

采集需求

网站：<http://www.ccqp-shanxi.gov.cn/view.php?app=&type=&nav=104&page=1>

我们需要采集两种类型的页面

- 两个页面都需要 `bid_url_id` 是为了将两块内容对应起来。

配置页面抽取规则

页面配置

[illegible]

1. 规则名称：必填，项目的名称
2. 页面正则：该类型页面 url 的正则表达式，必填
3. 代理配置：使用何种代理，nil 表示不采用代理
4. 使用Bot：暂时无用
5. 下载超时：自定义网页下载超时时间，默认 60
6. User-Agent：访问网站需要使用的 User-Agent，默认会随机生成
7. Referer：访问网页需要使用的 referer，默认为该页面 url
8. headers：字典，需要使用的 headers

9. cookies : 字典，需要使用的 cookies

下面我们开始定义抽取规则

行配置

根据上篇文章的建模，我们首先需要定义行，对于 html 网页来说，一般我们采用 XPath 定位。

打开对应的网址，点击 Xpath Generator，然后测试后，得到行的 XPath 表达式



如上图所示，我们得到了行的表达式：`//*[@style="height:30px;"]`。在泛采系统中点击新建行，并填入对应的表达式

新增行

填写好后如下图所示：

Row	新增行 删除行 重置行		
行名称	<input type="text" value="data"/>	描述	<input type="text"/>
存储配置	<input type="text" value="招投标列表存储"/>	定位方式	<input type="text" value="xpath"/>
表达式	<input type="text" value="//div[@style='height:30px;']"/>		

1. 行的名称自定：不影响存储，保证在该页面中唯一即可。
2. 存储配置：指定改行数据如何存储，我们暂时忽略这个配置，随后会说到
3. 定位方式：改行的定位方式，在这里我们选择 xpath
4. 表达式：定位的表达式，我们填写生成的 `//*[@style="height:30px;"]` 这个值

字段配置

好了，有了行的定义后，我们为这一行增加字段，点击增加字段按钮

新增字段

按照之前的方式，我们填写改行中需要抽取的字段

字段	新增字段		
字段名称	<input type="text" value="bid_title"/>	描述	<input type="text"/>
定位方式	<input type="text" value="xpath"/>	表达式	<input type="text" value="//a[@target='_blank']/@title"/>
数组	<input type="text" value="请选择"/>	保存HTML	<input type="text" value="请选择"/>
拓展URL	<input type="text" value="请选择"/>		
新增字段			

1. 字段名称：这个字段名称为 bid_title
2. 描述：可以留空

3. 定位方式：依然选择 xpath
4. 表达式：这里也是填 xpath 表达式，注意这个表达式必须是相对于行的表达式
5. 数组：表示抽取的字段是否是多个值
6. 保存 HTML：是否保留字段的 HTML 结构
7. 拓展URL：根据当前页面 url，补全抽取字段为绝对地址

依次配置好其他的字段。

测试抽取

当我们配置了一些字段以后，可以测试抽取效果：

测试用例

供子 测试 测试用例

方法 Url

Post Data

测试用例

点击新增例子按钮，添加一个例子。选择 HTTP 方法，输入链接，点击测试按钮

测试结果分为三部分，第一部分是日志，可以用来排查错误：

```

日志列表

skip_validate=true, 跳过验证环节

not global dedup key found, skip dedup

[load_rule] Rule already loaded for http://www.ccgp-shanxi.gov.cn/view.php?app=tsar=104page=1&type=
[preload] skip_cache is True, skip preload

[load_user] not loading users, skip

crawl_doc_proxy is adsl

[proxy_stage] proxy(namepage=adsl) not found

[download] using download_url = http://www.ccgp-shanxi.gov.cn/view.php?app=1&type=tsar=104page=1
[download] start downloading http://www.ccgp-shanxi.gov.cn/view.php?app=tsar=104page=1&type=
[download] download http://www.ccgp-shanxi.gov.cn/view.php?app=tsar=104page=1 is successful

[redirect] no meta refresh tag found

[validate_data] skip validate

decodes page is successful

[extract] using local extractor

[extract_data] doc crawl=10 for url=http://www.ccgp-shanxi.gov.cn/view.php?app=tsar=104page=1&type=

提取成功!文档 10

```

第二部分是结果列表，我们可以看到抽取到了想要的字段：

行名称	类型	字段名/链接规则	内容
data	字段	bid_title	晋关县住房保障和城乡建设管理局晋关县粮食仓库改造工程施工公告
data	字段	bid_url	http://www.ccgo-shanxi.gov.cn/view.php?id=506275
data	字段	bid_area	山西省-长治市-晋关县
data	字段	bid_date	2019-03-09
data	字段	bid_source	ccgo_shanxi
data	字段	bid_url_id	506275

