(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 109788050 B
(45)授权公告日 2021.08.20

(21)申请号 201811653799.9

(22)申请日 2018.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 109788050 A

(43)申请公布日 2019.05.21

(73)专利权人 奇安信科技集团股份有限公司
地址 100088 北京市西城区新街口外大街28号102号楼3层332号

(72)发明人 曾海波 陈筱牧

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 周天宇

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 29/12(2006.01)

(56)对比文件
CN 103595827 A,2014.02.19
CN 103634422 A,2014.03.12
CN 107465666 A,2017.12.12
CN 106453598 A,2017.02.22

审查员 李世成

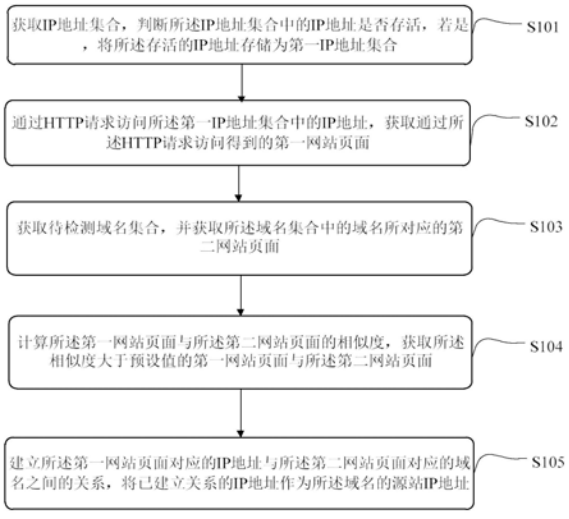
权利要求书2页 说明书6页 附图4页

(54)发明名称

一种获取源站IP地址方法、系统、电子设备和介质

(57)摘要

本申请提供了一种获取源站IP地址方法、系统、电子设备和介质。所述方法包括：获取IP地址集合，判断所述IP地址集合中的IP地址是否存活，若是，将所述存活的IP地址存储为第一IP地址集合；通过HTTP请求访问所述第一IP地址集合中的IP地址，获取通过所述HTTP请求访问得到的第一网站页面；获取待检测域名集合，并获取所述域名集合中的域名所对应的第二网站页面；计算所述第一网站页面与所述第二网站页面的相似度，获取所述相似度大于预设值的第一网站页面与所述第二网站页面；建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系，将已建立关系的IP地址作为所述域名的源站IP地址。



CN 109788050 B

1.一种获取源站IP地址方法,其特征在于,所述方法包括:

获取IP地址集合,判断所述IP地址集合中的IP地址是否存活,若是,将所述存活的IP地址存储为第一IP地址集合;

通过HTTP请求访问所述第一IP地址集合中的IP地址,获取通过所述HTTP请求访问得到的第一网站页面;

获取待检测域名集合,并获取所述域名集合中的域名所对应的第二网站页面;

计算所述第一网站页面与所述第二网站页面的相似度,获取所述相似度大于预设值的第一网站页面与所述第二网站页面;

建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系,将已建立关系的IP地址作为所述域名的源站IP地址;

其中,在所述判断所述IP地址集合中的IP地址是否存活之前,该方法还包括:

获取CDN节点IP地址;

判断所述IP地址集合中是否包含所述CDN节点IP地址,若是,将所述IP地址集合中与所述CDN节点IP地址相同的IP地址进行删除。

2.根据权利要求1所述的获取源站IP地址方法,其特征在于,所述获取IP地址集合,包括:获取预设条件下的IP地址,所述预设条件下包括预设区域、预设IP协议版本。

3.根据权利要求1所述的获取源站IP地址方法,其特征在于,所述判断所述IP地址集合中的IP地址是否存活,包括:

通过探测工具nmap进行判断所述IP地址集合中的IP地址是否存活。

4.根据权利要求1所述的获取源站IP地址方法,其特征在于,所述获取CDN节点IP地址,包括:

通过IPIP资料库或GeoV2资料库获取CDN节点IP地址。

5.根据权利要求1所述的获取源站IP地址方法,其特征在于,所述计算所述第一网站页面与所述第二网站页面的相似度,包括:

