

文章采集



用户手册



西安博达软件股份有限公司



目 录

1.		从网页系	采集功能说明	5
	1.1	从网页	页采集概述	5
	1.2	从网页	页采集主要功能特点	5
2.		从网页采	采集使用说明	7
	2.1	任务设	设置	7
		2.1.1	导入模板	7
		2.1.2	添加任务	
		2.1.3	删除任务	
		2.1.4	启用任务和停用任务	
		2.1.5	设置计划	
		2.1.6	采集记录管理	
	2.2	任务哲	控制台	
	2.3	任务日	日志	
	2.4	回收站	站	
3.		从数据戽	车采集管理的功能说明	
	3.1	从数据	据库采集管理概述	
	3.2	任务设	设置	
		3.2.1	基本信息设置	
		3. 2. 2	采集信息设置	
		3.2.3	属性映射设置	
		3.2.4	修改任务	
		3.2.5	删除任务	
		3.2.6	启用任务和停用任务	
		3. 2. 7	设置计划	
		3.2.8	采集记录管理	
	3.3	任务哲	控制台	
	3.4	任务日	日志	

西安博达软件股份有限公司 XI'AN WEBBER SOFTWARE CO., LTD. 地址: 中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 2 页 共 41 页



	3.5	回收站	41
1.		从网页采集功能说明	5
	1.1	从网页采集概述	5
	1.2	从网页采集主要功能特点	5
2.		从网页采集使用说明	7
	2.1	任务设置	7
		2.1.1 导入模板	7
		2.1.2 添加任务	
		2.1.2.1 认证	8
		2.1.2.2 基础信息	12
		2.1.2.3 列表页	13
		2.1.2.4 内容页	17
		2.1.2.5 附加	20
		2.1.2.6 RSS 文章列表采集任务设置	26
		2.1.3 删除任务	
		2.1.4 启用任务和停用任务	
		2.1.5 设置计划	29
		2.1.6 采集记录管理	29
		2.1.6.1 重置采集状态	30
		2.1.6.2 删除采集记录	30
		2.1.6.3 清空采集记录	31
	2.2	任务控制台	31
	2.3	任务日志	31
	2.4	回收站	32
3.		从数据库采集管理的功能说明	33
	3.1	从数据库采集管理概述	33
	3.2	任务设置	33
		3.2.1 基本信息设置	35
		 3. 4. 4 不果 信 思 叹 直 3. 9. 9. 9 6. 日本 7. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10	
		3.2.4 修改任条	<i>ا</i> د ۶۶
		3.2.5 删除任务	
		3.2.6 启用任务和停用任务	
		3.2.7 设置计划	39
西3 XI'AI	を博え N WEB	大软件股份有限公司 BER SOFTWARE CO., LTD. 地址: 中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:7 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.	10077 com



	3.2.8	采集记录管理	39
3.3	任务	控制台	40
3.4	任务	日志	40
3.5	回收	站	41



文章采集

1. 从网页采集功能说明

1.1 从网页采集概述

从网页采集,是为采集人员方便从不同的 Internet 站点获取所需的文章信息而定制的系统功能。

此系统可以在最短时间内,帮您把您关注的 Internet 站点上最新的文章实时采集,并根据您的需求提取对应的文章标题、发布时间、详细内容等页面信息,并将采集后的文章存储 在本地数据库中。

该系统功能可以将所跟踪网站的新闻资讯自动与后台进行同步更新,以您所定制的方式 在对文章进行分类和统一格式后,第一时间传递并展现给最需要它的终端用户,使得用户可 以方便的浏览到最新的社会动态,从而将网络传媒的文章实时性推向了极致。

1.2 从网页采集主要功能特点

从网页采集,具有智能、可定制、高扩展性的采集技术。该功能可以快速大量的收集互 联网信息,为用户提供最准确、最广泛、最具时效性的信息提供了坚实基础。它所具备的主 要功能如下:

1. 强大的从网页采集能力

此系统功能针对专业用户所要求的信息搜索深度、采集精度和采集速度等进行了专门的 优化,采用了分布式多线程并发指令执行体系结构,可以把采集目标的文字或者图片或者连 接地址的目标文件都可以采集到我们自己的网站上。

2. 智能化信息提取技术

当用户在最短时间内获取了其需要的海量的信息页面,其处理工作也就可想而知了,此 系统功能不但能在瞬间获取你所要的页面,还能快速同步的进行页面分析提取。具体特点如 下:

支持按位置提取、按关键字提取和按表单提取等多种不同的智能化信息提取技术,保证 对不同网站构建技术的信息提取通用性;

支持对信息页面的标识,及对信息页面中用户关注的信息内容字段的精确定义,使得信 息提取能够高效地获取用户所需的内容,并以结构化的数据项形式直接存入数据库,具有开



放性;

信息提取不依赖于具体的信息内容,支持多个不同信息类型的不同采集任务同时运行; 提供快捷的方式使得操作人员能够对该任务的提取结果进行浏览;

3. 方便化的信息管理技术

此系统提供定时性的任务采集技术,不需要用户手动的进行采集。同时还提供文章信息 的方便管理,使文章实时性的发布,保证文章的实效性。



文章采集

2. 从网页采集使用说明

从网页采集,主要应用在 VSB 网站群管理平台中。在使用此功能前,必须确保该站点中 含有相应的栏目文章源。在网站群管理平台的网站管理下,可以找到文章采集功能。

从网页采集,拥有:任务设置、任务控制台、任务日志、计划管理、回收站五个功能。 任务设置中有采集文章信息的相关设置;任务控制台中有任务的启动与停止;计划管理是设 置采集任务的定时采集文章计划;任务日志主要记录采集任务执行的操作日志;回收站存放 删除过得采集任务。

2.1 任务设置

在网站管理界面,点击"文章采集>>从网页采集>>任务设置",进入界面。如图 2-1:

	网站建设 网站管理		🔧 切换到公开	目录系统 📑切换到内容管理 📢	🕅 我的新消息 0) 📜 云文件库 🔟 退出
	备份恢复 用户和权限	网站统计 网站安全	文章采集 更多>>	应 预览网站>> 📔 🛃 发布网站	🦉访问网站 🗍当前网站: 🗣 wyx
文章采集	从网页采集				
₽ •	💊 当前位置:文章采集>	>从网页采集>>任务设置			
	导入模板 添加任务	删除任务 启用任务	§ 停用任务 设置	计划 复制 粘贴 所有栏目 ▼	所有任务状态 ▼
一一一日子拉制台	共0条 每页 15 ▼条 0/0	转到页			查询
	🔲 采集任务名称	任务状态 栏目名称		采集地址	采集记录条数 操作
	没有设置采集任务,请点	击 [添加任务] 按钮增加家	100采集任务。		

图 2-1

任务设置页面以列表形式显示了所有设置的采集任务。列表包括采集任务名称、任务状态、文章存放的栏目名称、采集地址、采集记录条数以及操作。

任务设置页面还包含了:导入模板、添加任务、删除任务、启用任务、停用任务、设置 计划等共10个功能项。

2.1.1 导入模板

在本系统中提供了下载采集模板功能,文章采集者可以将一些设置好的采集任务定制成 模板并下载保存到本地。当以后再使用时,直接上传导入就可以了。具体导入操作如下:



点击"导入模板"按钮,进入配置页面。如图 2-2:

	网站建设 网站管理 回切期	剩内容管理 🍃文件库 🗾退出
	备份恢复 用户和权限 网站统计 网站安全 文章采集 更多>> 🗾 预览网站 🛃 发布网站 🦉访)	可网站 🍵当前网站: 📬 demo 👻
文章采集	从网页采集	
₽ 文章采集	😪 当前位置:文章采集>>从网页采集>>任务设置>>导入模板	ø
	上传采集任务模板,可以略过繁琐的配置过程,快速建立自己的采集任务。	
·····································	任务名称 *	
	导入内容放入 选择栏目文章源 *	
	上传任务模板 浏览	
■ ■ 从数据库采集		
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一		
一一一一一日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日		

图 2-2

【任务名称】为导入的任务所定义的名称。

【导入内容放入】在此处可以选择采集下来的文章所放置的栏目文章源。

【上传任务模板】上传采集任务模板。

当这三项设置完后,点击" 导入模板 "。

2.1.2 添加任务

当采集者需要添加新的采集任务时,可以点击"^{添加任务}",在弹出的页面中添加采集 任务。文章采集的任务设置分为五步(认证、基础、列表页、内容页、附加),填写任务时必 须逐步填写。其中每一步中凡是有带*符号的填写项都为必填项。

2.1.2.1 认证

在建立采集任务的第一步认证页面中,其需要填写项。如图 2-3:



	网站建设 网站管理	📑 切換到内容管理 🚔文件库 💌 退出
	备份恢复 用户和权限 网站统计 更多>> 应 预览网站 🛃发布网站	🧭访问网站 🗍当前网站: 🗣 demo 🔻
文章采集 文章采集 一一一八小四页采集 一一一八小四页采集 一一一八小四页条集 一一一八小四页条集 一一一八小四页条集 一一一八小四页条集 一一一八小四页条集 一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八小四页条集 一一一一八四页条集 一一一一八四页条集 一一一一八四页条集 一一一一八四页条集 一一一一八四页条集 一一一一八一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	 从网页采集 当前位置:文章采集>>从网页采集>>任务设置>>第1步:认证 第1步:认证 第2步:基础 第3步:列去页 第4步:正文页 任务名称 * ● 启用 ● 停用 采集内容放入 法择栏目文章源 * 采集网站编码 GB2312/GBK ▼ 调试模式 ● 正常模式 ○ 调试模式 (采集时会输出更多的调试信息) 代理服务器: ● 不设置 ○ 设置代理服务器 是否需要认证: ● 不需要认证 ○ 需要认证 ○ 需要认证 	第5步 附加 保存 返回

图 2-3

【任务名称】本次文章采集任务的任务名称。当此任务在"启用"状态下,执行采集任务时,会执行此项任务;当在"停用"状态下,执行采集任务时,此项任务不执行。

【采集内容放入】在此处可以选择把采集下来的文章放到那个栏目文章源中,栏目文章 源的选择是根据用户的需求来定。选择时点击"选择栏目文章源",会弹出一个选择栏目文 章源页面,如图 2-4,在其页面中选择所要的栏目文章源,然后点击"选择栏目文章源"。 执行完采集任务后,所采集的文章就直接存放到指定的栏目文章源中了。



文章采集

选择栏目文章源	关闭
会約	
ᇦ₀┫康生活	
🖶 🛅 内地新闻	
🖶 🛑 综述分析	
●●●●中国军情	
● — — ● 美洲经济	
● — — — — — — — — — — — — — — — — — — —	
● — — — — — — — — — — — — — — — — — — —	
● <mark>●</mark> IT业界	
●─────────────────────	

图 2-4

【是否需要认证】这里有两种选择:不需要认证,需要认证。默认情况下是不需要认证 的,如果您需要采集的信息是需要登录认证的,这时必须选择需要认证。如图 2-5:

