熵简泛采系统使用文档

概述

泛采系统中可配置规则有三种，调度计划、页面规则、存储配置。简单理解一下：

调度计划：控制何时如何发布任务

页面规则：如何下载、解析、校验网页数据

存储配置：数据存储到哪里去

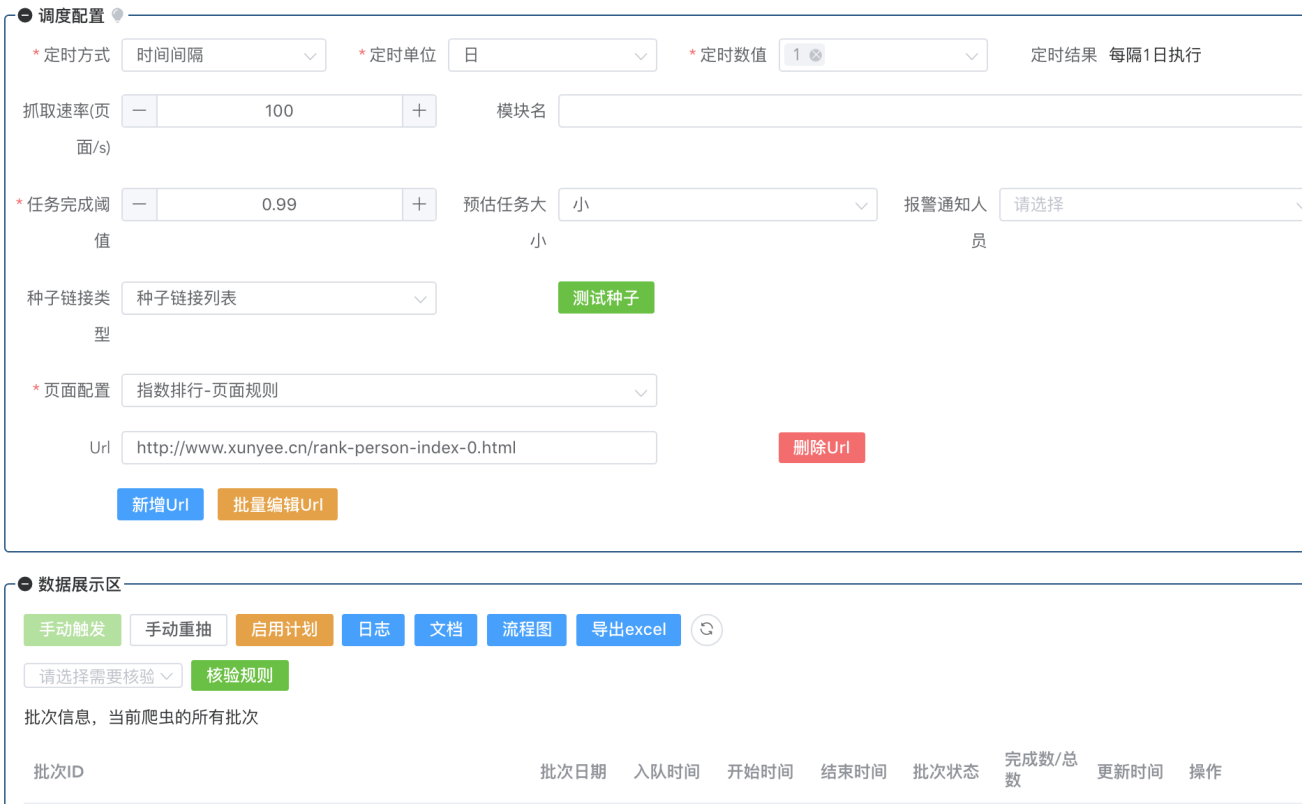
项目、组、规则层级结构

泛采系统采用目录结构组织其中配置的规则。

不同项目中间的规则不可互相关联。

规则配置说明

调度计划



概念

控制采集任务何时被发布，采集任务的启动可设置为周期或单次执行，在启动时会生成采集任务的种子链接。

若启动多次，则称每次启动为一个批次。

区域-基本属性

此区域配置调度计划的基本描述属性：如名称、描述

按钮说明：

保存：保存当前页面修改内容到数据库

查看源文件：查看当前页面的配置生成的yaml文件内容，此项为高级功能

区域-调度配置

此区域为调度计划执行属性，包含以下可配置项：

- 定时方式：

手动触发：非定时任务、仅在用户主动启动时此调度计划开始执行

时刻：每当时间满足所设置时刻时，调度计划开始执行。例如每日8点、每周一

时间间隔：每当经过所设置时间间隔时，调度计划开始执行。例如每1个小时、每1天

注意：若前一批次尚未完成，下一批次不会开始

- 抓取速率：设置爬虫采集速度。单位为页面每秒，默认1，即每秒1个页面。注意此速率为上限。
- 任务完成阈值：本批次任务最小完成度，默认0.99。用于判断本批次任务是否完成。
- 预估任务大小：预估单次采集任务重所下载的全部页面的数量量级，一般为小，若超过百万则可设置大
- 模块名：略
- 报警通知人员：略
- 页面配置：种子链接对应的解析页面规则
- 种子链接类型：

