

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 1 月 9 日 (09.01.2020)



(10) 国际公布号
WO 2020/006909 A1

(51) 国际专利分类号:
G06F 17/00 (2019.01)

越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(21) 国际申请号: PCT/CN2018/108715

(22) 国际申请日: 2018 年 9 月 29 日 (29.09.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810733716.0 2018 年 7 月 5 日 (05.07.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 熊庆昌(XIONG, Qingchang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司(SCIHEAD IP LAW FIRM); 中国广东省广州市

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

(54) Title: METHOD AND DEVICE FOR DEDUPLICATING URLS

(54) 发明名称: 一种 URL 去重方法及装置

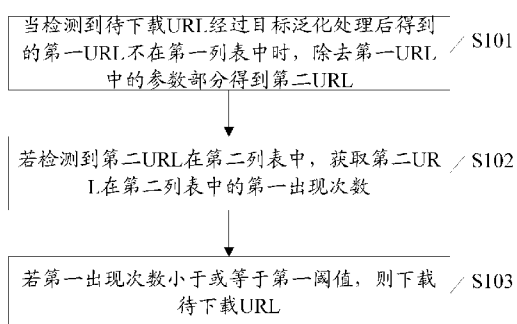


图 1

- S101 If it is detected that a first list does not contain a first URL obtained by performing target generalization processing on a URL to be downloaded, remove a parameter part from the first URL to obtain a second URL
- S102 If it is detected that a second list contains the second URL, acquire a first count of the second URL in the second list
- S103 If the first count is less than or equal to a first threshold, download the URL

(57) Abstract: Disclosed in an embodiment of the present invention are a method and device for deduplicating uniform resource locators (URLs). The method comprises: if it is detected that a first list does not contain a first URL obtained by performing target generalization processing on a URL to be downloaded, removing a parameter part from the first URL to obtain a second URL; if it is detected that a second list contains the second URL, acquiring a first count of the second URL in the second list; and if the first count is less than or equal to a first threshold, downloading the URL. The embodiment of the present invention enables reduction of duplicate URLs to be downloaded, and ensures deduplication accuracy while enhancing scanning efficiency of a web vulnerability scanning system.

(57) 摘要: 本申请实施例公开了一种统一资源定位符 URL 去重及装置, 其中方法包括: 在检测到待下载 URL 经过目标泛化处理后的第一 URL 不在第一列表中时, 除去该第一 URL 中的参数部分得到第二 URL, 又在检测到该第二 URL 在该第二列表中, 获取该第二 URL 在该第二列表中的第一出现次数, 当该第一出现次数小于或等于第一阈值, 就下载该待下载 URL。采用本申请实施例, 可以减少下载到重复 URL 的数量, 在保证去重准确率的同时提高 WEB 漏洞扫描系统的扫描效率。

本国际公布：

- 包括国际检索报告(条约第21条(3))。

一种 URL 去重方法及装置

本申请要求于 2018 年 7 月 5 日提交中国专利局、申请号为 2018107337160、申请名称为“一种统一资源定位符 URL 去重方法及装置”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。

技术领域

本申请涉及互联网技术领域，尤其涉及一种 URL 去重方法及装置。

背景技术

目前业内很多万维网（world wide web，WEB）漏洞扫描系统，都有自己的统一资源定位符（uniform resource locator，URL）爬虫，也有自己的 URL 去重方法。URL 爬虫指爬虫系统首先从互联网页面中精心选择一部分网页，以这些网页的链接地址作为种子 URL，将这些种子放入待抓取 URL 队列中，爬虫从待抓取 URL 队列依次读取，并将 URL 通过域名系统（domain name system，DNS）解析，把链接地址转换为网站服务器对应的 IP 地址。然后将其和网页相对路径名称交给网页下载器，网页下载器负责页面的下载。对于下载到本地的网页，一方面将其存储到页面库中，等待建立索引等后续处理；另一方面将下载网页的 URL 放入已抓取队列中，这个队列记录了爬虫系统已经下载过的网页 URL，以避免系统的重复抓取。

URL 去重指的是将重复抓取的 URL 去除，避免多次抓取同一网页。例如，对每一个给定的 URL，将其映射到某个物理地址上。当需要检测该给定 URL 是否重复时，只需判断该给定 URL 对应的物理地址是否已经存在，若存在，说明已经被下载过，则放弃下载，否则将该给定 URL 放入待抓取队列，等待下载。由于一个网站中的很多 URL 仅是参数部分不同，而这些仅是参数部分不同的 URL 很大可能已经被下载过，但这些 URL 经过映射后得到的物理地址是不相同的，这时仅通过 URL 对应的物理地址来判断这些 URL 是否重复将导致爬取到很多重复的 URL，影响 WEB 漏洞扫描系统的扫描效率。

发明内容

本申请实施例提供一种 URL 去重方法及装置，可减少下载到重复 URL 的数量，在保证去重准确率的同时提高 WEB 漏洞扫描系统的扫描效率。

本申请实施例提供了一种 URL 去重方法，该方法包括：

当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去该第一 URL 中的参数部分得到第二 URL，该目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符，该第一列表用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL；

若检测到该第二 URL 在该第二列表中，获取该第二 URL 在该第二列表中的第一出现次数，该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL；

若该第一出现次数小于或等于第一阈值，则下载该待下载 URL。

本申请实施例通过对 URL 进行分级去重，可以减少下载到重复 URL 的数量，在保证去重准确率的同时提高 WEB 漏洞扫描系统的扫描效率。

附图说明

- 图 1 是本申请实施例提供的 URL 去重方法的一示意图；
图 2a 是第一列表和出现次数的关系示意图；
图 2b 是第二列表和出现次数的关系示意图；
图 3 是本申请实施例提供的 URL 去重方法的另一示意图；
图 4 是本申请实施例提供的 URL 去重装置的一示意性框图；
图 5 是本申请实施例提供的终端的一示意性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都在本申请保护的范围。

