

# 小说规则捕捉器 V2.3.2



## 使用说明手册

2021 年 8 月 10 日

## 一、软件使用说明

功能简介

软件特色

软件界面

- 1、 主界面
- 2、 任务管理
- 3、 软件设置
- 4、 预设网站管理
- 5、 导入链接（任务导入）
- 6、 源代码查看
- 7、 向导、书籍搜索辅助工具
- 8、 章节重复过滤

操作步骤

- 1、 从预设网站中捕捉
- 2、 从浏览器中搜索书籍捕捉

## 二、名词定义

书籍页、目录页、内容页、章节分页：

预设网站：

入口网址，

捕捉类型：

规则：

捕捉任务：

HTML 代码：

源码分析的关键：

标签链接层数：

目录链接转换：

## 三、规则定制实例

实例一：多书籍规则定制

实例二：单书籍规则定制

实例三：使用正则表达式定制规则

实例四：章节分页捕捉实例

实例五：导入链接并定制规则（序号链接导入）

实例六：导入链接并定制规则（从文件导入）

# 一、软件使用说明

## 功能简介

本软件能通过小说网站的 html 网页源代码，分析关键信息的规则进行书籍捕捉，最终输出捕捉到的书籍（支持 txt、ePub、zip 格式输出）。

本软件即可说易用，也可说难用，如只是简单的从网站捕捉书籍，从自带的 100 多个预设网站中直接捕捉即可（需要通过浏览器查找要下的书籍，再把链接复制到入口网址处即可），不需要去分析复杂的源代码。针对逻辑思维能力较强的用户，可以通过分析小说网站的源代码，制定该网站的捕捉规则，基本能应对大部分小说网站。

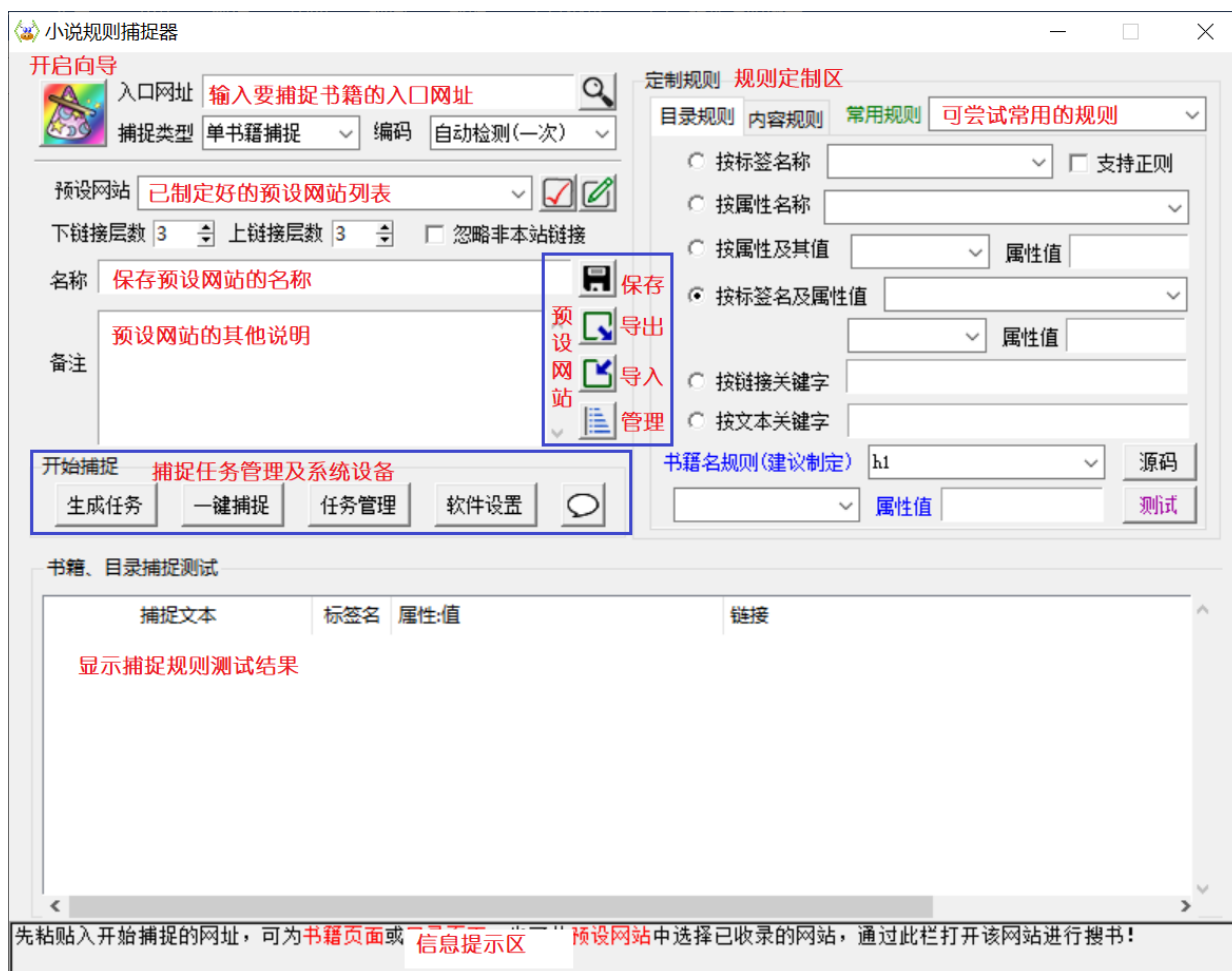
## 软件特色

- ✚ 自定义规则捕捉，能对大部分小说网进行文章捕捉，个别网站对书籍分类详细的，还支持多书籍捕捉；
- ✚ 自带大量预计网站，不会定义规则的用户可以直接套用，一样能捕捉到需要的小说；
- ✚ 提供常用规则选择，适用于大部分网站，方便快速定义规则；
- ✚ 支持多线程捕捉，能实现高速捕捉书籍；
- ✚ 自带源代码查看器，提供链接分析、关键定位、标签分段等工具；
- ✚ 自带操作向导提示功能，能提示下一步的操作；
- ✚ 捕捉前能对目标章节进行排序、删除、甚至是过滤重复等操作，让下载的书籍阅读起来更加顺畅；
- ✚ 针对大型小说，将任务暂存于数据库后可随意中断、恢复任务；

- ✚ 书籍提供多种输出方式：按章节文件、独立文本文件、压缩包、ePub 电子书籍等；
- ✚ 支持章节分页的捕捉，分页数量可以调整，支持动态分页情况（即分页数量不固定）；
- ✚ 支持任务导入，即从存有章节页面链接的文本文件、excel 文档中导入任务进行捕捉；
- ✚ 所有组件支持提示信息，即光标停置后会显示相关提示，大部分操作支持状态栏提示，让使用变得加容易；
- ✚ 支持预设网站的添加、修改、导入、导出、排序、删除；
- ✚ 提供书籍捕捉记录，可查看历史捕捉信息，方便重新捕捉；
- ✚ 附带小工具：ePub 电子书制作、分解工具，支持多以章节存放的书籍生成 ePub 文件，也可以将 ePub 文件分解为多章节的文本文件。

