 Google 获得。

在数据库用户名和数据库密码两栏分别填入所授权的用于登录数据库的用户名和密码。

填写完毕，点击提交按钮，服务器端会检验所填写内容是否有效，检验结果会进行弹窗提示。

注意，若此处设置内容无效，平台上的任何其他服务均无法正常使用。

## 1.2 平台配置

点击左栏的平台配置按钮，进入平台配置页面。


在基础工作路径栏中所填写的内容为该服务所有运行时数据和所下载数据所存储的根目录。该栏需根据服务所部署的服务器系统环境进行合理配置。

配置完成，点击提交按钮，服务器端会校验所填写内容是否正确、合理。对校验结果会进行弹窗提示。注意，若此处设置内容无效，则爬虫任务无法正常启动，无法爬取任何数据。

# 2. 采集功能

采集功能主要是为用户提供参数配置及任务调度的接口，根据数据结构的分类，将采集功能分为结构型数据采集功能及非结构型数据采集功能，以下将分别描述具体操作流程。

## 2.1 结构型数据采集

结构型数据采集，即表格类数据网站采集，其中结构型数据采集模式分为基于链接和基于接口，基于链接采集模式适用于静态更新页面内容的网站；基于接口采集模式适用于使用 api 接口更新页面内容的网站。以下将分别描述具体操作流程。

### 2.1.1 基于链接的结构型数据采集

基于链接的结构型数据采集（以扶贫网站数据采集为例）具体操作流程如下：

点击主菜单栏“采集”->点击任务确立 ->运行模式选中“结构型”->采集模式选中“基于链接”，打开结构型数据采集任务确立界面；

（1） 填写任务参数，参数具体信息如下：

任务名称：用户可以自定义。不可为空。 例：结构_基于链接_0515_test01

网站链接：待爬取数据所在的网址。不可为空。 例：http://121.194.104.120/jkld/show.html

工作路径后缀：存储文件的文件夹路径。不可以为空，不可包含/或\。 例：structure

填写完成后，点击“提交”，提交任务，将提示是否提交成功，点击确定，将进入 URL 解析规则配置页面。

（2） URL 解析规则配置页面将显示所有新建的任务及其相应信息，如下图所示；可进行修改和删除操作。点击对应任务的“修改”按钮，进入相应任务 URL 解析规则配置页面。点击对应任务的“删除”按钮，将删除与该任务相关的所有信息。


例：点击结构型任务编号 148 对应的“修改”，进入任务 URL 解析规则配置页面。

（3） URL 解析规则配置

分为 URL 参数配置、登录参数配置和下载参数配置，以下分别进行具体描述：

1）URL 参数配置


具体参数说明如下：

- 搜索栏链接：搜索页面链接前缀，即完整的搜索页面链接？前部分。

  例：http://121.194.104.120/jkld/Search?

- 关键词参数名称：所提交的搜索词的参数名称。

  例：识别标准,农户属性,主要致贫原因

- 分页参数名称：指定搜索结果第几页的参数名称

  例：curpage

- 开始页面号：开始页面号这一栏填入的是一个组合参数，由两个整数值构成，用英文的逗号隔开。第一个整数值代表搜索结果页面第一个分页的 URL 中分页参数的值。第二个整数值代表的是，在搜索结果的多个分页中，相邻分页的 URL 中分页参数的差值

  例：1,1

- 其他参数名称列表：搜索结果页面的 URL 中，除关键词参数和分页参数这两个变量外的其他必须携带的常量，则将常量参数的参数名称填入其他参数名称列表栏中，多个参数名称用英文逗号隔开

  例：Submit

- 其他参数值列表：对应于其他参数名称的值列表

  例：跳转

- 属性值：关键词参数的可选属性列表，不同参数属性用分号隔开，相同参数的属性用逗号隔开。

例：国家标准;一般贫困户,一般农户,低保户,低保户贫困户;因病,因残,因学,因灾,缺土地,缺水,缺技术,缺劳力,缺资金,交通条件落后,自身发展动力不足,其他,未填写

配置完毕，点击提交按钮提交配置值信息。

2）登录参数配置

下图为登录参数配置页面，若目标网站不需要登录即可进行爬取，则该部分参数可不填写，在本例中，不需要进行登录参数配置。


具体参数填写说明如下

- 登录界面链接：所爬取网站用户登录页面的网址
- 用户名输入框 Xpath：用户名输入框的 id，可在登录页面，通过浏览器右键“检查”进行获取。
- 密码输入框 Xpath：密码输入框的 id，方法同上。
- 登录按钮 Xpath：登录按钮的 id
- 用户名：有效的用户登录名
- 密码：有效的用户登录密码
-

