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

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。

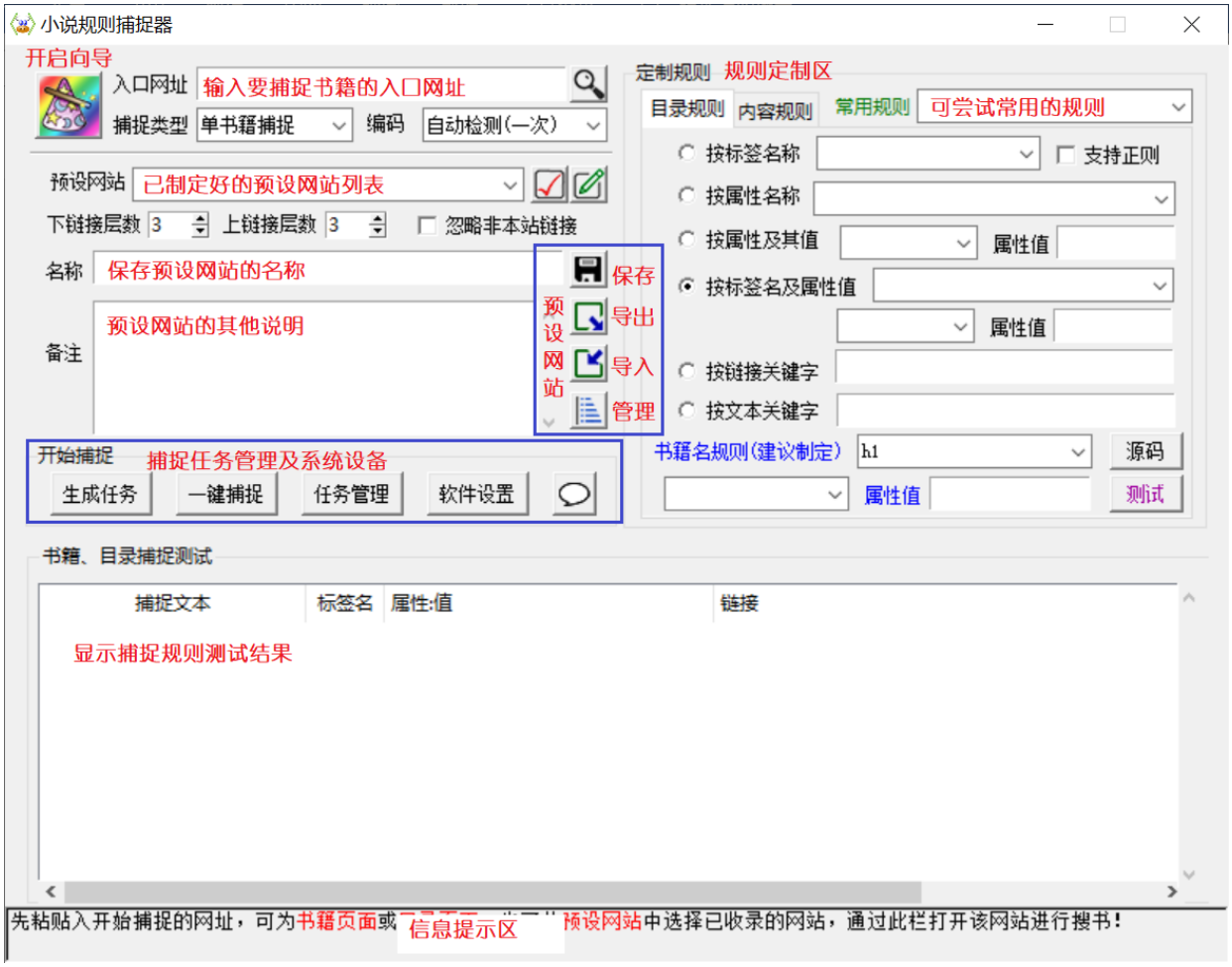
软件特色

- 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

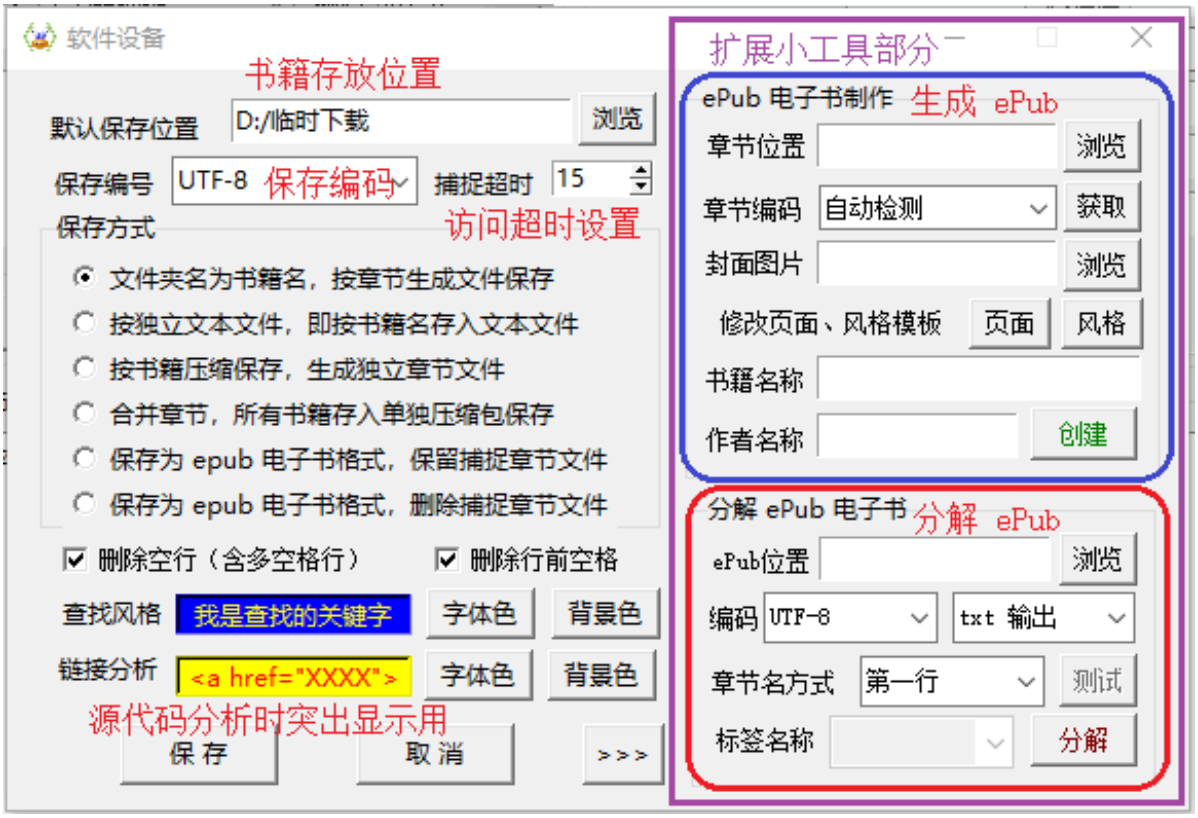
1、 主界面


小说规则捕捉器
开启向导
入口网址
输入要捕捉书籍的入口网址
捕捉类型
单书籍捕捉
编码
自动检测(一次)
预设网站
已制定好的预设网站列表
下链接层数 3
上链接层数 3
忽略非本站链接
名称
保存预设网站的名称
备注
预设网站的其他说明
保存
导出
导入
管理
预设网站
开始捕捉
捕捉任务管理及系统设备
生成任务
一键捕捉
任务管理
软件设置
定制规则
规则定制区
目录规则
内容规则
常用规则
可尝试常用的规则
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