## 软件界面

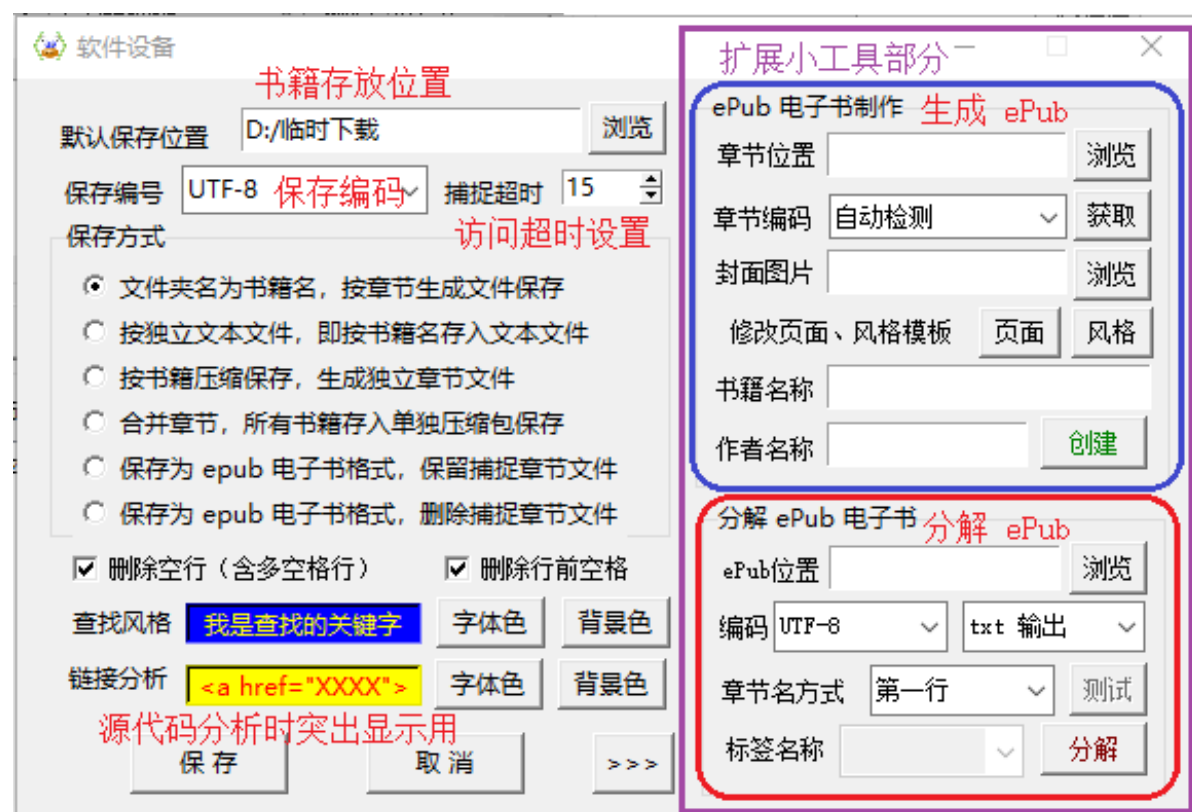
### 1、 主界面



## 2、任务管理



## 3、软件设置



#### 4、 预设网站管理



#### 5、 导入链接 (任务导入)

序号链接导入

从文件导入

链接导入方式

网址前段

开始

01

网址后段

结束

100

书籍名称

网页编码

自动检测

网址类型

目录级别

目录页规则

内容页规则

规则设定

标签名称

支持正则

属性名称

属性值

链接匹配

文本匹配

导入的链接检测

规则设定的测试

链接检测

规则测试

确定导入

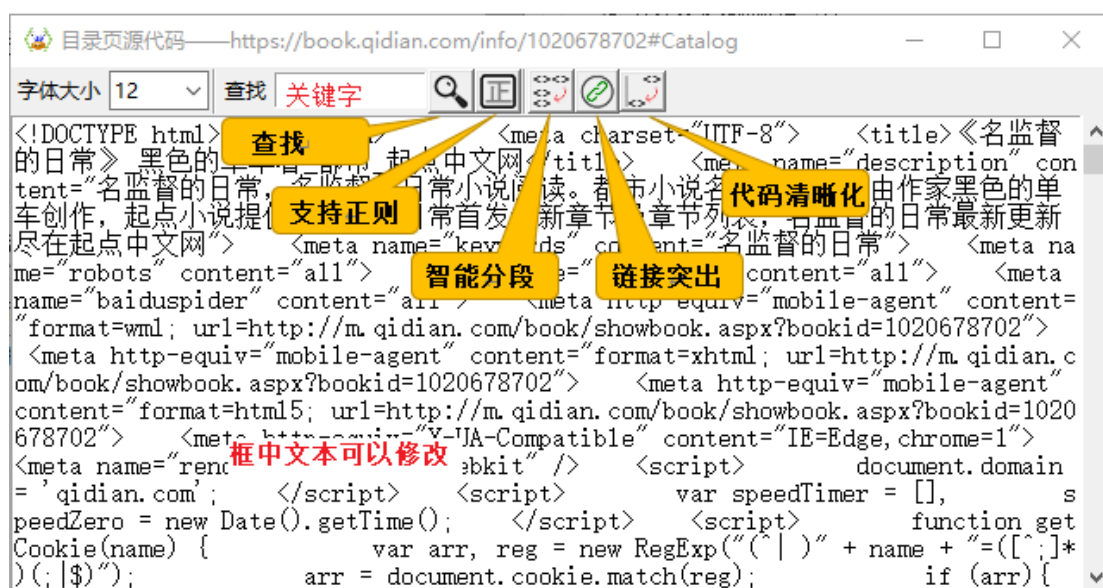
取消导入

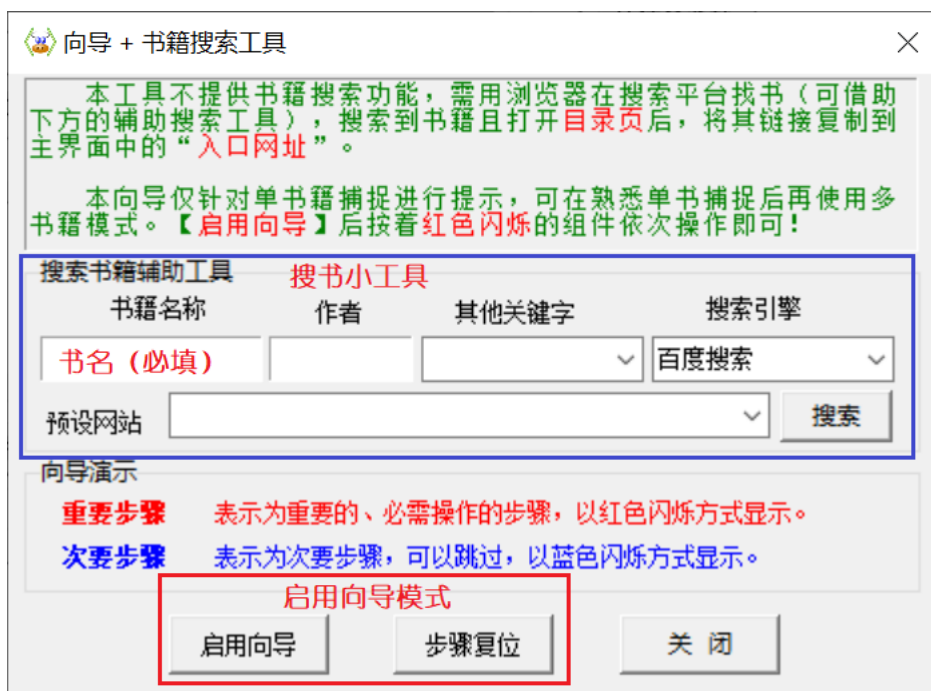
测试结果

请按下列步骤测试完规则后再进行导入：

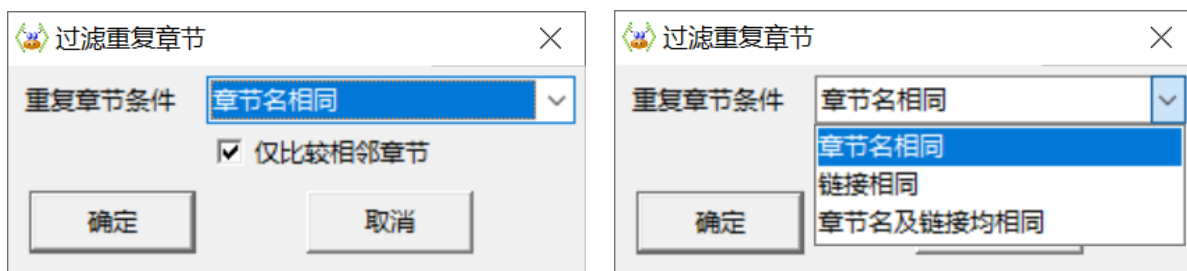
- 1、导入链接，可指定网址序号或从文件中导入。
  - 1.1、网址序号链接：网址前段 + 序号 + 网址后段；  
序号需要指定开始、结束两个数据；  
开始序号填“01”时表示从1开，不足位数则补0；  
不需要补0时直接填写“1”，也可写其他字符。
  - 1.2、文件导入支持Excel和文本文件格式；  
Excel数据需要在第一张工作表以行的方式添加。
- 2、按【链接检测】查看有效是否有效。
- 3、制定目录规则，按【规则测试】查看制定是否有效；  
能匹配到有效结果后，会自动跳到内容规则制定页；  
如导入的链接为内容页时，跳过本步骤。
- 4、制定内容规则，按【规则测试】查看制定是否有效。  
成功捕捉到章节内容后，可按【确定导入】开始导入。