通过PHASH算法计算所述第一网站页面与所述第二网站页面的相似度。

6.根据权利要求5所述的获取源站IP地址方法,其特征在于,所述获取通过所述HTTP请求访问得到的第一网站页面,包括:

判断所述通过所述HTTP请求访问的返回页面是否为空,若是,不获取所述空页面。

7.一种获取源站IP地址的系统,其特征在于,所述系统包括:

第一IP地址集合获取模块,用于获取IP地址集合,判断所述IP地址集合中的IP地址是否存活,若是,将所述存活的IP地址存储为第一IP地址集合;

第一网站页面获取模块,用于通过HTTP请求访问所述第一IP地址集合中的IP地址,获取通过所述HTTP请求访问得到的第一网站页面;

第二网站页面获取模块,用于获取待检测域名集合,并获取所述域名集合中的域名所对应的第二网站页面;

相似度计算模块,用于计算所述第一网站页面与所述第二网站页面的相似度,获取所述相似度大于预设值的第一网站页面与所述第二网站页面;

源站IP地址获取模块,用于建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系,将已建立关系的IP地址作为所述域名的源站IP地址;

其中,所述第一IP地址集合获取模块用于判断所述IP地址集合中的IP地址是否存活之前,还包括:

获取CDN节点IP地址;

判断所述IP地址集合中是否包含所述CDN节点IP地址,若是,将所述IP地址集合中与所述CDN节点IP地址相同的IP地址进行删除。

8.一种电子设备,其特征在于,所述设备包括:

处理器;

存储器,其存储有计算机可执行程序,该程序在被所述处理器执行时,使得所述处理器执行如权利要求1-6中任一项所述的获取源站IP地址方法。

9.一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1-6中任一项所述的获取源站IP地址方法。

一种获取源站IP地址方法、系统、电子设备和介质

技术领域

[0001] 本申请涉及一种获取源站IP地址方法、系统、电子设备和介质。

背景技术

[0002] 内容分发网络(Content Delivery Network,CDN),其基本思路是尽可能避开互联网上有可能影响数据传输速度和稳定性的瓶颈和环节,使内容传输得更快、更稳定。通过在网络各处放置节点服务器所构成的在现有的互联网基础之上的一层智能虚拟网络,CDN系统能够实时地根据网络流量和各节点的连接、负载状况以及到用户的距离和响应时间等综合信息将用户的请求重新导向离用户最近的服务节点上。核心目的就是使用户可就近访问网络,取得所需内容,解决网络拥挤的状况,明显提高用户访问网站的响应速度或者用户下载速度。

[0003] CDN是将网站内容存放在遍布全国乃至全球的CDN节点上,用户访问该网站,就可以就近获取CDN上缓存的内容,从而提升访问速度。开启CDN后的网站,不同地区用户访问会是不同的服务器,而网站的真实服务器(源服务器)一般只有CDN节点回去访问获取,全国各地的用户访问的CDN节点服务器,并不直接访问源服务器,这样就可以介绍网站服务器宽带资源,降低服务器压力。

[0004] 现在很多网站用了CDN技术,但采用CDN技术后,原来用来获取访问源的IP地址的程序已不能正常使用,它拿到的并不是访问源的真实IP地址,而是CDN节点的IP地址。例如,在没有使用CDN之前,我们如果想在网站程序里拿到浏览者的IP,只需要读取REMOTE_ADDR这个服务器变量就行了,而采用CDN后,REMOTE_ADDR这个变量的值并不是访问源的上网IP地址,而是CDN节点的IP地址。因此如何获取源IP地址已成为一个值得关注的问题。

发明内容

[0005] 本申请的一个方面提供了一种获取源站IP地址方法,所述方法包括:获取IP地址集合,判断所述IP地址集合中的IP地址是否存活,若是,将所述存活的IP地址存储为第一IP地址集合;通过HTTP请求访问所述第一IP地址集合中的IP地址,获取通过所述HTTP请求访问得到的第一网站页面;获取待检测域名集合,并获取所述域名集合中的域名所对应的第二网站页面;计算所述第一网站页面与所述第二网站页面的相似度,获取所述相似度大于预设值的第一网站页面与所述第二网站页面;建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系,将已建立关系的IP地址作为所述域名的源站IP地址。

