(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 110399546 B
(45)授权公告日 2022.02.08

(21)申请号 201910670803.0

(22)申请日 2019.07.23

(65)同一申请的已公布的文献号
申请公布号 CN 110399546 A

(43)申请公布日 2019.11.01

(73)专利权人 中南民族大学
地址 430074 湖北省武汉市洪山区民族大道182号中南民族大学

(72)发明人 雷建云 王锦群 郑禄 毛腾跃 孙翀 马尧 张蕾

(74)专利代理机构 深圳市世纪恒程知识产权代理事务所 44287
代理人 胡海国

(51)Int.Cl.
G06F 16/951(2019.01)
G06F 16/955(2019.01)

(56)对比文件
CN 108628871 A,2018.10.09
CN 107885777 A,2018.04.06
CN 109561163 A,2019.04.02
CN 106407485 A,2017.02.15
CN 107798106 A,2018.03.13
CN 110008419 A,2019.07.12
CN 108121706 A,2018.06.05
郁晨.一种高性能网络爬虫系统关键技术研究.《中国优秀硕士学位论文全文数据库 信息科技辑》.2018,
Weipeng Zhou等.An Improved Bloom Filter in Distributed Crawler.《2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation 》.2018,

审查员 轩海珍

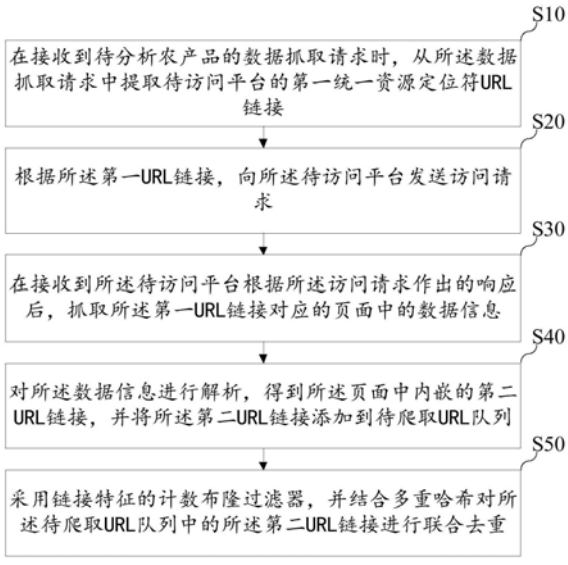
权利要求书3页 说明书13页 附图3页

(54)发明名称

基于网络爬虫的链接去重方法、装置、设备及存储介质

(57)摘要

本发明涉及互联网技术领域,公开了一种基于网络爬虫的链接去重方法、装置、设备及存储介质。该方法包括:在接收到待分析农产品的数据抓取请求时,从数据抓取请求中提取待访问平台的第一统一资源定位符URL链接;根据第一URL链接,向待访问平台发送访问请求;在接收到待访问平台根据访问请求作出的响应后,抓取第一URL链接对应的页面中的数据信息;对数据信息进行解析,得到页面中内嵌的第二URL链接,并将第二URL链接添加到待爬取URL队列;采用链接特征的计数布隆过滤器,并结合多重哈希对待爬取URL队列中的第二URL链接进行联合去重。本发明通过对链接去重方式的优化来改善网络爬虫的性能,从而保证网络爬虫能够快速的获取人们所需的信息,提升用户体验。



CN 110399546 B

1.一种基于网络爬虫的链接去重方法,其特征在于,所述方法包括以下步骤:

在接收到待分析农产品的数据抓取请求时,从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接;

根据第一URL链接,向所述待访问平台发送访问请求;

在接收到所述待访问平台根据所述访问请求作出的响应后,抓取所述第一URL链接对应的页面中的数据信息;

对所述数据信息进行解析,得到所述页面中内嵌的第二URL链接,并将所述第二URL链接添加到待爬取URL队列;

采用链接特征的计数布隆过滤器,并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重;

其中,所述采用链接特征的计数布隆过滤器,并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤,包括:

对所述待爬取URL队列进行遍历,获取遍历到的当前第二URL链接对应的整体特征URL链接;

采用链接特征的计数布隆过滤器对所述整体特征URL链接进行整体查重,得到所述整体特征URL链接对应的查重标志;

根据所述查重标志,对所述整体特征URL链接进行特征识别,得到多个特征片段;

根据预设的URL链接重组规则,对所述多个特征片段进行重组,得到N个重组URL链接片段,所述N为大于等于1的整数;

对N个重组URL链接片段进行多重哈希查重,得到所述当前第二URL链接对应的查重结果;

根据所述查重结果,对所述待爬取URL队列中的第二URL链接进行保留或丢弃操作。

2.如权利要求1所述的方法,其特征在于,所述采用链接特征的计数布隆过滤器,并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤之前,所述方法还包括:

对所述待爬取URL队列进行遍历,对遍历到的当前第二URL链接进行特征分析,提取所述当前第二URL链接的协议类型部分、路径部分和询问部分;

根据所述协议类型部分、所述路径部分和所述询问部分,得到所述当前第二URL链接对应的整体特征URL链接;

建立所述当前第二URL链接与所述整体特征URL链接之间的对应关系,并将所述对应关系更新到所述待爬取URL队列中。

3.如权利要求2所述的方法,其特征在于,所述根据预设的URL链接重组规则,对所述多个特征片段进行重组,得到N个重组URL链接片段的步骤之后,所述方法还包括:

基于MD5算法,对得到的N个重组URL链接片段分别进行压缩,得到N个重组URL链接片段对应的字符串密文;

将所述字符串密文替换掉对应的重组URL链接片段中的内容。

4.如权利要求3所述的方法,其特征在于,所述对N个重组URL链接片段进行多重哈希查重,得到所述当前第二URL链接对应的查重结果的步骤,包括:

提取N个重组URL链接片段对应的字符串密文,从N个字符串密文中选取任意一个字符

串密文进行K次哈希处理,得到K个哈希值,所述K为大于等于2的整数;

将K个哈希值散列到预先构建的位向量空间作为参考哈希值,并为每一个参考哈希值对应的空间可变计数器设置初始计数值;

分别对剩余N-1个字符串密文进行K次哈希处理,得到每一个剩余字符串密文对应的K个哈希值;

将每一个剩余字符串密文对应的K个哈希值随机散列到所述位向量空间,且与任意一个参考哈希值相邻;

采用头插法在相邻的参考哈希值对应的初始计数值前为每一个新散列到所述位向量空间的哈希值插入一位预设字符;

统计每一个参考哈希值对应的初始数值前预设字符的个数,根据所述预设字符的个数,确定所述当前第二URL链接对应的查重结果。

5.如权利要求1至4任一项所述的方法,其特征在于,所述采用链接特征的计数布隆过滤器,并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤之后,所述方法还包括:

基于MD5算法,对去重后的所述待爬取URL队列中的每一个第二URL链接进行压缩,得到每一个第二URL链接对应的字符串密文;

