火车采集器软件

用户使用手册

合肥乐维信息技术有限公司

前 言

火车采集器是一款专业的网页数据抓取、处理、分析、挖掘软件，可以灵活迅速地抓取网页中散乱分布的文本，图片等资源信息，然后通过一系列的分析处理，准确挖掘出所需数据。并可以选择发布到网站后台、导入数据库或者保存在本地 Excel，Word 等格式的文件中。

火车采集器历经十年的升级更新，凭借高效稳定的特性积累了大量用户和良好口碑，被誉为最经典的采集软件。

研发背景

- **从搜索引擎到网页数据采集**

以惊人的速度发展起来的网络时代，成就了万维网这个拥有着大量信息资源的宝藏，基于万维网信息资源而生的搜索引擎则实现了信息的有效查找和利用；但网络的与时俱进让我们对互联网信息产生了新的需求，不仅要搜索到信息，还要将所需数据信息快速地收集到目标中去，这个目标可能是一家网站，一个数据库，一间网店，一篇文档……所有需要数据的地方，正是这种对数据利用的强烈需求催生了网页数据采集技术。

- **从手动采集到软件采集**

对网页数据的采集需求最初通过人工手动采集来实现，我们把需要的数据复制下来，再粘贴到目标中去，就完成了一个最简单的采集过程。手动采集可以满足少量的采集需求，但网页数据是海量的，我们需要的数据往往也是大量而又复杂的，传统手动采集会耗费许多时间和精力，因此我们需要一个高效的采集工具，来帮助我们快速完成采集，在这种需求之下，火车采集器应运而生。

火车采集器实现了将数据从采集到处理到发布的一系列智能操作，能够快速稳定地应对大量的数据采集需求，取代手动采集模拟人工操作，大幅提升工作效率。

核心功能

- **分布式高速采集**

任务分配至多个客户端，同时运行采集，效率倍增。

- **多识别系统**

配备正文识别、中文分词识别、任意编码识别等多种识别系统，智能识别操作更轻松。

- **全自动运行**

无需人工值守操作，可计划任务运行，任务完成后自动关机。

- **无限级多页采集**

支持包含 ajax 请求数据在内的多个页面信息的无限级采集。

- **任意文件格式下载**

图片、压缩文件、视频等任意格式的文件都能轻松下载。

- **支持多数据库**
  支持 Access/MySQL/MsSQL/Sqlite/Oracle 多种类型的数据库保存及发布。
- **支持扩展**
  支持接口和插件扩展，满足各种采发需求。
- **采集监控系统**
  实时监控采集，确保数据的准确性。

术语解释

1.采集任务：采集任务是火车采集器中对于数据采集和数据发布任务的完整配置，包含采集规则和发布模块。

2.采集规则：即我们对如何采集和采集什么的问题给出一些设置让采集器按照设置的规则来执行，这个设置可以从火车采集器里面导出保存为.ljobx 文件，也可以再次导入火车采集器。

3.发布模块：在火车采集器中，发布模块是对“将数据发布到哪里”进行的设置。包括 WEB 在线发布模块和数据库发布模块，其设置分别可以导出保存为.wpm 文件和.dbm 文件，并可以再次导入火车采集器使用。

4.发布接口：发布接口是一个小型的页面程序，通常和 WEB 在线发布模块配合使用来满足用户的特定需求。即采集器将采集的数据发送到发布接口文件中，接口文件得到数据，并按照用户特定需求灵活地处理数据。

5.标签：是指用来提取某项内容信息的一个字段名字，由用户在编辑规则的时候指定，比如标题、手机号、作者，内容标签采集到的信息在发布模块中就可以通过该标签名对应获取到，格式为[标签：标签名]如[标签：标题]。标签在火车采集器里面有分为两种：分别为列表页标签和内容页标签，顾名思义列表页标签就是在获取列表页时（即采网址时）就获取到内容信息，内容页标签是在获取内容页或多页内容时（采内容）才获取内容信息。

注：这里所说的标签是软件采集标签，不同于网页 html 标签，类似于数据库字段。

6.（*）：火车采集器中变量的通用符号，如果我们只需要知道这个变量的变化规律，而不需要关心这个变量到底是什么，这时就可使用这个符号代替。

7.[参数]：用来匹配某项准备提取信息的标记标签，如想要在代码中提取组合出某种格式。以从代码"mClk(this,'108484','134217', '168475','1');"中提取组合出新的地址格式为例。

"mClk(this,'[参数]','[参数]', '[参数]','1'); ",按照次序，108484 参数就是参数 1，

依次类推。实际需要的地址为以下的地址格式：bbs/read.php?id=[参数 1]&sort=[参数 3]&action=[参数 2]，上面代码中的 3 个参数和下面地址中的 id，soft 和 action 参数要对应相应的值，次序不要颠倒。这样就组合成了新的地址格式。

8.起始网址：用来获取下级链接地址的入口网址，可以为一条或多条，可以通过添加起始网址向导添加同格式多条网址或导入文本网址。如果没有定义多级网址的获取方法，这些地址即作为内容页网址进行内容采集。

9.多级网址：依次根据列表里面的多级网址顺序采集分析地址，通过依次采集分析到最后一级得到内容页地址。多级网址的获取可以使用页面自动分析和手动获取的方法采集下级网址，在采集的过程中，可以同时采集列表分页及提取列表页附加参数。

10.Cookie：是在 http 请求访问中记录用户信息即登录信息的一段用于与服务器进行交互的字符串。在浏览器中使用时通常还会以文本形式记录到您的 IE 缓存目录中以便下次在有效期内不用输入用户信息即可继续访问验证权限的网页。

11.User-Agent：浏览器标识，是用来向服务器通知用户使用的客户端类型，在某些需要登录的网页可能需要同时验证 Cookie 和 User-Agent，所以需要将其设置为与本机浏览器同样的格式。

12.分页：列表或内容页面较长，分成多个页面显示，采集时需要将所有子页的内容组合起来，这样的子页面就是分页（列表分页或内容分页）。

13.多页：有些情况下，需要采集一个页面对应的网址，图片等内容时，需要另外打开一个新的页面才能采集到这些信息，这些另外打开的页面则称为多页。

14.网页编码：是网页中指定特定字符编码格式的库，例如一般网页中都有如下一句：<meta http-equiv="Content-Type" content="text/html; charset=gb2312">，这样的字句指示此网页的字符集编码是 gb2312。火车采集器对一般的网页可以做到自动识别，也罗列出了大部分的网页编码格式，可以直接在采集器中手动选择指定相应的编码格式。

15.代理：是指网络中的代理服务器，可以代理网络用户去取得所需要的网络信息。代理的功能有可以突破自身 ip 的访问限制访问国外站点，访问一些单位或团体内部资源，突破电信的 ip 封锁和隐藏真实的 ip 等。

16.插件：在火车采集器中，插件是指可以对采集到的数据进行特定处理的一个外部程序，编写好插件后，采集器可以把数据传递给插件，然后对数据进行处理，再把数据传给采集器。

17.任务网址库：采集器在文件夹 Data 下，该站点下的每一个任务都会生成一个独立或公用的网址库用来对比网址重复之用。

18.Http 请求：浏览器打开网页时实际就是发送一个又一个 Http 请求，火车采集器也一样，从指定的地址获取内容的过程就是发送 Http 请求，然后对根据请求得到的内容进行处理。
当浏览器向 web 服务器发送请求时，它向服务器传递了一个数据块，也就是请求信息。Http 请求信息由 3 部分组成 :请求方法 URI 协议/版本，请求头(Request Header ）和请求正文。

编程语言

火车采集器的编程语言是 C#，C#综合了 VB 简单的可视化操作和 C++的高运行效率，增强开发效率的同时也致力于消除编程中可能导致严重结果的错误，以其强大的操作能力、优雅的语法风格、创新的语言特性和便捷的面向组件编程的支持成为软件开发的首选语言。

软件资质

火车采集器软件是合肥乐维信息技术有限公司自主研发而成。火车采集器的源代码、布局、界面设计、电子文档等均已提交国家版权局登记备案，并已获得著作权审批。

火车采集器 V9 版软件著作权登记号：2015SR17313


中华人民共和国国家版权局
计算机软件著作权登记证书
证书号：软著登字第1060259号
软件名称：火车采集器软件
[简称：火车采集器]
V9
著作权人：合肥乐维信息技术有限公司
开发完成日期：2015年07月01日
首次发表日期：2015年07月01日
权利取得方式：原始取得
权利范围：全部权利
登记号：2015SR173173
根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。
中华人民共和国国家版权局
计算机软件著作权登记专用章
No. 00805414
2015年09月08日

目 录

第一章 登录与注册电脑

登录火车采集器网站（www.locoy.com）下载安装火车采集器软件后，进行登录和注册电脑即可运行软件。

一、软件登录

运行采集器主程序 LocoySpider.exe，软件将打开如图 1.1 登录界面，在火车采集器官网注册一个账号，在登录界面输入账号和密码后即可进行登录。

（图 1.1）

二、注册电脑

火车采集器支持多种登录验证方式，如果使用的是机器码版本或非加密狗版本，最新版本的火车采集器在第一次登录软件时会自动注册电脑（非最新版本会弹窗提示注册当前电脑，直接点击注册当前电脑即可）。如果为加密狗版本，需插入加密狗后才可运行软件。

自动授权版本可随时更换电脑，自动登录。更换方式：下载新版 V9 到新电脑直接登录，会提示客户端管理器，用户选中后禁用，即可禁用之前的采集器，并在新机器上运行了。

第二章 开始菜单

包括新建分组和任务，对采集过程中可能用的发布、插件、代理、任务计划等进行配置。

（图 2）

一、新建分组

新建一个任务分组，选择所属分组，确定分组名称和备注。

（图 2.1）

二、新建任务

新建一个任务，随后对任务进行规则配置，规则配置的部分就是火车采集器的核心操作，需要重点学习，第四章将详细讲解操作。

三、Web 发布配置

我们通过火车采集器采集数据之后，还可以使用采集器发布数据，Web 发布配置就是用来定义如何登录一个网站以及向该网站提交数据的。该功能主要涉及到登录信息的获取，网站编码设定，栏目列表的获取，以及使用数据测试发布效果。

