

网页数据处理方法、装置、计算机设备及计算机存储介质

本申请要求于 2017 年 12 月 30 日提交中国专利局、申请号为 201711487763.3、发明名称为“网页数据处理方法、装置、计算机设备及计算机存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及网络安全领域，特别是涉及一种网页数据处理方法、装置、计算机设备及存储介质。

背景技术

随着互联网科技的发展，在生活中，用户可以从网络上获取到越来越多的信息，因此，
10 有时会出现一些相关的高危漏洞或者于高危漏洞相关的高危信息，则从网页上获取到高危漏洞或者高危漏洞的相关信息是非常重要的。

传统地，是从当前已知的网页上查询相应的网页数据，从而对网页数据进行分析得到有关高危漏洞或者高危漏洞的相关信息，但是，仅从当前网页查询相应的网页数据会导致大量的网页数据出现遗漏，导致对网页数据分析不准确。

15

发明内容

根据本申请的各种实施例，提供一种网页数据处理方法、装置、计算机设备及计算机存储介质，解决了背景技术中所涉及的一个或多个问题。

一种网页数据处理方法，所述方法包括：

获取第一网页的第一网页数据，查询与所述第一网页数据关联的第二网页地址；

20 从所述第二网页地址中获取所述第二网页对应的网站的域名，提取所述第二网页对应的网站的域名的后缀；

当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址；

25 根据所述第二网页的网络地址对所述第二网页进行访问，并爬取所述第二网页上的第二网页数据；及

将所述第一网页数据与所述第二网页数据分别输出至对应的类别。

一种网页数据处理装置，所述装置包括：

查询模块，用于获取第一网页的第一网页数据，查询与所述第一网页数据关联的第二网页地址；

30 提取模块，用于从所述第二网页地址中获取所述第二网页对应的网站的域名，提取所述第二网页对应的网站的域名的后缀；

获取模块,用于当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时,则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址;

爬取模块,用于根据所述第二网页的网络地址对所述第二网页进行访问,并爬取所述第二网页上的第二网页数据;

5 输出模块,用于将所述第一网页数据与所述第二网页数据分别输出至对应的类别。

一种计算机设备,包括存储器和处理器,所述存储器存储有计算机可读指令,其特征在于,所述处理器执行所述计算机可读指令时实现以下步骤:

获取第一网页的第一网页数据,查询与所述第一网页数据关联的第二网页地址;

10 从所述第二网页地址中获取所述第二网页对应的网站的域名,提取所述第二网页对应的网站的域名的后缀;

当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时,则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址;

根据所述第二网页的网络地址对所述第二网页进行访问,并爬取所述第二网页上的第二网页数据;及

15 将所述第一网页数据与所述第二网页数据分别输出至对应的类别。

一个或多个存储有计算机可读指令的非易失性计算机可读存储介质,所述计算机可读指令被一个或多个处理器执行时,使得一个或多个处理器执行以下步骤:

获取第一网页的第一网页数据,查询与所述第一网页数据关联的第二网页地址;

20 从所述第二网页地址中获取所述第二网页对应的网站的域名,提取所述第二网页对应的网站的域名的后缀;

当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时,则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址;

根据所述第二网页的网络地址对所述第二网页进行访问,并爬取所述第二网页上的第二网页数据;及

25 将所述第一网页数据与所述第二网页数据分别输出至对应的类别。本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征和优点将从说明书、附图以及权利要求书变得明显。

附图说明

30 为了更清楚地说明本申请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申请的一些实施例,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。

图1为一实施例中网页数据处理方法的应用场景图;

图2为一实施例中网页数据处理方法的流程图;

图 3 为一实施例中网页数据处理装置的结构示意图；

图 4 为一实施例中计算机设备的结构示意图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用于解释本申请，并不用于限定本申请。

在详细说明根据本申请的实施例前，应该注意到的是，所述的实施例主要在于与网页数据处理方法、装置、计算机设备及存储介质相关的步骤和装置组件的组合。因此，所述装置组件和方法步骤已经在附图中通过常规符号在适当的位置表示出来了，并且只示出了与理解本申请的实施例有关的细节，以免因对于得益于本申请的本领域普通技术人员而言显而易见的细节模糊了本申请的公开内容。

在本文中，诸如左和右，上和下，前和后，第一和第二之类的关系术语仅仅用来区分一个实体或动作与另一个实体或动作，而不一定要求或暗示这种实体或动作之间的任何实际的这种关系或顺序。术语“包括”、“包含”或任何其他变体旨在涵盖非排他性的包含，由此使得包括一系列要素的过程、方法、物品或者设备不仅包含这些要素，而且还包含没有明确列出的其他要素，或者为这种过程、方法、物品或者设备所固有的要素。

请参见图 1，图 1 提供一网页数据处理方法的应用场景图，其中包括网页数据处理平台、第一网站服务器与第二网站服务器，当网页数据处理平台从第一网站服务器获取到第一网页的网页数据时，则查询第一网页数据对应的第二网页地址，进而获取第二网页对应的网站的域名，当获取到的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与标准域名对应的网络地址作为第二网页的网络地址，进而根据第二网页的网络地址，网页数据处理平台向第二网站服务器发送请求访问网络地址对应的第二网页，进而请求经第二网站服务器通过后，则网页数据处理平台访问第二网页，并爬取第二网页上的第二网页数据，进而将第一网页数据与第二网页数据输出。

