ZVING
泽 元 软 件

泽元网站内容管理系统（ZCMS）
网页采集操作手册

北京泽元迅长软件有限公司

2010 年 04 月

关于本文档

ZCMS是泽元软件出品的一款基于J2EE技术和AJAX技术的企业级网站内容管理软件，旨在帮助用户解决日益复杂与重要的Web内容的创建、维护、发布和应用。本文档概要地介绍了通过ZCMS快速采集其他网站内容的方法和步骤。

读者对象

本文档的读者为ZCMS的使用者。使用者应具备以下基础知识:

- 熟悉Microsoft Internet Explorer或Mozilla Firefox的使用;
- 熟悉Windows或Linux/Unix操作系统;
- 熟悉HTML基本知识和相关的HTML页面制作方法。

用户反馈

感谢您使用泽元软件的产品。如果您发现本文档中有错误或者产品运行不正常，或者您对本文档有任何意见和建议，请及时与泽元软件联系。您的意见将是我们做版本修订时的重要依据。

联系地址

北京泽元迅长软件有限公司:

北京市海淀区学院路30号北京科技大学国家科技园D座311

邮编: 100086

电话: （010）52752668

传真: （010）52752667

Email: support@zving.com

1. ZCMS 中的 Web 采集

ZCMS 中的 Web 采集功能是一个易用的功能强大的基于模板的内容采集和提取工具，支持自动采集文章列表分页、ASP.net 分页采集、自动采集 URL 转向后的内容、自动识别内容编码、自动识别网页修改日期、多线程采集、多层级 URL 采集等特性，并支持使用代理服务器和 URL 过滤、内容过滤。

采集完成后，ZCMS 将根据匹配块中的规则，提取文章的标题、内容等信息，并自动添加到指定的内容，以便于编辑人员进一步利用。

2. 填写采集基本设置

点击菜单“采集与分发”下的“从 Web 采集”子菜单，点击“新建”按钮，可以增加新的采集任务，如下图所示:

其中:

采集类别为**文档采集**时，采集程序将直接将网页转化成 ZCMS 中的文档，如果是**自定义采集**，则只采集数据，不进行转换，需要开发程序去读取采集回来的文本，并进行处理。**自定义采集**只用于 ZCMS 的二次开发。

内容页最大采集数表示本任务最多采集多少个文章内容页。

列表页最大采集数表示本任务最多采集多少个文章列表页。

采集线程数表示同时进行采集的线程个数，此数值越大，则采集速度越快，占用带宽也越多。一般情况使用 1 个线程即可，最多不超过 30 个。

超时等待时间表示如果目标网页所在服务器忙时，采集程序等待的秒数。默认是 30 秒，一般不应超过 120 秒。

发生错误时重试次数表示如果目标服务器没有响应或者响应出错，采集程序重试的次数。

发布日期格式表示从网页内容中提示出来的发布日期的格式，与 JAVA 中的日期格式一致，以 y 代表年，M 代表月，d 代表日，h 代表小时，m 代表分，s 代表秒。发布日期将用来排序采集到的文档，发布日期较晚的将会排在前面。

采集选项中的“**下载远程图片**”被勾选的话，采集程序会自动将内容中的图片下载到 ZCMS 服务器，并替换内容中的图片地址。

采集选项中的“**去掉内容中的链接**”被勾选的话，则采集程序会自动将内容中所有超链接变成纯文本。

采集到此栏目表示采集后的文档存放到哪个栏目

如果 ZCMS 所在服务器不能直接访问互联网或者目标网页必须通过特殊代理才能访问，则需要勾选“**使用代理服务器**”选项，并填写代理服务器的地址、端口、用户名以及密码。

3. 填写 URL 规则

填写完基本设置后，即可开始填写 URL 规则，以网易通信新闻为例，可以按如下步骤进行:

1）填写起始 URL，将网易通信新闻列表页 URL 填如，如下图所示:

2）填写下一层级 URL

通过观察列表页中的新闻链接，发现大部分新闻链接 URL 都和下面这个类似：

http://tech.163.com/10/0412/10/642IP0LO0000915BE.html

我们将此 URL 转化为 URL 通配符，如下所示：

http://tech.163.com/${D}/${D}/${D}/${A}.html

其中${D}表示此处允许是数字，${A}表示允许是任意字符。

但有一部分新闻链接 URL 不符合此规则，例如：

http://popme.163.com/link/7301-0412-3150.html

我们将此 URL 也转化为 URL 通配符，如下所示：

http://popme.163.com/link/${D}_${D}_${D}.html

然后点击按钮“**增加 URL 层级**”，并将上述两上 URL 通配符填入下一层级的文本框中，如下图所示：

3）如果列表页不能直接到达文章内容页，则可能需要填多个层级的 URL。整个 URL 处理的流程是：首选采集起始 URL（起始 URL 可以有多个），然后分析起始 URL 采集回来的 HTML 文本中的所有链接 URL，一一和 2 级 URL 通配符比较，如果 URL 和 2 级 URL 通配符中有一个符合则将其采集。待符合条件的所有 2 级 URL 采集完后，从 2 级 URL 采集回来的 HTML 中再次提取所有链接 URL，一一和 3 级 URL 通配符比较……，直到最后一级 URL。

4）有时候要求过滤掉一部分 URL，则需要勾选“**过滤 URL**”选项，并填写过滤 URL 通配符，其规则和普通 URL 通配符类似。采集程序会将 URL 和过滤 URL 通配符比较，如果发现和其中的一项通配符符合，则直接忽略不采集。

4. 填写内容匹配块

填写完基础信息后，开始填写内容匹配块。

首先打开一个文章内容页面，如下图所示:

我们看到发布日期的格式是 yyyy-MM-dd HH:mm:ss，如果此格式与我们前面填写的发布日期格式不一致，则需要将此格式填入到“基础信息”选项卡的“发布日期格式”中。

然后查看网页的源代码,找到其中包括标题、发布日期和内容的部位，如下图所示:

复制包括了标题和内容的 HTML 文本到常用的文本编辑顺，将其中的标题换成${A:Title}，内容换成${A:Content},发布日期换成${A:PublishDate}，替换后的字符串如下图所示:

接下再打开一个文章内容页，查看网页源代码，用相关的字符串替换标题、内容、发布日期，然后再和前一个比较，找出所有不一致的地方（有多余的空行以及行前行后的空格数不同不算不一致，不需要处理），并用${A}替换，替换后结果如下图所示:

此处的${A}和前面填写 URL 通配符时含义相同，表示允许任意字符。

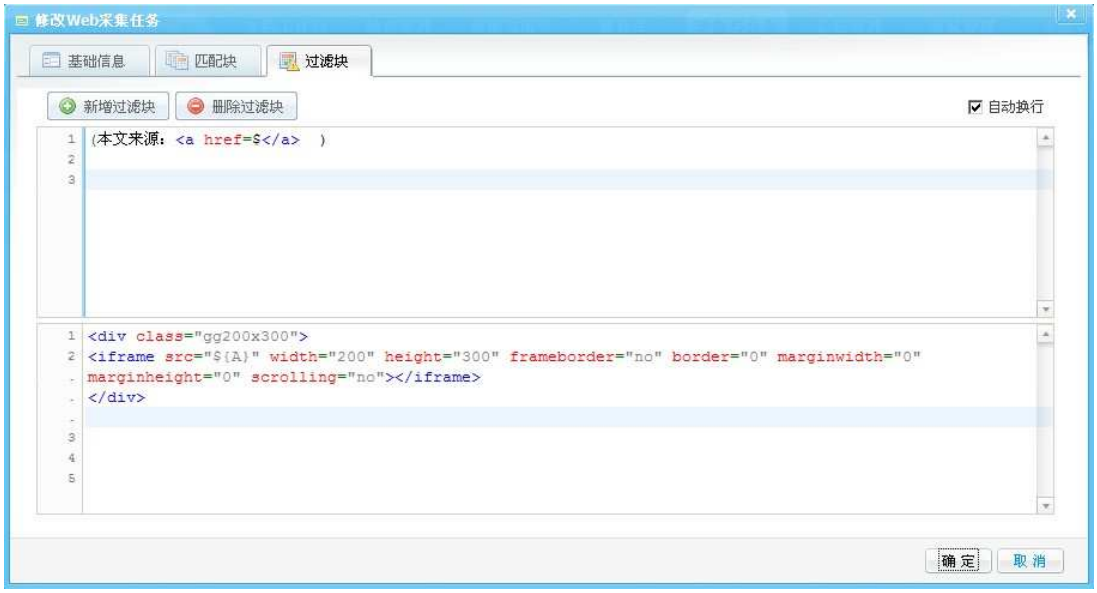
${A:TItle}中冒号以后的部分表示字段名，采集程序会将此名称和数据库中的文章表字段进行匹配。例如我们可以增加一个${A:Author}匹配符，则匹配到值会变成文章的作者字段的值。

