

火车采集器 7.0 功能简介

火车采集器 7.0 版进行了重新的设计，功能更加强大，使用更舒适。以下是新版增加的和增强的功能简介

1.数据采集平台

打开菜单 扩展 ，即可以看到扩展设置及平台软件仓库。



火车采集器新版基于于火车头数据采集平台，火车采集器只是平台中的一个扩展应用程序。在该平台中，提供了统一的程序管理配置界面，如计划任务，OCR，分词，入库，Web 发布，分组管理，任务管理等功能。其它任务使用统一的 API 进行开发，可以调用平台的功能。这将极大的方便软件的使用和开发：用户可以很方便的下载新软件，新软件开发也可以使用更少的时间和精力。如果用户是开发人员，可以在平台的基础上开发自己的应用。比如我们的论坛数据采集专家也将集成到该平台中。

2.新版火车头数据采集平台支持安装为系统服务，支持 httpServer，支持主从服务模式(服务端+客户端)

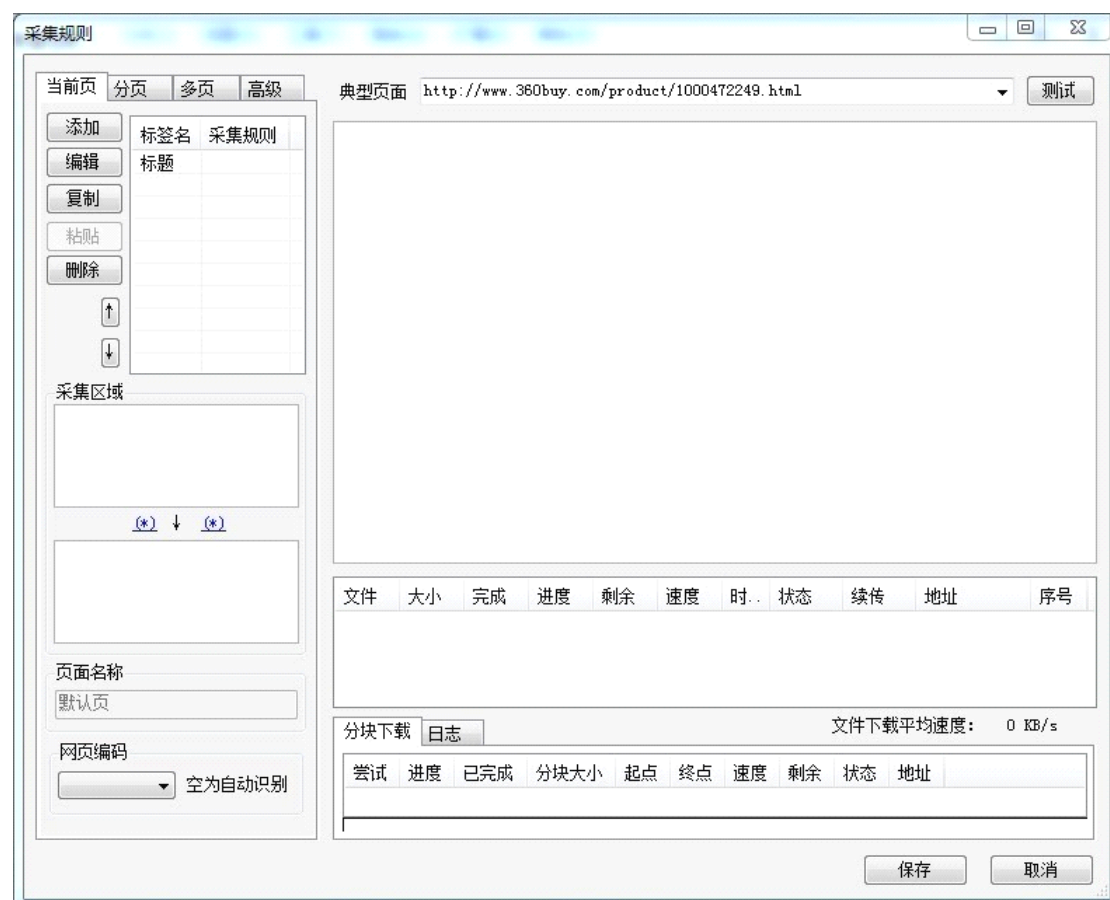
该功能可以让采集器由服务端分配采集任务，客户端进行数据采集，极大的方便了分布式采集大批量数据。该功能将在新版发布后开始开发。

3.无限级分页采集功能

通过在火车采集器任务编辑第二步数据采集，点击使用无限级分页采集模式



在采集规则部分，我们可以使用 2010 版的设置方式，也可以使用无限级页面编辑模式。新版本的采集内容规则是继承了无限级页面编辑模式。两种模式可无缝切换。



在无限级多页编辑模式下，我们可以在多页上再获取新的多页，对于每个多页页面，可以设置自己的采集区域，标签，分页，还可以设置页面编码，甚至 http 请求。通过多页设置功能，可以达到有多少多页采集多少多页的效果。

4. 计划任务支持分组和 Cron 表达式

在高级菜单中或是工具栏中。点击计划任务管理器，就可以打开计划任务界面



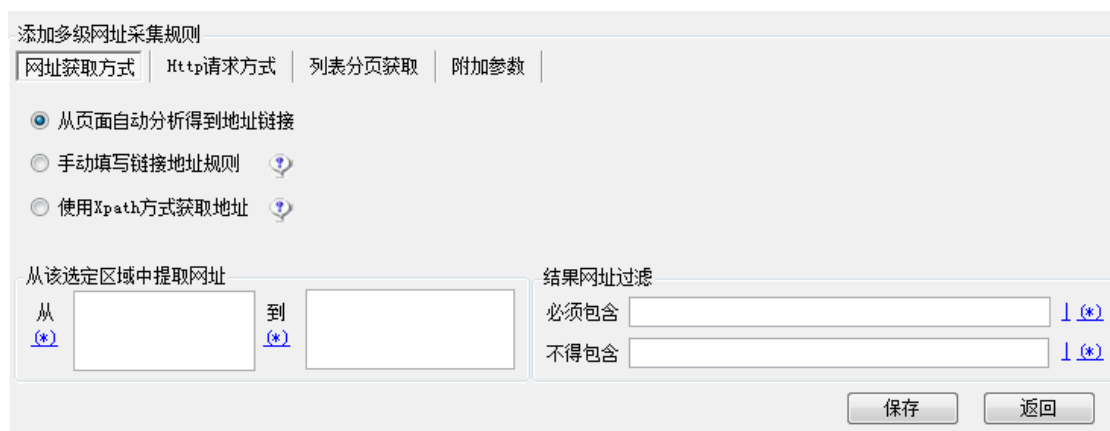
新版本计划任务支持分组，计划任务使用多种设置方式。还支持 Cron 表达式，可以方便用户设置更为复杂的运行计划。