3）下载参数配置


**该页面配置的爬虫运行过程中的一些系统性配置信息。**具体参数填写说明如下：

- 线程总数：线程总数，此参数不宜设置过大。设置过大会导致爬虫启动过多线程，当线程数量严重超越 CPU 核心数量时，会导致爬虫运行过程中 CPU 资源浪费在频繁的线程上下文切换中。对于个人电脑，此处建议值为 5，对于商业服务器，此处建议值为 20。具体值可参照操作系统参数中的 CPU 核心数量进行微调。 例：5
- 超时时间：最大响应时间，单位为 ms，如果链接的响应时间超过此时间，则放弃对该链接的爬取。对于网速较好的情况下，建议设为 3000。在网速较慢的情况下，建议酌情加大该值。正常情况下，该值宜大不宜小。 例：3000
- 编码方式：填入所爬取网站页面所使用的编码格式。该栏填写错误将导致所下载数据编码错误，数据打开为乱码。该值可在目标网站页面的 head 中查看。 例：utf-8
- 待爬取数据总量：所爬取网站背后的数据量，单位为文档，可以为空。例：47511

（5）内容抽取

内容抽取主要为用户提供数据抽取规则参数配置，分为两个部分，模板列表及新建模板，初次建立模板时选择新建模板，后可在模板列表中进行相应修改。页面如下：

结构型数据采集初次建立模板，点击“新建模板”->选中“结构型”页面，进行参数配置。

具体参数填写说明如下：

- 绑定任务：在下拉菜单中选择需配置模板的任务名称。 例：结构_基于链接_0515_test01
- 模板名称：填写模板名称（单个字母），不可为空。例：a
- 模板类型：模板类型，默认公式组合（formula）、用户自定义（userDefined）。例：formula
- 模板 XPath：表格在网站页面中的 XPath。多个 XPathyong’#’分割。

  例：/html/body/table
- 表格 formula：表格组合公式，模板类型选择 formula 时填写，否则填写 null。

  例：a
- 模板 headerXPath：模板类型选择 userDefined 时填写，模板在网站页面表头的 XPath。多个 XPath 用‘#’分割，否则填写 null。 例：null

配置完成页面如下，可进行修改及删除操作。

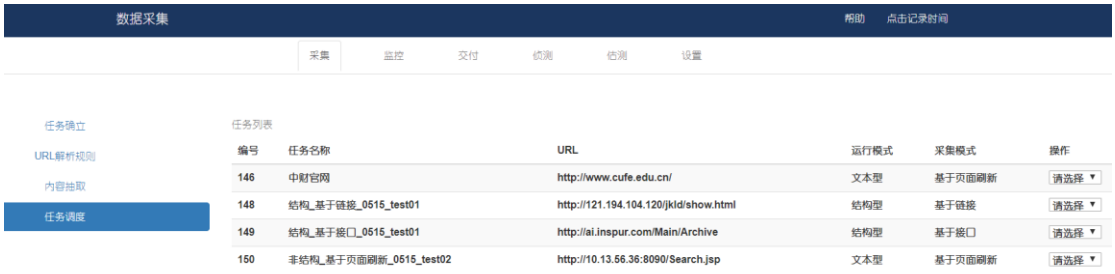

（6）任务调度

下图为可进行任务调度的任务列表，操作分为启动，停止和删除。点击相应任务启动按钮，将提示是否启动成功，启动成功，则在后端进行数据采集，可点击主菜单栏“监控”查看具体任务进度。

