

(19) 中华人民共和国国家知识产权局



## (12) 发明专利申请

(10) 申请公布号 CN 104965901 A

(43) 申请公布日 2015. 10. 07

(21) 申请号 201510375465. X

(22) 申请日 2015. 06. 30

(71) 申请人 北京奇虎科技有限公司

地址 100088 北京市西城区新街口外大街  
28号D座112室(德胜园区)

申请人 奇智软件(北京)有限公司

(72) 发明人 黄钊

(74) 专利代理机构 北京润泽恒知识产权代理有  
限公司 11319

代理人 苏培华

(51) Int. Cl.

G06F 17/30(2006. 01)

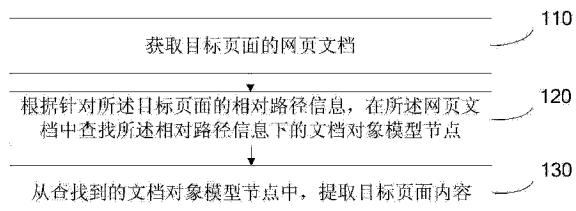
权利要求书2页 说明书14页 附图4页

### (54) 发明名称

一种目标页面内容抓取方法和装置

### (57) 摘要

本发明公开了一种目标页面内容抓取方法和装置,涉及网页技术领域。所述方法包括:获取目标页面的网页文档;根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建;从查找到的文档对象模型节点中,提取目标页面内容解决了字符串匹配方式下很难准确区分目标元素信息特点,很难对抓取后的结果进行解析的问题,以及绝对路径的匹配方式下由于网页面对JS脚本延迟加载的因素,导致绝对路径不统一而抓取过程无法正常执行的问题,取得了抓取目标页面内容时,不受页面小范围变动干扰,抓取规则配置简单,并且可以提高抓取效率的有益效果。



1. 一种目标页面内容抓取方法,包括 :

获取目标页面的网页文档 ;

根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点 ;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建 ;

从查找到的文档对象模型节点中,提取目标页面内容。

2. 根据权利要求 1 所述的方法,其特征在于,所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括 :

由至少一个文档对象模型节点的标签与属性构建的相对路径信息 ;

和 / 或者,由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签,以 XPATH 形式构建的相对路径信息。

3. 根据权利要求 1 所述的方法,其特征在于,所述获取目标页面的网页文档包括 :

根据列表页面的链接地址,获取列表页面的网页文档。

4. 根据权利要求 3 所述的方法,其特征在于,所述根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点,包括 :

根据针对列表页面中列表区域的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点。

5. 根据权利要求 4 所述的方法,其特征在于,所述从查找到的文档对象模型节点中,提取目标页面内容包括 :

从列表页面的列表区域对应的文档对象模型节点标签中,提取各个资源页面的链接地址。

6. 根据权利要求 1 或 5 所述的方法,其特征在于,所述获取目标页面的网页文档包括 :根据资源页面的链接地址,获取资源页面的网页文档。

7. 根据权利要求 6 所述的方法,其特征在于,所述根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点,包括 :

根据针对资源页面的各资源内容的相对路径信息,在所述网页文档中查找各相对路径信息下的文档对象模型节点标签。

8. 根据权利要求 1 所述的方法,其特征在于,所述从查找到的文档对象模型节点中,提取目标页面内容,包括 :

根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式,从所述文档对象模型节点中,提取目标页面内容。

9. 一种目标页面内容抓取装置,包括 :

网页文档获取模块,适于获取目标页面的网页文档 ;

节点查询模块,适于根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点 ;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建 ;

内容提取模块,适于从查找到的文档对象模型节点中,提取目标页面内容。

10. 根据权利要求 9 所述的装置,其特征在于,所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括 :

由至少一个文档对象模型节点的标签与属性构建的相对路径信息构建的相对路径信息；

和 / 或者，由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签，以 XPATH 形式构建的相对路径信息。

## 一种目标页面内容抓取方法和装置

### 技术领域

[0001] 本发明涉及网页技术领域，具体涉及一种目标页面内容抓取方法和装置。

### 背景技术

[0002] 随着互联网的发展，越来越多的用户通过互联网获取各种信息。而在互联网中，网站很多，并且网站内部的网页更为庞大，用户如果想要了解某方面的内容，可能需要访问多个网站的多个网页，才能浏览到其需要的内容。

[0003] 针对上述情况，为了方便用户对目标页面内容的访问，产生了通过一些采集其将各个网站的某方面的目标页面内容进行抓取，然后用户即可直接浏览采集到的内容。用户不用逐个访问页面浏览目标页面内容。

[0004] 但是，在先技术中，通常使用的采集工具如八爪鱼采集器、狂人采集器、CMS 系统 (Content Management System, 内容管理系统) 采集模块等，其采用的目标页面内容抓取时，需要预先设定对 html 文档代码的匹配方式，该匹配方式均是通过流量 html 文档代码确定，匹配方式大致包括两种：

[0005] 其一是配置字符串匹配方式，该种方式是直接对网页文档的目标元素的开始和结束字符进行匹配，但是实际中，网页页面随文章变动大，内容区域复杂，单纯依赖简单的字符串匹配很难准确区分目标元素信息特点，很难对抓取后的结果进行解析。

[0006] 其二是配置绝对路径的匹配方式，该种方式是从网页文档的 body 开始进行匹配，该种方式可以较为准确的抓取到目标页面内容，但是由于网页面对 JS 脚本存在延迟加载的因素，并且抓取脚本没有浏览器内核的解析功能，只能获取初始的 html 代码，无法执行 html 代码中的 JS 脚本，因而抓取脚本和实际浏览器获取的目标页面内容具有差异，导致绝对路径不统一而抓取过程无法正常执行。

### 发明内容

[0007] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的目标页面内容抓取装置和相应的目标页面内容抓取方法。

[0008] 依据本发明的一个方面，提供了一种目标页面内容抓取方法，包括：

[0009] 获取目标页面的网页文档；

[0010] 根据针对所述目标页面的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点；其中，所述相对路径信息基于文档对象模型节点的属性相关信息构建；

[0011] 从查找到的文档对象模型节点中，提取目标页面内容。

[0012] 优选的，所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括：

[0013] 由至少一个文档对象模型节点的标签与属性构建的相对路径信息；

[0014] 和 / 或者，由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签，以 XPATH 形式构建的相对路径信息。