5.采网址部分可无限级采集网址。

新版火车采集器在第一步采网址部分，对采网址级数进行了升级，可以支持更多深度的网址采集。点击添加多级网址，就可以看到



打开后，我们可以看到好几列的设置



在这里，可以设置每一级网址的请求方式，获取方式，还增加了列表分页的获取。特别需要说明的是还增加了一个附加参数功能。该功能可以获取列表页的一个参数，然后将结果合并到每一个最后获取的结果中去。可以实现获取栏信息等需求。

添加多级网址采集规则

网址获取方式 | Http请求方式 | 列表分页获取 | 附加参数

从 到 地址样式 [参数1] (*) ?

最多获取分页数, 0为不限 0 ☒ 自动识别分页 分页网址 [参数1] [参数N]

5.采内容和采网址支持 Xpath.

在标签编辑中, 选择 XPath 方式可以使用这种所见即所得的试获取网页数据, 更方便对于 html 不熟悉的用户使用。

标签编辑

标签名: 内容 ☐ 该标签循环匹配 ☐ 该标签在分页中匹配 ☐ 从网址中采集 ?

☒ 通过采集得到数据 ☐ 自定义固定格式的数据

提取数据方式

☐ 前后截取 ☐ 正则提取 ☒ 可视化提取 ☐ 正文提取 ☐ 标签组合 所属多页: 默认页 ?

XPath表达式

/html[1]/body[1]/div[5]/div[1]/div[1]/div[1]/div[2]/div[2]

选择节点属性

☒ InnerHtml ☐ InnerText ☐ OuterHtml ☐ Href

通过XPath浏览器获取

数据处理

添加 删除 清空

文件下载选项

☐ 将相对地址补全为绝对地址 ☒ 下载图片

☐ 探测文件真实地址但不下载 ☐ 探测文件并下载

文件地址必含 (*) |

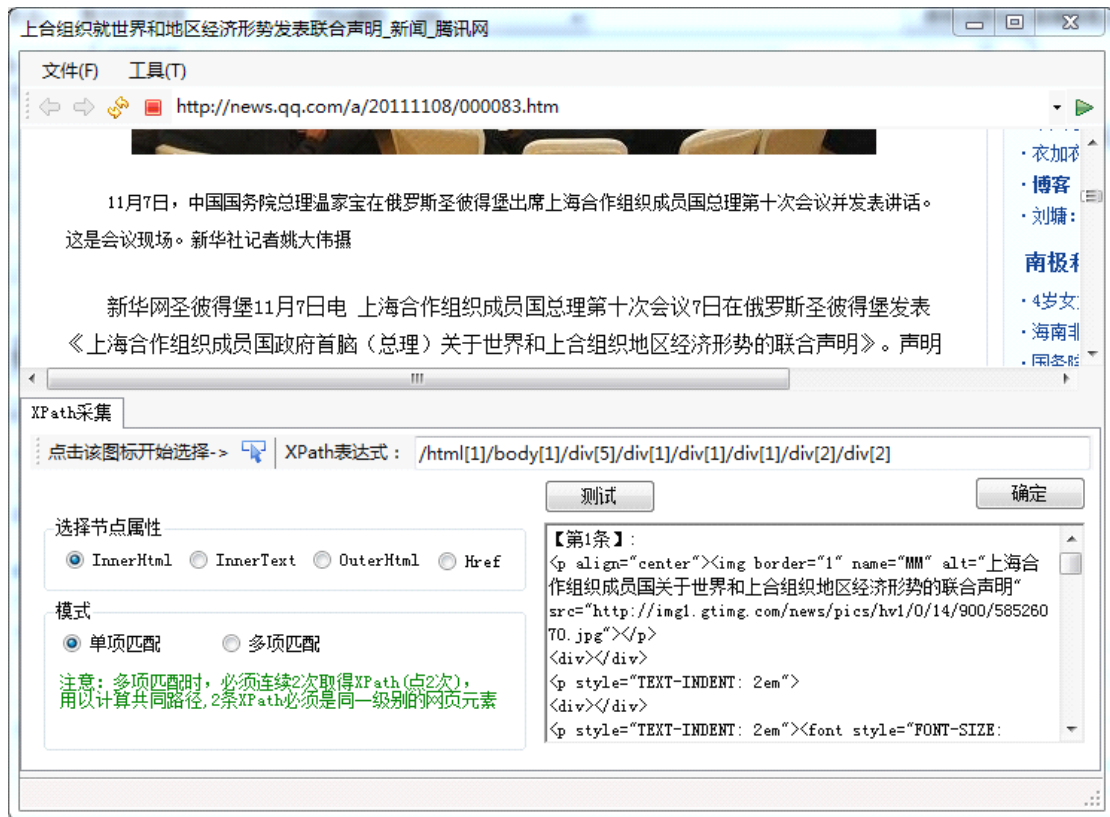
文件保存目录 yyyy\dghdmm ?

文件保存格式 /fdfd/[原文件名] ?