启动：开始对网站进行爬取

停止：暂停对网站的爬取，并将已爬取的链接保存

删除：删除所有与该任务相关的数据库数据及文件




## 2.1.2 基于接口的结构型数据采集

基于接口的结构型数据采集（以扶贫网站数据采集为例）具体操作流程如下：

（1）点击主菜单栏“采集”->点击任务确立 ->运行模式选中“结构型”->采集模式选中“基于接口”，打开结构型数据采集任务确立界面；


（2）填写任务参数，参数具体信息如下：

任务名称：用户可以自定义。不可为空。　　例：扶贫户数据

网站链接：待爬取数据所在的网址。不可为空。　例：http://ai.inspur.com/Main/Archive

工作路径后缀：存储文件的文件夹路径。不可以为空，不可包含/或\。 例：structure

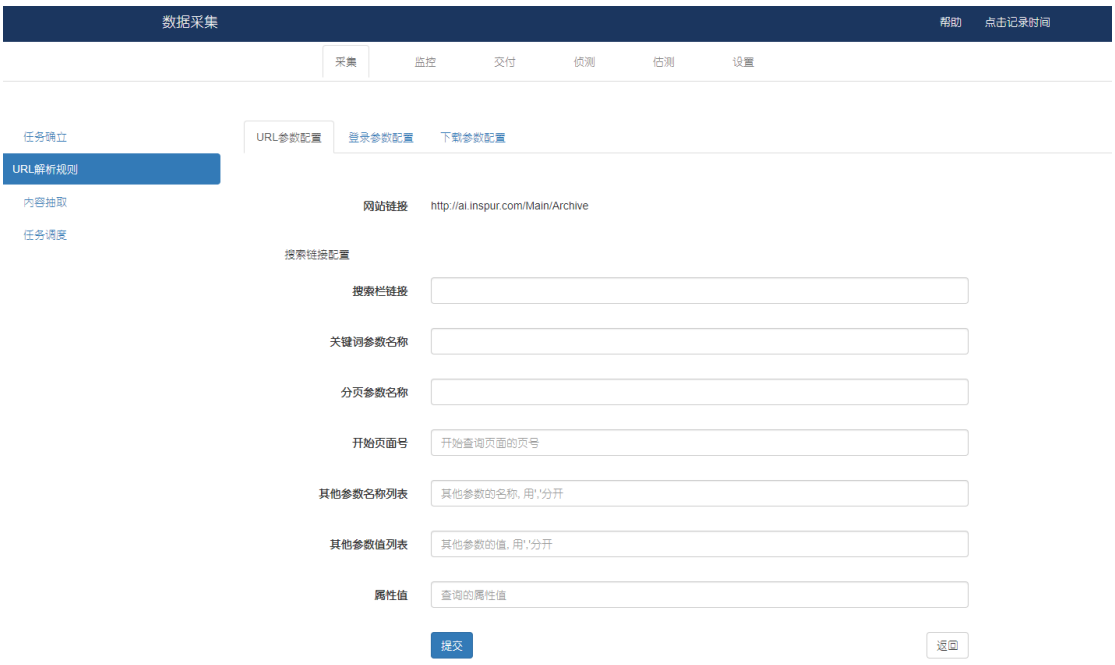
填写完成后，点击“提交”，提交任务，将提示是否提交成功，见下图，点击确定，将进入 URL 解析规则配置页面；

（3）URL 解析规则配置页面将显示所有新建的任务及其相应信息，如下图所示；点击对应任务的“修改”按钮，进入相应任务 URL 解析规则配置页面。

例：点击结构型任务编号 149 对应的“修改”，进入任务 URL 解析规则配置页面。


（4）URL 解析规则配置分为 URL 参数配置、登录参数配置和下载参数配置，以下分别进行具体描述：

1）URL 参数配置




具体参数信息填写说明如下：

- 搜索栏链接：目标网站的接口访问链接例：http://ai.inspur.com/Archive/PoorFamilyList-GetPoorFamilyData
- 关键词参数名称：所有查询关键词名称，多个用，隔开。例：poorproperty,poorcause,planOutPoor,realname,name6,basicArea,txtYear,Aad105,isHelp,isHelpPeople,isImmigrant,isPlan,AreaType,Aah006,Aad003,condition,membercondition,orders,sorts,poorFamilyType
- 分页参数名称：在与接口交互时，标识页数的参数名称。例：pagenumber
- 开始页面号：：此处分为两个部分：逗号之前参数表示开始的页数，逗号之后的参数表示查询的下一页页号数的增幅。例：1,1
- 其他参数名称列表：在访问目标网站时，固定不变的参数名称列表。例：isNull,pagesize
- 其他参数值列表: 在访问目标网站时，固定不变的参数值列表，与对应于其他参数名称。例：0,1000
- 页面数据大小：单个页面返回的数据条数。例：1000
- 数据总量地址：**填入指定在搜索接口返回的 JSON 数据中数据总量的位置。**例：/d/total
- Jsons 数据地址：**填入指定在搜索接口返回的 JSON 数据中内容数组的位置。**例：/d/rows
- 属性值：关键词参数的可选属性列表，不同参数属性用分号隔开，相同参数的属性用逗号隔开。

  例：一般贫困户,低保户,五保户,低保户贫困户,一般农户,五保贫困户;因病,因残,因学,因灾,缺土地,缺水,缺技术,缺劳力,缺资金,交通条件落后,自身发展动力不足,其他,未填写

2）登录参数配置

下图为登录参数配置页面，若目标网站不需要登录即可进行爬取，则该部分参数可不填写，在本例中，需要进行登录参数配置，具体参数填写说明如下：


- 登录界面链接：所爬取网站用户登录页面的网址　　例：http://ai.inspur.com/login
- 用户名输入框 Xpath：用户名输入框的 id，可在登录页面，通过浏览器右键“检查”进行获取。　　例： txtUserName
- 密码输入框 Xpath：密码输入框的 id，方法同上。　　例： txtPassword
- 登录按钮 Xpath：登录按钮的 id　　例： btnLogin
- 用户名：有效的用户登录名　　例： 431200000000
- 密码：有效的用户登录密码　　例： aaaaaa