1、操作指导

1.1 选择发布模块：新建或选取已设置好的模块，可对模块进行编辑、新建、删除、导入，导出等操作。（实施发布模式的编辑和新建时即转入第四节内容“web 发布模块”，详见本章第四节，这里不再次说明。）

（图 2.3.1.1）

1.2 网页编码 选择对应的与要发布网站一致的网页编码。如 GBK、GB2312、UTF-8，BIG5 等。

（图 2.3.1.2）

1.3 全局变量：可以在发布模块中所有位置使用，方便设置和修改某些参数。

1.4 登录操作

（1）网站地址：一般指网站域名，实际操作中根据发布模块里的地址做实际调整，需和模块里的地址组合成一个完整的绝对地址。

（2）登录方式：分 3 种，内置浏览器登录，数据包登录，不登录。

①内置浏览器登录：获取浏览器标识和用户信息，点击图 2.3.1.4-1 所示的“启动浏览器获取登录信息”按钮即可打开网页获取信息。

（图 2.3.1.4-1）

②数据包登录：填写用户名，密码以及获取到的验证码后登录。此种方法需要在所选择的发布模块里“网站自动登录” 有对应设置（见本章第四节）。

（图 2.3.1.4-2）

1.5 获取分类/栏目列表：可刷新出栏目 ID 和栏目名称。需要所选择的发布模块里“获取栏目列表”有对应设置（见第二章第四节）。

（图 2.3.1.5）

1.6 Web 发布配置列表：管理所有的 Web 发布配置。

1.7 测试当前发布：完成设置后可对发布进行测试。

（图 2.3.1.7）

四、Web 发布模块

火车采集器的在线发布数据功能如何实现？这就是 WEB 发布模块的用途：将我们手动在网站后台发布内容的整个过程包括登录网站后台、选择栏目、发布文章等步骤设置到火车采集器里面，从而通过采集器来模拟发布。

火车采集器采集到的值通过标签名传递给在线发布模块，实现数据提交到网站的目的。该功能需要定义网站自动登录，获取栏目列表，获取网页随机值，内容发布参数，以及上传文件，构造发布数据等高级功能。

1、操作指导

1.1 网站自动登录：即配置网站系统登录的步骤和方式。网站登录是以 post 数据的方式提交 form 数据，从而实现自动或半自动的登录。我们首先要打开并登录需要发布的后台，然后使用数据包抓取工具 fiddler 抓取登录这一步骤的相关数据，根据 fiddler 中的数据来填写以下内容。

（1）登录地址后缀：登录地址就是 post 地址，登录地址后缀即为 post 地址中除去域名和后台目录之后的后缀部分。

比如 fiddler 中数据为 **POST http://127.0.0.1:801/dede/dede/login.php**

则登录地址后缀填写如图 2.4.1.1 所示。

（2）来源页面后缀：来源页面即为 Referer,来源页面后缀同样为 Referer 除去域名和后台目录之后的后缀部分。

比如 fiddler 中数据为

Referer:http://127.0.0.1:801/dede/dede/login.php?gotopage=%2Fdede%2Fdede%2Findex.php

则来源页面后缀填写如图 2.4.1.1 所示。

（图 2.4.1.1）

（3）验证码地址：可在验证码上右击，复制地址查看填写。

（4）登录 post 数据：可对表单名和表单值进行添加、修改、删除等操作，其中表单的相关数据是可以通过自动抓取登录数据包、粘贴抓包获取的数据、提取 post 表单登录数据三种方式获取的。

使用粘贴转获取的数据提取结果如图 2.4.1.1。

（5）失败、成功标志：分别看看在网站发布正常的内容和发布失败的内容给出的提示（比如成功发布文章、标题不得为空、请选择栏目之类），然后写入模块设置，多个提示一行一个。如图 2.4.1.1。

注意事项：这里可以使用全局变量，该变量在正式发布时将被替换。该值在 Web 发布配置中设置。这里的参数名，是可以使用网页随机值的。

1.2 内容发布参数：该配置是设置网站发布内容的步骤和方式，也可以使用粘贴抓包获取的数据来自动填写。

在发布页面填写好需要发布的字段值，比如来源、标题、内容（先不要点击发布）；然后打开 fiddler(注意，如果有较乱的数据流，请先 Ctrl+X 清空数据流)；打开后点击发布，分析 fiddler 里抓取的发布这一步骤的数据。

发表地址后缀：这时数据中的 post 地址就是发表地址，发表地址后缀即为 post 地址中除去域名和后台目录之后的后缀部分，填写如图 2.4.1.2。

（图 2.4.1.2）

来源页面后缀：来源页面即为 Referer,来源页面后缀同样为 Referer 除去域名和后台目录之后的后缀部分。

发布 post 数据：可对表单名和表单值进行添加、修改、删除等操作，其中表单的相关数据是可以通过自动抓取登录数据包、粘贴抓包获取的数据、提取 post 表单登录数据三种方式获取的。使用粘贴转获取的数据提取结果如图 2.4.1.2。

1.3 获取栏目列表：获取网站的栏目，即获取分类 id 和分类名称的方式。获取到的栏目信息，可以在 Web 发布配置中调用，方便用户选项，该项也可以不用填写，在 Web 发布配置中手动指定。

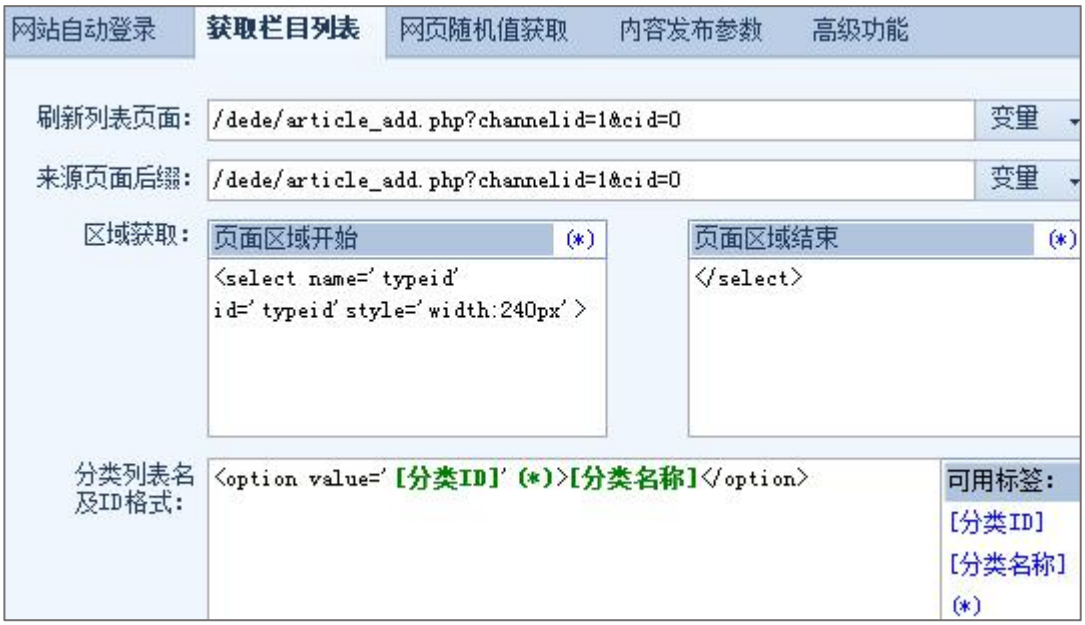
首先要确定我们的选择栏目列表是在哪个页面，最常见的一种，栏目选择就是在发布内容页面里，类似 DEDE 文章发布；特殊情况在其它页面下，不在发布内容页面。这里以常见情况为例讲解：

（1）刷新列表页面和来源页面后缀：把上述“内容发布参数” 中的来源页面后缀的设置拿过来直接使用即可。如图 2.4.1.3-2。

（2）区域获取：查看发布页面的源代码，找到刷新列表部分的源码如图 2.4.1.3-1，然后确定栏目列表的开始和结束代码，以及格式。

（图 2.4.1.3-1）

（3）分类列表名及 ID 格式：ID 用[分类 ID]替换；栏目名称用 [分类名称]替换；不规则出现的代码用 (*)通配符匹配。如图 2.4.1.3-2。



（图 2.4.1.3-2）

1.4 网页随机值获取：在发布文章或者登录的时候，会有些值时刻变化，这些值是由网站系统自动生成的，在此配置中获取这些特殊的值，可以在其它的地方，如网站登录，POST 提交内容，获取栏目分类信息等处使用。

比如系统时间，例如 pubdate 这个表单就是设置发布时间的。这个时间是可以在填写文章内容的页面上得到的。由于时间是会变动的，所以这种情况下不要勾选”每次请求都使用第一次获取的随机值 “，根据源代码可以填写如图 2.4.1.4-1。

（图 2.4.1.4-1）

（图 2.4.1.4-2）

需要注意的是这里的强制编码：图 2.4.1.4-2 中所示的强制编码使用后，整个发布过程都将使用该编码，这方便某个系统是单一编码的处理，如 wordpress 只有 utf8 编码。

完成上述操作之后还需要把内容发布参数“发布 POST 数据”里的值替换成标签。双击选中表单值，然后鼠标悬停在标签按钮上，对应选择要替换成的标签名即可，可选系统标签，常用标签，时间标签，替换为标签是为了后面的一系列操作做铺垫。

1.5 高级功能：文件上传及发布数据构造的设置，可以在文件上传时可以使用递增数字；发布数据构造可以对标签数据进行二次处理，比如对自动提取到的表单字段进行标签名修改等。

（图 2.4.1.5）

五、数据库发布配置

除了在线发布数据，火车采集器还可以将采集的数据发布到数据库，火车采集器可选 mysql、sqlserver、oracle、access 四种数据库类型。数据库发布配置定义了数据库链接信息的设置以及数据库模块的选择。

1、操作指导

1.1 选择发布模块：可对数据库模块进行新建、编辑、删除、导入，导出等操作。（实施数据库发布模式的编辑和新建时即转入本章第六节内容“数据库发布模块”，详见本章第六节，这里不再次说明。）

（图 2.5.1.1）

1.2 数据库链接信息：在选择模块时，相应的数据库链接信息配置会自动根据模块数据库类型显示在配置框，如图 2.5.1.2。除了 access，其他的数据库都需要配置登录信息，然后选择数据库，测试数据库连接成功与否。