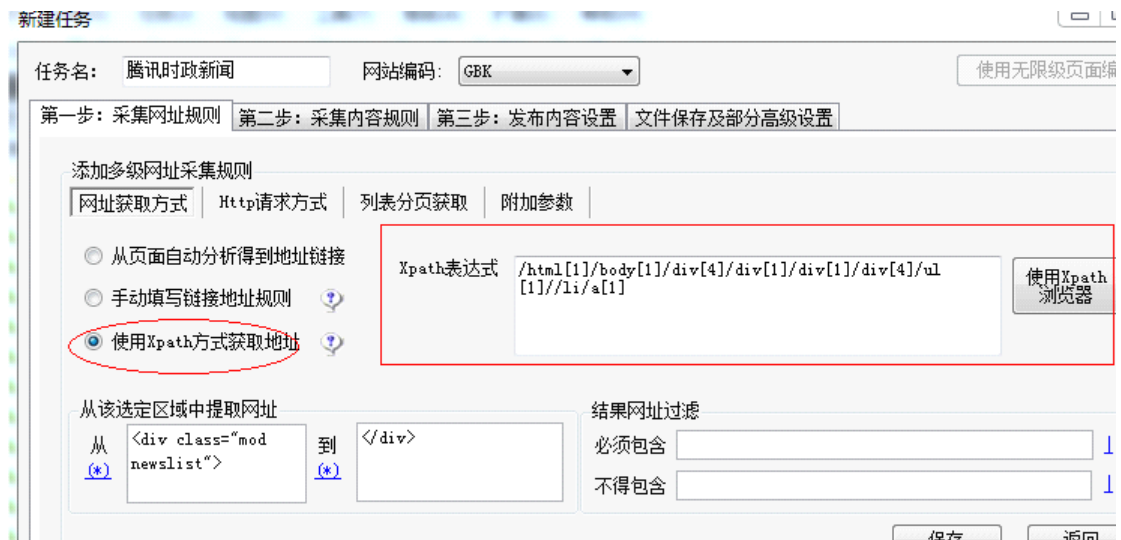
文件下载内容过滤

确定 取消

点击通过 XPath 获取 XPath 表达式, 我们可以看到这里已经提取到内容了, 点击确定即可。

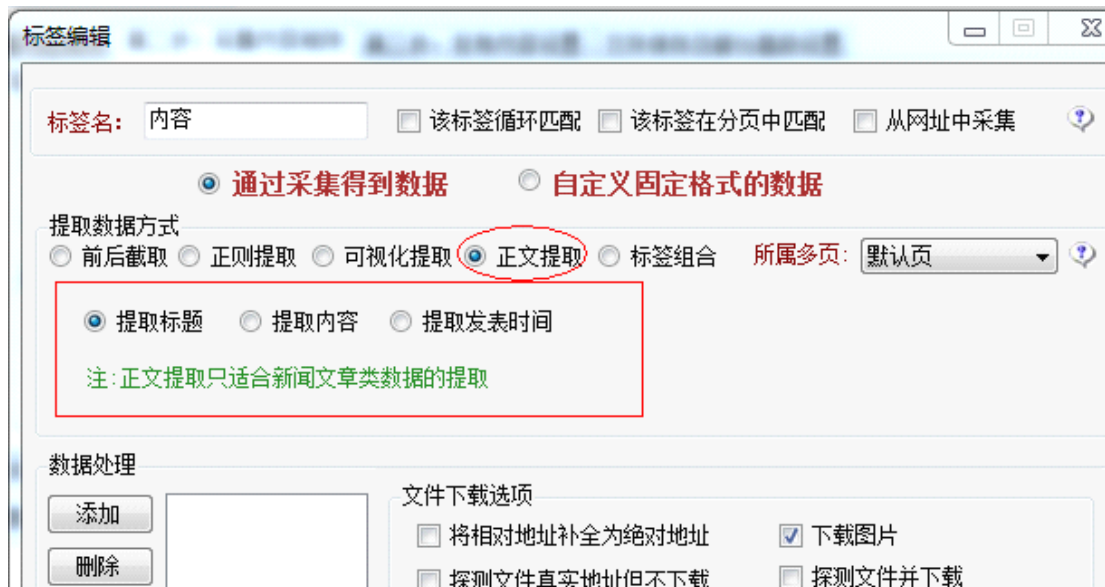


列表页用 XPath 获取内容页网址



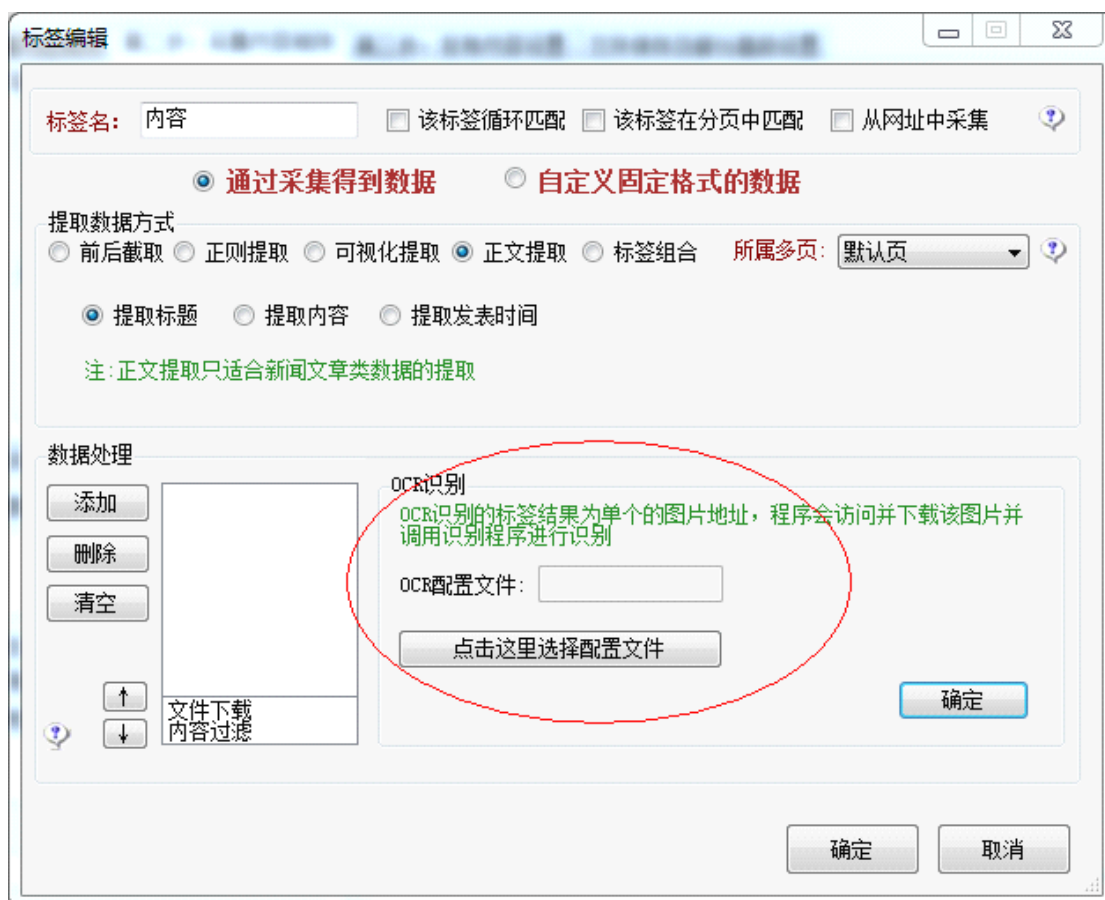
6.集成正文提取功能

在标签编辑中，选择数据提取方式为 正文提取 ，则可以用程序对文章的标题，内容，时间进行自动提取，该功能可以对大多数的文章闻类的网页进行准确的识别。



7.集成 ocr 图片识别功能

该版本集成了 OCR 自动识别模块，用户在标签编辑数据处理中，使最终数据为图片地址，然后添加 ocr 识别，程序将自动调用 OCR 配置对图片进行下载识别，并显示最终结果。



8.入库支持存储过程,Oracle 数据长度不限

MySQL,SqlServer 支持存储过程，用户可以使用更高级的方式对入库进行控制

access.dbm - 数据库模块编辑器

SQL语句: SQL语句最后不带分号, 多条语句之间请用回车换行分隔

Insert into data ([[title]]) values ('[标签:标题]')

系统标签: [\[标签:id\]](#) [\[采集页网址\]](#) [\[数据表前缀\]](#) [\[文章编号:表名XXX\]](#)

常用标签: [\[标签:XXX\]](#) [\[标签:标题\]](#) [\[标签:内容\]](#) [\[标签:作者\]](#) [\[标签:出处\]](#) [\[标签:时间\]](#)

时间转化: [\[时间转换:时间, yyyy-MM-dd\]](#) [\[系统时间转换: yyyy-MM-dd\]](#) [\[系统时间戳\]](#) [\[系统时间戳:时间\]](#)

数据库类型: SqlServer ☐ 是存储过程

使用说明: ddsdd

加载模块 新建/重置 保存模块

9.MongoDB 数据库支持

在扩展菜单下，火车采集器，更改数据保存数据库，就可以看到 Mongoddb 的配置界面

选择数据保存方式

选择数据库: MongoDb 当前数据库为Sqlite

服务器:

数据库路径: 浏览...