- [0015] 优选的，所述获取目标页面的网页文档包括：
- [0016] 根据列表页面的链接地址，获取列表页面的网页文档。
- [0017] 优选的，所述根据针对所述目标页面的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点，包括：
- [0018] 根据针对列表页面中列表区域的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点。
- [0019] 优选的，所述从查找到的文档对象模型节点中，提取目标页面内容包括：
- [0020] 从列表页面的列表区域对应的文档对象模型节点标签中，提取各个资源页面的链接地址。
- [0021] 优选的，所述获取目标页面的网页文档包括：
- [0022] 根据资源页面的链接地址，获取资源页面的网页文档。
- [0023] 优选的，所述根据针对所述目标页面的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点，包括：
- [0024] 根据针对资源页面的各资源内容的相对路径信息，在所述网页文档中查找各相对路径信息下的文档对象模型节点标签。
- [0025] 优选的，所述从查找到的文档对象模型节点中，提取目标页面内容，包括：
- [0026] 根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式，从所述文档对象模型节点中，提取目标页面内容。
- [0027] 依据本发明的另外一方面，提供了一种目标页面内容抓取装置，包括：
- [0028] 网页文档获取模块，适于获取目标页面的网页文档；
- [0029] 节点查询模块，适于根据针对所述目标页面的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点；其中，所述相对路径信息基于文档对象模型节点的属性相关信息构建；
- [0030] 内容提取模块，适于从查找到的文档对象模型节点中，提取目标页面内容。
- [0031] 优选的，所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括：
- [0032] 由至少一个文档对象模型节点的标签与属性构建的相对路径信息构建的相对路径信息；
- [0033] 和 / 或者，由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签，以 XPATH 形式构建的相对路径信息。
- [0034] 优选的，所述网页文档获取模块包括：
- [0035] 列表文档获取模块，适于根据列表页面的链接地址，获取列表页面的网页文档。
- [0036] 优选的，所述节点查询模块，包括：
- [0037] 列表节点查找模块，适于根据针对列表页面中列表区域的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点。
- [0038] 优选的，所述内容提取模块包括：
- [0039] 资源链接获取模块，适于从列表页面的列表区域对应的文档对象模型节点标签中，提取各个资源页面的链接地址。
- [0040] 优选的，所述网页文档获取模块包括：
- [0041] 资源文档获取模块，适于根据资源页面的链接地址，获取资源页面的网页文档。

[0042] 优选的，所述节点查询模块，包括：

[0043] 资源节点查询模块，适于根据针对资源页面的各资源内容的相对路径信息，在所述网页文档中查找各相对路径信息下的文档对象模型节点标签。

[0044] 优选的，所述内容提取模块，包括：

[0045] 再提取模块，适于根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式和 / 或前后表达式，从所述文档对象模型节点中，提取目标页面内容。

[0046] 根据本发明的目标页面内容抓取方法可以针对各个目标网页，基于文档对象模型节点的属性相关信息构建针对目标网页的相对路径信息，然后在所述网页文档中查找所述相对路径信息下的文档对象模型节点，从而可从查找到的文档对象模型节点中，提取目标页面内容，由此解决了字符串匹配方式下很难准确区分目标元素信息特点，很难对抓取后的结果进行解析的问题，以及绝对路径的匹配方式下由于网页面对 JS 脚本延迟加载的因素，导致绝对路径不统一而抓取过程无法正常执行的问题，取得了抓取目标页面内容时，不受页面小范围变动干扰，抓取规则配置简单，并且可以提高抓取效率的有益效果。

[0047] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

## 附图说明

[0048] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0049] 图 1 示出了根据本发明一个实施例的一种目标页面内容抓取方法的流程示意图；

[0050] 图 1A 示出了本发明一个实施例的 html 文档的 DOM 树结构示例。

[0051] 图 2 示出了根据本发明一个实施例的另一种目标页面内容抓取方法的流程示意图；

[0052] 图 3 示出了根据本发明一个实施例的另一种目标页面内容抓取方法的流程示意图；

[0053] 图 3A 示出了根据本发明一个实施例的列表页面的配置界面示例；

[0054] 图 3B 示出了根据本发明一个实施例的资源页面的配置界面示例；

[0055] 图 4 示出了根据本发明一个实施例的一种目标页面内容抓取装置的结构示意图；

[0056] 图 5 示出了根据本发明一个实施例的另一种目标页面内容抓取装置的结构示意图；以及

[0057] 图 6 示出了根据本发明一个实施例的另一种目标页面内容抓取装置的结构示意图。

## 具体实施方式

[0058] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围

完整的传达给本领域的技术人员。

[0059] 本发明的核心思想之一在于，目标页面内容抓取方法可以针对各个目标网页，基于文档对象模型节点的属性相关信息构建针对目标网页的相对路径信息，然后在所述网页文档中查找所述相对路径信息下的文档对象模型节点，从而可从查找到的文档对象模型节点中，提取目标页面内容。同时避免了字符串匹配方式和绝对路径的匹配方式产生的问题，本发明实施例不受页面小范围变动干扰，抓取规则配置简单，并且可以提高抓取效率。

[0060] 实施例一

[0061] 参照图 1，其示出了本发明一种目标页面内容抓取方法的流程示意图，具体可以包括：

[0062] 步骤 110，获取目标页面的网页文档；

[0063] 在本发明实施例中，通过在执行目标页面内容抓取过程之前，即启用抓取脚本之前，会配置针对目标页面的相关信息，比如目标页面的链接地址，针对目标页面的相对路径信息，该相对路径信息用于在目标页面的网页文档中查找目标页面内容所在位置。

[0064] 当然，本发明实施例中，本发明可以提供配置界面，该配置界面包括目标页面链接地址的配置栏，相对路径的配置栏等，在用户确定之后，生成最终的抓取脚本。

[0065] 在用户执行抓取脚本之后，首先根据目标页面的链接地址，去对应的服务器中获取目标页面的网页文档，如 html (Hypertext Markup Language, 超文本标记语言) 文档。然后抓取脚本对 html 文档中的代码进行后续匹配和提取过程。

[0066] 当然，本发明实施例中，目标页面的链接地址也可以通过其他方式获得，比如导入目标页面的链接地址，本发明不对其加以限制。

[0067] 步骤 120，根据针对所述目标页面的相对路径信息，在所述网页文档中查找所述相对路径信息下的文档对象模型节点；其中，所述相对路径信息基于文档对象模型节点的属性相关信息构建；

[0068] 在本发明实施例中，可预先配置针对所述目标页面的相对路径信息，该相对路径信息基于文档对象模型节点的属性相关信息构建。

[0069] 对于一个网页文档，其可以解析为 DOM (Document Object Model, 文档对象模型) 树，DOM 树中有各个 DOM 节点，而节点本身可设置属性，如设置 id 或 class 属性。如图 1A，其为一个 html 文档的 DOM 树解析图示例。以 html 的 html 为根节点，基于 html 语言的标签规则和标签之间的父子关系，逐层解析得到如图 1 的 DOM 树。其中如 body、div、ul、li、head、meta、title 等都是 DOM 节点，class 的值为相应 DOM 节点的属性。