种子链接生成规则类型，可选项如下。

1. 种子链接列表：手工填入种子链接，该链接支持拓展特性：

   expand a string with {0..100} to 100 strings
   expand a string with {0..100..1} to 50 strings
   expand a string with {2019-05-01..2019-05-21..1}
   expand a string with {2019-05-01..2019-05-21..2}
   expand a string with {20190501..20190521..1..%Y%m%d}
   expand a string with {now(%Y)} to 2019
   now(%Y-%m-%d)--->2019-05-21
   now(%Y%m%d)--->20190521
   now(%Y%m%d)-1 --->20190521
2. 脚本：编写python脚本生成种子链接

```
#
def start_requests(**kwargs):
    """
        yield
    :param kwargs:
                run_obj:
                logger:
                table_importer:   for field in table_importer(table_name): pass
    :return:
    """
    for page in range(1, 100):
        url = "http://www.boyar.cn/column/14.html?categoryid=1&page={}".format(page)
        yield {"url": url, "headers": {}, "data": b""}
```

3. 基础任务表：可选择从某个表中导入任务。但此表必须是某个页面规则中的新链接生成的。

区域-数据展示

此区域主要展示每个批次的执行情况

按钮说明：

手动触发：手动触发一次调度计划检查操作，调度计划是否会开始运行，取决于定时配置以及当前是否有未完成批次

~~手动重抽：废弃~~

停用计划：当处于停用状态时，定时配置不起效

日志：查看爬虫运行日志

文档：查看项目文档（自动生成）

流程图：查看项目流程图（各个规则之间的调用关系）

~~导出数据：废弃~~

刷新：刷新批次信息

批次记录操作说明：

缓存队列：查看当前批次未完成的任务

~~运行结果：废弃~~

废弃批次：废弃当前批次

删除批次：删除当前批次

页面抽取规则

概念

页面抽取规则用于指定某一类页面的下载方式和字段抽取规则。通过使用 XPath、CSS selector、正则表达式、JsonPath 等表达式，可以抽取 HTML、XML、JSON 等多种页面格式。