数据库列表: 获取列表

测试链接

开始转换

该数据库可以支持海量的数据采集，速度很快。

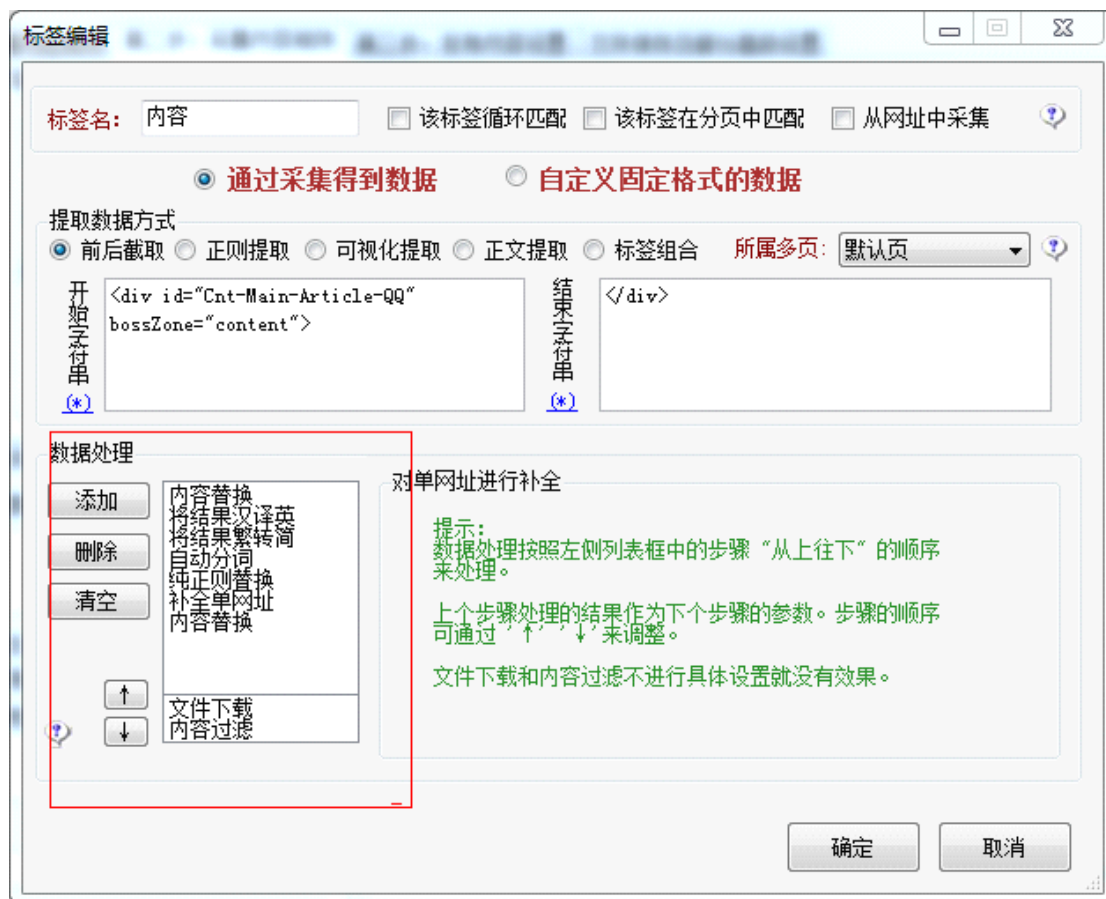
10.增加 Http 二级代理服务器，代理更方便



用户可以设置一个或多个二级代理服务器，在代理服务器中可以再设置更多的一级代理，从而达到单个任务使用多个代理的效果。

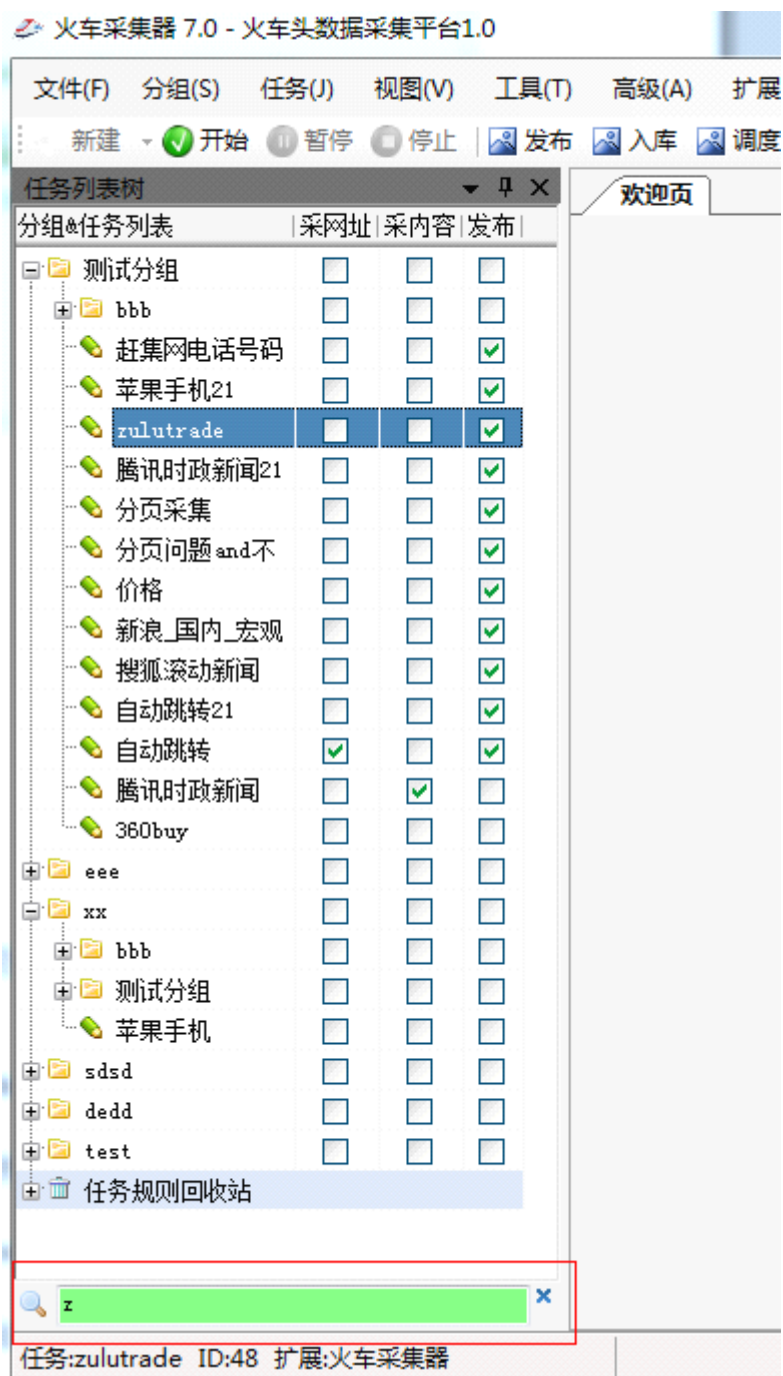
11.规则编辑中数据处理次序可随意调整随意添加

新版将原来的多个数据处理全部整合到一块，可以不限次序，不限次数的进行调用。



12.任务列表树支持搜索，无限级分级，拖拽

任务列中去掉了以前的站点概念，代之以分组，每个分组下面还可以有分组，分组下可以有任务，是无限级分组。而且每个任务都可以调整次序，只要拖拽即可，十分的方便。对于有大量任务需要搜索时，按 Ctrl+F 或是右键菜单中选择搜索，就可以打开搜索界面，如图。只要搜索到符合结果的任务，就会自动定位到该任务，使用回车可以定位到符合条件的下一个任务。



13.Web 发布模块支持插件和数据构造

新版 Web 支持插件功能，方便用户控制发布流程。同时支持数据构造功能，数据可以在发布前再被处理和组合一次。同时 Web 发布支持 json,xml 等格式数据的发布。

WEB发布模块编辑器

[网站自动登录](#)
[获取栏目列表](#)
[网页随机值获取](#)
[内容发布参数](#)
[高级功能](#)