按链接关键字
按文本关键字
书籍名规则(建议制定)
h1
源码
属性值
测试
书籍、目录捕捉测试
捕捉文本
标签名
属性:值
链接
显示捕捉规则测试结果
先粘贴入开始捕捉的网址，可为书籍页面或
信息提示区
预设网站中选择已收录的网站，通过此栏打开该网站进行搜书！

2、 任务管理

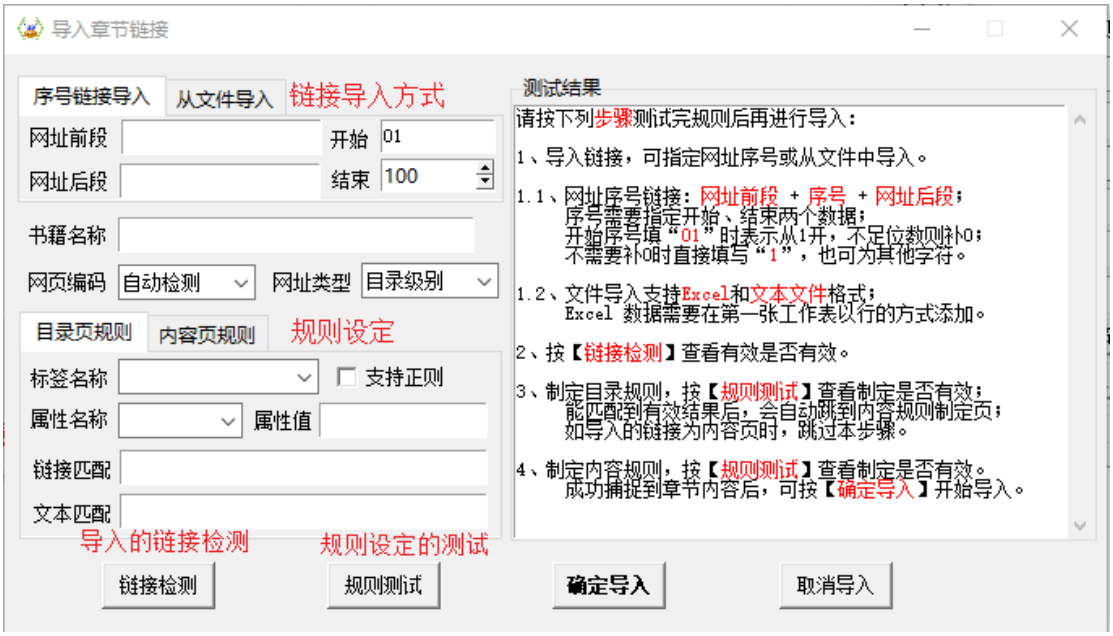

捕捉任务
开始
重置
恢复
标题
链接
删重
查看
过滤
捕捉
取消
反置
导入
关闭
开始：开始捕捉章节内容
重置：重新获取目录信息，
可在调整章节失误后用
恢复：中断后恢复捕捉用，
不需要重头捕捉
序号
尺寸
查看：显示已捕捉章节的文件尺寸
过滤：根据文件尺寸决定是否重新捕捉
捕捉：将选择中的章节设为重新捕捉标记
取消：不捕捉当前选中的章节
反置：将选中章节捕捉标记反置
标题：删除相邻的同名章节
链接：删除相邻的同链接章节
删重：删除重复章节的高级设置
导入：从文本、Excel文件中
导入任务
置顶
上移
下移
置底
删除
倒置
第4章大
第5章风
第6章有
第7章刻算师
第8章万众瞩目
已分析72个章节文件，其中未捕捉章节共68个，需要重新捕捉章节共68个，已正确捕捉章节共4个！