[0006] 可选地,所述获取IP地址集合,包括:获取预设条件下的IP地址,所述预设条件下包括预设区域、预设IP协议版本。

[0007] 可选地,所述判断所述IP地址集合中的IP地址是否存活,包括:通过探测工具nmap进行判断所述IP地址集合中的IP地址是否存活。

[0008] 可选地,在所述判断所述IP地址集合中的IP地址是否存活之前,还包括:获取CDN

节点IP地址;判断所述IP地址集合中是否包含所述CDN节点IP地址,若是,将所述IP地址集合中与所述CDN节点IP地址相同的IP地址进行删除。

[0009] 可选地,所述获取CDN节点IP地址,包括:通过IPIP资料库或GeoV2资料库获取CDN节点IP地址。

[0010] 可选地,所述计算所述第一网站页面与所述第二网站页面的相似度,包括:通过PHASH算法计算所述第一网站页面与所述第二网站页面的相似度。

[0011] 可选地,所述获取通过所述HTTP请求访问得到的第一网站页面,包括:判断所述通过所述HTTP请求访问的返回页面是否为空,若是,不获取所述空页面。

[0012] 本申请另一方面提供了一种获取源站IP地址的系统,所述系统包括:第一IP地址集合获取模块,用于获取IP地址集合,判断所述IP地址集合中的IP地址是否存活,若是,将所述存活的IP地址存储为第一IP地址集合;第一网站页面获取模块,用于通过HTTP请求访问所述第一IP地址集合中的IP地址,获取通过所述HTTP请求访问得到的第一网站页面;第二网站页面获取模块,用于获取待检测域名集合,并获取所述域名集合中的域名所对应的第二网站页面;相似度计算模块,用于计算所述第一网站页面与所述第二网站页面的相似度,获取所述相似度大于预设值的第一网站页面与所述第二网站页面;源站IP地址获取模块,用于建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系,将已建立关系的IP地址作为所述域名的源站IP地址。

[0013] 本申请的又一方面提供了一种电子设备,所述设备包括:处理器;存储器,其存储有计算机可执行程序,该程序在被所述处理器执行时,使得所述处理器执行如上文所述的获取源站IP地址方法。

[0014] 本申请的再以方面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如上文所述的获取源站IP地址方法。

附图说明

[0015] 为了更完整地理解本申请及其优势,现在将参考结合附图的以下描述,其中:

[0016] 图1示意性示出了本申请实施例提供的获取源站IP地址方法的方法流程图;

[0017] 图2示意性示出了本申请实施例提供的获取源站IP地址方法中在所述判断所述IP地址集合中的IP地址是否存活之前的步骤流程图;

[0018] 图3示意性示出了本申请实施例提供的获取源站IP地址方法系统框图;

[0019] 图4示意性示出了根据本申请实施例提供的电子设备框图。

具体实施方式

[0020] 以下,将参照附图来描述本申请的实施例。但是应该理解,这些描述只是示例性的,而并非要限制本申请的范围。在下面的详细描述中,为便于解释,阐述了许多具体的细节以提供对本申请实施例的全面理解。然而,明显地,一个或多个实施例在没有这些具体细节的情况下也可以被实施。此外,在以下说明中,省略了对公知结构和技术的描述,以避免不必要地混淆本申请的概念。

[0021] 在此使用的术语仅仅是为了描述具体实施例,而并非意在限制本申请。在此使用的术语“包括”、“包含”等表明了所述特征、步骤、操作和/或部件的存在,但是并不排除存在

