

# 从特定网址列表采集新闻信息的使用说明

## 1 登录后台

管理员身份登录进入网站管理后台

## 2 创建采集信息存放的栏目

新闻目录节点，是指采集到的新闻，存放于本站哪个栏目下，也可以新建一个栏目用于存放采集到的新闻，点网站运维，然后点首页，然后点创建文件夹，



在新建的节点主题处输入名称：

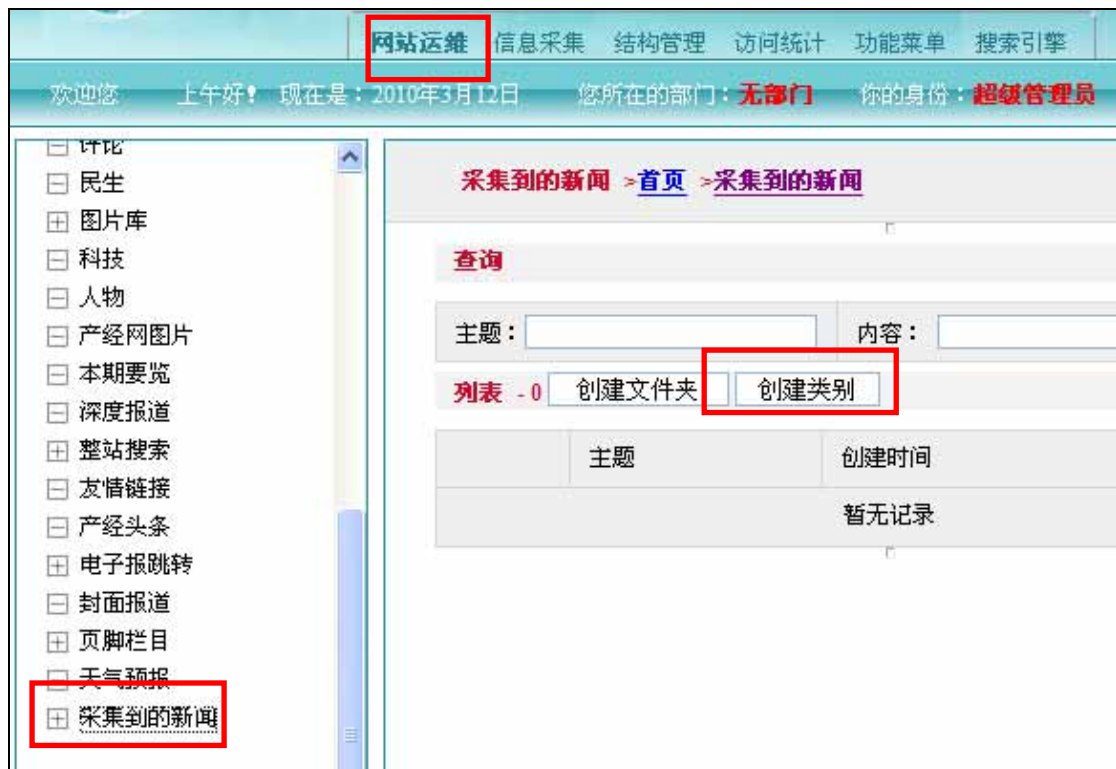
The screenshot shows the '创建文件夹' (Create Folder) dialog box. It has a '主题:' (Topic) input field with the text '采集到的新闻' (Collected News) entered. Below the input field, there are two buttons: '下一步' (Next Step) and '完成' (Finish).