请参见图 2，在其中一个实施例中，提供一网页数据处理方法的流程图，本实施例中以该方法应用到上述图 1 中的网页数据处理平台中来举例说明，该平台上运行有网页数据处理程序，通过该网页数据处理程序来实施网页数据处理。该方法包括如下步骤：

S202：获取第一网页的第一网页数据，查询与第一网页数据关联的第二网页地址。

具体地，第一网页是指存储有相应第一网页数据的网页，第一网页可以是直接由普通浏览器嵌入的搜索引擎直接搜索到的普通网页，第一网页可以是存储在第一网站服务器中的网页，网页数据处理平台可以直接通过开放的网络地址，查找到该服务器，进而通过该服务器访问第一网页获取第一网页上的第一网页数据。第一网页数据是指存储在第一网页上的网页内容，第一网页数据可以是文字数据、图片数据或数字数据等。第二网页是指存储有相应的第二网页数据的网页，第二网页可以是隐藏了网络地址的网页，该网页无法

通过直接由普通浏览器嵌入的搜索引擎直接搜索到的网页，例如，第二网页可以是深网或暗网等。网页地址是指在网络中，每个相应的网页都有对应的唯一的标识，例如，该网页地址可以是 URL（Uniform Resource Locator，统一资源定位器）地址，则第二网页地址是指第二网页的网页标识，第二网页地址可以第二网页 URL（Uniform Resource Locator，统一资源定位器）地址。进一步地，发送请求访问第一网页的请求，进而当该请求被验证通过时则访问第一网页，获取到第一网页的第一网页数据，根据第一网页数据，在数据查询库中获取到第一网页数据关联的第二网页地址，具体获取第一网页数据关联的第二网页地址的过程可以是通过第一网页数据在数据查询库中预存储的待匹配数据进行匹配，当匹配成功时，获取到该待匹配数据对应的第二网页地址作为第一网页数据对应的第二网页地址。例如，网页数据处理平台向第一网站服务器发送请求访问第一网页的请求，当该请求经第一网站服务器验证通过时则网页数据处理平台可以访问第一网页，进而获取第一网页的第一网页数据，根据第一网页数据，网页数据处理平台在数据查询库中获取到第一网页数据关联的第二网页。

需要说明的是，数据查询库是指存储有相应的网页数据以及网页数据关联的相应网页地址，可以是，存储有相应的网页数据，该网页数据关联有无法直接获取到的网页地址，如一些暗网或深网地址。

S204：从第二网页地址中获取第二网页对应的网站的域名，提取第二网页对应的网站的域名的后缀。

具体地，网站的域名是指相关网站的标识，网站的域名可以从网页地址中获取，例如，网站“百度”网站的域名为 baidu.com，该域名下可以有多个网页地址的网页，百度首页的网页地址为 www.baidu.com，进而，“百度”网站的域名可以从百度首页的网页地址中获取到。网站的域名的后缀是指根据网站的标识反应网站的类型的标记，网站的域名的后缀可以是国别域名，可以是通用域名等，例如，网站的域名的后缀可以是.com，可以是.cn 等。具体地，从获取到的第二网页地址中提取到第二网页对应的网站的域名，进而，从获取到的第二网页对应内的网站的域名中提取网站的域名的后缀。例如，网页数据处理平台根据获取到的第二网页地址，从第二网页地址中获取到第二网页对应的网站的域名，进而，网页数据处理平台从获取到的第二网页对应的网站的域名中提取该第二网页对应的网站的域名的后缀。

S206：当第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与标准域名对应的网络地址作为第二网页的网络地址。

具体地，标准域名是指预先存储的关联有可以访问相应的网页的网络地址的域名，标准域名可以是无法通过普通浏览器中嵌入的搜索引擎查询到的网页对应的网站的域名，例如，标准域名可以是一些深网或暗网的域名。网络地址是指可以唯一地标识网络中的计算机设备，该计算机与其他计算机进行通信时可以采用网络地址作为通信标识，相应的网页存储在计算机设备上也有对应该网络地址，例如，网络地址可以是 IP（Internet Protocol，

互联网协议)地址等。进一步地,当获取到的第二网页对应的网站的域名的后缀与预存的标准域名的后缀进行匹配,当第二网页对应的网站的域名的后缀与标准域名的后缀相同时,则第一级匹配成功,进而将第二网页对应的网站的域名的其他部分与标准域名的其他部分进行匹配,当匹配成功时则,则获取到该匹配成功的标准域名对应的网络地址作为第二网页的网络地址。例如,一些网站是有特定的域名后缀,可以是一些暗网或深网的网页对应的网站的域名的后缀为.onion。网页数据处理平台获取到第二网页对应的网站的域名的后缀与预存在域名存储库中的标准域名的后缀进行匹配,当第二网页对应的网站的域名的后缀与标准域名的后缀相同时,则第一级匹配成功,进而将第二网页对应的网站的域名的其他部分与标准域名的其他部分进行匹配,当其他部分也匹配成功时,则获取到标准域名对应的网络地址作为第二网页的网络地址。如,网页数据平台获取到第二网页对应的网站的域名为 abc.onion,则第二网页的域名的后缀为.onion,当该后缀与域名存储库中的标准域名的后缀相同时,则与标准域名的其他部分进行匹配,当其他部分也匹配成功时,则获取存储在域名存储库中的匹配成功的标准域名对应的网络地址作为第二网页的网络地址。需要说明的是,域名存储库是指存储有待匹配的标准域名以及与标准域名对应的网络地址的数据库。