3、 软件设置



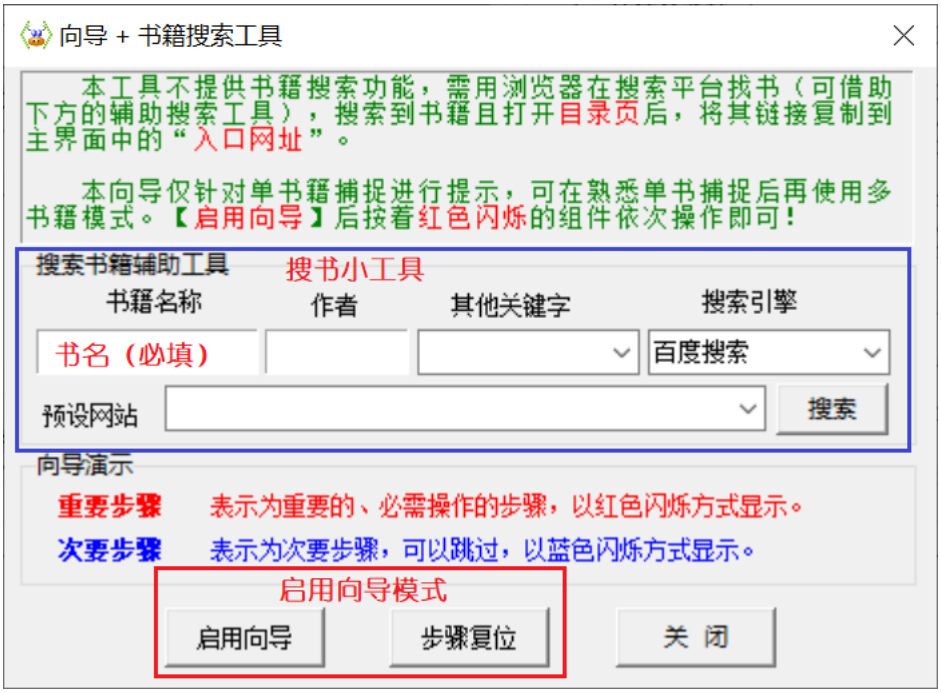
4、 预设网站管理

5、 导入链接（任务导入）

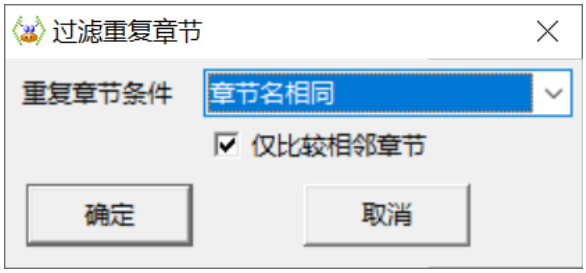


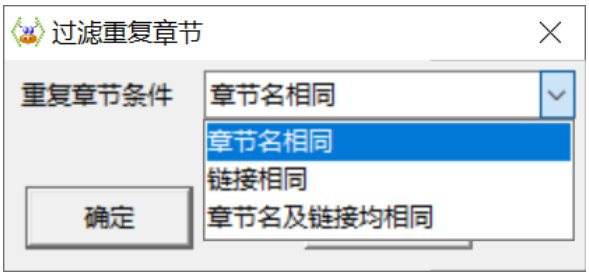
6、 源代码查看

7、 向导、书籍搜索辅助工具



8、 章节重复过滤



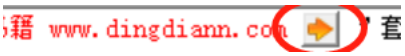


操作步骤

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  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（仅供参考）：

点击找到的小说进入到目录页（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：

文本信息

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

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**（即标签）存在多层关系，即在捕捉链接时不一定要匹配相关的<a>标签（也称为节点），可以是其附近的其他标签，层数范围可以调整（默认为 3 层，1 时表示匹配标签即是<a>标签），设置过大会稍微影响捕捉速度。

