(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 110929128 A
(43)申请公布日 2020.03.27

(21)申请号 201911266670.7

(22)申请日 2019.12.11

(71)申请人 北京启迪区块链科技发展有限公司
地址 100084 北京市海淀区中关村东路1号院8号楼12层A1201F

(72)发明人 王鸣鹿 郑羽 周雷皓

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06F 16/951(2019.01)
G06F 9/54(2006.01)

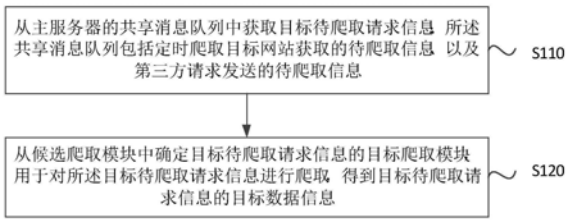
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种数据爬取方法、装置、设备和介质

(57)摘要

本发明公开了一种数据爬取方法、装置、设备和介质。该方法由分布式爬虫系统中的从服务器执行，其中，该方法包括：从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息；从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。本实施例的技术方案，增加了源目标网站的全面性和可读性，通过通用爬虫程序对不同目标源网站数据的进行抓取，满足了逐渐增多的不同源网站的抓取需求。



CN 110929128 A

1.一种数据爬取方法，其特征在于，由分布式爬虫系统中的从服务器执行，所述方法包括：

从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息；

从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。

2.根据权利要求1所述的方法，其特征在于，从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，包括：

将所述目标待爬取请求信息中统一资源定位符的属性信息与候选爬取模块的配置信息进行匹配，确定目标待爬取请求信息的目标爬取模块。

3.根据权利要求1所述的方法，其特征在于，在从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息之后，还包括：

确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新。

4.根据权利要求3所述的方法，其特征在于，确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新，包括：

确定目标待爬取请求信息与目标数据库中任一目标待爬取请求信息是否相同；

若相同，则比较所述目标待爬取请求信息的目标数据信息与所述目标数据库中任一目标待爬取请求信息的目标数据信息的哈希值是否相同；

若不相同，则根据所述目标待爬取请求信息的目标数据信息对所述目标数据库中任一目标待爬取请求信息的目标数据信息进行更新。

5.一种数据爬取装置，其特征在于，配置于分布式爬虫系统中的从服务器中，所述装置包括：

请求信息获取模块，用于从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息；

爬取模块确定模块，用于从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。

6.根据权利要求5所述的装置，其特征在于，所述爬取模块确定模块具体用于：

将所述目标待爬取请求信息中统一资源定位符的属性信息与候选爬取模块的配置信息进行匹配，确定目标待爬取请求信息的目标爬取模块。

7.根据权利要求5所述的装置，其特征在于，所述装置还包括：

更新模块，用于确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新。

8.根据权利要求7所述的装置，其特征在于，所述更新模块具体用于：

确定目标待爬取请求信息与目标数据库中任一目标待爬取请求信息是否相同；

若相同，则比较所述目标待爬取请求信息的目标数据信息与所述目标数据库中任一目标待爬取请求信息的目标数据信息的哈希值是否相同；

若不相同，则根据所述目标待爬取请求信息的目标数据信息对所述目标数据库中任一目标待爬取请求信息的目标数据信息进行更新。

9.一种设备，其特征在于，所述设备包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-4中任一所述的一种数据爬取方法。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-4中任一所述的一种数据爬取方法。

一种数据爬取方法、装置、设备和介质

技术领域

[0001] 本发明实施例涉及爬虫技术领域，尤其涉及一种数据爬取方法、装置、设备和介质。

背景技术

[0002] 数据爬取技术是按照一定的规则，自动抓取目标网站信息的程序。利用网络传输协议向服务器发起请求，接收服务器响应数据(请求状态、网页内容)，利用网页解析器将响应数据中所需的有效信息进行提取以及清洗，将信息以一种有效的方式进行存储。

[0003] 现有技术中，分布式爬虫系统中的源网站比较单一，针对不同源网站需要对应的启动不同的爬虫程序，不同爬虫程序启动过多不便于管理。

发明内容

[0004] 本发明提供一种数据爬取方法、装置、设备和介质，以增加了源目标网站的全面性和可读性，通过通用爬虫程序对不同目标源网站数据的进行抓取，满足了逐渐增多的不同源网站的抓取需求。