	网站建设	网站管理				🙎 切换到公开目录系统
	备份恢复 用	户和权限 区	羽站统计 网站	占安全 文章采集	┋ │ 其他设置	预览网站>
文章采集 ₽文章采集	────────────────────────────────────		、网页采集>>任务	;设置>>第1步: 认证	E	
	第1步: 认证	第2步:基	基础 (第3 步)	: 列表页 第4	步:正文页 第5步:	附加 保存
	任务名称		* (自用(停用		
	采集内容放入		选	择栏目文章源 🔹		
	采集网站编码	GB2312/GBK	✓			
	采集设备类型	◉电脑◯手	机OPAD			
	调试模式	●正常模式)调试模式 采算	町会輸出更多的 调	1试信息)	
	代理服务器:	●不设置○	设置代理服务器			
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	是否需要认证:	◉不需要认证	EO需要认证			
———————————————————————————————————						

图 2-5

【认证提交方式】: 提交方式分为两种分别是 POST 和 GET, 怎么样来选择这两种类型, 需要查看登录页面的源文件, 在登录页面右键选择查看源文件, 在代码页中查找 method, method 后对应的信息就是我们选择认证提交方式的依据, 例如 method="post", 那么我们就 选择 POST 。



【登录页面地址】: 在此处填入所要登录系统的登录页面地址, 例如: http://192.168.120.1:8888/login_page.php。

【认证页面地址】: 这里填写的和登录页面地址一样就可以了。

【登录页表单序号】:登录页面的表单序号就是在登录页面中登录部分所处的位置,如果 是第一个 form 就填写 1,如果是第二个 form 就填写 2,然后进行分析登录页,分析后会给出 登录需要的相关信息,在这里可以填写登录的相关信息,例如登录的用户名和密码。如果分 析不出来相关信息,也可以进行手动添加,"添加提交参数"可以进行到图 2-7 页面,也可 以对不必要的信息进行删除。如图 2-6:

	网站建设	网站管理	🔧 切换到公开目录系统			
备份恢复 用户和权限 网站统计 网站安全 <mark>文章采集</mark> 其他设置						
文章采集	从网页采集					
₽ 2	🔩 当前位置:	文章采集>>从网页采集>>任务设置>>第1步:认证				
	留1些,认证		但方			
	951 2V · MARE		म मा			
	任务名称	新闻中心 * 💿 启用 🔾 停用				
	采集内容放入	集团要闻 选择栏目文章源 *				
一一同计划管理	采集网站编码	GB2312/GBK 🗸				
	调试模式	● 正常模式 ○ 调试模式 (采集时会输出更多的调试信息)				
🖻 🔂 从数据库采集	代理服务器:	● 不设置 ○ 设置代理服务器				
任务设置	是否需要认证:	○ 不需要认证 ④ 需要认证				
	认证提交方式:	• POST O GET				
1211分日志	登录页面地址	http://192.168.120.1:8888/login_page.php *				
	认证页面地址	login.php *				
	登录页表单序号	1 分析登录页面 添加提交参数				
	参数名称 参数	汝值				
	username	刪除参数				
	password	刪除参数				
	perm_login	刪除金数				

图 2-6

名称	
属性	文本 🗸
值	
	保存取消
	图 2-7

当所需的项目设置完后,点击" 保存",然后进入第二步基础信息的设置。



2.1.2.2 基础信息

网站建设 网站管理 条统 副切						
1	备份恢复 用户	和权限 丨 网站安全 丨 文章采集 丨 其他设置 👘 🗾 预览网站>> 📔 🛃				
文章采集	从网页采集					
₽文章采集	💊 当前位置:	文章采集>>从网页采集>>任务设置>>第2步:基础				
	第1步:认证					
	任务名称					
	采集列表类型:	◎ 文章列表页面: 您采集的是包含文章列表的页面 ○ RSS 链接: 您采集的是RSS 格式的XML链接, 参考新浪的RSS频道				
·····································	采集列表地址	webber.com/list.jsp?urltype=tree.TreeTempUrl&wbtreeid=12529 * 预览列表页				
■ ■ 人数据库采集	任务详述	\bigcirc				
	包含关键词	<mark>说明:标题或正文包含</mark> 所设置关键词,多个 <u>关键词之间以","分隔</u> 必须包含所有关键词 ▼				
	不包含关键词	<mark>说明:标题或正文不能包含</mark> 所设置关键词,多个关键词之间以","分隔				
	内容位置	內容存入本地数据库 〇 链接到文章来源,不在本地保存內容				
	每次最多采集	5 条 -请选择条数>				
	采集顺序	从列表后面往前采集(默认) 🗸				
	采集记录	保留最近10 天 ,至少保留200 条采集记录 说明:已记录的采集地址不会被重新采集				

在建立采集任务的第二步基础信息页面中,其需要填写的项如图 2-8:



【任务名称】是在第一步填写的名称。

【采集列表类型】设置所要采集的文章类型,此处有两种选择类型,一种是采集文章列表页面,例如新浪网站上的文章栏目页面。另一种是 RSS 链接,这种采集列表类型采集的是 RSS 格式的 XML 链接,例如新浪网站中的 RSS 频道(http://rss.sina.com.cn/)。

【采集列表地址】此地址是采集某个网站中某个栏目下的文章所在的页面地址,此处地址的填写是和【采集列表类型】中选择的文章类型一一对应的。

例如:当选择了文章列表页面类型时,我们可以以新浪网站为例,选择采集其站中的文章栏目中的文章,把该栏目页面的链接地址填写到采集列表地址栏中;如果选择了 RSS 链接类型时,我们以新浪网站中的 RSS 频道为例,选择科技文章栏目下的焦点文章,把其 RSS 格式的 XML 链接地址填写到采集列表地址栏中。当填入了地址以后,可以点击"顶览列表页" 按钮进行预览所要采集的文章列表。

【任务详述】此处是填写关于此项采集任务的基本概况信息。

【包含关键词】在此处填入所要采集的文章标题或正文中所设置的关键词,当采集时, 会把包含此关键词的文章内容采集下来。

【不包含关键词】当采集者需要屏蔽带有某些关键词的文章时,只要把这些关键词填入



其中,采集文章时,具有这些关键词的文章就会不被采集。

【内容位置】此处有两个选项,一种是把采集下来的文章内容存入到本地数据库。一种 是把采集的文章直接链接到文章来源,不在本地保存文章内容。

【每次最多采集】在此处可以选择你每次所要采集文章的条数。

当所需的项目设置完后,点击" 保存",然后进入第二步列表页的设置。

2.1.2.3 列表页

在第三步列表页的设置页面中,主要是对从网页采集者所采集的特定文章列表部分的定位,此定位是通过此页面的源代码中的特定标记来设定的。需要注意的是:这些标记的选取必须是能够唯一的标识所采集的文章列表。下来我们做一详细的介绍。如图 2-9:

网站建设 网站管理 爱切换到公开目录系					
	备份恢复 用户和权限	限 网站统计 网站安全 <mark>文章采集</mark> 其他设置			
文章采集	从网页采集 🔍 当前位置:文章	章采集>>从网页采集>>任务设置>>第3步:列表页			
	第1步: <u>认证</u> 任务名称 제末提取在式	第2步:基础 第3步:列表页 第4步:正文页 第5步:附加 保存 新闻中心			
	>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	 ● 自动提取 ○ 从页面上采集 ● 自动提取 ○ 从页面上采集 ● 采集和本地系统中标题相同的文章 ○ 不采集和本地系统中标题相同的文章 			
	间隔时间:	每隔 0 毫秒采集一条文章			
	正文页链接	*			
→屋仕労日志 屋计划管理 屋回收站	列表下一页	● 无 ○ 获取下页链接 ○ 指定分页 ○ 获取分页 ○ 指定分页表单 ○ 指定正文页			

图 2-9

【任务名称】此处显示的名称是在第一步中定义的任务的名称。

【列表提取方式】此处有两个选项, **○**自动提取和 **○**从页面上采集,当选择【自动提取】时,系统会自动把文章列表页面中的所有文章采集下来。当选择【从页面上采集】,会 采集你所锁定的文章列表。选择后弹出两个填写框,分别是【列表起始标记】和【列表结束 标记】,如图 2-10:



	网站建设 网站管	管理 🥵 🕺 👘 👔 🕄
	备份恢复 用户和权	限 网站统计 网站安全 <mark>文章采集</mark> 其他设置 🛛 🗾 预览网站>> 🔹
文章采集	从网页采集	
₽ •	💊 当前位置:文章	采集>>从网页采集>>任务设置>>第3步:列表页
■ ■ ■ ↓ 网页采集	留1些,认证	
	第4少・Mur になった	
	11分名称	教用中心
	列表提取万式	○ 目动提取 ● 从页面上米集
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	列来把始标记	<table <="" border="0" cellspacing="0" id="ContentList" th=""></table>
一同收站	2042298410FWG	。 说明:请在列表页源代码内选取恰当的起始和结束标记,用以定位文章列表
	列表结束标记	<pre><div align="center">@nbsp;@nbsp;@nbsp;</div></pre>
	文章链接提取方式	● 自动提取 ○ 从页面上采集
	内容去重	●采集和本地系统中标题相同的文章 ○不采集和本地系统中标题相同的文章
	间隔时间:	每隔 0 毫秒采集—条文章
	正文页链接	http://www.jnszcg.com/content.aspx?contentid=1865&columnid= * 预览正文页 <mark>说明</mark> :请您选择列表页内指向文章正文的任一链接用以设置正文页参数。
	需要转换正文页链接	◎ 不需要 ○ 需要
Da	列表下一页	● 无 ○ 获取下页链接 ○ 指定分页 ○ 获取分页 ○ 指定分页表单 ○ 指定正文页

图 2-10

这两项的设置主要实现的是有针对性的文章采集,例如采集者想要采集文章列表页中"国内文章"列表中的文章时,只要在这两项中作相应的设置就可以实现。具体解释如下:

【列表起始标记】此标记是定位文章列表的起始位置,以采集新浪网文章栏目下的国内 文章为例,打开该文章栏目的页面(如 http://news.sina.com.cn/),查看其源文件,找出能 唯一标识国内文章列表起始的标记,我们取(<!---国内开始-->)。

【列表结束标记】此标记是定位文章列表的结束位置。我们取(<!-- div03 end -->)。 填好后,点击右侧的"^{预览列表}",预览看所选择的文章列表是否达到自己的要求。如果没 达到,则需要再另外定义标记。

【内容去重】在采集文章时,有可能本系统中已经存在一些文章,为了避免和这些文章 重复,可在此处进行设置,当需要采集和本地系统中标题相同的文章时,可以选择 〇 采集和 本地系统中标题相同的文章。当需要不采集和本地系统中标题相同的文章时,可以选择 〇 不 采集和本地系统中标题相同的文章。