将所述字符串密文替换掉对应的第二URL链接中的内容。

6.如权利要求1至4任一项所述的方法,其特征在于,所述采用链接特征的计数布隆过滤器,并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤之后,所述方法还包括:

判断去重后的所述待爬取URL队列中是否存在已访问的第二URL链接;

若所述待爬取URL队列中存在已访问的第二URL链接,则将所述已访问的第二URL链接从所述待爬取URL队列中删除。

7.一种基于网络爬虫的链接去重装置,其特征在于,所述装置包括:

提取模块,用于在接收到待分析农产品的数据抓取请求时,从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接;

发送模块,用于根据第一URL链接,向所述待访问平台发送访问请求;

抓取模块,用于在接收到所述待访问平台根据所述访问请求作出的响应后,抓取所述第一URL链接对应的页面中的数据信息;

解析模块,用于对所述数据信息进行解析,得到所述页面中内嵌的第二URL链接,并将所述第二URL链接添加到待爬取URL队列;

去重模块,用于采用链接特征的计数布隆过滤器,并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重;

所述去重模块,还用于对所述待爬取URL队列进行遍历,获取遍历到的当前第二URL链接对应的整体特征URL链接;

所述去重模块,还用于采用链接特征的计数布隆过滤器对所述整体特征URL链接进行整体查重,得到所述整体特征URL链接对应的查重标志;

所述去重模块,还用于根据所述查重标志,对所述整体特征URL链接进行特征识别,得到多个特征片段;

所述去重模块，还用于根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段，所述N为大于等于1的整数；

所述去重模块，还用于对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果；

所述去重模块，还用于根据所述查重结果，对所述待爬取URL队列中的第二URL链接进行保留或丢弃操作。

8.一种基于网络爬虫的链接去重设备，其特征在于，所述设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的基于网络爬虫的链接去重程序，所述基于网络爬虫的链接去重程序配置为实现如权利要求1至6中任一项所述的基于网络爬虫的链接去重方法的步骤。

9.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有基于网络爬虫的链接去重程序，所述基于网络爬虫的链接去重程序被处理器执行时实现如权利要求1至6任一项所述的基于网络爬虫的链接去重方法的步骤。

基于网络爬虫的链接去重方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及互联网技术领域，尤其涉及一种基于网络爬虫的链接去重方法、装置、设备及存储介质。

背景技术

[0002] 网络爬虫在进行网页爬取时不可避免的会碰到网页的重复下载的情况，为了防止因网络爬虫重复爬取导致的效率下降，浪费服务器资源，因此需要对统一资源定位符(Uniform Resource Locator，URL)进行过滤去重。目前常见的链接去重方式有：基于第五代报文摘要算法(message-digest algorithm 5，MD5)的链接压缩去重、基于哈希算法的存储去重、基于布隆过滤器的链接去重等方式对链接进行去重。

[0003] 虽然，基于MD5的链接压缩去重方式解决了统一资源定位符(Uniform Resource Locator，URL)占用很大存储空间的问题。但是，随着URL越来越多，内存空间占用率也会越来越高，并且冲突几率低的特性会降低查重的准确率，因此会严重影响网络爬虫的性能。

[0004] 而基于哈希算法的存储去重方式虽然查重速度快且准确率较高，但需要设计一个好的哈希函数，并且需要维护哈希表。此外，随着抓取网页的规模增大，耗用内存会过高，因此也会严重影响网络爬虫的性能。

[0005] 而基于布隆过滤器的链接去重方式虽然可以解决空间复杂度的问题，但是有一定的误判，且不能删除已有元素。也就是说，元素越多，误报率会越大，因此也会严重影响网络爬虫的性能。

[0006] 因此，亟需提供一种基于网络爬虫的链接去重方式，以提升网络爬虫的性能，使得网络爬虫能够快速的获取人们所需的信息，进而提升用户体验。

[0007] 上述内容仅用于辅助理解本发明的技术方案，并不代表承认上述内容是现有技术。

发明内容

[0008] 本发明的主要目的在于提供一种基于网络爬虫的链接去重方法、装置、设备及存储介质，旨在通过对链接去重方式的优化来改善网络爬虫的性能，从而保证网络爬虫能够快速的获取人们所需的信息，提升用户体验。

[0009] 为实现上述目的，本发明提供了一种基于网络爬虫的链接去重方法，所述方法包括以下步骤：

[0010] 在接收到待分析农产品的数据抓取请求时，从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接；

[0011] 根据所述第一URL链接，向所述待访问平台发送访问请求；

[0012] 在接收到所述待访问平台根据所述访问请求作出的响应后，抓取所述第一URL链接对应的页面中的数据信息；

[0013] 对所述数据信息进行解析，得到所述页面中内嵌的第二URL链接，并将所述第二

URL链接添加到待爬取URL队列；

[0014] 采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重。

[0015] 优选地，所述采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤之前，所述方法还包括：

[0016] 对所述待爬取URL队列进行遍历，对遍历到的当前第二URL链接进行特征分析，提取所述当前第二URL链接的协议类型部分、路径部分和询问部分；

[0017] 根据所述协议类型部分、所述路径部分和所述询问部分，得到所述当前第二URL链接对应的整体特征URL链接；

[0018] 建立所述当前第二URL链接与所述整体特征URL链接之间的对应关系，并将所述对应关系更新到所述待爬取URL队列中。

[0019] 优选地，所述采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤，包括：

[0020] 对所述待爬取URL队列进行遍历，获取遍历到的当前第二URL链接对应的整体特征URL链接；

[0021] 采用链接特征的计数布隆过滤器对所述整体特征URL链接进行整体查重，得到所述整体特征URL链接对应的查重标志；

[0022] 根据所述查重标志，对所述整体特征URL链接进行特征识别，得到多个特征片段；

[0023] 根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段，所述N为大于等于1的整数；

[0024] 对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果；

[0025] 根据所述查重结果，对所述待爬取URL队列中的第二URL链接进行保留或丢弃操作。

[0026] 优选地，所述根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段的步骤之后，所述方法还包括：

[0027] 基于MD5算法，对得到的N个重组URL链接片段分别进行压缩，得到N个重组URL链接片段对应的字符串密文；

[0028] 将所述字符串密文替换掉对应的重组URL链接片段中的内容。

[0029] 优选地，所述对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果的步骤，包括：

[0030] 提取N个重组URL链接片段对应的字符串密文，从N个字符串密文中选取任意一个字符串密文进行K次哈希处理，得到K个哈希值，所述K为大于等于2的整数；

[0031] 将K个哈希值散列到预先构建的位向量空间作为参考哈希值，并为每一个参考哈希值对应的空间可变计数器设置初始计数值；

[0032] 分别对剩余N-1个字符串密文进行K次哈希处理，得到每一个剩余字符串密文对应的K个哈希值；

[0033] 将每一个剩余字符串密文对应的K个哈希值随机散列到所述位向量空间，且与任意一个参考哈希值相邻；

[0034] 采用头插法在相邻的参考哈希值对应的初始计数值前为每一个新散列到所述位向量空间的哈希值插入一位预设字符；

[0035] 统计每一个参考哈希值对应的初始数值前预设字符的个数，根据所述预设字符的个数，确定所述当前第二URL链接对应的查重结果。

[0036] 优选地，所述采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤之后，所述方法还包括：

[0037] 基于MD5算法，对去重后的所述待爬取URL队列中的每一个第二URL链接进行压缩，得到每一个第二URL链接对应的字符串密文；

[0038] 将所述字符串密文替换掉对应的第二URL链接中的内容。

[0039] 优选地，所述采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的步骤之后，所述方法还包括：