在介绍本申请实施例之前，先介绍 URL 的数据结构。一个 URL 的结构通常可以为“协议://服务器名称(IP地址)/路径?参数”，例如：
`http://xxx.pingan.com/cgi-bin/index1.html?param1=value1¶m2=value2`，其中 `param1=value1¶m2=value2` 表示这个 URL 的参数字段(参数部分)，URL 的参数字段由参数名和参数值组成，`param1` 和 `param2` 表示参数名，`value1` 和 `value2` 表示参数值，参数值可为数字、字母(包括大小写)、特殊字符(指除数字、字母之外的字符)和/或它们的组合。问号“?”字符用于分隔 URL 中的路径和参数，逻辑与“&”字符表示 URL 中指定的参数间的分隔符。

下面将结合图 1 至图 5，对本申请实施例提供的 URL 去重方法及装置进行说明。

参见图 1，是本申请实施例提供的 URL 去重方法的一示意图。如图 1 所示，该 URL 去重方法可包括步骤：

S101，当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去第一 URL 中的参数部分得到第二 URL。

在一些可行的实施方式中，终端可以将爬取到的待下载 URL 的参数字段按照参数名进行排序，并对该待下载 URL 参数字段的参数值进行目标泛化处理，得到第一 URL。终端可以检测第一列表中是否存在与该第一 URL 相同的 URL，若不存在，说明该第一 URL 不在该第一列表中，则除去该第一 URL 中的参数部分得到第二 URL。其中，该目标泛化处理可以用于将 URL 中相同类型的多个连续字符替换为单个字符。例如，可将连续的数字如 145 替换为数字 1，连续的字母如 FK、aj、dgA 都替换为字母 A，特殊字符都替换为百分号%。该第一列表可以用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL。特殊字符指除数字和字母外的字符，如问号“?”、感叹号“!”等。因为经过目标泛化处理得到的第一 URL 只减少了待下载 URL 参数值中的部分变量，所以当第一 URL 不在第一列表中时，说明待下载 URL 未被下载过，终端再进行进一步去重处理，保证了去重的准确率。

例如，终端爬取到的待下载 URL 为 `http://xxx.pingan.com/cgi-bin/index1.html?param1=v167!ABD¶m2=val_ue2`，终端可以将待下载 URL 的参数部分按照参数名进行

排序,以保证待下载 URL 的参数部分顺序排列,再将待下载 URL 的参数值中连续的数字替换为预设的单个数字“1”、连续的字母替换为预设的单个字母“A”、特殊字符替换为预设的单个特殊字符“%”,得到 URL1 (第一 URL) 为 `http://xxx.pingan.com/cgi-bin/index1.html?param1=v1%A¶m2=A%A2`。假设第一列表中包括 URL4、URL6 以及 URL7 这三个 URL,终端检测第一列表(URL4、URL6 以及 URL7)中是否存在与 URL1 相同的 URL,即第一列表中是否存在 URL1,因为第一列表中不存在 URL1,说明 URL1 不在第一列表中,终端就删除 URL1 中的参数字段得到第二 URL 为 `http://xxx.pingan.com/cgi-bin/index1.html`。

在一些可行的实施方式中,终端可以将爬取到的待下载 URL 的参数字段按照参数名进行排序,并对该待下载 URL 参数字段的参数值进行目标泛化处理,得到第一 URL。终端可以利用消息-摘要算法 5 (message-digest algorithm 5, MD5) 计算上述第一 URL 的哈希(hash)值,并可以检测该第一 URL 的 hash 值是否在该第一列表中。若该第一 URL 的 hash 值不在第一列表中,终端可以除去该第一 URL 中的参数部分得到第二 URL。其中,该目标泛化处理可以用于将 URL 中相同类型的多个连续字符替换为单个字符。例如,可将连续的数字如 145 替换为数字 1,连续的字母如 FK、aj、dgA 都替换为字母 A,特殊字符都替换为百分号%。该第一列表可以用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL,再经过 MD5 计算后得到的 hash 值。特殊字符指除数字和字母外的字符,如问号“?”、感叹号“!”等。因为 hash 函数是将任意大小的数据转换成特定大小的数据的函数,且已下载列表(第一列表)中存储的是 URL 的 hash 值,而不是完整的 URL,可以减少存储空间(因为完整 URL 字符较多,而 hash 值是固定大小的数据)。

例如,终端爬取到的待下载 URL 为 `http://xxx.pingan.com/cgi-bin/index1.html?param1=v167!ABD¶m2=val_ue2`,终端可以将待下载 URL 的参数部分按照参数名进行排序,以保证待下载 URL 的参数部分顺序排列,再将待下载 URL 的参数值中连续的数字替换为预设的单个数字“1”、连续的字母替换为预设的单个字母“A”、特殊字符替换为预设的单个特殊字符“%”,得到 URL1 (第一 URL) 为 `http://xxx.pingan.com/cgi-bin/index1.html?param1=v1%A¶m2=A%A2`。终端可以利用 MD5 计算 URL1 的 hash 值为 03,假设第一列表中包括 hash 值 01、04 以及 07,终端检测 URL1 的 hash 值 03 是否在第一列表(01、04 以及 07)中,因为第一列表中不存在 hash 值 03,终端就删除 URL1 中的参数字段得到第二 URL 为 `http://xxx.pingan.com/cgi-bin/index1.html`。

在一些可行的实施方式中,终端可以将爬取到的待下载 URL 的参数字段按照参数名进行排序,并对该待下载 URL 参数字段的参数值进行目标泛化处理,得到第一 URL。终端可以检测第一列表中是否存在与该第一 URL1 相同的 URL,若存在,说明该第一 URL 在该第一列表中,终端可以将该第一 URL 在该第一列表中的出现次数加 1,得到该第一 URL 在该第一列表中的第二出现次数。终端可以比较该第二出现次数与预设的第二阈值之间的大小关系,若该第二出现次数小于或等于该第二阈值,则终端可以下载该待下载 URL。若该第二出现次数大于该第二阈值,则终端可以丢弃该待下载 URL,即不下载该待下载 URL。例如,如图 2a 所示,是第一列表与出现次数的关系示意图。其中,假设第一列表中的元素