5. 填与内容过滤块

有时候内容中可能会插入一些广告之类的不属于文章正文的文本，需要替换成字符串，因此需要填写内容过滤块。如果不需要过滤任何文本，则不需要填写此选项卡。

填写内容过滤块的规则和填写内容匹配块一样，符合内容过滤块规则的文本将会被替换成空字符串。允许填写多个过滤块，可以通过“**新增过滤块**”按钮增加新的过滤块。

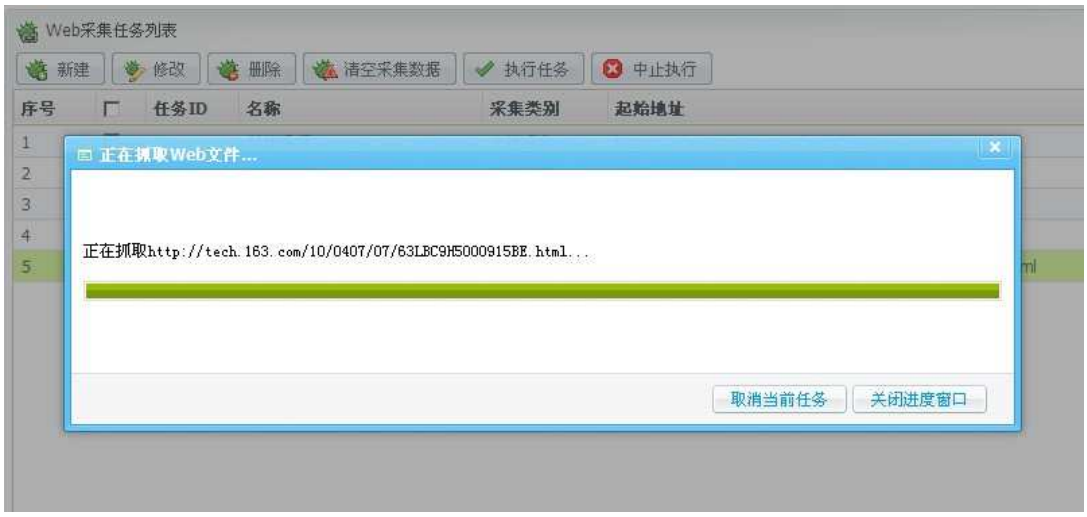
例如我们发现网页中的“本文来源”及后面的链接不需要，同时还发现有的页面中有 iframe 广告，因此我们都写入过滤块配置中，如下图所示:



6. 执行采集任务

填写完“**基础信息**”，“**匹配块**”，“**过滤块**”块后，点击“**确定**”按钮，系统将会添加一个新的采集任务，并显示在任务列表中，如下图所示：

选中刚才添加的任务，点“**执行任务**”按钮开始采集，如下图所示：

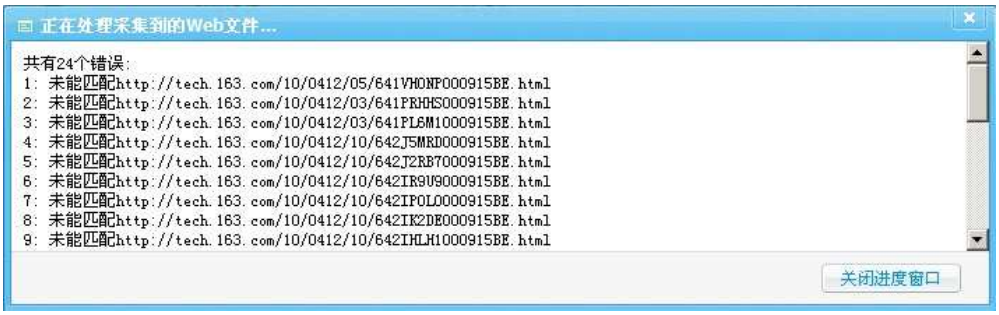


如果需要采集任务定时运行，请去“系统管理”菜单下的“定时计划”子菜单配置定时任务，如下图所示：

7. 采集后的处理

采集完成后系统会自动按匹配块中定义的规则提取文章内容和标题，并将提取成功的 URL 自动转化为指定栏目下的文章（文章状态为初稿），如下图所示：

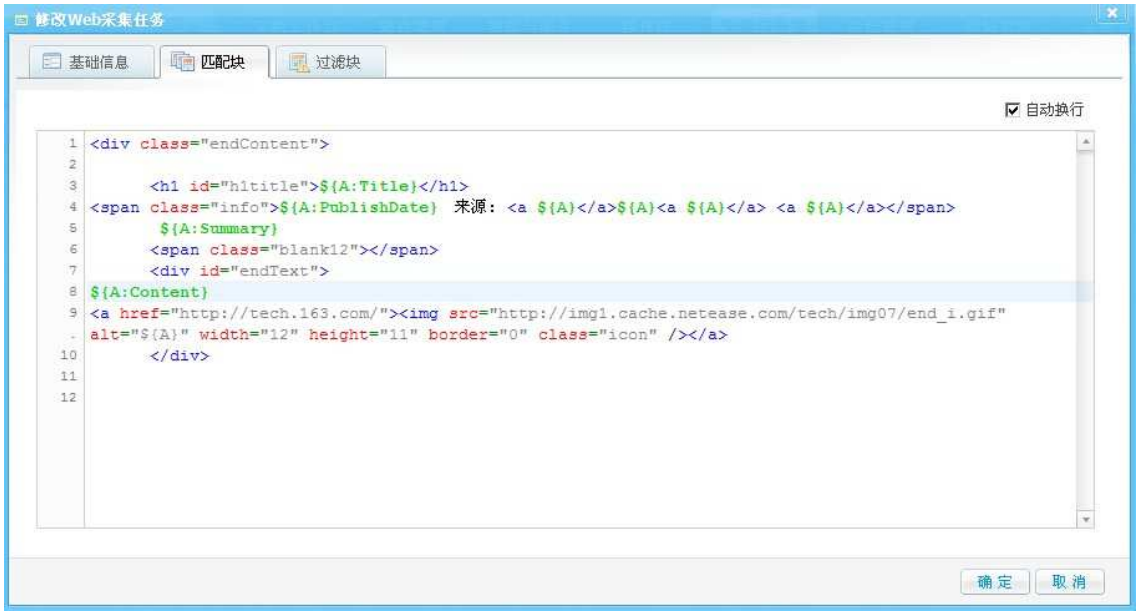
如果有未能提取成功的 URL，则会在最后显示一个未能转化的 URL 的列表，如下图所示：