区域-基本属性

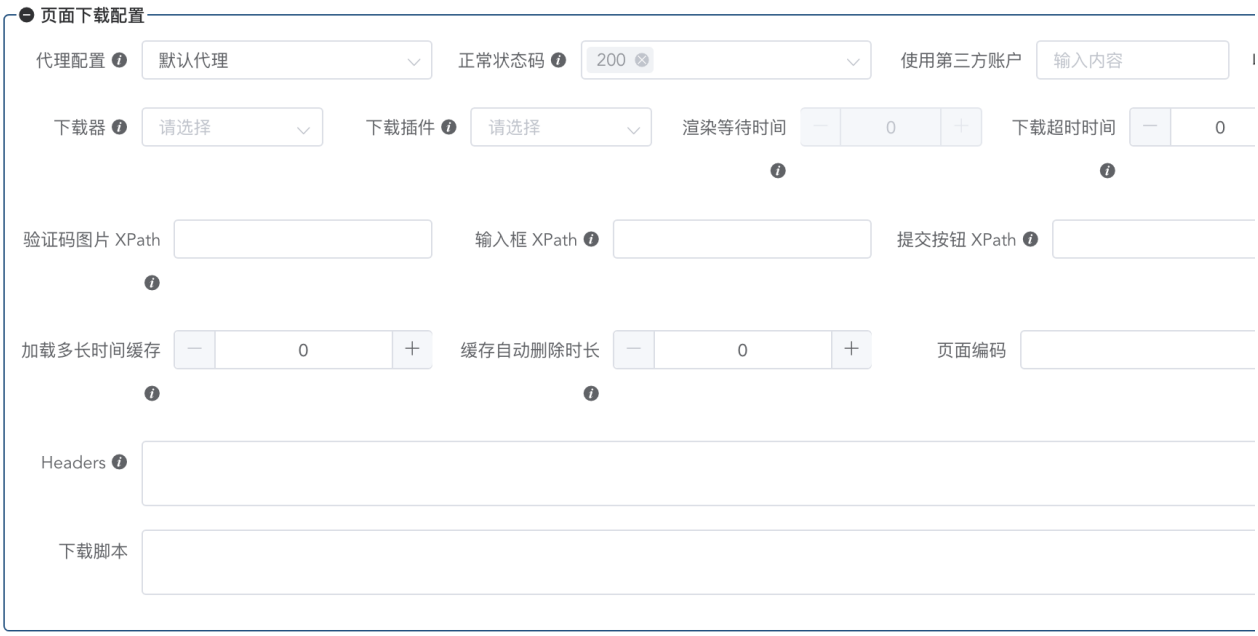
此区域配置调度计划的基本描述属性：如名称、描述

按钮说明：

保存：保存当前页面修改内容到数据库

查看源文件：查看当前页面的配置生成的 yaml 文件内容，此项为高级功能

区域-页面下载配置



代理配置：

用于指定爬虫下载页面时使用的代理，有n个选项：1.不使用代理 2.使用某种代理

对于爬取量小的页面类型（比如几百个到上千个网页）可以选择不使用代理。

对于爬取量较多的页面类型建议使用默认代理。另外，一般情况下，不使用代理访问速度会比使用代理更快一些。

正常状态码：

用于指定哪些 HTTP 状态码被认为是正常的。一般情况下，使用默认值 200 即可。但某些不规范的网站会在正常页面

返回 404、501 等错误状态码，这时需要添加上这些状态码以便爬虫正常运行。

~~使用第三方账户：废弃~~

页面编码：

用于指定页面的文本编码，比如 gbk，utf-8 等。

一般情况下无需填写，爬虫系统会根据网页内容智能识别页面的编码。

但某些特殊情况下，系统判断编码会失败，此时需要手动指定编码。

下载器：

用于选择使用普通下载还是浏览器渲染下载。

对于采用 Ajax 加载的动态网页需要使用浏览器渲染。

渲染等待时间：

用于配置使用浏览器渲染时，等待 JS 加载的时间。

下载超时时间：

用于配置下载时最长等待服务器响应的时间。

Headers：

用于填写下载该类页面需要的一些特殊头部，格式为JSON字典或行格式的。

下载脚本：

某些需要特殊处理的页面可能需要手动写脚本进行下载。需要定义download函数。参见 各种脚本默认代码

区域-页面去重配置

此区域定义是否将下载成功的页面加入去重库。

预估任务大小：用于选择不同类型的去重库。小：Redis的ZSET 大：Bloomfilter

区域-页面抽取配置

页面抽取配置用于定义从页面中抽取数据的规则。

关于行的定义

1.行可理解为一组相同类别的数据的集合。

2.我们一个网页定义为若干行的集合。

3.每个行可产生一条或多条数据，每条数据中包含一个或多个字段。

比如：对于一个新闻列表页，我们定义全部新闻标题所在区域的集合为一个行，故此行会产生很多条数据。

对一个新闻详情页，我们定义由此页面整体作为元素的集合为一个行，故此行仅产生一条数据。

行产生数据的数量取决于定义行时使用的定位方式所匹配到的位置的数量。

行操作

新增行：

点击新增行按钮可以定义一个新的行。

名称、描述：行的基本信息。

内容匹配正则和、URL匹配正则：使用正则匹配对应区域（源码或URL），用于当同一类页面拥有多种模板时区分不同的抽取规则，一般情况下无需填写。

存储配置：选择此行产生的数据应如何存储。

定位方式：用于选择如何定位本行在页面中的位置。

表达式：定位方式对应的参数。

导入行：

通过使用定义好的YAML规则生成行。（有配套按钮：导出行，用于跨配置复制粘贴）

~~批量新增行：废弃~~

行快捷操作：

	名称	描述	内容匹配正则	URL匹配正则	存储配置	定位方式	表达式	操作
>	数据行				指数排行-存储	Xpath	//a[@cla…	

上图中操作栏的按钮分别为：

编辑：编辑行属性。

删除：删除行。

创建表：使用行中的字段配置以及行的存储配置自动建表。

导出：将定义此行的YAML规则导出至剪切板。

字段操作

在建立行之后，可以点击新增字段按钮新增该行的字段。

名称、描述：行的基本信息，这两个信息在自动建表时会被用于数据库的字段名称和注释信息。

字段类型：用于标记数据库中的字段类型。

定位方式、表达式：和行中的类似，但是这里是相对行的表达式。

消重：表示该字段是否参与消重。

数组：表示该字段是否为一个会选取多个元素的字段，如果选取元素为多个，则存储为 JSON 数组的格式。

保存HTML：表示当选择的元素为 HTML 元素时是否保存 HTML 标签。

拓展 URL：表示如果该字段为 URL 是否根据当前页面地址将相对URL 拓展为绝对 URL。

附加字段：是否将此字段作为extra字段中的一个健值存入。在存储配置中会详细讲解。

允许无效表达式：指是否允许该表达式抽取不到字段。

下载资源：此字段抽取到的数据是否作为某个资源的URL并且是否需要被下载。

定位方式

当前页面：定位当前页面整体。

XPath：用于通过 XPath 表达式定位 HTML 页面中的元素。

多个XPath：某些需要选取位置不在同一区域，但需要合并为单条数据的页面的特殊处理。

抽取 json 字段：用于通过 JsonPath 定位 JSON 相应中的元素。

CSS：用于通过 CSS 表达式定位 HTML 页面中的元素。

多个CSS：某些需要选取位置不在同一区域，但需要合并为单条数据的页面的特殊处理。

URL 正则：用于通过正则表达式从当前页面URL中定位元素。

内容正则： 用于通过正则表达式定位页面中的元素。

URL 参数：用于获取当前页面URL中的参数。

填充默认值：直接将填充的值保存到数据库中。

列表页值：用于从父级页面附带的字段中抽取，表达式即为父级页面字段名称。

抽取区域内图片：抽取给定 XPath 区域内的所有 img 标签的链接。

脚本：用于使用 Python 脚本抽取对应的值。参见各种脚本默认代码

页面预处理器、处理器

页面预处理器：用于在抽取行之前对页面进行预处理

处理器：用于对抽取到的字段进行后续处理。

支持的处理方式有：

正则（表达式，分组）：使用正则表达式从目标中获取值

替换（源字符串，目标字符串）：略

去除前后空格（无参数）：略

格式化（目标字符串）：Python的format语法。例如：'标题是{}'