（图 2.5.1.2）

1.3 数据表前缀：有些数据表在创建的时候自带前缀用以区分其他表，比如一些 cms 在安装创建表的时候会自动带一个表的前缀。

1.4 模块说明：在模块使用说明文本框中进行相关使用说明的填写。

1.5 测试发布：以上步骤完成后，可以进行测试入库操作。标签列表会根据选择的模块，自动罗列对应标签，然后手动定义相应标签值。标签值定义完成后，点击测试入库按钮，采集器会显示对应的 sql 语句，然后显示对应 sql 语句的执行结果。若测试成功，在配置名文本框填写配置名称，然后点击保存配置按钮。

六、数据库发布模块

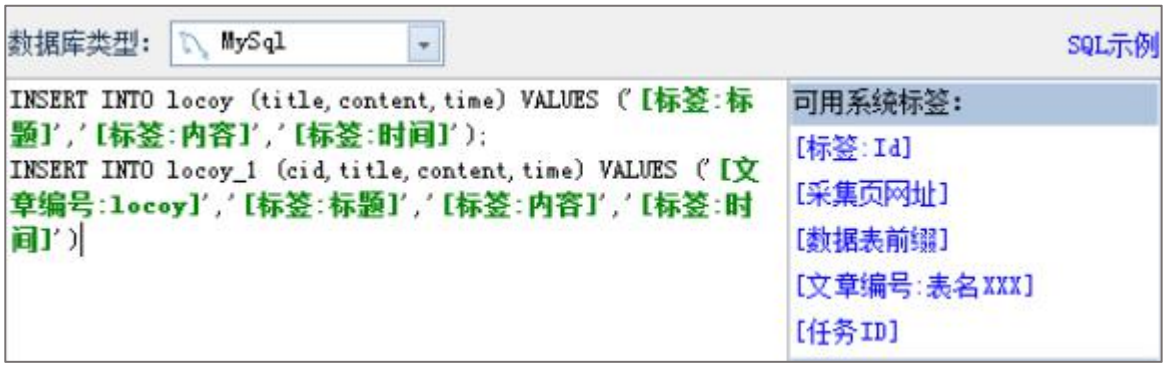
用于编辑数据库的发布模块，方便我们将数据发布到配置好的数据库中。火车采集器可选如图 2.6 所示 mysql、sqlserver、oracle、access 四种数据库类型，在文本输入框中填写 sql 语句，并可使用标签替换相应数据。也可在采集器模块文件夹中加载某一模块进行编辑。

（图 2.6）

1.操作指导

1.1 若是单表或多表无关联：则直接写 INSERT 语句即可。在 sql 语句文本框填写 sql 语句，相关数据用相应的标签进行替换。

注意：“数据表前缀”标签不能在其他标签使用，如在“文章编号：表名 XXX”标签中“表名 XXX”不能使用“数据表前缀”标签。“文章编号：表名 XXX”必须在插入某一数据的 sql 语句后使用，此 sql 语句插入数据成功后若返回一个自增 ID，文章编号就是此自增 ID，此 sql 语句后的 sql 语句若要用到此自增 ID，可用“文章编号：表名 XXX”标签进行替换，如图 2.6.1.1。



（图 2.6.1.1）

1.2 若是多表：且存在某字段相互关联，则用[文章编号:表名 XXX]来关联上一个表的自增 ID；自增 ID 字段和值需要删除，不需要写入 SQL 语句内。

1.3 备注说明：对入库模块进行相关使用说明，在“使用说明”文本框里填写相关信息即可。

1.4 加载：点击“加载”按钮可对从采集器模块文件夹加载选中的某一入库模块进行编辑。

七、计划任务

用于设置列表中采集任务的启动计划，可如图 2.7 所示每间隔、每天、每周、仅一次、或自定义 Cron 表达式。

（图 2.7）

Cron 表达式是一个由 6 或 7 个子表达式组成的字符串。每一个表达式代表一个域，每个域描述了一个单独的日程细节且每个域之间使用空格分隔，它由两种格式组成：Seconds Minutes Hours DayofMonth Month DayofWeek Year 或 Seconds Minutes Hours DayofMonth Month DayofWeek。具体可以查找相关资料学习了解，保存设置后，任务即可按照设置执行。

八、插件管理

火车采集器 V9 支持 PHP 源码、C#源码、C#类库、python 源码四种类型的插件，可用于扩展 http 请求、内容处理和文件下载的功能，并可以分别进行测试。

（图 2.8）

1、插件说明

这里以 C#源码类型的插件为例，进行插件说明。

1.1 HTTP 请求插件：可以修改 HTTP 请求前的请求数据（http header）和 HTTP 完成后的返回数据（response），这个插件包含了 2 个处理方法。

（1）BeforeRequest(Request Entry request)：这个方法会在所有 HTTP 请求前调用，包括网址采集、内容采集请求，可以通过修改请求来应对一些复杂的网站抓取。

参数介绍

• request 参数中包含 Url、Referer、Cookie、Headers、页面类型等，除 HTTP 基本属性外，还包含一些特殊值。

①request.Properties["PageType"]：此属性是页面类型，值为整数类型，包含 6 种类型“0：起始地址；1：列表页面； 2：列表页的分页； 3：内容页面；4：关联多页； 5：内容页的分页”。

②request.Properties["JobName"]：任务名称。

③request.Properties["JobID"]：任务 ID。

request.Properties 属性最好只做读取操作，不要修改，不然会造成无法预料的结果。其他的 RequestEntry 字段请参考官网的插件说明文档。

（2）AfterResponse(ResponseEntry response)：这个方法在所有 HTTP 请求完成后调用，可以修改为自己想要的数据，然后交给采集器来处理。

参数介绍

• response 中包含 HTTP 响应数据，如返回 HTML、响应 Header。

①response.RawText`：是返回的 HTML 代码。

②response.Url`：请求的 Url 地址。

和 request 一样，response 也包含 response.Properties["PageType"]、request.Properties["JobName"]、request.Properties["JobID"]，含义相同。

示例插件代码：如图 2.8.1.1。

```
public class Plugin1 : IHTTPTamper
{
    /// <summary>
    /// 处理下载前的request
    /// </summary>
    /// <param name="response"></param>
    public void BeforeRequest(RequestEntry request) {
        Console.WriteLine("BeforeRequest: "+request.Url);
    }
    /// <summary>
    /// 处理下载完成后的http响应,网址、默认页、多页、内容分页
    /// </summary>
    /// <param name="response"></param>
    public void AfterResponse(ResponseEntry response) {
        Console.WriteLine("AfterResponse: " + response.Url);
    }
}
```

（图 2.8.1.1）

1.2 内容采集插件：通过这个插件可以修改最终的标签数据结果。

（1）内容插件处理方法：Process(ResponseEntry response，Dictionary<string, string> result)

参数介绍：

①response：内容页的响应结果。

②result：是标签数据结果，键是标签名称，值是标签数据。

可以修改 result 中的值，但是不要删除、清空 result 的键 ，否则保存数据时可能会出错。

示例插件代码：如图 2.8.1.2。

```
public class Plugin2 : IResultProcesser
{
    /// <summary>
    /// 处理采集后的标签结果
    /// </summary>
    /// <param name="response">默认页面的响应</param>
    /// <param name="result">标签结果，键为标签名称，值为标签值</param>
    public void Process(ResponseEntry response, Dictionary<string, string> result)
    {
        Console.WriteLine("Process: " + response.Url);
        var m = Regex.Match(response.RawText, "<title>(?<t>[^<]*)");
        if (m.Success)
        {
            Console.WriteLine(response.Url);
        }
    }
}
```

（图 2.8.1.2）

1.3 文件下载插件：这个插件提供下载文件的相关信息，并且修改其中的某些值。

（1）处理方法：BeforeDownload(RequestEntry request, DownloadFile downloadFile)

参数介绍：

①request ：文件下载 HTTP 的请求数据。

②downloadFile：文件下载的相关属性，包括下载地址、保存路径、HTML 中的替换路径等。

示例插件代码：如图 2.8.1.3。

```
public class Plugin3 : IFileDownloader
{
    /// <summary>
    /// 文件下载前的处理方法
    /// </summary>
    /// <param name="request">文件下载的HTTP请求</param>
    /// <param name="downloadFile">下载文件</param>
    public void BeforeDownload(RequestEntry request, DownloadFile downloadFile)
    {
        if (downloadFile.DownlaodUrl.Contains("xxx"))
        {
            downloadFile.Cancel = true;//取消该文件的下载
        }
    }
}
```

（图 2.8.1.3）

在任务规则编辑过程中直接选用已设置完成的插件即可，选择之前可根据插件说明选取或编辑插件的源码并在插件管理工具中完成测试。

九、http 二级代理

网络中的代理服务器，可以代理网络用户去取得所需要的网络信息。代理的功能有可以突破自身 ip 的访问限制访问国外站点，访问一些单位或团体内部资源，突破电信的 ip 封锁和隐藏真实的 ip 等。火车采集器 V9 支持 http 代理、sockes4 和 sockes5 代理。

1、操作指导

1.1 二级代理：对二级代理进行添加、编辑、删除、验证、批量导入等操作。

（图 2.9.1.1-1）

（1）编辑：编辑代理的类型、地址、端口、用户名、密码，域等。

（2）批量导入：准备一个有 IP 地址的 TXT 文件，内容格式为 **ip:端口**，一行一个，如图 2.9.1.1-2。点击批量导入——浏览——选中 代理.txt 文件即可实现代理 IP 的导入如图 2.9.1.1-3。

（图 2.9.1.1-2）

（图 2.9.1.1-3）

1.2 页面缓存：使用二级代理采集时，同一个网址，多次的请求中，原页面可能并不存在任何的更新，所以直接调用缓存页面节约代理资源，提高了访问速度。通过设置网址必须包含和内容必须包含，则符合条件的内容会缓存在本地。

1.3 选项设置：二级代理验证设置或自动拨号设置。

（1）端口：默认是 8888。

（2）二级代理验证设置：根据访问地址设置，以 www.locoy.com 为例，查看此网页源代码，找到在正常访问时含有的某个字符串做标识（注意：当不正常访问时，比如封 IP 时，就不含有此字符），在这里可以根据 **<title>火车采集器软件**来判断，如图 2.9.1.3-1。