[0005] 第一方面，本发明实施例提供了一种数据爬取方法，由分布式爬虫系统中的从服务器执行，该方法包括：

[0006] 从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息；

[0007] 从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息；

[0008] 第二方面，本发明实施例还提供了一种数据爬取装置，配置于分布式爬虫系统中的从服务器中，该装置包括：

[0009] 请求信息获取模块，用于从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求方发送的待爬取信息；

[0010] 爬取模块确定模块，用于从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息；

[0011] 第三方面，本发明实施例还提供了一种设备，该设备包括：

[0012] 一个或多个处理器；

[0013] 存储装置，用于存储一个或多个程序，

[0014] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如本发明实施例中任一所述的一种数据爬取方法。

[0015] 第四方面，本发明实施例还提供了一种计算机可读介质，其上存储有计算机程序，该程序被处理器执行时实现如本发明实施例中任一所述的一种数据爬取方法。

[0016] 本发明通过分布式爬虫系统中的从服务器从主服务器中的共享消息队列中获取目标待爬取请求信息，从候选爬取模块中确定目标待爬取信息的目标爬取模块，对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。其中，共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取请求，将原本不同目标源网站的数据整合，增加了信息的全面性和可读性；由不同的爬取模块处理不同的目标待爬取请求信息，避免启动多个爬虫程序，方便管理。

附图说明

[0017] 图1为本发明实施例一提供的一种数据抓取方法的流程图；

[0018] 图2为本发明实施例一提供的一种分布式爬虫系统示意图；

[0019] 图3为本发明实施例二提供的一种数据抓取方法的流程图；

[0020] 图4为本发明实施例三提供的一种数据抓取装置的结构示意图；

[0021] 图5为本发明实施例四提供的一种设备的结构示意图。

具体实施方式

[0022] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部结构。

[0023] 实施例一

[0024] 图1为本发明实施例一提供的一种数据爬取方法的流程图，本实施例可适用于对数据进行爬取的情况，该方法可以由分布式爬虫系统的从服务器来执行，参照图1，具体包括如下步骤：

[0025] 步骤110、从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息。

[0026] 示例性的，本实施例中的分布式爬虫系统为Scrapy-Redis分布式爬虫框架，具体参照图2，该分布式爬虫系统包括一个主服务器Master和若干个从服务器Slave，主服务器中的Redis数据库中存储有共享消息队列，共享消息队列由若干统一资源定位符组成，若干个从服务器可以从共享消息队列中获取目标待爬取请求信息。

[0027] 进一步的，共享消息队列包括定时爬取目标网站获取的待爬取请求信息，以及第三方请求发送的待爬取信息。具体的，不同的待爬取目标网站相应的设置不同的爬取时间，并配置好对应的请求信息以及待更新标示、配置信息等附加信息。当到达对应的目标网站设置的爬取时间，程序会自动的对不同的待爬取目标网站进行爬取，将获取到的待爬取信息存入Redis共享消息队列中。

[0028] 此外，该分布式爬虫系统还实时的接收第三方请求发送的待爬取信息，上述两种请求当时获取的待爬取信息都会存入到Redis共享消息队列中，实现了将原本不同目标源网站的数据整合，增加了信息的全面性和可读性。

[0029] 步骤120、从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。

[0030] 本实施例中，通过通用爬虫来适配不同的目标源网站的数据抓取和存储。该通用

爬虫程序包含多个配置不同的爬虫模块，用于满足不同目标待爬取请求信息的爬取要求。

[0031] 具体的，从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，包括：

[0032] 将所述目标待爬取请求信息中统一资源定位符的属性信息与候选爬取模块的配置信息进行匹配，确定目标待爬取请求信息的目标爬取模块。

[0033] 其中，统一资源定位符(Uniform Resource Locator;URL)是WWW的统一资源定位标志，就是指网络地址。不同的URL对应不同的网页属性，例如时间属性、登录属性等，相应的，爬虫程序的不同爬取模块预先根据网页属性进行配置，以满足不同属性网页的爬取要求。

[0034] 本实施例中，在从共享消息队列中获取到目标待爬取请求信息之后，将目标待爬取请求信息中URL对应的网页属性信息与爬虫程序中各候选模块的配置信息进行匹配，确定目标待爬取请求信息的目标爬取模块，进而调用对应的下载器进行下载，下载后根据获取到的返回信息进行分析，调用对应的解析器对网页信息进行提取清洗并存储，得到目标待爬取请求信息的目标数据信息。通过统一的任务调度及多种下载器的分工合作，可实现通用爬虫来适配不同的目标源网站的数据抓取与存储，避免了当目标源网站增多时，同时开启多个不同的处理任务的弊端，方便了对待爬取请求的统一管理。