字典转换为数组：示例：{"a":1, "b":2} → [("a", 1), ("b", 2)]

URL反转义：解码 URL 中的 % 编码，无需参数

解析JSONP：将JSONP解析为JSON，无需参数

抽取JSON字段：使用Jsonpath语法抽取值

要去掉的多余节点：使用Xpath语法定义需要从DOM树中删除的节点

删除XML声明：删除HTML文档开头出现的XML版本声明。

取MD5：计算MD5值。

脚本：使用脚本处理。参见各种脚本默认代码

注意事项：

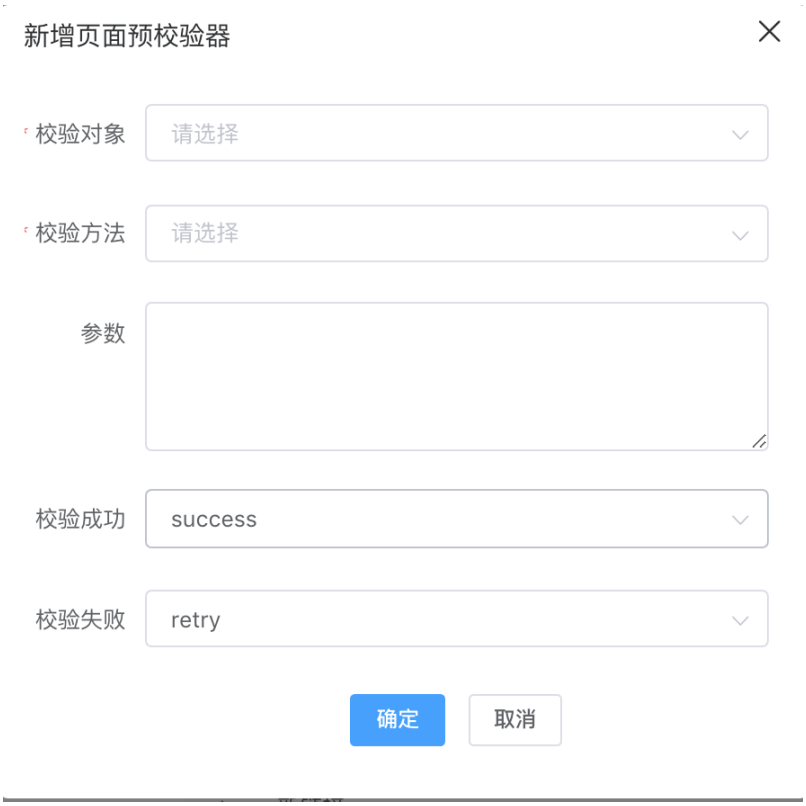
如果无需填充参数，那么直接留空即可。

如果需要填充两个参数，空格隔开即可。

如果参数中有空格，使用英文双引号把参数括起来

页面预校验器、验证

页面预校验器：对下载到的页面进行校验。根据校验结果选择处理还是抛弃当前页面。



校验对象：可选url、text、status_code、original_url、headers

校验方法：

长度：对指定对象做长度校验。参数为范围："1 100"、"-inf 1"、"100 +inf"

正则：规则为re.search，即存在即可

脚本（参见各种脚本默认代码）

参数：校验方法对应的参数

校验成功：即符合校验规则如何处理

校验失败：即不符合校验规则如何处理

#说明：符合校验规则不一定要保留页面，因为我校验的可能是不是一个错误页面。

#处理方式选择：success（继续运行）、retry（重新下载当前页面）、ignore（抛弃当前页面，但将任务状态改为已完成）

验证：对字段抽取到的结果记性验证。根据校验结果决定保留还是抛弃本条数据

支持以下几类校验：

方法：同上

参数：同上

校验失败：同上

区域-抽取结果校验

校验不通过则此页面重试

区域-测试用例

在配置了下载和抽取规则之后，可以填写一个测试链接，然后点击测试按钮，就可以测试下配置的规则是否正确。

其中是否存储选项表示此次测试的数据是否需要存储到数据库。

存储规则

存储规则用于指定一个数据库链接。

按照指定的规则填写即可。

各种脚本默认代码

调度-种子链接-脚本

```
def start_requests(**kwargs):
    """
        yield
    Args:
        **kwargs:
            run_obj:
            logger:
            table_importer:   for field in table_importer(table_name): pass

    Returns:
        Optional[Iterator]
    """
    for page in range(1, 100):
        url = "http://www.boyar.cn/column/14.html?categoryid=1&page={}".format(page)
        yield {"url": url}
```