先将第二网页地址对应的网站的域名的后缀与标准域名的后缀进行匹配,当匹配成功时再进行后续的匹配,节省时间提高效率。

S208: 根据第二网页的网络地址对网页进行访问,并爬取第二网页上的第二网页数据。

具体地,第二网页数据是指网页存储在第二网页上的网页内容,第二网页数据可以是文字数据、图片数据或数字数据等。当网页数据处理平台获取到第二网页的网络地址时,则根据第二网页的网络地址,从而查询到第二网页的网络地址对应的第二网站服务器,进而向第二网站服务器发送请求访问第二网页的访问请求,当访问请求通过时,则对第二网页进行访问,进而爬取第二网页上的第二网页数据。

S210: 将第一网页数据与第二网页数据分别输出至对应的类别。

具体地,网页数据处理平台将获取到的第一网页数据与第二网页数据共同输出。可以是,将第一网页数据与第二网页数据按照类别共同输出至数据库进行存储,也可以是,将第一网页数据与第二网页数据按照类别共同输出供用户查看等。进一步的,网页数据处理平台中可以存储有不同类别的网页数据,当网页数据处理平台将获取到的第一网页数据与第二网页数据时,则分别提取第一网页数据与第二网页数据的关键字,进而根据提取到的第一网页数据的关键字与第二网页数据的关键字,将第一网页数据与第二网页数据存储在提取出的关键字对应的类别下。例如,网页数据处理平台中可以存储有“安全漏洞”与“安全更新”的类别的网页数据,当提取到第一网页数据的关键字为漏洞时,则将该第一网页数据对应存储在“安全漏洞”类别下,当提取到第二网页数据的关键字为“补丁”时,则将该第二数据对应存储在“安全更新”的类别下。

本实施例中,网页数据处理平台获取到第一网页的第一网页数据,进而获取第一网页

数据对应的第二网页地址,根据第二网页地址获取第二网页对应的网站的域名,根据第二网页对应的网站域名的后缀获取到第二网页的网络地址,进而网页数据处理平台根据第二网页的网络地址访问第二网页,从而爬取第二网页数据,进而将第一网页数据与第二网页数据共同输出,第二网页可以为无法通过普通浏览器查询到的网页,第二网页数据存储在第二网页上,,进而通过本实施例的方法获取到第二网页数据,从而将获取到第一网页数据与第二网页数据,且将第一网页数据与第二网页数据输出至对应的类别下,防止网页数据的遗漏,提高数据分析的准确性。

在其中一个实施例中,步骤 S208,即根据第二网页的网络地址对第二网页进行访问,并爬取第二网页上的第二网页数据的步骤,可以包括:

当第二网页携带有限制访问标识时,则向代理服务器发送爬取第二网页上的网页数据的爬取指令。具体地,限制访问标识是指网站上携带的需要特定的计算机设备进行访问的标识,限制访问标识可以是字符标识等。代理服务器是指具有特定的访问权限的服务器,通过代理服务器可以访问携带有限制访问标识的第二网页。爬取指令是指访问指定的网页进而获取指定网页上的指定的网页数据的指令。进一步地,当第二网页携带有限制访问标识时,则需要采用特定的计算机设备进行访问,特定的计算机设备可以是代理服务器,进而网页数据处理平台则向代理服务器发送爬取指令,代理服务器可以根据爬取指令,访问第二网页进而爬取第二网页上的网页数据。

接收代理服务器返回的身份验证请求,并根据身份验证请求向代理服务器发送对应的身份标识。具体地,身份验证请求是指验证权限的请求,身份验证请求可以是文字数据、图片数据或数字数据等。身份标识是指表明具有对应操作权限的身份信息,身份标识可以是具有发送爬取指令权限的身份信息,例如,身份标识可以是与身份验证请求对应的文字数据、图片数据或数字数据等,如,身份标识可以是验证码,可以是账户密码等。进一步地,网页数据处理平台向代理服务器发送有爬取指令时,则接收到代理服务器返回的身份验证请求,进而网页数据处理平台根据该身份验证请求,向代理服务器发送相应的身份标识。可以是,网页数据平台向代理服务器发送爬取第二网页数据的爬取指令时,则代理服务器返回身份验证请求,进而在网页数据处理平台的界面上弹出相应的界面,显示有“请输入操作用户名和密码”,当用户在界面上完成用户名和密码的输入时,则网页数据平台将用户输入的用户名和密码也即身份标识发送至代理服务器。需要说明的是,代理服务器返回的身份验证请求还可以是对应的验证码,当用户根据网页数据处理平台的界面提示输入相应的验证码时,则网页数据平台向代理服务器发送用户输入的验证码,也即向代理服务器发送对应的身份标识。