[0035] 本实施例的技术方案，通过分布式爬虫系统中的从服务器从主服务器中的共享消息队列中获取目标待爬取请求信息，从候选爬取模块中确定目标待爬取信息的目标爬取模块，对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。其中，共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取请求，将原本不同目标源网站的数据整合，增加了信息的全面性和可读性；由不同的爬取模块处理不同的目标待爬取请求信息，避免启动多个爬虫程序，方便管理。

[0036] 实施例二

[0037] 图3为本发明实施例二提供的一种数据爬取方法。本发明实施例是在上述实施例的基础上，在步骤120之后还包括：确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新。参见图3，该方法具体包括：

[0038] 步骤210、从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息。

[0039] 步骤220、从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。

[0040] 步骤230、确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新。

[0041] 本实施例中，确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新，包括：

[0042] 确定目标待爬取请求信息与目标数据库中任一目标待爬取请求信息是否相同；

[0043] 若相同，则比较所述目标待爬取请求信息的目标数据信息与所述目标数据库中任一目标待爬取请求信息的目标数据信息的哈希值是否相同；

[0044] 若不相同，则根据所述目标待爬取请求信息的目标数据信息对所述目标数据库中任一目标待爬取请求信息的目标数据信息进行更新。

[0045] 具体的，在解析器对下载器下载的信息进行解析处理时，若目标待爬取请求信息已经存在于目标数据库中，则说明目标待爬取请求信息为已经爬取过的信息，进而将依据目标待爬取请求信息获取的目标数据信息的哈希值与目标数据库中存储的已经爬取获得的历史信息的哈希值进行比较，若哈希值相同，则说明两次爬取获取到的数据信息相同，不需要对目标数据库进行更新；若不哈希值相同，则将目标待爬取请求信息，以及依据目标待爬取请求信息获取的源网页对应模块结构存入mongodb数据库中。

[0046] 进一步的，若解析器提取的信息是网页链接时则将网页链接进行包装发送到Redis队列中等待下一轮的请求，当解析器提取的信息是数据时，则将数据信息对应存储到目标数据库中。

[0047] 本实施例的技术方案，通过增加更新策略比对哈希值，解决数据更新、源目标网站变更导致数据丢失等问题。

[0048] 实施例三

[0049] 图4为本发明实施例三提供的一种数据爬取装置的结构示意图。参见图4，该装置具体包括请求信息获取模块310和爬取模块确定模块320。

[0050] 其中，信息获取模块310，用于从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息。

[0051] 爬取模块确定模块320，用于从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。

[0052] 进一步的，爬取模块确定模块320具体用于：将所述目标待爬取请求信息中统一资源定位符的属性信息与候选爬取模块的配置信息进行匹配，确定目标待爬取请求信息的目标爬取模块。

[0053] 进一步的，该装置还包括更新模块，用于确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新。

[0054] 具体的，更新模块具体用于：确定目标待爬取请求信息与目标数据库中任一目标待爬取请求信息是否相同；

[0055] 若相同，则比较所述目标待爬取请求信息的目标数据信息与所述目标数据库中任一目标待爬取请求信息的目标数据信息的哈希值是否相同；

[0056] 若不相同，则根据所述目标待爬取请求信息的目标数据信息对所述目标数据库中任一目标待爬取请求信息的目标数据信息进行更新。

[0057] 本发明实施例所提供的一种数据爬取装置可执行本发明任意实施例所提供的一种数据爬取方法，具备执行方法相应的功能模块和有益效果，再次不再进行赘述。

[0058] 实施例四

[0059] 图5为本发明实施例四提供的一种设备的结构示意图。图5示出了适于用来实现本发明实施方式的示例性设备12的框图。图5显示的设备12仅仅是一个示例，不应对本发明实

施例的功能和使用范围带来任何限制。

[0060] 如图5所示，设备12以通用计算设备的形式表现。设备12的组件可以包括但不限于：一个或者多个处理器或者处理单元16，系统存储器28，连接不同系统组件(包括系统存储器28和处理单元16)的总线18。

[0061] 总线18表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(ISA)总线，微通道体系结构(MAC)总线，增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0062] 设备12典型地包括多种计算机系统可读介质。这些介质可以是任何能够被设备12访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

