WordPress 自动采集发布插件 WP-AutoPost 使用教程

WP-AutoPost 是目前最好用的 WordPress 自动采集发布插件，最大的特点是可以采集来自于任何网站的内容并自动发布到你的 WordPress 站点。不像其他大部分 WordPress 采集插件，只能根据 Feed 进行采集，使用 Feed 采集有很大的弊端，首先必须找到全文 Feed，而现在网上全文 Feed 很少，大部分只能采集到文章摘要，就算使用 Feed 采集到文章摘要也需要点击链接才能查看原文，相当于给别的网站做外链。做英文垃圾站用的比较多的 WP-Robot 也只有 20 几个采集来源，文章来源比较单一有限。而 WP-AutoPost 就没有上面这些劣势，真正做到可以采集任何网站的内容并自动发布，采集过程完全自动进行无需人工干预，并提供内容过滤、HTML 标签过滤、关键词替换、自动链接、自动标签、自动下载远程图片到本地服务器、自动添加文章前缀后缀、并且可以使用微软翻译引擎将采集的文章自动翻译为各种语言进行发布。

WP-AutoPost中文免费下载地址http://wp-autopost.org/zh

一、安装 WP-AutoPost

和安装其他 WordPress 插件一样，直接上传到插件目录，激活即可使用，无需再进行额外设置或修改代码。

二、创建采集任务

点击“新建任务”后，输入任务名称，即可创建新任务，创建好新任务之后可以在任务列表中查看到该任务，就可对该任务进行更多设置。

三、基本设置功能

在基本设置选项卡下，可以进行如下设置：

- **任务名称**：可以修改任务名称
- **分类目录**：该任务采集文章发布到的分类目录
- **作者**：该任务采集文章的发布作者，必须是 WordPress 里的注册用户
- **更新时间间隔**：间隔多长时间检测一次该采集任务下是否有新文章可以更新
- **字符集**：采集目标网站的字符集编码，默认为UTF8，如果目标网页字符集编码不是UTF8，抓取的网页会出现乱码，设置正确的字符集即可解决该问题（如何正确设置字符集）
- **下载远程图片**：如果该任务下采集的文章里包含图片，可以选择是否下载远程图片到本地服务器
    - 选择下载远程图片，可进一步选择是否将下载的图片信息保存到 WordPress 媒体库中
- **自动标签**：选择是否使用自动标签
    - **标签列表**：使用自动标签后如果文章包含列表里的关键词，将自动添加标签
    - **匹配完整的单词**：该设置对英文文章有效，中文文章请勿启用该设置

四、文章来源设置

在该选项卡下我们需要设置文章来源的 **文章列表网址** 及 **具体文章的匹配规则**
我们以采集“新浪互联网新闻”为例，文章列表网址为http://roll.tech.sina.com.cn/internet_worldlist/index.shtml 因此在 **手工指定文章列表网址** 中输入该网址即可，如下所示：

之后需要设置该文章列表网址下**具体文章网址的匹配规则**

五、文章网址匹配规则

文章网址匹配规则的设置非常简单，无需复杂设置，提供两种匹配模式，可以使用 URL 通配符匹配，也可以使用 CSS 选择器进行匹配，通常使用 URL 通配符匹配较为简单。

1. 使用 URL 通配符匹配

通过点击列表网址 http://roll.tech.sina.com.cn/internet_worldlist/index.shtml 上的文章，我们可以发现每篇文章的URL都为如下结构
http://tech.sina.com.cn/i/2013-06-27/16328485884.shtml

因此将 URL 中变化的数字或字母替换为通配符 **(*)** 即可，如：http://tech.sina.com.cn/i/**(*)**/**(*)**.shtml