## 6、源代码查看





## 8、章节重复过滤




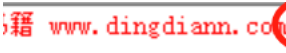

### 操作步骤

捕捉小说一般按分两种方式开始，一种是[预设网站中搜索小说](#)（需要浏览器器配合），找到小说的章节目录页后，把其网址复制到软件的“入口网址”中进生成捕捉任务，此方法不需要制定复杂的规则，适用大多数人。

还有一种方式是在[浏览器中搜索](#)到想要捕捉的小说，将其章节目录页网址复制到软件的“入口网址”后，再查询是否在预设网站内，如是则套用后可生成捕捉任务，如没有，则需要手动分析源代码制定规则（可随意将该网站添加为预设网站），此方法较为为复杂，仅适用于电脑高手。



## 1、 从预设网站中捕捉


先从“**预设网站**”中选择某一小说网（根据个人喜好选择），然后用右侧的  按键套用，在状态栏中点   套 按 键，软件将会在浏览器中打开该网站，在搜索栏中输入要找的小说名，如下图所示（仅供参考）：



点击找到小说进入到**目录页**（即显示所有章节的页面），再将目录页的网址复制到软件的“**入口网址**”中，即可开始捕捉。

建设先点“目录规则”框中的【**测试**】按键查看章节顺序是否正确（许多小说网站会把最后更新的一些章节添加到前面，不处理会影响阅读效果），如章节需要调整的，按【**生成任务**】按键，在生成任务列表后可对章节进行调整，最后再进行捕捉输出。

## 2、 从浏览器中搜索书籍捕捉

对于一些较新的小说书籍，预设网站可能会收录不及时，这时可以在百度中直接搜索书名，查找合适的网站进行捕捉。查到合适的小说**目录页**后，把目录页的网址复制到软件的“**入口网址**”中，再按右侧的  查询该网站是否在预设网站库中，查询到的套用后可直接**生成任务**进行捕捉，

查询不到的就需要手动**制定规则**进行捕捉，规则制定可查看[“规则制定实例”](#)。制定完规则后可生成任务再捕捉。

## 二、名词定义

### 书籍页、目录页、内容页、章节分页：

**书籍页**：指有多个书籍链接的页面，一般指书架网页；

**目录页**：有多个章节链接的页面，也称章节列表页；

**内容页**：小说最终显示内容的页面。

**章节分布**：个别网站对最终章节内容进行分页显示，即内容下方会出现类似“**下一页**”的导航按键，可进行简单的分页设定即可完整捕捉章节内容。

### 预设网站：

事先将部分网站的规则制定好并存以数据库中，可直接套用，也可进行添加、修改、删除。

### 入口网址，

即捕捉的起始网址，一般为**书籍列表页**或**书籍的目录页**。

### 捕捉类型：

区分于**多书籍捕捉**还是**单书籍捕捉**。

### 规则：

按 HTML 代码分析定位到相关的标签，再从匹配的标签附近捕捉**链接**（即捕捉到书籍目录、章节内容页的链接），或者捕捉到**文本内容**。较高级的规则设定需要掌握正则表达式（类似通配符），正则表达式非常适用于文字处理，建议大家学习掌握。

### 捕捉任务：

先生成捕获列表，因个别网站书籍的目录会有些错乱（常见为最后更新的章节会放在前面），需要手动调整，调整完后才开始捕捉。

如网站目录正常排序时可直接使用【**一键捕捉**】生成列表后自动进入捕捉状态。

如书籍较多或章节较多的，捕捉时间较长，为了防止意外出错导致捕捉中断，建议捕捉时按【保存并捕捉】先将捕捉任务存入数据库，在出现意外中断时能快速恢复捕捉（未存入数据库也能恢复捕捉，但需要输入入口网址，并花费些时间生产捕捉列表）。

## HTML 代码：

本软件没有直接提到 HTML 代码，但分析规则时需要用到，下面简单脑补一下 HTML 代码：

基本标签的形式<标签名 属性名 1=属性值 1 属性名 2=属性值 2>文本</标签名>

标签中的属性可为多个或无，</标签名>表示为该标签结束，标签或多层嵌套，如：

```
<div><a href="XXXX"><b>文本信息</b></a></div>
```

其中<b>标签为<div>标签的子标签（这里表示为第三层），而<div>标签为<a>标签的父标签（这里表示为第二层）

<a href="XXXX">为链接标签，其中的 XXXX 即为相关的链接，在书籍页、目录页中我们就是需要捕捉到这个链接，而内容页则需要捕捉到的是**文本信息**。

## 源码分析的关键：

捕捉以 HTML 标签特点进行区分，从网页中相关的书籍、章节链接附近标签中找到

1. 书籍页：要捕捉跳转到目录页的链接，即书籍目录页的链接；
2. 目录页：要捕捉跳转到内容页的链接，即章节内容页的链接；
3. 内容页：要捕捉章节的文本内容。

## 标签链接层数（可无视）：

规则制定时允许**匹配标签**与**链接标签**（即

表示为：向下链接层数（匹配标签与子标签之间的层数）、向上链接层数（匹配标签与父标签之间的层数）如：

### 向下链接层数

```
<li class=book><div ...><a href=xxxx>书名 1</a></div></li>
```

在制定规则时可为：标签名=li，属性名=class，属性值=book

<li class=book>为匹配标签，是第一层

<div>标签是第二层子标签，不含链接

<a>标签是第三层子标签，带链接

因此同下容错层数必须为 3 或更大时才能正确捕捉到链接。

### 向上链接层数

```
<a href=xxxx><li class=book>书名 1</li></a>
```

在制定规则时可为：标签名=li，属性名=class，属性值=book

<li class=book>为匹配标签，是第一层

<a>标签是第二层父标签，带链接

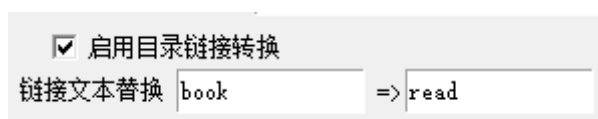
因此同上容错层数必须为 2 或更大时才能正确捕捉到链接。

## 目录链接转换：

部分网站在书籍页链接没有直接打开到目录页，而是到书籍简介页，需要再打开一链接才能转到目录页，幸好网站的这种链接关系是固定的，因此可以通过链接转换（替换）的方式进行指定正确的目录页。如：

书籍页的捕捉到的原链接为：/book/xxxxx.html，而目录页为

/read/xxxxx.html，这时我们可以按下图设定目录链接转换设定：



The image shows a configuration window for link conversion. At the top, there is a checkbox labeled '启用目录链接转换' (Enable directory link conversion) which is checked. Below this, there is a section labeled '链接文本替换' (Link text replacement). It contains two input fields: the first field contains the text 'book' and the second field contains the text 'read'. Between these two fields is a right-pointing arrow with an equals sign (=) indicating a replacement operation.

## 三、规则定制实例

### 实例一：多书籍规则定制

网站页面如下图，为多本书籍集中在同一页面中：


当前位置: 主页 > 图书读物 > 伍美珍小说

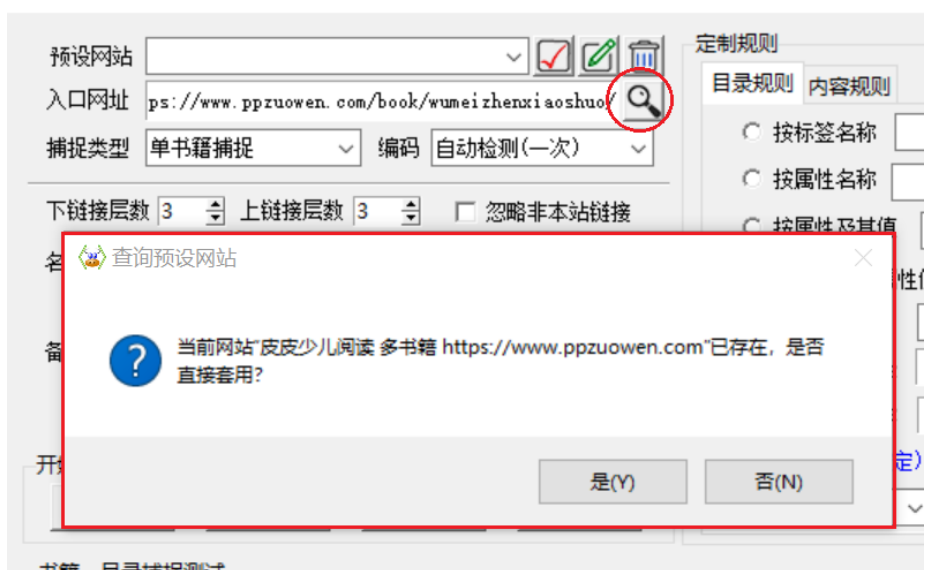
+ 分享到:  微信  QQ空间  新浪微博  腾讯微博  百度贴吧



把上面网页地址复制到软件的“入口网址”处:

<https://www.ppzuowen.com/book/wumeizhenxiaoshuo/> , 我们可以先按右侧的

【 查询】按键, 查询该网站是否已存在预设网站中, 如已存在, 可直接套用并开始捕捉了。



现以上网址做为实例，假设上面的网站未录入预设网站中，我们对其 HTML 源代码进行分析，并制定捕捉规则：

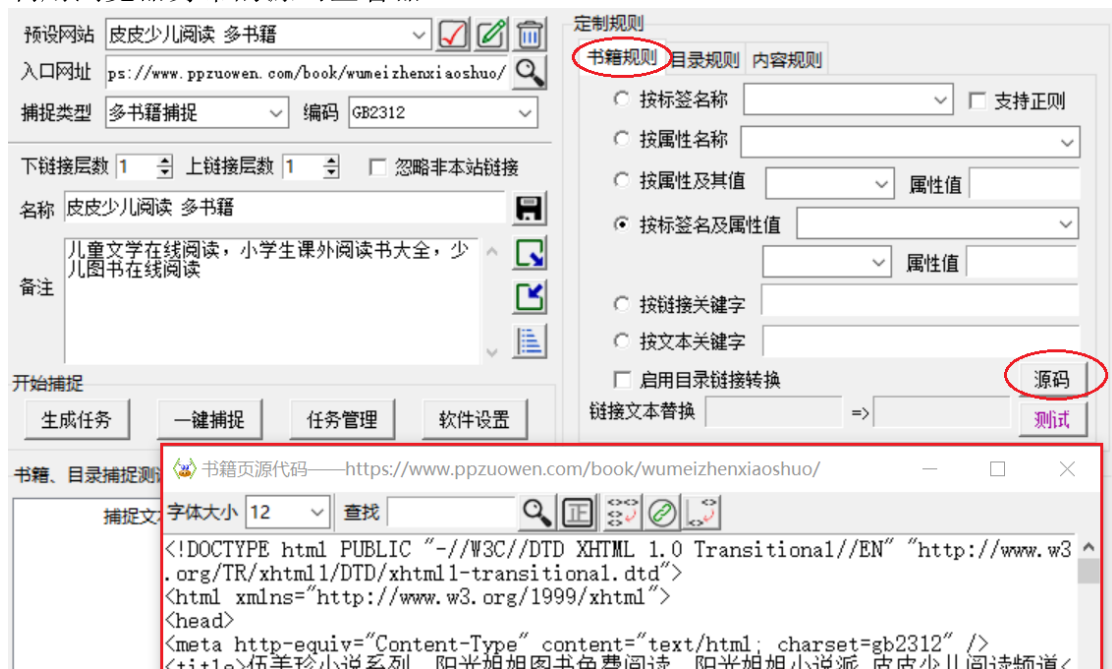
### 一、确定捕捉级别

因我们需要从书籍页开始捕捉，即同时捕捉多本书籍，因此在【捕捉类型】处选“多书籍捕捉”，如入口网址是一本书籍的目录页，则要选为“单书籍捕捉”。

### 二、分析页面源代码

本软件为规则制定软件，需要对 HTML 代码进行简单分析，找出书籍、目录、内容三种页面的规律，从而定制捕捉规则。

在书籍、目录、内容规则制定中，在右下角均有【源码】按键，可直接查看相关网页的源代码，也提供了一些便捷分析的小功能（详见后面）（当然也可能利用浏览器身带的源码查看器）



### 三、分析页面源代码，制定规则



这步需要有一定的 HTML 知识，后面附一些简单的教程，看过后就能顺利使用本软件了。

源码分析的关键点为：

1. 书籍页：要捕捉跳转到目录页的链接，即书籍目录页的链接；
2. 目录页：要捕捉跳转到内容页的链接，即章节内容页的链接；
3. 内容页：要捕捉章节的文本内容。

### 1、书籍页规则制定

我们先分析书籍页的规则，通过源码我们可以找到下面的代码：

```
<a href="/book/shanshanruorenai/" class="text" target="_blank" >闪闪惹人爱</a>
```



在浏览器中打开网址，在其中一本目标书籍（章节）上右键复制链接，再到源代码窗口的查找中粘贴入链接，注意得把链接前端的网站域名删除（源代码链接到本站是不需要前缀的，即把 “<https://www.ppzuowen.com/>” 文字删除）

通过上面分析，我们可以找到关键的规则：`class="text"`，我们可以试着在查找中输入 `class="text"`，看看是否能全部将目录链接全关联到。

关闭源代码窗口回到主界面，在“书籍规则”中做下图中的设定：

定制规则

书籍规则 目录规则 内容规则

☐ 按标签名称 ☐ 支持正则

☐ 按属性名称

☐ 按属性及其值 属性值

☒ 按标签名及属性值 a class 属性值 text

☐ 按链接关键字

☐ 按文本关键字

☐ 启用目录链接转换

链接文本替换 =>

源码 测试

按【测试】查看捕捉效果



捕捉文本	标签名	属性:值	链接
闪闪惹人爱	a	class:['text']、target:_blank	http://www.ppzuowen.com/book/shanshanruorer
在你鼻尖跳舞	a	class:['text']、target:_blank	http://www.ppzuowen.com/book/zainibijiantiaowu
我班流行写小说	a	class:['text']、target:_blank	http://www.ppzuowen.com/book/wobanliuxingxie
男生不许进	a	class:['text']、target:_blank	https://www.ppzuowen.com/book/nanshengbuxuji
我们班的狗仔队	a	class:['text']、target:_blank	https://www.ppzuowen.com/book/womenbandeg
惜城灵魂出窍记	a	class:['text']、target:_blank	https://www.ppzuowen.com/book/xichenglinghunc
青蛙王子副班长	a	class:['text']、target:_blank	https://www.ppzuowen.com/book/qingwawangzifu
我是便利贴女生	a	class:['text']、target:_blank	https://www.ppzuowen.com/book/woshibianlitenv

共捕捉到 41 条信息，如均为有用信息，可用右键选择其中一条作为**目录页的入口网址**，如仍有多余信息，建议重新定制规则！

捕捉正确后，选中其中一个目录页点鼠标右键，按“**引用到下一级页面**”，即将其链接做为目录页的网址，进行目录页的规则制定。

## 2、目录页规则制定

同书籍页的一样，先查看、分析目录页源代码，可找到下面代码：

```
<a href="/book/zainibijiantiaowu/160848.html" class="title"
target="_blank">第一章 人群中的孤独感</a>
```

在书籍级别的捕捉中，不需要定制书籍名称的规则，但在目录级别的捕捉中，为了能捕捉到正确的书名，建议定制该规则，方便以后捕捉用，如不保存为预设网站，也可不定制此规则。规则设定见下图：

定制规则

书籍规则

目录规则

内容规则

☐ 按标签名称
☐ 按属性名称
☐ 按属性及其值
☒ 按标签名及属性值

☐ 按链接关键字
☐ 按文本关键字

书籍名规则(建议制定)

书籍页面捕捉类型时不设此规则

查看测试效果：

书籍、目录捕捉测试

选中某一章节后，用鼠标右键把链接设成内容页的入口网址

捕捉文本	标签名	属性:值	链接
第一章人群中的孤独感	a	class:['title']	https://www.ppzuowen.com/book/zainibijiantiaow
同桌黄瓜很了解邱佳	a	class:['title']	https://www.ppzuowen.com/book/zainibijiantiaow
黄瓜的唠叨和牙痛	a	class:['title']	https://www.ppzuowen.com/book/zainibijiantiaow
人群中的孤独感	a	class:['title'], target_blank	https://www.ppzuowen.com/book/zainibijiantiaow
酷儿的情报和请求	a	class:['title'], target_blank	https://www.ppzuowen.com/book/zainibijiantiaow
第二章麻麻脸的新同桌	a	class:['title'], target_blank	https://www.ppzuowen.com/book/zainibijiantiaow
邱佳开始反悔了	a	class:['title'], target_blank	https://www.ppzuowen.com/book/zainibijiantiaow
原来酷儿是下了一个套	a	class:['title'], target_blank	https://www.ppzuowen.com/book/zainibijiantiaow