3）下载参数配置

同基于链接 2.1.1（4）3）

（5）内容抽取

基于接口的结构型数据不需要配置模板参数。

（6）任务调度

同基于链接（2.1.1（6））

## 2.2非结构型数据采集

非结构型数据采集，即针对文本类数据网站采集。文本类数据网站搜索结果页面的数据刷新方式主要是基于页面刷新。在基于页面刷新的模式中，不同的搜索结果页面将会处于不同的 URL 下，整个结果页面在服务器后端渲染完毕之后再发送到浏览器端。

在本应用中，针对这种模式会有其对应的爬虫程序对指定网站进行数据爬取。下面，以一个例子来阐述这种模式的配置和启动过程。

### 2.2.1 基于页面刷新

在本例子中，所爬取的目标网站为 中央财经大学官网，目标数据为官网上搜索引擎内部的数据。具体步骤如下：

**（1） 任务确立**

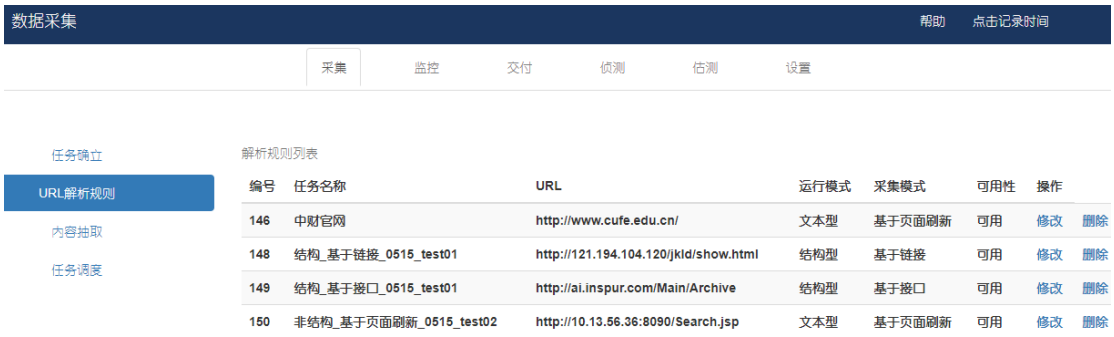

首先，在任务名称中填入本次爬取任务的名称，该名称是此任务的唯一标识。
其次，选择运行模式和数据刷新方式，在基于页面刷新的模式下，运行模式选择非结构型，数据刷新方式选择基于页面刷新。
之后，在网站链接栏填入所爬取目标网站的首页链接，爬虫将会从该链接所指向的目的页面中随机选取第一个用于搜索的关键词。
然后，在工作路径后缀栏中填入本任务在服务器文件系统中的工作路径后缀。请注意，工作路径前缀在网站部署时已经由管理员在顶栏的

设置栏中配置完毕，这里仅能配置工作路径后缀。

最后，点击提交按钮，提交此任务。

（2） URL 解析规则配置

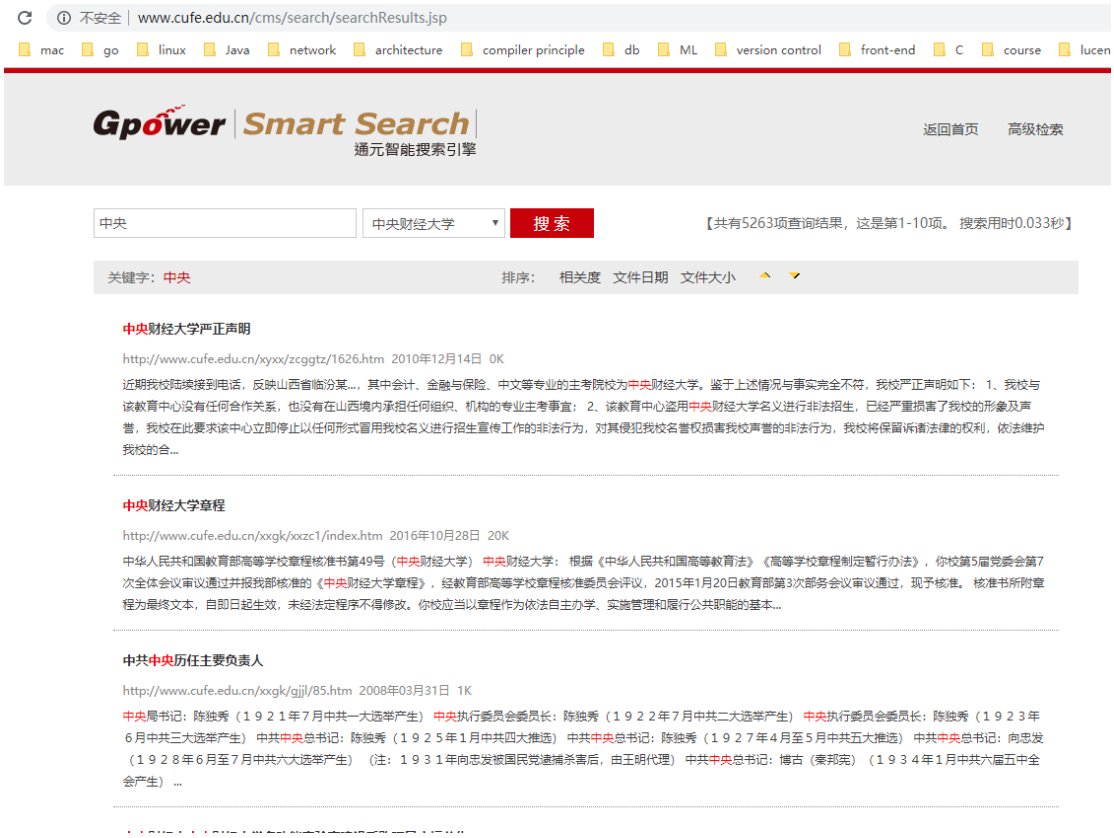
任务提交完成后，点击左侧的 URL 解析规则按钮，到达 URL 解析规则列表页面。如下：