（图 2.9.1.3-1）

接下来，进行代理列表的批量验证，然后删除失效代理后，留下来的就是有效代理了。

（图 2.9.1.3-2）

（3）自动拨号：对于用宽带拨号上网的用户，可以通过重新拨号来达到更换 ip 的效果。自动拨号有定时拨号和根据 http 响应字符来拨号两种选择方式：

①定时拨号，通过设置间隔时间：每隔多久重新拨号。

②根据 http 响应字符来重新拨号：比如当前 ip 被服务器禁止，服务返回包含“禁止访问”的 html 代码，则可以将特征字符设置为“禁止访问”，程序检测到该字符串则重新拨号。

（图 2.9.1.3-3）

二级代理及自动拨号功能的开启可在编辑任务的“采集规则——其他设置——代理设置——使用指定代理”处点击使用火车采集器二级代理开启，端口与之前的设置的端口保持一致即可。

1.4 代理服务：自动获取二级代理的代理服务功能。

（1）代理服务登录：如图 2.9.1.4 所示填写代理服务登录的用户们及密码后点击登录。

（图 2.9.1.4）

（2）代理获取配置：设置好每批获取，定时更换，定量更换，代理匿名度后，开启启用。注意：每批获取最多支持 60 个；定时更换，到间隔时间自动更换新的一批 ip；定量更换，平均每个 ip 使用多少次。要重新修改以上值，需要关闭启用。

1.5 运行日志：记录二级代理的运行日志。

十、http 模拟请求

可以设置如何发起一个 http 请求，包括设置请求信息，返回头信息等。并具有自动提交的功能。工具主要包含两大部分：一个 MDI 父窗体和请求配置窗体。

1、操作指导

（图 2.10.1）

1.1 请求地址：正确填写请求的链接。

1.2 请求信息：常规设置和更高级设置两部分。

（1）常规设置：

①来源页：正确填写请求页来源页地址。

②发送方式：get 和 post，当选择 post 时，请在发送数据文本框正确填写发布数据。

③客户端：选择或粘贴浏览器类型至此处。

④cookie 值：读取本地登录信息和自定义两种选择。

（2）高级设置：包含如图 2.10.1.2 所示系列设置，当不需要以上高级设置时，点击关闭按钮即可。

（图 2.10.1.2）

①网页压缩 :选择压缩方式，可全选，对应请求头信息的 Accept-Encoding。

②网页编码：自动识别和自定义两种选择，若选中自定义，自定义后面会出现编码选择框，在选择框选择请求的编码。

③Keep-Alive：决定当前请求是否与 internet 资源建立持久性链接。

④自动跳转：决定当前请求是否应跟随重定向响应。

⑤基于 Windows 身份验证类型的表单：正确填写用户名，密码，域即可，无身份认证时不必填写。

⑥更多发送头信息：显示发送的头信息，以列表形式显示更清晰直观的了解到请求的头信息。此处的头信息供用户选填的，若要将某一名称的头信息进行请求，勾选 Header 名对应的复选框即可，Header 名和 Header 值都是可以进行编辑的。

1.3 返回头信息：将详细罗列请求成功之后返回的头信息，如图 2.10.1.3。

（图 2.10.1.3）

1.4 源码：待请求完毕后，工具会自动跳转到源码选项，在此可查看请求成功之后所返回的页面源码信息。

1.5 预览：可在此预览请求成功之后返回的页面。

1.6 自动操作选项：可设置自动刷新/提交的时间间隔和运行次数，启用此操作后，工具会自动的按一定的时间间隔和运行次数向服务器自动请求，若想取消此操作，点击后面的停止按钮即可。

配置好上述信息后，点击“开始查看”按钮即可查看请求信息，返回头信息等，为避免填写请求信息，可以点击“粘贴外部监视 HTTP 请求数据”按钮粘贴请求的头信息，然后点击开始查看按钮即可。这种捷径是在粘贴的头信息格式正确的前提下，否则会弹出错误提示框。

第三章 工具菜单

（图 3）

一、任务批量编辑

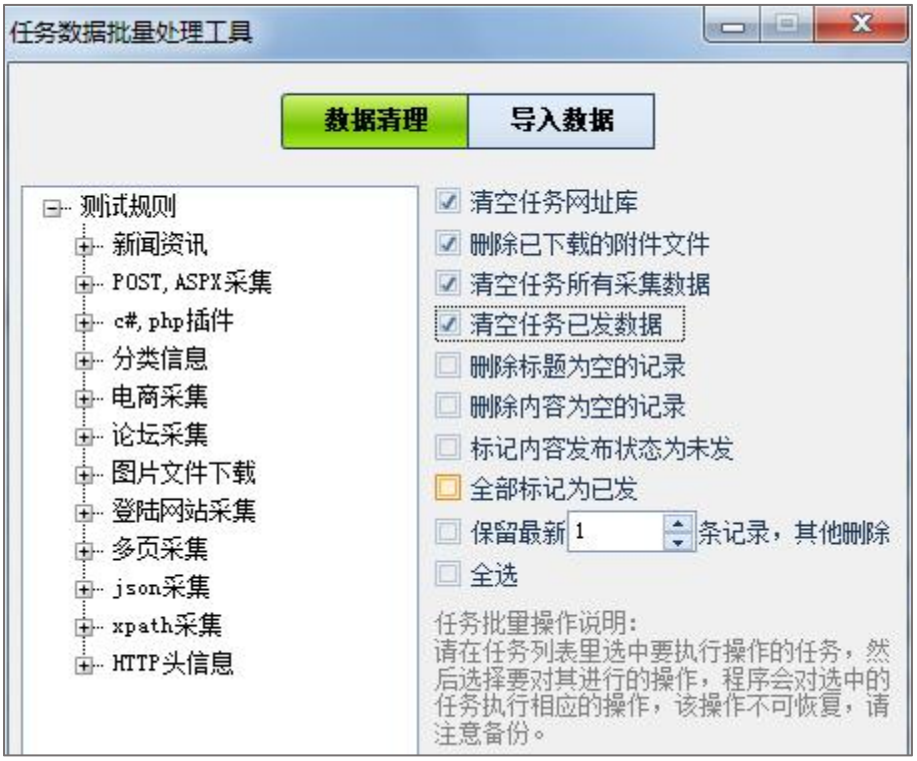
批量修改任务规则细节参数。

（图 3.1）

二、任务批量处理

1、数据清理

可对任务进行批量处理，包含清空任务网址库、删除已下载的附件、清空采集数据、清空已发数据等选项。



（图 3.2.1）

2、导入数据

即导入本地已经存在的 TXT 或者 EXCEL 内容到采集器规则的数据库内，用来发布等其他操作使用。

注意：导入 TXT 需确定原任务中有标题和内容标签；请确认文本编码：TXT 中 ANSI 编码对应 GBK，TXT 中 UTF-8 编码对应 UTF-8。导入 EXCEL 时表格第一行字段即为采集任务的标签。

（图 3.2.2）

三、远程管理

该功能可以通过 http 协议来对服务器上的采集器进行远程管理，即我们可以通过浏览器访问到我们的采集器来进行管理。

::: 任务列表 :::

ID	任务名	采内容	发布	状态	总数	已采	未采	已发	未发	操作	计划任务/查看数据
1	新浪_各地新闻	☑	☐	空闲						开始	查看数据 添加计划任务
2	网易国际新闻	☑	☐	空闲						开始	查看数据 添加计划任务
3	腾讯财经新闻	☑	☐	空闲						开始	查看数据 添加计划任务
4	POST分页地址	☑	☐	空闲						开始	查看数据 添加计划任务
5	中国低碳网-国内时政	☑	☐	空闲						开始	查看数据 添加计划任务
6	市交易中心施工工程	☑	☐	空闲						开始	查看数据 添加计划任务
7	网易新闻[翻译]	☑	☐	空闲						开始	查看数据 添加计划任务
8	58[图片号码]	☑	☐	空闲						开始	查看数据 添加计划任务
9	大众点评	☑	☑	空闲						开始	查看数据 添加计划任务

（图 3.3）

四、用户管理

该功能允许用户将自己写好的采集规则共享给一个或多个客户端用户。可以通过服务器用户管理设置用户帐号，用来限定用户下载规则的权限和允许用户访问的分组。客户端用户可以下载远程的采集规则，并可以有选择地更新任务,还支持一键更新所有远程的采集规则。对于一些非技术用户而言省去了写规则的麻烦，也帮助开设规则服务器的用户省去了远程指导的环节。

