(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109582850 B
(45)授权公告日 2021.07.02

(21)申请号 201811467095.2
(22)申请日 2018.12.03
(65)同一申请的已公布的文献号
申请公布号 CN 109582850 A
(43)申请公布日 2019.04.05
(73)专利权人 金瓜子科技发展(北京)有限公司
地址 100085 北京市海淀区清河安宁庄东路18号23号楼二层2356
(72)发明人 陈耽思
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 罗满
(51)Int.Cl.
G06F 16/951(2019.01)

(56)对比文件
CN 108595583 A,2018.09.28
CN 1797395 A,2006.07.05
CN 104933138 A,2015.09.23
US 8990200 B1,2015.03.24
CN 108595583 A,2018.09.28

审查员 董统传

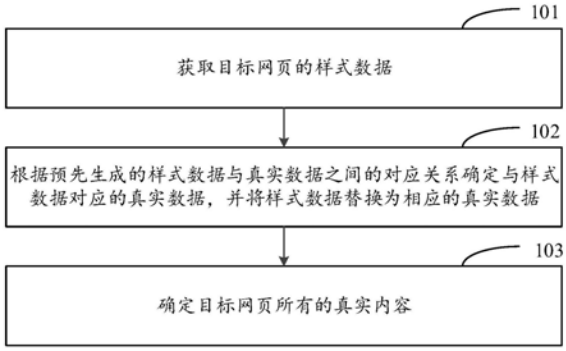
权利要求书2页 说明书11页 附图5页

(54)发明名称

一种网页爬取的方法、装置、存储介质及电子设备

(57)摘要

本发明提供了一种网页爬取的方法、装置、存储介质及电子设备，其中，该方法包括：获取目标网页的样式数据，样式数据为对目标网页的源数据基于反爬策略生成的数据；根据预先生成的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据，并将样式数据替换为相应的真实数据；确定目标网页所有的真实内容。通过本发明实施例提供的一种网页爬取的方法、装置、存储介质及电子设备，根据预先生成的样式数据与真实数据之间的对应关系确定与该样式数据对应的真实数据，从而可以快速、准确地获取网页的所有真实数据；该方式不需要重复使用图像识别技术识别网页中的数据，节省了大量的处理资源，大大提高了抓取速度和抓取效率。



CN 109582850 B

1.一种网页爬取的方法,其特征在于,包括:

获取目标网页的样式数据,所述样式数据为在所述目标网页的源数据中基于反爬策略生成的数据;

判断是否存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系,在存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时,根据预先生成的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据,并将所述样式数据替换为相应的真实数据;其中,所述样式数据与真实数据之间的对应关系表示某一样式数据实际对应的真实数据;

确定所述目标网页所有的真实内容,所述真实内容包括与所述样式数据对应的真实数据;

其中,所述判断是否存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系包括:

确定所述样式数据的文件名称,并判断是否存在与所述文件名称相匹配的历史文件名称,所述历史文件名称为解析过的历史样式数据的文件名称;

在存在相匹配的历史文件名称时,确定存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系。

2.根据权利要求1所述的方法,其特征在于,在所述获取目标网页的样式数据之后,还包括:

在不存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时,建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系。

3.根据权利要求1所述的方法,其特征在于,与所述样式数据相匹配的样式数据与真实数据之间的对应关系为:有效历史样式数据与基于所述有效历史样式数据的解析结果所确定的真实数据之间的对应关系;所述有效历史样式数据为与所述文件名称相匹配的历史文件名称所对应的历史样式数据。

4.根据权利要求1所述的方法,其特征在于,所述判断是否存在与所述文件名称相匹配的历史文件名称包括:

将所述文件名称和历史文件名称分别分为多个子字符串,并确定所述文件名称的每个子字符串在所述文件名称中的排列顺序、以及所述历史文件名称的每个子字符串在所述历史文件名称中的排列顺序;

从最后顺位的子字符串开始,判断所述文件名称的子字符串与所述历史文件名称的相对应的子字符串是否相同,在二者不同时确定所述文件名称与所述历史文件名称不匹配;

在二者相同时,倒序确定下一顺位的子字符串,并重复上述判断所述文件名称的子字符串与所述历史文件名称的相对应的子字符串是否相同的过程,直至确定所述文件名称与所述历史文件名称不匹配、或者确定所述文件名称的所有子字符串与所述历史文件名称的所有子字符串全部相匹配;在确定所述文件名称的所有子字符串与所述历史文件名称的所有子字符串全部相匹配时,确定所述文件名称与所述历史文件名称相匹配。

5.根据权利要求2所述的方法,其特征在于,所述建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系包括:

创建本地网页,并将所述目标网页的样式数据加载至所述本地网页中;

获取所述本地网页的网页图像，并识别所述网页图像，确定所述网页图像中的真实数据；

建立所述样式数据与识别出的相应的真实数据之间的对应关系。

6.根据权利要求2所述的方法，其特征在于，在所述建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系之后，还包括：

将与所述样式数据相匹配的样式数据与真实数据之间的对应关系存储至数据库中。

7.根据权利要求1所述的方法，其特征在于，所述样式数据包括文字样式数据和/或图片样式数据。

8.一种网页爬取的装置，其特征在于，包括：

获取模块，用于获取目标网页的样式数据，所述样式数据为在所述目标网页的源数据中基于反爬策略生成的数据；

处理模块，用于根据预先生成的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据，并将所述样式数据替换为相应的真实数据；其中，所述样式数据与真实数据之间的对应关系表示某一样式数据实际对应的真实数据；

确定模块，用于确定所述目标网页所有的真实内容，所述真实内容包括与所述样式数据对应的真实数据；

判断模块，用于判断是否存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系；在存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时，所述处理模块根据相匹配的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据；

其中，所述判断模块具体用于：确定所述样式数据的文件名称，并判断是否存在与所述文件名称相匹配的历史文件名称，所述历史文件名称为解析过的历史样式数据的文件名称；

在存在相匹配的历史文件名称时，确定存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系。

9.一种存储介质，其特征在于，所述存储介质存储有计算机可执行指令，所述计算机可执行指令用于执行权利要求1-7任意一项所述的网页爬取的方法。

10.一种电子设备，其特征在于，包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-7任意一项所述的网页爬取的方法。

一种网页爬取的方法、装置、存储介质及电子设备

技术领域

[0001] 本发明涉及网页爬取的技术领域，具体而言，涉及一种网页爬取的方法、装置、存储介质及电子设备。

背景技术

[0002] 传统爬虫从一个或若干个初始URL(统一资源定位符)开始，获得初始URL对应的网页上的URL以及其他内容，同时也将当前页面上获得的新的URL放入队列继续抓取，直到满足系统的一定停止条件。所有被爬虫抓取的内容将会被存储，按照关键字、文本、图片、音视频等进行分类、分析、过滤，并建立索引，以便之后的查询和检索。

