(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 103617228 A
(43)申请公布日 2014.03.05

(21)申请号 201310607851.8

(22)申请日 2013.11.25

(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 王智广

(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 赵娟

(51)Int.Cl.
G06F 17/30(2006.01)

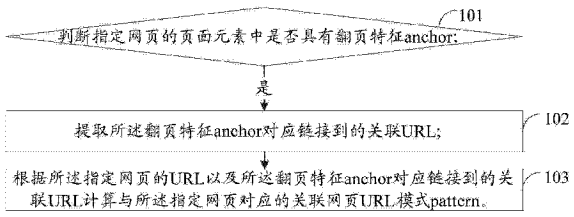
权利要求书2页 说明书12页 附图2页

(54)发明名称

一种计算关联网页URL模式pattern的方法和装置

(57)摘要

本发明公开了一种计算关联网页URL模式pattern的方法和装置,所述方法包括:判断指定网页的页面元素中是否具有翻页特征anchor;若是,则提取所述翻页特征anchor对应链接到的关联URL;根据所述指定网页的URL以及所述翻页特征anchor对应链接到的关联URL计算与所述指定网页对应的关联网页URL模式pattern。本发明采用翻页特征anchor识别关联网页,识别准确率高,基于指定网页的URL中和关联URL计算出关联网页URL模式pattern,计算效率高。



CN 103617228 A

1. 一种计算关联网页 URL 模式 pattern 的方法，包括：

判断指定网页的页面元素中是否具有翻页特征 anchor；若是，则提取所述翻页特征 anchor 对应链接到的关联 URL；

根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern。

2. 如权利要求 1 所述的方法，其特征在于，所述判断指定网页的页面元素中是否具有翻页特征 anchor 的步骤包括：

采用翻页特征 anchor 在当前网页的 DOM 树节点中进行匹配；

当匹配成功时，则判断当前网页具有翻页特征 anchor。

3. 如权利要求 1 所述的方法，其特征在于，所述翻页特征 anchor 对应链接到一个或多个关联 URL。

4. 如权利要求 1 或 2 或 3 所述的方法，其特征在于，所述根据所述指定网页的 URL 以及所述关联页 URL 计算所述关联网页 URL 模式 pattern 的步骤进一步包括：

使用通配字符替换指定网页的 URL 中的数字块，获得第一特征 URL 前缀；其中，所述数字块为被间隔标识分割出的单个数字或多个数字；

使用通配字符替换所述关联 URL 中的数字块，获得第二特征 URL 前缀；

当所述第一特征 URL 前缀与所述第二特征 URL 前缀相同时，将所述第一特征 URL 前缀或第二特征 URL 前缀作为关联网页 URL 模式 pattern。

5. 如权利要求 4 所述的方法，其特征在于，所述使用通配字符替换指定网页的 URL 中的数字块，获得第一特征 URL 前缀的步骤为：

采用相同的通配字符替换指定网页的 URL 中不同位置的数字块，获得第一特征 URL 前缀；

所述使用通配字符替换所述关联 URL 中的数字块，获得第二特征 URL 前缀的步骤为：

采用相同的通配字符替换所述关联 URL 中不同位置的数字块，获得第二特征 URL 前缀。

6. 如权利要求 5 所述的方法，其特征在于，所述使用通配字符替换指定网页的 URL 中的数字块，获得第一特征 URL 前缀的步骤为：

分别采用不同的通配字符，替换指定网页的 URL 中不同位置的数字块，获得第一特征 URL 前缀；

所述使用通配字符替换所述关联 URL 中的数字块，获得第二特征 URL 前缀的步骤为：

分别采用与第一特征 URL 相同的通配字符替换所述关联 URL 在相同位置的数字块，获得第二特征 URL 前缀。

7. 如权利要求 1 或 2 或 3 或 5 或 6 所述的方法，其特征在于，还包括：

通过对关联网页 URL 模式 pattern 中的共性部分进行结构分析，提取关联网页 URL 模式 pattern 中的翻页块，将所述翻页块替换为首页标识获得首页关联网页的 URL；其中，所述翻页块为多个关联网页 URL 模式 pattern 中位置相同但数字不同的数字块。

8. 如权利要求 7 所述的方法，其特征在于，所述首页标识包括 0、1 和 / 或当前关联网页中的最大数值。

9. 一种计算关联网页 URL 模式 pattern 的装置，包括：

翻页特征 anchor 判断模块，适于判断指定网页的页面元素中是否具有翻页特征

anchor ;若是,则调用关联 URL 提取模块 ;

URL 提取模块,适于提取所述翻页特征 anchor 对应链接到的关联 URL ;

关联网页 URL 模式 pattern 计算模块,适于根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern。

10. 如权利要求 9 所述的装置,其特征在于,所述翻页特征 anchor 判断模块还适于 :

采用翻页特征 anchor 在当前网页的 DOM 树节点中进行匹配 ;

当匹配成功时,则判断当前网页具有翻页特征 anchor。

一种计算关联网页 URL 模式 pattern 的方法和装置

技术领域

[0001] 本发明涉及数据处理技术领域，具体涉及一种计算关联网页 URL 模式 pattern 的方法、一种计算关联网页 URL 模式 pattern 的装置。

背景技术

[0002] 随着因特网的发展，愈来愈多的信息是通过网页方式呈现在因特网上供用户查询，同样的通过搜寻引擎在因特网中查询数据也成为最常使用的数据搜寻方法。

[0003] 搜索引擎收录网页时需要针对不同种类的网页采取不同的调度策略，网页种类的识别是一项基础工作，其中翻页(Page turning)网页的识别是一项比较关键的工作。所谓翻页网页，即查看分页文件的上一个页面、下一个页面或任意存在的非当前页面。翻页网页可以将实体书或者移动 Web 窗体中的内容进行改变，以观看不同内容。在互联网上运用时该机制还呈现可用于浏览到其他页的用户界面元素。