然后点完成

这样我们就创建了一个存放采集到新闻的文件夹，这个文件夹位于首页下面。



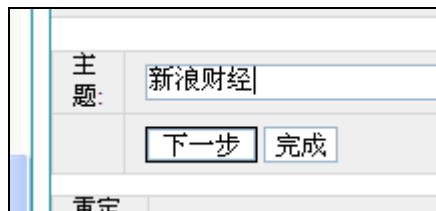
重新点击网站运维菜单：



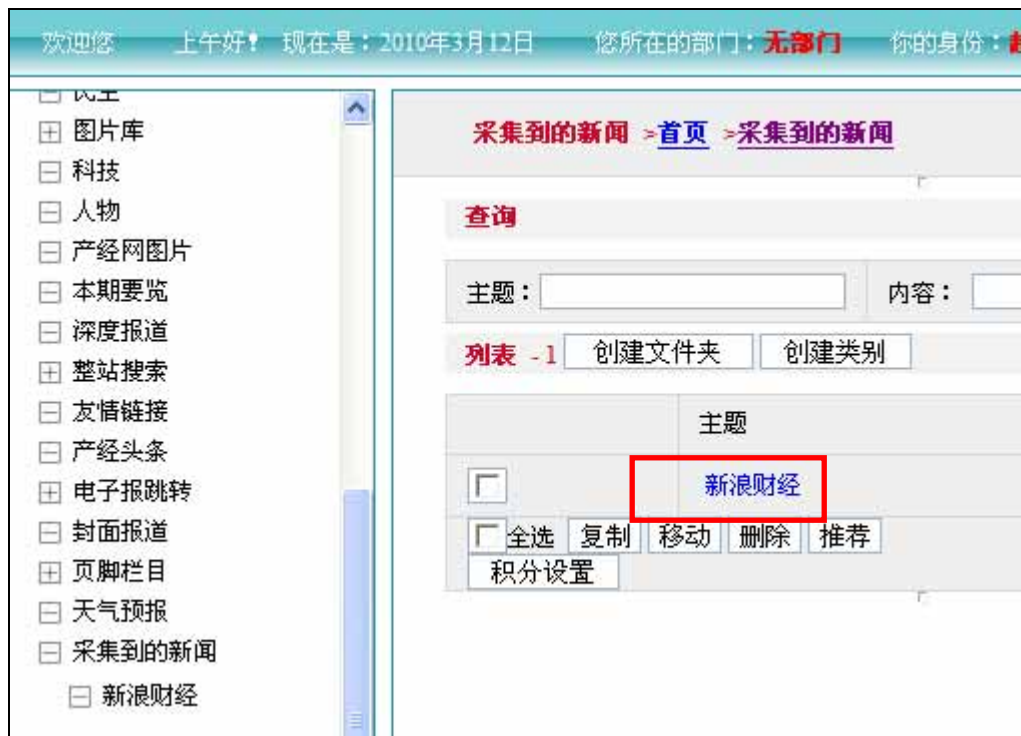
在左侧会出现新创建的采集到的新闻子栏目

点击进入这个栏目

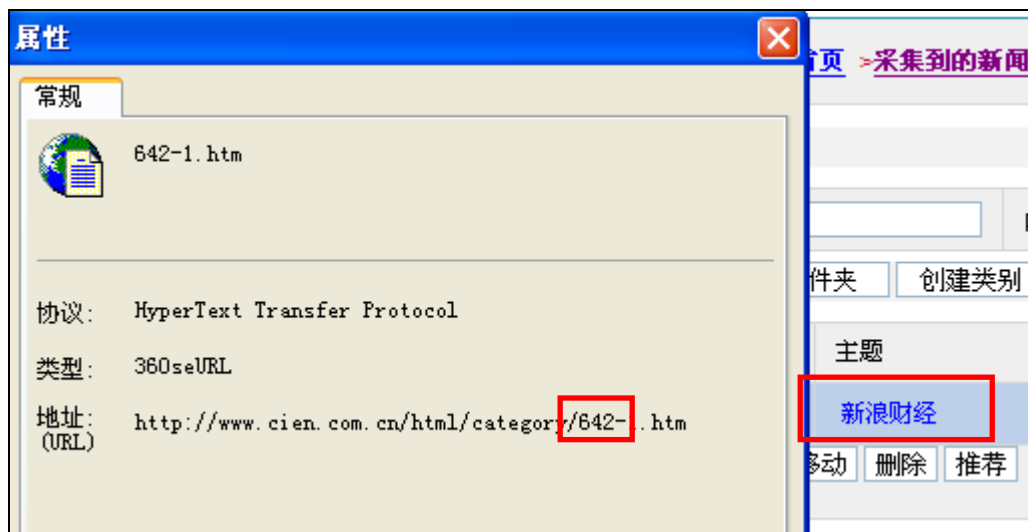
点创建类别，可以在这个文件夹下分类放置我们要将要采集的新闻了



比如我们建一个用来存放将要采集的新浪财经新闻的类别，在点击创建类别后，主题中输入新浪财经，点完成即可。



创建好的新浪财经的通过查看链接属性可以看到节点号



如上节点号为: 642

### 3 采集设置

找到信息采集菜单



点击采集设置中的新建

**信息采集设置**

**新建**

名称	采集时间	操作
----	------	----

**采集信息设置**

采集名称	<input type="text"/>
新闻目录节点	<input type="text"/>
网站编码	<input type="text" value="qb2312"/>
采集数量	<input type="text" value="20"/>
采集站点连接	<input type="text"/>
图片保存路径	<input type="text"/>
列表开始代码	<input type="text"/>
列表结束代码	<input type="text"/>
连接开始代码	<input type="text"/>
连接结束代码	<input type="text"/>
标题开始代码	<input type="text"/>
标题结束代码	<input type="text"/>
发布时间开始	<input type="text"/>
发布时间结束	<input type="text"/>
新闻来源开始	<input type="text"/>
新闻来源结束	<input type="text"/>
新闻正文开始	<input type="text"/>
新闻正文结束	<input type="text"/>
<input type="button" value="提交"/>	

采集名称，可以随意起一个名称，比如：新浪财经，  
新闻目录节点处填写之前创建好的新浪财经的存在栏目节点号——642

采集信息设置	
采集名称	新浪财经
新闻目录节点	642
网站编码	qb2312
采集数量	20
采集站点连接	<a href="http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml">http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml</a>
图片保存路径	/res/Home/1003/
列表开始代码	

网站编码输入该网站的编码方式

采集信息设置	
采集名称	新浪财经
新闻目录节点	642
网站编码	qb2312
采集数量	20
采集站点连接	<a href="http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml">http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml</a>
图片保存路径	/res/Home/1003/
列表开始代码	

采集数量写 20，表示每次采集前 20 个以前没采集过的。