[0003] 然而，有些网站采取了反爬虫措施，阻止爬虫获取网页源代码，从而爬虫无法准确完成对目标网页信息的获取。为了可以准确识别采取反爬策略的网页，一般方法是在打开网页后进行截图保存图片，并通过OCR(Optical Character Recognition，光学字符识别)识别图片来获取网页中的所有真实文本数据。但是使用OCR识别会占用大量的CPU资源和处理时间，网页抓取效率较低。

发明内容

[0004] 为解决上述问题，本发明实施例的目的在于提供一种网页爬取的方法、装置、存储介质及电子设备。

[0005] 第一方面，本发明实施例提供了一种网页爬取的方法，包括：

[0006] 获取目标网页的样式数据，所述样式数据为对所述目标网页的源数据基于反爬策略生成的数据；

[0007] 根据预先生成的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据，并将所述样式数据替换为相应的真实数据；

[0008] 确定所述目标网页所有的真实内容，所述真实内容包括与所述样式数据对应的真实数据。

[0009] 在一种可能的实现方式中，在所述获取目标网页的样式数据之后，还包括：

[0010] 判断是否存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系，在存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时，根据相匹配的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据；

[0011] 在不存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时，建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系。

[0012] 在一种可能的实现方式中，所述判断是否存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系包括：

[0013] 确定所述样式数据的文件名称，并判断是否存在与所述文件名称相匹配的历史文件名称，所述历史文件名称为解析过的历史样式数据的文件名称；

[0014] 在存在相匹配的历史文件名称时，确定存在与所述样式数据相匹配的样式数据与

真实数据之间的对应关系，且与所述样式数据相匹配的样式数据与真实数据之间的对应关系为有效历史样式数据与基于所述有效历史样式数据的解析结果所确定真实数据之间的对应关系；所述有效历史样式数据为与所述文件名称相匹配的历史文件名称所对应的历史样式数据。

[0015] 在一种可能的实现方式中，所述判断是否存在与所述文件名称相匹配的历史文件名称包括：

[0016] 将所述文件名称和历史文件名称分别分为多个子字符串，并确定所述文件名称的每个子字符串在所述文件名称中的排列顺序、以及所述历史文件名称的每个子字符串在所述历史文件名称中的排列顺序；

[0017] 从最后顺位的子字符串开始，判断所述文件名称的子字符串与所述历史文件名称的相对应的子字符串是否相同，在二者不同时确定所述文件名称与所述历史文件名称不匹配；

[0018] 在二者相同时，倒序确定下一顺位的子字符串，并重复上述判断所述文件名称的子字符串与所述历史文件名称的相对应的子字符串是否相同的过程，直至确定所述文件名称与所述历史文件名称不匹配、或者确定所述文件名称的所有子字符串与所述历史文件名称的所有子字符串全部相匹配，在确定所述文件名称的所有子字符串与所述历史文件名称的所有子字符串全部相匹配时，确定所述文件名称与所述历史文件名称相匹配。

[0019] 在一种可能的实现方式中，所述建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系包括：

[0020] 创建本地网页，并将所述目标网页的样式数据加载至所述本地网页中；

[0021] 获取所述本地网页的网页图像，并识别所述网页图像，确定所述网页图像中的真实数据；

[0022] 建立所述样式数据与识别出的相应的真实数据之间的对应关系。

[0023] 在一种可能的实现方式中，在所述建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系之后，该方法还包括：

[0024] 将与所述样式数据相匹配的样式数据与真实数据之间的对应关系存储至数据库中。

[0025] 在一种可能的实现方式中，所述样式数据包括文字样式数据和/或图片样式数据。

[0026] 第二方面，本发明实施例还提供了一种网页爬取的装置，包括：

[0027] 获取模块，用于获取目标网页的样式数据，所述样式数据为对所述目标网页的源数据基于反爬策略生成的数据；

[0028] 处理模块，用于根据预先生成的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据，并将所述样式数据替换为相应的真实数据；

[0029] 确定模块，用于确定所述目标网页所有的真实内容，所述真实内容包括与所述样式数据对应的真实数据。

[0030] 第三方面，本发明实施例还提供了一种存储介质，所述存储介质存储有计算机可执行指令，所述计算机可执行指令用于执行上述任意一项所述的网页爬取的方法。

[0031] 第四方面，本发明实施例还提供了一种电子设备，包括：

[0032] 至少一个处理器；以及，

[0033] 与所述至少一个处理器通信连接的存储器;其中,

[0034] 所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行上述任意一项所述的网页爬取的方法。

[0035] 本发明实施例上述第一方面提供的方案中,在爬取采用反爬策略的网页时提取网页中的样式数据,根据预先生成的样式数据与真实数据之间的对应关系确定与该样式数据对应的真实数据,从而可以快速、准确地获取网页的所有真实数据;该方式不需要重复使用图像识别技术识别网页中的数据,节省了大量的处理资源,大大提高了抓取速度和抓取效率。

[0036] 为使本发明的上述目的、特征和优点能更明显易懂,下文特举较佳实施例,并配合所附附图,作详细说明如下。

附图说明

[0037] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0038] 图1示出了本发明实施例所提供的一种网页爬取的方法的流程图;

[0039] 图2示出了本发明实施例所提供的网页爬取的方法中,判断是否存在对应关系的流程图;

[0040] 图3示出了本发明实施例所提供的网页爬取的方法中,建立与样式数据相匹配的样式数据与真实数据之间的对应关系的流程图;

[0041] 图4示出了本发明实施例所提供的另一种网页爬取的方法的流程图;

[0042] 图5示出了本发明实施例所提供的一种网页爬取的装置的结构示意图;

[0043] 图6示出了本发明实施例所提供的另一种网页爬取的装置的结构示意图;

[0044] 图7示出了本发明实施例所提供的网页爬取的装置中,解析模块的具体结构示意图;

[0045] 图8示出了本发明实施例所提供的用于执行网页爬取方法的电子设备的结构示意图。

具体实施方式

[0046] 在本发明的描述中,需要理解的是,术语“中心”、“纵向”、“横向”、“长度”、“宽度”、“厚度”、“上”、“下”、“前”、“后”、“左”、“右”、“竖直”、“水平”、“顶”、“底”“内”、“外”、“顺时针”、“逆时针”等指示的方位或位置关系为基于附图所示的方位或位置关系,仅是为了便于描述本发明和简化描述,而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作,因此不能理解为对本发明的限制。

[0047] 此外,术语“第一”、“第二”仅用于描述目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本发明的描述中,“多个”的含义是两个或两个以上,除非另有明确具体的限定。

[0048] 在本发明中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”、“固定”等术语应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本发明中的具体含义。