预处理器-字段处理器-脚本

```
def process(text: str):
    """


    Args:
        text: string

    Returns:
        str
    """
    return text
```

预校验器-字段校验器-脚本

```
def validate(context):
    """
     Ture or False
    :param context:

             {"url": url, "text": text, "status_code": 200, "original_url": "url", "headers": "headers"}

            context
    :return: bool
    """
    return True
```

新增行-新增字段-新链接-脚本

```
def extract(context):
    """

    Args:
        context:  {"url": url, "row": "", "body": ""}

    Returns:
        :[str, str, ...]

            new_link = {"url": "http://www.baidu.com", "method": "POST", "data": "q=1"}
            return [new_link]

    """
    return []
```

下载-脚本

```
# coding:utf8
"""

   download
"""
def download(req) -> dict:
    """

    Args:
        req:   DownloadRequest
                class DownloadRequest:
                    def __init__(self):
                        self.url: str = ""
                        self.url_id: str = ""
                        self.method: str = ""
                        self.data: str = ""
                        self.post_type: str = "form"
                        self.headers: dict = {}
                        self.cookies: dict = {}
                        self.download_url: str = ""
                        self.timeout: int = 0
                        self.proxy: str = "" # ip:port
                        self.allow_redirects = True

                        self.js_expr: str = ""
                        self.load_wait: int = 0
                        self.host: str = ""
                        self.host_id: int = 0
                        self.js_wait: int = 0

    Returns:
        {
            "redirect_url": "url str",
            "content": " bytes",
            "text": " str"",
            "status_code": " int",
            "headers": "headers  dict",
            "cookies": "cookies  dict",
            "status": "",
        }

              redirect_urlcontentstatus_code

    """
    import requests

    r = requests.get(req.url)
    return dict(redirect_url=r.url, content=r.content)
```