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 ...>书名 1

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

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

<li class=book>书名 1

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

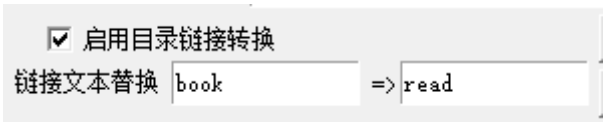
<a>标签是第二层父标签，带链接

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：

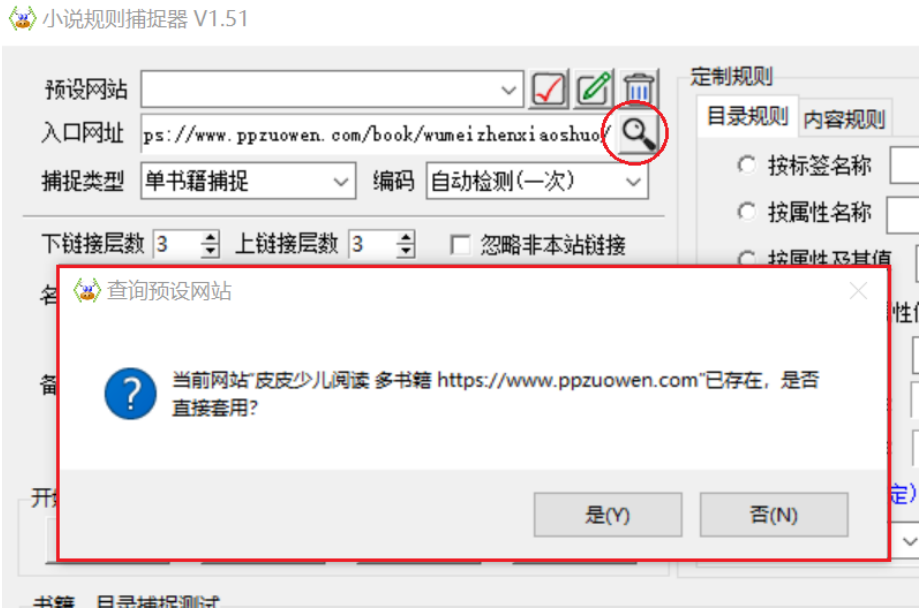


三、规则定制实例

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：

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：

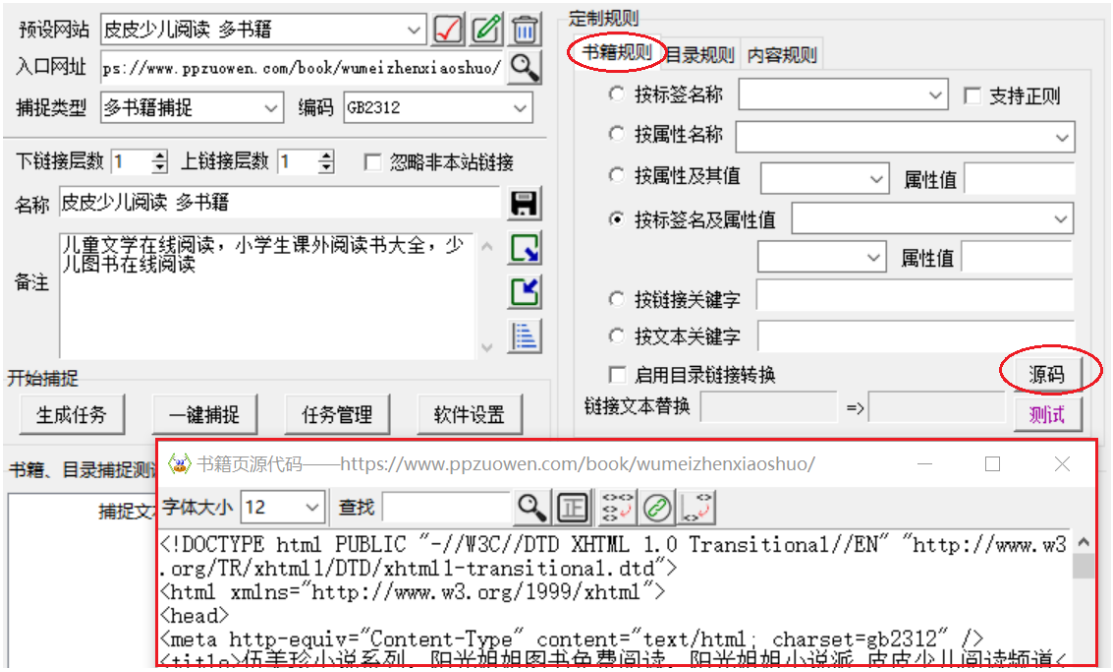
一、确定捕捉级别

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。

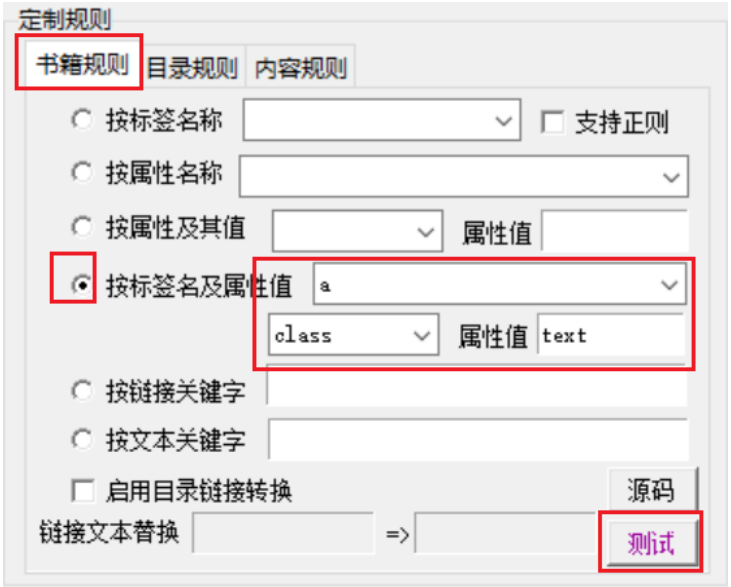
1、书籍页规则制定

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：

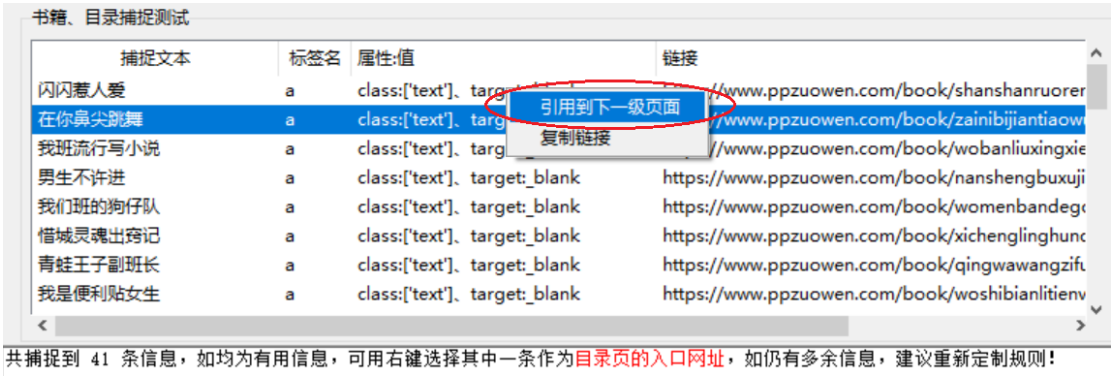
```
<a href="/book/shanshanruorenai/" class="text" target="_blank" >闪闪惹人爱</a>
```

> 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：



按【测试】查看捕捉效果



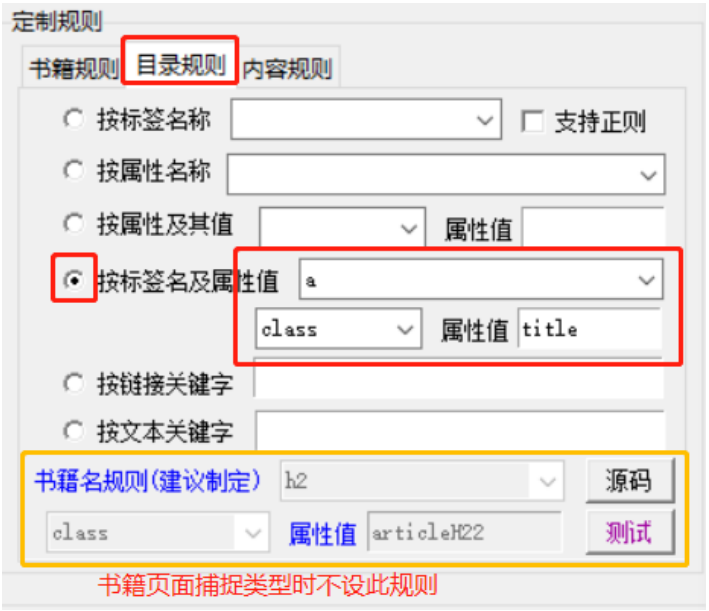
捕捉正确后，选中其中一个目录页点鼠标右键，按“**引用到下一级页面**”，即将其链接做为目录页的网址，进行目录页的规则制定。

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：

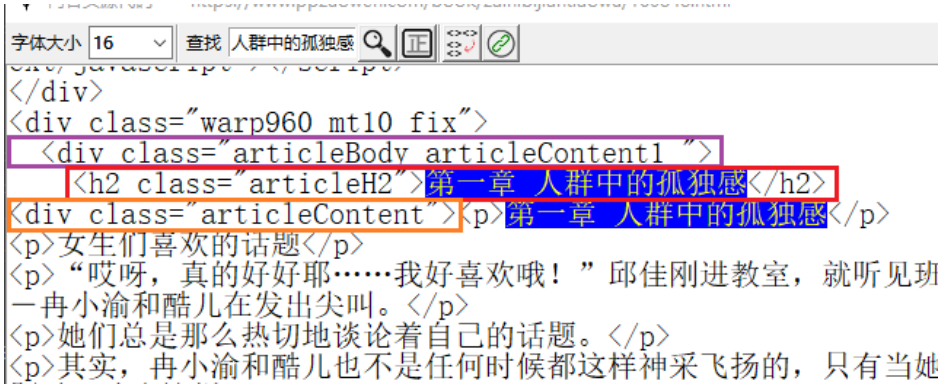


查看测试效果：

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:



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 标签，可先制定为：标签名=div，属性名=class，属性值=articleBody articleContent1