根据之前所填写的任务名称，在页面的显示列表中找到对应解析规则项，对于新创建的任务，其解析规则项的可用性将显示为不可用。点击该项的修改按钮，进入解析规则配置页面进行具体规则配置。若在任务确立过程中存在参数误填操作，可点击对应的删除按钮删除该任务。

1）URL 参数配置

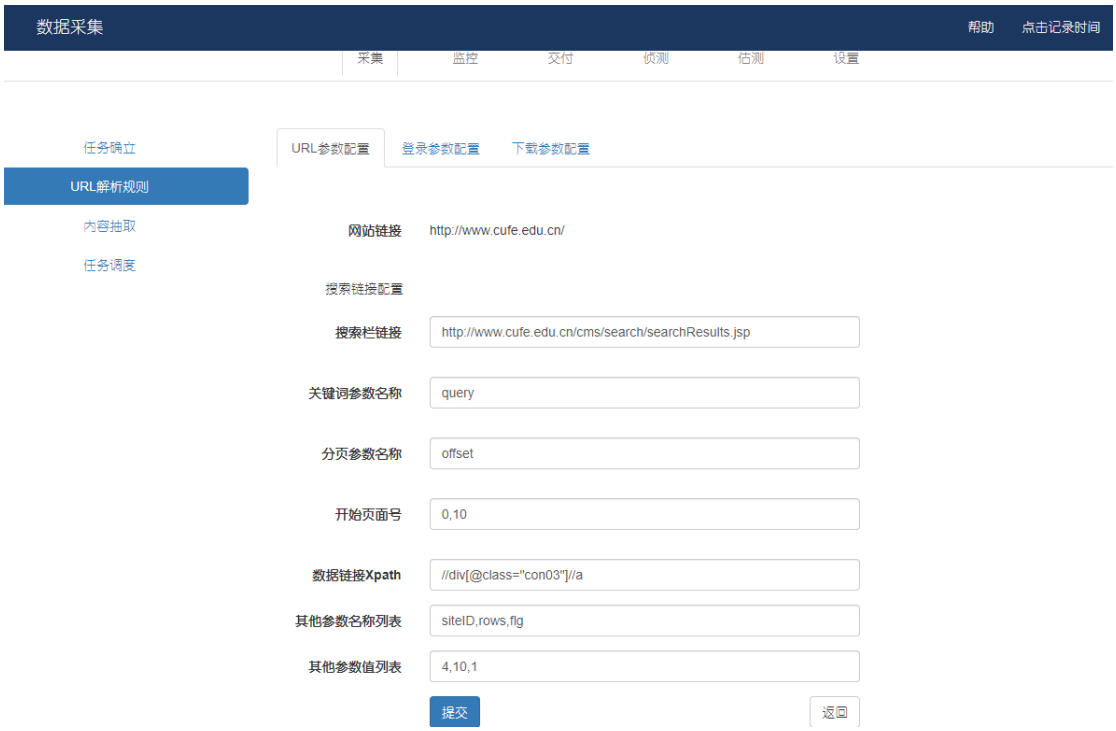
如图所示，下图是中财官网搜索页面：




其页面链接如下：

http://www.cufe.edu.cn/cms/search/searchResults.jsp

根据对该搜索引擎的仔细研究，对 URL 参数进行配置，如下：




首先，在搜索栏链接处填入搜索页面链接前缀：

http://www.cufe.edu.cn/cms/search/searchResults.jsp

从控制台观察完整的搜索请求，在关键词参数名称和分页参数名称两栏分别填入对应的参数。

关键词参数名称即所提交的搜索词的参数名称。

分页参数名称即指定搜索结果第几页的参数名称。

开始页面号这一栏填入的是一个组合参数，由两个整数值构成，用英文的逗号隔开。第一个整数值代表搜索结果页面第一个分页的 URL 中分页参数的值。第二个整数值代表的是，在搜索结果的多个分页中，相邻分页的 URL 中分页参数的差值。