[0070] 本发明实施例则直接以 DOM 节点的属性作为参照，构建针对上述目标页面的相对路径信息。

[0071] 在本发明实施例中，在配置相对路径信息之前，可以通过 DOM 解析插件，将 html 文档解析为可视化的 DOM 树，用户可以通过浏览该 DOM 树，确定需求的目标页面内容相关的 DOM 节点的 DOM 节点名和 DOM 节点属性，从而构建相对路径信息。

[0072] 优选的，所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括：

[0073] A1，由至少一个文档对象模型节点的标签与属性构建的相对路径信息构建的相对路径信息；

[0074] 在本发明实施例中 DOM 节点的属性相关信息，可以理解为 DOM 节点的标签和其具

体的属性，在本发明实施例中 DOM 节点的具体属性为 id 或 class 属性，比如 DOM 节点的代码为 `<ul class = "clearfix">`，那么其标签为 ul，其属性为 `class = "clearfix"`。

[0075] 在实际应用中，对于 html 文档中的目标页面内容所在 DOM 节点，其 DOM 节点的属性相关信息，即 DOM 节点的标签 + 属性，可能是唯一的，比如对于目标页面内容所在的 DOM 节点 `<ul class = "clearfix">`，在 html 文档中只存在一个，那么可直接以该 DOM 节点的标签 `ul` 和属性 `class = "clearfix"` 构建相对路径信息，即可定位该 DOM 节点。

[0076] 在本发明实施例中，可在相对路径配置界面中输入 `ul[class = clearfix]` 的相对路径信息。

[0077] 对于 html 文档中的目标页面内容所在 DOM 节点，其 DOM 节点的属性相关信息，即 DOM 节点的标签 + 属性，可能不是唯一的，其还有其他具备相同 DOM 节点的标签 + 属性的 DOM 节点记录了其他的目标页面内容。那么如果只以需求的 DOM 节点的 DOM 节点的标签 + 属性进行定位，则不够精确。那么本发明实施例，则基于目标页面内容所在的 DOM 节点，基于 DOM 节点的父子关系，向上游的父级子标签逐级确定父级 DOM 节点的标签 + 属性，然后构建一个相对路径信息。比如对于目标页面内容所在的 DOM 节点 `<ul class = "clearfix">`，在 html 文档中存在多个，那么向结合类似图 1A 的 DOM 树，可以确定 `<ul class = "clearfix">` 上一级的父 DOM 节点，比如为 `<div class = "fire">`，如果 `<DIV class = "fire">` 在 html 文档中是唯一的，则以该两个 DOM 节点的父子关系构建相对路径信息，如 `div[class = fire]-UI[class = clearfix]`；如果 `<div class = "fire">` 在 html 文档中不是唯一的，则继续获取向上一级的父 DOM 节点的标签 + 属性，类似构建相对路径信息。

[0078] 在实际应用中，一般各个 html 文档中承载具体的目标页面内容的 DOM 节点中，其 DOM 节点标签 + DOM 节点属性是唯一的，因此只用目标页面内容的 DOM 节点标签 + DOM 节点属性即可确定相对路径信息，配置过程简单。

[0079] 和 / 或者，A2，由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签，以 XPATH 形式构建的相对路径信息。其中，XPATH 是一种 W3C 的路径表达式构建方法。

[0080] 在本发明实施例中，可能某些 DOM 节点的属性相关信息只有 DOM 节点标签，而未有属性。比如图 1A 中的 `li` 标签，只有标签而没有属性，那么为了将 `li` 节点下的目标页面内容获取到，由于 `li` 没有具体的属性，则无法定位具体的 `li` 节点。因此，需要采用 XPATH 形式的命令，比如前述 `li`，由于有多个，其是在 `<ul class = "list">` 这个 DOM 节点之下，那么可以从 `li` 节点向上一级的父 DOM 节点一起构建相对路径信息，如 `ui[class = list]/li`，如此可以定位获取 `<ul class = "list">` 节点下的各个 `li` 节点。

[0081] 在本发明实施例中，目标页面内容所在的 DOM 节点的属性相关信息不同，如 A1、A2 的示例，本发明可以同时采用 A1、A2，也可以单独采用 A1 或 A2，具体根据实际情况选择。

[0082] 在实际应用中，html 文档的代码示例如下：

[0083] `<html>`

[0084] `<body>`

[0085] `<div class = my>`

[0086] `<h1>My First Heading</h1>`

[0087] `<p>My first paragraph.</p>`

[0088] </div>  
[0089] <div class = you>  
[0090] <h1>tutu</h1>  
[0091] </div>  
[0092] </body>  
[0093] </html>

[0094] 本发明实施例则根据针对该目标页面的相对路径信息,去 html 文档中匹配查询相应的 DOM 节点,比如相对路径信息 div[class = list],则匹配到上述 html 代码中的<div class = my>。

[0095] 可以理解,在本发明实施例中,用户可以针对不同的目标页面内容设置相应的相对路径信息,从而可以查找 html 文档中相应的文档对象模型节点。

[0096] 步骤 130,从查找到的文档对象模型节点中,提取目标页面内容。

[0097] 如前述 html 示例,根据 html 代码的规则,一个起始的标签,对应有个结束的标签,那么<div class = my>对应结束的标签为</div>,那么两者之间即为目标页面内容所在。从而本发明实施例可以从该 DOM 节点中提取目标页面内容。

[0098] 当然,本发明实施例中,对于提取的目标页面内容,还可以做进一步处理。优选的,还包括:

[0099] 步骤 140,根据预置的字符转换规则,将目标页面内容中的字符进行转换。

[0100] 比如设置将“多玩”转换为“360 手游网”的字符转换规则可为:“多玩”=>“360 手游网”。然后从目标页面内容中查找“多玩”,将多玩替换为“360 手游网”。

[0101] 上述过程可以将用户不想保留的字符进行替换,更灵活的适配用户的需求。

[0102] 本发明实施例的目标页面内容抓取方法可以针对各个目标网页,基于文档对象模型节点的属性相关信息构建针对目标网页的相对路径信息,然后在该网网页文档中查找与相对路径信息下的文档对象模型节点,从而可从查找到的文档对象模型节点中,提取目标页面内容,由此解决了字符串匹配方式下很难准确区分目标元素信息特点,很难对抓取后的结果进行解析的问题,以及绝对路径的匹配方式下由于网页页面对 JS 脚本延迟加载的因素,导致绝对路径不统一而抓取过程无法正常执行的问题,取得了抓取目标页面内容时,不受页面小范围变动干扰,抓取规则配置简单,并且可以提高抓取效率的有益效果。

[0103] 实施例二

[0104] 参照图 2,其示出了本发明一种目标页面内容抓取方法的流程示意图,具体可以包括:

[0105] 步骤 210,获取目标页面的网网页文档;

[0106] 步骤 220,根据针对所述目标页面的相对路径信息,在所述网网页文档中查找所述相对路径信息下的文档对象模型节点;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建;

[0107] 步骤 230,根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式,从所述文档对象模型节点中,提取目标页面内容。