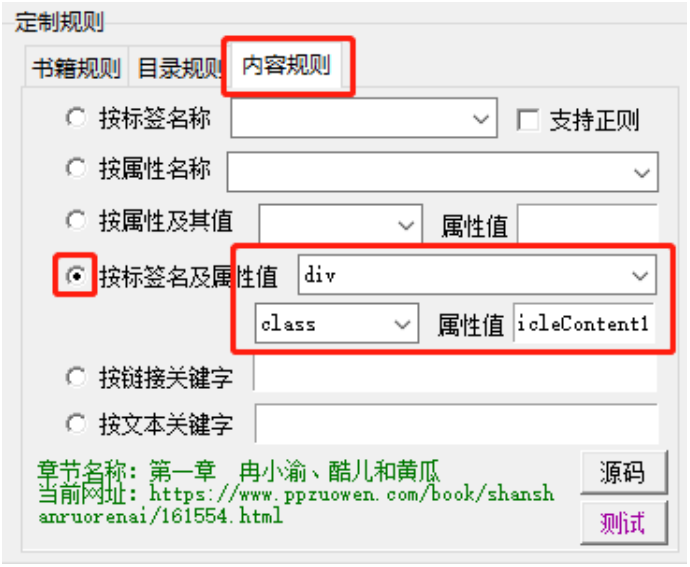
测试发现捕捉时多了一行章节名，不理想。

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：

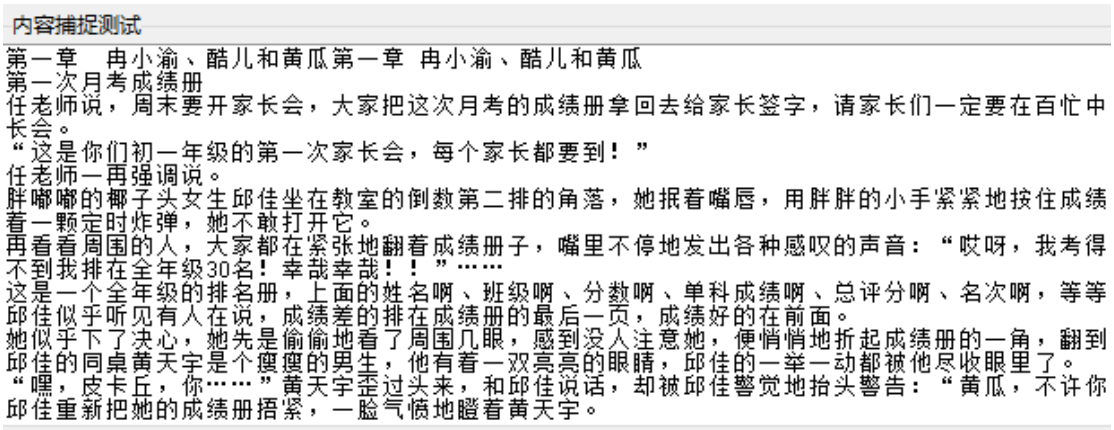
第一章 人群中的孤独感

第三行代码 标签，可先制定为：标签名=div，属性名=class，属性值=articleContent

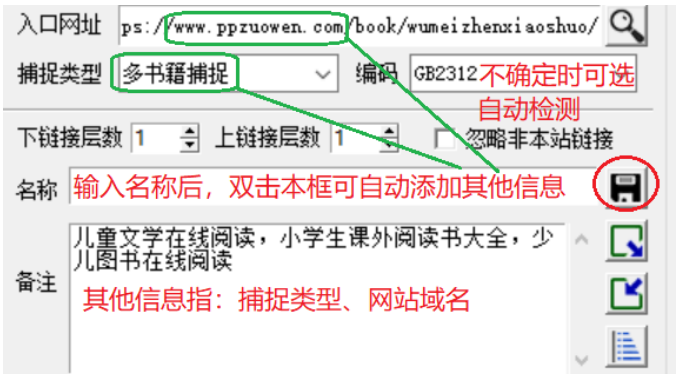
为较理想的结果，因此可将规则设以上述值，规则设定见下图：



测试结果与书籍、目录时不一样：



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：

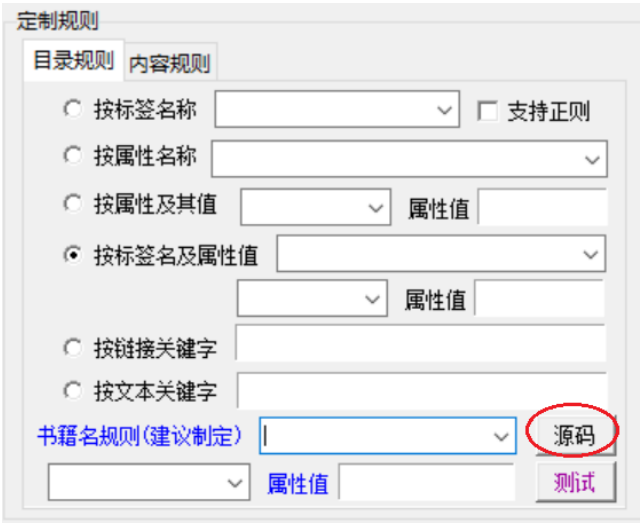


最后按【生成任务】开始捕捉书籍，如测试目录规则时发现章节顺序均正确也可以按【一键捕捉】直接开始捕捉。

实例二：单书籍规则定制

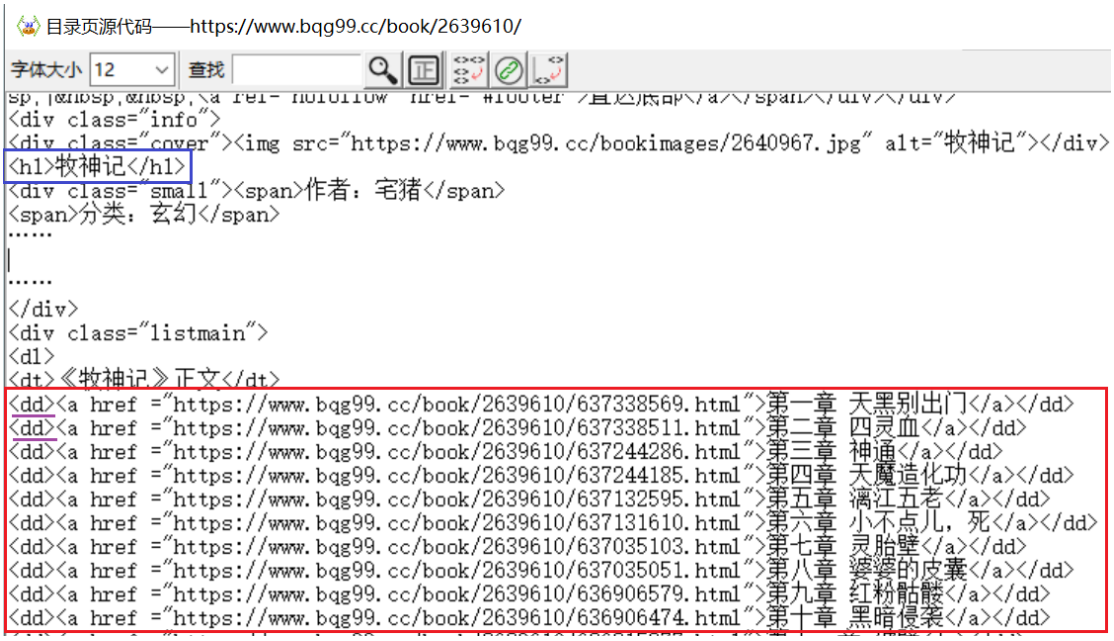
本例选用较常见的笔趣阁网站进行分析（网站笔趣阁网站非常多，就本软件默认收录的就多达 30 个），这类网站不建议使用多书籍捕捉模式，选用“**单书籍捕捉**”即可，直接在浏览器中搜索到相捕捉的书籍后，打开到该书籍的目录页，如下图：