或添加一个或多个其他特征、步骤、操作或部件。

[0022] 本申请的一个实施例提供了一种获取源站IP地址方法，参见图1，所述方法包括步骤S101～步骤S102的内容：

[0023] 步骤S101，获取IP地址集合，判断所述IP地址集合中的IP地址是否存活，若是，将所述存活的IP地址存储为第一IP地址集合。

[0024] 其中，所述获取IP地址集合，包括：获取预设条件下的IP地址，所述预设条件下包括预设区域、预设IP协议版本。预设区域可以示例性地为全世界范围内、全中国范围内等等。预设IP协议版本中常见的IP地址通过IP协议版本被分为IPv4与IPv6两大类。IPV4就是有4段数字，每一段最大不超过255。IPv6采用128位地址长度。另外预设条件还可以包括预设IP地址类别等，IP地址编址方案将IP地址空间划分为A、B、C、D、E五类，其中A、B、C是基本类，D、E类作为多播和保留使用。本申请实施例对该预设条件不作具体限定，其可以为上述多种预设条件之一，也可以为多个预设条件的组合，因此可以根据实际情况进行实际限定。

[0025] 另外，需要说明的是，所述判断所述IP地址集合中的IP地址是否存活，包括：通过探测工具nmap进行判断所述IP地址集合中的IP地址是否存活。也就是通过探测工具对IP地址进行过滤，将未存活的IP地址进行过滤，减少后续判断的复杂性。探测工具nmap即为Network Mapper。其基本功能有三个，一是探测一组主机是否在线；其次是扫描主机端口，嗅探所提供的网络服务；还可以推断主机所用的操作系统，其探测IP地址是否存活可以进行ping扫描，打印出对扫描做出响应的主机等方式进行实现，该探测工具nmap的使用为现有技术，在此不做详细赘述。并且本申请实施例同样不对探测IP地址是否存活的工具进行限制，其只需能实施对IP地址的存活进行判断即可。

[0026] 在一个可选的方式中，如图2所示，在所述判断所述IP地址集合中的IP地址是否存活之前，步骤S101还包括：

[0027] 步骤S101a，获取CDN节点IP地址。

[0028] 其中，该获取过程具体为通过IPIP资料库或GeoV2资料库获取CDN节点IP地址。例如通过进入IPIP.NET后获取已登记的CDN节点IP地址。该获取的过程为现有技术，本申请实施例在此不做详细赘述。

[0029] 步骤S101b，判断所述IP地址集合中是否包含所述CDN节点IP地址，若是，将所述IP地址集合中与所述CDN节点IP地址相同的IP地址进行删除。

[0030] 通过对包含所述CDN节点IP地址进行过滤，减少后续步骤的复杂性。由此，通过步骤S101得到的第一IP地址集合中的元素为存活的并且不包含已知CDN节点IP地址的。

[0031] 步骤S102，通过HTTP请求访问所述第一IP地址集合中的IP地址，获取通过所述HTTP请求访问得到的第一网站页面。

[0032] 其中，获取第一网站页面的过程，包括：判断所述通过所述HTTP请求访问的返回页面是否为空，若是，不获取所述空页面。当通过HTTP请求访问成功后，返回的网站页面应该是一个正常的页面，但若HTTP请求访问失败，可能会出现返回的是一个空页面的情况，因此，需要将这些空页面进行筛除。

[0033] 步骤S103，获取待检测域名集合，并获取所述域名集合中的域名所对应的第二网站页面。

[0034] 待检测域名集合是根据实际情况而言需要进行判断其源站IP地址是什么的域名

集合，因此需要根据实际需求进行获取，通过在搜索引擎中输入域名后进入与该域名对应的网站页面，并对这些网站页面进行获取存储。

[0035] 步骤S104，计算所述第一网站页面与所述第二网站页面的相似度，获取所述相似度大于预设值的第一网站页面与所述第二网站页面。

[0036] 在一个可行的方式中，在步骤S102和步骤S103中获取的第一网站页面和第二网站页面可以存储为图片的形式，所述计算所述第一网站页面与所述第二网站页面的相似度，包括：通过PHASH算法计算所述第一网站页面与所述第二网站页面的相似度。

[0037] 其中，PHASH算法是指感知哈希算法(Perceptual Hash Algorithm)，感知哈希算法是一类算法的总称，包括aHash平均值哈希、pHash感知哈希、dHash差异值哈希。PHASH是将图片转换为明汉距离，用汉明距离进行图片相似度检测。