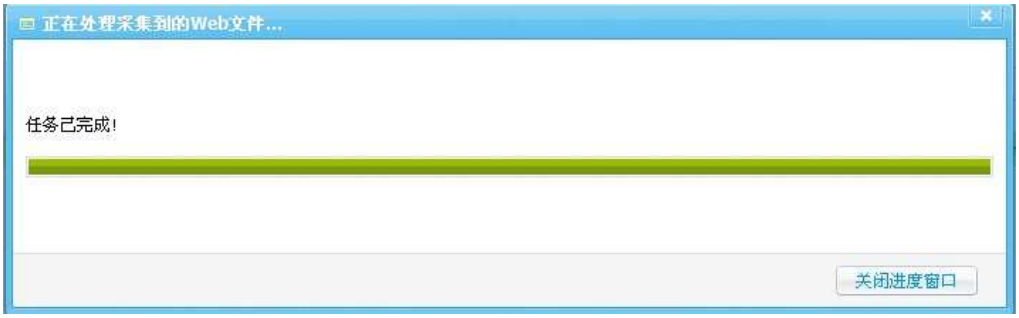
如果有未能提取成功 URL，一般是因为我们填写内容匹配块时有些情况未考虑（通常都会有一部分 URL 不能提取，除非我们特别熟悉目标网站的文章详细页的规则），这时候我们需要回头修改我们的内容匹配块，一般步骤是：

1）从未匹配的 URL 中复制一个到浏览器的地址栏，打开后查看源代码，按照填写内容匹配块的方法替换掉其中的标题，比较替换后的文本和内容匹配块的差异，例如我们打开 http://tech.163.com/10/0412/05/641VH0NP000915BE.html，如下图所示：

2）发现该页面和我们原来填写内容匹配块时不一致，多出来一块文章摘要，因此我们再次查看网页源代码，比照修改内容匹配块，以适应这种情况，修改后的内容匹配块如下所示：



3）然后点击“**处理数据**”按钮，再次运行数据提取程序。注意，此时不需要再次执行任务了，因为网页已经采集到了服务器。如果再次执行任务，将会尝试再次下载网页。再次处理的结果如下图所示：



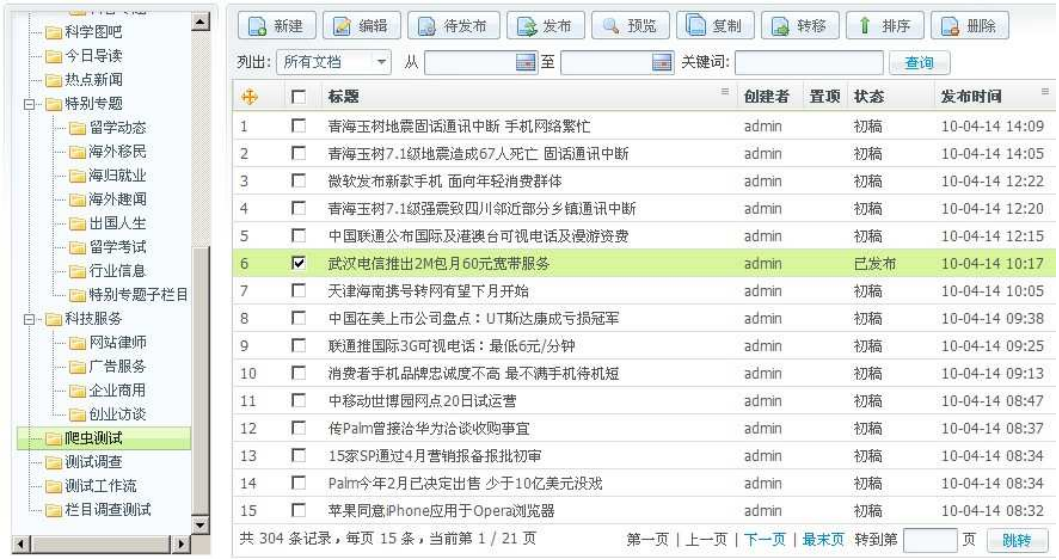
说明已全部转化为栏目下的文章，没有发生错误。

有时候可能需要多次重复这一步骤来提高匹配块的兼容性。在某些特别的情况下，各个文章内容页结构差异很大，可能需要建立多个采集任务才能将同一个 URL 下的所有文章转入指定栏目下。