第三部分是网页预览，如果抽取结果有问题，那么我们可以看一下在下载到的页面中是否真的有这些字段

项目内容

2020/3/30 3:35:21 PM 星期一

项目 - 结果公告		名称	地区	状态	时间
招标公告		普兰县内供保障城多建设保障县居民县供内农舍改造...	长治市壶关县	已结束	(2019-03-09)
招标公告		保康县人防工程防、供热公司与家湖路改造工程施工...	忻州市保德县	已结束	(2019-03-08)
招标公告	第一中标候选人	绵阳市妇幼保健院医疗服务中心门诊业务搬迁项目招标...	运城市闻喜县	已结束	(2019-03-08)
招标公告		武乡县部分“两院和服务中心”以“电代煤”采暖设备招采...	长治市沁源县	已结束	(2019-03-08)
招标公告		武乡县部分“两院和服务中心”以“电代煤”电力配置工程...	长治市武乡县	已结束	(2019-03-08)
招标公告		普兰县内供保障城多建设保障县居民县供内农舍改造...	长治市壶关县	已结束	(2019-03-08)
中标公告		介休市妇幼保健院迁建晋能集团中心医院设备采购项目招...	晋中市介休市	已结束	(2019-03-08)
中标公告		大同市新城北区民生教育医疗安全应急服务项目中标公告	大同市新城北区	已结束	(2019-03-08)
中标公告		濮阳市人民防空工程新建项目招标公告	濮阳市濮阳县	已结束	(2019-03-08)
中标公告		2017年五台县农村饮水保障建设采购项目成交公告	运城市临猗县	已结束	(2019-03-08)

第1/15776页 前一页， 最后一页， 转到页 go

【关闭】

在上面的结果中我们可以看到已经抽取到了想要的字段。 [点击保存。](#)

我们可以按照上面的步骤再配置详情页的抽取。**最后一定要点击保存**

链接配置

列表页除了包含字段以外，还包含了详情页的链接，我们需要配置新链接这样，才能让泛采系统在采集列表页之后继续采集详情页：

链接

新增链接

页面配置

山西政府采购网详情页

描述

定位方式

xpath

表达式

[@contains(@href, "view.php?id=506275")]

新增链接

新增链接

1. 页面配置指的是新产生的链接对应的哪个配置，可选的范围是当前项目中的所有页面配置，在这里我们选择刚刚配置好的详情页。
2. 描述选填
3. 定位方式依然选择 xpath
4. 表达式填写生成链接的表达式

再次点击测试，我们看到抽取到了新的链接

结果列表			
行名称	类型	字段名/链接规则	内容
data	字段	bid_title	晋关县住房保障和城乡建设管理局关县粮食库仓库改造工程成交公告
data	字段	bid_url	http://www.ccgp-shanxi.gov.cn/view.php?id=506275
data	字段	bid_area	山西省-长治市-壶关县
data	字段	bid_date	2019-03-09
data	字段	bid_source	ccgp-shanxi
data	字段	bid_url_id	506275
data	链接	山西政府采购网详情页	http://www.ccgp-shanxi.gov.cn/view.php?id=506275

抽取规则配置到此结束。

存储规则配置

存储规则指定了每一行中的数据如何存储。在每一行都可以指定一个存储规则，然后查找对应的存储规则并存入对应的数据库中。

基本属性

* 存储名称

招投标列表存储

MySQL配置

是否启用

是

实例名称

mysql_4

Host

Port

3306(无数字)

用户名

密码

数据库

spider_data_5

编码

utf8mb4

表

bid_list

OSS配置

是否启用

否

测试模式

否

ServiceLine

模块

表

bid_list

1. 在上面的配置中，我们指定了启用 MySQL 存储，并禁用 OSS 存储。
2. 实例名称可以从 mysql_1，2，3，4，5 这5个变量中选，采用内置的配置，不需填写 host port 等字段。
3. 数据库是对应的数据库的名称
4. 编码一般填 utf8mb4，注意一定不能填写 utf8，mysql 的 utf8 不是真的 utf8
5. 表指定要存的表的名称

MySQL 中的表现在还需要手动创建。

最后，返回到页面抽取配置中，选择对应的存储配置：

Row

创建表

删除表

删除行

行名称

data

存储配置

招投标列表存储

定义

表达式

//tr[@style='height:30px;']

调度计划配置

调度计划用于指定对整个网站的抓取节奏。

[218]山西政府采购网 [218] (调度计划)

保存 测试

计划配置

计划说明 山西政府采购网

抓取速率 5 定时时间 Cron表达式

页面配置

类型 url

页面配置 山西政府采购网列表页

url http://www.ccgp-shanxi.gov.cn/view.php?app=&type=&nav=104&page={1..15800} 新增url

执行计划

Offset 1 跳过缓存 否

批次日期 按 运行日期

执行计划 新增计划 重置计划

1. 抓取速率定义了每秒最大下载多少网页，我们这里设定为 5，因为政府网站配置普遍较差，以免影响对方服务器性能
2. 定时时间是一个 Cron 表达式，用于指定合适触发更新
3. 类型指的是种子链接的生成方式，我们这里选择 url，也就是直接指定种子链接
4. 页面配置指的是我们当前指定的种子链接对应的页面配置，在这里我们选择对应的列表页

url 我们填写了：<http://www.ccgp-shanxi.gov.cn/view.php?app=&type=&nav=104&page={1..15800}>

注意其中的 {1..15800}，这表示这个链接世界上代表了 15799 个链接，他会被自动拓展成为：

1. <http://www.ccgp-shanxi.gov.cn/view.php?app=&type=&nav=104&page=1>
2. <http://www.ccgp-shanxi.gov.cn/view.php?app=&type=&nav=104&page=2>
3. ...
4. <http://www.ccgp-shanxi.gov.cn/view.php?app=&type=&nav=104&page=15799>

其他模式的种子链接可以见高级教程。

手工触发

选择批次日期之后，点击执行计划按钮，可以手工触发当前计划执行，稍等片刻，就可以在 MySQL 中看到抓取的数据了