【间隔时间】是设置采集一条文章时彼此间隔的时间。

【内容页链接】指用户所定义的文章列表中文章的内容显示页的地址,因为一个文章列 表中所有的文章一般都会采用同一个内容显示页。所以我们点开某一个文章,把其地址填写 到内容页链接地址栏中。填好后,点击右侧的"^{预览正文页}"检测其内容页是否能显示出 来。

【列表下一页】此处设置了六个选项,分别是: 〇 无、〇 获取下页链接、〇 指定分页、



○ 获取分页、○ 指定分页表单、○ 指定正文页。当所采集的文章列表中只有一页时,就选择"○ 无";当所采集的文章列表中有"上一页"和"下一页"时,选择"○ 获取下页链接",此处有两个必填项,【下一页链接起始】和【下一页链接结束】,如图 2-11;

网站建设 网站管理 念切换到公开目录系统			
文章采集	从网页采集		
♪ 文章采集	💊 当前位置:文章	采集>>从网页采集>>任务设置>>第3步:列表页	
	第1步: 认证	第2步:基础 第3步:列表页 第4步:正文页 第5步: 附加 保存 正则小 :	
同任务控制台	任务名称	新闻中心	
	列表提取方式	○ 自动提取 ● 从页面上采集	
	列表起始标记	******访明:请在列表页源代码内选取恰当的起始和结束标记,用以定位文章列表	
	列表结束标记	<div align="center">@nbsp;@nbsp;@nbsp;@nbsp; 、 、 、</div>	
1011万 反立	文章链接提取方式	● 自动提取 ○ 从页面上采集	
	内容去重	●采集和本地系统中标题相同的文章 ○不采集和本地系统中标题相同的文章	
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	间隔时间:	每隔 0 毫秒采集一条文章	
	正文页链接	http://www.jnszcg.com/content.aspx?contentid=1865&columnid= * 预览正文页 <mark>说明</mark> :请您选择列表页内指向文章正文的任一链接用以设置正文页参数。	
	需要转换正文页链接	◎ 不需要 ○ 需要	
	列表下一页	○ 无 ● 获取下页链接 ○ 指定分页 ○ 获取分页 ○ 指定分页表单 ○ 指定正文页	
	下一页链接起始	上一页∜≥>	
	下一页链接结束		

图 2-11

例如:上一页下一页,可以取起始标签为:上一页,结束标签为:。填写完后可以点击"^{预览链接}",看链接是否正常。如果不正常,检查标签设置是否合适,网络是否畅通。

当所采集的文章列表中有好多页且具备页码时,选择"指定分页",此处有两个必填项, 【指定分页 URL】和【分页 ID 范围】。

【 指 定 分 页 URL 】 处 填 写 的 是 具 有 分 页 页 面 的 地 址 , 如 (http://ent.163.com/special/00031HI0/entnews1.html),【分页 ID 范围】填写的是分页的 页面地址的 ID 数以及增加的基数设置。如图 2-12:



	网站建设 网站	管理
	备份恢复 用户和权	
文章采集	从网页采集	
₽文章采集	💊 当前位置:文章	采集>>从网页采集>>任务设置>>第3步:列表页
		第2步:基础 第3步:列表页 第4步:正文页 第5步:附加 I 保存 I 正 侧
	任务名称	Ruduv Ruduv
	列表提取方式	○ 自动提取 ● 从页面上采集
	列表起始标记	strl =="width:100%; border=collense; collense; "> <mark>说明</mark> :请在列表页源代码内选取恰当的起始和结束标记,用以定位文章列表
	列表结束标记	<div align="center">@nbsp:@nbsp:@nbsp:@nbsp: @nbsp: 》</div>
「日本法制会	文章链接提取方式	● 自动提取 ○ 从页面上采集
	内容去重	● 采集和本地系统中标题相同的文章 ○ 不采集和本地系统中标题相同的文章
	间隔时间:	每隔 0 毫秒采集一条文章
	正文页链接	http://www.jnszcg.com/content.aspx?contentid=1865&columnid= * 预览正文页 <mark>说明</mark> :请您选择列表页内指向文章正文的任一链接用以设置正文页参数。
	需要转换正文页链接	◎ 不需要 ○ 需要
	列表下一页	○无 ○ 获取下页链接 ◎ 指定分页 ○ 获取分页 ○ 指定分页表单 ○ 指定正文页
	指定分页URL	例如:http://finance.sina.com.cn/money/list/list_\${ID}.htm, \${ID}是分页的数字
100	分页ID范围	0 到 0 指定为数字范围,例如:100-200 每次增加 1 <mark>增加基数,如基</mark> 数是10时链接是list_0.htm,list_10.htm,list_20.htm

图 2-12

当所采集的文章列表中设置了分页时,选择"获取分页",会自动弹出设置框。此处有两 个必填项,【下一页链接起始】和【下一页链接结束】。

例如:采集的页面中有"第12345页"时,可以取起始标签为:上一页第,结 束标签为:页。填写完后可以点击"^{预览链接}",看链接是否正常。如果不正常,检查 标签设置是否合适,网络是否畅通。如图 2-13:

网站建设 网站管理 爱尔族到公开目录系统 🔤切换到				📑 切换到内容管	
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~					
文章采集	从网页采集				
₽文章采集	🗣 当前位置:文章	采集>>从网页采集>>任务设置>>第3步:列表页		E	
	第1步: 认证	第2步:基础 第3步:列表页 第4步:正文页 第5步:附加	保存   ]	正则小工具 透	
	任务名称 列表提取方式	<b>新闻中心</b> ○ 自动提取 ◎ 从页面上采集			
	列表起始标记	<pre></pre>			
	列表结束标记	<pre><div align="center">@nbsp;@nbsp;@nbsp;</div></pre>	<b>页</b> 览列表		
	文章链接提取方式	● 自动提取 ○ 从页面上采集			
	内容去重	● 采集和本地系统中标题相同的文章 ○ 不采集和本地系统中标题相同的	文章		
	间隔时间:	每隔 0 毫秒采集一条文章			
	正文页链接	http://www.jnszcg.com/content.aspx?contentid=1865&columnid= * 预 说明:请您选择列表页内指向文章正文的任一链接用以设置正文页参数。	览正文页		
	需要转换正文页链接	● 不需要 ○ 需要			
100	列表下一页	○无 ○ 获取下页链接 ○ 指定分页 ● 获取分页 ○ 指定分页表单 ○	指定正文页		
	下一页链接起始	上一页(/a>			
	下一页链接结束	✓ ▲> ① 預送 请确保能预览分页标签。 例如: 8页295条记录 当前为第1页1 2 3 4 5 6	的链接		

图 2-13

**西安博达软件股份有限公司** 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 16 页 共 41 页



当所需的设置填写完后,点击" 保存",此时列表页中的配置项就设定好了。

#### 2.1.2.4 内容页

在第四步内容页的设置页面中。如图 2-14:

	网站建设 网站	管理	🔧 切换到公开目录系统	回切换到内容管
	备份恢复   用户和机	限   网站统计   网站安全   <mark>文章采集</mark>   其他设置	🔎 预览网站>>	📑 发布网站
文章采集	从网页采集			
▶文章采集	👊 当前位置:文	¥采集>>从网页采集>>任务设置>>第4步:正文页		í
	第1步: 试证	第2步:基础 第3步:列表页 <b>第4步:正文页</b> 第5步: 附加	保存   正	则小工具 🧾 🤃
	任务名称 标题提取方式	<mark>新闻中心</mark> ● 自动提取 ○从页面上采集 ○从页面上采集 必须采集到标题,否则 ●	该条资料作废)	
	正文提取方式	○ 自动提取为主  ● 页面采集为主 如果当前方式提取不出正文,则用另一	种方式提取	
<ul> <li>→ 一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一</li></ul>	正文结束标记		船正文	
	恢复HIML转义字符	○不恢夏 ● 恢复 将诸如@nbsp:@lt:@gt:的字符替换成原字符空格,'ぐ,'	×	
	一键排版格式:	<ul> <li>○常用样式-默认</li> <li>○常用样式-大字体</li> <li>○常用样式-代留更少格式</li> <li>○常用样式-大字体+保留更少格式</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体+保留更少格式</li> <li>●通知样式-大字体+保留更少格式</li> <li>●无样式</li> <li>○不排版</li> </ul>		
	正文的下一页	●不采集分页 ○ 获取下一页或分页的链接标签 ○ 指定分页URL		

图 2-14

此处主要是对从网页采集者所要采集的文章内容的定向,此定向同样是通过此页面的源 代码中的特定标记来设定的。具体设置如下:

【标题提取方式】此处有两个选项, **○**自动提取和 **○**从页面上采集,当选择【自动提取】时,系统会自动把文章内容页面中的文章标题采集下来。当选择【从页面上采集】,会采 集你所锁定的文章标题。选择后自动弹出两个填写框,分别是【标题起始标记】和【标题结 束标记】。如图 2-15:



<b>A</b>	第理 気切換到公开目录系统   ■切換到内容管理	
备(		艮   网站统计   网站安全   <mark>文章采集</mark>   其他设置 🗾 🍺 预览网站>>   🛃 发布网站   🦉
文章采集	从网页采集	
	👊 当前位置:文	章采集>>从网页采集>>任务设置>>第4步:正文页
	第1步: 认证	第2步: 基础 第3步: 列志页 第4步: 正文页 第5步: 附加   保存   正则小工具 返回
	任务名称	新闻中心
	标题提取方式	○ 自动提取 💿 从页面上采集 💿 从页面上采集 必须采集到标题,否则该条资料作废)
	标题起始标记	<pre><center class="aTitle"></center></pre>
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □	标题结束标记	≪/center》 ① 预览标题
	正文提取方式	○ 自动提取为主 ④ 页面采集为主 如果当前方式提取不出正文,则用另一种方式提取
	正文起始标记	
	正文结束标记	○ 预览正文
	恢复HTML转义字符	○不恢复 ● 恢复 将诸如@absp:@lt;@gt;的字符替换成原字符空格,'ぐ,'>
	一键排版格式:	<ul> <li>常用祥式-大字体</li> <li>常用祥式-大字体</li> <li>常用祥式-代留更少格式</li> <li>第用祥式-大字体+保留更少格式</li> <li>通知祥式-大字体</li> <li>通知祥式-大字体</li> <li>通知祥式-保留更少格式</li> <li>通知祥式-保留更少格式</li> <li>通知祥式-大字体+保留更少格式</li> <li>● 元梓式</li> <li>不排版</li> </ul>
	正文的下一页	●不采集分页 ○ 获取下一页或分页的链接标签 ○ 指定分页URL

图 2-15

【标题起始标记】和【标题结束标记】搭配起来采集文章的标题。

一般情况下标题起始标记用<title>,标题结束标记用</title>。填好后,点击 "**顶览标题**"可以预览所标记的标题。

【正文提取方式】分为: O 自动提取和 O 从页面上采集,选择【自动提取】时,系统 会自动把文章内容页面中的文章内容采集下来。选择【从页面上采集】,会采集你所锁定的文 章内容。选择后自动弹出两个填写框,分别是【正文起始标记】和【正文结束标记】,如图 2-16:



A	对站建设 网站管理	🔤 切換到内容管理   🛜 文件库   🗵 退
备	分恢复   用户和权限	—————————————————————————————————————
文章采集	从网页采集	
会文章采集	🔦 当前位置:文	章采集>>从网页采集>>任务设置>>第4步:正文页
	第1步: 认证	第2步:基础 第3步:列表页 第4步:正文页 第5步: 附加   保存   正则小工具 返回
	任务名称 标题提取方式	新闻中心 ○自动提取 ◎从页面上采集 ○从页面上采集 必须采集到标题,否则该条资料作废)
	标题起始标记	<pre><center class="aTitle"></center></pre>
■■↓↓数据库采集	标题结束标记	✓center> ⑦ 预览标题
	正文提取方式	○ 自动提取为主
	正文起始标记	
	正文结束标记	⟨/td>         ○ 预览正文
	恢复HTML转义字符	○不恢复 ●恢复 将诸如@mbsp;@lt;@gt:的字符替换成原字符空格/ ぐ, / У
	一键排版格式:	<ul> <li>常用祥式-紫认</li> <li>常用祥式-大字体</li> <li>常用祥式-保留更少格式</li> <li>常用祥式-快留更少格式</li> <li>通知祥式-大字体+保留更少格式</li> <li>通知祥式-保留更少格式</li> <li>通知祥式-大字体+保留更少格式</li> <li>通知祥式-大字体+保留更少格式</li> <li>通知祥式</li> </ul>
	正文的下一页	● 不采集分页 O 获取下一页或分页的链接标签 O 指定分页UBL



【正文起始标记】和【正文结束标记】搭配起来就可以采集文章的正文。注意的是,此 处的标记要能够唯一的标识文章内容。填好后,可以点击"^{预览正文}"进行预览。

【恢复 HTML 转义字符】此处有两个选项,分别是: 〇不恢复和 〇恢复。选择不恢复时, 采集下来的文章内容如果包含 <等字符,将不转换为原字符。选择恢复时,若采集 的文章内容如果包含 <等字符,将转换为原字符。

【一键排版格式】选择对应的选项,系统会自动采集到的文章进行对应的排版。

【内容的下一页】此处有两个选择项,当文章内容页中没有下一页时,就选择 ○ 不采集 分页。当文章内容页中有分页时,就选择 ○ 获取下一页或分页的链接标签,此处设定方法和 2.1.2.3 节列表页【列表下一页】设置项"获取下一页"中是一样。如图 2-17:



文章采集

	网站建设 网站管理		■切換到內容管理   🕞 文件库
	备份恢复   用户和权限	网站统计   网站安全   <mark>文章采集   更多&gt;&gt;                                   </mark>	站   🍘访问网站   🗍切换网站: 👊
文章采集	从网页采集		
文章采集	🗣 当前位置:文	章采集>>从网页采集>>任务设置>>第4步: 正文页	l l l l l l l l l l l l l l l l l l l
	第1步: 认证	第2步: 基础 第3步: <u>列表页</u> <b>第4步: 正文页</b> 第5步: 附加	保存   正则小工具 遊
	任务名称	新闻中心	
	标题提取方式	○自动提取 ◉从页面上采集 ○从页面上采集 必须采集到标题,否则该条资	料作废)
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	标题起始标记	<center class="aTitle"></center>	
同收站			
10000000000000000000000000000000000000	标题结束标记	√ center オ ⑦ 预览标	题
	正文提取方式	○ 自动提取为主 ④ 页面采集为主 如果当前方式提取不出正文,则用另一种方式	式提取
	正文起始标记		
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	正文结束标记	⟨/td>         ○ 預览正:	文
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	恢复HIML转义字符		
	<b>一键排</b> 版格式:	<ul> <li>○常用样式-默认</li> <li>○常用样式-大字体</li> <li>○常用样式-大字体</li> <li>○常用样式-大字体+保留更少格式</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体</li> <li>●通知样式-大字体</li> <li>●元样式</li> <li>●元样式</li> <li>○不排版</li> </ul>	
	正文的下一页	○不采集分页 · ● 获取下一页或分页的链接标签 · ○ 指定分页URL	
	下一页链接起始	0	
	下一页链接结束	▲ 予約1000000000000000000000000000000000000	

当此页面的设置项设定完后,点击" 保存",文章内容的采集就设置好了。

## 2.1.2.5 附加

在第五步附加的设置页面中,主要是针对某些网页中穿插一些图片、附件以及用户想要 采集一些附加的内容而设置的。此处的设置是可选项,如果用户有需求的话,可以进行设置, 否则不需要设置,采用默认状态。接下来我们对其页面中的设置项分别进行介绍。如图 2-18:



	网站建设 网站管理	<b>念切换到公开</b> 间
文章采集	从网页采集	
会会主要 (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998)	嗿 当前位置: 文章采集	>>从网页采集>>任务设置>>第5步: 附加
🖻 👼从网页采集		
	第1 <b>步:</b> <u>加</u> 止 第43	
	11分名称 	
──□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	友仲祆忿	
	内容所嵌的文档和图片	○不下载,链接到原来地址(推荐) ●下载并存放在本地系统中
一一一一一回收站	下载的文档和图片类型	jpg, jpeg, png, gif, bmp, swf 🔨
■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□		$\sim$
	需要转换文档和图片链接	◎ 不需要 ○ 需要
	设置图片链接	● 不设置图片链接 ○ 提取内容中的图片作为图片链接
	附件	
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	附件	● 不下载 ○ 下载并存放在本地系统中
同收站	附件链接提取方式	● 自动提取 ○ 从页面上采集 附件文件名编码: VIIF-8
	需要转换附件链接	<ul> <li>● 不需要</li> <li>○ 需要</li> </ul>
141	发布日期	
and a second sec	发布日期获取方式	●当前日期 ○在正文页采集 ○在列表后面采集 ○在列表前面采集
	浏览次数	
	浏览次数获取方式	<ul><li>●指定固定值</li><li>○在正文页采集</li><li>○在列表后面采集</li><li>○在列表前面采集</li></ul>
	指定固定值	
1000	摘要	
	摘要获取方式	● 从正文截取摘要 ○ 从页面上采集摘要 ○ 不采集
	获取正文前	字符作为摘要

图 2-18

【任务名称】显示的是此项采集任务的名称。

【发布状态】此处有两个选项, ○采集后需要审核后才能发布和 ○采集后立即发布。当 从网页采集者需要对采集下来的文章进行审核, 审核通过后才能发布时(前提条件是站点中 已经存在文章审核流), 可以选择 ○采集后需要审核后才能发布, 当从网页采集者对采集下 来的文章不需要做审核就让它发布时。可以选择 ○采集后立即发布。

【内容所嵌的文档和图片】一般的网页上的文章内容中都会镶嵌一些图片来美化页面, 文章采集者也希望采集系统能够采集文章内容的同时,连同内容中的图片也采集下来。

此处设置了两种选项供从网页采集者选择,当从网页采集者考虑本地磁盘空间,不想让 所采集页面中的图片下载到本地,只想链接该图片时,从网页采集者可以选择默认的 〇 不下 载,链接到原来地址(推荐)选项。当从网页采集者想要让所采集页面中的图片下载到本地, 已为方便使用时,可以选择 〇 下载并存放在本地系统中选项。

【需要转换文档和图片链接】对于内容所嵌的文档或图片,如不需要转换链接,选择 不需要即可。如需要转换链接,则选择 一需要,输入"原始链接模板"和"转换后链接模板" 即可。



【设置图片链接】如果想要把从文章内容页面中采集下来图片作为此条文章的链接,那 么在此处可以进行设置,选择这一设置项中的 记提取内容中的图片作为图片链接。如不需要 的话,可以选择 了不设置图片链接。

【附件】此处有两个选择项 ○ 不下载和 ○ 下载并存放在本地系统中。有的文章中除了显示内容外还带有附件,以供读者阅读,当采集者采集文章时,想要把附件也采集下来时,在此处可以选择 ○ 下载并存放在本地系统中。此时会在下方自动出现一个填写定向附件位置的标签填写项,然后在出现的两项填写处分别填入能够定位到附件的起始标记和结束标记。点击" 预览采集内容 "可以预览所要采集的附件信息。当采集者只需采集文章的内容,不想要附件时,可以选择 ○ 不下载。如图 2-19:

网站	建设 网站管理	📑 切换到内容管理   🚞 文件库   📧
备份恢		统计 丨 网站安全 📙 文章采集 📙 更多>> 👘 预览网站 📔 🗟 发布网站 📗 🦉 访问网站 📔 🗇 切换网站: 📬 de
文章采集	从网页采集	
₽  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2	👊 当前位置:文章采集	>>从网页采集>>任务设置>>第5步:附加
■ ■ ↓ 网页采集		
	第12・ <u>MIE</u> 第22	
	住方:名称 华女伊太	
	皮仰1/53 由変形(含約合称)の同時	
	内谷所做的又相和國方	○ 不下戰,閱接到原米地址(推存) ● 下戰开仔預任本地系统中
一一一一回收站	下载的文档和图片类型	jpg, jpeg, png, gif, bmp, swf
🖹 🕞 从数据库采集		$\checkmark$
	需要转换文档和图片链接	◎ 不需要 ○ 需要
	设置图片链接	● 不设置图片链接 ○ 提取内容中的图片作为图片链接
	附件	
	附件	<ul> <li>不下载</li> <li>○下载并存放在本地系统中</li> </ul>
同收站	附件链接提取方式	● 自动提取 ○从页面上采集 附件文件名编码: \UTF-8 ▼
	需要转换附件链接	◎不需要 ○需要
	发布日期	
	发布日期获取方式	●当前日期 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	浏览次数	
N.	浏览次数获取方式	● 指定固定值 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	指定固定值	
1.00	摘要	
	摘要获取方式	● 从正文截取摘要 ○ 从页面上采集摘要 ○ 不采集
	获取正文前	字符作为摘要

#### 图 2-19

【作者】当文章采集者想要为采集下来的文章内容添加一个自己所拟定的作者名称时, 可以选择 [•] 指定作者名称,然后在下边的指定作者名称为右边的输入框中填写所要拟定的作 者名称。当文章采集者想要直接获取原文章内容中作者的名称时,可以选择 • 从页面中采集, 此时会在下方自动出现一个填写定位到作者位置的标签填写项,分别填入能够定向作者位置 的起始标记和结束标记。如图 2-20:



	网站建设 网站管理	☎切换到公开目录系
文章采集	从网页采集	
会会主要 (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998) (1998)	💊 当前位置: 文章采集	;>>从网页采集>>任务设置>>第5步: 附加
➡■■↓₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩	Advertise and the second secon	
一一一一任务设置		
	任务名称	新闻中心
──────────────────────	发布状态	● 采集后需要审核才能发布 ○ 采集后立即发布
- 123 计划管理	内容所嵌的文档和图片	○ 不下载,链接到原来地址(推荐) 💿 下载并存放在本地系统中
一一一一一日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日	下我的立档和图片米刑	jpg, jpeg, png, gif, bmp, swf 🔨
🖻 👼 从数据库采集	19303天1119四月天至	✓
一层任务设置	需要转换文档和图片链接	● 不需要 ○ 需要
一一一一一日子子子子子子子子子子子子子子子子子子子子子子子子子子子子子子子子子	设置图片链接	● 不设置图片链接 ○ 提取内容中的图片作为图片链接
	附件	
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	附件	○ 不下载 ④ 下载并存放在本地系统中
	附件链接区域起始标记:	0
3/14	附件链接区域结束标记:	○ 预览采集内容
	附件链接提取方式	● 自动提取 ○ 从页面上采集 附件文件名编码: UTF-8 ▼
50	需要转换附件链接	◎ 不需要 ○ 需要
	发布日期	
	发布日期获取方式	●当前日期 ○在正文页采集 ○在列表后面采集 ○在列表前面采集
	浏览次数	
	浏览次数获取方式	● 指定固定值 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	指定固定值	

图 2-20

点击"^{预览作者}"可以预览所要采集的作者信息。

【发布日期】采集的文章,有的显示了文章的发布日期,如果希望采集过来的文章显示 当前日期,则选择 〇当前日期。如需要采集原文章的信息,则选择 〇从页面采集,在弹出的 指定位置处,输入标记即可。



Ā	网站建设 网站管理	🧏 切換到公开目录系统   📄
备		
文章采集	从网页采集	
₽  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2	💊 当前位置: 文章采集	;>>从网页采集>>任务设置>>第5步: 附加
□	第1步: 认证 第2步	: 基础 第3步: 列志页 第4步: 正立页 <b>第5步: 附加   保存</b>
日子在	任务名称	新浪健康
同任务日志	发布状态	○ 采集后需要审核才能发布 ④ 采集后立即发布
同计划管理	内容所嵌的文档和图片	● 不下载,链接到原来地址(推荐) ○下载并存放在本地系统中
····同回收站 ····同人数据库采集	下载的文档和图片类型	jpg, jpeg, png, gif, bmp, swf
■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	需要转换文档和图片链接	◎ 不需要 ○ 需要
	设置图片链接	○ 不设置图片链接 ● 提取内容中的图片作为图片链接
	附件	
	附件	◎ 不下载 ○下载并存放在本地系统中
一同收站	附件链接提取方式	● 自动提取 ○ 从页面上采集 附件文件名编码: UTF-8 ▼
	需要转换附件链接	◎ 不需要 ○ 需要
	发布日期	
	发布日期获取方式	○当前日期 ○ 在正文页采集 ④ 在列表后面采集 ○ 在列表前面采集
- N/4.	日期起始标记:	<pre><span style="color:#4D4D4D;">(</span></pre>
	日期结束标记:	)√/span≫/li> ⑦ 预览日期
	日期格式:	MM月43日 H91:mm y=年, M=月, d=日, H=小时, m=分钟, s=秒,采集日期格式应当匹配此格式
	浏览次数	
	浏览次数获取方式	● 指定固定值 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	指定固定值	
	摘要	

【文章来源】一般的文章中都会标识其内容的来源,我们在此也特意添加了采集资料来 源的功能,以方便文章采集者获取来源信息。

如果采集者想要为采集下来的文章内容添加一个自己所设定的资料来源信息时,可以选择 f 指定固定值,填写所要设定的资料来源信息即可。当文章采集者想要直接获取原文章内 容资料来源信息时,可以选择 f 从页面上采集,此时会在下方自动出现一个填写定向资料来 源位置的标签填写项,分别填入定位到资料来源位置的起始标记和结束标记。点击

"**预览来源**"可以预览所要采集的文章资料的来源信息。如图 2-21:



文章采集	😪 当前位置: 文章采集	>>从网页采集>>任务设置>>第5步:附加
会文章采集	□ 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1	
	任务之称	Side
	发布状态	
	及101/20 由密码定的支撑的图理	
	的合则取到又有种国方	◎ 不下戰,捉接到原來地址(推存)  ○ 下戰并行與往本地系統中
一一一一日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日	下载的文档和图片类型	jpg, jpeg, png, gif, bmp, swf
		~
🖻 🐻 从数据库采集	需要转换文档和图片链接	◎ 不需要 ○ 需要
	设置图片链接	○ 不设置图片链接 ● 提取内容中的图片作为图片链接
	附件	
	附件	◎ 不下载 ○下载并存放在本地系统中
──□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	附件链接提取方式	◎ 自动提取  〇 从页面上采集  附件文件名编码: □□TF-8
	需要转换附件链接	◎ 不需要 ○ 需要
	发布日期	
	发布日期获取方式	◎ 当前日期 〇 在正文页采集 〇 在列表后面采集 〇 在列表前面采集
	浏览次数	
	浏览次数获取方式	● 指定固定值 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	指定固定值	
	摘要	
	摘要获取方式	○从正文截取摘要 ○ 从页面上采集摘要 ● 不采集
	作者	● 不采集 〇 指定固定值 ○ 从页面上采集
	关键字	● 不采集 〇 指定固定值 ○ 从页面上采集
	回复次数	● 不采集 〇 指定固定值 ○ 从页面上采集
	文章来源	○不采集 ○ 指定固定值 ● 从页面上采集
	文章来源起始标记:	<pre><span id="media_name"></span></pre>
	文章来源结束标记:	○ 预览文章来源

#### 图 2-21

【摘要获取方式】一般的文章中都会附带一些内容摘要,为文章提供内容概括,我们在 此也特意添加了采集摘要的功能,当采集者想要直接从文章内容的开始处截取一部分内容作 为此文章内容的摘要时,就选择 〇从正文截取摘要,输入需要截取的内容的字符数就即可。 如采集者想要直接获取原文章内容摘要信息,可以选择 〇从页面上采集摘要,此时会在下方 自动出现定为到内容摘要位置的标签填写项,分别填入能够定向内容摘要位置的起始标记和 结束标记。点击" 顶宽摘要 "可以预览所要采集的文章内容摘要的信息。如果在采集的文 章中不配置摘要,可以选择 〇不采集。如图 2-22:



	网站建设 网站管理	🧝 🥵 🎜 🎽 🏂 🏂 🏂 🏂 🏂 🏂 🏂 🏂 🏂 🏂 🏂 🏂
	备份恢复   用户和权限	
文章采集	从网页采集	
全文章采集   一□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	State 217	
	用1步: <u>以此</u> 用25 任圣夕称	→: <u>単磁</u>
	发布状态	● 采集后需要审核才能发布 ○ 采集后立即发布
	内容所嵌的文档和图片	○ 不下载, 锐接到原来地址(推荐) ● 下载并存放在本地系统中
	下载的文档和图片类型	jpg, jp_k, png, gif, bmp, swf
一一一一日子子	需要转换文档和图片链接	◎ 不需要 ○ 需要
	设置图片链接	● 不设置图片链接 ○ 提取内容中的图片作为图片链接
	附件	
	附件	● 不下载 ○ 下载并存放在本地系统中
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	附件链接提取方式	●自动提取 ○从页面上采集 附件文件名编码: UTF-8 ▼
	需要转换附件链接	● 不需要 ○ 需要
	发布日期	
	发布日期获取方式	●当前日期 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	浏览次数	
Service Contraction	浏览次数获取方式	● 指定固定值 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	指定固定值	
	摘要	
	摘要获取方式	○ 从正文截取摘要 ● 从页面上采集摘要 ○ 不采集
	摘要起始标记:	$\bigcirc$
	摘要结束标记:	☆ 預定摘要

图 2-22

当所需的设置项都填完以后,点击"**保存**",此页面填写的信息就设置成功了。 以上四步全部设置完以后,一个完整的文章列表内容页文章的采集任务就设置好了。

## 2.1.2.6 RSS 文章列表采集任务设置

RSS 是站点用来和其他站点之间共享内容的一种简易方式(也称为聚合内容),通常被用 于文章和其他按顺序排列的网站。RSS 已经成为目前最成功的 XML 应用,搭建了信息迅速传 播的一个技术平台,使得每个人都成为潜在的信息提供者。目前在很多大型的网站中都提供 RSS 支持,基于这种情况我们也特意开发了自动采集 RSS 文章的功能。接下来我们详细介绍 如何采集 RSS 文章。

RSS 文章采集也和文章列表页文章的采集方式一样。

【第一步:认证】此任务的功能和设置和选择采集文章列表页的一样。

【第二步:基础】此任务的基础设置页面和采集文章列表页文章的设置页面是一个页面,如图 2-3。提供的设置项大部分相同,唯一不同点,在"采集列表类型项"中要选择 CRSS 链接。在【采集列表地址】中要填入所要采集文章的 RSS 格式的 XML 链接地址,例如新浪的 RSS 频道中文章中心下的文章要闻列表(http://rss.sina.com.cn/news/marquee/ddt.xml), **西安博达软件股份有限公司** XI'AN WEBBER SOFTWARE CO., LTD. 地址:中国•西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 26 页 共 41 页



其他项的功能和设置和选择采集文章列表页的一样。在此不多做介绍。

【第三步:列表页】此页面的主要有"内容去重"和"正文页链接"。内容页链接的设置 是为下一步采集具体文章内容做准备。

具体操作是在 RSS 文章列表页中随便打开一条文章,把此条文章内容页中的地址输入正 文页链接地址栏中,点击"______"按钮可以进行预览。这样设置的原因和普通文章 列表的文章内容显示页一样,都会采用同样一个内容显示页。如图 2-23:



图 2-23

【第四步:正文页】此页面的设置和采集文章列表页中第四步正文页设置一样,具体操 作也相同。如图 2-24:

网边	站建设 网站	管理	🔧 切换到公开目录系统   📑 切换到内容
备份			🔎 预览网站>>   📑 发布网站
文章采集 ()	从网页采集		
₽文章采集	🛓 当前位置: 文:	\$P\$采集>>从网页采集>>任务设置>>第4步:正文页	
	第1步: 认证	第2步:基础 第3步:列表页 第4步:正文页 第5步: 附加	保存
	务名称	新闻中心	
	题提取方式	○自动提取 💿 从页面上采集 💿 从页面上采集 必须采集到标题,否则该条	(资料作废)
	题起始标记	<div id="Title"></div>	
□·□□□收站 □·□□□从数据库采集 标题	题结束标记	① 预览	标题
	文提取方式	○ 自动提取为主 💿 页面采集为主 如果当前方式提取不出正文,则用另一种?	<b>方式提</b> 取
	文起始标记	<pre>(div id="Content")</pre>	
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	文结束标记	〈/font〉〈/div〉	正文
	夏HIML转义字符	◯ 不恢复 💿 恢复 将诸如@mbsp:@lt:@gt:的字符替换成原字符空格,′ぐ,′≻	
-1	建排版格式:	<ul> <li>○常用样式-默认</li> <li>○常用样式-大字体</li> <li>○常用样式-大字体+保留更少格式</li> <li>○通知样式-大字体+保留更少格式</li> <li>○通知样式-大字体</li> <li>○通知样式-大字体+保留更少格式</li> <li>●通知样式-大字体+保留更少格式</li> <li>●无样式</li> <li>○ 无样式</li> </ul>	
E	文的下一页	●不采集分页 ○ 获取下一页或分页的链接标签 ○ 指定分页URL	