[0108] 本发明实施例中,在从查找到的文档对象模型节点中,提取目标页面内容时,为了更精确的提取用户需要的内容,过滤掉不需要的内容。

[0109] 在本发明实施例中,可通过针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式,从查找到的文档对象模型节点中,提取目标页面内容。

[0110] 具体的,上述正则匹配表达式又称正规表示法、常规表示法,正则匹配表达式使用单个字符串来描述、匹配一系列符合某个句法规则的字符串。

[0111] 比如,如果要匹配 DOM 节点中的链接,避免获取 DOM 节点中的其他无关项。比如 DOM 节点代码为:

[0112] <div><a href = "http://ng.d.cn/daotachuanqi/news/detail\_413625\_1.html" target = "\_blank" title = "刀塔传奇仙女龙使用及克制仙龙女看过来">刀塔传奇仙女龙使用及克制仙龙女看过来</a></div>

[0113] 如果只想提取链接地址,则可以设置正则匹配表达式  $(\w+:\//\w+\.\w+.\w+\/\w+\/\w+\/\w+_\w+_\w+.\w+)$ ,其中 \w+ 表示匹配任何字类字符, \: 表示匹配“:”, \/ 表示匹配“/”, \. 表示匹配“.”, \\_ 表示匹配“\_”。如此,上述正则匹配表达式则能匹配链接地址 http://ng.d.cn/daotachuanqi/news/detail\_413625\_1.html。

[0114] 具体的,上述前后匹配表达式,可以理解为以一个字符串为起点和一个字符串终点进行匹配对象。如上述 DOM 节点,如果只想提取链接地址,可以设置前后匹配表达式为 [http]-[html],其中起点的字符串为 http,终点的字符串为 html。那么从 DOM 节点的内容中开始匹配后,首先找到 http 所在位置,那么记录 http 及之后的字符串,直到匹配到 html 结束。

[0115] 当然,本发明实施例中,由于提取的 html 文档中 DOM 节点可能有多个,对提取时采用的方式可能也有多种,比如对某个 DOM 节点直接提取节点内容,对另外一个 DOM 节点根据预置针对目标页面内容的正则匹配表达式提取节点内容,对另外一个 DOM 节点根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式提取节点内容。从而,本发明可以在一次抓取过程中使用上述提取方式中的一种或者多种。

[0116] 本发明实施例的目标页面内容抓取方法可以针对各个目标网页,基于文档对象模型节点的属性相关信息构建针对目标网页的相对路径信息,然后在所述网页文档中查找所述相对路径信息下的文档对象模型节点,从而可从查找到的文档对象模型节点中,提取目标页面内容,并且在提取目标页面内容时可通过正则匹配表达式和 / 或前后匹配表达式提取。由此解决了字符串匹配方式下很难准确区分目标元素信息特点,很难对抓取后的结果进行解析的问题,以及绝对路径的匹配方式下由于网页页面对 JS 脚本延迟加载的因素,导致绝对路径不统一而抓取过程无法正常执行的问题,取得了抓取目标页面内容时,不受页面小范围变动干扰,抓取规则配置简单,并且可以提高抓取效率的有益效果,并且通过正则匹配表达式和 / 或前后匹配表达式可以进一步过滤不需要的内容,提高目标页面内容的精确度。

[0117] 实施例三

[0118] 参照图 3,其示出了本发明一种目标页面内容抓取方法的流程示意图,具体可以包括:

[0119] 步骤 310,根据列表页面的链接地址,获取列表页面的网页文档;

[0120] 在本步骤中,述及的列表页面为实施例一及的目标页面为列表页面。

[0121] 在本发明实施例中步骤 310 之前,可以先配置列表页面的链接地址、针对列表页

面中列表区域的相对路径信息,对DOM节点的提取规则等。如图3A,用户可以在图3A的列表配置页中,配置网站名称“站点”,网站的栏目名称“专区”,该栏目的列表页的网址http://ng.d.cn/wushuangjianji/news/list\_walkthrough\_1.html。和资源页面的网址链接的相对路径信息ul[class = znewsList]。

[0122] 在本发明实施例中,上述相对路径信息的配置,可以通过查看http://ng.d.cn/wushuangjianji/news/list\_walkthrough\_1.html的html代码或者DOM树确定资源页面的网址链接所在的DOM节点<ul class = "znewsList">,然后将ul[class = znewsList]写入图3的配置页的“列表”栏中。

[0123] 在本发明实施例中,列表配置页的“列表”栏中,设置了属性、XPATH、正则、前后匹配选择按钮。选择属性按钮时,“列表”栏对应接收用户输入的由至少一个文档对象模型节点的标签与属性构建的相对路径信息,如ul[class = znewsList]。选择XPATH属性按钮时,“列表”栏对应接收用户输入的由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签,以XPATH形式构建的相对路径信息,如ui[class = list]/li。选择正则按钮时,“列表”栏对应接收针对目标页面内容的正则匹配表达式,用于对从所述文档对象模型节点中的内容进行匹配,如(\w+\:\.\w+\.\w+\.\w+\.\w+\.\w+\.\w+\.\w+\.\w+)。选择前后匹配按钮时,“列表”栏对应接收针对目标页面内容的前后匹配表达式,用于对从所述文档对象模型节点中的内容进行匹配,如[http]-[html]。

[0124] 如果用户不输入相应内容,则默认为空。

[0125] 在用户确定输入后,相应按钮下的输入内容,则通过接口将输入内容赋予相应的执行函数,在后续执行逻辑时调用相应执行函数进行计算。

[0126] 当然,在实际应用中,列表页面可能存在分页,那么本发明还可配置列表分页的相关信息,从而可以自动识别列表分页的信息。

[0127] 步骤320,根据针对列表页面中列表区域的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点;

[0128] 比如,根据前述ul[class = znewsList],在http://ng.d.cn/wushuangjianji/news/list\_walkthrough\_1.html的文档中查找DOM节点<ul class = "znewsList">……</ul>的代码。

[0129] 步骤330,从列表页面的列表区域对应的文档对象模型节点标签中,提取各个资源页面的链接地址。

[0130] 在本步骤中,述及的各个资源页面的链接地址为实施例一中的目标页面内容。

[0131] 然后,即可从<ul class = "znewsList">……</ul>的代码中,提取各个资源页面的链接地址。如图3A中,当用户点击测试,则在右侧自动提取了各个资源页面的连接地址。当然,如果链接地址很多,那么图3A中可显示部分链接地址。

[0132] 在本发明实施例中,上述资源页面如具体的文章页面,图片页面等详细介绍内容页面。

[0133] 步骤340,根据资源页面的链接地址,获取资源页面的网页文档。

[0134] 在本步骤中,述及的资源页面为实施例一及的目标页面为列表页面。

[0135] 在本发明实施例中,在前述步骤自动获取了资源页面的链接地址后,可逐个获取资源页面的网页文档,然后进行资源内容的抓取。

[0136] 如前述右侧的 [http://ng.d.cn/wushuangjianji/news/detail\\_406707\\_1.html](http://ng.d.cn/wushuangjianji/news/detail_406707_1.html), 可以获取其网网页文档。