采集信息设置	
采集名称	新浪财经
新闻目录节点	642
网站编码	qb2312
采集数量	20
采集站点连接	http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml
图片保存路径	/res/Home/1003/
列表开始代码	

采集站点连接输入新闻列表页 如：

http://roll.finance.sina.com.cn/finance/zq1/gsjy/index.shtml

访问列表如下图：

- [股市在线：2月24日股市早班车](#) (02月24日 08:55)
- [八大机构：短期调整基本到位](#) (02月24日 07:29)
- [金百灵投资：主题投资活跃市场 反抽趋势或延续](#) (02月24日 07:00)
- [午后盘升有乾坤 突破方向仍是谜](#) (02月24日 04:07)
- [无惧调整 反弹还将继续](#) (02月24日 03:16)
- [两会概念预热 三大主线孕育机会](#) (02月24日 03:15)
- [中国平安领跌 题材继续唱戏](#) (02月24日 01:44)
- [申银万国：先抑后扬 企稳回升](#) (02月23日 19:58)
- [视频：势之所趋 利之所在](#) (02月23日 18:06)
- [视频：平安减持能否改变市场走势](#) (02月23日 18:05)

图片保存路径 设置本机存在的虚拟路径，如在 res 下建立一个 temp 目录  
该路径填写 /res/home/1003

采集数量	20
采集站点连接	http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml
图片保存路径	/res/Home/1003/
列表开始代码	<div class="hs01"></div> <ul class="list_009">

采集信息设置	
采集名称	新浪财经
新闻目录节点	642
网站编码	qb2312
采集数量	20
采集站点连接	http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml
图片保存路径	/res/Home/1003/
列表开始代码	

列表开始代码数据新闻列表开始的代码：查看欲采集网页的源代码在列表开始前始代码。

```

<div class="hs01"></div>
<ul class="list_009">
    <li><a
href="http://finance.sina.com.cn/stock/jsy/20100224/08557447093.shtml"
target="_blank">股市在线：2月24日股市早班车</a><span>(02月24日 08:55)</span></li>
    <li><a
href="http://finance.sina.com.cn/stock/jsy/20100224/07297446427.shtml"
target="_blank">八大机构：短期调整基本到位 </a><span>(02月24日 07:29)</span></li>
    <li><a
href="http://finance.sina.com.cn/stock/jsy/20100224/07007446078.shtml"
target="_blank">金百灵投资：主题投资活跃市场 反抽趋势或延续</a><span>(02月24日
07:00)</span></li>