[0049] 本发明实施例提供的一种网页爬取的方法，参见图1所示，包括步骤101-103：

[0050] 步骤101：获取目标网页的样式数据，样式数据为对目标网页的源数据基于反爬策略生成的数据。

[0051] 本发明实施例中，目标网页为待爬取的网页，该目标网页采用了反爬策略，目标网页基于反爬策略对部分文字进行了特殊处理，并生成了样式数据。具体的，可以爬取得到目标网页的源数据，该源数据可以为html源码，其中包含目标网页的各个节点，比如meta节点、body节点等；由于目标网页基于反爬策略进行了特殊处理，目标网页的源数据中还包含基于反爬策略生成的样式数据，该样式数据可以为某个样式文件，或者是样式链接。该样式数据具体可以包括文字样式数据、图片样式数据等，其中，文字样式数据指的是将原网页的文字替换为另一文字或另一种格式(或字体)的文字，此时采用传统的爬取方法可能获取到的是错误的数据；例如，目标网页的页面显示某车辆的行驶里程是“12.66万公里”，但是目标网页源码中的行驶里程是“45.22万公里”，目标网页通过对源码进行格式处理，使得显示出真实的行驶里程12.66万公里；此处的行驶里程即为一种文字样式数据。

[0052] 图片样式数据指的是将原网页的文字替换为相应的图片、或者将原网页的图片替换为另一种图片，例如将具有隐私性的电话号码中的每个数字或多个数字的组合替换为相应的图片等，传统的爬取方法并不能正确获取到该电话号码，除非采用图像识别技术。

[0053] 步骤102：根据预先生成的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据，并将样式数据替换为相应的真实数据。

[0054] 本发明实施例中，预先生成样式数据与真实数据之间的对应关系，该对应关系可以表示某一样式数据实际对应的真实数据(真实文本或真实图片)，比如，某个图片所对应的真实文字等。根据该样式数据与真实数据之间的对应关系即可快速确定与该目标网页的样式数据对应的真实数据，之后将目标网页中的样式数据替换为可识别的真实数据，方便后续对目标网页的识别。

[0055] 步骤103：确定目标网页所有的真实内容，该真实内容包括与样式数据对应的真实数据。

[0056] 本发明实施例中，在将目标网页中的样式数据替换为相应的真实数据，即可快速获取该目标网页中所有的真实内容，爬取到该目标网页中的所有的真实数据，其包括在步骤102中确定的与样式数据对应的真实数据。

[0057] 本发明实施例提供的一种网页爬取的方法，在爬取采用反爬策略的网页时提取网页中的样式数据，根据预先生成的样式数据与真实数据之间的对应关系确定与该样式数据对应的真实数据，从而可以快速、准确地获取网页的所有真实数据；该方式不需要重复使用图像识别技术识别网页中的数据，节省了大量的处理资源，大大提高了抓取速度和抓取效率。

[0058] 在上述实施例的基础上，在步骤101“获取目标网页的样式数据”之后，还包括判断

是否存在对应关系的过程，参见图2所示，该过程具体包括步骤201-203：

[0059] 步骤201：判断是否存在与样式数据相匹配的样式数据与真实数据之间的对应关系，在存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，继续步骤202；在不存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，继续步骤203。

[0060] 步骤202：根据相匹配的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据。

[0061] 步骤203：建立与样式数据相匹配的样式数据与真实数据之间的对应关系。

[0062] 本发明实施例中，在获取到目标网页的样式数据，需要判断是否存在与样式数据相匹配的样式数据与真实数据之间的对应关系，即需要判断是否已经设置了样式数据与真实数据之间的对应关系，或者说该样式数据是否曾经被解析过。当存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，即可根据该样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据。其中，本申请中的步骤202与步骤102中的“根据预先生成的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据”本质相同，即在存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，可以继续执行步骤102。

[0063] 在不存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，说明该目标网页的样式数据从未被解析过，此时需要对该样式数据进行解析，建立与样式数据相匹配的样式数据与真实数据之间的对应关系；即本次通过对样式数据进行解析来确定目标网页的真实数据，并建立样式数据与真实数据之间的对应关系供后续爬取过程中使用。

[0064] 在上述实施例的基础上，通过样式数据的文件名称来判断是否存在相匹配的对应关系。具体的，上述步骤201“判断是否存在与样式数据相匹配的样式数据与真实数据之间的对应关系”包括：确定样式数据的文件名称，并判断是否存在与文件名称相匹配的历史文件名称，该历史文件名称为解析过的历史样式数据的文件名称。

[0065] 本发明实施例中，样式数据具体可以为一个样式文件或者一个样式链接，样式数据的文件名称即为样式文件的名称或者样式链接的地址。由于样式文件的名称一般是长字符串(比如64位字符串)，两个样式数据采用相同名称的概率极低，而不同的链接地址也指向不同的网络资源，故本发明实施例中，不同的文件名称对应不同的样式数据，通过判断文件名称是否被解析过来确定是否存在相应的对应关系。

[0066] 具体的，每解析一个样式数据后则将该样式数据作为历史样式数据，并记录该历史样式数据的历史文件名称；在当前需要解析某个样式数据时，判断历史文件名称中是否存在与该待解析的样式数据的文件名称相一致的历史文件名称，若存在，则说明该待解析的样式数据曾经被解析过，即存在样式数据与真实数据之间的对应关系；否则不存在对应关系。比如，文件名称为f1、f2、f3的三个样式数据已经被解析过，则该三个样式数据f1、f2、f3即为三个历史样式数据；若当前需要解析一个文件名称为f2的样式数据，则根据文件名称即可知该样式数据曾经被解析过，此时继续步骤202即可；若当前需要解析一个文件名称为f4的样式数据，而不存在文件名称为f4的历史样式数据，此时需要解析该文件名称为f4的样式数据，即需要继续步骤203。

[0067] 具体的，在存在相匹配的历史文件名称时，确定存在与样式数据相匹配的样式数据与真实数据之间的对应关系，且该与样式数据相匹配的样式数据与真实数据之间的对应

关系为有效历史样式数据与基于有效历史样式数据的解析结果所确定真实数据之间的对应关系;有效历史样式数据为与文件名称相匹配的历史文件名称所对应的历史样式数据。

[0068] 具体的,在历史解析过程中,每解析一个历史样式数据,则可确定该历史样式数据的解析结果,此时将该解析结果作为与该历史样式数据对应的真实数据即可。即,在解析历史样式数据之后,将该历史样式数据的文件名称(即历史文件名称)标为被解析状态,同时将历史样式数据与解析结果之间的对应关系作为样式数据与真实数据之间的对应关系。在当前阶段,当判断存在与文件名称相匹配的历史文件名称时,说明存在与该样式数据相匹配的样式数据与真实数据之间的对应关系,此时确定与该文件名称相匹配的历史样式数据,即有效历史样式数据,并确定有效历史样式数据与解析结果之间的对应关系,该有效历史样式数据与解析结果之间的对应关系即为与当前的样式数据相匹配的样式数据与真实数据之间的对应关系。在确定样式数据与真实数据之间的对应关系之后,即可继续执行上述的步骤202。