[0137] 当然,在步骤 340 之前或者在步骤 310 之前,可以先配置针对资源页面各资源内容的相对路径信息,针对资源页面各资源内容的 DOM 节点的提取规则等。如图 3B,对于文章类型的资源页面,针对标题、作者、来源、简介、标签、内容分别设置属性、XPATH、正则、前后匹配选择按钮。选择属性按钮时,相应输入栏对应接收用户输入的由至少一个文档对象模型节点的标签与属性构建的相对路径信息。选择 XPATH 属性按钮时,相应输入栏对应接收用户输入的由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签,以 XPATH 形式构建的相对路径信息。选择正则按钮时,相应输入栏对应接收针对目标页面内容的正则匹配表达式,用于对从所述文档对象模型节点中的内容进行匹配。选择前后匹配按钮时,相应输入栏对应接收针对目标页面内容的前后匹配表达式,用于对从所述文档对象模型节点中的内容进行匹配。如果用户不输入相应内容,则默认为空。

[0138] 如图 3B,比如对 [http://ng.d.cn/wushuangjianji/news/detail\\_406707\\_1.html](http://ng.d.cn/wushuangjianji/news/detail_406707_1.html) 的网网页文档的 DOM 树的分析,确定标题所在的 DOM 节点为“`<div class = "article">`

[0139] `<h1>无双剑姬刷资源技巧攻略怎么快速刷资源</h1>`

[0140] `.....</div>`

[0141] 由于 `<h1>` 没有具体属性,因此从 `<h1>` 的上以及 DOM 节点构造 XPATH 的相对路径信息,即 `div[class = article]/h1`。

[0142] 对于具体内容所在的 Dom 节点为 `<div class = "articleText">.....</div>`。那么对内容所在节点的相对路径信息在属性按钮下的 `div[class = articleText]`。

[0143] 其他节点的以此类推。

[0144] 当然,对于每个目标内容,还可设置选择正则按钮设置正则匹配表达式,和 / 或选择前后匹配按钮设置前后匹配表达式。

[0145] 当然,在本发明实施例中还可在图 3A 中预置字符转换规则,然后抓取脚本在对提取目标页面内容之后,可根据预置的字符转换规则,将目标页面内容中的字符进行转换。

[0146] 比如设置将“多玩”转换为“360 手游网”的字符转换规则可为:“多玩”=>“360 手游网”。然后从目标页面内容中查找“多玩”,将多玩替换为“360 手游网”。

[0147] 步骤 350,根据针对资源页面的各资源内容的相对路径信息,在所述网网页文档中查找各相对路径信息下的文档对象模型节点标签。

[0148] 那么基于图 3B 设置的规则,可对 [http://ng.d.cn/wushuangjianji/news/detail\\_406707\\_1.html](http://ng.d.cn/wushuangjianji/news/detail_406707_1.html) 的网网页文档进行查询,分别查询 `div[class = article]/h1` 和 `div[class = articleText]` 对应的 DOM 节点。

[0149] 步骤 360,从查找到资源页面的各目标页面内容的文档对象模型节点标签中,提取目标资源内容。

[0150] 在本步骤中,述及的目标资源内容为实施例一中的目标页面内容。

[0151] 然后可根据配置的对目标资源内容的提取规则,比如正则匹配表达式的提取规则,前后匹配表达式的提取规则,从前述 DOM 节点中提取相应的目标资源内容。

[0152] 在本发明实施例中,所述目标资源内容,比如文章内容,图片等。

[0153] 本发明实施例的目标页面内容抓取方法可以先采用针对列表页面中列表区域的

相对路径信息,去列表页面的网页文档中查找文档对象模型节点,从而从该文档对象模型节点中提取所需的各个资源页面的链接地址;然后采用针对资源页面的各资源内容的相对路径信息,从这些资源链接的网页文档中查找文档对象模型节点,从而从这些中查找文档对象模型节点提取目标资源内容。由此解决了字符串匹配方式下很难准确区分目标元素信息特点,很难对抓取后的结果进行解析的问题,以及绝对路径的匹配方式下由于网页页面对JS脚本延迟加载的因素,导致绝对路径不统一而抓取过程无法正常执行的问题,取得了抓取目标页面内容时,不受页面小范围变动干扰,抓取规则配置简单,并且可以提高抓取效率的有益效果,并可以自动从列表页面中识别资源页面的链接地址,从而进行具体的资源内容的提取,进一步减少了用户的操作过程。

[0154] 实施例四

[0155] 参照图4,其示出了本发明一种目标页面内容抓取装置的结构示意图,具体可以包括:

[0156] 网页文档获取模块410,适于获取目标页面的网页文档;

[0157] 节点查询模块420,适于根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建;

[0158] 内容提取模块430,适于从查找到的文档对象模型节点中,提取目标页面内容。

[0159] 优选的,所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括:

[0160] 由至少一个文档对象模型节点的标签与属性构建的相对路径信息构建的相对路径信息;

[0161] 和/或者,由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签,以XPATH形式构建的相对路径信息。

[0162] 优选的,所述网页文档获取模块包括:

[0163] 列表文档获取模块,适于根据列表页面的链接地址,获取列表页面的网页文档。

[0164] 优选的,所述节点查询模块,包括:

[0165] 列表节点查找模块,适于根据针对列表页面中列表区域的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点。

[0166] 优选的,所述内容提取模块包括:

[0167] 资源链接获取模块,适于从列表页面的列表区域对应的文档对象模型节点标签中,提取各个资源页面的链接地址。

[0168] 优选的,所述网页文档获取模块包括:

[0169] 资源文档获取模块,适于根据资源页面的链接地址,获取资源页面的网页文档。

[0170] 优选的,所述节点查询模块,包括:

[0171] 资源节点查询模块,适于根据针对资源页面的各资源内容的相对路径信息,在所述网页文档中查找各相对路径信息下的文档对象模型节点标签。

[0172] 优选的,所述内容提取模块,包括:

[0173] 再提取模块,适于根据预置针对目标页面内容的正则匹配表达式和/或前后匹配表达式和/或前后表达式,从所述文档对象模型节点中,提取目标页面内容。

[0174] 实施例五

[0175] 参照图 5,其示出了本发明一种目标页面内容抓取装置的结构示意图,具体可以包括:

[0176] 网页文档获取模块 510,适于获取目标页面的网页文档;

[0177] 节点查询模块 520,适于根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建;

[0178] 内容提取模块 530,适于从查找到的文档对象模型节点中,提取目标页面内容,具体包括:

[0179] 第二提取模块 532,适于根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式和 / 或前后表达式,从所述文档对象模型节点中,提取目标页面内容。