```

从图中我们可以看到列表开始的代码是：<div class="hs01"></div>

<ul class="list\_009"> ,最好是在源文件里面查一下是否这句代码是否是唯一的。如果是唯一的,则可以在设置条件的框里面填上。如果不是唯一的,则可以扩大代码的范围,一定要保证代码的唯一性。条件框里输入的数据要从网页原代码中复制,将前后的空格和回车去掉。如很难保证唯一性,要保证该代码在列表前首次出现。

采集数量	20
采集站点连接	http://roll.finance.sina.com.cn/finance/zq1/qsisv/index.shtml
图片保存路径	/res/Home/1003/
列表开始代码	<div class="hs01"></div> <ul class="list_009">



从图上看新闻标题前面是:

```
<h1 id="artibodyTitle" pid="31" tid="1" did="7446427" fid="1554">
```

但此处需要注意的是 pid="31" tid="1" did="7446427" fid="1554"这些代码在每个新闻中可能不同，所以标题开始只填能表示该标题前的标签的一部分即可：

此处填<h1 id="artibodyTitle"

结束 代码填写</h1>

连接结束代码	<input type="text" value="&lt;/a&gt;"/>
标题开始代码	<input artibodytitle"="" type="text" value="&lt;h1 id="/>
标题结束代码	<input type="text" value="&lt;/h1&gt;"/>

发布时间、来源保证开始结束的代码唯一，并保证每条新闻的开始和结束相同

发布时间开始	<input date"&gt;"="" pub="" type="text" value="&lt;span id="/>
发布时间结束	<input type="text" value="&lt;/span&gt;"/>
新闻来源开始	<input media="" name"&gt;"="" type="text" value="&lt;span id="/>
新闻来源结束	<input type="text" value="&lt;/span&gt;"/>

发布时间、来源，如果不填，就是采集不到这两项，不会影响其他项的采集。

正文内容保证开始代码唯一，保证结束代码在开始代码后首次出现。

如开始代码填写 <!-- 正文内容 begin -->

结束填写 <div class="blkComment

新闻正文开始	<input type="text" value="&lt;!-- 正文内容 begin --&gt;"/>
新闻正文结束	<input blkcomment"="" type="text" value="&lt;div class="/>
<input type="button" value="提交"/>	

然后点提交

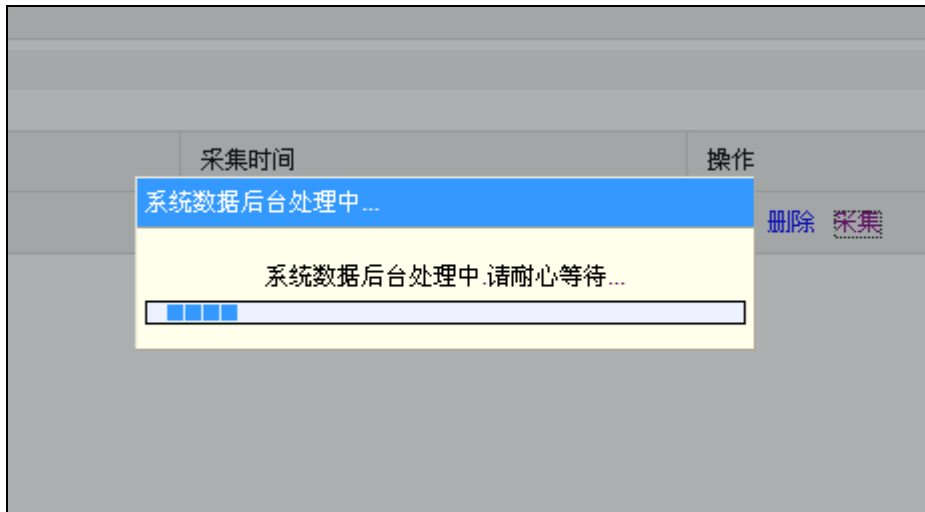
完成采集站点的设置

#### 4 信息采集

信息采集设置		
新建		
名称	采集时间	操作
新浪股市及时雨	未采集	<a href="#">编辑</a> <a href="#">删除</a> <a href="#">采集</a>