为 URL1、URL4 以及 URL7, 第一 URL 为 URL1, 第二阈值为 7。终端检测第一列表(URL1、URL4 以及 URL7) 中是否存在与 URL1 相同的 URL, 即第一列表中是否存在 URL1, 因为第一列表中存在 URL1, 终端就将 URL1 在第一列表中的出现次数 1 加 1, 得到 URL1 在第一列表中的第二出现次数 2。因为第二出现次数 2 小于第二阈值 7, 故终端可以下载该待下载 URL。

在一些可行的实施方式中, 若上述待下载 URL 为某个网站中爬取到的第一个 URL, 那么第一列表可以为空(因为此时没有已下载的 URL)。

S102, 若检测到第二 URL 在第二列表中, 获取第二 URL 在第二列表中的第一出现次数。

在一些可行的实施方式中, 终端可以检测上述第二 URL 是否在第二列表中, 若是, 则可以将该第二 URL 在该第二列表中的出现次数加 1, 得到该第二 URL 在该第二列表中的第一出现次数。若不是, 即上述第二 URL 不在第二列表中, 终端可以直接下载上述待下载 URL。其中, 该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL。

在一些可行的实施方式中, 终端可以利用 MD5 计算上述第二 URL 的 hash 值, 并可以检测该第二 URL 的 hash 值是否在该第二列表中, 若是, 则可以将该第二 URL 的 hash 值在该第二列表中的出现次数加 1, 得到该第二 URL 的 hash 值在该第二列表中的第一出现次数。若不是, 即上述第二 URL 的 hash 值不在第二列表中, 终端可以直接下载上述待下载 URL。

例如, 如图 2b 所示, 是第二列表与出现次数的关系示意图。其中, 假设第二列表中的元素为 02 和 06, 第二 URL 为 URL2, 终端计算 URL2 的 hash 值为 02, 终端检测到 URL2 的 hash 值 02 在第一列表中, 终端将 URL2 的 hash 值 02 在第一列表中的出现次数 2 加 1, 得到 URL2 的 hash 值 02 在第二列表中的第一出现次数 3。

在一些可行的实施方式中, 终端在检测出该第二 URL 不在该第二列表中时, 终端可以将该第二 URL 加入该第二列表以形成新的第二列表, 即更新该第二列表, 同时可以将该第二 URL 在该第二列表中的出现次数置为 1, 并可以下载上述待下载 URL。终端在下载上述待下载 URL 或丢弃上述待下载 URL 之后, 终端利用最新的第二列表检测下一个待下载 URL 是否被下载过。终端通过不断更新第二列表, 可以更准确地过滤掉重复的 URL, 并且可以进一步提高扫描效率。

在一些可行的实施方式中, 若上述待下载 URL 为某个网站中爬取到的第一个 URL, 那么第二列表可以为空(因为此时没有已下载的 URL)。

S103, 若第一出现次数小于或等于第一阈值, 则下载待下载 URL。

在一些可行的实施方式中, 终端可以比较上述获取到的第一出现次数与预设的第一阈值之间的大小关系。若该第一出现次数小于或等于该第一阈值, 则终端可以直接下载上述待下载 URL。若该第一出现次数大于该第一阈值, 则终端可以直接丢弃上述待下载 URL, 即不下载该待下载 URL。其中, 该第一阈值可以小于上述第二阈值, 该第一阈值可以为大于或等于 0 的整数。终端通过判断除去第一 URL 中的参数部分得到的第二 URL 是否已经被下载过, 从而判断出是否需要下载待下载 URL, 可以过滤掉参数部分不相同的待下载 URL, 即减少了待下载 URL 中的变量, 丢弃的待下载 URL 多, 进而减少了下载到重复 URL 的数量, 在保证准确率的同时进一步提高了扫描效率, 处理效率更高。

本申请实施例通过在检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时,除去该第一 URL 中的参数部分得到第二 URL,该目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符,该第一列表用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL,又在检测到该第二 URL 在该第二列表中,获取该第二 URL 在该第二列表中的第一出现次数,该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL,当该第一出现次数小于或等于第一阈值,就下载该待下载 URL,可以减少下载到重复 URL 的数量,在保证去重准确率的同时提高 WEB 漏洞扫描系统的扫描效率。

参见图 3,是本申请实施例提供的 URL 去重方法的另一示意流程图。如图 3 所示,该 URL 去重方法可包括步骤:

S301,对待下载 URL 进行第一泛化处理得到第一泛化处理后的 URL1。

S302,当第一 URL 不在第一列表中时,除去第一 URL 参数部分中的参数值得到除去参数值的 URL2。

在一些可行的实施方式中,为便于描述,URL1 可以表示本申请实施例中的第一 URL。其中 URL1、URL2 表示经过不同处理后的 URL。终端可以将待下载 URL 的参数部分按照参数名进行排序,并对该待下载 URL 参数部分的参数值进行第一泛化处理,得到 URL1。终端可以检测第一列表中是否存在与上述 URL1 相同的 URL,若不存在,说明上述 URL1 不在第一列表,则可以删除 URL1 参数部分中的参数值得到 URL2。若存在,说明该 URL1 在该第一列表中,则终端可以将该 URL1 在该第一列表中的出现次数加 1,得到该 URL1 在该第一列表中的第二出现次数,并可以判断该第二出现次数是否小于或等于第二阈值,若是(即该第二出现次数小于或等于该第二阈值),则下载上述待下载 URL,若否(即该第二出现次数大于该第二阈值),则丢弃上述待下载 URL,即不下载该待下载 URL。终端通过判断经过第一泛化处理得到的 URL1 是否已经被下载过,从而判断出是否需要下载待下载 URL,可以过滤掉参数值中具有相同数据格式的待下载 URL,减少了下载到重复 URL 的数量,又因为只减少了待下载 URL 参数值中的部分变量,所以保证了去重的准确率。其中,该第一泛化处理用于将该待下载 URL 中相同类型的多个连续字符替换为单个字符。例如,可将连续的数字如 145 替换为数字 1,连续的字母如 FK、aj、dgA 都替换为字母 A,特殊字符都替换为符号%。其中,特殊字符指除数字和字母外的字符,如问号“?”、感叹号“!”等。该第一列表可以用于存储历史记录中已下载的 URL 经过上述第一泛化处理后得到的 URL。