共捕捉到 15 条信息，如均为有用信息，可用右键选择其中一条作为内容页的入口网址，如仍有多余信息，建议重新定制规则！

查看是否显示为章节信息，链接是否齐全，如无多余信息，可选中其中一个章节页用右键引用成下一级网址。

### 3、内容页规则制定

同前面页的一样，先查看、分析内容页源代码，可找到下面代码，可以看到比较关键的几处标签：

```

</div>
<div class="warp960 mt10 fix">
  <div class="articleBody articleContent1">
    <h2 class="articleH2">第一章 人群中的孤独感</h2>
    <div class="articleContent"><p>第一章 人群中的孤独感</p>
    <p>女生们喜欢的话题</p>
    <p>“哎呀，真的好好耶……我好喜欢哦！”邱佳刚进教室，就听见班
    一冉小渝和酷儿在发出尖叫。</p>
    <p>她们总是那么热切地谈论着自己的话题。</p>
    <p>其实，冉小渝和酷儿也不是任何时候都这样神采飞扬的，只有当她
  
```

从源码中可以看出下面较关键的部分代码：

```
<div class="articleBody articleContent1">
```

```
<h2 class="articleH2">第一章 人群中的孤独感</h2>
```

```
<div class="articleContent"><p>第一章 人群中的孤独感</p>
```

分析：

第一行代码<div>标签，可先制定为：标签名=div，属性名=class，

属性值=articleBody articleContent1

测试发现捕捉时多了一行章节名，不理想。

第二行代码<h2>标签，但在章节名后面就结束了（</h2>），实际捕捉结果为：

第一章 人群中的孤独感

第三行代码<div>标签，可先制定为：标签名=div，属性名=class，

属性值=articleContent

为较理想的结果，因此可将规则设以上述值，规则设定见下图：

定制规则

书籍规则 目录规则 内容规则

☐ 按标签名称 ☐ 支持正则

☐ 按属性名称

☐ 按属性及其值 属性值

☒ 按标签名及属性值 属性值 属性值 icleContent1

☐ 按链接关键字

☐ 按文本关键字

章节名称: 第一章 冉小渝、酷儿和黄瓜  
当前网址: <https://www.ppruowen.com/book/shanshanruorenai/161554.html>

源码 测试

测试结果与书籍、目录时不一样：

内容捕捉测试

第一章 冉小渝、酷儿和黄瓜第一章 冉小渝、酷儿和黄瓜

第一次月考成绩册

任老师说，周末要开家长会，大家把这次月考的成绩册拿回去给家长签字，请家长们一定要在百忙中

长会。

“这是你们初一年级的第一次家长会，每个家长都要到！”

任老师一再强调说。

胖嘟嘟的椰子头女生邱佳坐在教室的倒数第二排的角落，她抿着嘴唇，用胖胖的小手紧紧地按住成绩

着一颗定时炸弹，她不敢打开它。

再看看周围的人，大家都在紧张地翻着成绩册子，嘴里不停地发出各种感叹的声音：“哎呀，我考得

不到我排在全年级30名！幸哉幸哉！！”……

这是一个全年级的排名册，上面的姓名啊、班级啊、分数啊、单科成绩啊、总评分啊、名次啊，等等

邱佳似乎听见有人在说，成绩差的排在成绩册的最后一页，成绩好的在前面。

她似乎下了决心，她先是偷偷地看了周围几眼，感到没人注意她，便悄悄地折起成绩册的一角，翻到

邱佳的同桌黄天宇是个瘦瘦的男生，他有着一双亮亮的眼睛，邱佳的一举一动都被他尽收眼里了。

“嘿，皮卡丘，你……”黄天宇歪过头来，和邱佳说话，却被邱佳警觉地抬头警告：“黄瓜，不许你

邱佳重新把她的成绩册捂紧，一脸气愤地瞪着黄天宇。

如以后还准备在该网站捕捉其他书籍，可将此网站与规则设定保存为预设

网站，先填写预设网站的名称（建议用域名+中文名，方便调用时查看）后保

存了：

入口网址

捕捉类型  编码

下链接层数  上链接层数  ☐ 忽略非本站链接

名称

备注

其他信息指：捕捉类型、网站域名

最后按【生成任务】开始捕捉书籍，如测试目录规则时发现章节顺序均正确也可以按【一键捕捉】直接开始捕捉。

开始捕捉

## 实例二：单书籍规则定制

本例选用较常见的笔趣阁网站进行分析（网站笔趣阁网站非常多，就本软件默认收录的就多达 30 个），这类网站不建议使用多书籍捕捉模式，选用“**单书籍捕捉**”即可，直接在浏览器中搜索到相捕捉的书籍后，打开到该书籍的目录页，如下图：

笔趣阁 > 牧神记最新章节列表 加入书架 | 直达底部



**牧神记**

作者：宅猪 分类：玄幻 状态：连载 字数：3360271

更新时间：2019-12-15 11:16:37 最新章节：宅猪新书，《临渊行》已经上传

简介：大墟的祖训说，天黑，别出门。大墟残老村的老弱病残们从江边捡到了一个婴儿，取名秦牧，含辛茹苦将他养大。这一天夜幕降临，黑暗笼罩大墟，秦牧走出了家门.....做个春风中荡漾的反派吧！瞎子对他说。秦牧的反派之路，正在崛起！书友群：600290060，624672265，VIP群：663057414（有验证）普通群：424940671

推荐阅读：伏天氏、我真不是欧皇、后手、修真原理、牧神记、戏闹初唐、我的刀剑世界、西游大妖王、空间悍女：种田吧，王爷！、重生之傲娇军嫂

《牧神记》正文

第一章 天黑别出门	第二章 四灵血	第三章 神通
第四章 天魔造化功	第五章 漓江五老	第六章 小不点儿，死
第七章 灵胎壁	第八章 婆婆的皮囊	第九章 红粉骷髅
第十章 黑暗侵袭	第十一章 破壁	第十二章 无双战技
第十三章 敲死	第十四章 元气淬体	第十五章 水上行走
第十六章 庙中幼女	第十七章 灵胎破壁	第十八章 坏孩子
第十九章 霸体觉醒	第二十章 人形灵胎	第二十一章 药力催人猛

在入口网站中粘贴入网址：<https://www.bqg99.cc/book/2639610/>  
先查看网页的源代码，见下图

定制规则

目录规则 内容规则

☐ 按标签名称
☐ 按属性名称
☐ 按属性及其值
☒ 按标签名及属性值
☐ 按链接关键字
☐ 按文本关键字

☐ 支持正则

书籍名规则(建议制定)

属性值

源码

测试

也可以用浏览器直接查看源代码，部分网站会屏蔽右键，可点击状态栏中的“复制”键，再到浏览器的地址栏中粘贴后回车，可能直接查看源代码。


入口网址: <https://www.bqg99.cc/book/2639610/>，按  键复制查看源代码链接，再到浏览器地址栏中粘贴也可以查看网页源代码！

通过分析，可以找到一些关键的标签，如下图，蓝色的<h1>标签表示为书名（单书籍捕捉时最好制定书籍名称规则，这样导出的书籍也会按书名来命名，方便管理）；红色部分表示为章节部分的代码，每个章节的标签为<dd>。

目录页源代码——<https://www.bqg99.cc/book/2639610/>

字体大小 12 查找

```

sp,|enbsp,embsp,\a rel="nofollow" href="#"> 键复制查看源代码链接，再到浏览器地址栏中粘贴也可以查看网页源代码！
<div class="info">
<div class="cover"></div>
<h1>牧神记</h1>
<div class="small"><span>作者: 宅猪</span>
<span>分类: 玄幻</span>
.....
|
.....
</div>
<div class="listmain">
<dl>
<dt>《牧神记》正文</dt>
<dd><a href="https://www.bqg99.cc/book/2639610/637338569.html">第一章 天黑别出门</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637338511.html">第二章 四灵血</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637244286.html">第三章 神通</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637244185.html">第四章 天魔造化功</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637132595.html">第五章 漓江五老</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637131610.html">第六章 小不点儿，死</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637035103.html">第七章 灵胎壁</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/637035051.html">第八章 婆婆的皮囊</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/636906579.html">第九章 红粉骷髅</a></dd>
<dd><a href="https://www.bqg99.cc/book/2639610/636906474.html">第十章 黑暗侵袭</a></dd>