2. 使用 CSS 选择器进行匹配

使用CSS选择器进行匹配，我们只需要设置**文章网址的CSS选择器**即可（不知道CSS选择器为何物，一分钟学会如何设置CSS选择器），通过查看列表网址http://roll.tech.sina.com.cn/internet_worldlist/index.shtml的源代码即可轻松设置，找到该列表网址下具体文章的超链接的代码，如下所示：

```
<!-- 列表 begin -->
        <div class="contList">
        <ul>
                <li><a href="http://tech.sina.com.cn/i/2013-06-
11/06478431716.shtml" target=_blank>Rovio任命前诺基亚高管为COO</a><span class="c_date">
(新浪科技 06月11日 06:47)</span></li>
                                <li><a href="http://tech.sina.com.cn/i/apple/2013-
06-11/06308431671.shtml" target=_blank>苹果公司推音乐流播放服务：或成Pandora杀手</a><span
class="c_date"> (新浪科技 06月11日 06:30)</span></li>
                                <li><a href="http://tech.sina.com.cn/i/2013-06-
11/05548431658.shtml" target=_blank>每日风投速递：谷歌或13亿美元鲸吞Waze</a><span
class="c_date"> (IT桔子 06月11日 05:54)</span></li>
```

可以看到，文章的超链接 a 标签在 class 为“contList”的标签内部，因此**文章网址的 CSS 选择器**只需要设置为 **.contList a** 即可，如下所示：

设置完成之后，不知道设置是否正确，可以点击上图中的测试按钮，如果设置正确，将列出该列表网址下所有文章名称和对应的网页地址，如下所示：

六、文章抓取设置

在该选项卡下，我们需要设置文章标题和文章内容的匹配规则，提供两种方式进行设置，推荐**使用CSS选择器**方式，使用该方式更为简单，精确。（不知道CSS选择器为何物，一分钟学会如何设置CSS选择器）

我们只需要设置**文章标题 CSS 选择器**和**文章内容 CSS 选择器**，即可准确抓取文章标题和文章内容。

在文章来源设置里，我们以采集”新浪互联网新闻“为例，这里还是以该例子讲解，通过查看列表网址http://roll.tech.sina.com.cn/internet_worldlist/index.shtml下某一篇文章的源代码即可轻松设置，例如，我们通过查看某篇具体文章http://tech.sina.com.cn/n/i/2013-06-10/06308430630.shtml的源代码，如下所示：

```
          <div class="blkContainerSblk">
                        <h1 id="artibodyTitle" pid="2" tid="1" did="8430630"
fid="2108">遥遥无期？Google重返中国市场无望</h1>

             <div class="artInfo">
```

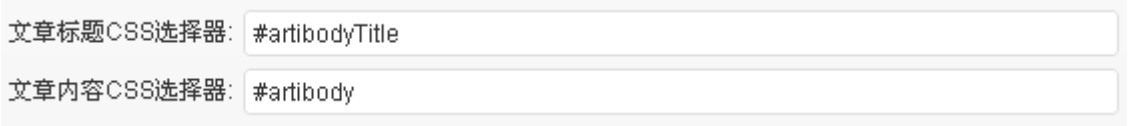
可以看到，文章标题在 id 为“artibodyTitle”的标签内部，因此**文章标题 CSS 选择器**只需要设置为 **#artibodyTitle** 即可；