必备技能

参见 爬虫基础培训大纲

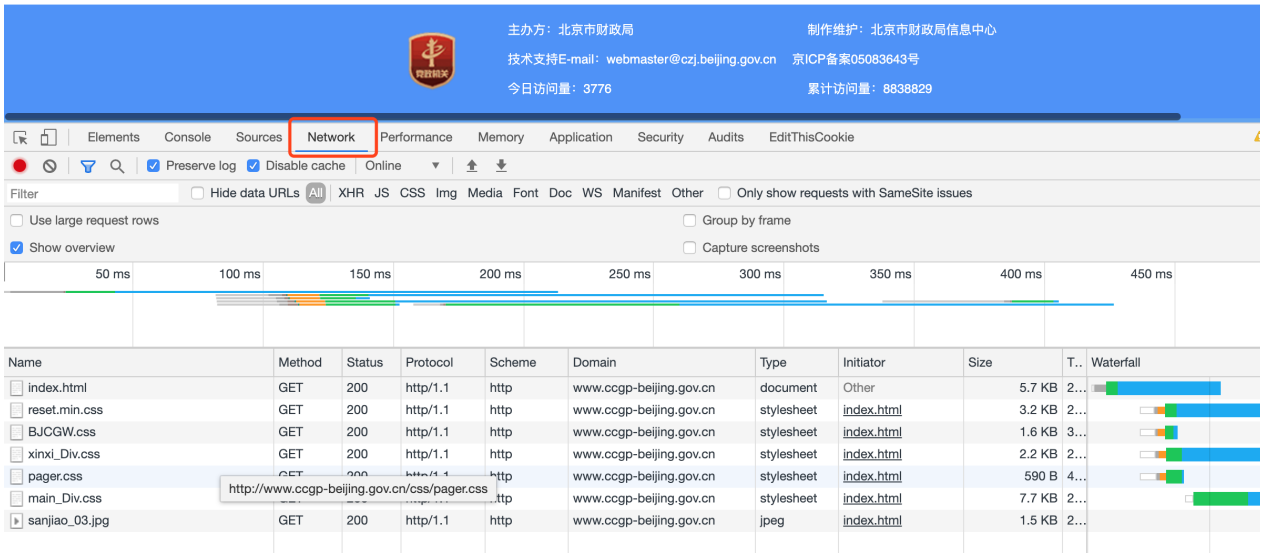
网络流量抓包：

需要用户掌握浏览器的审查元素功能，可通过F12开启控制台。

初级需要能看懂Network面板属性。

如图：

1 2 3 4 5 下一页 下5页



字段抽取工具

需掌握XPath、正则表达式、JsonPath基本使用方式。

Xpath：https://www.w3school.com.cn/xpath/index.asp

正则表达式：https://www.cnblogs.com/dyfblog/p/5880728.html

JsonPath: 略

FAQ

1、IP 池的支持方式及维护方式？

系统支持接入第三方的ip代理服务，由系统的管理员来维护ip池。目前，熵简内部的ip池支持以下几类ip代理：

讯代理（国内代理）

服务商：http://www.xdaili.cn/usercenter/order

套餐类型：混拨代理

蘑菇代理（国内代理）

服务商：http://www.moguproxy.com/usercenter

白名单：基础5个

太阳代理（国内优质代理）

服务商：http://http.taiyangruanjian.com/ucenter/

简介：每天提取上限 15000个，单ip有效期5-1440分钟

自建海外代理

服务商：阿里云EIP+ECS

目前一共20组（EIP+ECS）：香港30、美国50、东京50

luminati-海外代理服务

服务商：https://luminati-china.biz/cp/dashboard

套餐类型：数据中心 50个非中国地区共享IP 不限带宽

2、系统的并发抓取量？

系统中的各个组件如爬虫调度服务，下载解析服务等组件均以单独的 Docker 容器形式部署于容器编排引擎之上。基于容器编排引擎对于Docker 镜像弹性伸缩的机制，理论上可以实现无限伸缩。

一般而言，如果机器配置按照1核2G内存，计算单台机器并发请求量一般为15/秒（普通网站带宽、爬虫集群带宽、CPU处理能力等综合考虑）

5 机器的集群可以支持每天小时的抓取量为 15 * 3600 * 5 = 27万
10 台机器的集群可以支持每天小时的抓取量为 54万
20 台机器的集群可以支持每天小时的抓取量为 108万

案例1：北京政府采购网

配置前准备：

需要确定采集目标和采集范围。

确定采集目标：

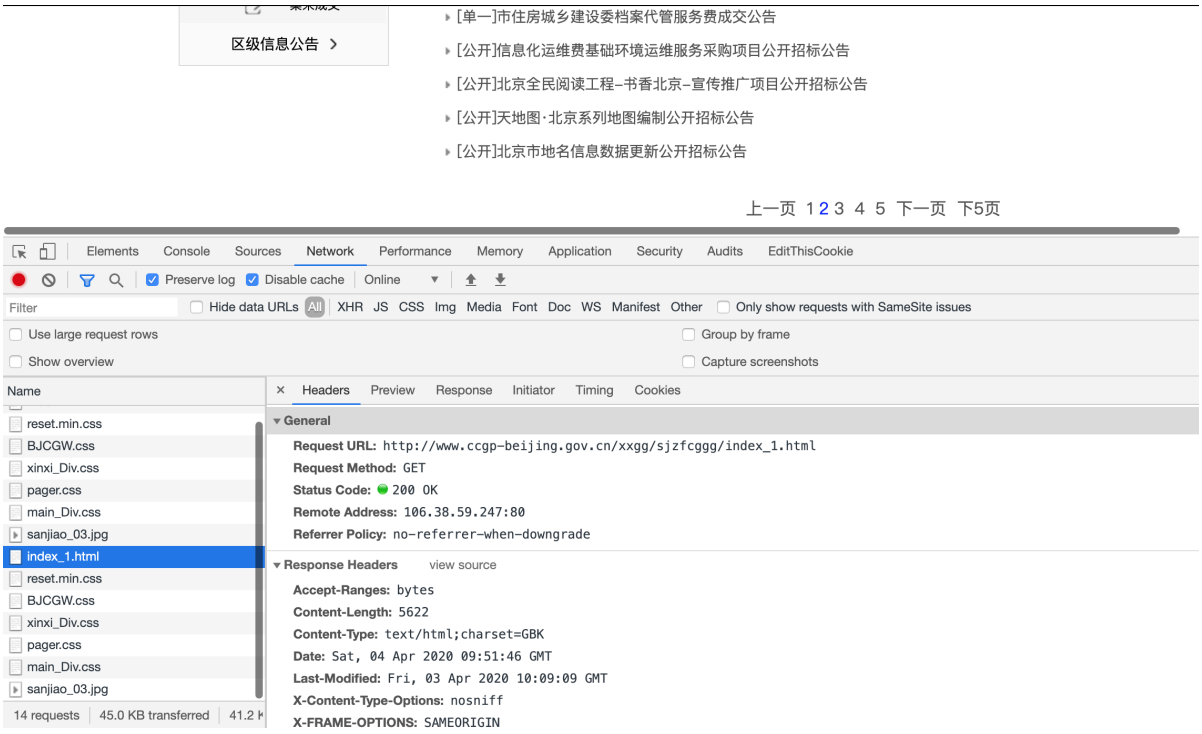
采集北京市政府采购网中，市级信息公告一栏下的所有中标公告。

网址：http://www.ccgp-beijing.gov.cn/xxgg/index.html?city=shi&name=shiji

如下图所示：

通过点击翻页发现当前页面URL无变化，可确定是通过AJAX请求获取下一页内容。

通过F12开发者工具查看后



可获取每一页的真实URL为：http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/index_1.html