[0063] 系统存储器28可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器(RAM)30和/或高速缓存存储器32。设备12可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统34可以用于读写不可移动的、非易失性磁介质(图5未显示，通常称为“硬盘驱动器”)。尽管图5中未示出，可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器，以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下，每个驱动器可以通过一个或者多个数据介质接口与总线18相连。存储器28可以包括至少一个程序产品，该程序产品具有一组(例如至少一个)程序模块，这些程序模块被配置以执行本发明各实施例的功能。

[0064] 具有一组(至少一个)程序模块42的程序/实用工具40，可以存储在例如存储器28中，这样的程序模块42包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块42通常执行本发明所描述的实施例中的功能和/或方法。

[0065] 设备/终端/服务器12也可以与一个或多个外部设备14(例如键盘、指向设备、显示器24等)通信，还可与一个或者多个使得用户能与该设备12交互的设备通信，和/或与使得该设备12能与一个或多个其它计算设备进行通信的任何设备(例如网卡，调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口22进行。并且，设备12还可以通过网络适配器20与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器20通过总线18与设备12的其它模块通信。应当明白，尽管图中未示出，可以结合设备12使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0066] 处理单元16通过运行存储在系统存储器28中的程序，从而执行各种功能应用以及数据处理，例如实现本发明实施例所提供的一种数据抓取方法。

[0067] 实施例五

[0068] 本发明实施例五还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现本发明实施例中任一所述的一种数据爬取方法，其中，所述方法包括：从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息；

[0069] 从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待

爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息。

[0070] 本发明实施例的计算机存储介质，可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0071] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0072] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括——但不限于无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0073] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如”C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0074] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

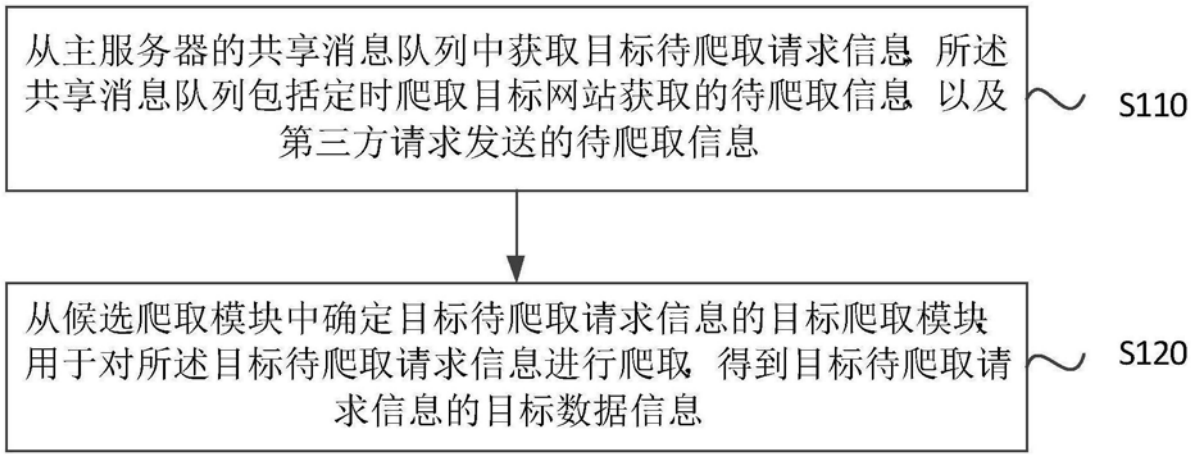

从主服务器的共享消息队列中获取目标待爬取请求信息，所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息
S110
从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息
S120