```

目录规则设定如下：

定制规则

目录规则 内容规则

☒ 按标签名称 dd ☐ 支持正则

☐ 按属性名称

☐ 按属性及其值 属性值

☐ 按标签名及属性值

☐ 按链接关键字

☐ 按文本关键字

书籍名规则(建议制定) h1

属性值

源码

测试

可以手动输入其他标签名, 可支

后在测试后发现存在其他无用的信息（即网页中有一些内容也使用<dd>标签的），按上例中我们也可以根据链接内容进行过滤，如上例中内容页的链接为 .../book/数字/数字.html

在测试结果中选任意一章节引用为内容页，如下图操作：

书籍、目录捕捉测试	捕捉文本	标签名	属性值	链接
	牧神记	h1		【书籍名称】
	第一章天黑别出门	dd		https://www.bqq99.cc/book/2639610/637338569.html
	第二章四灵血	dd		https://www.bqq99.cc/book/2639610/637338511.html
	第三章神通	dd		https://www.bqq99.cc/book/2639610/637244286.html
	第四章天魔造化功	dd		https://www.bqq99.cc/book/2639610/637244185.html
	第五章漓江五老	dd		https://www.bqq99.cc/book/2639610/637132595.html
	第六章小不点儿，死	dd		https://www.bqq99.cc/book/2639610/637131610.html

打开内容页的源代码，可以看到下面的信息：

```

</div>牧神记</a></span>
<span>作者：宅猪</span>
<span>字数：3212</span>
<span>更新时间：2017-06-20 20:00:00</span></div>
<div id="content" class="showtxt">
    司婆婆拉着他兴冲冲的往村里走，笑
    别看了，快点过来，今天是你的人日子！村长，马爷，都出来！<br /><br />
    里燃起了篝火，村长又被人用担架抬了出来，沉声道：“四灵都找到了？”<br />
    > “都找到了。”<br /><br /> 独臂的马爷拖来了一条几丈长的大蛇，
    也还活着，泛着腥气，只是被马爷单手捏住七寸，动弹不得。<br /><br />
    匠则捉来了一头大鸟，那头大鸟比哑巴还要高大一些，但是被绑住了翅膀和双脚
    鸟挣扎时，羽毛中竟然有火光飞出，噼里啪啦作响，很是吓人。<br /><br />
    则搬来了一只比桌子面还要大的巨龟，这头巨龟不知活了多少年，龟壳都翻起了：
    。巨龟四肢都缩在壳里，时不时偷偷的探出一只爪子，秦牧看到它的爪子探出壳
  
```

可以分析到关键标签为<div>，且其属性名、属性值可作为关键的规则，内容页的规则制定见下图：



测试后能捕捉到章节内容说明规则设定成功，可将本次规则定制保存成预设网站，以便下次能直接从本网站捕捉其他书籍。

最后按【生成任务】开始捕捉书籍，如测试目录规则时发现章节顺序均正确也可以按【一键捕捉】直接开始捕捉。

### 实例三：使用正则表达式定制规则

本例演示按正则表达式来匹配链接中的关键字，同样先设入口网址：<https://www.xuanshu.com/book/53183/>（选书网），分析目录页的源代码：

```
... <li><a href="19618048.html">楔子</a></li>
<li><a href="19618049.html">第一章 人称 789</a></li>
<li><a href="19618050.html">第二章 林不易</a></li>
<li><a href="19618051.html">第三章 翻车了</a></li>
<li><a href="19618052.html">第四章 风云台</a></li>
<li><a href="19618053.html">第五章 首次降临</a></li>
<li><a href="19618054.html">第六章 卢某人要会会他</a></li>
<li><a href="19618055.html">第七章 完美任务</a></li>
<li><a href="19618056.html">第八章 龙亦澜</a></li>
<li><a href="19618057.html">第九章 白衣</a></li>
<li><a href="19618058.html">第十章 再次第一</a></li>
<li><a href="19618059.html">第十一章 灵界巡弋者</a></li>
```

```
<li><a href="19618060.html">第十二章 无敌的猎敌之锋</a></li>
<li><a href="19626860.html">第十三章 破晓军团</a></li>
<li><a href="19633458.html">第十四章 白衣惊世</a></li>
<li><a href="19646042.html">第十五章 大丰收</a></li>
<li><a href="19657714.html">第十六章 猎妈之锋</a></li>
<li><a href="19674980.html">第十七章 丛林绿甲</a></li>
<li><a href="19682593.html">第十八章 我爱工作</a></li>
<li><a href="19696179.html">第十九章 全成就开启</a></li>
<li><a href="19708209.html">第二十章 亡灵法师</a></li>
<li><a href="19722362.html">第二十一章 风声鹤唳</a></li>
<li><a href="19731185.html">第二十二章 狗头人之王</a></li> ...
```

从代码可找到<li>标签较合适，但源代码中还有许多不相关的地方也使用<li>

标签，因此只能从<a>标签中的链接下手，从链接可看到均为数字串，因此可

根据此情况进行正则匹配，规则设定见下图（蓝框内的设置和**红框内**的设置效

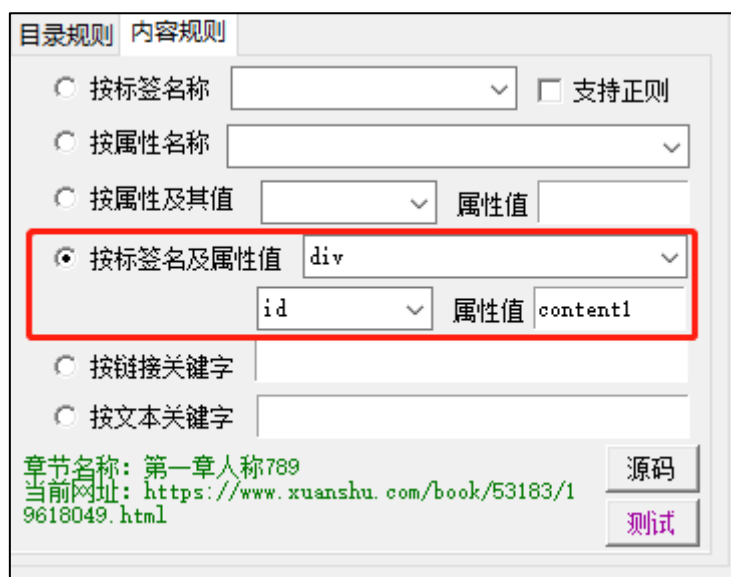
果一样）：

正则表达式：**`^\\d+\\.html`**

^ 表示从开始位置开始匹配，\\d+ 表示任意数字串（如能确定数字的数量，可直接用 \\d{8} 表示 8 位数字）\\.html 表示匹配链接后端的字符，如不清楚的就百度**正则表达式**。

内容页源代码的分析就不细说了，规则设定见下图：





目录规则 内容规则

☐ 按标签名称 ☐ 支持正则

☐ 按属性名称

☐ 按属性及其值 属性值

☒ 按标签名及属性值 div id 属性值 content1

☐ 按链接关键字

☐ 按文本关键字

章节名称: 第一章人称789  
当前网址: https://www.xuanshu.com/book/53183/19618049.html

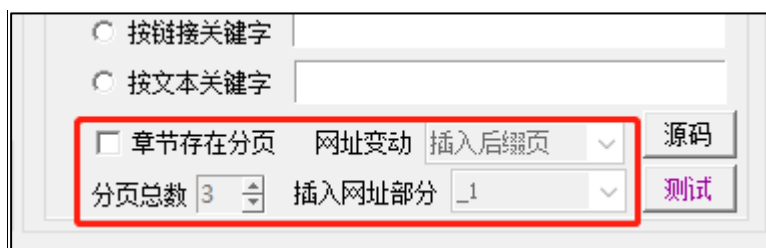
源码 测试

## 实例四：章节分页捕捉实例

部分网站会将章节的内容进行分页显示，即同一章节会在多个网址中分页显示，从而让捕捉难度增加（也许网站的目的是担心章节内容太多，用分页显示来减轻浏览器负担 \*—\*），要区分网站是否分页，只需要随便打开一章节即可了解！如下图：



只要章节内容中显示类似上图的“下一页”导航条，均属章节分页类型网站。对于分页类型的网站，我们要在内容规则页中进行简单的设定，界面如：



☐ 按链接关键字

☐ 按文本关键字

☒ 章节存在分页 网址变动 插入后页

分页总数 3 插入网址部分 \_1

源码 测试

这里我用引用顶点小说网 <https://www.w23us.com> 的其中一本书来作演示：《南禅》-唐酒卿： <https://www.w23us.com/book/14811/>

对于分页类型的网站，我们要注意的就是对比该章节第一页（即从目录中链接过去的网址），与第二页以及后面页网址之间的区别。经分别点击后确定它们的区别（网址前缀省略）：

页数	网址	说明
章节首页	.../book/14811/79.html	即从目录链接过去的网址
第 2 页	.../book/14811/79_1.html	点击下一页时的网址
第 3 页	.../book/14811/79_2.html	再次点击下一页时的网址

从上表中可看出分页网址的变化规律，即为插入式的递增方式，因此在内容规则页的分页处作如下设定即可：

☒ 章节存在分页

网址变动

插入后缀页

分页总数 3

插入网址部分 \_1

网址变动选“插入后缀页”，经查看，大部分章节分为 3 页显示，因此分布总数设为 3，插入网址部分设为“-1”，第 3 页工具将自动递增。其他规则参考前面的方法：目录规则为：按链接关键字【\d+/\d+\.html】，内容规则为：名称【div】属性【id】属性值【content】。按下“测试”可捕捉到章节内容：

【章节第 1 页内容】长度：1419

《防采集章节，与小说内容无关，请勿阅读！！正确的内容在 6 - 9 - 书 - 吧）

察的引导下有序撤离。”

这话说的，好像你和警察是一伙的.....于明疑问：“

.....