在一些可行的实施方式中,终端在得到上述 URL1 之后,终端可以利用 MD5 计算上述 URL1 的 hash 值,并可以检测该 URL1 的 hash 值是否在该第一列表中,若是,则可以删除该 URL1 的参数部分中的参数值得到 URL2。若否,可以将该 URL1 的 hash 值在该第一列表中的出现次数加 1,得到该 URL1 的 hash 值在该第一列表中的第二出现次数,并可以判断该第二出现次数是否小于或等于第二阈值,若是(即该第二出现次数小于或等于该第二阈值),则下载上述待下载 URL,若否(即该第二出现次数大于该第二阈值),则丢弃上述待下载 URL,即不下载该待下载 URL。其中,第一列表可以用于存储历史记录已下载的

URL 经过上述第一泛化处理后得到的 URL 的哈希值。该第二阈值为大于 0 的整数。因为 hash 函数是将任意大小的数据转换成特定大小的数据的函数，且已下载列表（第一列表）中存储的是 URL 的 hash 值，而不是完整的 URL，可以减少存储空间（因为完整 URL 字符较多，而 hash 值是固定大小的数据），同时在检测 URL1 是否在第一列表中时，可以提高处理效率。

在一些可行的实施方式中，终端在检测出该 URL1 不在该第一列表中时，终端可以将该 URL1 加入该第一列表以形成新的第一列表，即更新该第一列表，同时可以将该 URL1 在该第一列表中的出现次数置为 1，并可以删除该 URL1 的参数部分中的参数值得到 URL2。终端在下载上述待下载 URL 或丢弃上述待下载 URL 之后，终端利用最新的第一列表检测下一个待下载 URL 是否被下载过。终端通过不断更新第一列表，可以更准确地过滤掉重复的 URL，并且可以进一步提高扫描效率。例如，第一列表包括 URL5、URL6 以及 URL7，此时，URL1 不在第一列表中，终端可以将该 URL1 加入第一列表，此时更新后的第一列表包括 URL1、URL5、URL6 以及 URL7。同时，终端也可以将 URL1 在第一列表中的出现次数也置为 1。

S303，当检测到 URL2 在第二列表时，检测 URL2 在第二列表中的第一出现次数是否小于或等于第一阈值，若是，则下载待下载 URL。

S304，当检测到 URL2 不在第二列表中时，除去 URL2 中的参数部分得到除去参数部分的 URL3。

在一些可行的实施方式中，为便于描述，URL3 可以表示本申请实施例中的第二 URL。其中 URL2、URL3 表示经过不同处理后的 URL。终端可以检测第二列表中是否存在与上述 URL2 相同的 URL，若存在，说明该 URL2 在该第二列表中，则可以将该 URL2 在该第二列表中的出现次数加 1，得到该 URL2 在该第二列表中的第一出现次数，并可以检测该第一出现次数是否小于或等于第一阈值，若是（即该第一出现次数小于或等于该第一阈值），则下载上述待下载 URL，若否（即第一出现次数大于该第一阈值），则丢弃上述待下载 URL，即不下载该待下载 URL。若终端检测到该第二列表中不存在该 URL2，说明该 URL2 不在该第二列表中，则可以删除该 URL2 中的参数部分得到 URL3。其中，第二列表可以用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL。终端通过判断除去 URL1 的参数部分中的参数值得到的 URL2 是否已经被下载过，从而判断出是否需要下载待下载 URL，可以过滤掉参数值不相同的待下载 URL，进一步减少了下载到重复 URL 的数量，在保证准确率的同时提高了扫描效率。需要说明的是，该第一阈值小于或等于上述第二阈值，该第一阈值可以为大于 0 的整数。因为第一泛化处理仅是减少了待下载 URL 的参数值中的变量，此时 URL1 中的变量还较多，过滤掉的 URL 就少，下载的 URL 就较多，URL1 在第一列表中出现的次数就多。因此第一阈值小于第二阈值可以保证步骤 S301-步骤 S302 未过滤掉的 URL，除去参数部分中的参数值之后可以被过滤掉，达到分级去重的目的。

在一些可行的实施方式中，终端可以利用 MD5 计算上述 URL2 的 hash 值，并可以检测该 URL2 的 hash 值是否在该第二列表中，若是，可以将该 URL2 的 hash 值在该第二列表中的出现次数加 1，得到该 URL2 的 hash 值在该第二列表中的第一出现次数，并可以判断该第一出现次数是否小于或等于第一阈值，若是（即该第一出现次数小于或等于该第一阈

值), 则下载上述待下载 URL, 若否 (即该第一出现次数大于该第一阈值), 则丢弃上述待下载 URL, 即不下载该待下载 URL。若否, 则可以删除 URL2 中的参数部分 (包括参数值和参数名) 得到 URL3。其中, 第二列表可以用于存储历史记录中已下载的 URL 除去参数值后得到的 URL, 再经过 MD5 计算后得到的 hash 值。因为 hash 函数是将任意大小的数据转换成特定大小的数据的函数, 且已下载列表 (第二列表) 中存储的是 URL 的 hash 值, 而不是完整的 URL, 可以进一步减少存储空间, 同时在检测 URL2 是否在第二列表时, 可以进一步提高处理效率。