图1

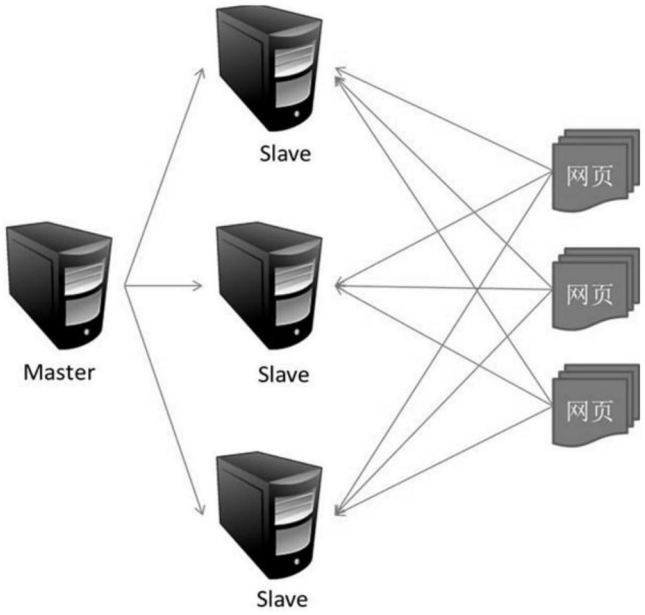

Slave
网页
网页
网页
Master
Slave
Slave

图2

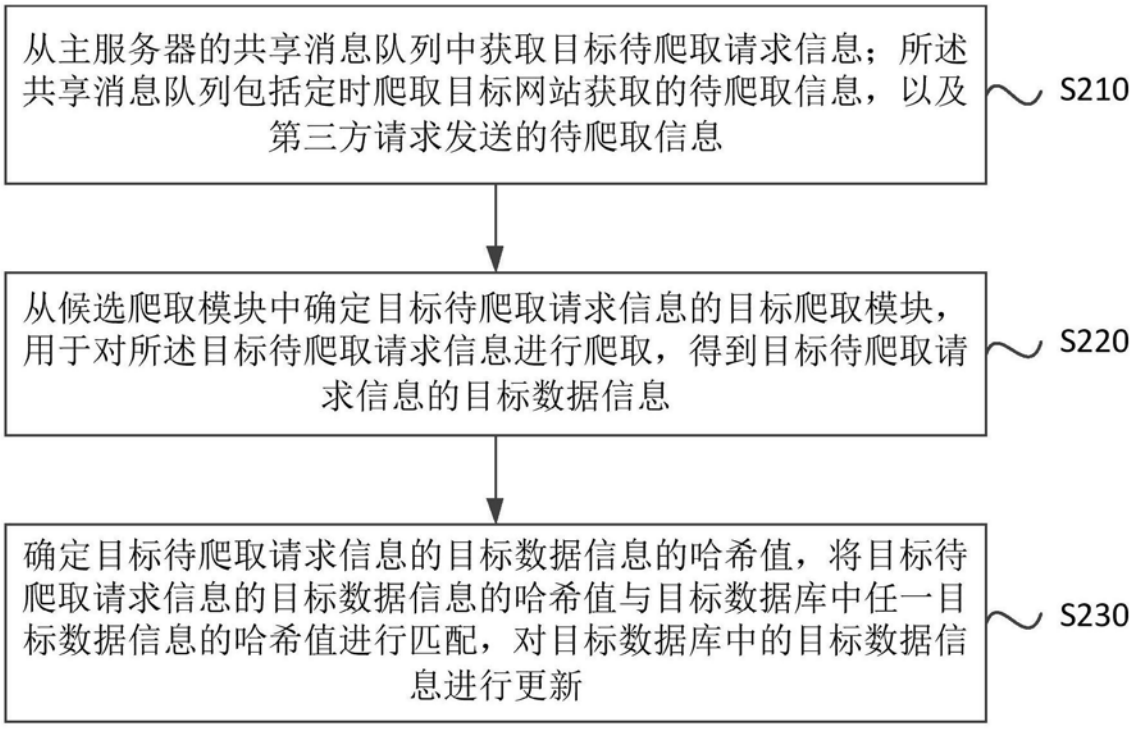

从主服务器的共享消息队列中获取目标待爬取请求信息；所述共享消息队列包括定时爬取目标网站获取的待爬取信息，以及第三方请求发送的待爬取信息
S210
从候选爬取模块中确定目标待爬取请求信息的目标爬取模块，用于对所述目标待爬取请求信息进行爬取，得到目标待爬取请求信息的目标数据信息
S220
确定目标待爬取请求信息的目标数据信息的哈希值，将目标待爬取请求信息的目标数据信息的哈希值与目标数据库中任一目标数据信息的哈希值进行匹配，对目标数据库中的目标数据信息进行更新
S230