五、运行统计

用于统计用户运行的任务，可以以天、周、月或选择的时间段来查询，包括采集到的网址、重复网址、采集成功、采集失败、发布成功、发布失败等统计数据。

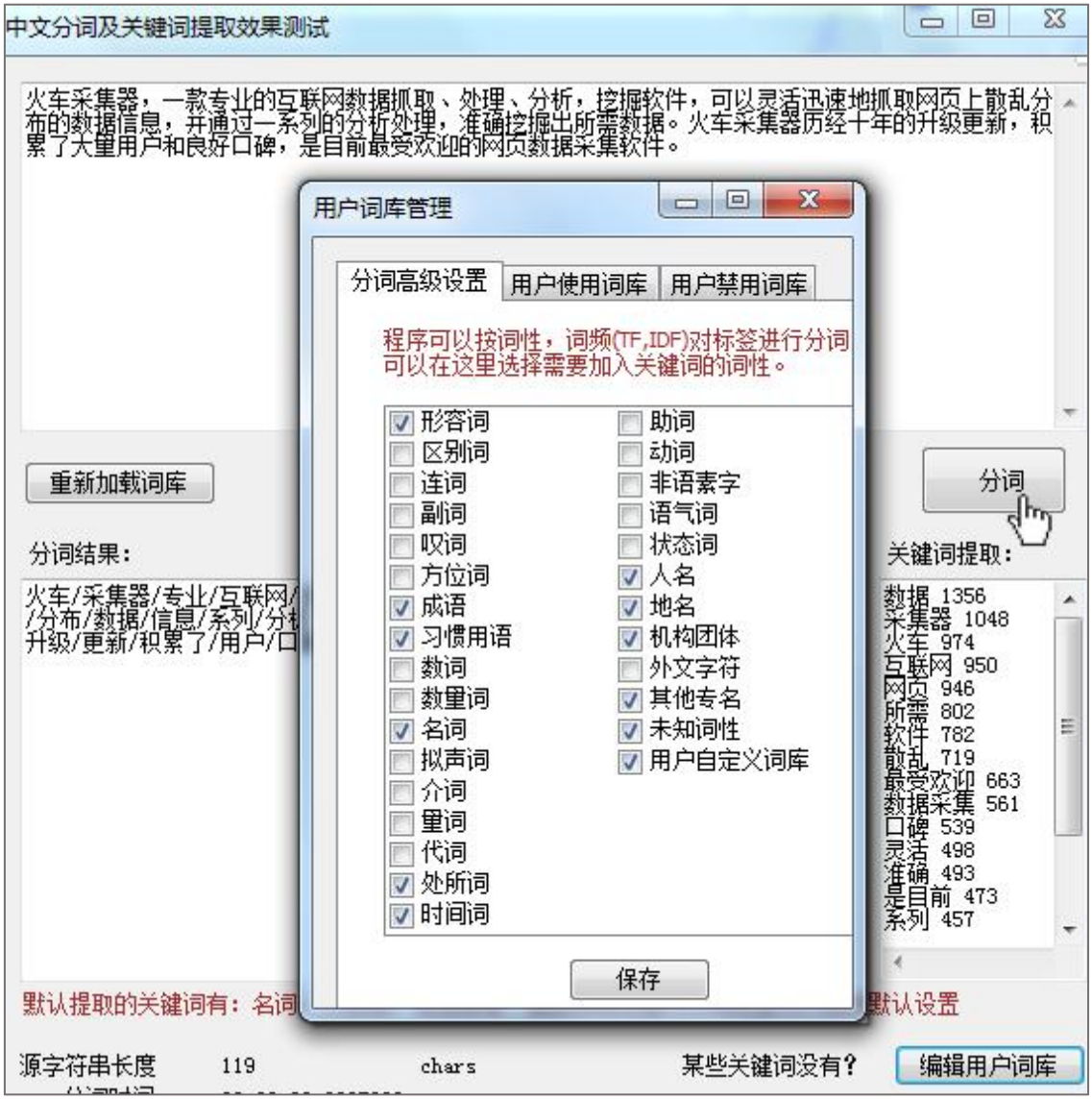
六、同义词替换

此功能可以将采集内容中的词语进行替换，自定义替换规则。（注意：两个同义词之间要用英文半角逗号连接。)比如将快乐替换成高兴，如图 3.6 所示“快乐,高兴” “美丽,漂亮”，然后保存，即可在第四章第三节的数据处理过程中选择对应词库使用替换功能。

（图 3.6）

七、中文分词

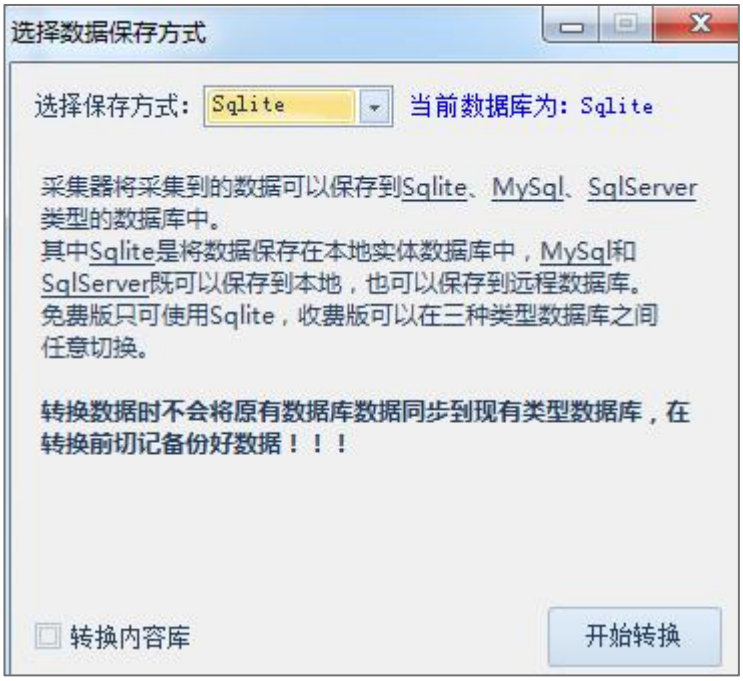
用来测试中文分词以及关键词提取的效果。可通过编辑用户词库，设置词性，词频，允许词，禁用词，来影响最终效果。



（图 3.7）

八、数据转换

数据采集下来后可选择保存到 sqlite、mysql、sqlserver 三种类型的数据库中。默认保存为 sqlite 数据库，如图 3.8 可转换为其他类型，其中 sqlite 是可以保存在本地数据库的。mysql、sqlserver 既可以保存在本地数据库，也可以保存到远程数据库。



（图 3.8）

九、数据同步

数据同步功能是将当前采集器中存在的任务保存到云端，以便在需要的时候进行下载和恢复。数据同步功能是默认开启，如果不需要使用此功能可以在：工具——选项——偏好设置中将其关闭，如图 3.9 的云端规则同步配置。

（图 3.9）

1、选项说明：

每个采集器同步数据后都会在云端建立一个属于此采集器的任务库。

1.1 开启云端任务同步上传：此选项勾选后开启云端任务同步功能。

1.2 云端获取任务同时获取对应的分组信息： 此选项勾选后将云端任务同步到本地时，也会将包含任务的分组同步到本地采集器中。

1.3 本机规则库备份规则，名称不同时直接覆盖：关闭自动同步之后，本地有改动的情况下，会导致同一个 ID 的任务规则，在云端存储的和本地存储的名称不同。那么同步的时候，勾选的话，就会将云端的覆盖本地。不勾选，下载下来的就是一个新的任务。（注：每个任务都有一个 ID 号，采集器区分任务是根据 ID 号来区分的。）

1.4 其他规则库的同名规则，直接覆盖本机规则：此选项勾选后同步其他的数据库中的任务到本地时，如果有其他任务库的任务名称和本地任务的名称相同，会将本地的同名任务覆盖。

选项设置完成后并不是立即生效，而是在下次启动后生效。

2、同步功能说明：同步功能在每次启动采集器的时候会与云端数据库进行对比更新。打开数据同步界面后可以看到当前采集器同步到云端的任务，以及其它相同用户名登录的采集器同步到云端的任务，如图 3.9.2。

（图 3.9.2）

2.1 数据列表：显示任务的分组和名称，及相应的选项框。勾选数据列表的选项框可进行清除当前数据的操作。

2.2 创建时间：显示任务的创建时间。

2.3 同步：显示同步选项框。

如需清除采集器在云端的任务，勾选数据列表栏的选项框后，点击清除当前云任务，可进行清除当前任务的操作。删除操作只能对当前采集器同步的任务进行操作，无法对其他采集器同步的任务进行操作。