[0180] 实施例六

[0181] 参照图 6,其示出了本发明一种目标页面内容抓取装置的结构示意图,具体可以包括:

[0182] 网页文档获取模块 610,具体包括:

[0183] 列表文档获取模块 612,适于根据列表页面的链接地址,获取列表页面的网页文档;进入列表节点查找模块 622;

[0184] 资源文档获取模块 614,适于根据资源页面的链接地址,获取资源页面的网页文档;进入资源节点查询模块 624;

[0185] 节点查询模块 620,具体包括:

[0186] 列表节点查找模块 622,适于根据针对列表页面中列表区域的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点;进入资源链接获取模块 632;

[0187] 资源节点查询模块 624,适于根据针对资源页面的各资源内容的相对路径信息,在所述网页文档中查找各相对路径信息下的文档对象模型节点标签;进入资源内容获取模块 634。

[0188] 内容提取模块 630,具体包括:

[0189] 资源链接获取模块 632,适于从列表页面的列表区域对应的文档对象模型节点标签中,提取各个资源页面的链接地址;进入资源文档获取模块 614;

[0190] 资源内容获取模块 634,适于从查找到资源页面的各目标页面内容的文档对象模型节点标签中,提取目标资源内容。

[0191] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述,构造这类系统所要求的结构是显而易见的。此外,本发明也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本发明的内容,并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0192] 在此处所提供的说明书中,说明了大量具体细节。然而,能够理解,本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中,并未详细示出公知的方法、结构和技术,以便不模糊对本说明书的理解。

[0193] 类似地,应当理解,为了精简本公开并帮助理解各个发明方面中的一个或多个,在

上面对本发明的示例性实施例的描述中,本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下意图:即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说,如下面的权利要求书所反映的那样,发明方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本发明的单独实施例。

[0194] 本领域那些技术人员可以理解,可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件,以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和 / 或过程或者单元中的至少一些是相互排斥之外,可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述,本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0195] 此外,本领域的技术人员能够理解,尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征,但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如,在下面的权利要求书中,所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0196] 本发明的各个部件实施例可以以硬件实现,或者以一个或者多个处理器上运行的软件模块实现,或者以它们的组合实现。本领域的技术人员应当理解,可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的目标页面内容抓取设备中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上,或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到,或者在载体信号上提供,或者以任何其他形式提供。

[0197] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制,并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中,不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中,这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

[0198] 本发明公开了 A1、一种目标页面内容抓取方法,包括:

[0199] 获取目标页面的网页文档;

[0200] 根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建;

- [0201] 从查找到的文档对象模型节点中,提取目标页面内容。
- [0202] A2、如 A1 所述的方法,所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括 :
- [0203] 由至少一个文档对象模型节点的标签与属性构建的相对路径信息 ;
- [0204] 和 / 或者,由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签,以 XPATH 形式构建的相对路径信息。
- [0205] A3、如 A1 所述的方法,所述获取目标页面的网页文档包括 :
- [0206] 根据列表页面的链接地址,获取列表页面的网页文档。
- [0207] A4、如 A3 所述的方法,所述根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点,包括 :
- [0208] 根据针对列表页面中列表区域的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点。
- [0209] A5、如 A4 所述的方法,所述从查找到的文档对象模型节点中,提取目标页面内容包括 :
- [0210] 从列表页面的列表区域对应的文档对象模型节点标签中,提取各个资源页面的链接地址。
- [0211] A6、如 A1 或 A5 所述的方法,所述获取目标页面的网页文档包括 :
- [0212] 根据资源页面的链接地址,获取资源页面的网页文档。
- [0213] A7、如 A6 所述的方法,所述根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点,包括 :
- [0214] 根据针对资源页面的各资源内容的相对路径信息,在所述网页文档中查找各相对路径信息下的文档对象模型节点标签。
- [0215] A8、如 A1 所述的方法,所述从查找到的文档对象模型节点中,提取目标页面内容,包括 :
- [0216] 根据预置针对目标页面内容的正则匹配表达式和 / 或前后匹配表达式,从所述文档对象模型节点中,提取目标页面内容。
- [0217] 本发明还公开了 B9、一种目标页面内容抓取装置,包括 :
- [0218] 网页文档获取模块,适于获取目标页面的网页文档 ;
- [0219] 节点查询模块,适于根据针对所述目标页面的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点 ;其中,所述相对路径信息基于文档对象模型节点的属性相关信息构建 ;
- [0220] 内容提取模块,适于从查找到的文档对象模型节点中,提取目标页面内容。
- [0221] B10、如 B9 所述的装置,所述基于文档对象模型节点的属性相关信息构建的相对路径信息包括 :
- [0222] 由至少一个文档对象模型节点的标签与属性构建的相对路径信息构建的相对路径信息 ;
- [0223] 和 / 或者,由至少一个文档对象模型节点的标签与属性和至少一个文档对象模型节点的标签,以 XPATH 形式构建的相对路径信息。
- [0224] B11、如 B9 所述的装置,所述网页文档获取模块包括 :

- [0225] 列表文档获取模块,适于根据列表页面的链接地址,获取列表页面的网页文档。
- [0226] B12、如B11所述的装置,所述节点查询模块,包括:
- [0227] 列表节点查找模块,适于根据针对列表页面中列表区域的相对路径信息,在所述网页文档中查找所述相对路径信息下的文档对象模型节点。
- [0228] B13、如B12所述的装置,所述内容提取模块包括:
- [0229] 资源链接获取模块,适于从列表页面的列表区域对应的文档对象模型节点标签中,提取各个资源页面的链接地址。
- [0230] B14、如B9或B13所述的装置,所述网页文档获取模块包括:
- [0231] 资源文档获取模块,适于根据资源页面的链接地址,获取资源页面的网页文档。
- [0232] B15、如B14所述的装置,所述节点查询模块,包括:
- [0233] 资源节点查询模块,适于根据针对资源页面的各资源内容的相对路径信息,在所述网页文档中查找各相对路径信息下的文档对象模型节点标签。
- [0234] B16、如B9所述的装置,所述内容提取模块,包括:
- [0235] 再提取模块,适于根据预置针对目标页面内容的正则匹配表达式和/或前后匹配表达式和/或前后表达式,从所述文档对象模型节点中,提取目标页面内容。