[0069] 可选的,本发明实施例中,上述“判断是否存在与文件名称相匹配的历史文件名称”具体包括:

[0070] 步骤A1:将文件名称和历史文件名称分别分为多个子字符串,并确定文件名称的每个子字符串在该文件名称中的排列顺序、以及历史文件名称的每个子字符串在该历史文件名称中的排列顺序

[0071] 步骤A2:从最后顺位的子字符串开始,判断文件名称的子字符串与历史文件名称的相对应的子字符串是否相同,在二者不同时,继续步骤A3;在二者相同时,继续步骤A4。

[0072] 步骤A3:确定文件名称与历史文件名称不匹配。

[0073] 步骤A4:倒序确定下一顺位的子字符串,并重复上述判断文件名称的子字符串与历史文件名称的相对应的子字符串是否相同的过程,即重复步骤A2,直至确定文件名称与历史文件名称不匹配、或者确定文件名称的所有子字符串与历史文件名称的所有子字符串全部相匹配,在确定文件名称的所有子字符串与历史文件名称的所有子字符串全部相匹配时,确定文件名称与历史文件名称相匹配。

[0074] 本发明实施例中,由于样式数据的文件名称为长字符串或者链接地址,而链接地址一般也为长字符串,直接判断两个长字符串(文件名称的长字符串和历史文件名称的长字符串)是否相同会存在处理量过大的问题,故本实施例将文件名称分为多段的子字符串,依次分段判断文件名称和历史文件名称的子字符串是否相同,若二者不同,则说明文件名称和历史文件名称一定不相同,此时不需要判断其他的子字符串,从而减少处理量;若二者相同,则可以继续判断下一个子字符串是否相同,直至确定某个子字符串不同(说明文件名称和历史文件名称不同)或者所有的子字符串完全相同(说明文件名称和历史文件名称相同)。

[0075] 同时,为了统一,同一网站中所用的样式数据的文件名称的前一部分可能相同,区别点主要在名称后面的部分,故本实施例中通过倒序的顺序进行判断,从而可以更高概率确定不同的子字符串,从而进一步提高处理效率。

[0076] 具体的,在将一个文件名称的长字符串分为多个子字符串后,按照子字符串在文件名称中的位置即可对所有的子字符串进行排序,之后倒序判断子字符串是否相同。例如,文件名称为“aabbccddeeff”,历史文件名称为“aabbccddyeff”,首先将文件名称和历史文

件名称分段，比如三个字符为一组，分为多个子字符串，文件名称包括四个子字符串，依次为“aab”、“bcc”、“dde”、“eff”，历史文件名称的子字符串为“aab”、“bcc”、“ddy”、“eff”。

[0077] 在倒序判断时，先判断最后顺位的一个子字符串是否相同，即判断文件名称的“eff”和历史文件名称的“eff”相同，若二者不同，则说明文件名称与历史文件名称不同，不需要判断二者其他的子字符串是否相同；若二者相同，则倒序选取下一顺位的子字符串，继续判断两个子字符串是否相同，即选取下一顺位的子字符串“dde”与“ddy”，判断二者是否相同，如此重复倒序判断的过程，直至确定文件名称与历史文件名称是否相同。本实施例中，通过将文件名称分为多个子字符串，以子字符串为单位判断文件名称与历史文件名称是否相同，在确定子字符串不同时不需要执行后续的判断过程，可以减少处理量，提高处理效率；同时，基于文件名称的特点，采用倒序判断的方式可以更高概率定位到不同的子字符串，进一步提高了处理效率。

[0078] 在上述实施例的基础上，参见图3所示，步骤203“建立与样式数据相匹配的样式数据与真实数据之间的对应关系”包括步骤2031-2033：

[0079] 步骤2031：创建本地网页，并将目标网页的样式数据加载至本地网页中。

[0080] 本发明实施例中，当需要对样式数据进行解析时，在本地创建一个网页文件，即本地网页，之后将目标网页中未解析的样式数据加载至本地网页中，即通过本地网页加载目标网页样式数据的方式来提取、并可视化显示该样式数据。

[0081] 可选的，当目标网页中包含多个样式数据时，可以将所有的样式数据统一加载至同一个本地网页，可以提高对样式数据的解析效率。此外，该本地网页中可以为格式化网页，用于为多个样式数据添加唯一的标记，以在同一个本地网页中准确区分不同的样式数据；例如在本地网页中添加序号①、②、③等，并将每个样式数据加载至与序号相对应的位置，比如序号的后方。

[0082] 步骤2032：获取本地网页的网页图像，并识别网页图像，确定网页图像中的真实数据。

[0083] 步骤2033：建立样式数据与识别出的相应的真实数据之间的对应关系。

[0084] 本发明实施例中，本地网页中只包含所显示的样式数据，此时识别本地网页的网页图像时，可以准确识别出样式数据所对应的文本，具体可采用OCR识别技术识别该网页图像。例如，目标网页源代码中的行驶里程45.22万公里设有特殊样式处理，则该行驶里程为一样式数据，此时将该样式数据加载至本地网页中后，本地网页中可视化显示的行驶里程为12.66万公里，通过识别本地网页的网页图像即可确定该真实数据为“12.66万公里”，此时即可建立样式数据与真实数据之间的对应关系，即该行驶里程的样式数据“45.22万公里”与真实数据“12.66万公里”是对应的。其中，通过打开本地网页并截图即可得到本地网页的网页图像，也可采用其他获取方式，本实施例对此不做限定。

[0085] 本发明实施例中，建立另一个本地网页并加载样式数据，之后利用网页图像来解析样式数据，进而可以获得样式数据与真实数据之间的对应关系；该方式只需要识别目标网页的样式数据，不需要图像识别目标网页的所有内容，可以减少图像识别的处理量，提高处理效率；同时，建立样式数据与真实数据之间的对应关系之后，后续再爬取具有该样式数据的网页时不需要进行图像识别，根据样式数据与真实数据之间的对应关系即可方便快速地确定真实文本，极大提高了抓取效率。

[0086] 在上述实施例的基础上，在步骤203“建立与样式数据相匹配的样式数据与真实数据之间的对应关系”之后，该方法还包括：将与样式数据相匹配的样式数据与真实数据之间的对应关系存储至数据库中。

[0087] 本发明实施例中，建立数据库来存储解析出的样式数据与真实数据之间的对应关系，每当需要爬取网页时即可通过查询数据库的方式来确定该待爬取网页中的样式数据是否已经被解析过。同时，通过数据库存储样式数据与真实数据之间的对应关系，方便对该对应关系进行管理，比如添加、删除或更新样式数据与真实数据之间的对应关系等。