.....

这道亮光和时报广场的那片灯光相遇。”

凌晨两点，如两人所预料一样，亮光终于抵达了时报广场，两道亮光汇合，亮光朝北面运动，当亮光离开时报广场时候，时报广场的人失望的叹气，几乎同时，爆炸发生了。一面巨大的电子屏幕发生剧烈爆炸，碎玻璃等物品从十几层的高楼扑洒向楼下的人群，下面的人纷纷躲避，但是-->>本章未完，点击下一页继续阅读

【章节第 2 页内容】长度：1395

发生的太快，加上时报广场驻留的人员很多，这一下造成了重大伤亡。而同时，因为爆炸声，附近的民众开始恐慌逃跑，践踏事故不可避免的发生，现在一片呼爹喊娘声。

.....

.....

测试捕捉到分页的章节内容后，可按“**生成任务**”打开任务窗口开始捕捉。

### 实例五：导入链接并定制规则（序号链接导入）

有一些网站的目录存在多页情况，即把所有章节分别显示在多个页面中，但章节的网址通常都有一定规律（一般为递增的数字），我们可以利用 excel 的填充功能自动生成所有章节的链接，并导入到软件中生成任务。下面以“风雨小说网”（<https://m.44pq.cc/>）作为例子进行演示。

打开“导入章节链接”窗口，步骤为：按【任务管理】打开“捕捉任务窗口”，再按【导入】。




在浏览器中打开该网站的某一书籍，如：[https://m.44pq.cc/chapters\\_1\\_52878/](https://m.44pq.cc/chapters_1_52878/)，这类网站适合用导入链接方法捕捉，章节页面见下图：

第25章：新的功能开启 包裹空间
第26章：大唐第一僧
第27章：授香
第28章：满座皆惊
第29章：前无古人 十二香疤
第30章：我要还俗（求推荐票）
第1/41页[30章/页] 输入页数 跳转

把光标移到章节链接上时，可发现其链接为数字递增方式，但并不是全部都连续（查看最后一章与第一章的网址序号数字相减，如与章节编号相差较大的为不连续），因此我们不能按**内容页**的方式导入。再看看目录列表页，试着按**下一页**，可以发现目录列表的分页有一定规律，因此可以按**目录页**的方式进行导入。从浏览器中可以看到目录的列表页共有 41 页，第一页的网址为：[https://m.44pq.cc/chapters\\_152878/1](https://m.44pq.cc/chapters_152878/1)（如网址为.....chapters\_152878/01，即开始时数字前需要补 0 的，这时需要在“开始”栏中输入“01”字样），最后一页为：[https://m.44pq.cc/chapters\\_152878/41](https://m.44pq.cc/chapters_152878/41)，是按序号方式递增，因此导入设置可如下图设置：



改变的序号在网址的末端，因此“**网址后段**”不需填写，按【**链接检测**】

按键可查看导入结果，在测试结果中按  打开源代码分析规则：

```
内容页源代码 https://m.44pq.cc/chapters_152878/6
字体大小 12 查找
<a class="cur" href="/chapters_152878/">[倒序]</a>
</div>
<ol class="last9">
  <li class="title"><a href="/book_152878/" class="back">返回《打穿西游的唐僧》简介</a></li>
  <li class="even"><a href="/book_152878/58807781.html">第151章：现实现实的猴子（求订阅，求月票）</a></li>
  <li class=""><a href="/book_152878/58807782.html">第152章：师徒联手斗鸟巢（求订阅，求月票）</a></li>
  <li class="even"><a href="/book_152878/58807783.html">第153章：鸟巢的惊人之言（求订阅，求月票）</a></li>
  <li class=""><a href="/book_152878/58807784.html">第154章：鸟巢和如来的对话（求订阅，求月票）</a></li>
  <li class="even"><a href="/book_152878/58807785.html">第155章：孜然是用来做香囊的吗？（求订阅，求月票）</a></li>
  <li class=""><a href="/book_152878/58807786.html">第156章：馋了半夜的黄风大王（求订阅，求月票）</a></li>
  <li class="even"><a href="/book_152878/58807787.html">第157章：群妖汇聚寻辣椒（求订阅，求月票）</a></li>
  <li class=""><a href="/book_152878/58807788.html">第158章：惊慌逃跑的黄风怪</a></li>
  <li class="even"><a href="/book_152878/58807789.html">第159章：灵山鼠患成灾吗？（求订阅，求月票）</a></li>
  <li class=""><a href="/book_152878/58807790.html">第160章：求生欲很强的老鼠（求订阅，求月票）</a></li>
  <li class="even"><a href="/book_152878/58807791.html">第161章：香辣火锅知道是什么不？（求订阅，求月票）</a></li>
</ol>
```

章节都有<li>标签，如用<li>标签进行匹配，也会将“返回”捕捉到（紫色框），但其链接与章节的链接还是有区别，因此可以根据链接进行匹配，规则设定见下图：

导入章节链接

序号链接导入 从文件导入

网址前段  开始

网址后段  结束

书籍名称

网页编码  网址类型

目录页规则 内容页规则

标签名称  ☒ 支持正则

属性名称  属性值

链接匹配

文本匹配

测试结果 些按键可查看源代码

第1章：授香之礼

第2章：杀怪爆东西

第3章：少女高阳

第4章：这个世界上有妖？

第5章：组队

第6章：升级 技能书

第7章：口吐人言的妖怪

第8章：妖卒

第9章：精良品质装备

第10章：待我还俗，娶你可好？

第11章：连孩子的名字都想好了

第12章：怀璧其罪

第13章：非系统技能 罗汉拳

第14章：阴影

第15章：人心险恶

链接检测

规则测试

确定导入

取消导入


需要启用正则表达式支持，因此要将“支持正则”选钩，章节链接的格式为：[/book\\_152878/58807687.html](/book_152878/58807687.html)，正则表达可写为：[/book\\_\d+/\d+](#)

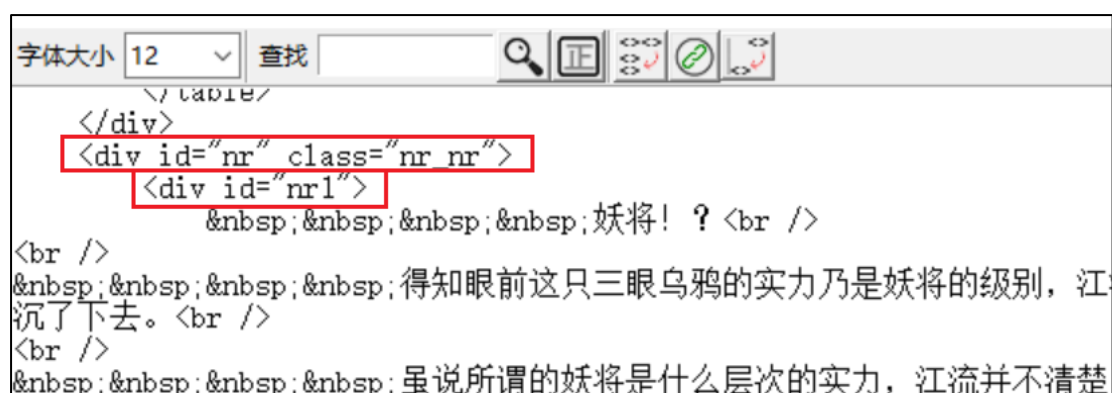
其中 /book\_ 为原样表达，\d+ 表示为第一组、第二组的数字串

上述表达式为化简版，完整版的为：/book\_\d+/\d+\.html

下面的规则设定与上面效果一样

目录页规则		内容页规则	
标签名称	a	<input checked="" type="checkbox"/>	支持正则
属性名称	href	属性值	/book_\d+/\d+