当身份标识经代理服务器验证成功时,则接收代理服务器返回的从第二网页上爬取到的网页数据。具体地,当网页数据处理平台向代理服务器发送的身份标识经过代理服务器验证成功时,则通过验证即有向代理服务器发送爬取指令的权限,则代理服务器可以根据该爬取指令,向第二网站服务器发送访问第二网页的请求,当该访问请求经过第二网站服

务器验证成功时,则代理服务器访问第二网页,进而爬取该第二网页的数据,从而,网页数据处理平台接收代理服务器爬取到的第二网页数据。

需要说明的是,本实施例中,代理服务器可以采用 ss 系统 (shadowsocks 系统),进而通过 ss 系统实现上述步骤从而对第二网页数据进行爬取。

- 5 本实施例中,当第二网页携带有限制访问标识时,则通过代理服务器对第二网页数据进行爬取,增强适用性,且代理服务器在对第二网页数据进行爬取时需要对当前操作的身份进行验证,保证第二网页数据传输与交互的安全性。

在其中一个实施例中,步骤 S208,即根据第二网页的网络地址对网页进行访问,并爬取第二网页上的第二网页数据的步骤,还可以包括:

- 10 当第二网页未携带有限制访问标识时,则根据第二网页地址,获取第二网页对应的爬取逻辑与通信协议。具体地,爬取逻辑是指爬取网页上的网页数据时采用的爬取规则,爬取逻辑可以包含有网页的地址,还存储有网页的待爬取网页数据的位置,例如,可以是网页的待爬取网页数据的数据行数,还可以是网页的待爬取网页数据所在网页的显示区域的坐标等,爬取逻辑还可以包含获取网页数据的数量。通信协议是指在网络通信时,网页数据
15 数据处理平台与网站服务器遵守的相应的通信规则或通信协议。通信协议可以是采用 HTTP 通信协议的通信方式,通信协议可以是采用 FTP 通信协议的通信方式等。进一步地,当第二网页上未携带有限制访问标识时,则可以直接通过网页数据处理平台访问第二网页,则网页数据处理平台获取预先存储的爬取第二网页的网页数据的爬取逻辑,进而获取预先存储的第二网页对应的通信协议。

- 20 根据第二网页对应的通信协议访问第二网页并遍历第二网页的第二网页数据。具体地,当网页数据处理平台获取到第二网页对应的通信协议时,则网页数据处理平台将第二网页对应的通信协议以及访问请求发送到第二网页对应的第二网站服务器,当第二网页服务器接收到该第二网页对应的通信协议以及访问请求时,经过对接收到的通信协议以及访问请求的验证成功时,则允许网页数据处理平台访问第二网页,进而网页数据处理平台遍
25 历第二网页上的网页数据,网页数据处理平台可以对网页数据中的文字数据进行逐行逐个字符进行查询,直至查询至第二网页上的网页数据的最后的字符,即遍历第二网页的网页数据,也可以是对网页数据中的图片数据逐个图片的进行查询,直至查询至第二网页上最后一幅图片完成遍历第二网页的第二网页数据。

- 当遍历到与爬取逻辑对应的第二网页数据时,则爬取与爬取逻辑对应的第二网页数据。具体地,爬取逻辑中可以预设有待爬取网页数据的位置,待爬取网页数据的数据关键字,以及当查询到待爬取网页的数据关键字时获取的数据的量,如当第二网页数据为文字数据时,则爬取逻辑预设爬取文字数据的位置为全部网页数据或前五
30 行网页数据等,设置待爬取网页数据的关键字,进而查询到待爬取网页数据中包含的关键字时,获取包含关键字的网页数据的数量具体可以为包含该关键字的网页数据的前五行,全部网页数据等。网
35 页数据处理平台遍历当前第二网页的第二网页数据,当遍历到与爬取逻辑对应的第二网页

数据时,则将与爬取逻辑对应的第二网页数据进行爬取。可以是,爬取逻辑中预设有待爬取网页的网页数据的位置为全部网页数据,且设置有待爬取网页的网页数据的关键字为“平安银行”,当网页数据处理平台遍历第二网页的第二网页数据时,遍历全部的第二网页数据,且当查询到有“平安银行”对应的数据,则将该第二网页的所有网页数据进行爬取。

- 5 本实施例中,当第二网页未携带有限制访问标识时,则直接采用网页数据处理平台对第二网页的第二网页数据进行爬取,提高效率,且根据爬取规则对第二网页上的第二网页数据进行爬取,爬取数据准确,保证第二数据获取准确。

在其中一个实施例中,步骤 S210,即将第一网页数据与第二网页数据分别输出至对应的类别的步骤,可以包括:

- 10 将第一网页数据携带的网页标识以及第二网页数据携带的网页标识分别与已存储的网页标识进行匹配。网页标识是指对应的网页数据来源的网页的标识,该网页标识可以将网页数据来源的网页与其他网页进行区分,网页标识可以是网页对应的网站的名称,网页标识可以是网页地址,网页标识也可以是网页对应的网站的网站域名等。例如,网页标识可以是网页的 URL 地址,可以是网页的 URL 地址对应的网站的网站域名等。进一步地,
- 15 网页数据处理平台获取到的第一网页数据携带有对应的第一网页的网页标识,第二网页数据携带有对应的第二网页的网页标识,进而,网页数据处理平台将第一网页的网页标识和第二网页的网页标识分别与已存储的网页标识进行逐一匹配,可以是,在主线程中先将第一网页数据上携带的网页标识与已存储的网页标识逐一进行匹配,当第一网页数据上携带的网页标识与已存储的网页标识匹配完成时,则在主线程中将第二网页数据上携带的网页
- 20 标识与已存储的网页标识逐一进行匹配;也可以是,在主线程中将第一网页数据上携带的网页标识与已存储的网页标识逐一进行匹配,进而在与主线程互为异步的另一线程中将第一网页数据上携带的网页标识与已存储的网页标识逐一进行匹配。例如,网页数据处理平台获取到的第一网页数据携带有对应的第一网页的 URL 地址,第二网页数据携带有对应的第二网页的 URL 地址,进而网页数据处理平台将第一网页数据的携带的第一网页的
- 25 URL 地址与第二网页数据携带的第二网页的 URL 地址与已存储的 URL 地址逐一进行匹配。

- 当第一网页数据携带的网页标识以及第二网页数据携带的网页标识其中至少一个与已存储的网页标识不匹配时,则提取不匹配的网页数据的关键字。具体地,当网页数据处理平台将第一网页数据携带的网页标识与第二网页数据携带的网页标识与已存储的网页
- 30 标识进行逐一匹配时,第一网页数据携带的网页标识以及第二网页数据携带的网页标识其中至少一个与已存储的网页标识匹配不成功时,则未匹配成功的网页数据未被存储,则提取不匹配的网页数据的关键字。可以是,当第一网页数据携带的网页标识未与已存储的网页标识匹配成功时,则第一网页数据未被存储,则提取第一网页数据的关键字。也可以是,第二网页数据携带的网页标识未与已存储的网页数据标识匹配成功,则第二网页数据未被
- 35 存储,则提取第二网页数据的关键字。还可以是,当第一网页数据携带的网页标识和第二

网页携带的网页标识均未与已存储的网页标识匹配成功时，则第一网页数据与第二网页数据均未被存储，则提取第一网页数据的关键字与第二网页数据的关键字。

将不匹配的网页数据输出至与关键字对应的存储类别下。具体地，网页数据处理平台中存储有不同类别的网页数据，当经过上述步骤识别出未存储的网页数据时，则提取网页数据的关键字，进而根据关键字，将不匹配的网页数据进行输出，存储在关键字对应的存储类别中。例如，网页数据处理平台中存储有不同类别的网页数据可以是行业要闻、安全漏洞、安全更新、漏洞利用、国际咨询、推荐阅读等，例如，而行业要闻对应的关键字有金融、银行、保险、证券、信用卡、支付、swift、bank、banks 等，安全漏洞对应的关键字有每日安全资讯、CVE（Common Vulnerabilities & Exposures 公共漏洞和暴露）或漏洞等，安全更新对应的关键字有 update、补丁、安全更新或升级等，第一网页数据未存储，则提取第一网页数据的关键字，如第一网页数据的关键字为“补丁”，则将该第一网页数据进行输出，对应存储在“安全漏洞”，当前第一网页数据的关键字均不为行业要闻、安全漏洞、安全更新、漏洞利用、国际咨询对应的关键字，则将第一网页数据输出，存储在对应的推荐阅读类别中。当第二网页数据未存储或第一网页数据与第二网页数据均未被存储，则按照如上步骤将未被存储的网页数据输出并进行存储在对应的存储类别下，在此不再赘述。

需要说明的是，当获取到第一网页数据与第二网页数据时，第一网页数据与第二网页数据中可能存在有一些特殊字符，如下划线、空格或乱码等，当第一网页数据与第二网页数据中存在有特殊字符时，则选取第一网页数据与第二网页数据对应的转换逻辑，根据转换逻辑，将第一网页数据与第二网页数据进行转换，即可以删除下划线、删除空格或删除乱码等。其中，转换逻辑是指将网页数据转换为特定显示格式或特定显示数据的规则。

本实施例中，先将获取到的第一网页数据携带的网页标识与第二网页数据携带的网页标识与已存储的网页数据进行匹配，保证不会重复存储网页数据，提高存储效率，进而将未存储的网页数据存储在与相应的类别下，方便后续查找，增强适用性。

在其中一个实施例中，上述方法还可以包括：

获取预设的接收第一网页数据与第二网页数据的邮箱地址。具体地，网页数据处理平台可以将存储的第一网页数据与第二网页数据进行推送，且接收第一网页数据与第二网页数据的邮箱可以是预先设置并进行存储的，则网页数据处理平台获取预设的接收第一网页数据与第二网页数据的邮箱地址。