同样的，找到文章内容的相关代码：

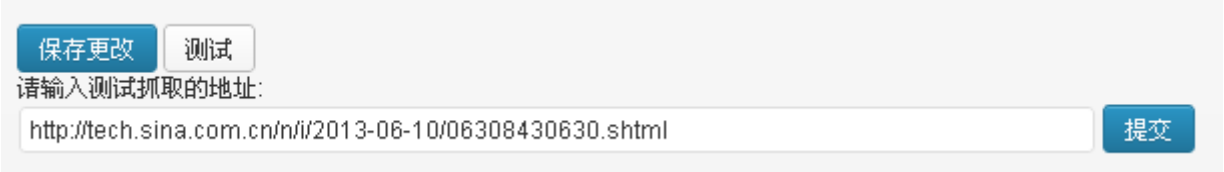
```
                        <!-- 正文内容 begin -->
            <div class="blkContainerSblkCon BSHARE_POP" id="artibody">
                        <!--视频轮播start-->

                                    <!--视频轮播end-->
<!-- 内容模块：单图 begin -->

<div class="img_wrapper">

                  <center></center></div>
<!-- 内容模块：单图 end -->
<!-- publish_helper name='原始正文' p_id='2' t_id='1' d_id='8430630' f_id='2' -->

<p>    谷歌应用其实在平板上随处可见，作为占据平板半数操作平台的安卓，得到如此待遇理所应当。不过，应用虽有，但却并不好用，原理了大陆市场多年多年的谷歌服务何时能够再次顺手起来成了大家关心的问题。</p>

<div class="img_wrapper"><img alt="遥遥无期？Google重返中国市场无望" style="border:1px
solid #000;" src="http://i1.sinaimg.cn/IT/cr/2013/0610/3340307132.jpg" /></div>

<p>    不知不觉 已是三年</p>

<p>    近日，Google执行董事长Eric Schmidt在日前的公司股东大会上透露，短期内，Google不会重新返回中国市场。而原因，则与当年推出别无二致，其表示“持续不断提供审查报告并且监控用户这类事情，是谷歌无法接受的，也是其不愿向中国市场投入更多资源的原因。”</p>

<p>    距离谷歌退出中国市场已有3年，尽管这期间中国网民总数增长到了5.6亿，但这仍没有改变谷歌的态度，不过索性谷歌仍旧给大陆留下了一个线索。用户访问Google.cn后将被指向Google.com.hk，在这个域名上，Google提供未经审查的简体中文搜索结果。</p>

<p>    编者评论：作为安卓的实际持有者，谷歌退出中国市场无疑是其大数据收集上的巨大损失，并且如今看来也没有任何回转的可能。因此说来谷歌当年之举，看来难免冲动了些。</p>
<!-- publish_helper_end -->

                        </div>
```

可以看到，文章内容在 id 为“artibody”的标签内部，因此**文章内容 CSS 选择器**只需要设置为 **#artibody** 即可；

如下所示：

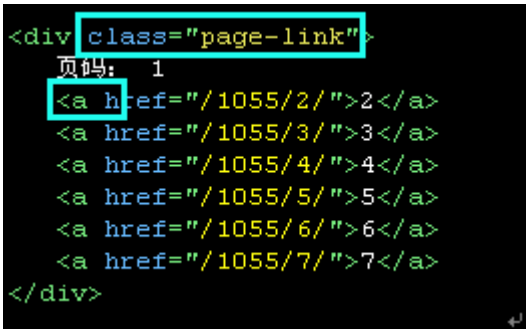


设置完成之后，不知道设置是否正确，可点击测试按钮，输入测试地址，如果设置正确，将显示出文章标题和文章内容，方便检查设置

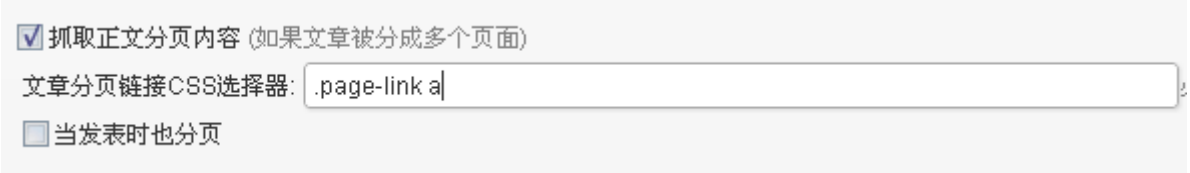


七、抓取文章分页内容

如果文章内容过长，有多个分页同样可以抓取全部内容，这时需要设置**文章分页链接 CSS 选择器**，通过查看具体文章网址源代码，找到分页链接的地方，例如某篇文章分页链接代码如下：



可以看到，分页链接 A 标签在 class 为 “page-link” 的标签内部
因此，**文章分页链接 CSS 选择器** 设置为 **.page-link a** 即可，如下所示：