设定好规则后，按【规则测试】查看规则设定效果，正确后可用  分析源代码，见下图：



源代码编辑器截图，显示HTML代码。代码中两处被红框标出：  
1. `<div id="nr" class="nr nr">`  
2. `<div id="nr1">`

从上代码可看到两处符合条件的规则，下面设定均可作捕捉规则：

第一个红框

- ① 标签名称：div      属性名称：id      属性值：nr
- ② 标签名称：div      属性名称：class      属性值：nr\_nr

第二个红框

- ③ 标签名称：div      属性名称：id      属性值：nr1

设定好规则后，按【规则测试】查看内容规则设定效果，如无误后就按【确定导入】键进行书籍导入生成任务。

导入章节链接

序号链接导入

从文件导入

网址前段

https://m.44pq.cc/ch

开始

1

网址后段

结束

41

书籍名称

《打穿西游的唐僧》

网页编码

自动检测

网址类型

目录级别

目录页规则

内容页规则

标签名称

div

规则设定

☐

支持正则

属性名称

id

属性值

nr1

章节名称标签名称

如导入链接为内容页，此处需要设定规则

属性名称

章节名称规则

属性值

链接检测

规则测试

确定导入

取消导入

测试结果

“江流，这是一把粗盐，你今天晚上的任务是自己把它提纯成食用盐，既然你让我跟着我学厨艺，你就先从调料开始，我不管外面那些餐馆是怎样揽客的，我告诉你，三流的厨师才放乱七八糟的一大堆调味料，一流的顶尖厨师只放盐……”，一个高大的男子，穿着洁白的厨师服，头戴高高的厨师帽，神色严厉。

“粗盐到食用盐的方式，需要提纯，过滤杂质……”，少年面前摆着一些滤纸、烧杯、玻璃杯之类的工具，手中捧着智能机，搜索着自己想要的讯息，恰在此时，一段语音发过来：“流哥，赶紧上线，今天晚上攻略终极BOSS牛魔王，少不了你”。

某个昏暗的网吧，桌子上摆着三个空空如也的矿泉水瓶，少年双目布满血丝，却精神亢奋的盯着电脑屏幕，看着游戏中的大BOSS血条几乎空了，无比激动。

终于，当BOSS倒下的一刻，无数的光芒四散而出，少年紧绷的神经突然放松，双眼一黑，失去了意识。

……

咯咯咯！

半睡半醒之间，隐约间传来雄鸡报晓之声，江流缓缓的睁开了双眼，坐起身来。

目光掠过窗户，看着外面依旧昏暗的景色，隐约间能看到山体的轮廓和树木，心中暗叹一声：来到这个世界已经快半个月了，心中依旧希望着睁开眼的时候，已经回到了现代吗？

摸黑起身，点燃一盏油灯，微弱的灯光驱散了禅房内的昏暗

内容正确后可开始导入

## 实例六：导入链接并定制规则（从文件导入）

我们一样用实例四的网站进行演示，即把链接用 Excel 软件的序号填充功能生成并导入。需要生成的链接：

[https://m.44pq.cc/chapters\\_152878/1](https://m.44pq.cc/chapters_152878/1)

[https://m.44pq.cc/chapters\\_152878/2](https://m.44pq.cc/chapters_152878/2)

.....

[https://m.44pq.cc/chapters\\_152878/41](https://m.44pq.cc/chapters_152878/41)

打开 Excel 软件，在 A1 单元格内把网址粘贴上去（软件自动容错为 10 行、10 列，即保证第 1 个链接的位置在 A1:J10 之间的区域内均可，超出后将提示找不到有效链接，建议大家在 A1 单元格中粘贴），然后利用 Excel 的自动填充功能补充余下的章节链接：

	A	B	C	D
1	<a href="https://m.44pq.cc/chapters_152878/1">https://m.44pq.cc/chapters_152878/1</a>			
2				
3				
4				
5				
6				
7				

单元格右下解绿色小方块为填充柄，  
按住向下拖动可实现余下链接的填充



选中粘贴网址的单元格，按住单元格右下角的填充柄向下拉，即可填充余下的章节。

	A	B	C	D
1	<a href="https://m.44pq.cc/chapters_152878/1">https://m.44pq.cc/chapters_152878/1</a>			
2				
3				
4				
5				
6				
7				
8				
9				

注意此数字达到  
最后链接的数字  
时即可松手

如章节效多，手动填充不方便，这里教大家一招，先算出最后章节出现的行数按下图操作可快速生成序号链接。

①	A	②	A	B	③	A	B	C	D
1658	先翻到要填充	1	<a href="https://m.44pq.cc/chapters_152878/1">https://m.44pq.cc/chapters_152878/1</a>		1	<a href="https://m.44pq.cc/chapters_152878/1">https://m.44pq.cc/chapters_152878/1</a>			
1659	的行号，选中	2	1		2	<a href="https://m.44pq.cc/chapters_152878/2">https://m.44pq.cc/chapters_152878/2</a>			
1660	后用	3	回到首行粘贴入第一个		3	<a href="https://m.44pq.cc/chapters_152878/3">https://m.44pq.cc/chapters_152878/3</a>			
1661	Shift+Home	4	链接，再双击该单元格		4	<a href="https://m.44pq.cc/chapters_152878/4">https://m.44pq.cc/chapters_152878/4</a>			
1662	再随便输入一	5	的填充柄，会自动向下		5	<a href="https://m.44pq.cc/chapters_152878/5">https://m.44pq.cc/chapters_152878/5</a>			
1663	些字符，用	6	填充，链接序号也将自		6	<a href="https://m.44pq.cc/chapters_152878/6">https://m.44pq.cc/chapters_152878/6</a>			
1664	Ctrl+回车	7	动递增		7	<a href="https://m.44pq.cc/chapters_152878/7">https://m.44pq.cc/chapters_152878/7</a>			
1665	全部填充	8			8	<a href="https://m.44pq.cc/chapters_152878/8">https://m.44pq.cc/chapters_152878/8</a>			
1666		9			9	<a href="https://m.44pq.cc/chapters_152878/9">https://m.44pq.cc/chapters_152878/9</a>			
1667		10			10	<a href="https://m.44pq.cc/chapters_152878/10">https://m.44pq.cc/chapters_152878/10</a>			
1668		11			11	<a href="https://m.44pq.cc/chapters_152878/11">https://m.44pq.cc/chapters_152878/11</a>			
1669		12			12	<a href="https://m.44pq.cc/chapters_152878/12">https://m.44pq.cc/chapters_152878/12</a>			
1670		13			13	<a href="https://m.44pq.cc/chapters_152878/13">https://m.44pq.cc/chapters_152878/13</a>			

保存后回到本软件的导入窗口，选“从文件导入”，按【浏览】指定刚保存的 Excel 文档，再按【链接检测】获取链接。



导入章节链接

序号链接导入 **从文件导入**

指定文件 D:/《打穿西游的唐僧》.xlsx 浏览

书籍名称 《打穿西游的唐僧》

网页编码 自动检测 网址类型 目录级别

目录页规则 内容页规则

标签名称 支持正则

属性名称 属性值

链接匹配

文本匹配

链接检测 规则测试 确定导入 取消导入

测试结果

检测到文件中有效的链接共 41 条!

<https://m.44pq.cc/chapters/152878/1>

<https://m.44pq.cc/chapters/152878/2>

<https://m.44pq.cc/chapters/152878/3>

<https://m.44pq.cc/chapters/152878/4>

<https://m.44pq.cc/chapters/152878/5>

<https://m.44pq.cc/chapters/152878/6>

<https://m.44pq.cc/chapters/152878/7>

<https://m.44pq.cc/chapters/152878/8>

<https://m.44pq.cc/chapters/152878/9>

<https://m.44pq.cc/chapters/152878/10>

<https://m.44pq.cc/chapters/152878/11>

<https://m.44pq.cc/chapters/152878/12>

<https://m.44pq.cc/chapters/152878/13>

<https://m.44pq.cc/chapters/152878/14>

后面的规则设定与实例四一样，就不再重复了。