数据连接 Xpath 指的是搜索结果页面中有效数据链接的 Xpath 值。该值用于帮助爬虫快速从搜索结果页面中筛选出有效链接。

相对于关键词参数和分页参数这两个变量，搜索结果页面的 URL 中如果有其他必须携带的常量，则将常量参数的参数名称填入其他参数名称列表栏中，多个参数名称用英文逗号隔开，同理，将参数值填入其他参数值列表栏中。参数名称和参数值的顺序一一对应。

配置完毕，点击提交按钮提交配置值信息。

2）登录参数配置

URL 参数配置填写并提交完毕，点击顶栏的登录参数配置按钮，进入登录参数配置页面。如下：


若对所爬取数据的爬取过程中无需登录验证，则该页面内容可不填写。在对诏安县官网的观察中发现无需在爬取过程中进行登录验证，因此在本例中无需填写登录参数配置。

点击顶栏的下载参数配置按钮，进入下载参数配置页面。

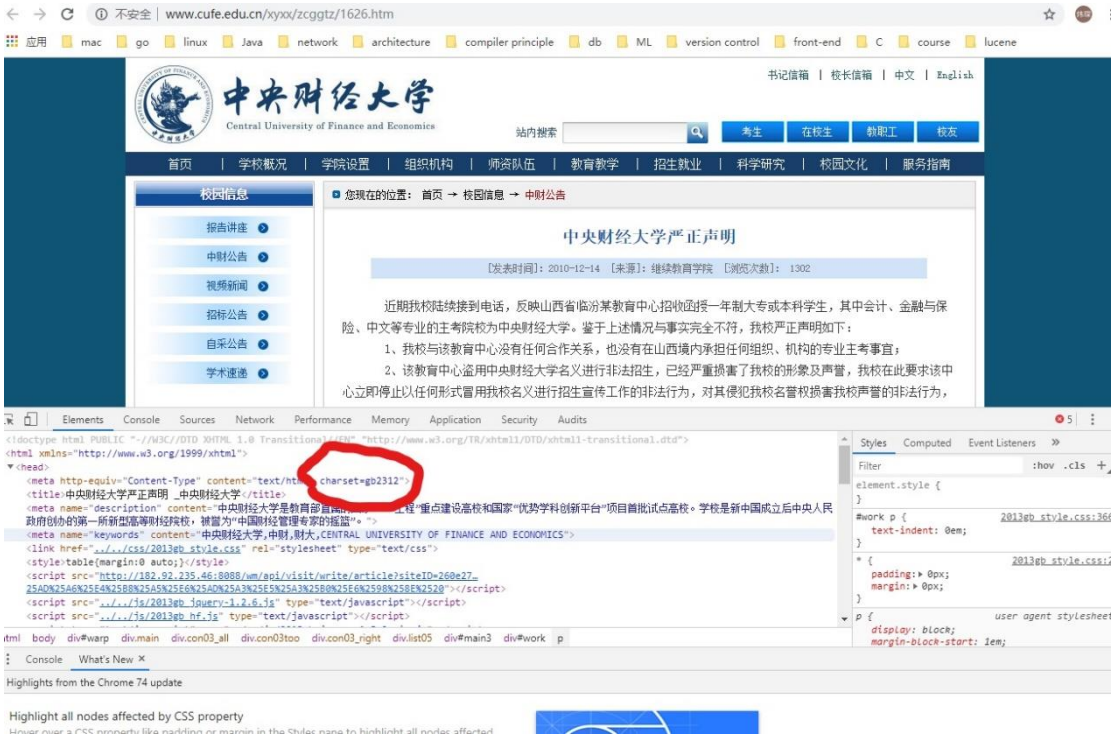
3）下载参数配置


该页面配置的爬虫运行过程中的一些系统性配置信息。

首先在线程总数一栏填入分配给改任务的线程数量。此参数不宜设置过大。设置过大会导致爬虫启动过多线程，当线程数量严重超越 CPU 核心数量时，会导致爬虫运行过程中 CPU 资源浪费在频繁的线程上下文切换中。对于个人电脑，此处建议值为 5，对于商业服务器，此处建议值为 20。具体值可参照操作系统参数中的 CPU 核心数量进行微调。

接着在超时时间一栏填入爬虫使用的 HTTP 连接中的超时时间。对于网速较好的情况下，建议设为 3000。在网速较慢的情况下，建议酌情加大该值。正常情况下，该值宜大不宜小。