如需将云端任务同步到本地，勾选数据列表栏的选项框或勾选同步列表栏的选项框，点击同步所选任务，即可将任务同步到本地采集器。刷新按钮可刷新云端任务。

十、选项

可对全局选项、采集设置、配置备份和使用偏好等进行设置，如图 3.10。

（图 3.10）

十一、自动关机

如勾选该项，则在任务运行完毕后自动关机。

（图 3.11）

第四章 操作指南

一、任务列表树

任务列表的分组可以进行无限级设置，分组下可设置分组也可设置任务。如

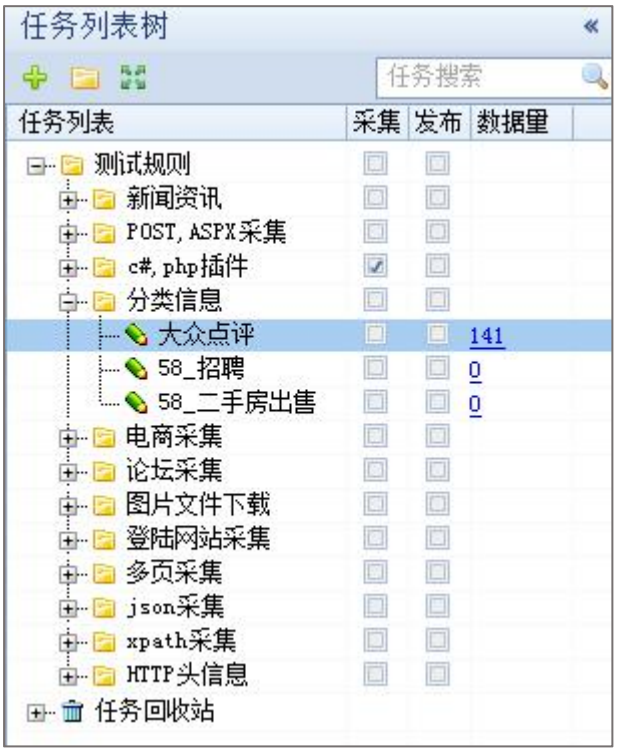


图 4.1 右键点击任务可进行进程控制、编辑、复制、清空数据等多种操作。

（图 4.1）

二、新建分组

新建一个分组，在分组下进行新建任务的操作。

（图 4.2）

三、新建任务

在任务列表树中或在开始菜单中新建一个任务，并对任务进行相关设置。

1、网址采集规则

我们在采集内容之前，需要将内容所在的页面网址获取到，在火车采集器中，为了获取网址而设置的规则称之为网址采集规则，也是实施采集的第一步。

1.1 起始网址

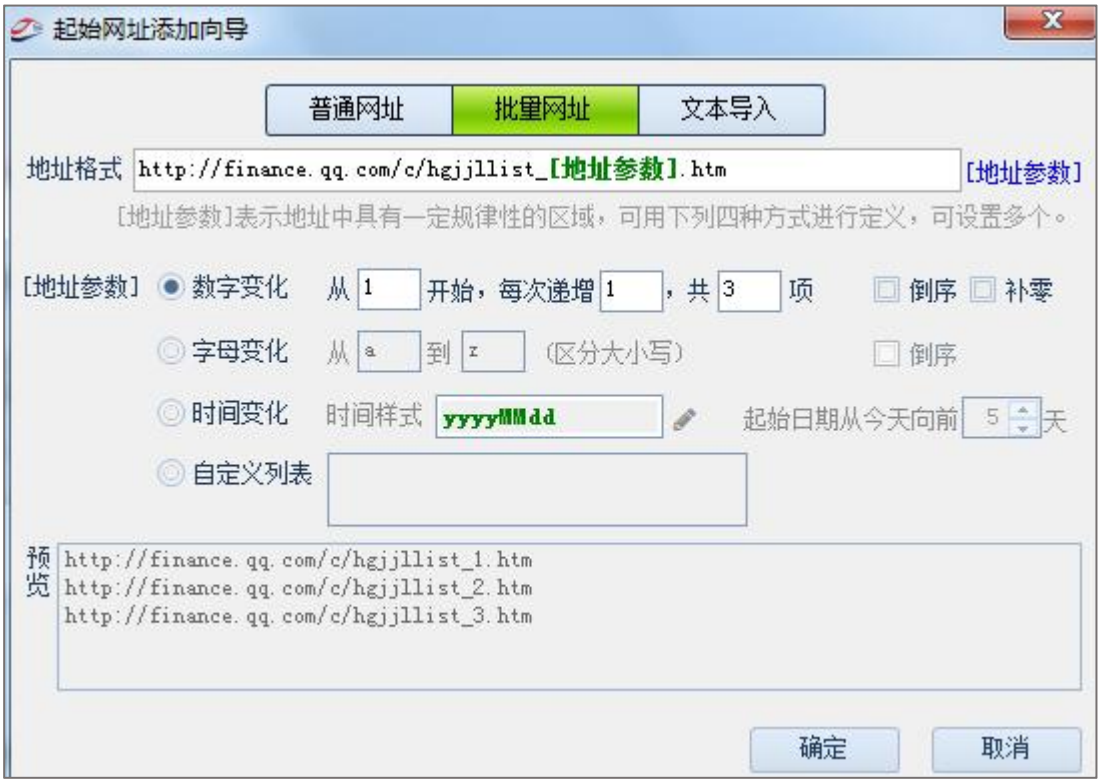
用来获取下级链接地址的入口网址，可以为一条或多条，可以通过添加起始网址向导添加同格式多条网址或导入文本网址。如果没有定义多级网址的获取方法，这些地址即作为内容页网址进行内容采集。

（图 4.3.1.1-1）

（1）向导添加

①普通网址：手动输入单条或多条网址 URL（一行一个，以 http：//或 https://开头）。

②批量网址：以通用的表达式批量生成网址。如图 4.3.1.1-2，可对有规律性数字变化的网址匹配数字递增表达式。



（图 4.3.1.1-2）

③文本导入：将文本中的网址导入采集器中，文本中网址需为一行一个。

（2）添加一行：对起始网址进行添加操作。

（3）清空：清空起始网址列表。

1.2 获取内容网址

（1）常规模式：该模式默认抓取一级地址，即从起始页源代码中获取到内容页 A 链接。它有自动获取地址链接和手动设置规则获取两种方式。

①自动获取地址链接 :选择此方式将会自动获取该级列表页中所有的 a 标签即<a href="URL">内的 URL 链接。为了使链接更加准确，还可以通过查看网页源代码，分析源码设置区域和链接过滤，如图 4.3.1.2-1。

（图 4.3.1.2-1）

②手动设置规则获取：对于有些由脚本生成的网址如图 4.3.1.2-2，采集器不能自动识别，此时就要手动设置规则获取了。手动设置规则获取设置原理是编写脚本规则，去和源代码里的内容匹配，获取到自己设置的参数即可。其中提取规则里的[参数]，(*) ， [标签:XXX] 都是通配符，可以通配任意字符，区别在于[参数]有返回值，一般用于拼接地址，(*)没有返回值，[标签:XXX]有返回值，返回值给标签。

```
<li><a href="http://news.sina.com.cn/c/nd/2015-10-10/doc-ifxirmpy1472664.shtml"
target="_blank">山西公布政府部门责任清单</a><span>(10 月 10 日 20:20)</span></li>
<li><a href="http://news.sina.com.cn/c/nd/2015-10-10/doc-ifxirwnr6902154.shtml"
target="_blank">河南登封市长被举报建寺涉贪</a><span>(10 月 10 日 20:14)</span></li>
<li><a href="http://news.sina.com.cn/c/nd/2015-10-10/doc-ifxirmqc5006034.shtml"
target="_blank">张家界国土局副局长涉严重违纪</a><span>(10 月 10 日 19:45)</span></li>
```

（图 4.3.1.2-2）

此时，我们可以取其中的一条代码作为循环匹配，把我们要获取的链接替换成[参数]，需要采集到的值替换成标签。 如如 4.3.1.2-3 所示。

（图 4.3.1.2-3）

（2）高级模式：该模式对 0 级，多级，POST 类型网址的抓取有效。即起始网址就是内容页网址； 或者需要对多级列表网址采集才能得到最终内容页链接； 或者是 post 网址类型抓取等情况下使用高级模式。

（图 4.3.1.2-4）

①多级列表：需要设置多级网址步骤后，才能得到最终内容页链接。可添加或删除列表页，每设置一级，则点击添加一次，然后在对应的网址获取选项中按照常规模式进行网址获取方式设置。

（注：多级列表为空或者大于 1 级，不能切换到常规模式。可以先添加/删除至一级列表规则，再进行切换。）

②网址获取选项：同上述常规模式。

③分页设置：这里需要明白何为分页，列表页面较长，分成多个页面显示，像是新闻列表的第一页、第二页这样的页面（分页包括列表分页和内容分页，这里操作讲解的网址采集所需获取的列表分页）。

分页设置可选的 http 请求方式有 GET、POST、ASPX POST 三种，这里可以根据页面的请求方式来决定，比如是用 post 提交翻页请求的，那么页面只进行局部刷新，地址栏中的 URL 不变。所以处理此类采集时的思路就是用抓包工具 fiddler（抓包教程可参考 http://bbs.locoy.com/spider-107387-1-1.html），截取请求时提交的内容找出共同特点，用“分页”变量进行替换并给定值范围，这样在采集时会自动提交请求内容得到新的内容列表进行采集。

GET：设置区域开始字符串、区域结束字符串，如图 4.3.1.2-5 所示采集器则会在此区域内提取分页地址，同时设置地址样式、分页地址等帮助识别地址。

（图 4.3.1.2-5）

POST：如图 4.3.1.2-6 所示，post 方式请求时主要填写发送的数据，发送数据由指定字符数据和“[分页]”、“[POST 随机值 X]”构成。“[分页]”标签可以指定某一区域值，开始数值必须小于结束数值；“[POST 随机值 X]”从随机值列表获取；随机值是通过前后字符串从采集到的内容里获取的。

（图 4.3.1.2-6）

ASPX POST：此种请求可自动识别大多数 ASPX 网站的 POST 分页。但是需要注意的是：此方法并不能保证全部解决这类分页地址不变的网站。

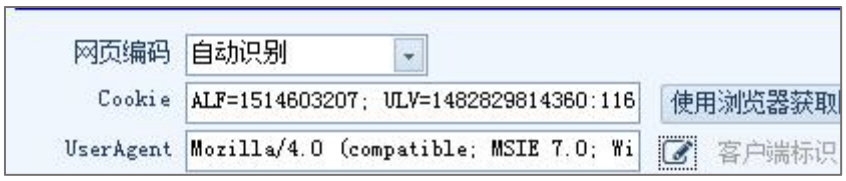
（图 4.3.1.2-7）

④列表页独立标签：从整体列表页源码中独立获取的标签，不参与内容网址获取的循环，将直接复制填充到每条记录。

（这里区别于列表页标签：指列表页下每个内容页地址的循环源码块中的标签，与内容页获取的标签共同组合成一条完整的记录。）

1.3 登录采集

如果采集的网站需要登录才能访问，或者是需要指定的浏览器才能访问，可点击页面上的提示进行设置。Cookie 可通过使用火车采集器内置的浏览器来获取；userAgent 可选本地 IE 浏览器、Google Chrome 浏览器、火狐 firefox 浏览器、百度蜘蛛，手机客户端等。



（图 4.3.1.3）

1.4 采集网址测试

完成规则设置后，测试网址采集的准确性。

（图 4.3.1.4）

2、内容采集规则

2.1 标签列表

对需要采集的内容进行标签设定，方便后续对不同的内容实施采集。如新闻的标题、内容，来源等，如图 4.3.2.1-1。

（图 4.3.2.1-1）

（1）数据获取方式：对标签中数据的获取方式进行定义。

①从源码中获取数据 :从源码中获取数据可精确地设置标签的来源是从默认页的源码、返回头信息和网页地址中，或者是分页、循环分块、多页中。

其数据提取方式包括前后截取、正则提取、正文提取、Xpath 提取和 JSON 提取五种，并可对数据进行过滤、替换等一系列处理。

前后截取：如图 4.3.2.1-2 所示通过设置开始字符串和结束字符串，来获取中间的字符，可以在开始和结束字符串中设置通配符（*）。

（图 4.3.2.1-2）

正文提取 :注意这种方式只适合格式较为规则的多文字数据提取，例如新闻文章，可智能分析提取文章正文，文章标题，以及发布时间。

（图 4.3.2.1-3）

正则提取：支持两种正则，一个纯正则，一个参数正则。先介绍纯正则，举个例子：如：前字符串 (?<content>[\s\S]*?)后字符串，这个正则其实效果跟前后截取一样，程序提取匹配到的 content 组的内容（注意：content 是程序约定的组名，必须包含）；关于参数正则，是通过参数组合，来生成内容。比如说要匹配如下内容：标题：正则表示式 30 分钟教学视频，我们想要得到的字符如下，正则表示式 XXX 视频可以这样写，内容部分：[参数]30 分钟教学[参数]，组合结果部分：[参数 1] XXX[参数 2]。此功能运用需有一定的正则基础。

（图 4.3.2.1-4）

Xpath 提取：通过 Xpath 表达式来获取数据，比如//div[@id=’ content’]，就是获取 id 为 content 的 div 可指定要获取 html 节点的属性，比如 Innerhtml、Outerhtml、Innertext，Href 属性。（注意：这种有一定的局限性，对于部分 html 标签不规范的页面无法解析。）