提取邮箱地址对应的部门标识，并获取与部门标识对应的存储类别。具体地，部门标识是指不同的机构的识别标识，部门标识可以是部门名称，可以是部门代码等。具体地，当网页数据处理平台获取到预设的接收第一网页数据与第二网页数据的邮箱地址时，则提取邮箱地址对应的部门标识，根据部门标识，获取到该部门对应的存储类别，即获取到该部门接收网页数据的类别。可以是，邮箱地址中包含有对应的部门标识，如部门代码。则网页数据处理平台从邮箱地址中直接提取对应的部门标识，根据该部门标识，网页数据处

理平台获取到该部门接收网页数据的类别。也可以是，当获取到邮箱地址时，网页数据处理平台根据邮箱地址与预存储的邮箱地址进行匹配，当匹配成功时，则获取预存储的匹配成功的邮箱地址对应的部门标识作为该邮箱地址的部门标识，根据该部门标识，获取到该部门接收网页数据的类别。例如，网页数据处理平台提取到该邮箱地址对应的部门标识为行业分析部门，则获取到行业分析部门对应的存储类别为行业要闻。

将所获取的存储类别下的第一网页数据与第二网页数据发送至邮箱地址对应的邮箱中。具体地，当网页数据处理平台获取到与邮箱地址对应的部门标识时，则获取部门标识对应的存储类别，进而网页数据处理平台将获取到的存储类别下的第一网页数据与第二网页数据都发送至邮箱地址对应的邮箱中，进而，将发送完成的第一网页数据与第二网页数据添加发送完成标签。例如，网页数据处理平台提取到该邮箱地址对应的部门标识为行业分析部门，则获取到行业分析部门对应的存储类别为行业要闻，进而将行业要闻下存储的第一网页数据与第二网页数据都发送至邮箱地址对应的邮箱中，进而将发送完成的第一网页数据与第二网页数据都添加有发送完成标签。需要说明的是，可以预设有发送时间，当网页数据处理平台检测到系统时间为预设的发送时间时，则将获取的存储类别下的第一网页数据与第二网页数据发送至邮箱地址对应的邮箱中。

本实施例中，可以根据邮箱地址对应的部门标识，获取部门标识对应的存储类别，并将存储类别对应的第一网页数据与第二网页数据发送给网页邮箱对应的邮箱，即根据部门标识将该部门感兴趣的第一网页数据与第二网页数据进行推送，增强适用性，且当将第一网页数据与第二网页数据进行发送后添加已发送标签，避免重复推送，提高效率增强适用性。

在其中一个实施例中，上述实施例中的根据第二网页的网络地址对网页进行访问，并爬取第二网页上的第二网页数据的步骤，可以包括：

预设爬取第二网页的第二网页数据的爬取时间。具体地，网页数据处理平台设置有对第二网页的第二网页数据的爬取时间，爬取时间的设置可以是固定的时间，也可以设置为间隔的时间段等，例如，爬取时间的设置可以是整点，如早晨 8 点，早晨 10 点，也可以是间隔半小时或间隔一小时等。

当到达爬取时间时，则从网络地址库中随机选取可用的爬取网络地址。爬取网络地址是指用于爬取第二网页数据时采用的与对方进行通讯时的通讯标识，例如，爬取网络地址可以是网页数据处理平台获取的 IP 地址等。网络地址库是预先设置在网页数据处理平台中的可以存储不同的网络地址的数据库，例如网络地址库中可以存储有第一 IP 地址、第二 IP 地址等不同的 IP 地址。进一步地，当网页数据平台检测到到达预设的爬取时间时，则网页数据处理平台从网络地址库中随机选取可用的爬取网络地址，如选取到第一 IP 地址作为爬取网络地址时，可以将该被选取的第一 IP 地址进行标记，被标记的第一 IP 地址是正在使用的网络地址，当下一次网页爬取平台从网络地址库中选取网络地址时，则从未标记的网络地址中选取网络地址，当被标记的网络地址，即第一 IP 地址使用完成时，则

将该网络地址的标记进行删除。

通过爬取网络地址访问第二网页，并爬取第二网页上的第二网页数据。具体地，当网页数据处理平台获取到爬取网络地址时，则向第二网站服务器发送第二网页对应的通信协议以及访问请求，此时通信协议以及访问请求携带有爬取网络地址，当爬取网络地址被第二网站服务器验证成功时，进而第二网站服务器验证通信协议以及访问请求，当通信协议与访问请求都被验证成功时，则网页数据处理平台访问第二网页，并根据爬取逻辑爬取第二网页上的第二网页数据。

本实施例中，当对第二网页进行访问并爬取第二网页上的网页数据时，网页数据处理平台从网络地址库中存储网络地址中随机获取一个网络地址，进而完成后续的对第二网页上的第二网页数据的爬取，避免爬取反复使用相同的网络地址而触发第二网页的风控机制从而导致爬取第二网页数据不成功，增强适用性。