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



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。

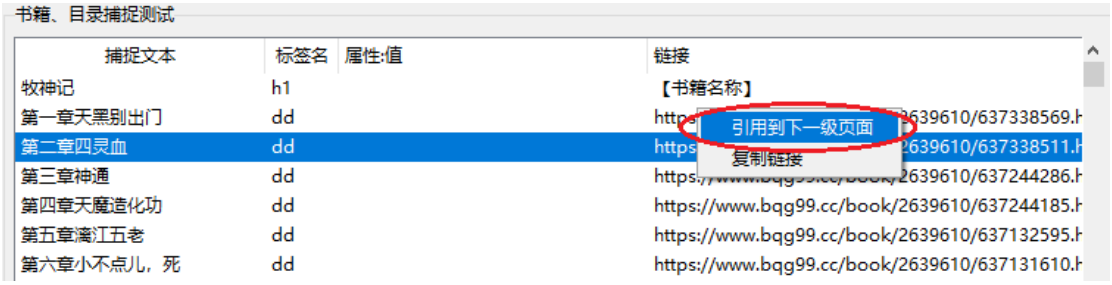
通过分析，可以找到一些关键的标签，如下图，蓝色的<h1>标签表示为书名（单书籍捕捉时最好制定书籍名称规则，这样导出的书籍也会按书名来命名，方便管理）；红色部分表示为章节部分的代码，每个章节的标签为<dd>。



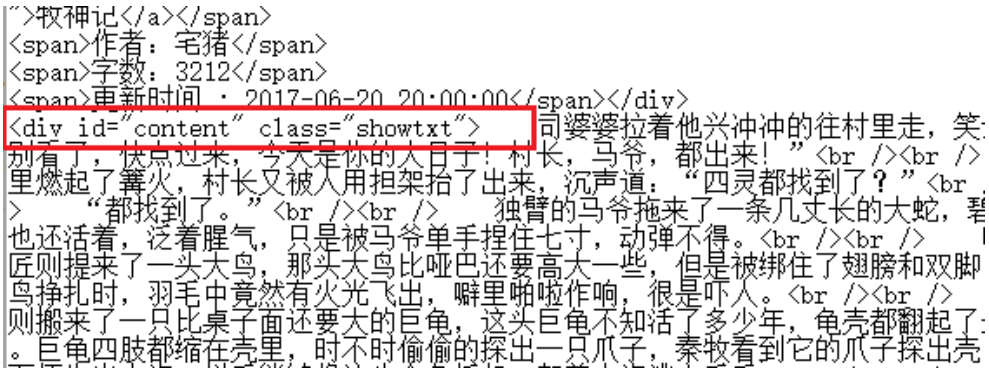
目录规则设定如下：

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：



打开内容页的源代码，可以看到下面的信息：



可以分析到关键标签为 ，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。

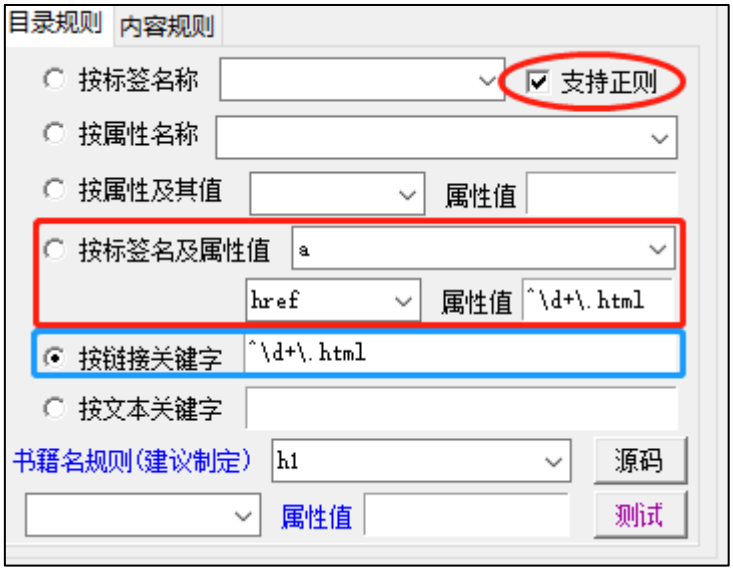
实例三：使用正则表达式定制规则

本例演示按正则表达示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：

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：

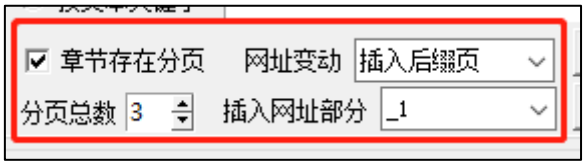
只要章节内容中显示类似上图的“下一页”导航条，均属章节分页类型网站。对于分页类型的网站，我们要在内容规则页中进行简单的设定，界面如：

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：



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)
察的引导下有秩序的撤离。”
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。