[0004] 现有的翻页网页的识别方法是根据网页的 URL(Uniform Resource Locator，统一资源定位符)所包含的关键词来识别是否是索引页。例如，当 URL 包含有 page、pn、p 等关键词以及关键词后面有数字时，判断该 URL 对应的网页为翻页网页。

[0005] 但是，这种识别方法召回率低，并且很多网站的翻页是不具有这些关键词的，比如“http://cq.ABC.com/lvshi/o12/”、“http://bbs.BCA.com/t661_10”、“http://china.BCD.com/product/20110617/2647”，但是这些网页依然是翻页，使得这些识别方法容易造成误操作，实用性低。

发明内容

[0006] 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一种计算关联网页 URL 模式 pattern 的方法和相应的一种计算关联网页 URL 模式 pattern 的装置。

[0007] 依据本发明的一个方面，提供了一种计算关联网页 URL 模式 pattern 的方法，包括：

[0008] 判断指定网页的页面元素中是否具有翻页特征 anchor；若是，则提取所述翻页特征 anchor 对应链接到的关联 URL；

[0009] 根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern。

[0010] 可选地，所述判断指定网页的页面元素中是否具有翻页特征 anchor 的步骤包括：

[0011] 采用翻页特征 anchor 在当前网页的 DOM 树节点中进行匹配；

[0012] 当匹配成功时，则判断当前网页具有翻页特征 anchor。

[0013] 可选地，所述翻页特征 anchor 对应链接到一个或多个关联 URL。

[0014] 可选地，所述根据所述指定网页的URL以及所述关联页URL计算所述关联网页URL模式 pattern 的步骤进一步包括：

[0015] 使用通配字符替换指定网页的 URL 中的数字块，获得第一特征 URL 前缀；其中，所述数字块为被间隔标识分割出的单个数字或多个数字；

[0016] 使用通配字符替换所述关联 URL 中的数字块，获得第二特征 URL 前缀；

[0017] 当所述第一特征 URL 前缀与所述第二特征 URL 前缀相同时，将所述第一特征 URL 前缀或第二特征 URL 前缀作为关联网页 URL 模式 pattern。

[0018] 可选地，所述使用通配字符替换指定网页的 URL 中的数字块，获得第一特征 URL 前缀的步骤为：

[0019] 采用相同的通配字符替换指定网页的 URL 中不同位置的数字块，获得第一特征 URL 前缀；

[0020] 所述使用通配字符替换所述关联 URL 中的数字块，获得第二特征 URL 前缀的步骤为：

[0021] 采用相同的通配字符替换所述关联 URL 中不同位置的数字块，获得第二特征 URL 前缀。

[0022] 可选地，所述使用通配字符替换指定网页的 URL 中的数字块，获得第一特征 URL 前缀的步骤为：

[0023] 分别采用不同的通配字符，替换指定网页的 URL 中不同位置的数字块，获得第一特征 URL 前缀；

[0024] 所述使用通配字符替换所述关联 URL 中的数字块，获得第二特征 URL 前缀的步骤为：

[0025] 分别采用与第一特征 URL 相同的通配字符替换所述关联 URL 在相同位置的数字块，获得第二特征 URL 前缀。

[0026] 可选地，还包括：

[0027] 通过对关联网页 URL 模式 pattern 中的共性部分进行结构分析，提取关联网页 URL 模式 pattern 中的翻页块，将所述翻页块替换为首页标识获得首页关联网页的 URL；其中，所述翻页块为多个关联网页 URL 模式 pattern 中位置相同但数字不同的数字块。

[0028] 可选地，所述首页标识包括 0、1 和 / 或当前关联网页中的最大数值。

[0029] 根据本发明的另一方面，提供了一种计算关联网页 URL 模式 pattern 的装置，包括：

[0030] 翻页特征 anchor 判断模块，适于判断指定网页的页面元素中是否具有翻页特征 anchor；若是，则调用关联 URL 提取模块；

[0031] URL 提取模块，适于提取所述翻页特征 anchor 对应链接到的关联 URL；

[0032] 关联网页 URL 模式 pattern 计算模块，适于根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern。

[0033] 可选地，所述翻页特征 anchor 判断模块还适于：

[0034] 采用翻页特征 anchor 在当前网页的 DOM 树节点中进行匹配；

[0035] 当匹配成功时，则判断当前网页具有翻页特征 anchor。

[0036] 可选地，所述翻页特征 anchor 对应链接到一个或多个关联 URL。

[0037] 可选地，所述关联网页 URL 模式 pattern 计算模块包括：

[0038] 第一特征URL前缀获得子模块，适于使用通配字符替换指定网页的URL中的数字块，获得第一特征URL前缀；其中，所述数字块为被间隔标识分割出的单个数字或多个数字；

[0039] 第二特征URL前缀获得子模块，适于使用通配字符替换所述关联URL中的数字块，获得第二特征URL前缀；

[0040] 关联网页URL模式pattern获得模块，适于在所述第一特征URL前缀与所述第二特征URL前缀相同时，将所述第一特征URL前缀或第二特征URL前缀作为关联网页URL模式pattern。

