三、使用教程——以招投标网站为例

采集需求

采集山西政府采购网的所有标书信息。

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

名称	地区	状态	时间
襄汾县公安局襄汾县公安局 2019年度“天眼工程”...	临汾市襄汾县	已结束	[2019-03-05]
交口县财政局交口县财政授权支付代理银行项目中标公告	吕梁市交口县	已结束	[2019-03-05]
运城市盐湖区魏风小学教学楼改扩建建设项目中标公告	运城市盐湖区	已结束	[2019-03-05]
运城市盐湖区科技实验小学绿化工程中标公告	运城市盐湖区	已结束	[2019-03-05]
运城市盐湖区魏风小学后稷校区绿化工程中标公告	运城市盐湖区	已结束	[2019-03-05]
云冈区机关一食堂餐饮委托管理服务采购中标公告	大同市南郊区	已结束	[2019-03-05]
潞城市农业委员会旱作节水农业和地膜减量增效推广技术...	长治市潞城市	已结束	[2019-03-05]
黎城县污水处理中心提标改造工程PPP项目预中标结果...	长治市黎城县	已结束	[2019-03-05]
沁源县“四好农村路”韩洪乡两条道路改造工程监理项目...	长治市沁源县	已结束	[2019-03-05]
沁源县2018年赤石桥乡倪庄村等四村土地开发项目中...	长治市沁源县	已结束	[2019-03-05]

第1/15762页 后一页 › 最后一页 » 转到第 页 GO

我们需要采集两种类型的页面

1. 列表页，所需字段及字段名：
    a. 名称：bit_title
    b. 链接：bid_url
    c. 地区：bid_area
    d. 时间：bid_date
    e. 站内 ID：bid_url_id
2. 详情页，所需字段
    a. 标书内容 bid_content
    b. 站内 ID：bid_url_id

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

首先我们来配置列表页的抽取规则。在泛采系统中新建项目和分组，并创建页面规则配置。创建之后如下图所示：

图片中的页面级别字段包括：

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，选填
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

填写好后如下图所示：

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

按照之前的方式，我们填写改行中需要抽取的字段

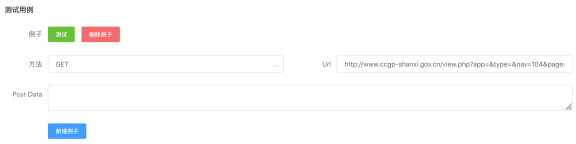
1. 字段名称：这个字段名称为 bid_title
2. 描述：可以留空

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：



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：

第二部分是结果列表，我们可以看到抽取到了想要的字段：

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

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：

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

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。

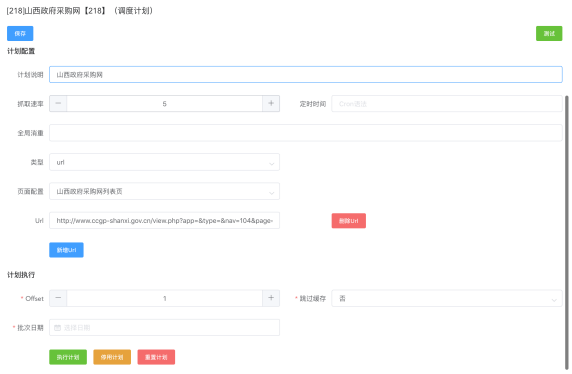
1. 在上面的配置中，我们指定了启用 MySQL 存储，并禁用 OSS 存储。
2. 实例名称可以从 mysql_1，2，3，4，5 这5个变量中选，采用内置的配置，不需填写 host port 等字段。
3. 数据库是对应的数据库的名称
4. 编码一般填 utf8mb4，注意一定不能填写 utf8，mysql 的 utf8 不是真的 utf8
5. 表指定要存的表的名称

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：

调度计划配置

调度计划用于指定对整个网站的抓取节奏。



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