图3

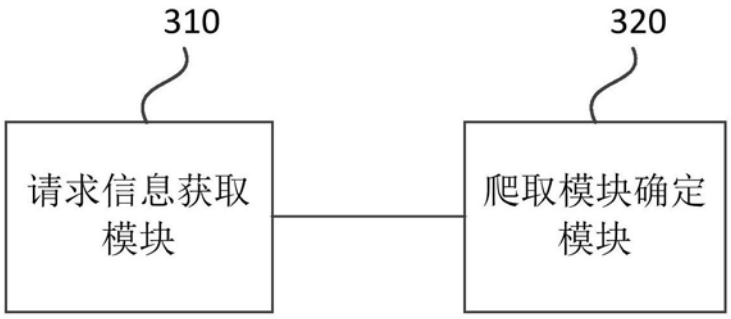

310
320
请求信息获取模块
爬取模块确定模块

图4

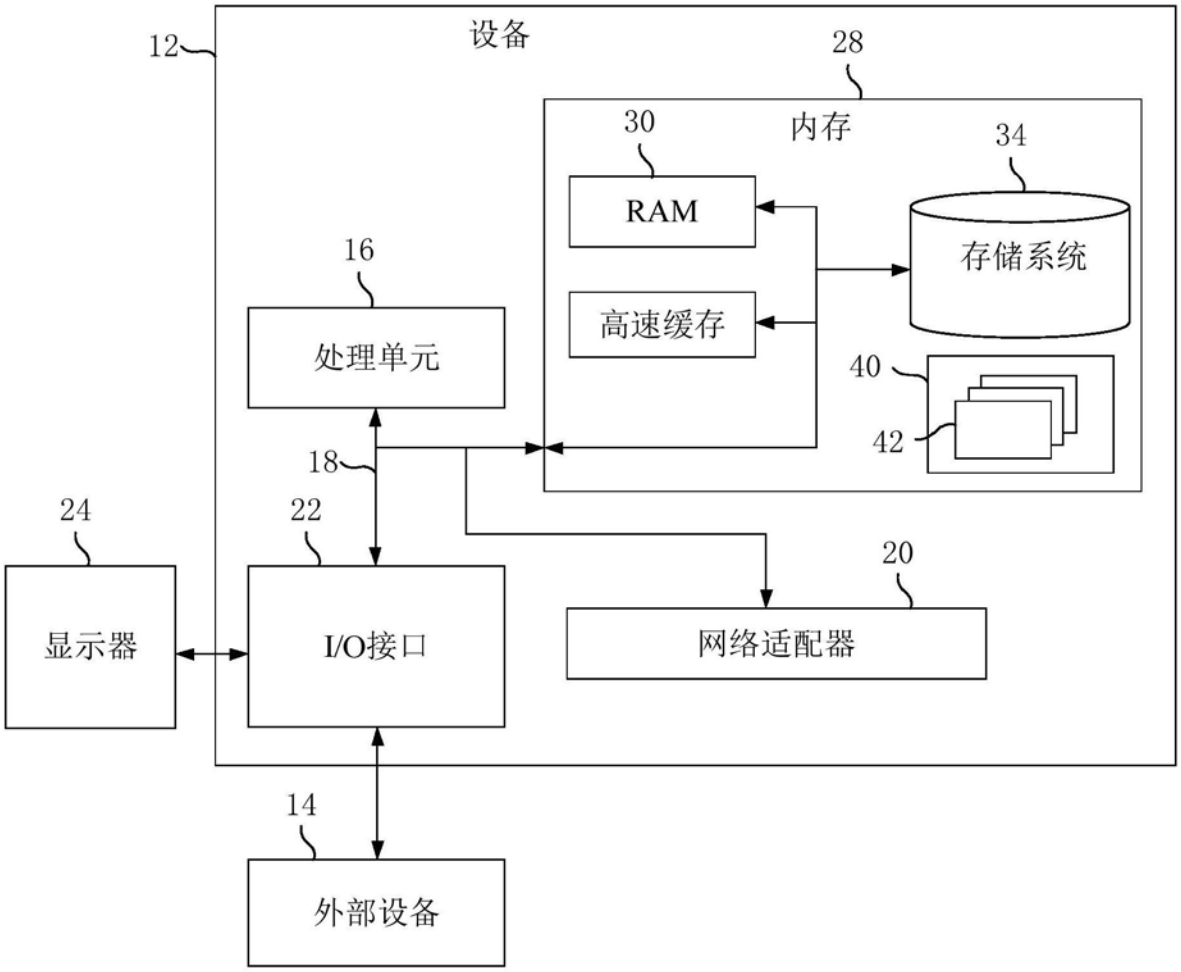

12
设备
28
30
内存
34
RAM
存储系统
16
高速缓存
处理单元
40
42
18
24
22
20
显示器
I/O接口
网络适配器
14
外部设备

图5