[0041] 可选地，所述第一特征URL前缀获得子模块还适于：

[0042] 采用相同的通配字符替换指定网页的URL中不同位置的数字块，获得第一特征URL前缀；

[0043] 所述第二特征URL前缀获得子模块还适于：

[0044] 采用相同的通配字符替换所述关联URL中不同位置的数字块，获得第二特征URL前缀。

[0045] 可选地，所述第一特征URL前缀获得子模块还适于：

[0046] 分别采用不同的通配字符，替换指定网页的URL中不同位置的数字块，获得第一特征URL前缀；

[0047] 第二特征URL前缀获得子模块还适于：

[0048] 分别采用与第一特征URL相同的通配字符替换所述关联URL在相同位置的数字块，获得第二特征URL前缀。

[0049] 可选地，其特征在于，还包括：

[0050] 首页关联网页URL获得模块，适于通过对关联网页URL模式pattern中的共性部分进行结构分析，提取关联网页URL模式pattern中的翻页块，将所述翻页块替换为首页标识获得首页关联网页的URL；其中，所述翻页块为多个关联网页URL模式pattern中位置相同但数字不同的数字块。

[0051] 可选地，所述首页标识包括0、1和/或当前关联网页中的最大数值。

[0052] 本发明采用翻页特征anchor识别关联网页，识别准确率高，基于指定网页的URL中和关联URL计算出关联网页URL模式pattern，计算效率高。

[0053] 本发明使用通配字符替换数字块获得第一特征URL前缀和获得第二特征URL前缀，当所述第一特征URL前缀与所述第二特征URL前缀相同时，将所述第一特征URL前缀或第二特征URL前缀作为关联网页URL模式，本发明采用URL的共性部分进行匹配，进一步提高了关联网页的识别准确率，使得召回率大幅提高，在实际应用中可以识别90%以上的关联网页。

[0054] 本发明将关联网页URL模式pattern的翻页块替换为首页标识获得首页关联网页的URL，同理，也可以将翻页块替换为其他挂链网页标识获得其他关联网页的URL，从而增加了关联网页的覆盖率，使得能够获取更加全面的关联网页，进而实现了细颗粒度的操作。

[0055] 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

[0056] 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：

[0057] 图 1 示出了根据本发明一个实施例的一种计算关联网页 URL 模式 pattern 的方法实施例 1 的步骤流程图；

[0058] 图 2 示出了根据本发明一个实施例的一种网页结构示例图；

[0059] 图 3 示出了示出了本发明一个实施例的一种翻页块的示例图；

[0060] 图 4 示出了根据本发明一个实施例的一种计算关联网页 URL 模式 pattern 的方法实施例 2 的步骤流程图；

[0061] 图 5 示出了根据本发明一个实施例的一种计算关联网页 URL 模式 pattern 的装置实施例 1 的结构框图；以及，

[0062] 图 6 示出了根据本发明一个实施例的一种计算关联网页 URL 模式 pattern 的装置实施例 2 的结构框图。

具体实施方式

[0063] 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。

[0064] 参照图 1，示出了本发明一个实施例的一种计算关联网页 URL 模式 pattern 的方法实施例 1 的步骤流程图，具体可以包括如下步骤：

[0065] 步骤 101，判断指定网页的页面元素中是否具有翻页特征 anchor；若是，则执行步骤 102；

[0066] 网页按照功能可以划分为多个区域，以某一个论坛（Bulletin Board System，BBS）的页面为例，如图 2 所示，该页面可以划分为导航块（1）、垃圾块（2、4）、翻页块（3）、标题块（5）、作者信息块（6）、发表日期块（7）、正文块（8）。其中，导航块可以位于网页页眉顶部，或者 banner（网页的横幅广告）下部，用于指向网页的信息栏目。垃圾块可以为与网页主题相关度很低的页面元素所在的区域，例如“发帖”、“回复”等功能按钮。翻页块可以为指示翻页的区域。标题块可以为网页主题的标题（例如图 2 所示的“安全浏览器聚集黑色星期四”）所在的区域。作者信息块为记载该网页主题作者信息的区域。正文块为记载该网页主题正文的区域。

[0067] 参照图 3，示出了示出了本发明一个实施例的一种翻页块的示例图。

[0068] 如图 3 所示，翻页块主要可以由翻页特征 anchor 组成，翻页特征 anchor 即翻页特征字符串，其可以为用于标识翻页的页面元素。

[0069] 在具体实现中，翻页特征 anchor 可以包括以下的一种或多种：

[0070] [<<]、[>>]、[‹‹]、[››]、[«]、[»]、[>]、[<]、[下一页]、[上一页]、[上一]、[下一]、[next]、[末页]、[尾页]、[前页]、[后页]、[<上一页]、[<上一]、[下一>]、

[下一页 >]、[1...]。

[0071] 当然，上述翻页特征 anchor 只是作为示例，在实施本发明实施例时，可以根据实际情况设置其他翻页特征 anchor，本发明实施例对此不加以限制。

[0072] 在本发明的一种优选实施例中，所述步骤 101 具体可以包括如下子步骤：

[0073] 子步骤 S11，采用翻页特征 anchor 在当前网页的 DOM 树节点中进行匹配；

[0074] 子步骤 S12，当匹配成功时，则判断当前网页具有翻页特征 anchor。

[0075] DOM（文件对象模型，DocumentObjectModel）是处理可扩展置标语言的标准编程接口。DOM 可以以一种独立于平台和语言的方式访问和修改一个文档的内容和结构，是表示和处理一个 HTML 或 XML 文档的常用方法。

[0076] DOM 实际上是以面向对象方式描述的文档模型。DOM 定义了表示和修改文档所需的对象、这些对象的行为和属性以及这些对象之间的关系。可以把 DOM 认为是页面上数据和结构的一个树形表示，不过页面当然可能并不是以这种树的方式具体实现。

[0077] 通过 JavaScript 可以重构整个 HTML 文档，可以添加、移除、改变或重排页面上的项目。

[0078] 要改变页面的某个东西，JavaScript 就需要获得对 HTML 文档中所有元素进行访问的入口。这个入口，连同对 HTML 元素进行添加、移动、改变或移除的方法和属性，都是通过文档对象模型来获得的（DOM）。