**西安博达软件股份有限公司** 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. 単话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 27 页 共 41 页



#### 图 2-24

【第五步:附加】此页面的设置方法与采集文章列表页一样。如图 2-25:

	网站建设 网站管理	常切換到公开目录系统   回切換到内容管理
	<b>昔份恢复   用户和权限</b>	网站统计   网站安全   安皇亲儒   其他说罢 👘 极斑网站>>   🛃 发布网站
文章采集	下载的文档和图片类型	Jpg. Jp+g. png. gif, hep, swf
少文章采集 ⇒ ○从网页采集 → ○任务控制台 → ○任务控制台 → ○任务控制 → ○任务控制 → ○日教管理 → ○日教管理 → ○日教法	雷要转换文档和图片链接 设置图片链接 開件 附件 附件链接提取方式 雷要转换附件链接 发 <b>会目期</b>	<ul> <li>● 不霉要 ○ 霉要</li> <li>● 不设置图片链接 ○ 提取內容中的图片作为图片链接</li> <li>● 不下载 ○ 下载并存放在本地系统中</li> <li>● 自动提取 ○ 从页面上采集 附件文件名编码: [JTP-8 ]</li> <li>● 不需要 ○ 需要</li> </ul>
○ QM資源除米属 同任务设置	发布日期获取方式	○当前日期 ●在正文页采集 ○在列表后最采集 ○在列表前面采集
同任务控制自	日期起始标记:	(tr) (th aligns' senter' selector' ()
一〇日の日本	日期结束标记:	来禮: 〇 粉炎日期
一〇日代は	日期格式	yyyy ===-84 y=年、=月、は日、昨小時、=分钟・=秒・系集日期格式应当匹配比格式
	浏览次数 测范次数获取方式 描定固定值	●指定固定值 ○ 在正文页采集 ○ 在列表后面采集 ○ 在列表前面采集
	新客	
	接要获取方式	※从正文献取損要 〇 从页面上采集損要 〇 不采集
	録歌正文前	字符作为编奏
	8- <b>8</b>	●不采集 ○ 指定固定值 ○ 从页面上采集
	关键字	●不采集 ○指定固定值 ○从页面上采集
	回夏大教	●不采集○指定固定值 ○从页面上采集

图 2-25

通过以上步骤的设置后,一个完整的 RSS 文章采集任务就完成了,接下来可以采集 RSS 文章了。

#### 2.1.3 删除任务

## 2.1.4 启用任务和停用任务

为了对采集任务的启用和停用操作方便,我们特意在任务列表的上部增设了这两项功能,如图 2-1。当采集管理者想对某些任务启用时,只要选中要启用的任务,然后点击"^{信用任务}", 任务就启用了;当采集管理者想对已经启用的一些任务停止时,选中要停用的任务,点击

" ^{停用任务}"就可以了。 西安博达软件股份有限公司 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077

西安博达软件股份有限公司 | 地址: 中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. | 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 28 页 共 41 页



## 2.1.5 设置计划

设置计划主要是为采集任务定时启动而设置的。此功能为从网页采集者提供了智能化的 采集方案,方便了采集者的维护。同时也提高了文章的时效性。具体操作是,点击" 设置计划 ", 弹出如图 2-26 界面:

设定计划	>	ĸ
任务名: 任务描述: 计划开始日期: 执行时间:	信息采集 采集:商院新闻 年月日 0时▼0分▼	]
	<ul> <li>○ 每天</li> <li>○ 每周 周- ▼</li> <li>○ 每月 1日 ▼</li> <li>○ 每隔 0天 ▼ 0小时 ▼ 0分 ▼</li> <li>○ 一次性执行</li> </ul>	
	确定取消	

图 2-26

然后在弹出框中详细填写其中的配置项,点击"^{确定}",采集计划就添加成功了。在此 需要注意的是:配置项中的任务名指的是采集任务的任务名,这个必须填写正确,以确保任 务执行正确。

## 2.1.6 采集记录管理

在每一个采集任务的名称旁边都配置了采集记录管理的功能,以方便从网页采集者对采 集的记录进行维护。点击查看记录,进入采集记录管理页面,此页面有三个设置项,分别是: 重置采集状态、删除采集状态、清空采集状态。在其下面的列表页中记录了采集的文章的来 源信息,此功能是方便从网页采集者查阅所采集下来的文章的出处,并且每条记录都有链接。 如图 2-27:



文章采集

	网站建设 <b>网站管理</b>	🔧 切打	奂到公开目录系统   📑 切换到	内容管理   🐠
	备份恢复   用户和权限   网站统计   网站安全   <mark>文章采集</mark>   其他设置		🔎 预览网站>> 📔 🛃 发布	词网站   🌽访问
文章采集	采集记录管理			
	🔩 当前位置:文章采集>>从网页采集>>任务设置>>采集记录管理			E
	重置采集状态 删除采集记录 清空采集记录			返
	共10条 每页 10 ∨ 条 1/1 上页 1 下页 转到 页			查
	□ 标题	采集标志	抓取日期	栏目名称
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	□ 运动防癌该掌握什么样的度	已采集	2014-03-04 17:05:20	健康生活
一一一一回收站	□ 这四类人不适合爬山	已采集	2014-03-04 17:05:18	健康生活
■■人数据库采集	□ 爬山四注意事项三要领	已采集	2014-03-04 17:05:15	健康生活
一一一件务设置	□ 春天爬山有什么好处	已采集	2014-03-04 17:05:13	健康生活
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	□ 登山运动的健康步法	已采集	2014-03-04 17:05:11	健康生活
	□ 初春多做五种养生活动	已采集	2014-03-04 17:05:08	健康生活
「「「「「「」」と「「「」」」と「「「「」」」と「「「「」」」」	□ 4个好习惯让你做个健身达人	已采集	2014-03-04 17:05:06	健康生活
	□ 如意美人操打造美人脸	已采集	2014-03-04 17:05:03	健康生活
	□ 坐着让你瘦腰的简单小动作	已采集	2014-03-04 17:05:01	健康生活
	□	口采住	2014-03-04 17:04:59	健康生活

如图 2-27

## 2.1.6.1 重置采集状态

每个采集任务执行后,所有的采集记录都会在此页面中列出来,当从网页采集者发现某 条采集下来的文章的显示格式或者内容不正确或者没有达到自己的需求,需要对任务的设置 改动,然后再对此条文章重新采集,但要求不对其他采集下来的文章再重新采集时,从网页 采集者可以选中此条记录,点击"重置采集状态",此条记录的采集标志会变为"未采集"。 接下来,采集者就可以改动任务的设置了,改动后,然后再重新采集,此刻需要重新采集的 那条文章就采集下来,并保存到了指定的文章组件中。如果采集者不想要原来那个不尽意的 文章时,可以在栏目文章源中把原来的文章的删掉。

## 2.1.6.2 删除采集记录



## 2.1.6.3 清空采集记录

当采集者想要对某个任务中的文章重新采集时,那么必须在此页面中,把此任务中所有的记录删除。为了删除方便,我们提供了清空采集记录功能,只要在此页面中点击 "^{清空采集记录}",所有的记录都会删除掉。

## 2.2 任务控制台

任务控制台中主要提供了对采集任务的启动和停止管理。任务工作区会显示出采集任务 过程中的相关信息。所有的任务想要执行,都要在任务控制台进行启动。在此需要说明的是, 如果从网页采集者建立了好多采集任务,但是想要让指定的任务执行时,必须在任务设置中, 把其他不执行的任务的状态改为停用状态,然后采集。如果在采集的过程中,想要停止采集, 直接点击" 停止采集 ",采集任务就会终止。控制界面如图 2-28:

	网站建设 网站管理		🟂 切换到公开目录系统	🖹 切换到内容管理   🐗
		全   <mark>文章采集</mark>   其他设置	🔎 预览网站>>	🛃 发布网站   🌌 访]
文章采集	从网页采集			
文章采集 文章采集 二人任务设置 二人任务设置 二人任务时志 二人代教理》 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人代教授教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の代教 二人の一人の一人の 二人の一人の 二人の一人の 二人の一人の 二人の一人の 二人の一人の 二人の一人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の 二人の	<ul> <li></li></ul>	■ まま	07125439.shtml六种实用 06125438.shtml每天走路 29125415.shtml运动的十 53125285.shtml活家庭约 15124976.shtml初春多做 15125016.shtml长期坚持 43124969.shtml5个强劲/ 43124968.shtml4个好习/ 30124959.shtml瑜伽减肥	的散步方式 OK. ·20分钟延寿5年 ·大好处 OK. 东习操舒缓颈肩疲 ·五种养生活动 ·跑步对身体有13 然脂运动 OK. 贯让你做个健身达 ·的4大好处 OK.
	[17:05:44]新浪健康(列表)采集条数已 [17:05:44]新浪健康(列表)内容链接采: [17:05:44]新浪健康(列表)执行完成	达到任务设置的最大限制,任务完成 集完成,共采集20条	<u>گ</u> ۰	
	[17.05:46]共有 0 个采集任务,当前有 [17:05:46]***采集任务全部结束。 [17:05:46]任务[1018260793]完成。	0在运行,有0在等待执行		

图 2-28

## 2.3 任务日志

此处的功能主要是详细记录采集任务的操作情况。

**西安博达软件股份有限公司** 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 31 页 共 41 页



例如: 启用一个任务, 然后执行, 此处会记录执行任务过程中出现的操作。如图 2-29:

	网站建设 网站管理	🔏 初狭到公井日求条统 📗 副初狭到内谷官理 🛛 🦏 花的
音	备份恢复   用户和权限   网站统计   网站安全	文章采集   其他设置 🗾 🗾 预览网站>>   🖬 发布网站   🖉 访问网站
文章采集	从网页采集	
会 文章采集	💊 当前位置: 文章采集>>从网页采集>>任务日期	5 🦻
■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■		
	共203956条 每页 15 ∨条 1/13598 上页 1 2	2 3 4 5 6 13598 下页 转到 页 查询
	日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日	事件
	2014-03-04 17:05:44	新浪健康 (列表)执行完成
	2014-03-04 17:05:44	新浪健康 (列表)内容链接采集完成,共采集20条
■ ■ → → → → → → → → → → → → → → → → → →	2014-03-04 17:05:44	新浪健康 (列表)采集条数已达到任务设置的最大限制,任务完成。
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	2014-03-04 17:04:46	开始采集 新浪健康(列表),共有24个链接
	2014-03-04 17:04:46	新浪健康(列表) 下一页链接为:http://roll.health.sina.com.cn/health_jianshen/index_1.shtml
	2014-03-04 17:04:46	新浪健康(列表) 获取列表下一页链接
「日本日本	2014-03-04 17:04:46	开始采集 新浪健康(列表):http://roll.health.sina.com.cn/health_jianshen/index_1.shtml
	2013-11-21 11:18:48	开始采集 新浪健康(列表),共有24个链接
	2013-11-21 11:18:48	新浪健康(列表) 下一页链接为:http://roll.health.sina.com.cn/health_jianshen/index_97.shtml
	2013-11-21 11:18:48	新浪健康(列表) 获取列表下一页链接
	2013-11-21 11:18:47	新浪健康 (列表)采集下一页:http://roll.health.sina.com.cn/health_jianshen/index_96.shtml
	2013-11-21 11:18:47	新浪健康(列表)内容链接采集完成,共采集2271条
N/14	2013-11-21 11:17:04	开始采集 新浪健康(列表),共有24个链接
690	2013-11-21 11:17:04	新浪健康(列表) 下一页链接为:http://roll.health.sina.com.cn/health_jianshen/index_96.shtml
	2013-11-21 11:17:04	新浪健康(列表)获取列表下一页链接