[0088] 下面通过一个实施例详细介绍该网页爬取的方法流程。

[0089] 本发明实施例中，在获取样式数据后判断该样式数据是否被解析过，之后执行相应的处理流程。参见图4所示，该网页爬取的方法流程包括步骤401-408：

[0090] 步骤401：获取目标网页的样式数据，样式数据为基于反爬策略生成的数据。

[0091] 步骤402：判断是否存在与样式数据相匹配的样式数据与真实数据之间的对应关系，在存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，继续步骤403；在不存在与样式数据相匹配的样式数据与真实数据之间的对应关系时，继续步骤404。

[0092] 步骤403：根据相匹配的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据，并继续步骤407。

[0093] 步骤404：创建本地网页，并将目标网页的样式数据加载至本地网页中。

[0094] 步骤405：获取本地网页的网页图像，并识别网页图像，确定网页图像中的真实数据。

[0095] 步骤406：根据样式数据以及识别出的相应的真实数据建立样式数据与真实数据之间的对应关系，并将识别出的真实数据作为与该样式数据对应的真实数据。

[0096] 步骤407：将样式数据替换为相应的真实数据。

[0097] 步骤408：确定目标网页所有的真实内容。

[0098] 本发明实施例提供的一种网页爬取的方法，在爬取采用反爬策略的网页时提取网页中的样式数据，根据预先生成的样式数据与真实数据之间的对应关系确定与该样式数据对应的真实数据，从而可以快速、准确地获取网页的所有真实数据；该方式不需要重复使用图像识别技术识别网页中的数据，节省了大量的处理资源，大大提高了抓取速度和抓取效率。在解析样式数据时，只需要识别目标网页的样式数据即可，不需要图像识别目标网页的所有内容，可以减少图像识别的处理量，提高处理效率；同时，建立样式数据与真实数据之间的对应关系之后，后续再爬取具有该样式数据的网页时不需要进行图像识别，根据样式数据与真实数据之间的对应关系即可方便快速地确定真实文本，极大提高了抓取效率。利用样式数据文件名称不重复的特点，基于文件名称来判断该样式数据是否被解析过，且通过对文件名称进行分段以及倒序判断的方式，可以进一步提高判断该样式数据是否被解析过的处理效率。

[0099] 以上详细介绍了网页爬取的方法流程，该方法也可以通过相应的装置实现，下面详细介绍该装置的结构和功能。

[0100] 本发明实施例还提供了一种网页爬取的装置，参见图5所示，包括：

[0101] 获取模块51，用于获取目标网页的样式数据，所述样式数据为对所述目标网页的源数据基于反爬策略生成的数据；

[0102] 处理模块52，用于根据预先生成的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据，并将所述样式数据替换为相应的真实数据；

[0103] 确定模块53，用于确定所述目标网页所有的真实内容，所述真实内容包括与所述样式数据对应的真实数据。

[0104] 在一种可能的实现方式中，参见图6所示，该装置还包括判断模块54和解析模块55；

[0105] 在所述获取模块51获取目标网页的样式数据之后，所述判断模块54用于判断是否存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系；

[0106] 在存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时，所述处理模块52用于根据相匹配的样式数据与真实数据之间的对应关系确定与所述样式数据对应的真实数据；

[0107] 在不存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系时，所述解析模块55用于建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系。

[0108] 在一种可能的实现方式中，所述判断模块54具体用于：确定所述样式数据的文件名称，并判断是否存在与所述文件名称相匹配的历史文件名称，所述历史文件名称为解析过的历史样式数据的文件名称；

[0109] 在存在相匹配的历史文件名称时，确定存在与所述样式数据相匹配的样式数据与真实数据之间的对应关系，且与所述样式数据相匹配的样式数据与真实数据之间的对应关系为有效历史样式数据与基于所述有效历史样式数据的解析结果所确定真实数据之间的对应关系；所述有效历史样式数据为与所述文件名称相匹配的历史文件名称所对应的历史样式数据。

[0110] 在一种可能的实现方式中，所述判断模块54判断是否存在与所述文件名称相匹配的历史文件名称的步骤包括：

[0111] 将所述文件名称和历史文件名称分别分为多个子字符串，并确定所述文件名称的每个子字符串在所述文件名称中的排列顺序、以及所述历史文件名称的每个子字符串在所述历史文件名称中的排列顺序；

[0112] 从最后顺位的子字符串开始，判断所述文件名称的子字符串与所述历史文件名称的相对应的子字符串是否相同，在二者不同时确定所述文件名称与所述历史文件名称不匹配；

[0113] 在二者相同时，倒序确定下一顺位的子字符串，并重复上述判断所述文件名称的子字符串与所述历史文件名称的相对应的子字符串是否相同的过程，直至确定所述文件名称与所述历史文件名称不匹配、或者确定所述文件名称的所有子字符串与所述历史文件名称的所有子字符串全部相匹配，在确定所述文件名称的所有子字符串与所述历史文件名称的所有子字符串全部相匹配时，确定所述文件名称与所述历史文件名称相匹配。

[0114] 在一种可能的实现方式中，参见图7所示，所述解析模块55包括：

[0115] 预处理单元551，用于创建本地网页，并将所述目标网页的样式数据加载至所述本地网页中；

[0116] 识别单元552，用于获取所述本地网页的网页图像，并识别所述网页图像，确定所述网页图像中的真实数据；

[0117] 确定单元553，用于建立所述样式数据与识别出的相应的真实数据之间的对应关系。

[0118] 在一种可能的实现方式中，参见图6所示，该装置还包括存储模块56；

[0119] 在所述解析模块55建立与所述样式数据相匹配的样式数据与真实数据之间的对应关系之后，所述存储模块56用于将与所述样式数据相匹配的样式数据与真实数据之间的对应关系存储至数据库中。

[0120] 在一种可能的实现方式中，所述样式数据包括文字样式数据和/或图片样式数据。

[0121] 本发明实施例提供的一种网页爬取的装置，在爬取采用反爬策略的网页时提取网页中的样式数据，根据预先生成的样式数据与真实数据之间的对应关系确定与该样式数据对应的真实数据，从而可以快速、准确地获取网页的所有真实数据；该方式不需要重复使用图像识别技术识别网页中的数据，节省了大量的处理资源，大大提高了抓取速度和抓取效率。在解析样式数据时，只需要识别目标网页的样式数据即可，不需要图像识别目标网页的所有内容，可以减少图像识别的处理量，提高处理效率；同时，建立样式数据与真实数据之间的对应关系之后，后续再爬取具有该样式数据的网页时不需要进行图像识别，根据样式数据与真实数据之间的对应关系即可方便快速地确定真实文本，极大提高了抓取效率。利用样式数据文件名称不重复的特点，基于文件名称来判断该样式数据是否被解析过，且通过对文件名称进行分段以及倒序判断的方式，可以进一步提高判断该样式数据是否被解析过的处理效率。

[0122] 本发明实施例还提供了一种存储介质，所述存储介质存储有计算机可执行指令，其包含用于执行上述网页爬取的方法的程序，该计算机可执行指令可执行上述任意方法实施例中的方法。