[0079] 可以将 HTML 文档视作树结构，而这种结构被称为节点树（HTML DOM）。通过 HTMLDOM，树中的所有节点均可通过 JavaScript 进行访问。所有 HTML 元素（节点）均可被修改，也可以创建或删除节点。

[0080] 节点树中的节点彼此拥有层级关系。可以采用父（parent）、子（child）和同胞（sibling）等术语用于描述这些关系。其中，父节点拥有子节点。同级的子节点被称为同胞（兄弟或姐妹）。在节点树中，顶端节点被称为根（root）。每个节点都有父节点、除了根（它没有父节点）。一个节点可拥有任意数量的子，同胞是拥有相同父节点的节点。

[0081] 具体可以通过若干种方法在节点树来查找希望操作的网页元素：

[0082] 例如，可以通过使用 getElementById() 和 getElementsByTagName() 方法进行查找。

[0083] 又例如，可以通过使用一个元素节点的 parentNode、firstChild 以及 lastChild 属性。

[0084] 其中，getElementById() 和 getElementsByTagName() 这两种方法，可查找整个 HTML 文档中的任何 HTML 元素。而这两种方法会忽略文档的结构。假如查找文档中所有的 <p> 元素，getElementsByTagName() 会把它们全部找到，不管 <p> 元素处于文档中的哪个层次。同时，getElementById() 方法也会返回正确的元素，不论它被隐藏在文档结构中的什么位置。这两种方法会提供任何所需要的 HTML 元素，不论它们在文档中所处的位置。

[0085] 此外，getElementById() 可通过指定的 ID 来返回网页元素。

[0086] 在具体实现中，可以通过识别该网页的 HTML 文本 DOM 树中超链接 <a>（anchor，锚点）标识是否包括 [<<]、[>>]、[«]、[»]、[«]、[»]、[>]、[<]、[下一页]、[上一页]、[上一]、[下一]、[next]、[末页]、[尾页]、[前页]、[后页]、[< 上一页]、[< 上一]、[下一 >]、[下一页 >]、[1...] 中的一种或多种，若是，则判断当前网页具有翻页特征 anchor。

[0087] 其中，<a> 可以用于把当前位置的文本或图片连接到其他的页面、文本或图像等。

[0088] ＜ a ＞标识的基本语法结构可以如下：

[0089] ＜ a

[0090] class=type

[0091] id ＝ value

[0092] href ＝ reference

[0093] name ＝ value

[0094] rel=same|next|parent|previous

[0095] rev ＝ value

[0096] target ＝ window

[0097] style ＝ value

[0098] title=title

[0099] onclick ＝ function

[0100] onmouseout ＝ function

[0101] onMouseOver=function ＞显示文字或者图片的代码＜ /a ＞

[0102] 例如以下一种 HTML 文本中 <a> 标识的内容为：

[0103] <divid=" pgt" class=" bm bw0 pgs cl" >

[0104] <spanid=" fd_page_top" >

[0105] <divclass=" pg" >

[0106] <a

[0107] href=" forum-99-1.html" class=" prev" >

[0108] <a

[0109] href=" forum-99-1.html" >12<>

[0110] <a

[0111] href=" forum-99-3.html" >3

[0112] <a

[0113] href=" forum-99-4.html" >4

[0114] <a

[0115] href=" forum-99-5.html" >5

[0116] <a

[0117] href=" forum-99-6.html" >6

[0118] <a

[0119] href=" forum-99-7.html" >7

[0120] <a

[0121] href=" forum-99-8.html" >8

[0122] <a

[0123] href=" forum-99-9.html" >9

[0124] <a

[0125] href=" forum-99-10.html" >10

[0126] <a

[0127] href=" forum-99-1000.html" class=" last" >...2107

[0128] <label>

[0129] <input type=" text" name=" custompage" class=" px" size=" 2" title=" 输入页码,按回车快速跳转" value=" 2" onkeydown=" if(event.keyCode==13){window.location= ' forum.php?mod=forumdisplay&fid=99&page= ' +this.value;doane(event);}" />

[0130] <spantitle=" 共 1000 页" >/1000 页

[0131] </label>

[0132] <a

[0133] href=" forum-99-3.html" class=" nxt" > 下一页

[0134]

[0135]

[0136] 通过 HTML 文本中 <a> 标识的匹配,可以判断该网页具有一个或多个翻页特征 anchor。

[0137] 步骤 102,提取所述翻页特征 anchor 对应链接到的关联 URL ;

[0138] 在实现应用中,所述翻页特征 anchor 可以对应链接到一个或多个关联 URL。

[0139] 具体地,在识别出该一个或多个翻页特征 anchor 之后,提取该一个或多个翻页特征 anchor 链接的一个或多个关联 URL,该一个或多个关联 URL 指向其他的与当前网页关联的翻页网页。

[0140] 步骤 103,根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern。

[0141] 关联网页 URL 模式 Pattern,可以为长相或者功能类似的 URL/ 网页聚在一起形成的集合。

[0142] 在本发明的一种优选实施例中,所述步骤 103 具体可以包括如下子步骤 :

[0143] 子步骤 S21,使用通配字符替换指定网页的 URL 中的数字块,获得第一特征 URL 前缀 ;其中,所述数字块为被间隔标识分割出的单个数字或多个数字 ;

[0144] 子步骤 S31,使用通配字符替换所述关联 URL 中的数字块,获得第二特征 URL 前缀 ;

[0145] 需要说明的是,通配字符可以为任意字符,本发明实施例对此不加以限制。间隔标识可以为 URL 中用于间隔的符号,例如“/”、“.”、“-”、“?”、“ :”等等。数字块需要为间隔标识中连续的数字,例如“123ABC”不为数字块。

[0146] 在本发明实施例的一种优选示例中,所述子步骤 S21 进一步可以包括如下子步骤 :