其中，index_ 后面的数字即为翻页参数，通过改变此参数，即可实现翻页。

确定采集范围：

通过不断变化页码，可测试出最大页码为143。

配置步骤：

项目、组创建

进入爬虫管理系统，点击创建项目。输入项目名称“招投标项目”后，即可创建一个新的项目组

选中 招投标项目 后，点击右侧 创建组，并输入名称“北京市政府采购网”，即可创建一个爬虫项目。

规则创建

选中 北京市政府采购网 后，点击右侧 创建规则，分别创建如下4个规则：

1. 规则名称：北京市政府采购网调度计划，规则类型：调度计划

2.规则名称：北京市政府采购网目录页面规则，规则类型：页面规则

3.规则名称：北京市政府采购网详情页面规则，规则类型：页面规则

4.规则名称：北京市政府采购网存储配置，规则类型：存储配置

规则配置

调度计划配置

本项目中，我们选择每天运行一次，因此：定时方式为：时间间隔；定时单位：日；定时数值1

根据项目需求和实际网站情况，配置 抓取速率、任务完成阈值、预估任务大小，一般情况下，采用默认配置即可。

种子链接类型选择：种子链接列表

页面配置选择：北京市政府采购网目录页面规则

Url：http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/index_{1..144..1}.html

配置完之后的截图如下：

回到页面上方，点击保存。

列表页面规则配置

此页面配置的目标是从 http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/index_1.html 等页面抽取出每一条标书的链接，并作为后续爬虫的任务。

页面下载配置：代理配置选择默认代理，其他配置默认即可

页面去重配置：默认即可

页面抽取配置：

1.点击新增行，在弹出框中依次填入所需信息：

名称：单条标书

描述：

内容匹配正则、URL匹配正则：空缺即可；

存储配置：不使用存储。这个页面是为了抽取链接，并充当后续爬虫的任务，因此不需要存入用来存放数据的数据库。

定位方式：Xpath

表达式：.//*[@class="xinxi_ul"]//a

配置完成如图：

2.由于本页面的目标是抽取链接，因此点击行内的新增链接，填入如下配置：

页面配置：北京市政府采购网详情页面规则。选择本页面抽取出来的连接应该用什么页面规则来处理。

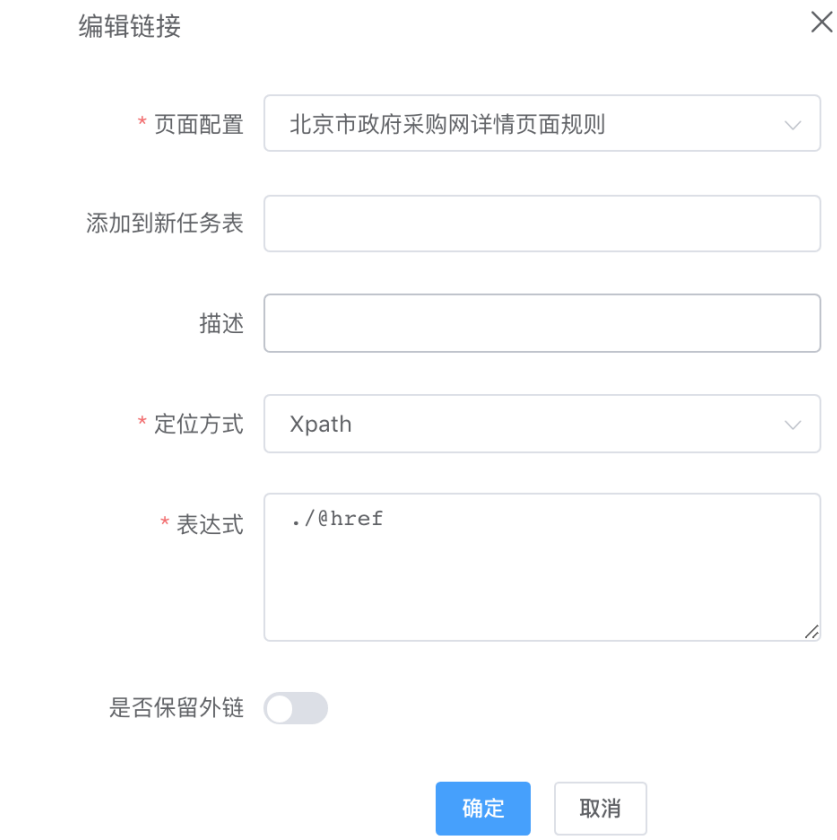
因为本页面抽取出来的是单条标书的链接，因此应该选择 北京市政府采购网详情页面规则 来处理

添加到新任务表、描述：留空即可

定位方式：Xpath

表达式：./@href

配置完成如下：


编辑链接
* 页面配置
北京市政府采购网详情页面规则
添加到新任务表
描述
* 定位方式
Xpath
* 表达式
./@href
是否保留外链
确定
取消

整体截图如下：


页面抽取配置
页面预处理器
新增页面预处理器
页面预校验器
新增页面预校验器
抽取行
新增行
导入行
批量新增行
名称
描述
内容匹配正则
URL匹配正则
存储配置
定位方式
单条标书
不使用存储
Xpath
字段
新增字段
导入字段
链接
新增链接
页面配置
添加到新任务表
描述
定位方式
表达式
北京市政府采购网详情页面规则
Xpath
./@href
处理器
新增处理器