例如, 假设第二列表中的 hash 值为 07 和 09, URL2 为: <http://xxx.pingan.com/cgi-bin/index1.html?param1¶m2>, 终端计算 URL2 的 hash 值为 04, 终端检测到 URL2 的 hash 值 04 不在第二列表中, 终端就直接删除 URL2 中的参数部分得到 URL3, URL3 为: <http://xxx.pingan.com/cgi-bin/index1.html>。

在一些可行的实施方式中, 终端在检测出该 URL2 不在该第二列表中时, 终端可以将该 URL2 加入该第二列表形成新的第二列表, 即更新该第二列表, 同时可以将该 URL2 在该第二列表中的出现次数置为 1, 并可以除去该 URL2 中的参数部分得到 URL3。终端在下载上述待下载 URL 或丢弃上述待下载 URL 之后, 终端利用最新的第二列表检测下一个待下载 URL 是否被下载过。终端通过不断更新第二列表, 可以更准确地过滤掉重复的 URL, 并且可以进一步提高扫描效率。

在一些可行的实施方式中, 若上述待下载 URL 为某个网站中爬取到的第一个 URL, 那么第二列表可以为空 (因为此时没有已下载的 URL)。

S305, 当检测到 URL3 在第三列表中时, 检测 URL3 在第三列表中的第三出现次数是否小于或等于第三阈值, 若是, 则下载待下载 URL。

S306, 当检测到 URL3 不在第三列表中时, 对 URL3 进行第二泛化处理得到第二泛化处理后的 URL4。

在一些可行的实施方式中, 为便于描述, URL3 可以表示本申请实施例中的第二 URL。其中 URL3、URL4 表示经过不同处理后的 URL。终端可以检测第三列表中是否存在与上述 URL3 相同的 URL, 若存在, 说明该 URL3 在该第三列表中, 则可以将该 URL3 在该第三列表中的出现次数加 1, 得到该 URL3 在该第三列表中的第三出现次数, 并可以检测该第三出现次数是否小于或等于第三阈值, 若是 (即该第三出现次数小于或等于该第三阈值), 则下载上述待下载 URL, 若否 (即第三出现次数大于该第三阈值), 则丢弃上述待下载 URL, 即不下载该待下载 URL。若终端检测到该第三列表中不存在该 URL3, 说明该 URL3 不在该第三列表中, 则可以对该 URL3 的路径部分进行第二泛化处理得到 URL4。该第二泛化处理用于将 URL3 的路径部分中目标类型的至少一个字符替换为目标字符, 如将该 URL3 的路径部分中的一个或多个数字替换为预设的单个数字 “1”。其中, 第三列表可以用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL。终端通过判断除去 URL2 参数部分得到的 URL3 是否已经被下载过, 从而判断出是否需要下载待下载 URL, 可以过滤掉参数部分不相同的待下载 URL, 即减少了待下载 URL 中的变量, 丢弃的待下载 URL 多, 进而减少了下载到重复 URL 的数量, 在保证准确率的同时进一步提高了扫描效率。需要说明的是, 该第三阈值可以小于或等于上述第一阈值, 上述第一阈值小于或等于上述第二阈

值,该第三阈值可以为大于0的整数。因为除去参数部分中的参数值仅是减少了待下载URL参数部分的变量,此时URL2中的变量还较多,过滤掉的URL就较少,下载的URL就较多,URL2在第二列表中出现的次数就较多。因此第三阈值小于第一阈值可以保证步骤S303-步骤S304未过滤掉的URL,除去参数部分后可以被过滤掉,达到分级去重的目的。

在一些可行的实施方式中,终端可以利用MD5计算上述URL3的hash值,并可以检测该URL3的hash值是否在该第三列表中,若是,可以将该URL3的hash值在该第三列表中的出现次数加1,得到该URL3的hash值在该第三列表中的第三出现次数,并可以判断该第三出现次数是否小于或等于第三阈值,若是(即该第三出现次数小于或等于该第三阈值),则下载上述待下载URL,若否(即该第三出现次数大于该第三阈值),则丢弃上述待下载URL,即不下载该待下载URL。若否,则可以对该URL3进行第二泛化处理得到URL4。该第二泛化处理用于将URL3的路径部分中目标类型的至少一个字符替换为目标字符,如将该URL3的路径部分中的一个或多个数字替换为预设的单个数字“1”。其中,第三列表可以用于存储历史记录中已下载的URL除去参数部分后得到的URL,再经过MD5计算后得到的hash值。因为hash函数是将任意大小的数据转换成特定大小的数据的函数,且已下载列表(第三列表)中存储的是URL的hash值,而不是完整的URL,可以进一步减少存储空间,同时在检测URL3是否在第三列表中时,可以进一步提高处理效率。

在一些可行的实施方式中,终端在检测出该URL3不在该第三列表中时,终端可以将该URL3加入该第三列表以形成新的第三列表,即更新该第三列表,同时可以将该URL3在该第三列表中的出现次数置为1,并可以对该URL3进行第二泛化处理得到URL4。终端在下载上述待下载URL或丢弃上述待下载URL之后,终端利用最新的第三列表检测下一个待下载URL是否被下载过。终端通过不断更新第三列表,可以更准确地分级过滤掉重复的URL,并且可以进一步提高扫描效率。

在一些可行的实施方式中,若上述待下载URL为某个网站中爬取到的第一个URL,那么第三列表可以为空(因为此时没有已下载的URL)。

S307,当检测到URL4在第四列表中时,检测URL4在第四列表中的第四出现次数是否小于或等于第四阈值,若是,则下载待下载URL。

S308,当检测到URL4不在第四列表中时,下载待下载URL。

在一些可行的实施方式中,终端可以检测第四列表中是否存在与上述URL4相同的URL,若存在,说明该URL4在该第四列表中,则可以将该URL4在该第四列表中的出现次数加1,得到该URL4在该第四列表中的第四出现次数,并可以检测该第四出现次数是否小于或等于第四阈值,若是(即该第四出现次数小于或等于该第四阈值),则下载上述待下载URL,若否(即第四出现次数大于该第四阈值),则丢弃上述待下载URL,即不下载该待下载URL。若终端检测到该第四列表中不存在该URL4,说明该URL4不在该第四列表中,则可以直接下载上述待下载URL。其中,第四列表可以用于存储历史记录中已下载URL除去参数部分并经过上述第二泛化处理后得到的URL。终端通过判断除去参数部分且经过第二泛化处理后得到的URL4是否已经被下载过,从而判断出是否需要下载待下载URL,可以过滤掉路径部分不同的待下载URL,减少了待下载URL中的变量,进而丢弃的待下载URL多,下载到的重复URL少,在保证准确率的同时进一步提高了扫描效率。