[0123] 其中，所述存储介质可以是计算机能够存取的任何可用介质或数据存储设备，包括但不限于磁性存储器(例如软盘、硬盘、磁带、磁光盘(MO)等)、光学存储器(例如CD、DVD、BD、HVD等)、以及半导体存储器(例如ROM、EPROM、EEPROM、非易失性存储器(NAND FLASH)、固态硬盘(SSD))等。

[0124] 图8示出了本发明的另一个实施例的一种电子设备的结构框图。所述电子设备1100可以是具备计算能力的主机服务器、个人计算机PC、或者可携带的便携式计算机或终端等。本发明具体实施例并不对电子设备的具体实现做限定。

[0125] 该电子设备1100包括至少一个处理器(processor)1110、通信接口(Communications Interface)1120、存储器(memory array)1130和总线1140。其中，处理器1110、通信接口1120、以及存储器1130通过总线1140完成相互间的通信。

[0126] 通信接口1120用于与网元通信，其中网元包括例如虚拟机管理中心、共享存储等。

[0127] 处理器1110用于执行程序。处理器1110可能是一个中央处理器CPU，或者是专用集成电路ASIC(Application Specific Integrated Circuit)，或者是被配置成实施本发明实施例的一个或多个集成电路。

[0128] 存储器1130用于可执行的指令。存储器1130可能包含高速RAM存储器，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。存储器1130也可以是存储器阵列。存储器1130还可能被分块，并且所述块可按一定的规则组合成虚拟卷。存储器1130存储的指令可被处理器1110执行，以使处理器1110能够执行上述任意方法实施例中

的网页爬取的方法。

[0129] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应所述以权利要求的保护范围为准。

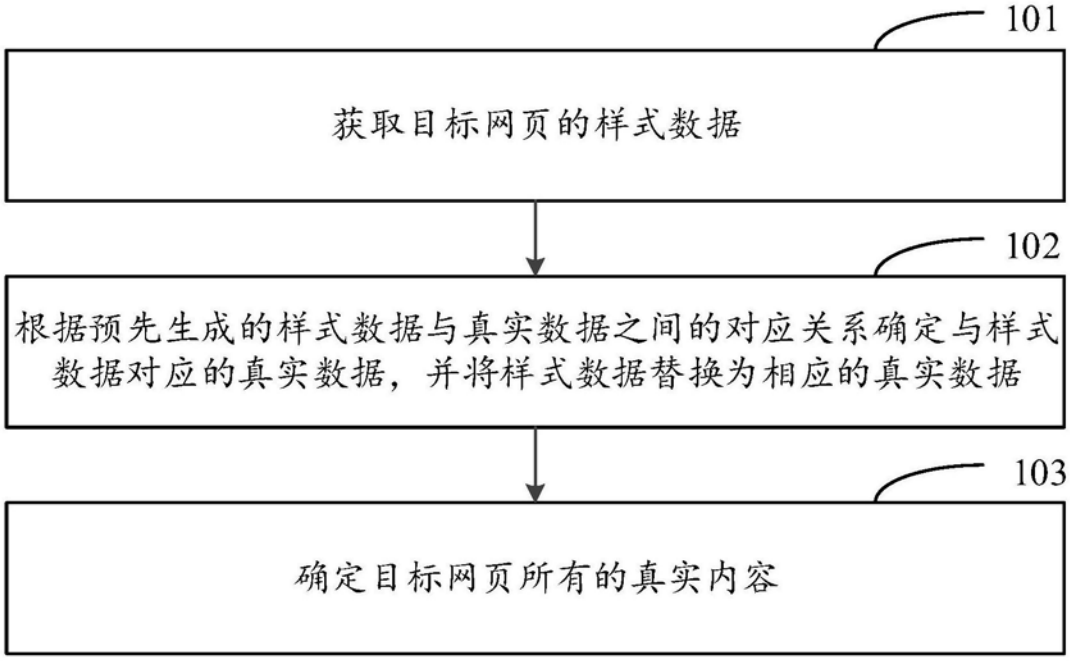

101
获取目标网页的样式数据
102
根据预先生成的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据，并将样式数据替换为相应的真实数据
103
确定目标网页所有的真实内容

图1

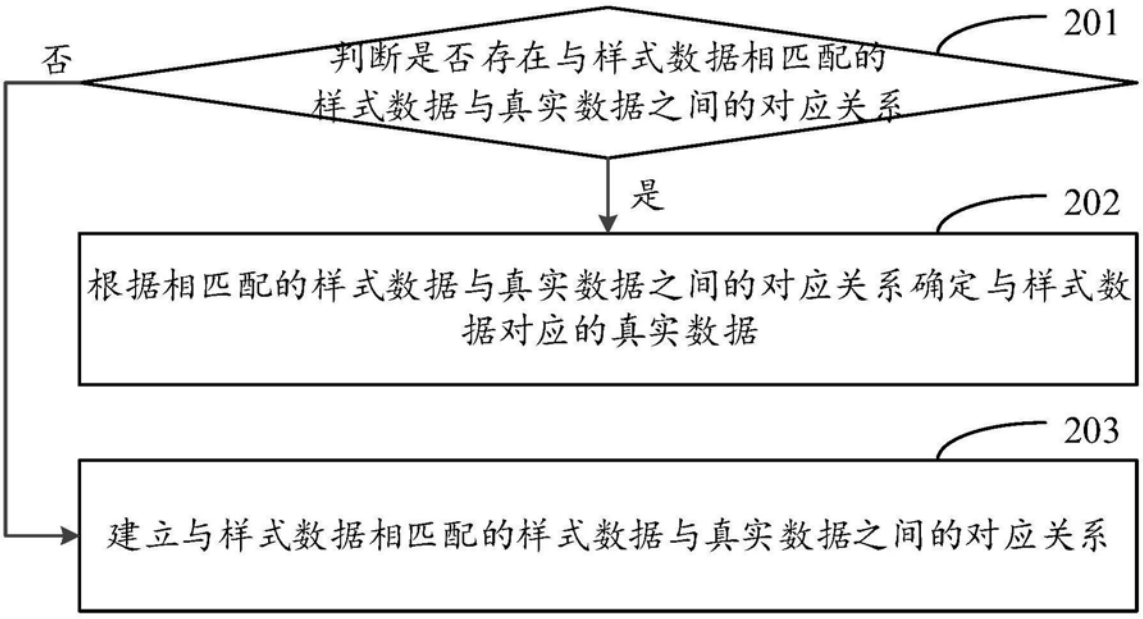

201
判断是否存在与样式数据相匹配的样式数据与真实数据之间的对应关系
否
是
202
根据相匹配的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据
203
建立与样式数据相匹配的样式数据与真实数据之间的对应关系