[0038] 但本申请实施例对第一网站页面与所述第二网站页面的相似度计算不作具体限定，例如，在另一个可行的方式中，在步骤S102和步骤S103中获取的第一网站页面和第二网站页面可以存储为词汇组合的形式，将第一网站页面和第二网站页面中的主体文字、文章等进行获取，并分别进行分词，得到一系列特征向量，然后通过simhash算法计算特征向量之间的距离(可以计算它们之间的欧氏距离、海明距离或者夹角余弦等等)，从而通过距离的大小来判断两个网页的相似度。

[0039] 示例性地，步骤S104中的相似度大于预设值，可将该预设值设置为80％。

[0040] 步骤S105，建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系，将已建立关系的IP地址作为所述域名的源站IP地址。

[0041] 因此，当一个第一网站页面与另一个第二网站页面的相似度大于预设值时，说明可以建立该第一网站页面对应的IP地址与该第二网站页面对应的域名之间的关系，即两者是一一对应的。由此即可以实现找到待检测域名的源站IP地址。

[0042] 综上所述，本申请实施例通过HTTP访问IP地址段，获取网站页面，将待判断的域名网页与所述网站页面进行相似度匹配，由此建立IP与域名之间的关系，实现找到待检测域名的源站IP地址。解决了现有技术中采用CDN技术后，原来用来获取访问源的IP地址的程序已不能正常使用，它拿到的并不是访问源的真实IP地址，而是CDN节点的IP地址的问题。

[0043] 参见图3，图3示例性示出了根据本申请实施例的一种获取源站IP地址的系统，所述系统300包括：第一IP地址集合获取模块301，用于获取IP地址集合，判断所述IP地址集合中的IP地址是否存活，若是，将所述存活的IP地址存储为第一IP地址集合；第一网站页面获取模块302，用于通过HTTP请求访问所述第一IP地址集合中的IP地址，获取通过所述HTTP请求访问得到的第一网站页面；第二网站页面获取模块303，用于获取待检测域名集合，并获取所述域名集合中的域名所对应的第二网站页面；相似度计算模块304，用于计算所述第一网站页面与所述第二网站页面的相似度，获取所述相似度大于预设值的第一网站页面与所述第二网站页面；源站IP地址获取模块305，用于建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系，将已建立关系的IP地址作为所述域名的源站IP地址。

[0044] 根据本申请的实施例的模块、子模块、单元、子单元中的任意多个、或其中任意多个的至少部分功能可以在一个模块中实现。根据本申请实施例的模块、子模块、单元、子单元中的任意一个或多个可以被拆分成多个模块来实现。

[0045] 图4示意性示出了根据本申请实施例的电子设备的框图。

[0046] 如图4所示，电子设备400包括处理器401和存储器402。该电子设备400可以执行根据本申请实施例的方法。

[0047] 具体地，处理器401例如可以包括通用微处理器、指令集处理器和/或相关芯片组和/或专用微处理器(例如，专用集成电路(ASIC))，等等。处理器401还可以包括用于缓存用途的板载存储器。处理器401可以是用于执行根据本申请实施例的方法流程的不同动作的单一处理单元或者是多个处理单元。

[0048] 存储器402，例如可以是能够包含、存储、传送、传播或传输指令的任意介质。例如，可读存储介质可以包括但不限于电、磁、光、电磁、红外或半导体系统、装置、器件或传播介质。可读存储介质的具体示例包括：磁存储装置，如磁带或硬盘(HDD)；光存储装置，如光盘(CD-ROM)；存储器，如随机存取存储器(RAM)或闪存；和/或有线/无线通信链路。其存储有计算机可执行程序，该程序在被所述处理器执行时，使得所述处理器执行如上文所述的直播间标签的添加方法。

[0049] 本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备/装置/系统中所包含的；也可以是单独存在，而未装配入该设备/装置/系统中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被执行时，实现根据本申请实施例的方法。

[0050] 根据本申请的实施例，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或者是上述两者的任意组合。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可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(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。在本申请中，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。而在本申请中，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。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、有线、光缆、射频信号等等，或者上述的任意合适的组合。

[0051] 本领域技术人员可以理解，本申请的各个实施例和/或权利要求中记载的特征可以进行多种组合或/或结合，即使这样的组合或结合没有明确记载于本申请中。特别地，在不脱离本申请精神和教导的情况下，本申请的各个实施例和/或权利要求中记载的特征可以进行多种组合和/或结合。所有这些组合和/或结合均落入本申请的范围。