[模块说明/保护](#)

文件上传设置

标签名	表单名	起始数字

新建

修改

删除

标签名

表单名

起始数字

递增数字

保存

发布数据构造

引用名	操作	值	参数

新建

修改

删除

操作类型

UrlEncode

数据值

参数

保存

插件设置

浏览

删除

加载模块

新建/重置

系统名称:

版本号:

保存模块

新建模块

说明: 所有的地址不带CMS系统安装路径, 并以反斜杠/开头

14.Web 发布支持多站点乱序发布

[第一步: 采集网址规则](#)
[第二步: 采集内容规则](#)
[第三步: 发布内容设置](#)
[文](#)

方式一: Web在线发布到网站

☒ 启用
 [→修改Web在线发布设置](#)

网站发布配置	分类名称	版块/分类ID
dede57	—cccc	2

添加发布配置

修改发布栏目

删除发布配置

发布方式

☐ 正序发布
 ☒ 倒序发布
 ☐ 乱序发布
 ☐ 多网站乱序发布

该功能更类似于一个简单的站群软件。

15.起始网址支持等差等比数列表网址格式，支持 rss 网址提取功能

添加开始采集地址

单条网址

批量/多页

文本导入

Rss地址

地址格式: (*)

☒ 等差数列

首项 1

项数 5

公差 1

☐ 补零

☐ 倒序

☐ 等比数列

首项 1

项数 5

公比 2

☐ 补零

☐ 倒序

☐ 字母变化

从 a

到 z

(区分大小写)