（图 4.3.2.1-5）

JSON 提取 通过对 JSON 形式的数据格式化操作，写表达式来获取其节点数据。Json 的相关操作可自行学习一下，或参考采集器官网教程。

（图 4.3.2.1-6）

②生成固定格式的数据：可选择生成固定的字符串、系统时间、随机字符串等。

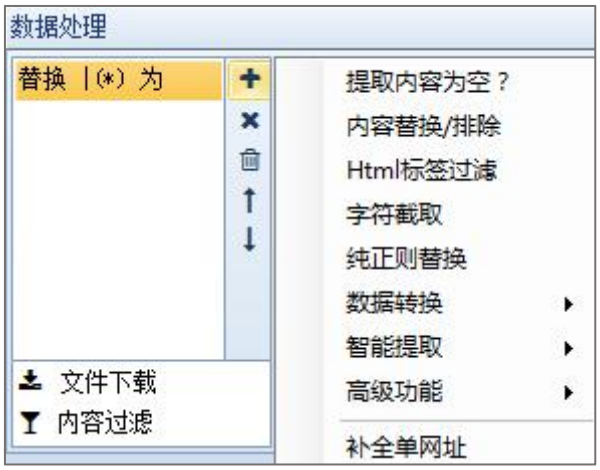
（图 4.3.2.1-7）

③已有标签组合：可对编辑后的标签进行组合，将按照组合格式显示数据如图 4.3.2.1-8。

（图 4.3.2.1-8）

（2）数据处理

对从内容页面提取的数据进行进一步处理，可以同时添加多个操作，如图 4.3.2.1-9 所示按照从上到下的顺序来执行，也就是说，上个步骤的结果会作为下个步骤的参数。



（图 4.3.2.1-9）

①提取内容为空：如果提取内容为空，则使用正则匹配从原始页面中再次提取。

②内容替换/排除：将采集到的内容进行字符串替换，如需排除，则替换为空字符串即可。

③html 标签过滤：过滤指定 html 标签，比如<a ， <font。

④字符截取：通过开始和结束字符串对内容进行截取。

⑤纯正则替换：通过强大的正则表达式进行复杂的替换。

⑥数据转换：包括将结果汉译英、将结果简转繁、将结果繁转简、自动转化为拼音和时间修正转化，共计五项处理。

⑦智能提取：包括提取第一张图片、智能提取时间、智能提取邮箱、智能提取手机号码、智能提取电话号码。

⑧高级功能：包括自动摘要、自动分词、Http 头信息提取、Http 请求、字符编码转换、同义词替换、空内容缺省值、内容加前后缀、随机插入、运行 C# 代码、批量内容替换，统计标签字符串长度等一系列功能。

⑨补全单网址：将当前内容作为一个网址进行补全。

（3）文件下载：可以自动探测并下载文件，可设置下载路径和文件名样式，如图 4.3.2.1-10 所示。

（图 4.3.2.1-10）

注意：文件下载中所指下载图片是源代码里有标准样式标签的图片地址。

如是一个直接的图片地址 http://www.locoy.com/logo.gif ,或者不规则的图片源码，采集器将会视为文件下载。

①将相对地址补全为绝对地址 :勾选后会把标签采集到的相对地址补全为绝对地址。

②下载图片 :勾选后源代码里的含标准样式的代码图片将被下载。

③探测文件真实地址但不下载 :有时候采集到的是附件下载地址，而非真实的下载地址，点击后会有跳转。这种情况下勾选此项会将真实地址采集出来，但是只是得到下载地址并不下载。

④探测文件并下载：勾选后可以把采集到的任何格式的文件附件下载下来。

（4）内容过滤：对于一些不符合条件的记录，可以通过设置内容过滤来删除或标记为未采。内容过滤有以下几个处理方法：

（图 4.3.2.1-11）

①内容不得包含和内容必须包含：可以设置多个词，支持选择所有条件都必

须满足或满足其中一个条件即可，如图 。

②采集结果不得为空：该功能可以让某个字段不出现空内容。

③采集结果不得重复：该功能可以让某个字段不出现重复内容。设置此项前请确保没有采集过数据，或者需先清空采集数据。

④当内容长度小于(大于，等于，不等于)N 时过滤：一个符号或一个字母或一个数字或一个汉字都计作一个。

2.2 内容分页

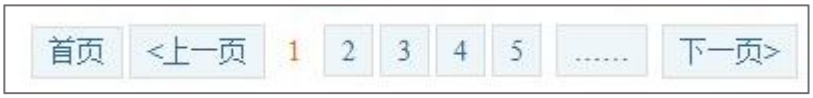
获取内容分页时包含首页全部列出，上下页模式两种列出模式。根据开始和结束字符串确定分页网址的提取区域，在此区域提取分页链接时可选择自动识别和手动设置规则。

（1）列出模式

①全部列出：适用于分页地址全部列出的类型，如图 4.3.2.2-1。

（图 4.3.2.2-1）

②上下页模式：如果分页地址只列出一部分可以使用此种模式，如图 4.3.2.2-2。（注：此种模式全部列出的类型也同样适用。）



（图 4.3.2.2-2）

（2）链接提取

①自动识别：采集器可在设置的范围内，自动匹配到分页地址。

②手动设置规则：遇到无法自动识别或者识别不是非常准确的情况，可以手动补充上分页的格式，来确保识别分页的正确性如图 4.3.2.2-3。

（图 4.3.2.2-3）

2.3 循环设置

对于标签循环的匹配结果可以根据个人需求选择添加为新记录或用分割符连接在上条记录后，分割符可自定义，效果如如 4.3.2.3。

（图 4.3.2.3）

2.4 关联多页

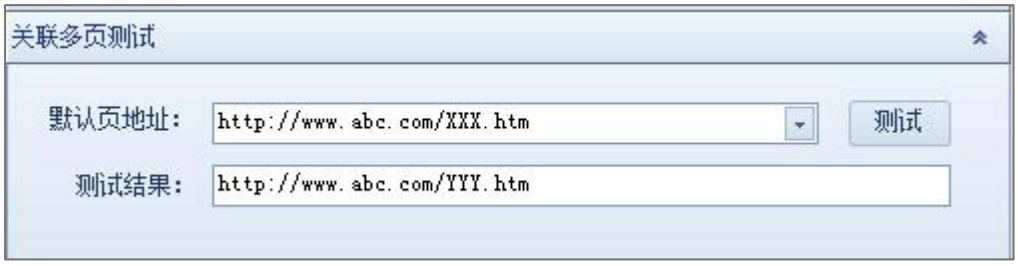
当采集的信息不在当前默认页，而在当前默认页某一个链接的所在页时，则需要找出关联的多页。

（1）多页地址的获取有两种方式：①页面地址替换②源码中截取

①页面地址替换：如多页与默认页存在相似的部分，那么可以使用此种方式将地址中不同的部分替换从而获取多页地址。

（图 4.3.2.4-1）

具体的替换内容依据网页地址的规律而定，这里也支持正则表达式替换，有正则基础的用户可以使用此功能。图 4.3.2.4-1 所示的替换测试如下图 4.3.2.4：



（图 4.3.2.4-2）

②源码中截取：如默认页面中包含多页地址，那么我们采用源码截取的方式找出多页地址，这时根据字符串截取到的参数就会出现在组合结果中。

（图 4.3.2.4-3）

（2）指定多页源码区域：如果需要的是多页页面中的部分内容，则需要指定源码区域来设置采集区域，如不设置则默认返回多页整个页面的源代码。

2.5 其他设置

（1）最大内容分页数量

当采集页面有分页时，可以在此设置分页最大数，若采集到的当前分页数超过该值，则不会继续采集超出该值后面的分页。若设置为 0，则为不限。

（图 4.3.2.5）

（2）每次最大采集数量

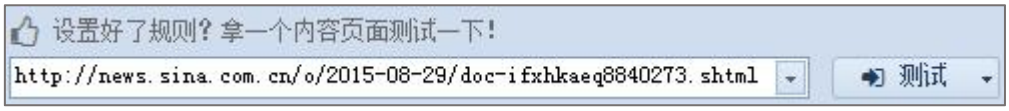
设置当前任务采集的最大数，若当前采集到的记录条数超过此值时，停止采集。设置为 0 时，当前任务采集不限制，默认为 0。

（3）分页链接代码

当设置了标签在分页中匹配，且没有设置循环匹配的两个页面间的内容，会以分页内容连接代码合并。例：一个标签“内容”有分页，在分页一中采集到的内容为“内容 1”，在分页二中采集到的内容为“内容 2”，分页内容链接代码为“p#p”，则最后的内容为：内容 1 p#p 内容 2。

2.6 规则测试

填写好内容采集规则后，点击测试按钮即可进行某一页面的测试。若规则里有下载文件或下载图片设置时，采集器会在测试时用采集器自带的下载工具下载相关文件或图片。

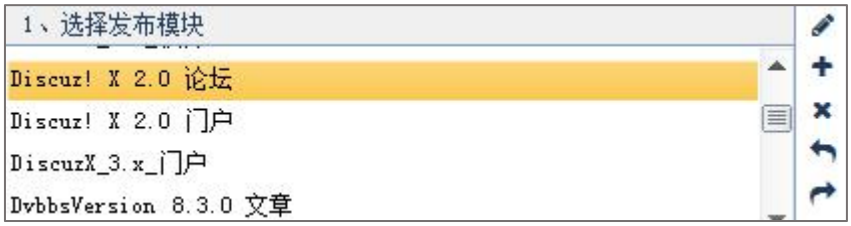


（图 4.3.2.6）

3、内容发布规则

3.1Web 在线发布

（1）选择发布模块：新建或选取 web 发布模块中已编辑好的模块（发布模块编辑在第二章第四节 **web 发布模块**中已详细说明），并可对模块进行编辑等操作。



（图 4.3.3.1-1）

（2）网页编码：选择对应的网页编码。GBK、GB2312、UTF-8，BIG5 等。

（图 4.3.3.1-2）

（3）全局变量：全局变量可以在发布模块中所有位置使用，方便设置和修改某些参数。

（4）登录操作：可通过内置浏览器登录，数据包登录，无需登录则选择不登录。

①内置浏览器登录：获取浏览器标识和用户信息。

（图 4.3.3.1-3）

②数据包登录：填写用户名，密码以及获取到的验证码后登录。此种方法需要所选择的发布模块里“网站自动登录”有对应设置，具体参见第二章第四节。

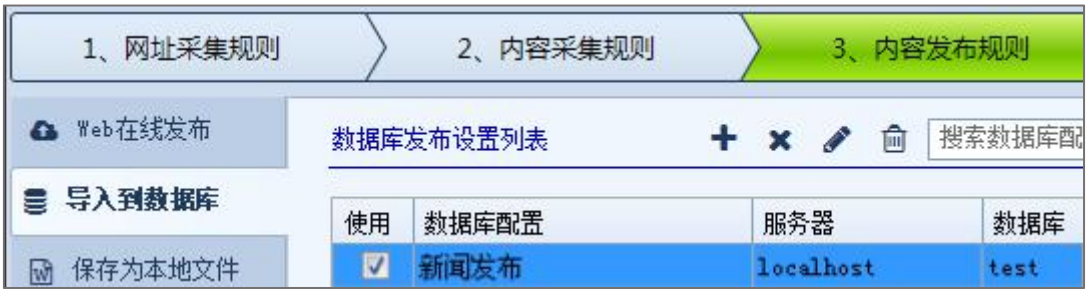
（5）获取分类栏目列表：可刷新出栏目 ID 和栏目名称。需要上文所选择的发布模块里“获取栏目列表”有对应设置。