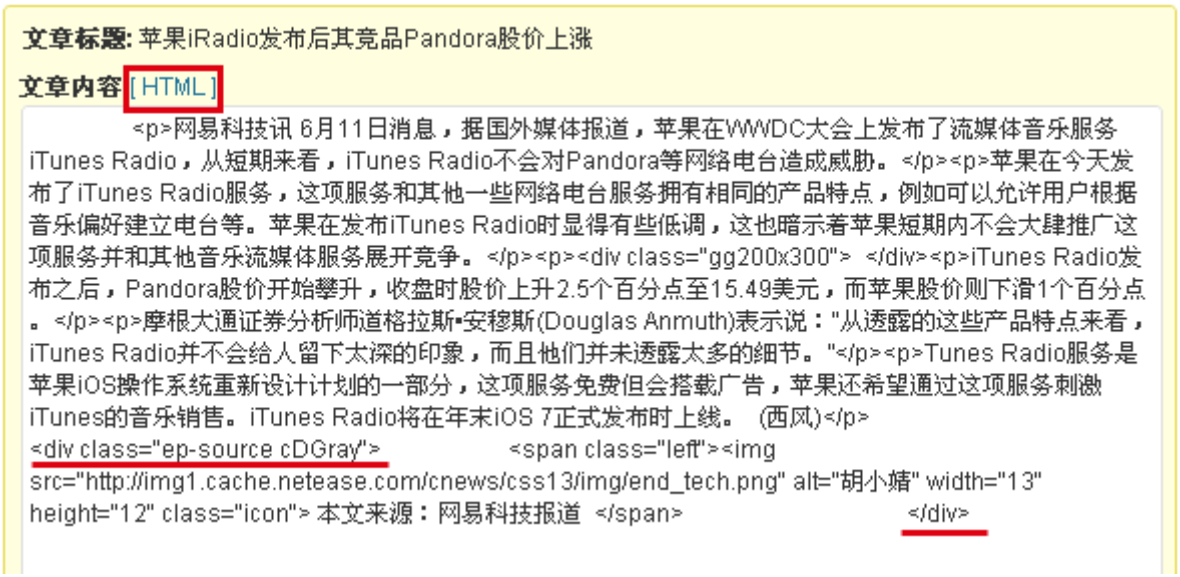


如果勾选 **当发表时也分页** 时，发表文章也将同样被分页，如果你的 WordPress 主题不支持 <!- - nextpage - -> 标签，请勿勾选。

八、文章内容过滤功能

文章内容过滤功能，可过滤掉正文中不希望发布的内容（如广告代码，版权信息等），可设置两个关键词，删除掉两个关键词之间的内容，关键词 2 可以为空，表示删除掉关键词 1 之后的所有内容。

如下所示，我们通过测试抓取文章后发现文章里有不希望发布的内容，**切换到 HTML 显示**，找到该内容的 HTML 代码，分别设置两个关键词即可过滤掉该内容。



如上所示，如果我们希望过滤掉上面 和 之间的内容，添加如下设置即可

如果需要过滤掉多处内容，可以添加多组设置。

九、HTML 标签过滤功能

HTML 标签过滤功能，可过滤掉采集文章中的超链接（a 标签），<script>和<style>等标签下不必要的代码。
下面是几个例子：

- 如果需要过滤掉文章中的超链接，只需输入 a 即可，是否删除标签内容 **选择否**；
- 如果要过滤掉文章中包含的<script>或<style>不必要的代码，只需输入对应标签名称，是否删除标签内容 **选择是**；

十、关键词替换、自动添加自定义超链接

可以将正文或标题中的关键词进行替换，同时也支持正文自定义超链接；

例如：需要将文章里的 wordpress 自动添加链接到 *http://wordpress.org/*，只需设置 *wordpress* 替换为 *wordpress* 即可

如下所示：

十一、解决乱码问题

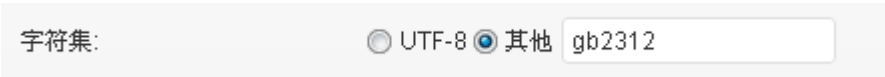
为什么产生乱码

WP-AutoPost 采集网页时默认字符集为 UTF-8，如果目标网页的字符集不是 UTF-8，采集的内容可能会出现乱码。

解决乱码问题

解决方法非常简单，只需查看目标网页的字符集，设置正确的字符集即可。

打开目标网页源代码，一般在最前面的位置，找到有 charset 的代码部分：

```
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=gb2312" />
<meta http-equiv="syssn" content="df-3r-12" />
```

如上所示，该网页的目标编码为 gb2312，设置相同的字符集即可：