[0040] 判断去重后的所述待爬取URL队列中是否存在已访问的第二URL链接；

[0041] 若所述待爬取URL队列中存在已访问的第二URL链接，则将所述已访问的第二URL链接从所述待爬取URL队列中删除。

[0042] 此外，为实现上述目的，本发明还提出一种基于网络爬虫的链接去重装置，所述装置包括：

[0043] 提取模块，用于在接收到待分析农产品的数据抓取请求时，从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接；

[0044] 发送模块，用于根据所述第一URL链接，向所述待访问平台发送访问请求；

[0045] 抓取模块，用于在接收到所述待访问平台根据所述访问请求作出的响应后，抓取所述第一URL链接对应的页面中的数据信息；

[0046] 解析模块，用于对所述数据信息进行解析，得到所述页面中内嵌的第二URL链接，并将所述第二URL链接添加到待爬取URL队列；

[0047] 去重模块，用于采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重。

[0048] 此外，为实现上述目的，本发明还提出一种基于网络爬虫的链接去重设备，所述设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的基于网络爬虫的链接去重程序，所述基于网络爬虫的链接去重程序配置为实现如上文所述的基于网络爬虫的链接去重方法的步骤。

[0049] 此外，为实现上述目的，本发明还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有基于网络爬虫的链接去重程序，所述基于网络爬虫的链接去重程序被处理器执行时实现如上文所述的基于网络爬虫的链接去重方法的步骤。

[0050] 本发明提供的基于网络爬虫的链接去重方案，通过采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中缓存的第二URL链接进行联合去重，尽可能的降低了计数布隆过滤器的误判率，显著的改善了网络爬虫的性能，从而保证网络爬虫能够快速的获取人们所需的信息，提升用户体验。

附图说明

[0051] 图1是本发明实施例方案涉及的硬件运行环境的基于网络爬虫的链接去重设备的

结构示意图；

[0052] 图2为本发明基于网络爬虫的链接去重方法第一实施例的流程示意图；

[0053] 图3为本发明基于网络爬虫的链接去重方法第二实施例的流程示意图；

[0054] 图4为本发明基于网络爬虫的链接去重装置第一实施例的结构框图。

[0055] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0056] 应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0057] 参照图1，图1为本发明实施例方案涉及的硬件运行环境的基于网络爬虫的链接去重设备结构示意图。

[0058] 如图1所示，该基于网络爬虫的链接去重设备可以包括：处理器1001，例如中央处理器(Central Processing Unit，CPU)，通信总线1002、用户接口1003，网络接口1004，存储器1005。其中，通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏(Display)、输入单元比如键盘(Keyboard)，可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如无线保真(WIreless-FIdelity，WI-FI)接口)。存储器1005可以是高速的随机存取存储器(Random Access Memory，RAM)存储器，也可以是稳定的非易失性存储器(Non-Volatile Memory，NVM)，例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

[0059] 本领域技术人员可以理解，图1中示出的结构并不构成对基于网络爬虫的链接去重设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

[0060] 如图1所示，作为一种存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及基于网络爬虫的链接去重程序。

[0061] 在图1所示的基于网络爬虫的链接去重设备中，网络接口1004主要用于与网络服务器进行数据通信；用户接口1003主要用于与用户进行数据交互；本发明基于网络爬虫的链接去重设备中的处理器1001、存储器1005可以设置在基于网络爬虫的链接去重设备中，所述基于网络爬虫的链接去重设备通过处理器1001调用存储器1005中存储的基于网络爬虫的链接去重程序，并执行本发明实施例提供的基于网络爬虫的链接去重方法。

[0062] 本发明实施例提供了一种基于网络爬虫的链接去重方法，参照图2，图2为本发明一种基于网络爬虫的链接去重方法第一实施例的流程示意图。

[0063] 本实施例中，所述基于网络爬虫的链接去重方法包括以下步骤：

[0064] 步骤S10，在接收到待分析农产品的数据抓取请求时，从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接。

[0065] 具体的说，本实施例的执行主体为任意部署或安装有网络爬虫系统的终端设备。

[0066] 值得一提的是，在本实施例中，为了尽可能提高待分析农产品对应的数据的抓取速度、解析速度等操作，本实施例中所说的网络爬虫系统优选分布式网络爬虫系统。

[0067] 此外，应当理解的是，在实际应用中所述终端设备可以是客户端设备，也可以是服务器端设备，此处不做限制。

[0068] 此外，上述所说的待访问平台在实际应用中可以是展示有待分析农产品的网络商

城。

[0069] 相应地,所说统一资源定位符(Uniform Resource Locator,URL)即为访问所述网络商城所需的网络地址。

[0070] 此外,应当理解的是,上述所说的待分析农产品只是对目前常见的各种农产品的一个统称,在实际应用中待分析农产品可以是茶产品、果蔬产品、粮食产品等等,此处不再一一列举,对此也不做任何限制。

[0071] 步骤S20,根据所述第一URL链接,向所述待访问平台发送访问请求。

[0072] 具体的说,在实际应用中,网络爬虫可以采用基于传输控制协议/因特网互联协议(Transmission Control Protocol/Internet Protocol,TCP/IP协议)来传输数据的超文本传输协议(HyperText Transfer Protocol,HTTP)向所述待访问平台(实质为该平台的服务器)发送访问请求。

[0073] 应当理解的是,以上给出的仅为一种向所述待访问平台发送访问请求的具体实现方式,对本发明的技术方案并不构成任何限定,在实际应用中,本领域的技术人员可以根据需要进行设置,此处不做限制。

[0074] 步骤S30,在接收到所述待访问平台根据所述访问请求作出的响应后,抓取所述第一URL链接对应的页面中的数据信息。

[0075] 应当理解的是,在实际应用中,如果向所述待访问平台发送的访问请求成功,并且所述待访问平台对所述访问请求中携带的第一URL链接验证成功后,并会作出成功的响应,并反馈所述第一URL链接对应的页面中的数据信息。此时,网络爬虫并可以抓取所述待访问平台反馈的针对所述第一URL链接对应的页面中的数据信息。

[0076] 步骤S40,对所述数据信息进行解析,得到所述页面中内嵌的第二URL链接,并将所述第二URL链接添加到待爬取URL队列。

[0077] 应当理解的是,在实际应用中,第一URL链接对应的页面中除了会显示与所述待分析农产品相同的数据信息,还可能会显示多个与所述数据信息相关的URL链接,为了便于区分此处称为第二URL链接。