建好的站点可以编辑，可以删除，可以进行信息采集

点采集，即可按设置好的条件把信息采集到新浪财经栏目（642 节点）下，



新采进来的信息为未审核状态，对新浪财经栏目（642 节点）

采集信息执行					
采集到的信息					
共采集到新信息4条，采集到的信息已放置到“新浪财经”栏目下，点击 <a href="#">这里</a> 去进行管理。 <a href="#">返回</a> 继续采集其他信息					
标题	发布时间	来源	正文		
新浪财经3月15日收盘播报文字实录	2010-3-15 22:56:47	新浪财经	点击查看最新行情 您创造财...	各位新浪网友大家好!	主持人乔旒：为
新浪财经3月15日收盘播报文字实录	2010-3-15 22:56:47	新浪财经	点击查看最新行情 您创造财...	各位新浪网友大家好!	主持人乔旒：为
两市缩量下跌 后市仍有探底	2010-3-15 22:56:47	申银万国	点击查看最新行情	申银万国证券研究所 钱启敏	盘中...
3000点关口失守后应如何应对	2010-3-15 22:56:47	中国证券网	点击查看最新行情 股大面积下跌...	提要：受加息预期冲击，周一深沪大盘在个	

如果信息源从上次采集至今没有新闻更新，则会提示没有采集到信息

采集信息执行
采集到的信息
源站点没有信息更新，未能采集到新信息。 <a href="#">返回</a> 继续采集其他信息

## 5 管理采集到的信息

点击链接去管理可直接去管理采集到的信息，  
也可以之后通过网站运维菜单

欢迎您 上午好! 现在是: 2010年3月12日 您所在的部门: 无部门 你的身份: 超级管理员

首页 > 首页

查询

主题:  内容:  GO

列表 - 30 创建文件夹 创建类别

	主题	创建时间	
<input type="checkbox"/>	采集到的新闻	2010-03-12	编辑 删除

找到采集到的新闻  
再找到下面的新浪财经栏目

网站导航 信息采集 结构管理 访问统计 功能菜单 搜索引擎

欢迎您 夜里好! 现在是: 2010年3月15日 您所在的部门: 无部门 你的身份: 超级管理员

新浪财经 > 首页 > 采集到的新闻 > 新浪财经

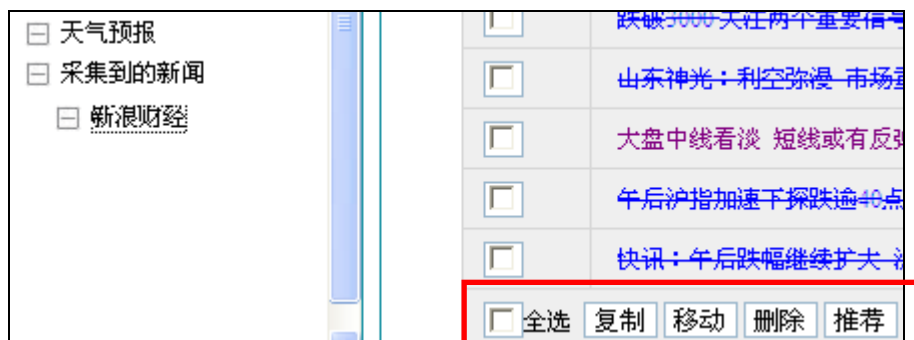
查询

主题:  内容:  GO

列表 - 59 创建

	主题	创建时间	
<input type="checkbox"/>	3000点关口失守后应如何应对	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	两市缩量下跌 后市仍有探底	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	新浪财经3月15日收盘播报文字实录	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	沪指跌逾1% 失3000点大关 逼近年线	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	权重板块领跌 资金净流出约61亿	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	升值预期或趋强 如何布局	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	临近牛熊分界线 主力如何博弈	2010-03-15	编辑 删除 批准 拒绝
<input type="checkbox"/>	加息预期影响可能越来越弱	2010-03-15	编辑 删除 批准 拒绝

新采集到的信息默认是未审核的（未审核的信息在前台不显示），有审核权限的人进入后审核采集到的信息，  
也可以对采集到的信息进行编辑修改，或删除。  
并可移动，复制信息到其它栏目，



也可以通过编辑后采用一稿多投的形式，让采集到的信息显示到指定的位置。