此部分内容在第二章第三节中已具体说明，完成相应配置后在列表中对应勾选即完成了在线发布设置。

（图 4.3.3.1-4）

3.2 导入到数据库

此功能用于将采集的数据发布到自定义的数据库里面。新建或选取数据库发布模块中已加载好的某一模块，如图 4.3.3.1-5，新建模块的操作过程在第二章第五节、第六节中已做详细介绍。



（图 4.3.3.2）

3.3 保存为本地文件

支持保存为本地 Excel、Txt、Word、Html、Csv 类型的文件。启用本地文件保存后，填写保存文件格式、保存方式、保存位置、文件模板、文件名格式、文件编码等信息后点击保存即可。

（图 4.3.3.3）

注意：1、文件模板中的标签必须与规则里的标签名相对应，否则将保存出错。

2、文件模版的编码需与软件此处设置的编码保存一致，否则发布后显示乱码。

4、其他设置

4.1 任务运行线程及时间

火车采集器 V9 支持多线程采集和发布，加快采集发布速度，但也要考虑到电脑配置以及网络状况等不定因素，并非设置得越多越好，因此设置一个合适的线程数和间隔时间十分重要。

单位：1000ms（毫秒） = 1 秒（该项的设置也可在任务运行时进行动态更改，实时生效）

（图 4.3.4.1）

4.2 HTTP 请求设置

在发送一个 http 请求前，先在此界面设置好相关配置，如图 4.3.4.2。列表页、内容页、分页、关联多页均使用该配置。

（1）默认设置

①网页编码：默认自动识别，也可根据采集网站自定义编码。

②Cookie 和 UserAgent：获取浏览器标识和网页登录信息。支持使用浏览器获取网页登录信息，或者使用抓包工具 fiddler 获取放入。

③Referer：正确填写请求页的来源页地址，如未填写则默认使用起始网址或上一级页面。

④AutoRedirect：决定当前请求是否应跟随重定向响应。

⑤Keep-Alive：决定当前请求是否与 internet 资源建立持久性链接。

⑥Accept-Encoding：决定是否支持 gzip、deflate 内容压缩编码。

⑦Accept-Language：接受何种语言。

⑧http 超时时间：设置 http 超时时间。

（图 4.3.4.2）

（2）基于 Windows 身份认证：正确填写用户名，密码，域即可，无身份认证时不必填写。

（3）http 请求头列表特殊设置：显示发送的头信息，以列表形式显示更清晰直观的了解到请求的头信息若要将某一名称的头信息进行请求，在列表添加并下拉选择对应的 Header 或设置名以及所属页，Header 或设置值是可以进行编辑的。

4.3 文件相关下载

（1）普通文件：设置普通文件的保存目录、链接地址前缀、下载模式等。

（2）图片文件：对下载的图片进行水印选项和设置，如图 4.3.4.3-2 所示设置水印图片。

（图 4.3.4.3-1）

（图 4.3.4.3-2）

4.4 FTP/SFTP 文件上传

此项包含不上传，使用 FTP 上传，使用 SSH 上传三种模式。

（1）FTP 上传：如果使用 FTP 软件上传，需填写服务器、用户名、密码信息。

文件上传根目录填写 FTP 软件登录显示的路径，如/www/images/20150912/那么根目录则填写/www/如图 4.3.4.4-2，不需要重复把 images/20150912/ 写上，因为在标签编辑的文件下载设置里已经定义了如图 4.3.4.4-1 所示的文件保存路径，软件会自动创建，并识别要发布到哪里。

（图 4.3.4.4-1）

（图 4.3.4.4-2）

（2）SSH 上传：同样的，需要填写服务器、客户端用户名和密码信息。但需要注意的是，当使用 SSH 上传时，文件上传的根目录建议填写绝对路径，例如/root/pic/xxx。

4.5 代理设置

有些网站的防采集策略是禁止 IP 的访问，这时需要设置 IE 浏览器代理或指定代理才能采集。代理设置时填写正确的且与代理选项设置中一致的代理服务器端口和用户名等其他信息保存即可。

（图 4.3.4.5）

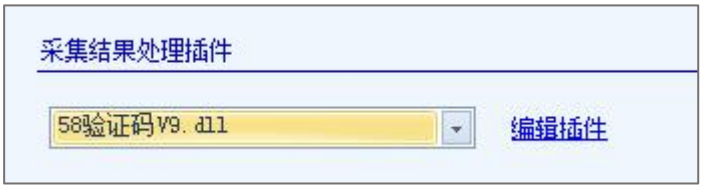
火车采集器中此项包含不使用代理，使用 IE 浏览器代理，使用指定代理三种模式，如图 4.3.4.5 所示。其中指定代理模式可以设置固定的一个代理或者二级代理随机切换 IP 采集，此部分的代理配置内容在第二章第九节中有具体说明，点击使用采集器二级代理即可开启之前配置的二级代理功能。

4.6 插件

火车采集器 V9 支持 PHP 源码、C#源码、C#类库三种类型的插件，可用于扩展 http 请求、内容处理和文件下载的功能。下面以 58 插件为例演示该功能的用法：

（1）首先我们需要把插件 **58 验证码 V9.dll** 放入到采集器的 **Plugins** 目录中；

（2）然后在采集结果处理插件一栏中选择此插件；



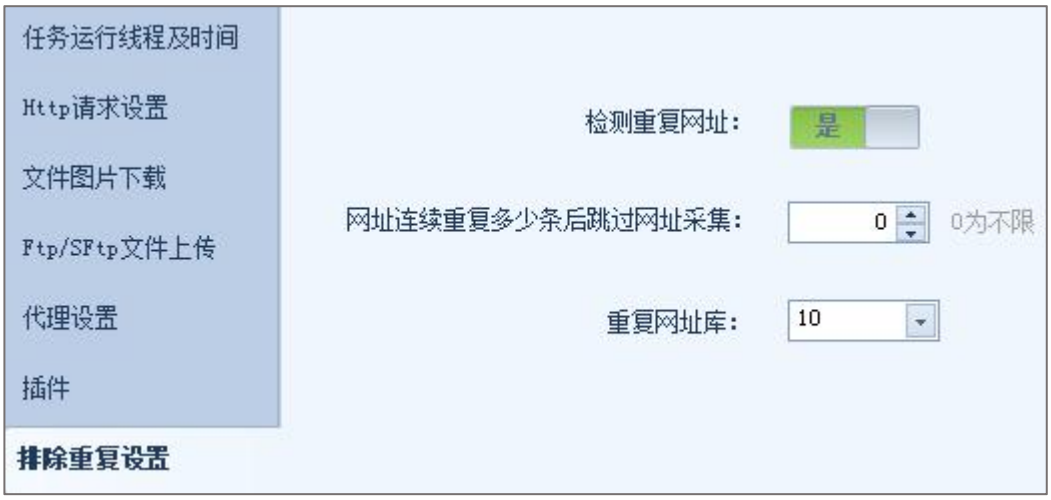
（图 4.3.4.6-1）

（3）最后我们需要在内容标签列表中建立一个名为 “手机号码” 的标签，此标签采集 58 手机号码的图片地址，在任务运行的时候，火车采集器就会自动调用此插件来将图片转义成数字文本的形式输出。

（图 4.3.4.6-2）

4.7 排除重复设置

设置是否检测网址重复，以及网址连续重复多少条后跳过网址采集，并支持选择网址库。



（图 4.3.4.7）

4.8 发布相关

与发布相关的一些细节设置，包括发布结束后是否删除，清空，是否边采集边发布，web 发布顺序方式等，如图 4.3.4.8 所示。

（图 4.3.4.8）

（1）每次最大发布记录条数：设置每次发布条数。

（2）边采集边发布：采集一条发布一条，采集的数据不入本地数据库，直接发布到网站，提高采集发布效率。

（3）Web 发布方式：支持正序，倒序，乱序，多网站乱序。其中多网站乱序反复，需要添加多个发布配置，此功能可以将多条内容随机分配发布。

四、运行管理

任务运行管理区域可以显示每个任务的运行界面。有多个任务在运行时，会有多个任务运行界面显示。可以查看运行日志、实时数据、文件下载，可以任务详情中实时调节任务线程，也可以控制任务的开始、暂停，停止。

（图 4.4）

第五章 软件适应性

一、运行环境

系统环境：
win2003/Win2008/Win2012/Windows xp/windows vista/Win7 / Win 8/Win10
框架支持：要求电脑安装.NET4.0 框架支持
平台选项：Anycpu X64 位开发，兼容 32 位及 64 位系统

二、授权方式

火车采集器采用网络认证进行授权，通过用户名和密码登录，在使用时需绑定电脑，不同版本支持绑定的电脑数量(即可允许同时在线使用的电脑数)不同，可按需选择。

三、软件升级

火车采集器官方会定期对软件进行功能的扩充和更新升级，用户在所选择的服务年限内可通过官网发布的地址下载最新版本的软件，如超出服务年限可选择续费后更新，如无需更新，原版本仍旧可以使用。

四、适应性服务

1、专属技术支持渠道
2、按需制定培训计划
3、定期视频培训课程
4、可进行 OEM 定制
5、支持自行扩展开发插件

五、技术支持

官网地址：http://www.locoy.com
企业 QQ：800019423
技术交流 QQ 群：183936276
官方论坛：http://bbs.locoy.com
文档教程：http://www.locoy.com/index/guide
视频教程：http://www.locoy.com/video
FAQ 系统：http://faq.locoy.com
技术邮箱：admin@locoy.com