在其中一个实施例中，上述实施例中根据第二网页的网络地址对第二网页进行访问，并爬取第二网页上的第二网页数据的步骤，包括：

根据第二网页的网络地址访问第二网页并查询第二网页是否渲染完成。具体地，渲染是指第二网页上的部分数据在显示时是被隐藏的状态，则将被隐藏的数据显示完成的方式。网页数据处理平台访问第二网页时，则检测第二网页上是否有隐藏的第二网页数据，可以是，网页数据处理平台检测第二网页上是否有数据携带有隐藏标签，当携带有隐藏标签时，则第二网页未渲染完成，也可以是网页数据处理平台检测第二网页上的第二网页数据是否需要进行特定的操作，当需要进行特定的操作则第二网页未渲染完成，特定的操作可以是需要用户点击提示信息“显示全文”的操作，进而第二网页根据用户点击提示信息后将隐藏的数据进行显示。

当第二网页未渲染完成时，则根据第二网页地址获取第二网页对应的渲染逻辑。具体地，渲染逻辑是指将网页上隐藏的数据进行完全显示的规则，当网页数据处理平台查询到第二网页未渲染完成时，则网页数据处理平台根据第二网地址，选取第二网页对应的渲染规则。

根据第二网页对应的渲染逻辑对第二网页进行渲染。具体地，当网页数据处理平台查询到第二网页未完成渲染时，则根据第二网页地址选取第二网页对应的渲染逻辑，进而根据第二网页对应的渲染逻辑对第二网页进行渲染，当第二网页完成渲染时，则第二网页上的第二网页数据显示完成。

爬取渲染完成的第二网页上的第二网页数据。具体地，根据上述步骤，当对第二网页进行渲染完成时，则第二网页的第二网页数据显示完成，则网页数据处理平台爬取经过渲染完成的第二网页上的第二网页数据。

上述实施例中，当第二网页未渲染完成时，根据第二网页地址选取第二网页的渲染逻辑，根据第二网页的渲染逻辑，将第二网页进行渲染完成时再爬取第二网页上的第二网页数据，保证全面爬取到第二网页的网页数据，避免数据有遗漏。

应该理解的是，虽然图 2 的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图 2 中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

在其中一个实施例中，请参见图 3，提供一网页数据处理装置的结构示意图，网页数据处理装置 300 可以包括：

10 查询模块 310，用于获取第一网页的第一网页数据，查询与第一网页数据关联的第二网页地址。

提取模块 320，用于从第二网页地址中获取第二网页对应的网站的域名，提取第二网页对应的网站的域名的后缀。

15 获取模块 330，用于当第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与标准域名对应的网络地址作为第二网页的网络地址。

爬取模块 340，用于根据第二网页的网络地址对第二网页进行访问，并爬取第二网页上的第二网页数据。

输出模块 350，用于将第一网页数据与第二网页数据分别输出至对应的类别。

在其中一个实施例中，爬取模块 340 可以包括：

20 发送单元，用于当第二网页携带有限制访问标识时，则向代理服务器发送爬取第二网页上的网页数据的爬取指令。

第一接收单元，用于接收代理服务器返回的身份验证请求，并根据身份验证请求向代理服务器发送对应的身份标识。

25 第二接收单元，用于当身份标识经代理服务器验证成功时，则接收代理服务器返回的从第二网页上爬取到的网页数据。

在其中一个实施例中，爬取模块 340 还可以包括：

获取单元，用于当第二网页未携带有限制访问标识时，则根据第二网页地址，获取第二网页对应的爬取逻辑与通信协议。

30 遍历单元，用于根据第二网页对应的通信协议访问第二网页并遍历第二网页的第二网页数据。

第二网页数据爬取单元，用于当遍历到与爬取逻辑对应的第二网页数据时，则爬取与爬取逻辑对应的第二网页数据。

在其中一个实施例中，输出模块 350 可以包括：

35 匹配单元，用于将第一网页数据携带的网页标识以及第二网页数据携带的网页标识分别与已存取的网页标识进行匹配。

提取单元,用于当第一网页数据携带的网页标识以及第二网页数据携带的网页标识其中至少一个与已存储的网页标识不匹配时,则提取不匹配的网页数据的关键字。

存储单元,用于将不匹配的网页数据输出至与关键字对应的存储类别下。

在其中一个实施例中,输出模块 350 还可以包括:

- 5 邮箱地址获取单元,用于获取预设的接收第一网页数据与第二网页数据的邮箱地址。

存储类别获取单元,用于提取邮箱地址对应的部门标识,并获取与部门标识对应的存储类别。

数据发送单元,用于将所获取的存储类别下的第一网页数据与第二网页数据发送至邮箱地址对应的邮箱中。

- 10 在其中一个实施例中,爬取模块 340 还可以包括:

爬取时间预设单元,用于预设爬取第二网页的第二网页数据的爬取时间。

网络地址选取单元,用于当到达爬取时间时,则从网络地址库中随机选取可用的爬取网络地址。

访问单元,用于通过爬取网络地址访问第二网页,并爬取第二网页上的第二网页数据。

- 15 在其中一个实施例中,爬取模块 340 还可以包括:

渲染查询单元,用于根据第二网页的网络地址访问第二网页并查询第二网页是否渲染完成。

渲染逻辑获取单元,用于当第二网页未渲染完成时,则根据第二网页地址获取第二网页对应的渲染逻辑。

- 20 渲染单元,用于根据第二网页对应的渲染逻辑对第二网页进行渲染。

渲染数据爬取单元,用于爬取渲染完成的第二网页上的第二网页数据。

- 25 上述关于网页数据处理装置的具体限定可以参见上文中关于网页数据处理方法的限定,在此不再赘述。上述网页数据处理装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。该处理器可以为中央处理单元(CPU)、微处理器、单片机等。上述网页数据处理装置可以实现为一种计算机可读指令的形式,计算机可读指令可在如图 1 所示的网页数据处理平台设备上运行。

- 30 本申请实施例提出了一种计算机设备,该计算机设备可以是服务器,其内部结构图可以如图 4 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为非易失性存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储网页数据处理数据。该计算机设备的网络接口用于与
35 外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种网页数据处

理方法。

本领域技术人员可以理解，图4中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

5 计算机可读指令计算机可读指令计算机可读指令其中，处理器执行该计算机可读指令时实现以下步骤：获取第一网页的第一网页数据，查询与第一网页数据关联的第二网页地址。从第二网页地址中获取第二网页对应的网站的域名，提取第二网页对应的网站的域名的后缀。当第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与标准域名对应的网络地址作为第二网页的网络地址。根据第二网页的网络地址对第二网页进行访问，并爬取第二网页上的第二网页数据。将第一网页数据与第二网页数据分别输出至对

10 应的类别。

在其中一个实施例中，处理器执行计算机可读指令时所实现根据第二网页的网络地址对第二网页进行访问，并爬取第二网页上的第二网页数据的步骤，可以包括：当第二网页携带有限制访问标识时，则向代理服务器发送爬取第二网页上的网页数据的爬取指令。接收代理服务器返回的身份验证请求，并根据身份验证请求向代理服务器发送对应的身份标识。当身份标识经代理服务器验证成功时，则接收代理服务器返回的从第二网页上爬取到的网页数据。

15

在其中一个实施例中，处理器执行计算机可读指令时所实现根据第二网页的网络地址对网页进行访问，并爬取第二网页上的第二网页数据的步骤，可以包括：当第二网页未携带有限制访问标识时，则根据第二网页地址，获取第二网页对应的爬取逻辑与通信协议。根据第二网页对应的通信协议访问第二网页并遍历第二网页的第二网页数据。当遍历到与爬取逻辑对应的第二网页数据时，则爬取与爬取逻辑对应的第二网页数据。

20

在其中一个实施例中，处理器执行计算机可读指令时所实现将第一网页数据与第二网页数据分别输出至对应的类别的步骤，包括：第一网页数据携带的网页标识以及第二网页数据携带的网页标识分别与已存取的网页标识进行匹配。当第一网页数据携带的网页标识以及第二网页数据携带的网页标识其中至少一个与已存储的网页标识不匹配时，则提取不匹配的网页数据的关键字。将不匹配的网页数据输出至与关键字对应的存储类别下。

25

在其中一个实施例中，所述处理器执行所述计算机可读指令时还可以实现：获取预设的接收第一网页数据与第二网页数据的邮箱地址。提取邮箱地址对应的部门标识，并获取与部门标识对应的存储类别。将所获取的存储类别下的第一网页数据与第二网页数据发送至邮箱地址对应的邮箱中。

30

在其中一个实施例中，处理器执行计算机可读指令时所实现根据第二网页的网络地址对网页进行访问，并爬取第二网页上的第二网页数据的步骤，还可以包括：预设爬取第二网页的第二网页数据的爬取时间。当到达爬取时间时，则从网络地址库中随机选取可用的爬取网络地址。通过爬取网络地址访问第二网页，并爬取第二网页上的第二网页数据。

35

在其中一个实施例中,处理器执行计算机可读指令时所实现根据第二网页的网络地址对第二网页进行访问,并爬取第二网页上的第二网页数据的步骤,可以包括:根据第二网页的网络地址访问第二网页并查询第二网页是否渲染完成。当第二网页未渲染完成时,则根据第二网页地址获取第二网页对应的渲染逻辑。根据第二网页对应的渲染逻辑对第二网页进行渲染。爬取渲染完成的第二网页上的第二网页数据。

上述关于计算机设备的具体限定可以参见上文中关于网页数据处理方法的限定,在此不再赘述。

在其中一个实施例中,请继续参见图4,提供一种存储有计算机可读指令的非易失性计算机可读存储介质,所述计算机可执行指令被一个或多个处理器执行,使得所述一个或多个处理器执行以下步骤:获取第一网页的第一网页数据,查询与第一网页数据关联的第二网页地址。从第二网页地址中获取第二网页对应的网站的域名,提取第二网页对应的网站的域名的后缀。当第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时,则获取与标准域名对应的网络地址作为第二网页的网络地址。根据第二网页的网络地址对第二网页进行访问,并爬取第二网页上的第二网页数据。将第一网页数据与第二网页数据分别输出至对应的类别。

上述关于计算机存储介质的具体限定可以参见上文中关于网页数据处理方法的限定,在此不再赘述。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机可读指令来指令相关的硬件来完成,所述的计算机可读指令可存储于一非易失性计算机可读存储介质中,该计算机可读指令在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

以上所述实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范

围。因此，本申请专利的保护范围应以所附权利要求为准。