[0078] 比如说,在第一URL链接对应的页面中显示的是包括所述待分析农产品的一个网络商城主页,在该主页中主要显示有农产品A、农产品B、农产品C以及农产品D等四大类农产品信息,同时每一大类农产品又对应有一个第二URL链接,该第二URL链接对应的页面中主要显示有对应农产品包括的小类农产品。

[0079] 比如,农产品A对应的第二URL链接对应的页面中主要显示有农产品A-1、农产品A-2和农产品A-3;农产品B对应的第二URL链接对应的页面中主要显示有农产品B-1和农产品B-2;农产品C对应的第二URL链接对应的页面中主要显示有农产品C-1、农产品C-2、农产品C-3和农产品C4;农产品D对应的第二URL链接对应的页面中主要显示有农产品D-1和农产品D-2。

[0080] 应当理解的是,以上仅为举例说明,对本发明的技术方案并不构成任何限定,在实际应用中,本领域的技术人员可以根据需要进行设置,此处不做限制。

[0081] 此外,在本实施例中,之所以要将所述页面中内嵌的第二URL链接添加到待爬取URL队列,是因为在实际应用中网络爬虫爬取的数据较多,因而解析出来的第二URL链接数量相对庞大。而每爬取、解析一个第二URL链接均会消耗不少时间,因而大量的第二URL链接

往往不能短时间内访问完，故需要将每次获取到的第二URL链接添加到待爬取URL队列中。

[0082] 此外，上述所说的“第一URL链接”中的“第一”，以及“第二URL链接”中的“第二”仅仅是用于区别待访问平台对应的URL链接与该URL链接对应的页面中内嵌的URL链接，并不对URL链接本身造成限定。在实际应用中，任意一个“第二URL链接”相对于其对应的页面中内嵌的URL链接都可以看作是一个“第一URL链接”。

[0083] 步骤S50，采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重。

[0084] 具体的说，上述所说的采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行的联合去重，主要分为对所述URL链接对应的整体特征URL链接去重和对URL链接片段去重。

[0085] 而URL链接片段则是根据整体特征URL链接得到的，因而为了保证上述联合去重的操作能够顺利进行，需要先确定第二URL链接与整体特征URL链接之间的对应关系。

[0086] 为了便于理解，本实施例给出一种确定第二URL链接与整体特征URL链接之间对应关系的具体实现方式，大致如下：

[0087] (1)对所述待爬取URL队列进行遍历，对遍历到的当前第二URL链接进行特征分析，提取所述当前第二URL链接的协议类型部分、路径部分和询问部分。

[0088] 具体的说，由于在实际应用中URL链接是用于唯一标识网络上的资源的。并且，一般来说，一个URL链接通常会包含如下五个组成部分：协议类型部分(通常用Protocol表示)、服务器地址部分(通常用户Host表示)、端口号部分(通常用Port表示)、路径部分(通常用Path表示)和询问部分(通常用Fragment表示)。

[0089] 其中，协议类型部分、路径部分和询问部分这三个部分通常就可以体现一个URL链接的特征。

[0090] 因而，本实施例通过对所述待爬取URL队列进行遍历，并对遍历到的当前第二URL链接进行特征分析，进而提取出当前第二URL链接的协议类型部分(为了便于后续说明以下用户p_1表示)、路径部分(为了便于后续说明以下用户p_2表示)和询问部分(为了便于后续说明以下用户p_3表示)。

[0091] (2)根据所述协议类型部分、所述路径部分和所述询问部分，得到所述当前第二URL链接对应的整体特征URL链接。

[0092] 具体的说，由于p_1、p_2和p_3这三部分就可以体现当前第二URL链接的全部特征，因而通过对p_1、p_2和p_3进行组合便可以得到当前第二URL链接对应的整体特征URL链接，以下用$p_1p_2p_3$表示每个第二URL链接对应的整体特征URL链接。

[0093] (3)建立所述当前第二URL链接与所述整体特征URL链接之间的对应关系，并将所述对应关系更新到所述待爬取URL队列中。

[0094] 具体的说，本实施例中之所以要建立所述当前第二URL链接与所述整体特征URL链接之间的对应关系，并将所述对应关系更新到所述待爬取URL队列中是为了方便后续对第二URL链接去重过程中，能够该对应关系快速找到当前第二URL链接对应的整体特征URL链接，进而根据整体URL链接得到当前第二URL链接对应的URL链接片段。

[0095] 此外，在实际应用中，也可以不把所述对应关系更新到所述待爬取URL队列中，而是单独存放。当对待爬取URL队列中的第二URL链接进行联合去重时，根据遍历到的当前第

二URL链接从单独存放的对应关系表中查找当前第二URL链接对应的整体特征URL链接即可。

[0096] 应当理解的是，以上仅为举例说明，对本发明的技术方案并不构成任何限定，在实际应用中，本领域的技术人员可以根据需要进行设置，此处不做限制。

[0097] 进一步地，在得到上述对应关系以及每个第二URL链接对应的整体特征URL链接之后，上述所说的采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的操作，具体可以如下所述：

[0098] (1)对所述待爬取URL队列进行遍历，获取遍历到的当前第二URL链接对应的整体特征URL链接。

[0099] 具体的说，获取遍历到的当前第二URL链接对应的整体特征URL链接即为根据上述所说的对应关系获取。

[0100] (2)采用链接特征的计数布隆过滤器对所述整体特征URL链接进行整体查重，得到所述整体特征URL链接对应的查重标志。

[0101] 具体的说，本实施例中所采用的计数布隆过滤器并非现有进行链接去重时采用的计数布隆器，而是基于URL链接的链接特征的计数布隆过滤器。

[0102] 也就是说，本实施例的计算布隆过滤器在对链接进行去重时，具体是通过对待爬取URL队列中每一个第二URL链接对应的整体特征URL链接进行特征识别，然后根据识别到的特征进行整体查重，即在去重时是对每个第二入了链接进行特征对比，进而实现整体查重。

[0103] 并且，为了方便识别后续根据特征片段重组后的URL链接片段，还会为整体特征URL链接分配对应的查重标志。

[0104] (3)根据所述查重标志，对所述整体特征URL链接进行特征识别，得到多个特征片段。

[0105] 具体的说，仍以整体特征URL链接为$p_1p_2p_3$为例，通过对所述整体特征URL链接进行特征识别后，得到的多个特征片段具体可以是分别包括协议类型部分、路径部分和询问部分的片段，即对特征片段p_1、特征片段p_2和特征片段p_3。

[0106] (4)根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段。

[0107] 应当理解的是，由于一个整体特征URL链接是由协议类型部分、路径部分和询问三部分组成的，因而至少会得到1个重组URL链接片段，故在本实施例中N为大于等于1的整数。

[0108] 此外，在实际应用总，所述URL链接重组规则可以由本领域的技术人员根据需要进行设置，比如规定重组后的URL链接片段必须包括特征片段p_1，或者重组后的URL链接片段不能包括特征片段p_3等，此处不再一一列举，对此也不做任何限制。