图2

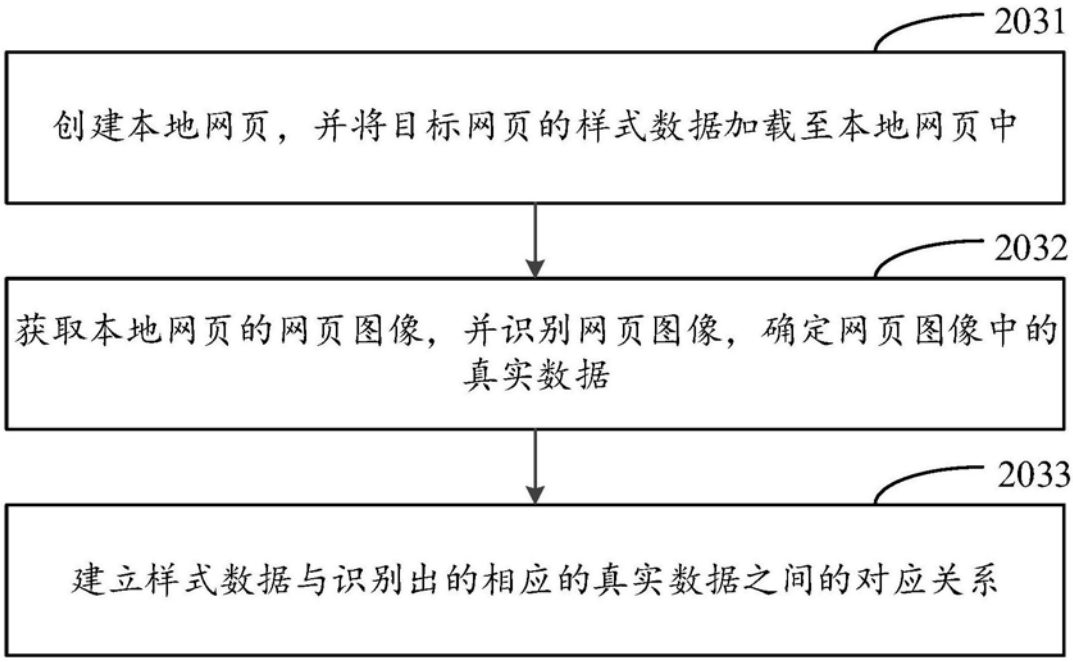

2031
创建本地网页，并将目标网页的样式数据加载至本地网页中
2032
获取本地网页的网页图像，并识别网页图像，确定网页图像中的真实数据
2033
建立样式数据与识别出的相应的真实数据之间的对应关系

图3

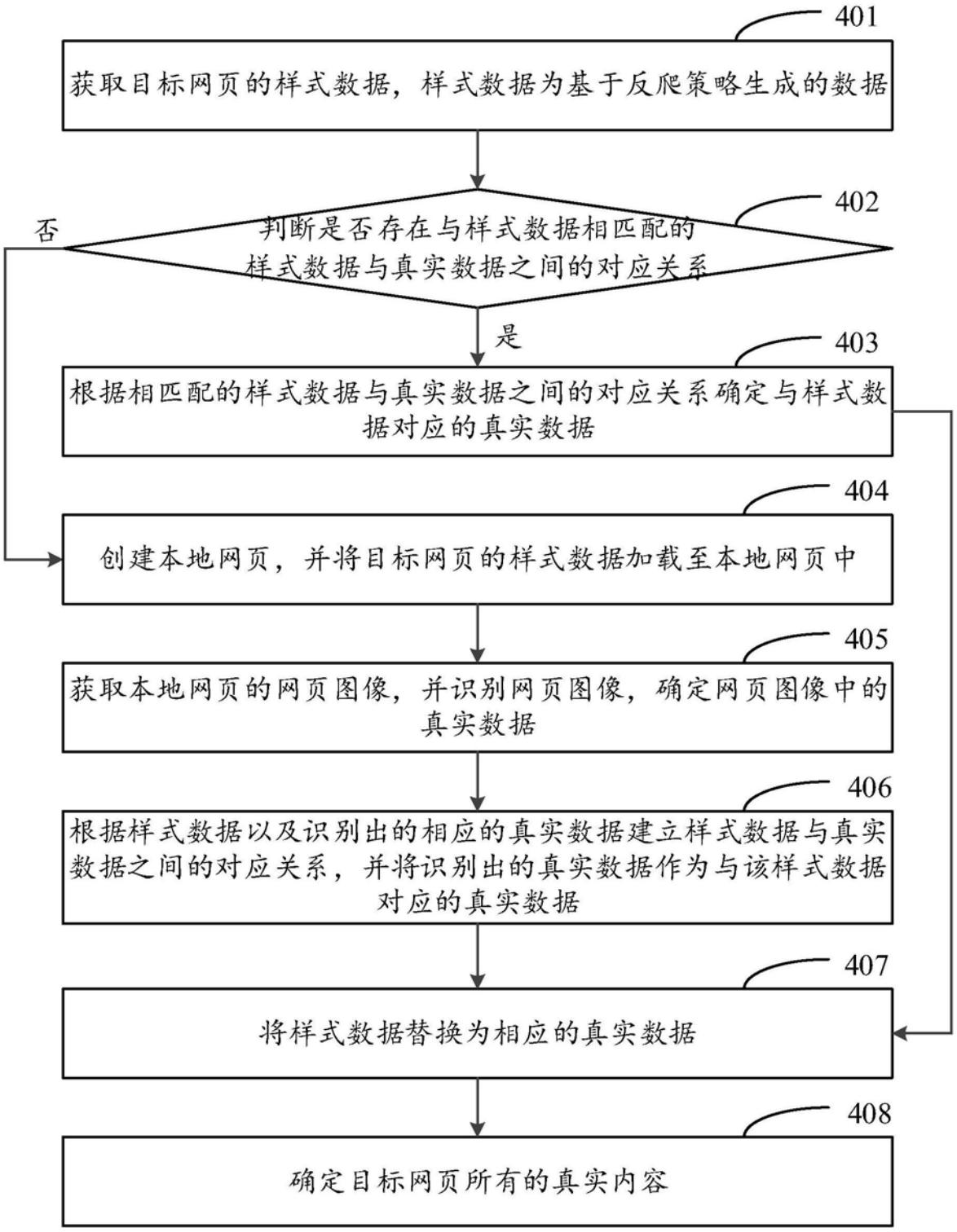

401
获取目标网页的样式数据，样式数据为基于反爬策略生成的数据
402
判断是否存在与样式数据相匹配的样式数据与真实数据之间的对应关系
否
是
403
根据相匹配的样式数据与真实数据之间的对应关系确定与样式数据对应的真实数据
404
创建本地网页，并将目标网页的样式数据加载至本地网页中
405
获取本地网页的网页图像，并识别网页图像，确定网页图像中的真实数据
406
根据样式数据以及识别出的相应的真实数据建立样式数据与真实数据之间的对应关系，并将识别出的真实数据作为与该样式数据对应的真实数据
407
将样式数据替换为相应的真实数据
408
确定目标网页所有的真实内容