同样地，过滤块也可能有些情况没有考虑到，导致过滤不完全，需要按照和内容匹配块类似的方式进行修改。

8. 采集效果

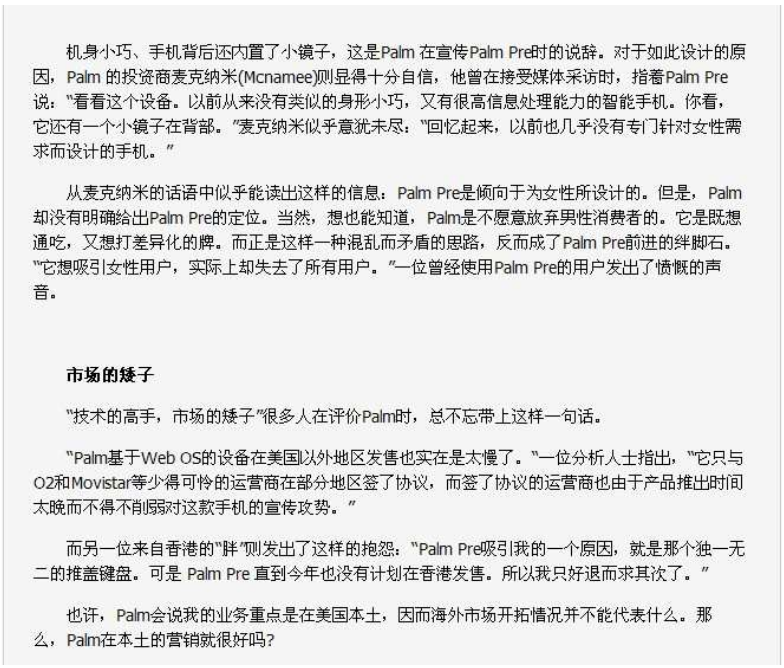
经过以上步骤后，目标网站上的文章数据将会出现在指定的栏目下，如图所示：



如果勾选了“**下载远程图片**”，则会自动将图片下载，并加入到图片库中，如下所示：

如果目标网页文章内有分页，则会自动分并成一篇文章，如下图所示:

原始网页



机身小巧、手机背后还内置了小镜子，这是Palm 在宣传Palm Pre时的说辞。对于如此设计的原因，Palm 的投资商麦克纳米(Mcnamee)则显得十分自信，他曾在接受媒体采访时，指着Palm Pre说："看看这个设备。以前从来没有类似的身形小巧，又有很高信息处理能力的智能手机。你看，它还有一个小镜子在背部。"麦克纳米似乎意犹未尽："回忆起来，以前也几乎没有专门针对女性需求而设计的手机。"

从麦克纳米的话语中似乎能读出这样的信息：Palm Pre是倾向于为女性所设计的。但是，Palm却没有明确给出Palm Pre的定位。当然，想也能知道，Palm是不愿意放弃男性消费者的。它是既想通吃，又想打差异化的牌。而正是这样一种混乱而矛盾的思路，反而成了Palm Pre前进的绊脚石。"它想吸引女性用户，实际上却失去了所有用户。"一位曾经使用Palm Pre的用户发出了愤慨的声音。

市场的矮子

"技术的高手，市场的矮子"很多人在评价Palm时，总不忘带上这样一句话。

"Palm基于Web OS的设备在美国以外地区发售也实在是太慢了。"一位分析人士指出，"它只与O2和Movistar等少得可怜的运营商在部分地区签了协议，而签了协议的运营商也由于产品推出时间太晚而不得不削弱对这款手机的宣传攻势。"

而另一位来自香港的"胖"则发出了这样的抱怨："Palm Pre吸引我的一个原因，就是那个独一无二的推盖键盘。可是 Palm Pre 直到今年也没有计划在香港发售。所以我只好退而求其次了。"

也许，Palm会说我的业务重点是在美国本土，因而海外市场开拓情况并不能代表什么。那么，Palm在本土的营销就很好吗?

采集后自动合并内容

同时我们也可以看到，采集后的内容中已经没有了网易上的广告了。