[0147] 子步骤 S211,采用相同的通配字符替换指定网页的 URL 中不同位置的数字块,获得第一特征 URL 前缀 ;

[0148] 与子步骤 S211 相对应地,所述子步骤 S31 进一步可以包括如下子步骤 :

[0149] 子步骤 S311,采用相同的通配字符替换所述关联 URL 中不同位置的数字块,获得第二特征 URL 前缀。

[0150] 在具体实现中，指定网页的URL和关联URL可以具有一个或多个数字块，为减少替换的操作步骤和系统的资源占用，可以用相同的通配字符替换数字块。

[0151] 例如，指定网页的URL为http://bbs.XXX.com/forum-99-2.html，关联URL为http://bbs.XXX.com/forum-99-3.html，其中“99”、“2”被识别出为数字块，以“(\d+)”作为通配字符的一种示例，则第一特征URL前缀可以为http://bbs.XXX.com/forum-(\d+)-(\d+).html，第二特征URL前缀可以为http://bbs.XXX.com/forum-(\d+)-(\d+).html。

[0152] 在本发明的一种实施例中，所述子步骤S21进一步可以包括如下子步骤：

[0153] 子步骤S212，分别采用不同的替换字符，替换指定网页的URL中不同位置的数字块，获得第一特征URL前缀；

[0154] 与子步骤S212相对应地，所述步骤103具体可以包括如下子步骤：

[0155] 子步骤S312，分别采用与第一特征URL相同的通配字符替换所述关联URL在相同位置的数字块，获得第二特征URL前缀。

[0156] 在具体实现中，指定网页的URL和关联URL可以具有一个或多个数字块，为提高后续第一特征URL前缀与第二特征URL是否相同的判断以及对数字块的标识的效率，可以采用不同的通配字符替换数字块。

[0157] 例如，指定网页的URL为http://bbs.XXX.com/forum-99-2.html，关联URL为http://bbs.XXX.com/forum-99-3.html，其中“99”、“2”被识别出为数字块，以“(\d+)”、“(\e+)”作为通配字符的一种示例，则第一特征URL前缀可以为http://bbs.XXX.com/forum-(\d+)-(\e+).html，第二特征URL前缀可以为http://bbs.XXX.com/forum-(\d+)-(\e+).html。

[0158] 子步骤S41，当所述第一特征URL前缀与所述第二特征URL前缀相同时，将所述第一特征URL前缀或第二特征URL前缀作为关联网页URL模式pattern。

[0159] 在实际应用中，当第一特征URL前缀与第二特征URL前缀相同时，可以判定指定网页的和关联URL对应的网页为关联的翻页网页。

[0160] 因为第一特征URL前缀和第二特征URL相同，则以第一特征URL前缀或第二特征URL前缀作为关联网页URL模式Pattern均可。

[0161] 本发明采用翻页特征anchor识别关联网页，识别准确率高，基于指定网页的URL中和关联URL计算出关联网页URL模式pattern，计算效率高。

[0162] 本发明使用通配字符替换数字块获得第一特征URL前缀和获得第二特征URL前缀，当所述第一特征URL前缀与所述第二特征URL前缀相同时，将所述第一特征URL前缀或第二特征URL前缀作为关联网页URL模式，本发明采用采用URL的共性部分进行匹配，进一步提高了关联网页的识别准确率，使得召回率大幅提高，在实际应用中可以识别90%以上的关联网页。

[0163] 参照图4，示出了本发明一个实施例的一种计算关联网页URL模式pattern的方法实施例2的步骤流程图，具体可以包括如下步骤：

[0164] 步骤401，判断指定网页的页面元素中是否具有翻页特征anchor；若是，则执行步骤402；

[0165] 步骤402，提取所述翻页特征anchor对应链接到的关联URL；

[0166] 步骤 403，根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern ；

[0167] 步骤 404，通过对关联网页 URL 模式 pattern 中的共性部分进行结构分析，提取关联网页 URL 模式 pattern 中的翻页块，将所述翻页块替换为首页标识获得首页关联网页的 URL ；

[0168] 其中，所述翻页块为多个关联网页 URL 模式 pattern 中位置相同但数字不同的数字块。

[0169] 在实际应用中，URL 可以包括以下的一种或多种结构 ：

[0170] 1、protocol (协议)：指定使用的传输协议，最常用的是 HTTP 协议，它也是目前 WWW 中应用最广的协议。具体地，传输协议包括 file 协议(资源是本地计算机上的文件，格式为 file:///)、ftp 协议(通过 FTP 访问资源，格式为 FTP://)、gopher (通过 Gopher 协议访问资源)、http 协议(通过 HTTP 访问资源，格式为 HTTP://)、https 协议(通过安全的 HTTPS 访问资源，格式为 HTTPS://) 等等。

[0171] 2、hostname (主机名)：指存放资源的服务器的域名系统(DNS) 主机名或 IP 地址。有时，在主机名前也可以包含连接到服务器所需的用户名和密码(格式为 username:password)。

[0172] 3、port (端口号)：省略时使用方案的默认端口，各种传输协议都有默认的端口号，如 http 的默认端口为 80。如果输入时省略，则使用默认端口号。有时候出于安全或其他考虑，可以在服务器上对端口进行重定义，即采用非标准端口号，此时，URL 中就不能省略端口号这一项。

[0173] 4、path (路径)：由零或多个“/”符号隔开的字符串，一般用来表示主机上的一个目录或文件地址。

[0174] 5、parameters (参数)：可以用于指定特殊参数的可选项。

[0175] 6、query(查询)：可以用于给动态网页(如使用 CGI、ISAPI、PHP/JSP/ASP/ASP.NET 等技术制作的网页)传递参数，可有多个参数，用“&”符号隔开，每个参数的名和值用“=”符号隔开。