[0052] 尽管已经参照本申请的特定示例性实施例示出并描述了本申请，但是本领域技术人员应该理解，在不背离所附权利要求及其等同物限定的本申请的精神和范围的情况下，可以对本申请进行形式和细节上的多种改变。因此，本申请的范围不应该限于上述实施例，

而是应该不仅由所附权利要求来进行确定,还由所附权利要求的等同物来进行限定。

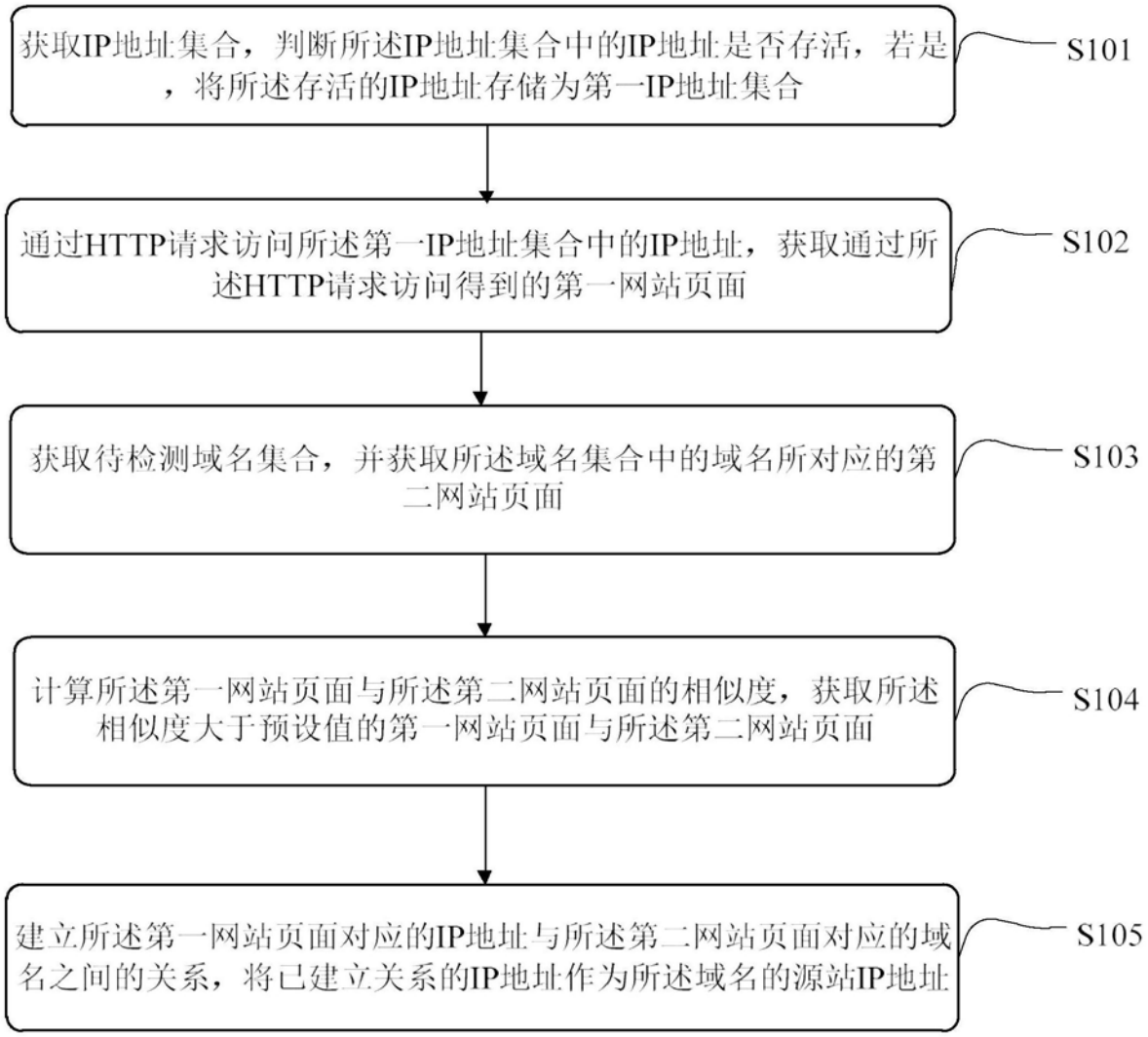

获取IP地址集合，判断所述IP地址集合中的IP地址是否存活，若是，将所述存活的IP地址存储为第一IP地址集合
S101
通过HTTP请求访问所述第一IP地址集合中的IP地址，获取通过所述HTTP请求访问得到的第一网站页面
