保密资料
禁止外传

用户手册

——文章采集

西安博达软件股份有限公司

目 录

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。

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：

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：

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：

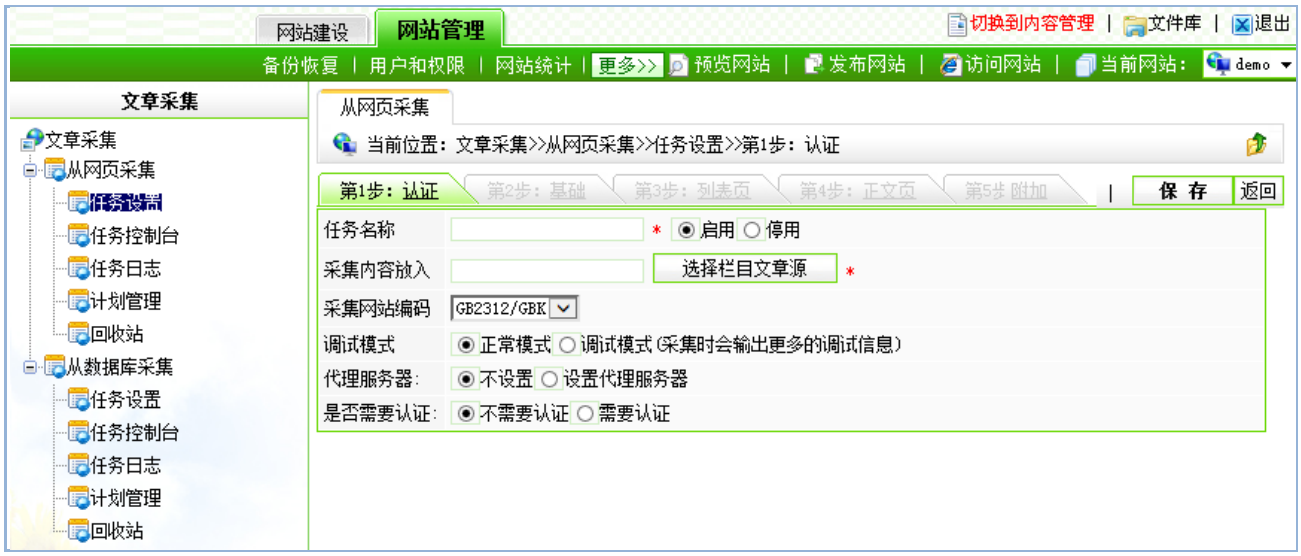

图 2-3

【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。

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：

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：

图 2-6

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

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：

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“ 预览列表页 ”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：

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：

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 ○ 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 ○ 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：○ 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。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：

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：

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 1 2 3 4 5 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：

图 2-13

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：

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：

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“ 预览标题 ”可以预览所标记的标题。

【正文提取方式】分为： 自动提取和 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：

图 2-16

【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：不恢复和 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：

图 2-17

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：

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 提取内容中的图片作为图片链接。如不需要的话，可以选择 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：

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：

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 当前日期。如需要采集原文章的信息，则选择 从页面采集，在弹出的指定位置处，输入标记即可。

【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 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 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：

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 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 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“ 预览摘要 ”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 不采集。如图 2-22：

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 RSS 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)，

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“ 预览正文页 ”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：