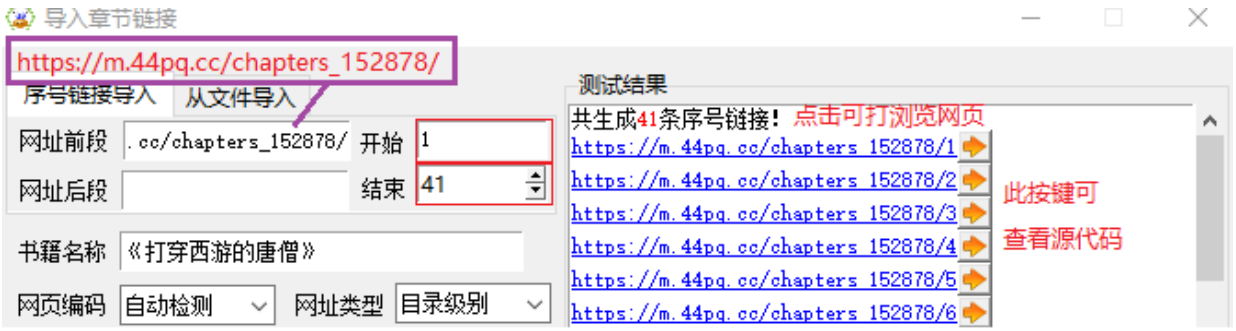
实例五：导入链接并定制规则（序号链接导入）

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。

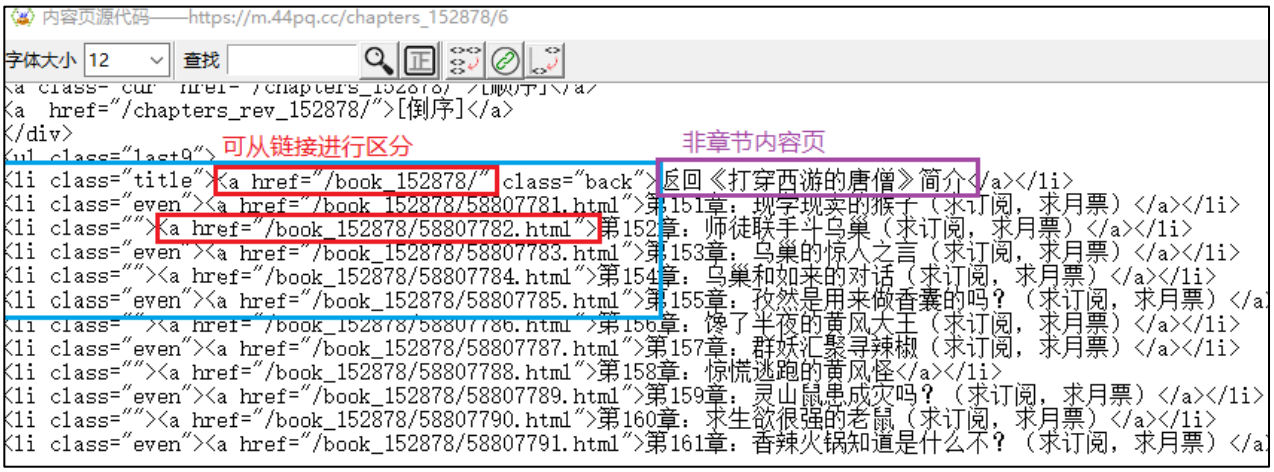
打开“导入章节链接”窗口，步骤为：按【任务管理】打开“捕捉任务窗口”，再按【导入】。

在浏览器中打开该网站的某一书籍，如：*https://m.44pq.cc/chapters_152878/*，这类网站适合用导入链接方法捕捉，章节页面见下图：

把光标移到章节链接上时，可发现其链接为数字递增方式，但并不是全部都的连续（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（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，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：



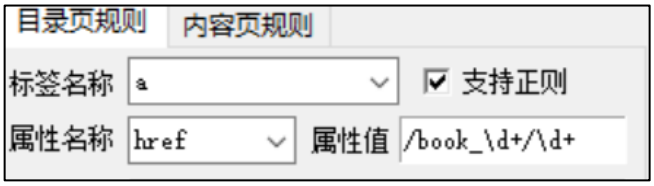
章节都有<li>标签，如用<li>标签进行匹配，也会将“**返回**”捕捉到（紫色框），但其链接与章节的链接还是有区别，因此可以根据链接进行匹配，规则设定见下图：