S102
获取待检测域名集合，并获取所述域名集合中的域名所对应的第二网站页面
S103
计算所述第一网站页面与所述第二网站页面的相似度，获取所述相似度大于预设值的第一网站页面与所述第二网站页面
S104
建立所述第一网站页面对应的IP地址与所述第二网站页面对应的域名之间的关系，将已建立关系的IP地址作为所述域名的源站IP地址
S105

图1

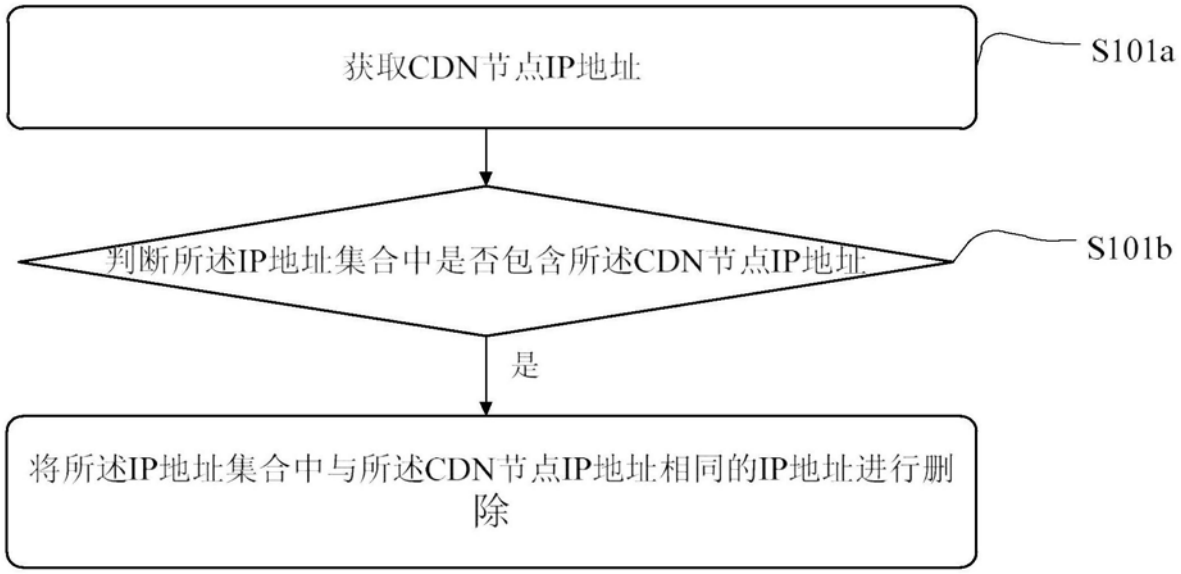

获取CDN节点IP地址
S101a
判断所述IP地址集合中是否包含所述CDN节点IP地址
S101b
是
将所述IP地址集合中与所述CDN节点IP地址相同的IP地址进行删除