需要说明的是，该第四阈值小于或等于上述第三阈值，上述第三阈值可以小于或等于上述第一阈值，上述第一阈值小于或等于上述第二阈值，该第四阈值为大于或等于 0 的整数。第四阈值小于第三阈值可以保证步骤 S305-步骤 S306 未过滤掉的 URL，经过第二泛化处理后可以被过滤掉，达到分级去重的目的。

在一些可行的实施方式中，终端可以利用 MD5 计算上述 URL4 的 hash 值，并可以检测该 URL4 的 hash 值是否在该第四列表中，若是，可以将该 URL4 的 hash 值在该第四列表中的出现次数加 1，得到该 URL4 的 hash 值在该第四列表中的第四出现次数，并可以判断该第四出现次数是否小于或等于第四阈值，若是（即该第四出现次数小于或等于该第四阈值），则下载上述待下载 URL，若否（即该第四出现次数大于该第四阈值），则丢弃上述待下载 URL，即不下载该待下载 URL。若否，则可以直接下载上述待下载 URL。其中，第四列表可以用于存储历史记录中已下载的 URL 除去参数部分且经过上述第二泛化处理后得到的 URL，再经过 MD5 计算后得到的 hash 值。因为 hash 函数是将任意大小的数据转换成特定大小的数据的函数，且已下载列表（第四列表）中存储的是 URL 的 hash 值，而不是完整的 URL，可以进一步减少存储空间，同时在检测 URL4 是否在第四列表时，可以进一步提高处理效率。

在一些可行的实施方式中，终端在检测出该 URL4 不在该第四列表中时，终端可以将该 URL4 加入该第四列表以形成新的第四列表，即更新该第四列表，同时可以将 URL4 在第四列表中的出现次数置为 1，并可以下载上述待下载 URL。终端在下载上述待下载 URL 或丢弃上述待下载 URL 之后，终端利用最新的第四列表检测下一个待下载 URL 是否被下载过。终端通过不断更新第四列表，可以更准确地过滤掉重复的 URL，并且可以进一步提高扫描效率。

在一些可行的实施方式中，若上述待下载 URL 为某个网站中爬取到的第一个 URL，那么第四列表可以为空（因为此时没有已下载的 URL）。

本申请实施例通过对待下载 URL 进行分级去重，逐级减少待下载 URL 中的变量，并根据逐级减少变量后的 URL 判断待下载 URL 是否已下载，若在该级就判断出待下载 URL 已下载，则丢弃待下载 URL，若该级就确定待下载 URL 未下载，则下载待下载 URL，若该级无法判断出待下载 URL 是否已下载，则进入下一级判断，直至判断出待下载 URL 已下载或未下载。通过更细化的分级去重方案，不仅能够减少下载到重复 URL 的数量，提高 WEB 漏洞扫描系统的扫描效率，还可以提高去重的准确率。

参见图 4，是本申请实施例提供的 URL 去重装置的一示意性框图。本申请实施例提供的 URL 去重装置包括：

处理模块 10，用于当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去该第一 URL 中的参数部分得到第二 URL。其中，该目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符，该第一列表用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL。

第一获取模块 20，用于当检测到该第二 URL 在该第二列表中时，获取该第二 URL 在该第二列表中的第一出现次数。其中，该第二列表用于存储历史记录中已下载的 URL 除去

参数部分后得到的 URL。

下载模块 30，用于当该第一出现次数小于或等于第一阈值时，则下载该待下载 URL。
在一些可行的实施方式中，该装置还包括：

第二获取模块 40，用于当检测到该待下载 URL 经过该目标泛化处理后得到的该第一 URL 在该第一列表中时，获取该第一 URL 在该第一列表中的第二出现次数；上述下载模块 30，还用于当该第二出现次数小于或等于第二阈值时，下载该待下载 URL。其中，该第二阈值大于或等于该第一阈值，该第一阈值为大于或等于 0 的整数。

在一些可行的实施方式中，上述处理模块 10 具体用于：对该待下载 URL 进行目标泛化处理得到第一 URL；利用目标哈希函数计算该第一 URL 的哈希值；检测该第一 URL 的哈希值是否在第一列表中；当该第一 URL 的哈希值不在该第一列表中时，除去该第一 URL 中的参数部分得到第二 URL。其中，该第一列表用于存储历史记录中已下载的 URL 经过目标泛化处理后得到的 URL 的哈希值。

在一些可行的实施方式中，该第一 URL 包括协议、服务器名称、路径以及参数字段；上述处理模块具体用于当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，删除该第一 URL 中的参数字段得到该第二 URL。

在一些可行的实施方式中，上述第一获取模块 20 具体用于：利用目标哈希函数计算该第二 URL 的哈希值；检测该第二 URL 的哈希值是否在该第二列表中；若该第二 URL 的哈希值在该第二列表中，获取该第二 URL 的哈希值在该第二列表中的该第一出现次数。其中，该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL 的哈希值。

在一些可行的实施方式中，上述下载模块 30 还用于当该第二 URL 不在该第二列表中时，下载该待下载 URL。

具体实现中，上述 URL 去重装置可通过上述各个模块执行上述图 1 或者图 3 所提供的实现方式中各个步骤所提供的实现方式，实现上述各实施例中所实现的功能，具体可参见上述图 1 或图 3 所示的方法实施例中各个步骤提供的相应描述，在此不再赘述。