图 2-29

此功能的作用是,当采集任务出现问题时,便于我们查找问题、进行排错。当采集管理 者不想要某条日志在此处显示时,可以选中日志,点击"^{删除日志}"就可以了。如果管理 者想要删除所有的日志时,点击"^{清空日志}",此刻页面中的所有日志就被清空了。除此之 外,此处还配置了模糊查询功能。以方便管理者查询记录。

## 2.4 回收站

显示删除任务信息记录,在这里选中记录点击"^{恢复任务}"按钮,就可以对已经删除的记录进行恢复操作了。如果确定某条记录要彻底删除,点击"^{彻底删除任务}"就可以彻底删除该采集任务了。

	网站建设 网站管理			🛪 切换到公开目录系统	🔄 切换到内容管理 丨 📢
	备份恢复   用户和权限		文章采集   其他设置	☑ 预览网站>>	🛃发布网站   🌌访
文章采集	从网页采集				
●文章采集	💊 当前位置:文章采集>>	从网页采集>>回收站			É
■·□○ M M D R 集 □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	恢复任务 彻底删除任	务			
一同任务控制台	共1条 每页 15 ✔条 1/1	上页 1 下页 转到	页		查
	□ 采集任务名称	任务状态	采集	也址	采集/扫描氛
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	□ 美洲经济	停用	http://roll.finance.sina.com.cn/	finance/gjcj/mzjj/index_1. shtml	0/0条
🖃 🐻 从数据库采集					
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一					
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一					

图 2-31

**西安博达软件股份有限公司** 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 32 页 共 41 页



# 3. 从数据库采集管理的功能说明

## 3.1 从数据库采集管理概述

将外部数据库中的数据信息按照用户的设定方式提取出来,存放到本产品产品所使用的 数据库中,并将这些数据以文章的形式展现出来(即就是将任意数据库中表的数据按照用户 指定方式采集到本产品的文章表中)。

其功能特点:可以对任意数据库进行操作,可以进行多表的数据联合采集。操作性简单, 用户自定义任务,提取数据。可以设置计划,定时、定量的从要求的数据库中提取数据,展 示在本产品的前台网站文章页中。

## 3.2 任务设置

	网站建设 网站管理			🔧切换	到公开目录系统   📑切换到
	备份恢复   用户和权限   网站约	除计   网站安全   <mark>文章</mark>	章采集   其他设置		🝺 预览网站>> 📔 🛃 发布
文章采集	从数据库采集				
₽文章采集	🗣 当前位置: 文章采集>>从数据	露车采集>>任务设置			¢
● ■ ↓ 网页采集		4条	沿署计划 「百制」		
──□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□			双血时刻 美明		
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	共1条 毎页 15 ✔条 1/1 上页	1 下页 转到 了	页		查
	▼ 采集任务名称	任务状态 林	栏目名称	任务创建时间	操作
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	□ 文件下载	启用(	健康生活	2014-03-04 17:12:54	[任务设置] [查看记录]
一同收站					
- 〒从数据库采集					
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一					

图 3-1

当需要添加新的从数据库采集任务时,可以点击"^{添加任务}",在弹出的页面中添加采 集任务。从数据库采集的任务设置分为三步(基本信息设置、采集信息设置、表属性映射设 置),填写任务时必须逐步填写。其中每一步中凡是有带*符号的填写项都为必填项。

前提条件是: 您需要导入的数据所在的数据需要在系统管理下的数据库连接池中进行配置, 您可以根据需要链接不同的数据库, 例如 access、oracle、mssq12000、sybase、db2、mysq1等. 以下以 mssq12000 数据库为例,介绍一下数据库的配置操作,如图 3-2 所示:



	站群管理	用户组织	安全防护	系统设置	环境与部署	授权与更新	
	部署配置	反向代理设置	远程诊断				
▶ 部署配置 ▶ 新考配置 ▶ 玩貨配置 ▶ ひけま)	🗣 当前(	立置: 部署配置)	>数据库连接池				
·····································						数据库连接池列表	
					jbpm		vsb
	Sec. 12. 1. 1. 1.	L				连接池参数设置	
	连接池名利	β œ=		数据	5年类型 [ORACLE	✓ 数据库应用名:	vsb 🗸
	JUBC WZZD	柱序					
	TIRC OKT	市と早					
	刻据库登/	「「「「「「」」					
	<b>数据库登</b> /	(留明			则11注接	新瑄/1朱仔	

图 3-2

首先在连接池名称后输入名称,接着选择数据库类型,数据库应用名使用默认的就可以 了,再接着须在 JDBC URL 字段中格根据提示输入相应的配置信息,Host 指连接数据库的主 机名或 ip, port 是指端口号,为各种数据库的默认值(mssq12000为1433),dbname 即为所 要连接的数据库名。输入数据库登入帐号,若有密码请输入密码。测试链接,成功链通之后,

点击"新增/保存""按钮即可保存以上设置。设置后的页面如图 3-3 所示:



文章采集

	站群管理用户组织	安全防护 系统设置	环境与部署	授权与更新	☑退出
	部署配置   反向代理设	置   远程诊断			
🌗 部署配置 🌗 环境配置	🗣 当前位置:部署配	置>>数据库连接池			1
·····································			数据库连接池列	表	
→ 3860123C96800 		defaul t	jbpm	ysh	
			5-2-	120	
the floor					
			连接池参数设置		
203	连接池名称	default 数据库类型	MSSQL2000/2005/200	18 💙 数据库应用名 default 💙	
	JDBC 驱动程序	com.microsoft.sqlserve	r.jdbc.SQLServerDriv	er	
	JDBC URL	jdbc:sqlserver://192.1 host 是数据库主机名或ij	68.0.214:5000;databa p 地址, port 默认1433	seName=data;responseBu 3. dbname 是指定的数据库名	
10010	数据库登入帐号	vsb_pgsql			
	数据库登入密码		测试连接 新增/	保存删除	

#### 图 3-3

## 3.2.1 基本信息设置

在建立从数据库采集任务的第一步基础信息页面中,其需要填写的项如图 3-4:

	网站建设 网站管理	<b>Z</b> t	川换到公开目录系统
	备份恢复   用户和权限   网	网站统计   网站安全   <mark>文章采集</mark>   其他设置	🔎 预览网站>
文章采集	从数据库采集		
₽文章采集	💊 当前位置:文章采集>>	>从数据库采集>>任务设置>>第1步:基本信息设置	
⇒ ■ ↓ 网页采集	第1 <u>年</u> ,甘太信自迟罢		1
	第4少・ <u>華平信息収益</u>		1
	任务名称	* ● 启用 ○ 停用	
	信息类型	● 文章资料 ○ 公开信息	
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	文章资料目标栏目	选择栏目文章资料源 *	
一一一一一回收站			
➡─────────────────────────────────────	任务描述	$\square$	
	新堀底佐接油・		
	每次最多采集	条 │─请选择条数─	
计划管理	发布状态	● 采集后需要审核才能发布 ○ 采集后立即发布	
回收站			

#### 图 3-4

【任务名称】本次文章采集任务的任务名称。当此任务在"启用"状态下,执行采集任 务时,会执行此项任务;当在"停用"状态下,执行采集任务时,此项任务不执行。

【目标栏目】在此处可以选择把采集下来的文章或公开信息放到指定的栏目文章源或公



开信息源中,栏目文章源或公开信息源的选择是根据用户的需求来定。例如:点击 "选择栏目文章资料源",会弹出一个选择页面组件的页面。在其页面中选择所要的栏目 文章源,点击"选择栏目文章源"。执行完采集任务后,所采集的文章就直接存放到指定的 栏目文章源中了。如图 3-5:

选择栏目文章源	×
选择栏目文章源	关闭
<ul> <li>● 資料库</li> <li>● ▲ 単目</li> <li>● ● 単目</li> <li>● ● ● 単目</li> </ul>	

图 3-5

【数据库连接池】源数据库和新数据库都需在系统管理下的数据库链接池中配好。这里的数据库链接池应设置为数据来源所在的连接池,我们选择以前我们配置的连接池 default。如图 3-6:

	V
数据库连接池:	
每次最多采集	jbpm 请选择条数- ✓
发布状态	●采集后需要审核才能发布 ○采集后立即发布

图 3-6

【每次最多采集】可以选择你每次所要采集文章的条数。

【发布状态】此处有两个选项: ○采集后需要审核后才能发布和 ○采集后立即发布。当 信从数据库采集者需要对采集下来的文章进行审核,审核通过后才能发布时,可以选择 ○采 集后需要审核后才能发布,当从数据库采集者对采集下来的文章不需要做审核就让它发布时。 可以选择 ○采集后立即发布。当所需的项目设置完后,点击 **保存**,然后进入第二步列



表页的设置。

## 3.2.2 采集信息设置

在第二步采集信息设置页面中填写一条 sql 查询语句。如图 3-7:

	网站建设 网站管理			乙切秧	到公开日求系统   📑 切秧到内谷官
	备份恢复   用户和权限   网	网站统计   网站安全	文章采集   其他设置		应 预览网站>> 📔 🗟 发布网站
文章采集	从数据库采集				
▶文章采集	💊 当前位置:文章采集>>	从数据库采集>>任务设置	置>>第2步:采集信息设置		ø
一一一人一人一人一人一人一人一人一人一人一人一人一人一人一人一人一人一人一人一人		答o止, 必使信白迟罕	答:"你是,其居姓咖啡识罢	1 12 15	
	第4少・ <u>華平信忌反</u> 回	第4少・太差信忌収立	第3少・辺區11時州12日		
	任务名称	文件下载			
		请输入一个SQL宣间语句	]: 		
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	被米集数据的条件设置	Select * from Abreas		0	
同收站					
- 🕞 从数据库采集					
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一					
-					

图 3-7

【任务名称】此处显示的名称是在第一步中定义的任务的名称。

【被采集数据的条件设置】从数据库采集者可以根据自己的需要在上图中输入 sql 语句, 但是一定要确保 sql 的正确性,否则点击保存后会出现错误提示信息。如图 3-8:

	网站建设 网站管理	🔀 切换到公开目录系统 📗 切换到内容管理	載 我的新
	备份恢复   用户和权限	网站统计 🛛 网站安全 🗋 文章采集 📕 其他设置 👘 🗾 预览网站>> 📔 发布网站 📔	∕❷访问网站
文章采集	从数据库采集		
	👊 当前位置:文章采集>>/	\数据库采集>>任务设置>>第2步:采集信息设置	ø
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	第1步:基本信息设置	第2步: 采集信息设置 第3步: 去屋性缺封设置   保存	
	任务名称	文件下载	
	被采集数据的条件设置	请输入一个SQL查询语句: select * from wbnew	
		不合法的SQL语句: org.postgresql.util.PSQLException: 错误: 关系 "wbnew" 不存在	位置: 15
10000000000000000000000000000000000000			
日子中国王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王王			
任务日志			
一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一			

图 3-8

## 3.2.3 属性映射设置

在第三步表属性映射设置的设置页面,主要是从数据库采集者根据源数据库的实际情况 (要了解表结构),设置文章的标题字段、正文字段、采集文本格式、唯一标识字段、作者、 资料来源、摘要、关键字、加入时间。点击保存,任务设置完毕。如图 3-9: 西安博达软件股份有限公司 XI'AN WEBBER SOFTWARE CO., LTD. 第 37 页共 41 页



	网站建设 网站管理		🤧 切换到公开目录系统 丨 📑 切换到
			🗾 预览网站>>   📑发布
文章采集	从数据库采集		
文章采集 → 文章采集 → 一○任务设置 → 一○任务投制台 → 一○任务日志 → 一○计划管理 → 回收站 → ○仲务控制台 → ○任务控制台 → ○任务控制台 → ○任务控制台 → ○任务控制台 → ○任务控制台 → ○行务控制台 → ○行务控制台 → ○行务控制台	<ul> <li>▲ 分恢复</li> <li>由戶和稅限</li> <li>从数据库采集</li> <li>▲ 当前位置:文章采</li> <li>第1步:基本信息设置</li> <li>任务名称</li> <li>标题字段:</li> <li>正文字段</li> <li>采集文本格式</li> <li>发布时间</li> </ul>	図話読计 図話安全 文章朱集 其他设置   ※外数据库采集>>任务设置>>第3步:表属性映射设:   第2步: £集信息设置 第3步:表属性映射设:   第2步: £集信息设置 第3步:表属性映射设:   》如bnewsid ●   ● 纯文本 超文本   》妙date ●   》が成本: 超文本   》がdate ●   》ががらいます。超文本   》がすいます。超文本   》がすいます。   》がすい	● 预凭网站>> ● 发布       置       設置       保存       設置       保存
	唯一标识字段:	wbtop     wbgotopdate       wbforwardtitle     wbeditor       wbshowpic     wbpagetitle       wbenddate     wbextable       wbfirstplan     wbfirststartdate	wbtopdate wbusername wbuuid wbfirst wbfirstenddate

图 3-9

## 3.2.4 修改任务

当从数据库采集者想要对已经创建的任务中的设置信息重新做修改时,只要点击要更改的任务记录后的"^[任务设置]",在其弹出的任务设置界面中作修改就可以了。

## 3.2.5 删除任务

当从数据库采集者由于某种原因,不需要采集某个文章列表中的文章,想删除时,可以 选中要删除的采集任务,点击" **#**除任务 ",此采集任务就被删除了。

## 3.2.6 启用任务和停用任务

为了对采集任务的启用和停用操作方便,我们特意在任务列表的上部增设了这两项功能,如图 3-1。当从数据库采集管理者想对某些任务启用时,只要选中要启用的任务,然后点击 " 启用任务 ",任务就启用了;当从数据库采集管理者想对已经启用的一些任务停止时,选



中要停用的任务,点击" 停用任务 "就可以了。

## 3.2.7 设置计划

设置计划主要是为采集任务定时启动而设置的。此功能为从数据库采集者提供了智能化的采集方案,方便了从数据库采集者的维护。同时也提高了文章的时效性。具体操作是,点击"设置计划",弹出提示界面。如图 3-10:

设置计划		×
任务名: 任务描述:	数据库采集 采集:	
计划开始日期: 执行时间:	年 月 日 回时 ▼ 0分 ▼ ◎ 毎天 ◎ 毎周 周- ▼ ◎ 毎月 1日 ▼ ◎ 毎隔 0天 ▼ 0小时 ▼ 0分 ▼ ● 一次性执行 确定 取消	

图 3-10

然后在弹出框中详细填写其中的配置项,点击"^{确定}",采集计划就添加成功了。

## 3.2.8 采集记录管理

当从数据库采集者想要对已经创建的任务进行采集的记录管理时,只要点击任务记录后的^[查看记录],就进入到采集记录管理界面。在该页面中可以删除和清空记录,注意:在这里删除的记录不会影响已经采集到栏目文章源的记录。如图 3-11:



	网站建设 网站管理	念切	换到公开目录系统   📑 切换到	副内容管理   •
	备份恢复   用户和权限   网站统计   网站安全   <mark>文章采集</mark>   其他设置		🔎 预览网站>> 📔 🛃 发科	佈网站   🌌 [:]
文章采集	从数据库采集			
₽ 2 2 2 2 3 4 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 <	🗣 当前位置: 文章采集>>从数据库采集>>任务设置>>采集记录管理			ß
◎ ■ → → → → → → → → → → → → → → → → → →				25
				1×
	共14条 每页 10 ∨条 1/2 上页 1 2 下页 转到 页			查
	□ 标题	采集标志	抓取日期	栏目名称
	□ 物产中大私洽特斯拉 酝酿并购浙商基金	已采集	2014-03-04 17:49:20	IT业界
一一一一回收站	□ 苹果CarPlay是天使与恶魔的共同体	已采集	2014-03-04 17:49:17	IT业界
□□↓□↓↓数据库采集	□ 福布斯全球富豪榜出炉:盖茨重夺首富桂冠	已采集	2014-03-04 17:49:14	IT业界
	□ 微软确认托尼-贝茨等离职: 埃洛普或也离职	已采集	2014-03-04 17:49:11	IT业界
■「日本応制台	□ 微软将向企业推OneDrive for Business服务	已采集	2014-03-04 17:49:07	IT业界
	□ 微软语音助手Cortana界面曝光: WP8.1将集成	已采集	2014-03-04 17:48:58	IT业界
	□ 三星奥斯卡推广费达2000万美元:含植入广告	已采集	2014-03-04 17:48:55	IT业界
	□ Office365将引入社交功能和机器学习技术	已采集	2014-03-04 17:48:47	IT业界
	□ 微软3月8日起通过弹窗通知提示XP用户升级	已采集	2014-03-04 17:48:44	IT业界
	□ 微软CEO发邮件谈管理层调整: 要团结一心	已采集	2014-03-04 17:48:41	IT业界
14	注意:采集过的内容将在这里保持记录,改变采集记录的状态将导致系统重新采集内	内容。		

图 3-11

# 3.3 任务控制台

任务控制台中主要提供了对采集任务的启动和停止管理,。所有的任务想要执行,都要在 任务控制台进行启动。在此需要说明的是,如果信息采集者建立了好多采集任务,但是想要 让指定的任务执行时,必须在任务设置中,把其他不执行的任务的状态改为停用状态,然后 采集。如果在采集的过程中,想要停止采集,直接点击"^{停止采集}",采集任务就会终止。 如图 3-12:

	网站建设 <b>网站管理</b>	🧏 切换到公开目录系统   📑 切换到内容管
	备份恢复   用户和权限   网站统计   网站安全   <mark>文章采集</mark>   其他设置	🔎 预览网站>>   📑 发布网站
<b>文章采集</b> → 文章采集 → 一、从网页采集 → 一、分子设置 → 一、分子设置 → 一、分子支制台 → 一、一、分子支制台 → 一、一、分子支制台 → 一、一、分子支制台 → 一、一、分子支制台 → 一、一、分子支制台 → 一、一、分子支制台 → 一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、		💆 预览网站>>   💽 发布网站

图 3-12

# 3.4 任务日志

此处的功能主要是详细记录采集的操作情况,例如: 启用一个任务并执行,此处会记录

**西安博达软件股份有限公司** 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 40 页 共 41 页



## 执行任务过程中出现的操作。如图 3-13:

	网站建设 网站管理	🧏 切换到公开目录系统 📗 📑
	备份恢复   用户和权限   网站统计   网站安全	│ <mark>文章采集</mark> │ 其他设置
文章采集	从数据库采集	
会议章采集	💊 当前位置:文章采集>>从数据库采集>>任务设置	2
■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□		
	共6条 每页 50 ∨条 1/1 上页 1 下页 转到	页
	日 时间	事件
	2014-03-04 16:51:18	文件下载 (列表)执行完成
一一一回收站	2014-03-04 16:51:18	文件下载 (列表)内容链接抓取完成,共抓取29条
	2014-03-04 16:50:55	文件下载 (列表)抓取条数已达到任务设置的最大限制,任务完成。
	2014-03-04 16:50:55	开始抓取 文件下載 (列表)
	2014-03-04 16:50:55	文件下载 (列表)执行完成
	2014-03-04 16:50:54	文件下载 (列表)内容链接抓取完成,共抓取49条

图 3-13

此功能的作用是,当采集任务出现问题时,便于我们查找问题、进行排错。当采集管理 者不想要某条日志在此处显示时,可以选中日志,点击"^{删除日志}"就可以了。如果管理 者想要删除所有的日志时,点击"^{清空日志}",此刻页面中的所有日志就被清空了。除此之 外,此处还配置了模糊查询功能。以方便管理者查询记录。

## 3.5 回收站

在如下图所示的页面中显示删除任务信息记录,在这里可以对已经删除的记录进行恢复操作只有选中记录点击"^{恢复任务}"按钮就可以了,如果确定某条记录要彻底删除在下图中点击"^{彻底删除任务}"就对该记录进行了彻底删除操作。

	网站建设 网站管理		🕱切换到公开目录系统	📑 切换到内容管理 丨 🐗 🗄
	备份恢复   用户和权限   网站统计   网站安全	:  文章采集   其他设置	▶️预览网站>>	🗟 发布网站   🦉 访问
文章采集	从数据库采集			
₽ 2 2 2 2 3 4 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 <	🗣 当前位置:文章采集>>从数据库采集>>回收站			
	恢复任务 彻底删除任务			
		で生たをちち		12 12 1 1 <del>1</del>
		米集性分泌杯		壮穷状态
□-□-□				
──────────────────────────────────────				
一一一〇日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日				

图 3-14

**西安博达软件股份有限公司** 地址:中国・西安市高新区锦业路 125 号第 201 幢 13 层 01 号 邮编:710077 XI'AN WEBBER SOFTWARE CO., LTD. 电话:+86-29-81021800 传真:+86-29-81021820 网址:www.chinawebber.com 第 41 页 共 41 页