[0109] 相应地，如果URL链接重组规则为重组后的URL链接片段必须包括特征片段p_1，则得到的重组URL链接片段大致包括仅包括p_1特征片段的URL链接片段、仅包括p_1特征片段和p_2特征片段的URL链接片段，以及仅包括p_1特征片段和p_3特征片段的URL链接片段。

[0110] 如果URL链接重组规则为重组后的URL链接片段不能包括特征片段p_3，则得到的重组URL链接片段大致包括仅包括p_1特征片段的URL链接片段和仅包括p_1特征片段和p_2特征片段的URL链接片段。

[0111] 应当理解的是，以上仅为举例说明，对本发明的技术方案并不构成任何限定，在实际应用中，本领域的技术人员可以根据实际需要进行设置，此处不做限制。

[0112] (5)对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果。

[0113] 值得一提的是，由于在实际应用中，缓存在待爬取URL队列中的第二URL链接可能有大量，因而重组后得到的URL链接片段会更加多。因此，在本实施例中，为了尽可能降低对待爬取URL队列中缓存的第二URL链接对存储空间的占用，在根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段之后，可以先基于MD5算法，对得到的N个重组URL链接片段分别进行压缩，进而得到N个重组URL链接片段对应的字符串密文，最终将所述字符串密文替换掉对应的重组URL链接片段中的内容。

[0114] 应当理解的是，以上给出的仅为一种具体的压缩方式，对本发明的技术方案并不构成任何限定，在实际应用中，本领域的技术人员可以根据实际需要选取合适的压缩方法，此处不做限制。

[0115] 相应地，上述对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果的操作，具体为：

[0116] (5-1)提取N个重组URL链接片段对应的字符串密文，从N个字符串密文中选取任意一个字符串密文进行K次哈希处理，得到K个哈希值。

[0117] 应当理解的是，由于本实施例提供的基于网络爬虫的链接去重方案，在对链接进行联合去重时具体结合的是多重哈希，即对一个字符串密文至少需要进行2次哈希处理，故上述所说的K为大于等于2的整数。

[0118] (5-2)将K个哈希值散列到预先构建的位向量空间作为参考哈希值，并为每一个参考哈希值对应的空间可变计数器设置初始计数值。

[0119] 具体的说，在本实施例中每个参考哈希值对应的空间可变计数器上显示的初始计数值用“0”表示。

[0120] (5-3)分别对剩余N-1个字符串密文进行K次哈希处理，得到每一个剩余字符串密文对应的K个哈希值。

[0121] (5-4)将每一个剩余字符串密文对应的K个哈希值随机散列到所述位向量空间，且与任意一个参考哈希值相邻。

[0122] 具体的说，为了便于确定新散列到所述位向量空间中的哈希值究竟与那一个参考哈希值相邻，可以预先设置一个确定标准，比如在相邻两个参考哈希值之间插入新的哈希值时，可以选取距离新插入的哈希值最近的参考哈希值作为相邻的参考哈希值。

[0123] 应当理解的是，以上仅为举例说明，对本发明的技术方案并不构成任何限定，在实际应用中，本领域的技术人员可以根据实际需要进行设置，此处不做限制。

[0124] (5-5)采用头插法在相邻的参考哈希值对应的初始计数值前为每一个新散列到所述位向量空间的哈希值插入一位预设字符。

[0125] 具体的说，在本实施例中所述预设字符选用“1”表示。

[0126] 比如说，对于一个参考哈希值，其对应的空间可变计数器上显示的初始计数值为“0”。当有一个新的哈希值散列到与其相邻的位置时，就需要采用头插法在“0”的前面插入一位预设字符“1”，此时空间可变计数器上显示的计数值变为“10”。

[0127] 相应地，如果有两个新的哈希值散列到该参考哈希值的想了位置，则需要采用头插法在“0”的前面插入两位预设字符“1”，此时空间可变计数器上显示的计数值变为“110”。

[0128] (5-6)统计每一个参考哈希值对应的初始数值前预设字符的个数，根据所述预设字符的个数，确定所述当前第二URL链接对应的查重结果。

[0129] 具体的说，确定的查重结果可以为：

[0130] 若初始计数值“0”前面的预设字符“1”的个数大于1，则确定所述重组URL片段重复，需要丢弃；

[0131] 否则，确定所述重组URL片段不重复，可以保留。

[0132] (6)根据所述查重结果，对所述待爬取URL队列中的第二URL链接进行保留或丢弃操作。

[0133] 应当理解的是，以上给出的仅为一种联合去重的具体实现方式，对本发明的技术方案并不构成任何限定，在实际应用中，本领域的技术人员可以根据需要合理调整，此处不做限制。

[0134] 此外，在实际应用中，为了进一步地降低对存储空间的占用，在采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重之后，还可以基于MD5算法，对去重后的所述待爬取URL队列中的每一个第二URL链接进行压缩，进而得到每一个第二URL链接对应的字符串密文；最后将所述字符串密文替换掉对应的第二URL链接中的内容，从而尽可能的压缩待爬取URL队列中的第二URL链接，降低对存储空间的占用。

[0135] 通过上述描述不难看出，本实施例提供的基于网络爬虫的链接去重方法，通过采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中缓存的第二URL链接进行整体和部分的联合去重，从而尽可能的降低了计数布隆过滤器的误判率，有效改善了网络爬虫的性能，使得网络爬虫能够快速的获取人们所需的信息，尽可能的提升了用户体验。

[0136] 此外，在去重过程中，通过基于压缩算法，如MD5算法对URL链接进行压缩，从而尽可能的降低了对存储空间的占用。

[0137] 参考图3，图3为本发明一种基于网络爬虫的链接去重方法第二实施例的流程示意图。

[0138] 基于上述第一实施例，本实施例基于网络爬虫的链接去重方法在所述步骤S50之后，还包括：

[0139] 步骤S60，判断去重后的所述待爬取URL队列中是否存在已访问的第二URL链接。

[0140] 具体的说，若通过判断，确定去重后的所述待爬取URL队列中存在已访问的第二URL链接，即网络爬虫已经根据该第二URL链接访问了该第二URL链接对应的页面，并抓取了该页面中的数据信息，为了避免网络爬虫再次访问该第二URL链接，导致重复抓取相同数据，浪费网络爬虫的资源，需要执行步骤S70的操作；否则，继续执行步骤S60。

[0141] 步骤S70，将所述已访问的第二URL链接从所述待爬取URL队列中删除。

[0142] 具体的说，在实际应用中，可以再检测到有一个第二URL链接被访问的情况下就执行一次删除操作，也可以先对已被访问的第二URL链接进行标记，然后再标记的被访问第二URL链接达到预定数量，或者预定删除时间时，将当前被标记的所有第二URL链接一起删除。