[0176] 7、fragment (信息片断)：可以用于指定网络资源中的片断。例如一个网页中有多个名词解释，可使用 fragment 直接定位到某一名词解释。

[0177] 在具体实现中，通过对多个关联网页 URL 模式中的共性部分进行结构分析，提取关联网页 URL 模式中的翻页块，然后将所述翻页块替换为首页标识获得首页关联网页的 URL。

[0178] 例如，对于上述示例的关联网页 URL 模式 -http://bbs.XXX.com/forum-(\d+)-(\e+).html，在识别出 (\e+) 为翻页块，然后将翻页块替换为首页标识后，获得首页关联网页的 URL-http://bbs.XXX.com/forum-99-1.html。

[0179] 在本发明实施例的一种优选示例中，所述首页标识可以包括 0、1 和 / 或当前关联网页中的最大数值。

[0180] 在具体实现中，关联网页中的首页关联网页一般会记载有重要的内容，例如图 3 所示的正文块，因此首页关联网页的重要性比较高，因此获知首页关联网页具有比较重要的意义。而不同的网站会采用不同的翻页结构，造成了首页关联网页的不同。例如，某些网

站会采用第0页作为首页关联网页，某些网站会采用第1页作为首页关联网页，某些网站会采用最大页(例如图3所示的2100)作为首页关联网页，等等。

[0181] 当然，上述首页关联网页只是作为示例，在实施本发明实施例时，可以根据实际情况将数字快替换为任一关联网页的标识获取对应的关联网页，本发明实施例对此不一一加以详述。

[0182] 本发明将关联网页URL模式pattern的翻页块替换为首页标识获得首页关联网页的URL，同理，也可以将翻页块替换为其他挂链网页标识获得其他关联网页的URL，从而增加了关联网页的覆盖率，使得能够获取更加全面的关联网页，进而实现了细颗粒度的操作。

[0183] 对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[0184] 参照图5，示出了本发明一个实施例的一种计算关联网页URL模式pattern的装置实施例1的结构框图，具体可以包括如下模块：

[0185] 翻页特征anchor判断模块501，适于判断指定网页的页面元素中是否具有翻页特征anchor；若是，则调用关联URL提取模块502；

[0186] URL提取模块502，适于提取所述翻页特征anchor对应链接到的关联URL；

[0187] 关联网页URL模式pattern计算模块503，适于根据所述指定网页的URL以及所述翻页特征anchor对应链接到的关联URL计算与所述指定网页对应的关联网页URL模式pattern。

[0188] 在本发明的一种优选实施例中，所述翻页特征anchor判断模块501还可以适于：

[0189] 采用翻页特征anchor在当前网页的DOM树节点中进行匹配；

[0190] 当匹配成功时，则判断当前网页具有翻页特征anchor。

[0191] 在本发明的一种优选实施例中，所述翻页特征anchor可以对应链接到一个或多个关联URL。

[0192] 在本发明的一种优选实施例中，所述关联网页URL模式pattern计算模块503具体可以包括如下子模块：

[0193] 第一特征URL前缀获得子模块，适于使用通配字符替换指定网页的URL中的数字块，获得第一特征URL前缀；其中，所述数字块为被间隔标识分割出的单个数字或多个数字；

[0194] 第二特征URL前缀获得子模块，适于使用通配字符替换所述关联URL中的数字块，获得第二特征URL前缀；

[0195] 关联网页URL模式pattern获得模块，适于在所述第一特征URL前缀与所述第二特征URL前缀相同时，将所述第一特征URL前缀或第二特征URL前缀作为关联网页URL模式pattern。

[0196] 在本发明的一种优选实施例中，所述第一特征URL前缀获得子模块还可以适于：

[0197] 采用相同的通配字符替换指定网页的URL中不同位置的数字块，获得第一特征URL前缀；

[0198] 所述第二特征URL前缀获得子模块还可以适于：

[0199] 采用相同的通配字符替换所述关联 URL 中不同位置的数字块，获得第二特征 URL 前缀。

[0200] 在本发明的一种优选实施例中，所述第一特征 URL 前缀获得子模块还可以适于：

[0201] 分别采用不同的通配字符，替换指定网页的 URL 中不同位置的数字块，获得第一特征 URL 前缀；

[0202] 第二特征 URL 前缀获得子模块还可以适于：

[0203] 分别采用与第一特征 URL 相同的通配字符替换所述关联 URL 在相同位置的数字块，获得第二特征 URL 前缀。

[0204] 对于图 5 的装置实施例而言，由于其与图 1 的方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0205] 参照图 6，示出了本发明一个实施例的计算一种关联网页 URL 模式 pattern 的装置施例 2 的结构框图，具体可以包括如下模块：

[0206] 翻页特征 anchor 判断模块 601，适于判断指定网页的页面元素中是否具有翻页特征 anchor；若是，则调用关联 URL 提取模块 602；

[0207] URL 提取模块 602，适于提取所述翻页特征 anchor 对应链接到的关联 URL；

[0208] 关联网页 URL 模式 pattern 计算模块 603，适于根据所述指定网页的 URL 以及所述翻页特征 anchor 对应链接到的关联 URL 计算与所述指定网页对应的关联网页 URL 模式 pattern；

[0209] 首页关联网页 URL 获得模块 604，适于通过对关联网页 URL 模式 pattern 中的共性部分进行结构分析，提取关联网页 URL 模式 pattern 中的翻页块，将所述翻页块替换为首页标识获得首页关联网页的 URL；其中，所述翻页块为多个关联网页 URL 模式 pattern 中位置相同但数字不同的数字块。

[0210] 在本发明实施例的一种优选示例中，所述首页标识可以包括 0、1 和 / 或当前关联网页中的最大数值。

[0211] 对于图 6 的装置实施例而言，由于其与图 4 的方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0212] 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。