本申请实施例中，URL 去重装置可通过在检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去该第一 URL 中的参数部分得到第二 URL，该目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符，该第一列表用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL，又在检测到该第二 URL 在该第二列表中，获取该第二 URL 在该第二列表中的第一出现次数，该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL，当该第一出现次数小于或等于第一阈值，就下载该待下载 URL，可以减少下载到重复 URL 的数量，在保证去重准确率的同时提高 WEB 漏洞扫描系统的扫描效率。

参见图 5，是本申请实施例提供的一种终端的示意性框图。如图 5 所示，本申请实施例中的终端可以包括：一个或多个处理器 501 和存储器 502。上述处理器 501 和存储器 502 通过总线 503 连接。存储器 502 用于存储计算机程序，所述计算机程序包括程序指令，处理器 501 用于执行存储器 502 存储的程序指令。其中，处理器 501 被配置用于调用该程序指令执行：

当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时, 除去该第一 URL 中的参数部分得到第二 URL, 该目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符, 该第一列表用于存储历史记录中已下载的 URL 经过该目标泛化处理后得到的 URL;

若检测到该第二 URL 在该第二列表中, 获取该第二 URL 在该第二列表中的第一出现次数, 该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL;

若该第一出现次数小于或等于第一阈值, 则下载该待下载 URL。

应当理解, 在一些可行的实施方式中, 所称处理器 501 可以是中央处理单元 (central processing unit, CPU), 该处理器还可以是其他通用处理器、数字信号处理器 (digital signal processor, DSP)、专用集成电路 (application specific integrated circuit, ASIC)、现成可编程门阵列 (field-programmable gate array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

该存储器 502 可以包括只读存储器和随机存取存储器, 并向处理器 501 提供指令和数据。存储器 502 的一部分还可以包括非易失性随机存取存储器。例如, 存储器 502 还可以存储设备类型的信息。

具体实现中, 本申请实施例中所描述的处理器 501 可执行本申请实施例提供的统一资源定位符 URL 去重方法中所描述的实现方式, 也可执行本申请实施例所描述的 URL 去重装置的实现方式, 在此不再赘述。

本申请实施例还提供一种计算机可读存储介质, 该计算机可读存储介质存储有计算机程序, 该计算机程序包括程序指令, 该程序指令被处理器执行时实现图 1 或图 3 所示的统一资源定位符 URL 去重方法, 具体细节请参照图 1 或图 3 所示实施例的描述, 在此不再赘述。

上述计算机可读存储介质可以是前述任一实施例所述的统一资源定位符 URL 去重装置或电子设备的内部存储单元, 例如电子设备的硬盘或内存。该计算机可读存储介质也可以是该电子设备的外部存储设备, 例如该电子设备上配备的插接式硬盘, 智能存储卡 (smart media card, SMC), 安全数字 (secure digital, SD) 卡, 闪存卡 (flash card) 等。进一步地, 该计算机可读存储介质还可以既包括该电子设备的内部存储单元也包括外部存储设备。该计算机可读存储介质用于存储该计算机程序以及该电子设备所需的其他程序和数据。该计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的数据。

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以所述权利要求的保护范围为准。

权利要求

1、一种统一资源定位符 URL 去重方法，其特征在于，包括：

当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL，所述目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符，所述第一列表用于存储历史记录中已下载的 URL 经过所述目标泛化处理后得到的 URL；

若检测到所述第二 URL 在所述第二列表中，获取所述第二 URL 在所述第二列表中的第一出现次数，所述第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL；

若所述第一出现次数小于或等于第一阈值，则下载所述待下载 URL。

2、根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

当检测到所述待下载 URL 经过所述目标泛化处理后得到的所述第一 URL 在所述第一列表中时，获取所述第一 URL 在所述第一列表中的第二出现次数；

若所述第二出现次数小于或等于第二阈值，下载所述待下载 URL；

其中，所述第二阈值大于或等于所述第一阈值，所述第一阈值为大于或等于 0 的整数。

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL，包括：

对所述待下载 URL 进行目标泛化处理得到第一 URL；

利用目标哈希函数计算所述第一 URL 的哈希值；

检测所述第一 URL 的哈希值是否在所述第一列表中，所述第一列表用于存储历史记录中已下载的 URL 经过目标泛化处理后得到的 URL 的哈希值；

当所述第一 URL 的哈希值不在所述第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL。

4、根据权利要求 1-3 任意一项所述的方法，其特征在于，所述第一 URL 包括协议、服务器名称、路径以及参数字段；所述除去所述第一 URL 中的参数部分得到第二 URL，包括：

删除所述第一 URL 中的参数字段得到所述第二 URL。

5、根据权利要求 1-4 任意一项所述的方法，其特征在于，所述若检测到所述第二 URL 在所述第二列表中，获取所述第二 URL 在所述第二列表中的第一出现次数，包括：

利用目标哈希函数计算所述第二 URL 的哈希值；

检测所述第二 URL 的哈希值是否在所述第二列表中，所述第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL 的哈希值；

若所述第二 URL 的哈希值在所述第二列表中，获取所述第二 URL 的哈希值在所述第二列表中的所述第一出现次数。

6、根据权利要求 1-4 任意一项所述的方法，其特征在于，所述方法还包括：

若所述第二 URL 不在所述第二列表中，则下载所述待下载 URL。

7、根据权利要求 1-4 任意一项所述的方法，其特征在于，所述获取所述第二 URL 在所述第二列表中的第一出现次数，包括：

将所述第二 URL 在所述第二列表中的出现次数加 1 得到第一出现次数。

8、一种 URL 去重装置，其特征在于，包括：

处理模块，用于当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL，所述目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符，所述第一列表用于存储历史记录中已下载的 URL 经过所述目标泛化处理后得到的 URL；