[0143] 应当理解的是，以上仅为举例说明，对本发明的技术方案并不构成任何限定，在实际应用中，本领域的技术人员可以根据需要进行设置，此处不做限制。

[0144] 通过上述描述不难看出，本实施例提供的基于网络爬虫的链接去重方法，通过定时或实时检测所述待爬取URL队列中第二URL链接的访问情况，并在检测到所述待爬取URL队列中存在已经被访问的第二URL链接时，将被访问的第二URL链接从待爬取URL队列中删除，从而可以保证所述待爬取URL队列中缓存的第二URL链接均为未被访问的第二URL链接，避免了网络爬虫根据同一个第二URL链接重复进行相同数据的爬取，进一步提升了网络爬虫的性能。

[0145] 此外，本发明实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有基于网络爬虫的链接去重程序，所述基于网络爬虫的链接去重程序被处理器执行时实现如上文所述的基于网络爬虫的链接去重方法的步骤。

[0146] 参照图4，图4为本发明基于网络爬虫的链接去重装置第一实施例的结构框图。

[0147] 如图4所示，本发明实施例提出的基于网络爬虫的链接去重装置包括：提取模块4001、发送模块4002、抓取模块4003、解析模块4004和去重模块4005。

[0148] 其中，提取模块4001，用于在接收到待分析农产品的数据抓取请求时，从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接；发送模块4002，用于根据所述第一URL链接，向所述待访问平台发送访问请求；抓取模块4003，用于在接收到所述待访问平台根据所述访问请求作出的响应后，抓取所述第一URL链接对应的页面中的数据信息；解析模块4004，用于对所述数据信息进行解析，得到所述页面中内嵌的第二URL链接，并将所述第二URL链接添加到待爬取URL队列；去重模块4005，用于采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重。

[0149] 需要说明的是，本实施例中所涉及到的各模块均为逻辑模块，在实际应用中，一个逻辑单元可以是一个物理单元，也可以是一个物理单元的一部分，还可以以多个物理单元的组合实现。此外，为了突出本发明的创新部分，本实施例中并没有将与解决本发明所提出的技术问题关系不太密切的单元引入，但这并不表明本实施方式中不存在其它的单元。

[0150] 此外，值得一提的是，本实施例中去重模块4005在采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重时，具体分为对所述URL链接对应的整体特征URL链接去重和对URL链接片段去重。

[0151] 而URL链接片段则是根据整体特征URL链接得到的，因而为了保证去重模块4005能够顺利执行上述操作，需要先确定第二URL链接与整体特征URL链接之间的对应关系。

[0152] 关于确定第二URL链接与整体特征URL链接之间对应关系的方式，大致可以如下所述：

[0153] 首先，对所述待爬取URL队列进行遍历，对遍历到的当前第二URL链接进行特征分析，提取所述当前第二URL链接的协议类型部分、路径部分和询问部分；

[0154] 然后，根据所述协议类型部分、所述路径部分和所述询问部分，得到所述当前第二URL链接对应的整体特征URL链接；

[0155] 最后，建立所述当前第二URL链接与所述整体特征URL链接之间的对应关系，并将所述对应关系更新到所述待爬取URL队列中。

[0156] 相应地，在得到上述对应关系之后，所述去重模块4005执行的操作，具体为：

[0157] 首先，对所述待爬取URL队列进行遍历，获取遍历到的当前第二URL链接对应的整体特征URL链接；

[0158] 然后，采用链接特征的计数布隆过滤器对所述整体特征URL链接进行整体查重，得到所述整体特征URL链接对应的查重标志；

[0159] 接着，根据所述查重标志，对所述整体特征URL链接进行特征识别，得到多个特征片段；

[0160] 接着，根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段；

[0161] 接着，对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果；

[0162] 最后，根据所述查重结果，对所述待爬取URL队列中的第二URL链接进行保留或丢弃操作。

[0163] 需要说明的是，在本实施例中，上述所说的N为大于等于1的整数。

[0164] 此外，应当理解的是，以上给出的仅为一种确定第二URL链接与整体特征URL链接之间对应关系，以及采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重的具体实现方式，对本发明的技术方案并不构成任何限定，在具体应用中，本领域的技术人员可以根据需要进行设置，本发明对此不做限制。

[0165] 进一步地，在实际应用中，为了尽可能降低对待爬取URL队列中缓存的第二URL链接对存储空间的占用，在根据预设的URL链接重组规则，对所述多个特征片段进行重组，得到N个重组URL链接片段之后，可以先基于MD5算法，对得到的N个重组URL链接片段分别进行压缩，进而得到N个重组URL链接片段对应的字符串密文，最终将所述字符串密文替换掉对应的重组URL链接片段中的内容。

[0166] 相应地，所述对N个重组URL链接片段进行多重哈希查重，得到所述当前第二URL链接对应的查重结果的操作，具体为：

[0167] 首先，提取N个重组URL链接片段对应的字符串密文，从N个字符串密文中选取任意一个字符串密文进行K次哈希处理，得到K个哈希值；

[0168] 然后，将K个哈希值散列到预先构建的位向量空间作为参考哈希值，并为每一个参考哈希值对应的空间可变计数器设置初始计数值；

[0169] 接着，分别对剩余N-1个字符串密文进行K次哈希处理，得到每一个剩余字符串密文对应的K个哈希值；

[0170] 接着，将每一个剩余字符串密文对应的K个哈希值随机散列到所述位向量空间，且与任意一个参考哈希值相邻；

[0171] 接着，采用头插法在相邻的参考哈希值对应的初始计数值前为每一个新散列到所述位向量空间的哈希值插入一位预设字符；

[0172] 最后，统计每一个参考哈希值对应的初始数值前预设字符的个数，根据所述预设字符的个数，确定所述当前第二URL链接对应的查重结果。

[0173] 需要说明的是，在本实施例中，上述所说的K为大于等于2的整数。

[0174] 此外，应当理解的是，以上给出的仅为一种获取当前第二URL链接对应的查重结果的具体实现方式，对本发明的技术方案并不构成任何限定，在具体应用中，本领域的技术人

员可以根据需要进行设置，本发明对此不做限制。

[0175] 此外，在实际应用中，为了进一步地降低对存储空间的占用，在对所述待爬取URL队列中的第二URL链接进行联合去重之后，还可以基于MD5算法，对去重后的所述待爬取URL队列中的每一个第二URL链接进行压缩，进而得到每一个第二URL链接对应的字符串密文；最后将所述字符串密文替换掉对应的第二URL链接中的内容，从而尽可能的压缩待爬取URL队列中的第二URL链接，降低对存储空间的占用。

[0176] 通过上述描述不难看出，本实施例提供的基于网络爬虫的链接去重装置，通过采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中缓存的第二URL链接进行整体和部分的联合去重，从而尽可能的降低了计数布隆过滤器的误判率，有效改善了网络爬虫的性能，使得网络爬虫能够快速的获取人们所需的信息，尽可能的提升了用户体验。

[0177] 此外，在去重过程中，通过基于压缩算法，如MD5算法对URL链接进行压缩，从而尽可能的降低了对存储空间的占用。

[0178] 需要说明的是，以上所描述的工作流程仅仅是示意性的，并不对本发明的保护范围构成限定，在实际应用中，本领域的技术人员可以根据实际的需要选择其中的部分或者全部来实现本实施例方案的目的，此处不做限制。