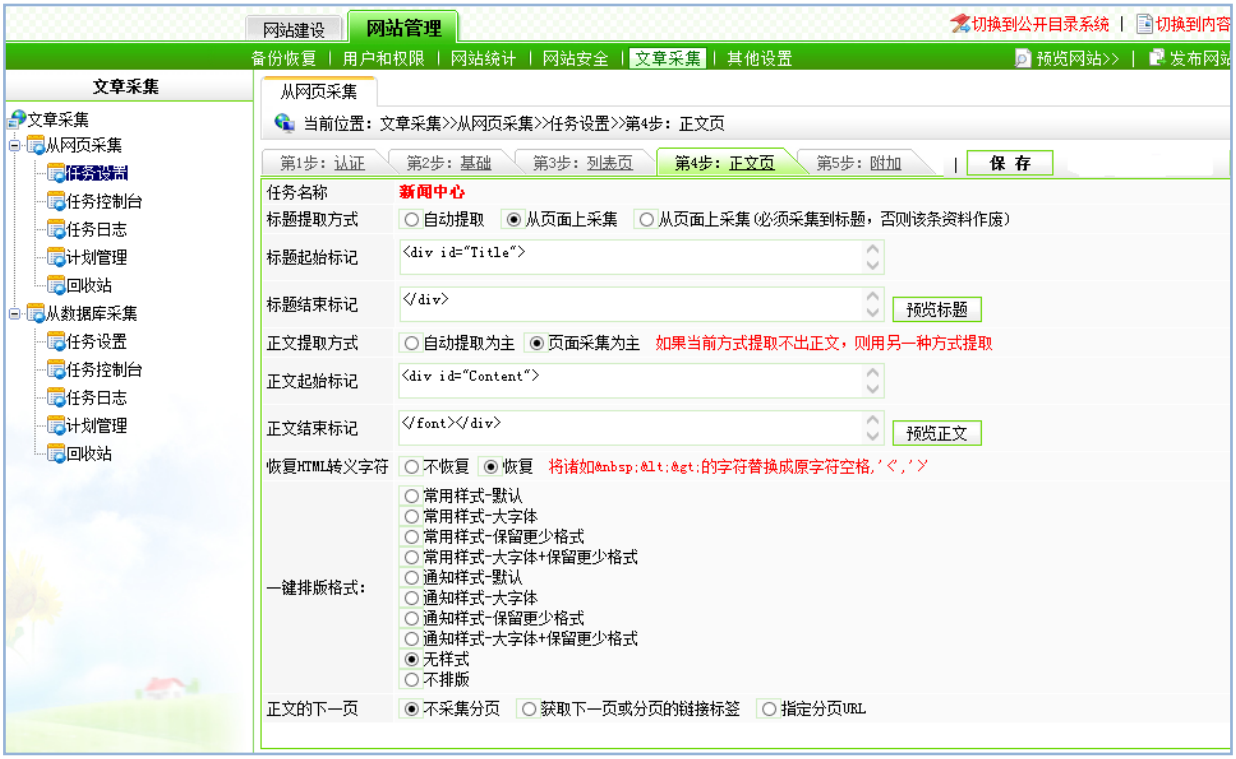

图 2-24

【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：

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

当从网页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“ 删除任务 ”，此采集任务就被删除了。

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“ 启用任务 ”，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。

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：

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:

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

由于本系统执行采集任务时，都会在采集记录管理中记录，如果下次再执行此任务时，采集记录管理中有采集记录的，就会跳过此记录不做采集。所以如果采集者想要重新采集某条或者某些条记录中的文章时，必须在此处把这些采集记录删除，再执行任务，系统会把这些记录中的文章重新再采集下来。具体的操作是，选中要删除的记录，点击“ 删除采集记录 ”，记录就删除掉了。

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：

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。

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：

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。

图 2-31

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

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、mssql2000、sybase、db2、mysql 等. 以下以 mssql2000 数据库为例，介绍一下数据库的配置操作，如图 3-2 所示：

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（mssql2000 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：