☐ 倒序

添加

预览

全部地址 (从上面多种方式添加，一次性加入起始地址，编辑请在上面右击)

完成

添加开始采集地址

单条网址

批量/多页

文本导入

Rss地址

Rss地址: 测试 添加

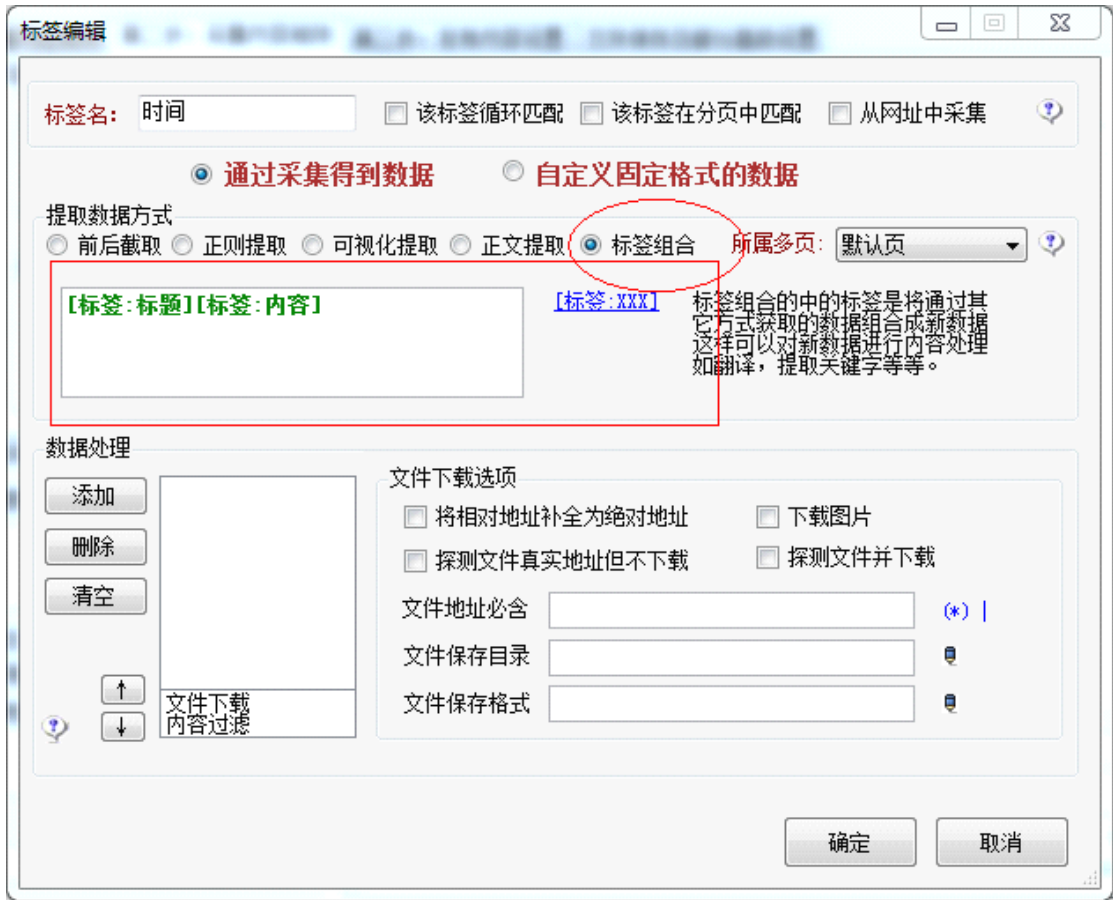
http://

预览

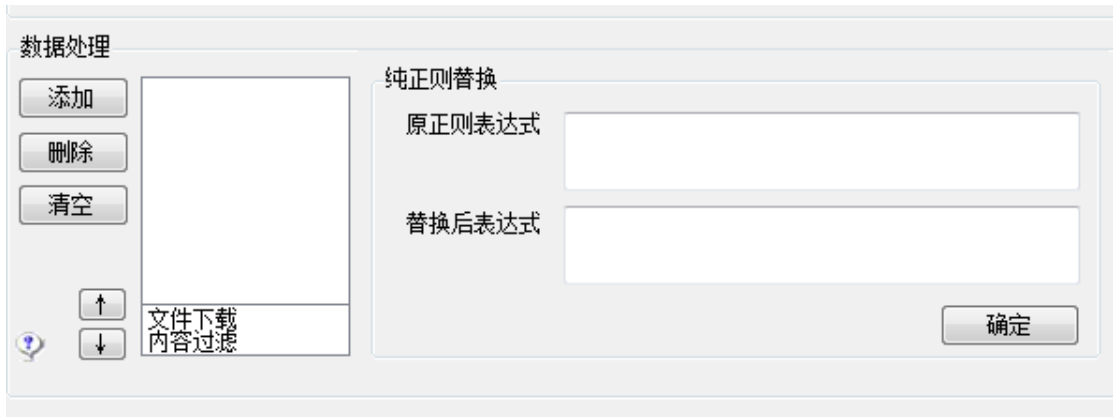
全部地址 (从上面多种方式添加，一次性加入起始地址，编辑请在上面右击)