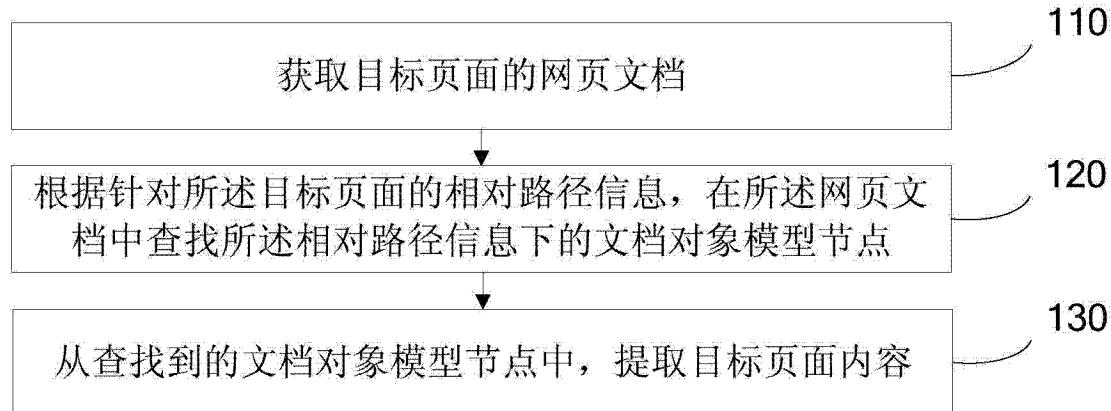


图 1

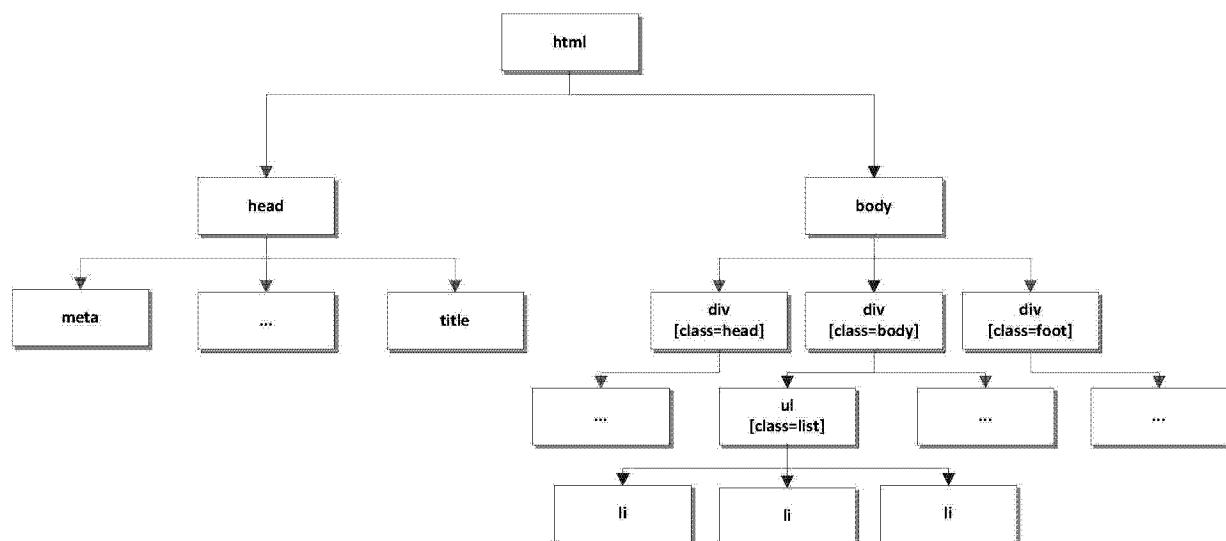


图 1A

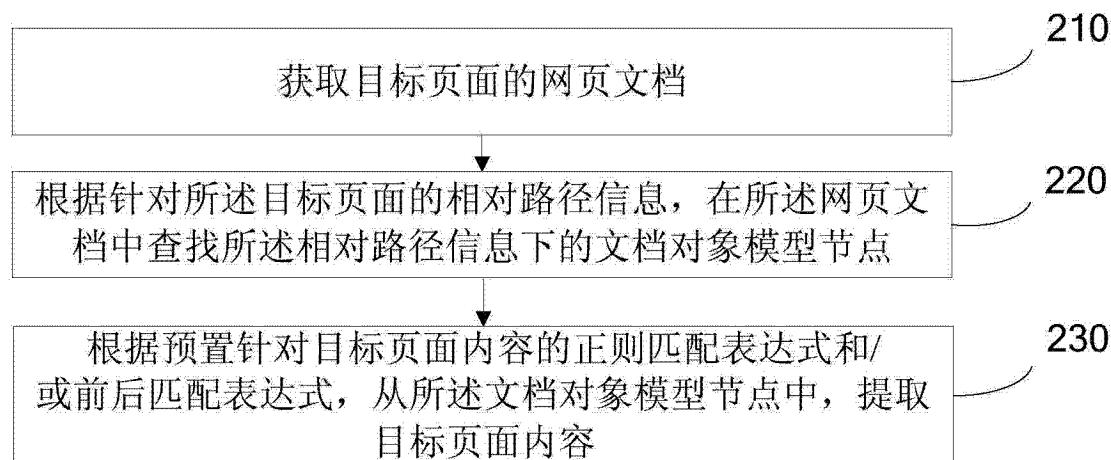


图 2

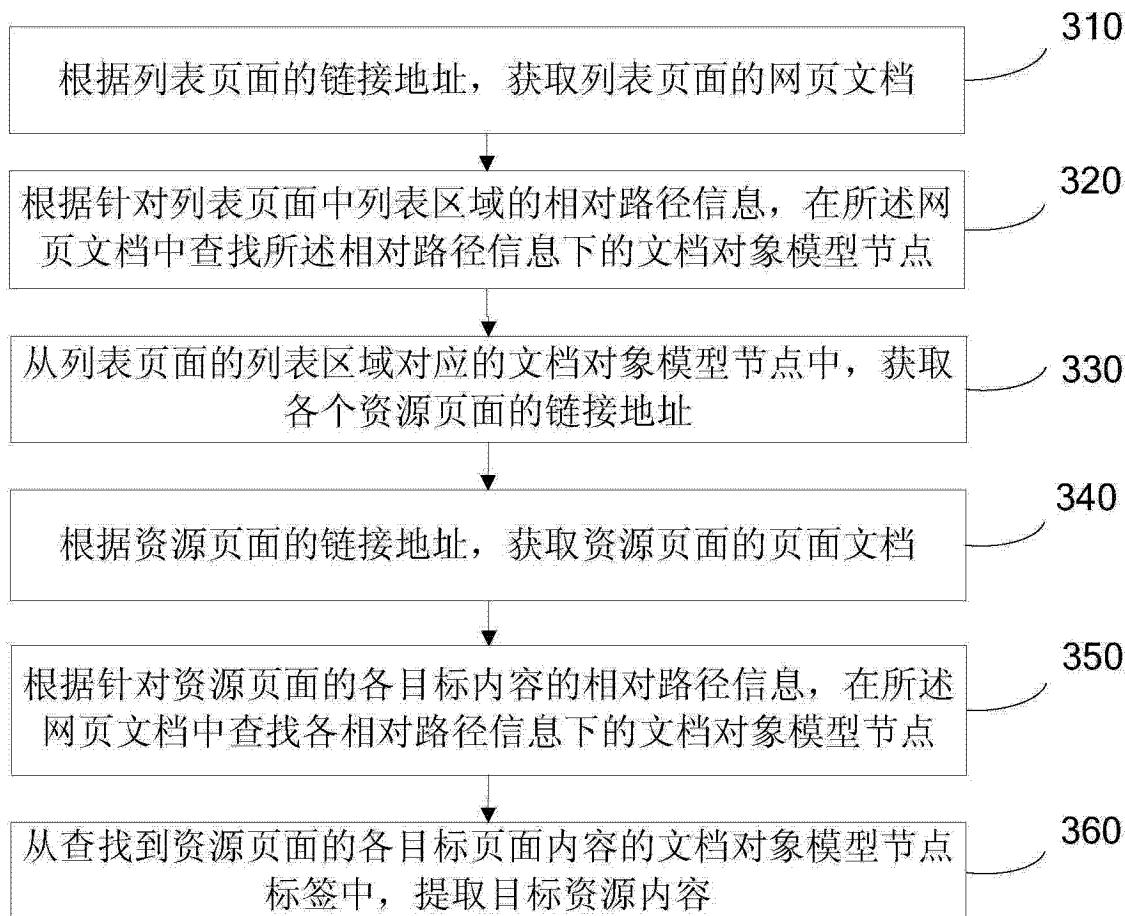


图 3



图 3A

标题	<code>div[class=article]/h1</code>
	④ 属性 ③ XPATH ② 正则 ⑤ 前后匹配
作者	<code>div[class=article] p</code>
	④ 属性 ③ XPATH ② 正则 ⑤ 前后匹配
来源	<code>div[class=article] p</code>