然后，在编码格式一栏填入所爬取网站页面所使用的编码格式。该栏填写错误将导致所下载数据编码错误，数据打开为乱码。如下，该值可在目标网站页面的 head 中查看。




在数据总量一栏填入所爬取数据的数据总量。该值用于衡量爬取进度。若无法获知该值，可填入 0。

最后，点击提交按钮，提交所配置信息。

### （3） 内容抽取

配置完 URL 解析规则，点击左栏中的内容抽取，进入模板列表页面。


该页面显示内容为为当前所配置的所有数据抽取模板。

一个爬取任务在爬取过程中会建立对所爬取任务的内容索引。每个爬取任务可以拥有多个模板，每个模板对应索引中的一个字段，字段名称即模板的名称。默认情况下非结构型爬取方式所建立的索引中都会有一个 fulltext 字段，也就是会有一个 fulltext 模板，该字段的内容是所爬取数据的全文。该模板默认不显示在模板列表中。

注意，在本例中仅使用默认的 fulltext 模板，因此下述内容仅作为功能介绍，无需真实配置。

若想在任务中新增索引字段，可点击顶栏的新建模板按钮跳转到新建模板页面，如下：


根据所建立模板将要绑定的任务类型对运行模式进行选择。在绑定任务的下拉列表中选择所要绑定的任务。对于非结构型任务的模板，其模板名称即所新建的索引字段的名称。

在模板 Xpath 栏填入所要打入字段的值在 HTML 页面中的 Xpath 路径。

填写完毕，点击提交按钮提交配置信息。

若想删除某个模板，可在模板列表页面中点击对应按钮进行删除操作。

（4） 任务调度

到此，对新建任务所需要的配置信息已经填写完毕。可点击左侧任务调度按钮进入任务调度列表页面，如下：


任务调度列表中仅显示配置完毕、可进行调度的任务。

选择所要操作的任务项。点击右侧操作栏中的启动连接，即可启动对应的爬取任务，如下：


同理，在爬取任务运行过程中，点击停止按钮，即可停止所启动的爬虫进程。

在爬取结束后，若想对对应任务进行删除，可点击删除按钮进行操作。注意删除操作将删除该任务的配置信息和所有已下载数据。

# 3. 监控

对已经配置好的任务进行监控，显示爬取任务的状态、采集轮数、下载数据量和爬取比例。让用户对采集任务可以实时观察采集任务的状态。

## 3.1 任务采集监控：

实时监控任务采集状态，(点击监控会跳到任务采集监控模块)


编号：前面步骤已经配置的网站编号

任务名称：前面步骤已经配置的任务名称

所处轮次：爬取任务所处的爬取的轮次(任务爬取会一轮一轮进行爬取)

已下载数据量：任务已经下载的数据量

爬取比例：已爬取数据量占网站数据总量的比例

任务状态：当前任务的状态（如果任务已经启动则显示启动，任务停止则显示未启动。

# 4. 侦测

用于对已经配置好的网站进行接口侦测，从网站首页不断进行遍历，获取网站所有含有搜索接口的链接，便于用户利用接口进行深网数据爬取,需要注意的是此功能只能侦测 form 形式的搜索接口。

## 4.1 侦测接口

点击侦测->会自动跳到侦测接口界面。


编号：前面步骤已经配置的网站编号

网站名称：前面步骤已经配置的网站名称

url：前面步骤已经配置的网站链接

已爬接口数目：显示已经侦测到有搜索接口的网站子链接数目，此数目会随着侦测过程实时刷新。

已测链接数量：显示已经侦测过的网站子链接数目（这些子链接可能有搜索接口，也可能没有搜索接口），此数目会随着侦测过程实时刷新。

状态：此网站是否处正在侦测（如果在进行侦测，则显示 start，如果未进行侦测，则显示 stop）。

展示：点击查看结果可以查看此网站已经侦测到含有搜索接口的子链接。

操作：对网站进行启动和停止侦测操作。点击启动可以对网站进行开始侦测操作，点击停止可以对网站进行停止操作。

例 ：如果想对编号 122 网站名称为“诏安县政府官网”的网站进行接口侦测。

点击启动按钮->出现操作成功提示


已爬接口数目和已测链接数目会不断刷新，如启动一段时间，显示变成：


## 4.2 侦测结果

->点击查看结果，会自动跳到侦测结果页面，并显示已经侦测到有搜索接口的子链接。


点击右上角 x 符合返回侦测网站列表界面

# 5. 估测功能

估测功能主要是对已经配置好网站参数的网站进行估测，从而确定网站的搜索接口背后的索引文档的数量。下面描述估测功能的操作流程。

## 5.1 参数配置

### 5.1.1 非结构化网站

估测功能的参数配置需要在网站参数配置完成的基础上进行，没有配置网站参数的估测需求不会显示在估测列表中。配置网站参数需要在采集功能下进行：先在任务确立模块进行任务确立，然后在 URL 解析规则中点击相对应的［修改］按钮进行配置。