完成

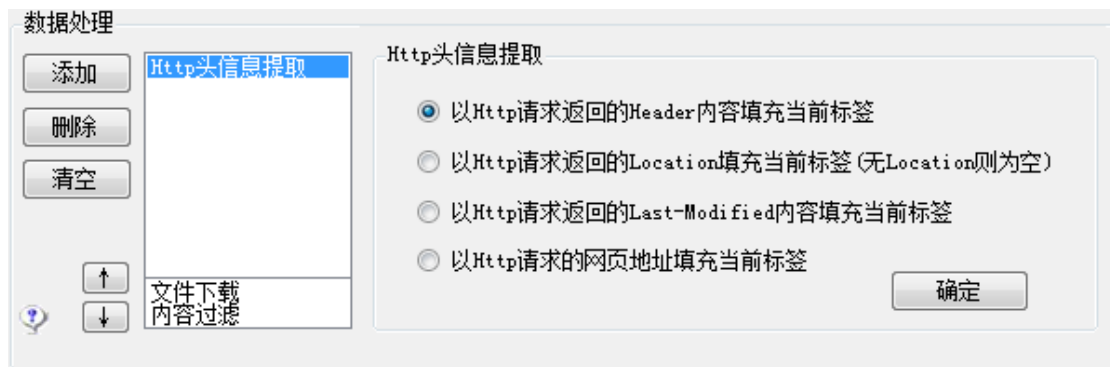
17.规则支持标签组合功能，新建的某个标签可以用其它的标签合并生成新数据。



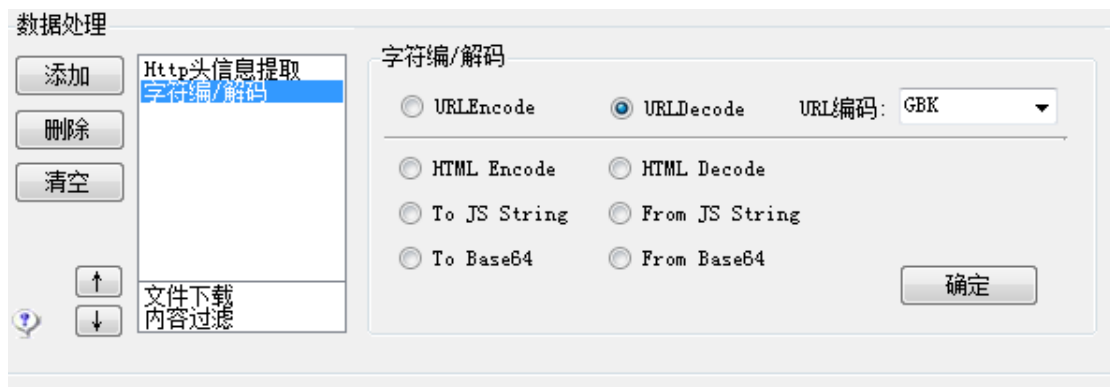
18.规则的数据处理中支持纯正则替换



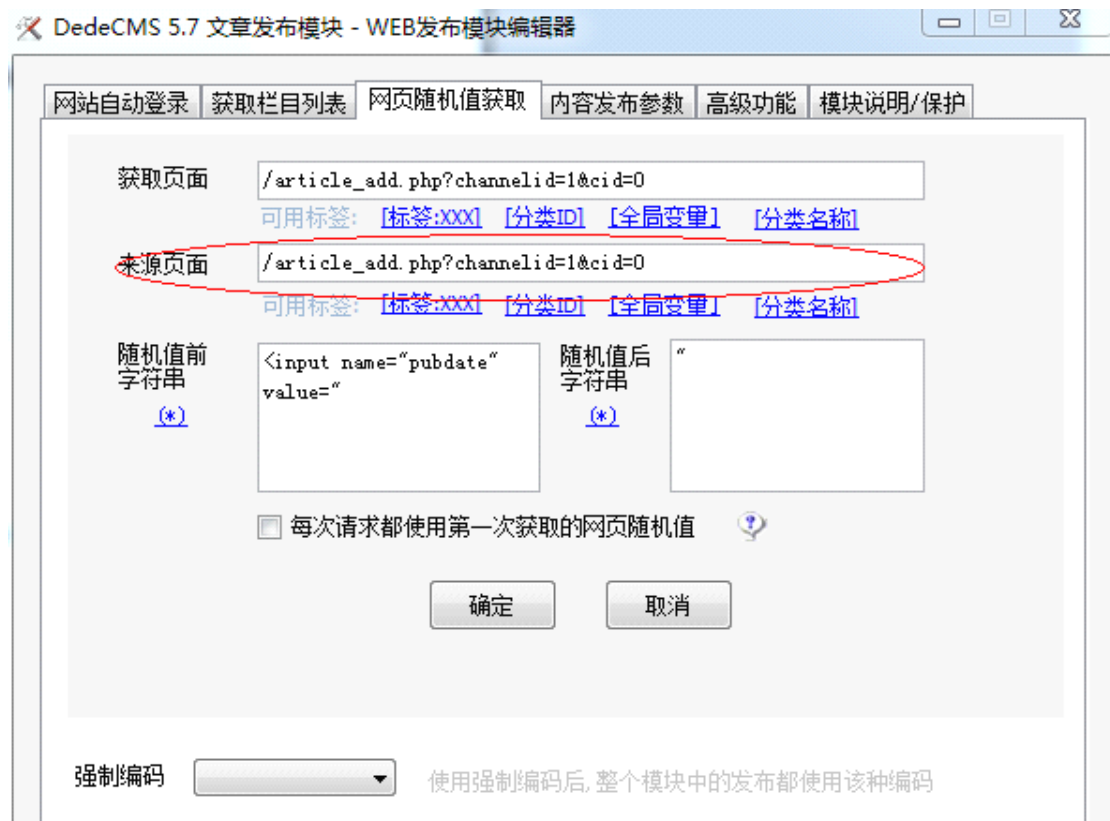
19.支持提取 http 头信息中的数据



20.支持对采集的数据进行编码，如 urlencode,htmlencode,to js String



21.发布模块中，网页随机值增加来源页



22.发布模块中，使用表单名，表单值的方式设置发布的数据

DedeCMS 5.7 文章发布模块 - WEB发布模块编辑器

网站自动登录 获取栏目列表 网页随机值获取 内容发布参数 高级功能 模块说明/保护

发表地址后缀: /article_add.php 可用标签

来源页面后缀: /article_add.php 可用标签

Post数据内容

自动抓取发布数据包

粘贴抓包获取的数据

提取POST表单发布数据

表单名	表单值
channelid	1
dopost	save
title	[标签:标题]
shorttitle	
redirecturl	
tags	
weight	13
picname	
source	
writer	

新建表单项

修改表单项

删除表单项

清空表单项

发表错误标志码: 标题不能为空
您没有此权限
档案主栏目必须选择
请选择档案的主类别
请选择所属栏目

成功标志码: 成功发布文章

加载模块 新建/重置

系统名称: DedeCMS

版本号: 5.7

保存模块

状态: 编辑模块DedeCMS 5.7 文章发布模块.说明: 所有的地址不带CMS系统安装路径, 并以反斜杠/开头

23.ftp 支持主动模式上传文件

FTP同步文件上传

FTP服务器: 192.168.0.111 端口: 21

☒ 启用FTP 用户名: rq204 密码: *****

☐ 匿名

☐ 被动模式

文件上传根目录: /FileUpload/ (以/开头和结尾)