	④ 属性 ③ XPATH ② 正则 ⑤ 前后匹配
简介	<code>div[class=article] p</code>
	④ 属性 ③ XPATH ② 正则 ⑤ 前后匹配
标签	<code>div[class=article]</code>
	④ 属性 ③ XPATH ② 正则 ⑤ 前后匹配
* 内容	<code>div[class=articleText]</code>
	④ 属性 ③ XPATH ② 正则 ⑤ 前后匹配
字符转换	<code>统一编码，转：UTF-8/GB2312/GB18030</code>

图 3B

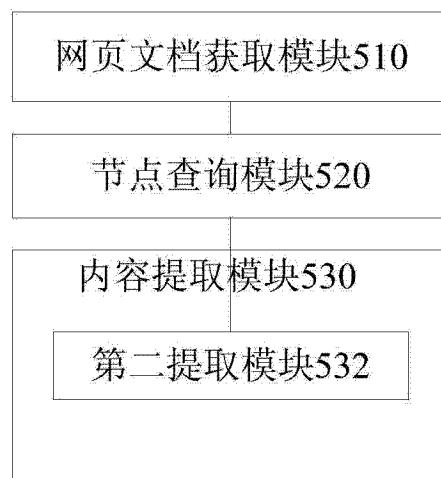


图 5

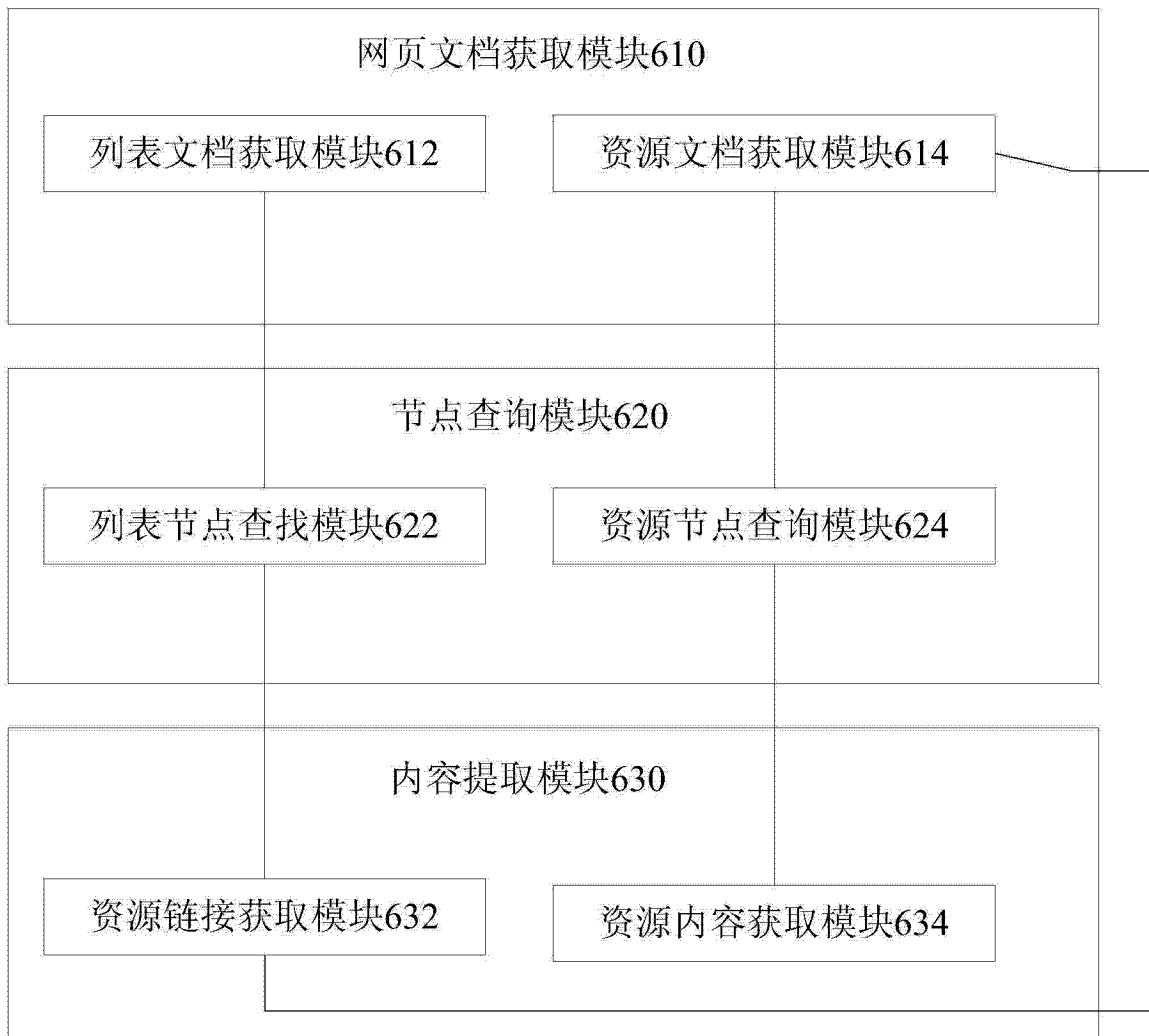


图 6