图4

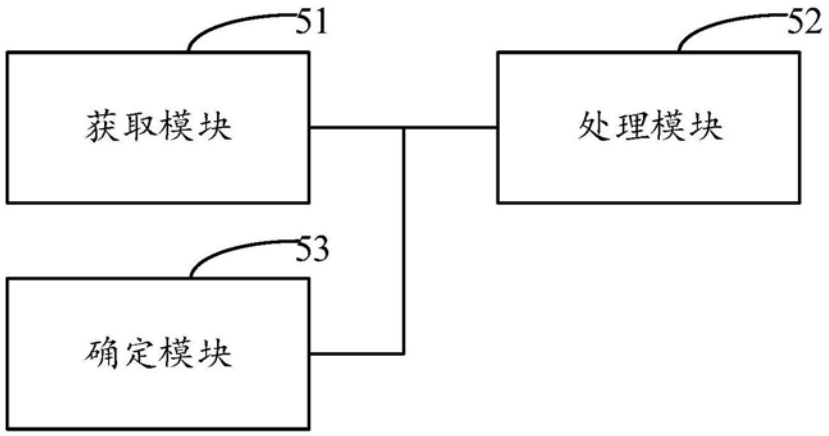

51
获取模块
52
处理模块
53
确定模块

图5

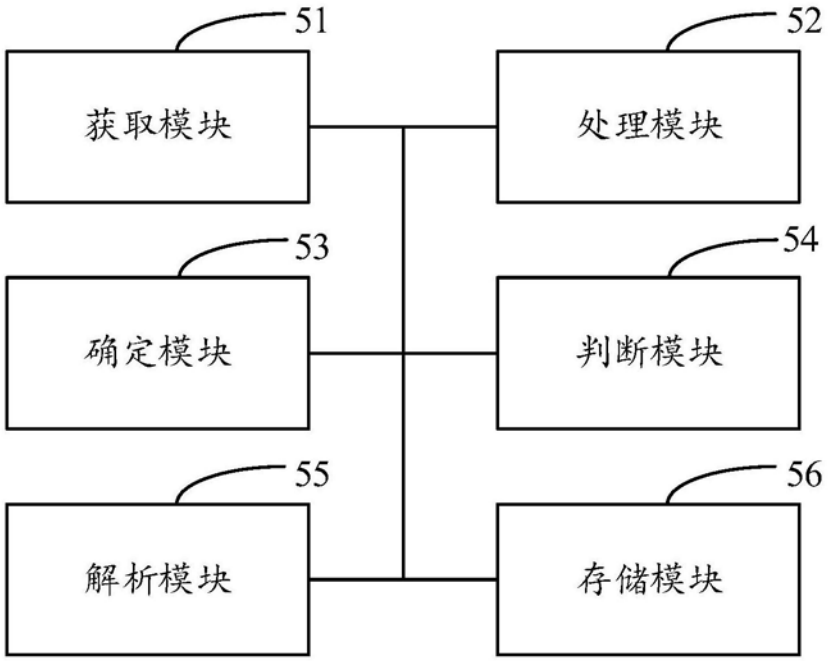

51
获取模块
52
处理模块
53
确定模块
54
判断模块
55
解析模块
56
存储模块

图6

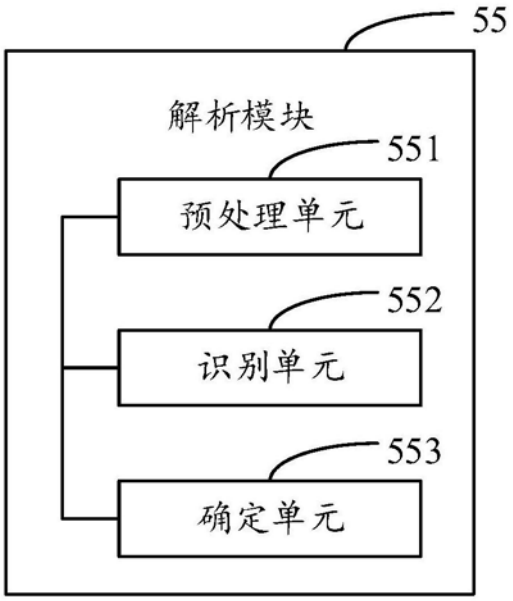

55
解析模块
551
预处理单元
552
识别单元
553
确定单元

图7

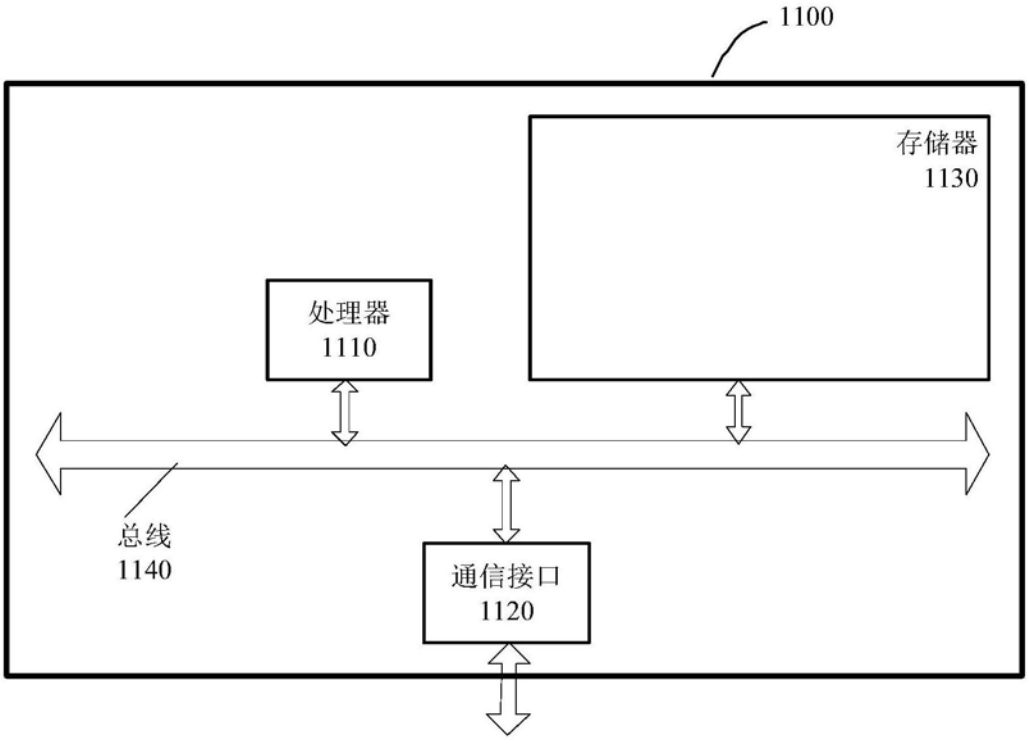

1100
存储器
1130
处理器
1110
总线
1140
通信接口
1120

图8