上传次序: ☐ 先发布数据 ☒ 先上传文件

☐ 文件上传成功后删除本地文件

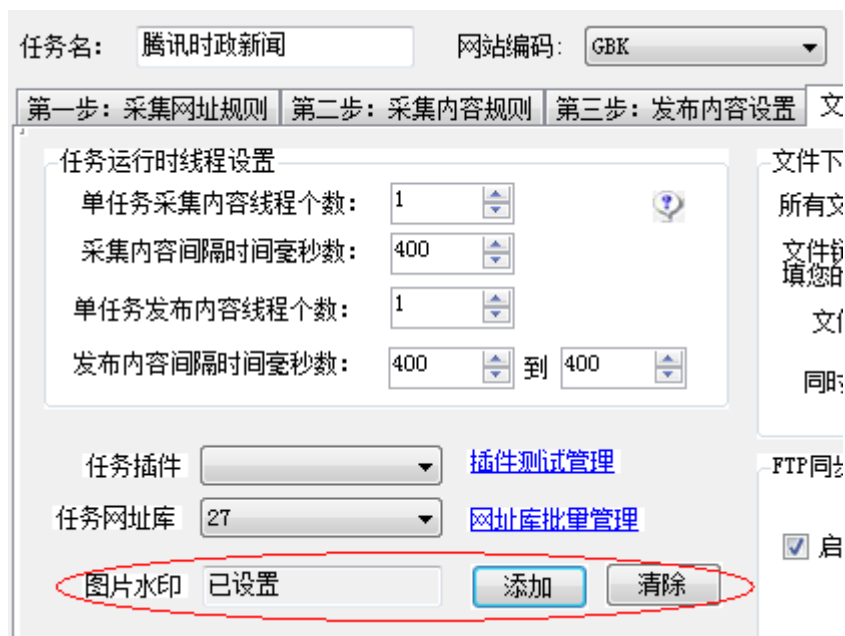
测试登录并创建文件夹 测试上传文件Test.zip

24.获取 cookie 功能增强，可以获取 discuz,phpwind 论坛的 cookie 信息。



如图中所示，以前版本是无法获取这个参数而导致网站不能登录，现在可以无误的抓取到这些信息了。

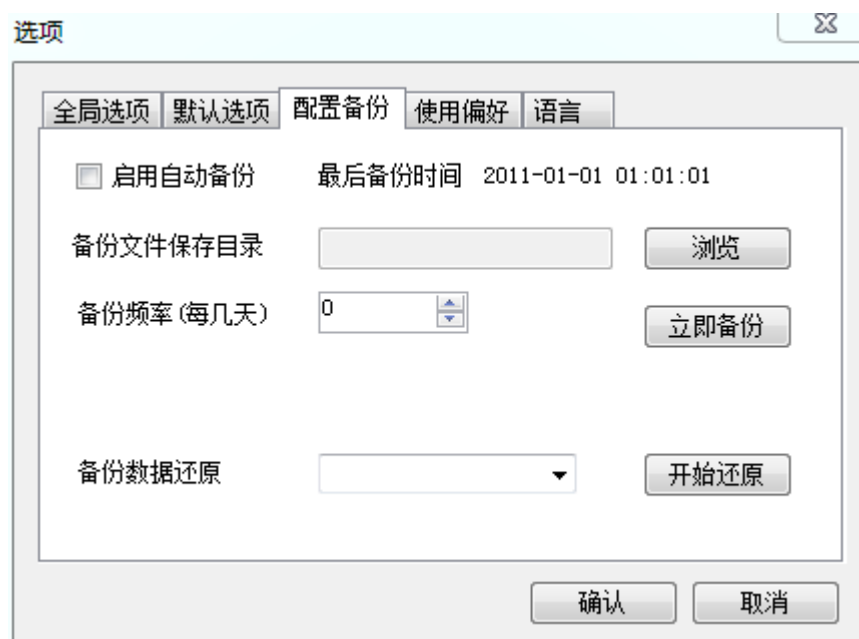
25. 图片水印



点击添加，可以看到



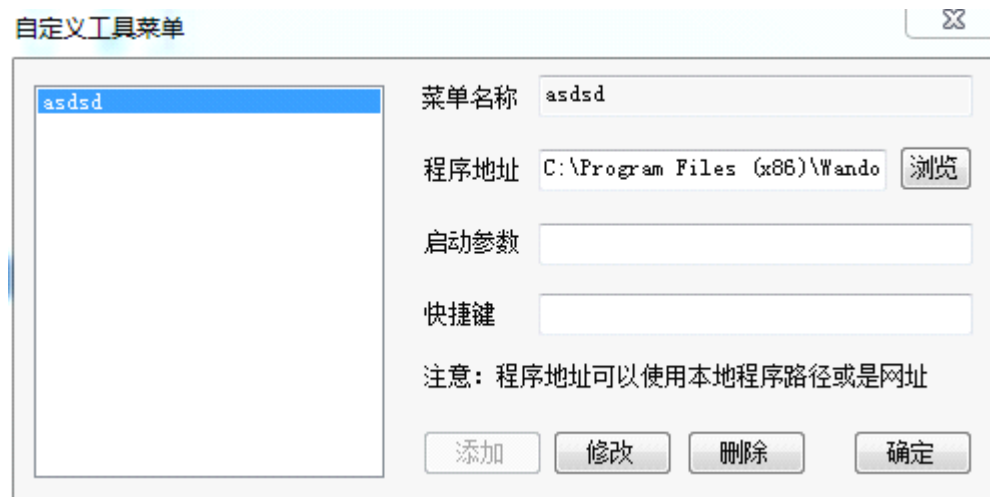
26.自动备份



采集器会自动将 Configuration 目录进行备份，防止误删除。

27.自定义菜单

在工具菜单中，选自定义菜单，则可以将自己常用的软件加到采集器菜单中



- 28.aspx 分页自识别
- 29.一键转载
- 30.其它用户体验和功能的改进。

合肥乐维信息技术有限公司
RQ204
2011-12-10