抽取结果校验：

此页面仅获取了新链接，因此进行如下配置：

最小数据数：0

最小新链接数：1

测试用例：

通过此功能可以很方便的测试 页面下载配置 和 页面抽取 配置是否正确；

测试用例配置如下：

方法：选择 GET

是否存储：否

Url：http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/index_1.html

然后点击测试，如果有如下结果，则表明 页面下载配置 和 页面抽取 配置都没问题。

结果列表

行：单条标书_0（共14条）

link:19
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1209041.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208971.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208965.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208960.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208961.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208959.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208952.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208950.html
http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20200403_1208899.html

请求信息

回到页面上方，点击保存。

详情页面规则配置

此页面配置的目标是从标书详情页面抓取标题、发布日期、标书正文等信息。

页面下载配置：同列表页面配置

页面抽取配置：

1.点击新增行，在弹出框中依次填入所需信息：

名称：单条标书详情

描述、内容匹配正则、URL匹配正则 空缺即可；

存储配置：北京市政府采购网存储配置。抽取的数据通过此配置，存入数据库中。

定位方式：当前页面

2.新增字段

由于本页面的目标是抽取所需要的标书信息，因此点击 新增字段，每一个字段对应一个配置信息，

由于需要抽取 三个字段 + 一个页面 url，因此需要新增四个字段，各字段配置如下：

标书标题：

名称：title（此为数据库中的字段名）

描述：标题（此为数据库中的字段说明）

字段类型：varchar

定位方式：Xpath

表达式：.//*[@style="font-size: 20px;font-weight: bold"]

其他选项，采用默认配置即可

标书日期：

名称：bid_date

描述：发布日期

字段类型：varchar

定位方式：Xpath

表达式：.//*[@class="datetime"]

其他选项，采用默认配置即可

标书正文：

名称：bid_content

描述：标书正文

字段类型：text

定位方式：Xpath

表示：.//*[@style="width: 1105px;margin:0 auto"]//div[position()>2]

保存HTML：打开

其他选项，采用默认配置即可

标书 url：

名称：url

描述：标书链接

字段类型：varchar

定位方式：URL正则

表达式.*

其他选项，采用默认配置即可

抽取结果校验：

此页面目标为抽取标书详情数据。因此进行如下配置

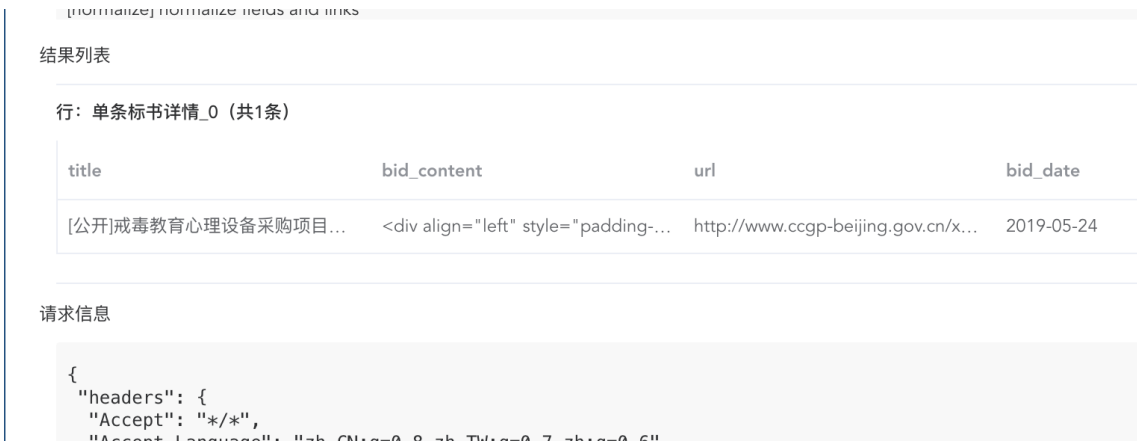
最小数据数：1

最小新链接数：0

测试用例：

Url：http://www.ccgp-beijing.gov.cn/xxgg/sjzfcggg/sjzbjggg/t20190524_1118159.html

点击测试，若结果如下图所示，则说明上述配置无误：



回到页面上方，点击保存。

存储规则配置

根据项目需求，可选择将采集到的数据存入 mysql 或者 Oracle 中，本示例以 mysql 为例：

1.在 MySQL 配置下，点击 是否启用；

2.实例名称：选择自定义，依次填入 mysql 相应的配置信息，如 Host，port，用户名等；

3.回到页面上方，点击保存。

启动爬虫项目

点击进入 北京市政府采购网调度计划；

在数据展示区，点击 启用计划 即可。

系统将根据用户所配置的运行时间和频率，调用相应的组件，根据前面的页面配置，进行网页下载、数据抽取和数据入库等操作。