[0213] 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。

[0214] 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。

[0215] 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。

[0216] 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。

[0217] 本发明的各个部件实施例可以以硬件实现，或者以在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计算关联网页 URL 模式 pattern 的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。

[0218] 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。

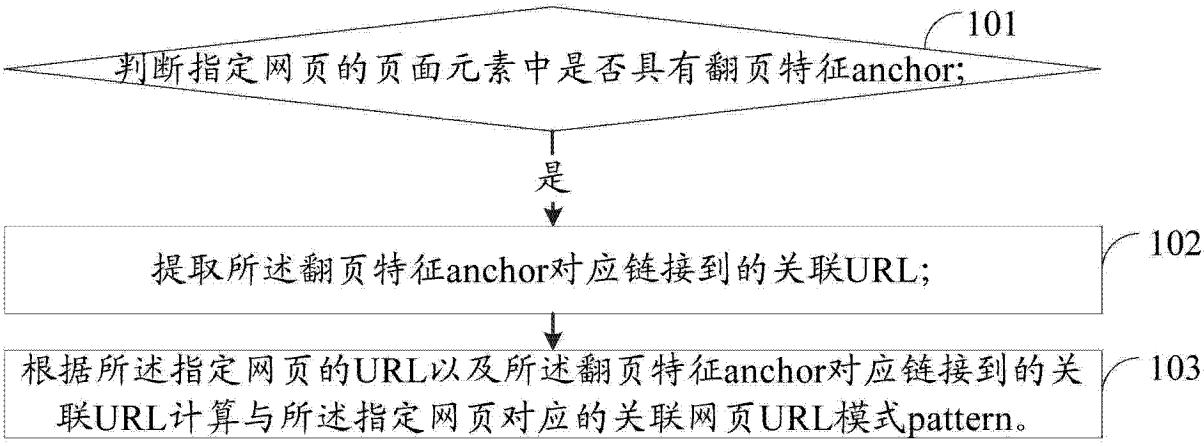

101
判断指定网页的页面元素中是否具有翻页特征anchor;
是
102
提取所述翻页特征anchor对应链接到的关联URL;
103
根据所述指定网页的URL以及所述翻页特征anchor对应链接到的关联URL计算与所述指定网页对应的关联网页URL模式pattern。

图 1

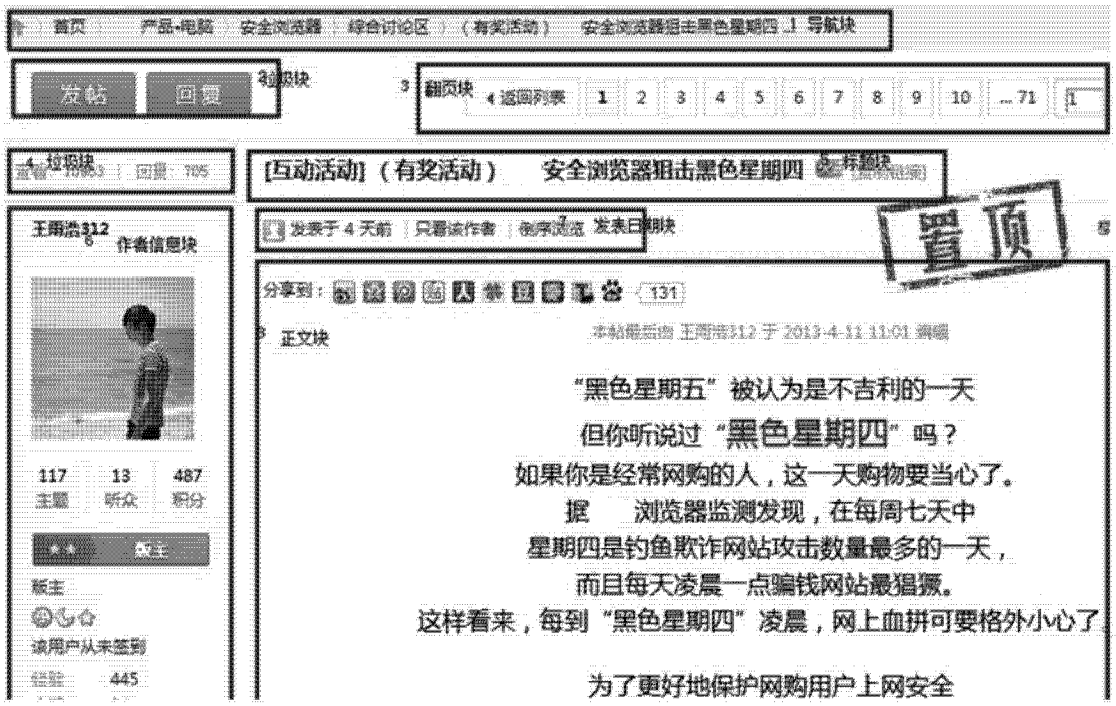

首页 产品-电脑 安全浏览器 综合讨论区 （有奖活动） 安全浏览器狙击黑色星期四 1 导航块
发帖 回复 2 功能块
3 翻页块 返回列表 1 2 3 4 5 6 7 8 9 10 ... 71
4 垃圾块
[互动活动]（有奖活动） 安全浏览器狙击黑色星期四 6 标题块
王雨浩312 6 作者信息块
发表于 4 天前 只看该作者 倒序浏览 7 发表日期块
置顶
分享到：
正文块
"黑色星期五"被认为是不吉利的一天
但你听说过"黑色星期四"吗？
如果你是经常网购的人，这一天购物要当心了。
据 浏览器监测发现，在每周七天中
星期四是钓鱼欺诈网站攻击数量最多的一天，
而且每天凌晨一点骗钱网站最猖獗。
这样看来，每到"黑色星期四"凌晨，网上血拼可要格外小心了
为了更好地保护网购用户上网安全
117 13 487
主题 听众 积分
板主
445