[0179] 另外，未在本实施例中详尽描述的技术细节，可参见本发明任意实施例所提供的基于网络爬虫的链接去重方法，此处不再赘述。

[0180] 基于上述基于网络爬虫的链接去重装置的第一实施例，提出本发明基于网络爬虫的链接去重装置第二实施例。

[0181] 在本实施例中，所述基于网络爬虫的链接去重装置还包括删除模块。

[0182] 具体的说，所述删除模块，用于判断去重后的所述待爬取URL队列中是否存在已访问的第二URL链接。

[0183] 相应地，若所述待爬取URL队列中存在已访问的第二URL链接，则将所述已访问的第二URL链接从所述待爬取URL队列中删除；否则，继续监测所述待爬取URL队列中的第二URL链接，并判断是否存在已访问的第二URL链接。

[0184] 需要说明的是，本实施例中所涉及到的各模块均为逻辑模块，在实际应用中，一个逻辑单元可以是一个物理单元，也可以是一个物理单元的一部分，还可以以多个物理单元的组合实现。此外，为了突出本发明的创新部分，本实施例中并没有将与解决本发明所提出的技术问题关系不太密切的单元引入，但这并不表明本实施方式中不存在其它的单元。

[0185] 此外，应当理解的是，以上仅为举例说明，对本发明的技术方案并不构成任何限定，在具体应用中，本领域的技术人员可以根据需要进行设置，本发明对此不做限制。

[0186] 通过上述描述不难看出，本实施例提供的基于网络爬虫的链接去重装置，通过定时或实时检测所述待爬取URL队列中第二URL链接的访问情况，并在检测到所述待爬取URL队列中存在已经被访问的第二URL链接时，将被访问的第二URL链接从待爬取URL队列中删除，从而可以保证所述待爬取URL队列中缓存的第二URL链接均为未被访问的第二URL链接，避免了网络爬虫根据同一个第二URL链接重复进行相同数据的爬取，进一步提升了网络爬虫的性能。

[0187] 需要说明的是，以上所描述的工作流程仅仅是示意性的，并不对本发明的保护范

围构成限定，在实际应用中，本领域的技术人员可以根据实际的需要选择其中的部分或者全部来实现本实施例方案的目的，此处不做限制。

[0188] 另外，未在本实施例中详尽描述的技术细节，可参见本发明任意实施例所提供的基于网络爬虫的链接去重方法，此处不再赘述。

[0189] 此外，需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

[0190] 上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

[0191] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如只读存储器(Read Only Memory，ROM)/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本发明各个实施例所述的方法。