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：

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：

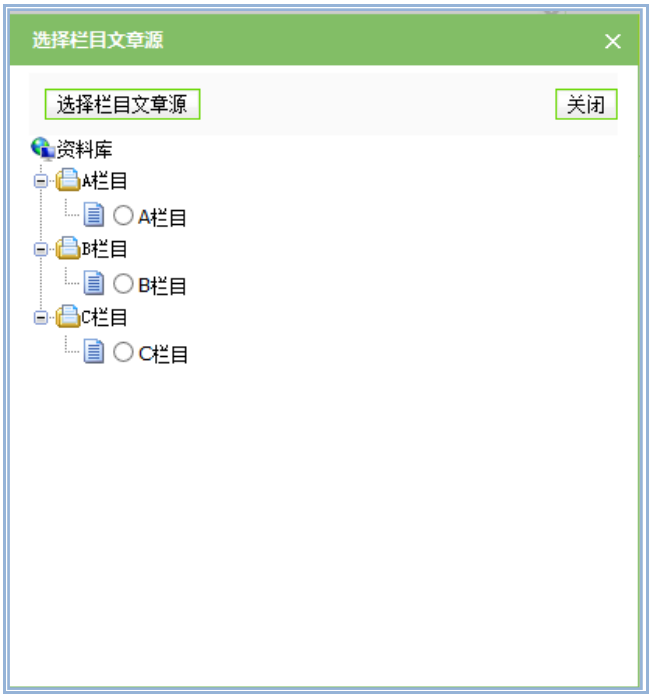

图 3-5

【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：

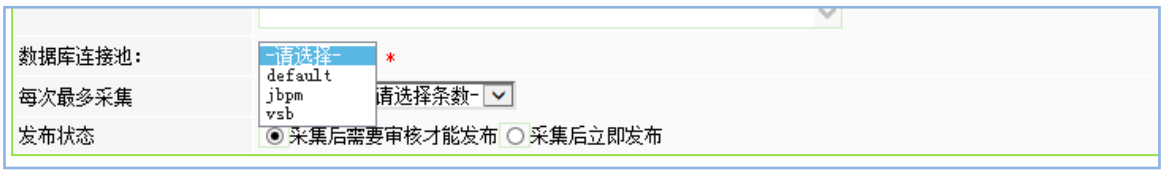

图 3-6

【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：

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：

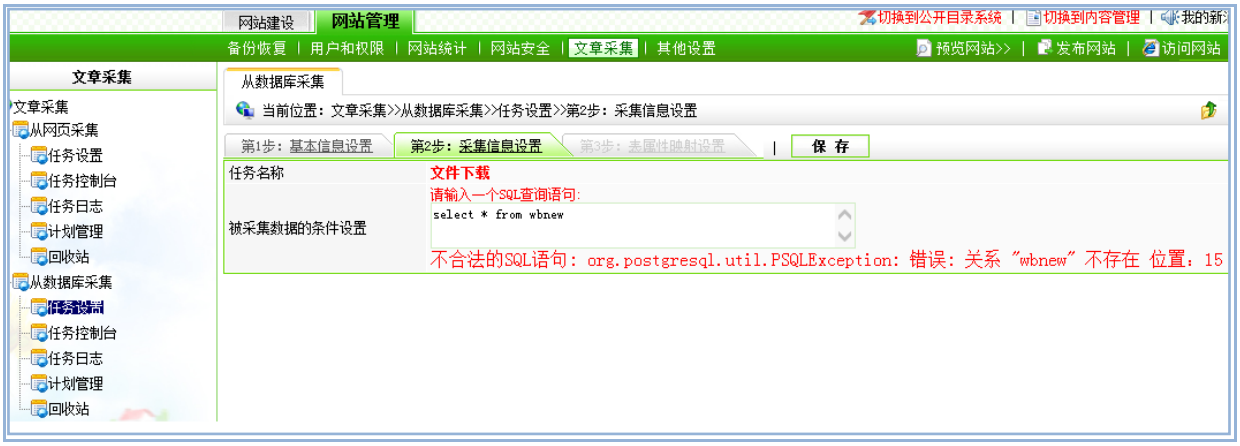

图 3-8

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：

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“ 删除任务 ”，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：

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：

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：

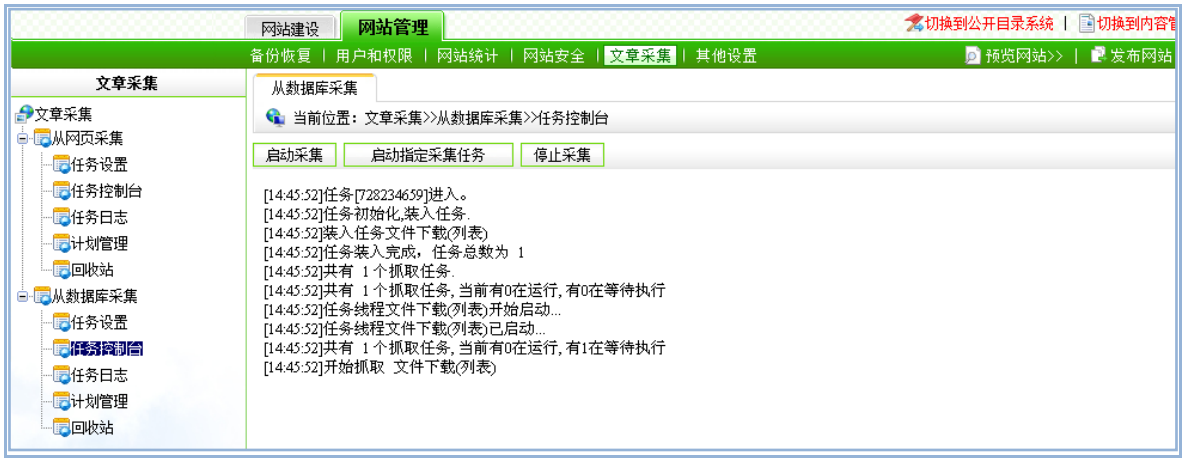

图 3-12

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

执行任务过程中出现的操作。如图 3-13：

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。

图 3-14