网站参数配置完成后，任务出现在估测列表中，点击修改按钮，出现如下界面。估测参数配置输入后点击提交，完成配置。


下面以某学校官网为例，说明估测参数的配置方法：

1. 结果链接 Xpath：

使用该网站的搜索接口进行任意关键词的搜索，出现的结果页面由一组可以点击的超链接组成，点击 F12 键，选中超链接区域，即可查看相应的 Xpath，也就是结果链接 Xpath。

如下图所示，查看该学校搜索页面的结果链接 Xpath，对应的 class 是 con03，

因此将 con03 填入“结果链接 Xpath”。


2. 文档内容 Xpath

任意点击一个文档链接，用上述同样的方法，在浏览器中查看文档内容的 Xpath，填入表中即可。

如下图，显示该网站的文档内容 Xpath 是 list05。


3. 种子关键词

种子关键词的填写一般根据具体的搜索接口而定，在填写之前可以对搜索接

口进行尝试，

一般使种子关键词产生的搜索结果大于一页。此处填写“的”字作为种子关键词。

4. 游走次数

一般情况下，游走次数越多，估测结果越精确，但同时也会使耗时越长。建议的游走次数为 1500。

5. 文档位置

文档位置默认为正文，但对于部分仅对标题索引的网站，需要选择标题，以减少游走过程中发送请求返回为空的情况。

6. 请求方式

一般而言，网站的搜索接口发送请求方式分为 GET 请求方式和 POST 请求方式。此处默认为 GET,但对于使用 POST 发送请求的接口，需要改填为 POST。

估测参数配置完成后如下图所示，点击提交按钮进行提交。


### 5.1.2 结构化网站

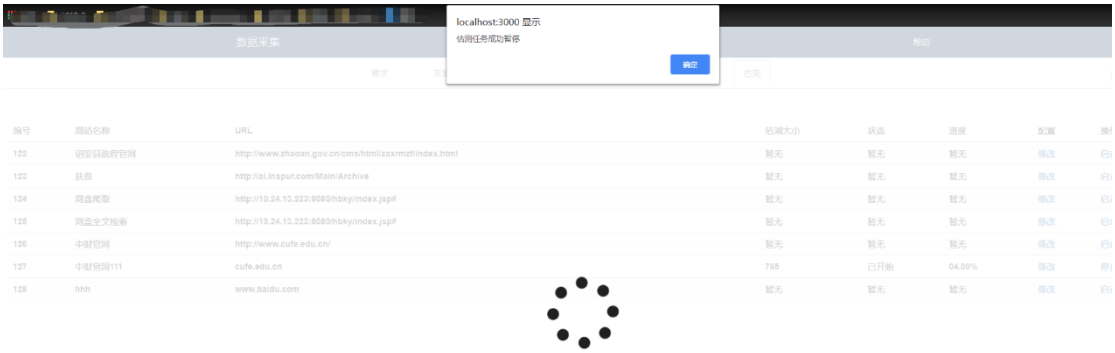
对于结构化网站，不需要在估测模块进行额外的参数配置，只需要在采集模块中配置完成即可。

## 5.2 估测控制

配置完成后，点击启动按钮，估测任务便进入[已启动]状态。估测列表中实时显示当前的估测进度，估测结果和估测状态。


若要终止估测任务，点击停止按钮，估测任务便会终止。并保存终止时刻的估测结果，仍然显示在列表中。

如下图，停止估测任务，系统提示“估测任务成功暂停”。




# 6. 交付

将任务已经爬取的数据交付给用户，其中结构型数据的交付方式暂时为让用户在线查看已爬取数据，文本型数据的交付方式为让用户下载已经爬取的数据。


点击交付：

编号：前面步骤已经配置的网站编号

交付任务：前面步骤已经配置的任务名称

运行模式：任务运行的模式

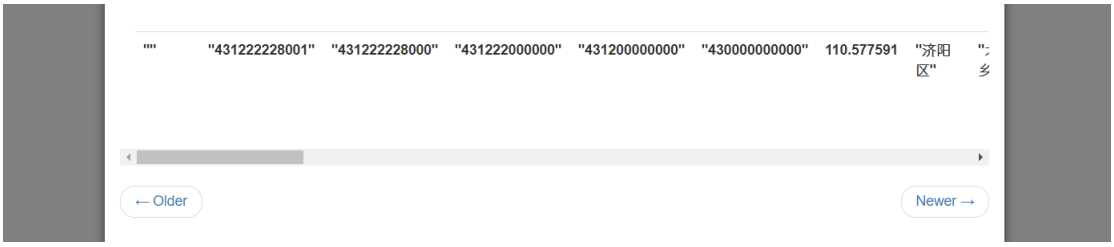
采集模式：任务采集数据的模式

查看：点击即可查看已经爬取的数据

例：如果是结构型数据：

点击编号为 132 的任务的查看，显示任务的数据


将进度条下移，发现有按钮 older 和 newer，点击可查看前一页的数据和后一页的**数据**



如果是文本型数据，将直接下载数据压缩包