[0192] 以上仅为本发明的优选实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

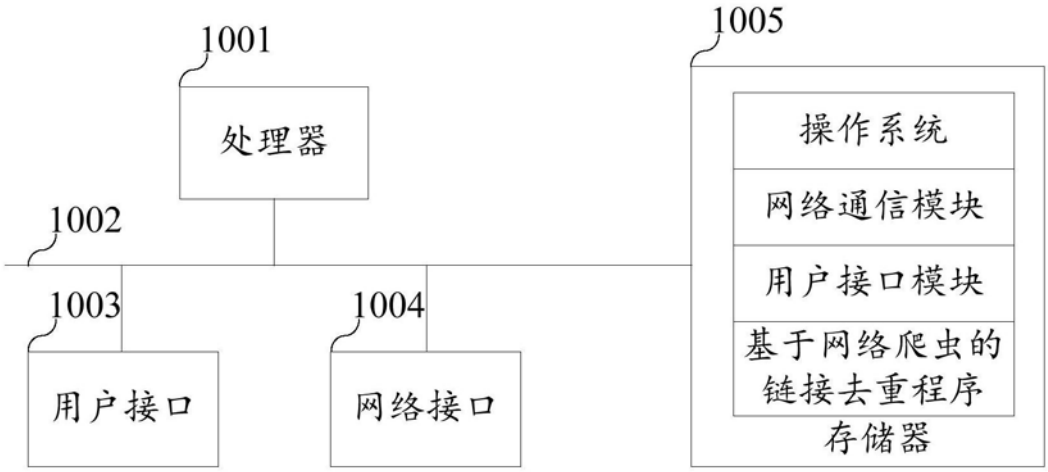

1001
处理器
1002
1003
用户接口
1004
网络接口
1005
操作系统
网络通信模块
用户接口模块
基于网络爬虫的链接去重程序
存储器

图1

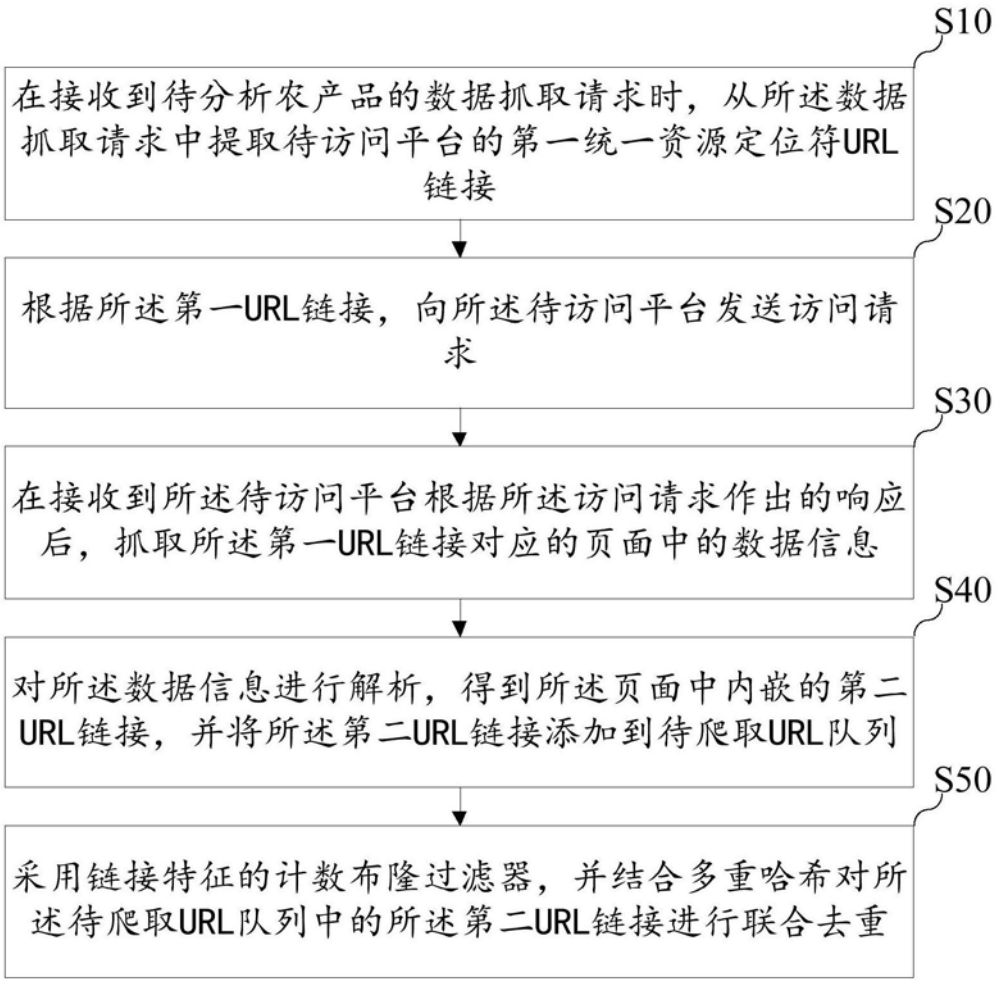

S10
在接收到待分析农产品的数据抓取请求时，从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接
S20
根据所述第一URL链接，向所述待访问平台发送访问请求
S30
在接收到所述待访问平台根据所述访问请求作出的响应后，抓取所述第一URL链接对应的页面中的数据信息
S40
对所述数据信息进行解析，得到所述页面中内嵌的第二URL链接，并将所述第二URL链接添加到待爬取URL队列
S50
采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重

图2

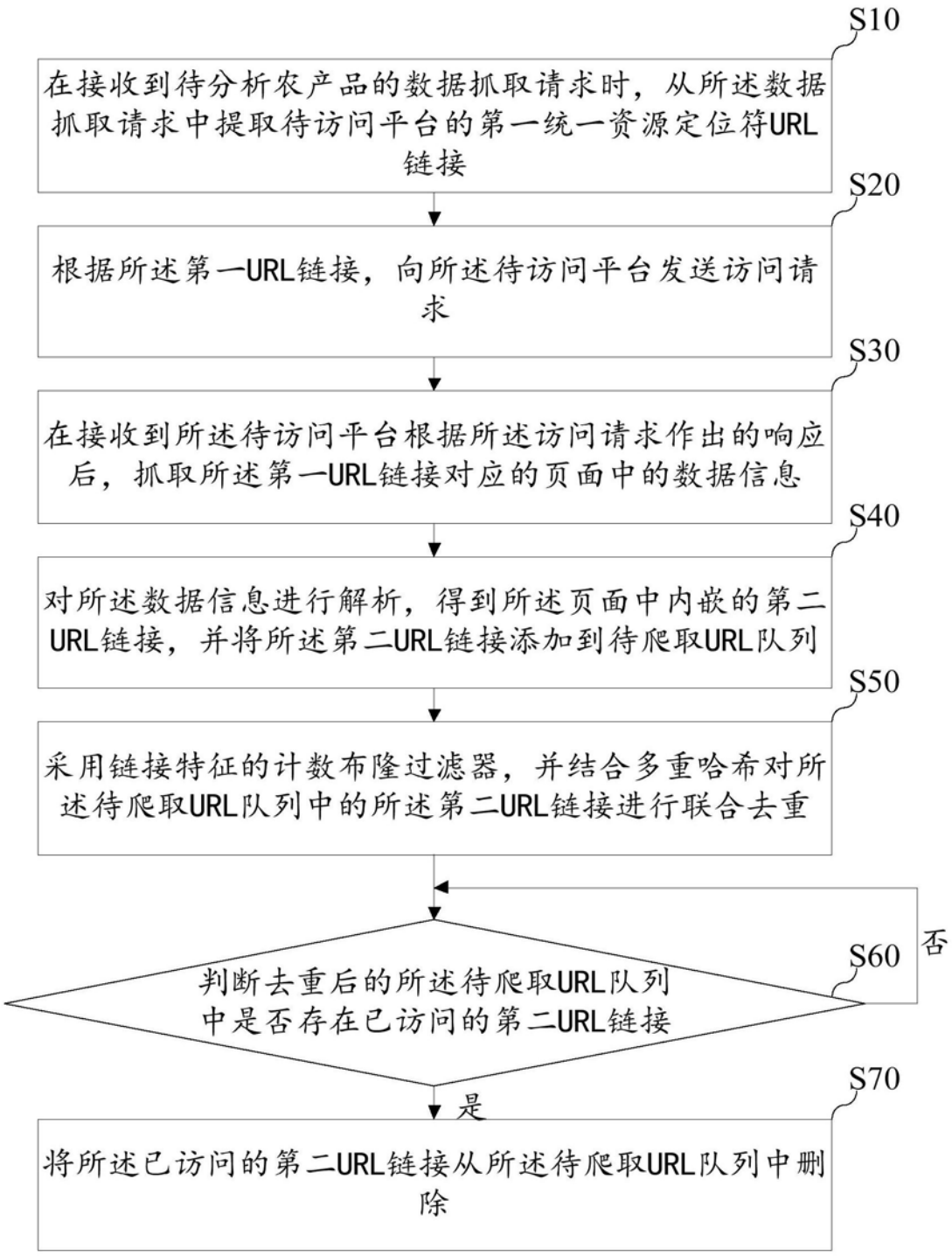

S10
在接收到待分析农产品的数据抓取请求时，从所述数据抓取请求中提取待访问平台的第一统一资源定位符URL链接
S20
根据所述第一URL链接，向所述待访问平台发送访问请求
S30
在接收到所述待访问平台根据所述访问请求作出的响应后，抓取所述第一URL链接对应的页面中的数据信息
S40
对所述数据信息进行解析，得到所述页面中内嵌的第二URL链接，并将所述第二URL链接添加到待爬取URL队列
S50
采用链接特征的计数布隆过滤器，并结合多重哈希对所述待爬取URL队列中的所述第二URL链接进行联合去重
S60
判断去重后的所述待爬取URL队列中是否存在已访问的第二URL链接
否
是
S70
将所述已访问的第二URL链接从所述待爬取URL队列中删除

图3

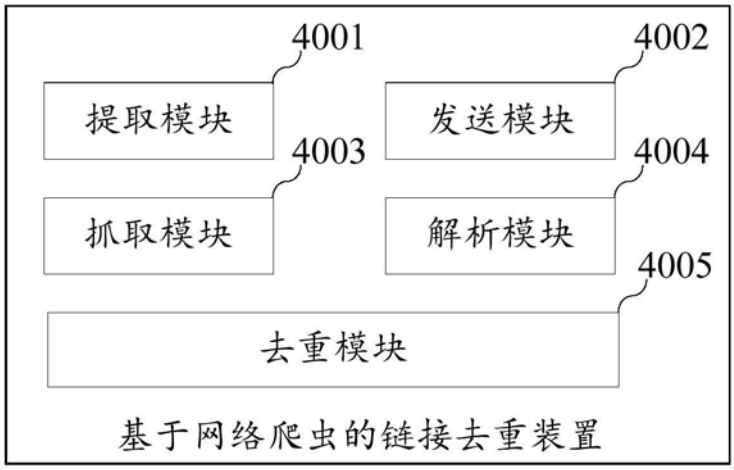

4001
提取模块
4002
发送模块
4003
抓取模块
4004
解析模块
4005
去重模块
基于网络爬虫的链接去重装置

图4