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，正则表达可写为：/book_\d+/\d+

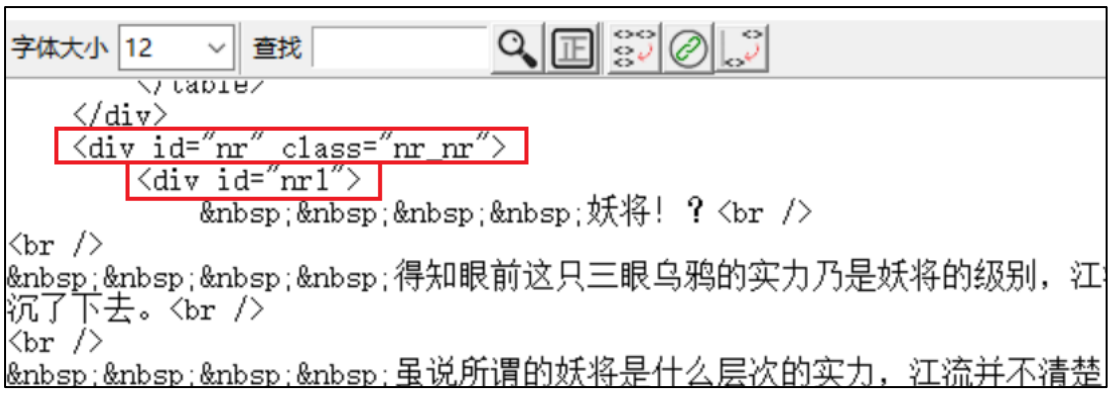
其中 /book_ 为原样表达，\d+ 表示为第一组、第二组的数字串

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



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：



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

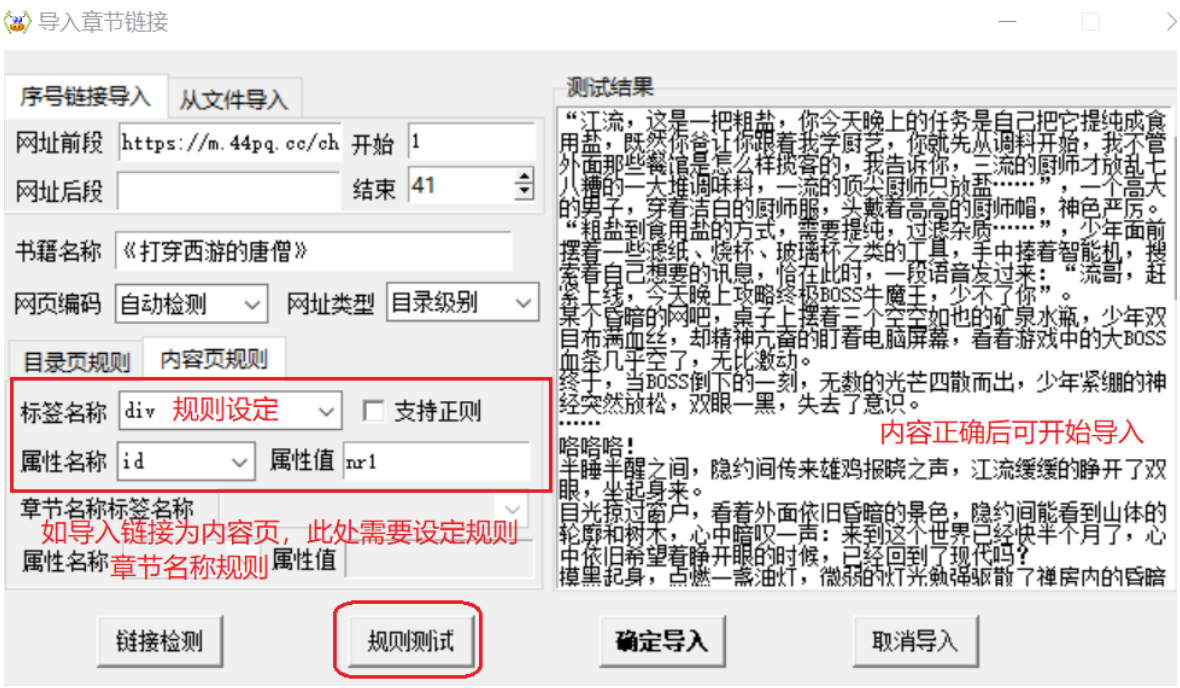
① 标签名称：**div**　　属性名称：**id**　　属性值：**nr**

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。

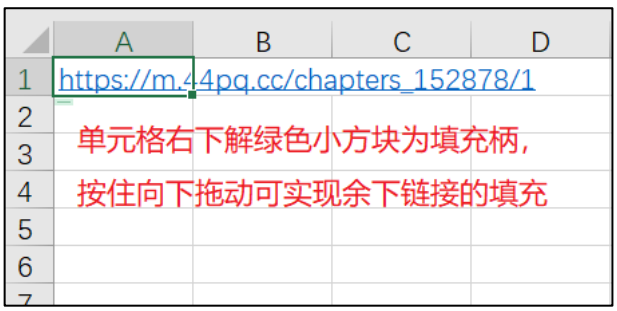


实例六：导入链接并定制规则（从文件导入）

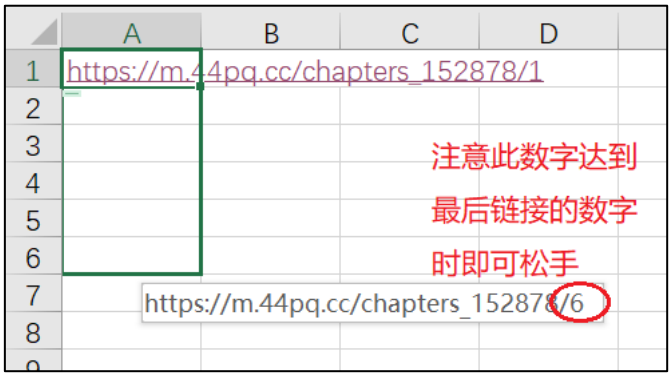
我们一样用实例四的网站进行演示，即把链接用 Excel 软件的序号填充功能生成并导入。需要生成的链接：

https://m.44pq.cc/chapters_152878/1
https://m.44pq.cc/chapters_152878/2
……
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：

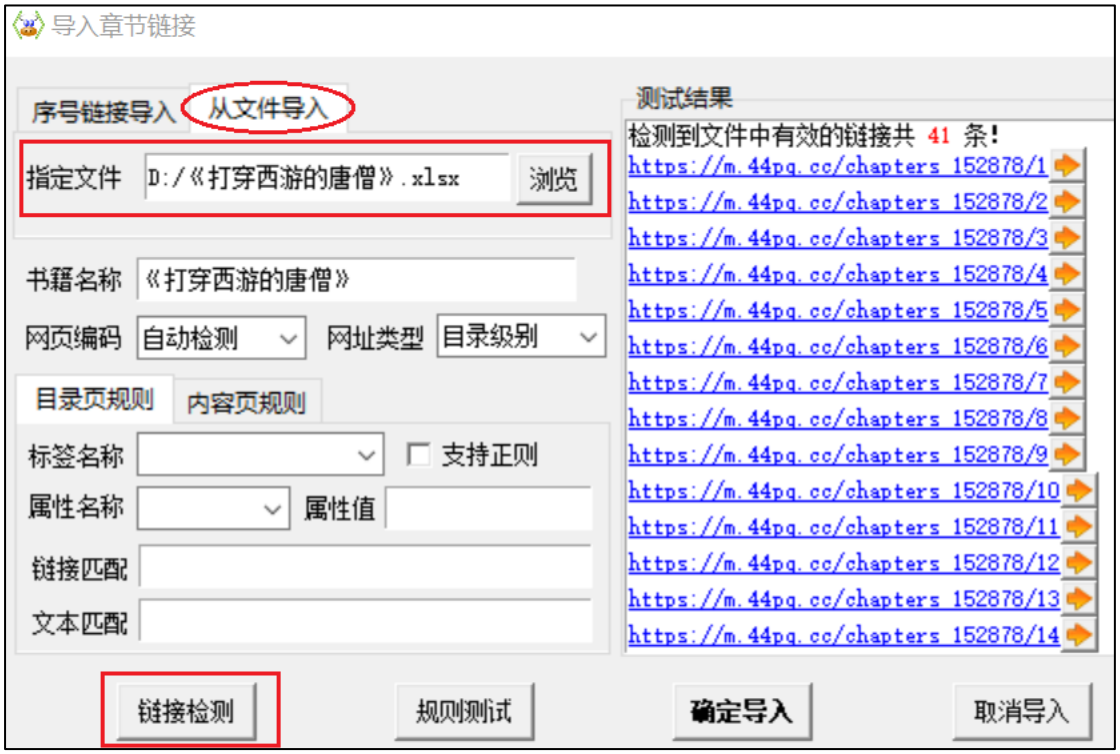


选中粘贴网址的单元格，按住单元格右下角的填充柄向下拉，即可填充余下的章节。



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。

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。



后面的规则设定与实例四一样，就不再重复了。