图 2

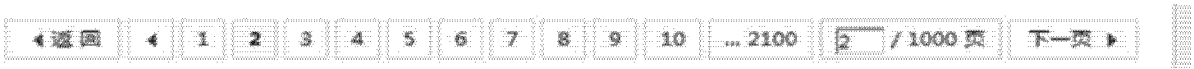

返回 1 2 3 4 5 6 7 8 9 10 ... 2100 / 1000 页 下一页

图 3

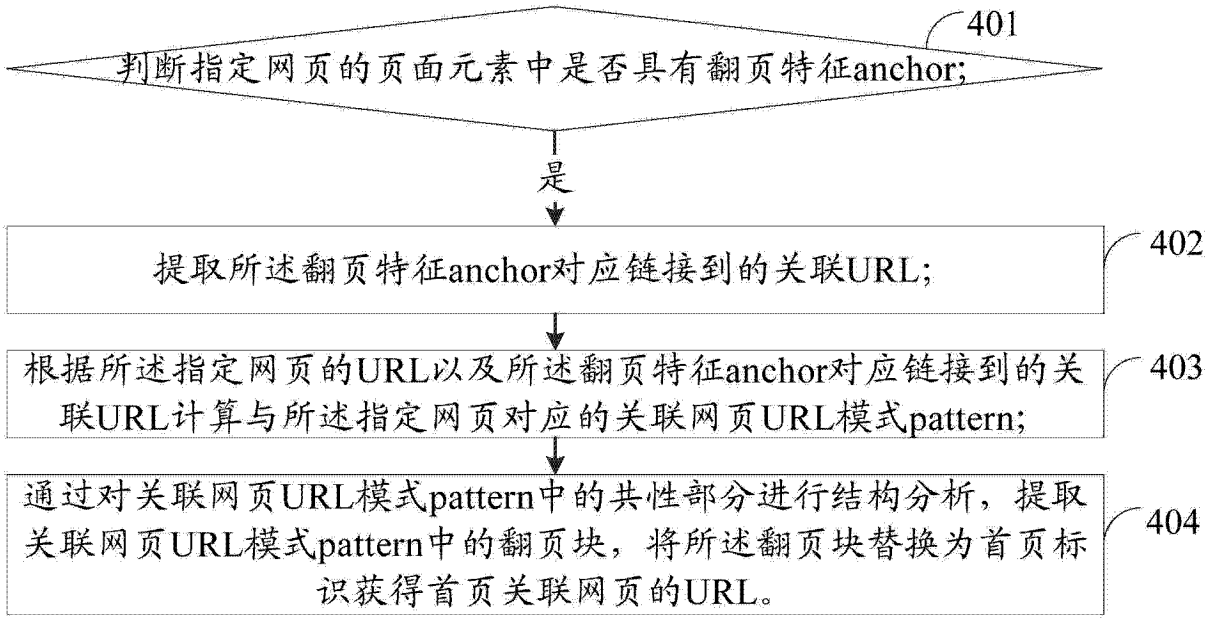

401
判断指定网页的页面元素中是否具有翻页特征anchor;
是
提取所述翻页特征anchor对应链接到的关联URL;
402
根据所述指定网页的URL以及所述翻页特征anchor对应链接到的关联URL计算与所述指定网页对应的关联网页URL模式pattern;
403
通过对关联网页URL模式pattern中的共性部分进行结构分析，提取关联网页URL模式pattern中的翻页块，将所述翻页块替换为首页标识获得首页关联网页的URL。
404

图 4

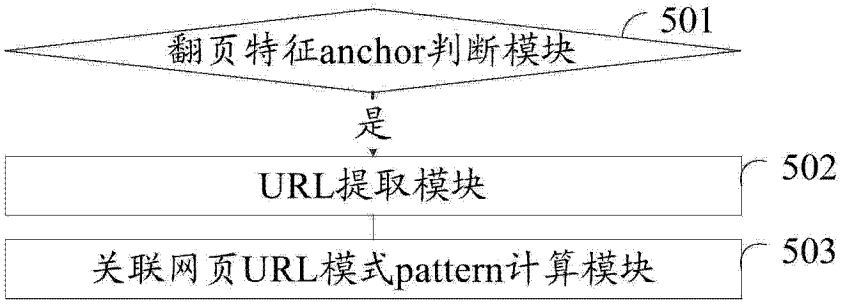

501
翻页特征anchor判断模块
是
URL提取模块
502
关联网页URL模式pattern计算模块
503

图 5

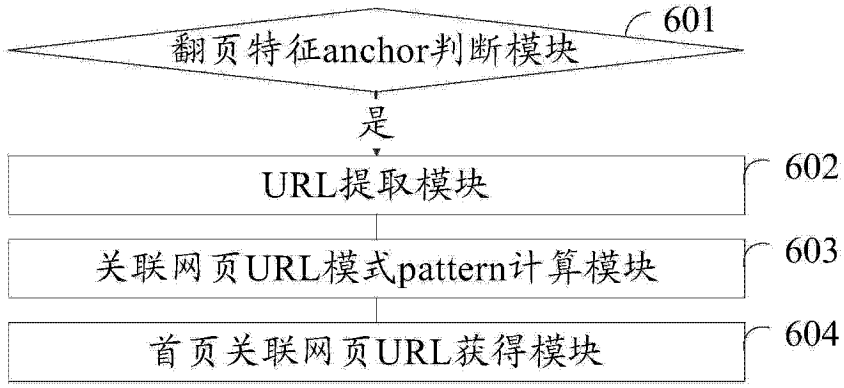

601
翻页特征anchor判断模块
是
URL提取模块
602
关联网页URL模式pattern计算模块
603
首页关联网页URL获得模块
604

图 6