图2

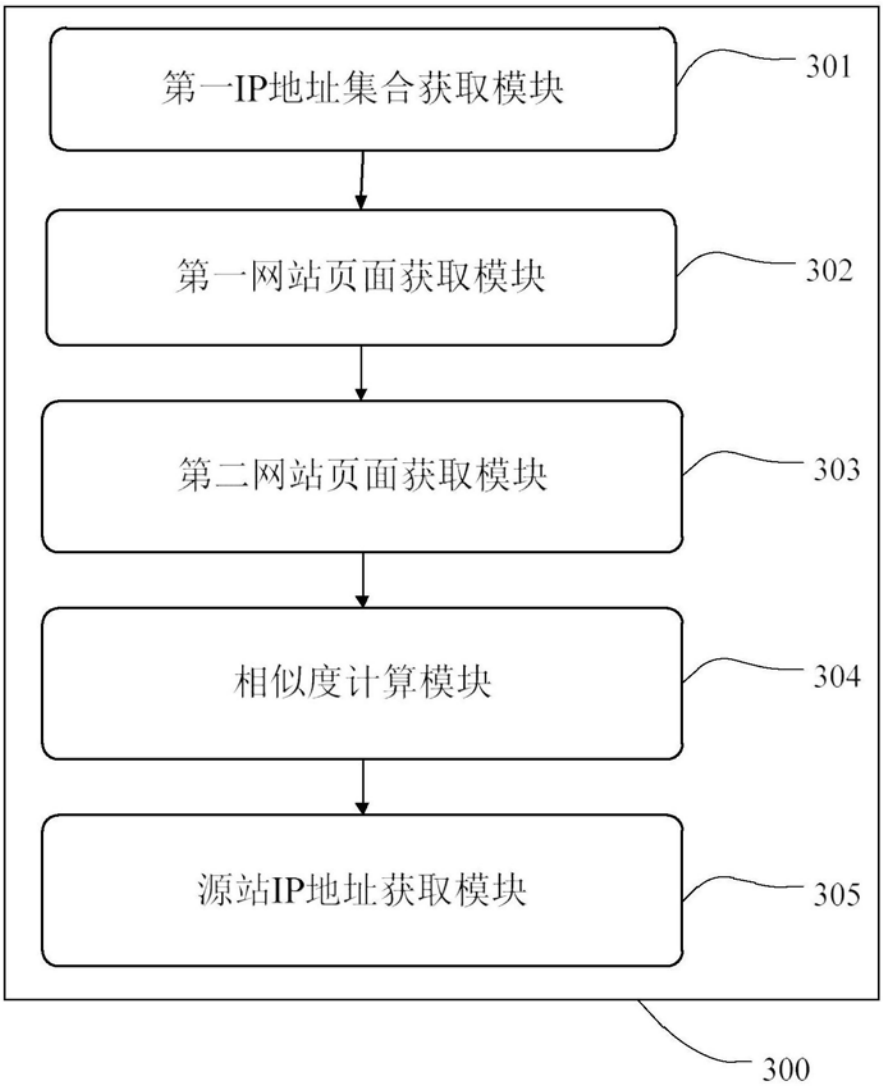

第一IP地址集合获取模块
301
第一网站页面获取模块
302
第二网站页面获取模块
303
相似度计算模块
304
源站IP地址获取模块
305
300

图3

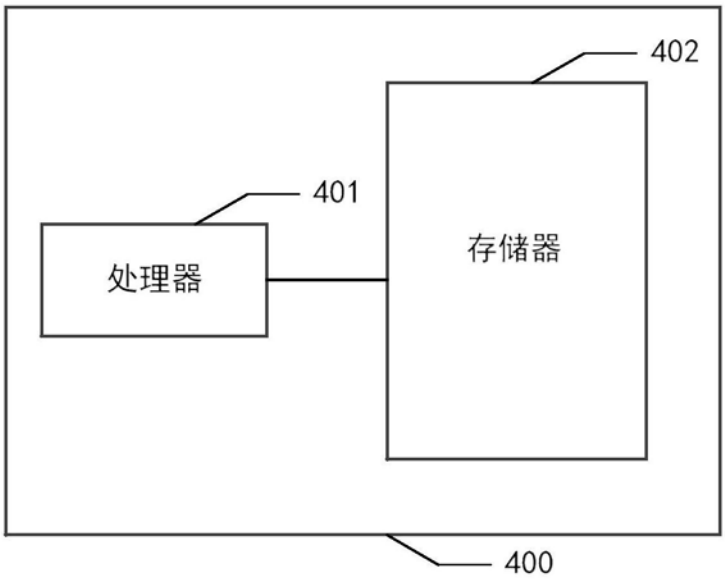


图4