第一获取模块，用于当检测到所述第二 URL 在所述第二列表中时，获取所述第二 URL 在所述第二列表中的第一出现次数，所述第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL；

下载模块，用于当所述第一出现次数小于或等于第一阈值时，则下载所述待下载 URL。

9、根据权利要求 8 所述的装置，其特征在于，所述装置还包括：

第二获取模块，用于当检测到所述待下载 URL 经过所述目标泛化处理后得到的所述第一 URL 在所述第一列表中时，获取所述第一 URL 在所述第一列表中的第二出现次数；

所述下载模块，还用于当所述第二出现次数小于或等于第二阈值时，下载所述待下载 URL；

其中，所述第二阈值大于或等于所述第一阈值，所述第一阈值为大于或等于 0 的整数。

10、根据权利要求 8 或 9 所述的装置，其特征在于，所述处理模块具体用于：

对所述待下载 URL 进行目标泛化处理得到第一 URL；

利用目标哈希函数计算所述第一 URL 的哈希值；

检测所述第一 URL 的哈希值是否在第一列表中，所述第一列表用于存储历史记录中已下载的 URL 经过目标泛化处理后得到的 URL 的哈希值；

当所述第一 URL 的哈希值不在所述第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL。

11、根据权利要求 8-10 所述的装置，其特征在于，所述第一 URL 包括协议、服务器名称、路径以及参数字段；

所述处理模块具体用于当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，删除该第一 URL 中的参数字段得到该第二 URL。

12、根据权利要求 8-11 所述的装置，其特征在于，所述第一获取模块具体用于：利用目标哈希函数计算该第二 URL 的哈希值；检测该第二 URL 的哈希值是否在该第二列表中；若该第二 URL 的哈希值在该第二列表中，获取该第二 URL 的哈希值在该第二列表中的该第一出现次数。其中，该第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL 的哈希值。

13、根据权利要求 8-11 所述的装置，其特征在于，所述下载模块还用于当该第二 URL 不在该第二列表中时，下载该待下载 URL。

14、根据权利要求 8-11 所述的装置，其特征在于，所述第一获取模块具体用于：

当检测到所述第二 URL 在所述第二列表中时，将所述第二 URL 在所述第二列表中的

出现次数加 1 得到第一出现次数。

15、一种终端，其特征在于，包括处理器和存储器，所述处理器和存储器相互连接，其中，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行：

当检测到待下载 URL 经过目标泛化处理后得到的第一 URL 不在第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL，所述目标泛化处理用于将 URL 中相同类型的多个连续字符替换为单个字符，所述第一列表用于存储历史记录中已下载的 URL 经过所述目标泛化处理后得到的 URL；

若检测到所述第二 URL 在所述第二列表中，获取所述第二 URL 在所述第二列表中的第一出现次数，所述第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL；

若所述第一出现次数小于或等于第一阈值，则下载所述待下载 URL。

16、根据权利要求 15 所述的终端，其特征在于，所述处理器还用于：

当检测到所述待下载 URL 经过所述目标泛化处理后得到的所述第一 URL 在所述第一列表中时，获取所述第一 URL 在所述第一列表中的第二出现次数；

若所述第二出现次数小于或等于第二阈值，下载所述待下载 URL；

其中，所述第二阈值大于或等于所述第一阈值，所述第一阈值为大于或等于 0 的整数。

17、根据权利要求 15 或 16 所述的终端，其特征在于，所述处理器具体用于：

对所述待下载 URL 进行目标泛化处理得到第一 URL；

利用目标哈希函数计算所述第一 URL 的哈希值；

检测所述第一 URL 的哈希值是否在所述第一列表中，所述第一列表用于存储历史记录中已下载的 URL 经过目标泛化处理后得到的 URL 的哈希值；

当所述第一 URL 的哈希值不在所述第一列表中时，除去所述第一 URL 中的参数部分得到第二 URL。

18、根据权利要求 15-17 任意一项所述的终端，其特征在于，所述第一 URL 包括协议、服务器名称、路径以及参数字段；

所述处理器具体用于：删除所述第一 URL 中的参数字段得到所述第二 URL。

19、根据权利要求 15-18 任意一项所述的终端，其特征在于，所述处理器具体用于：

利用目标哈希函数计算所述第二 URL 的哈希值；

检测所述第二 URL 的哈希值是否在所述第二列表中，所述第二列表用于存储历史记录中已下载的 URL 除去参数部分后得到的 URL 的哈希值；

若所述第二 URL 的哈希值在所述第二列表中，获取所述第二 URL 的哈希值在所述第二列表中的所述第一出现次数。

20、一种计算机可读存储介质，其特征在于，所述计算机存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行如权利要求 1-7 任一项所述的方法。

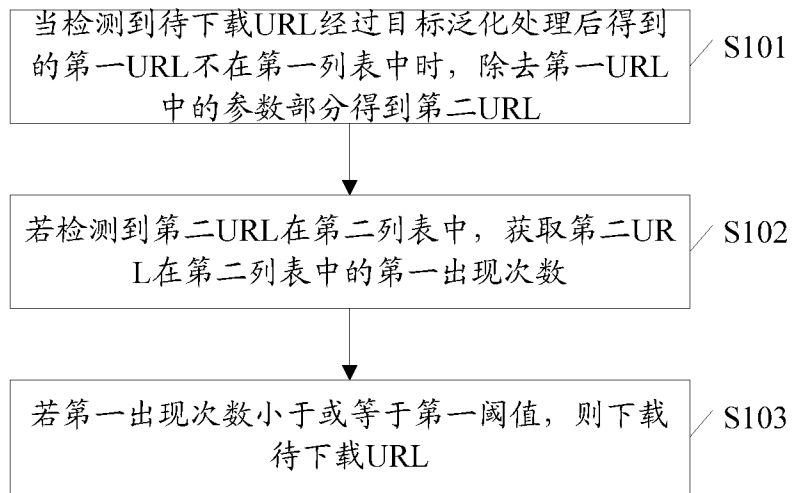


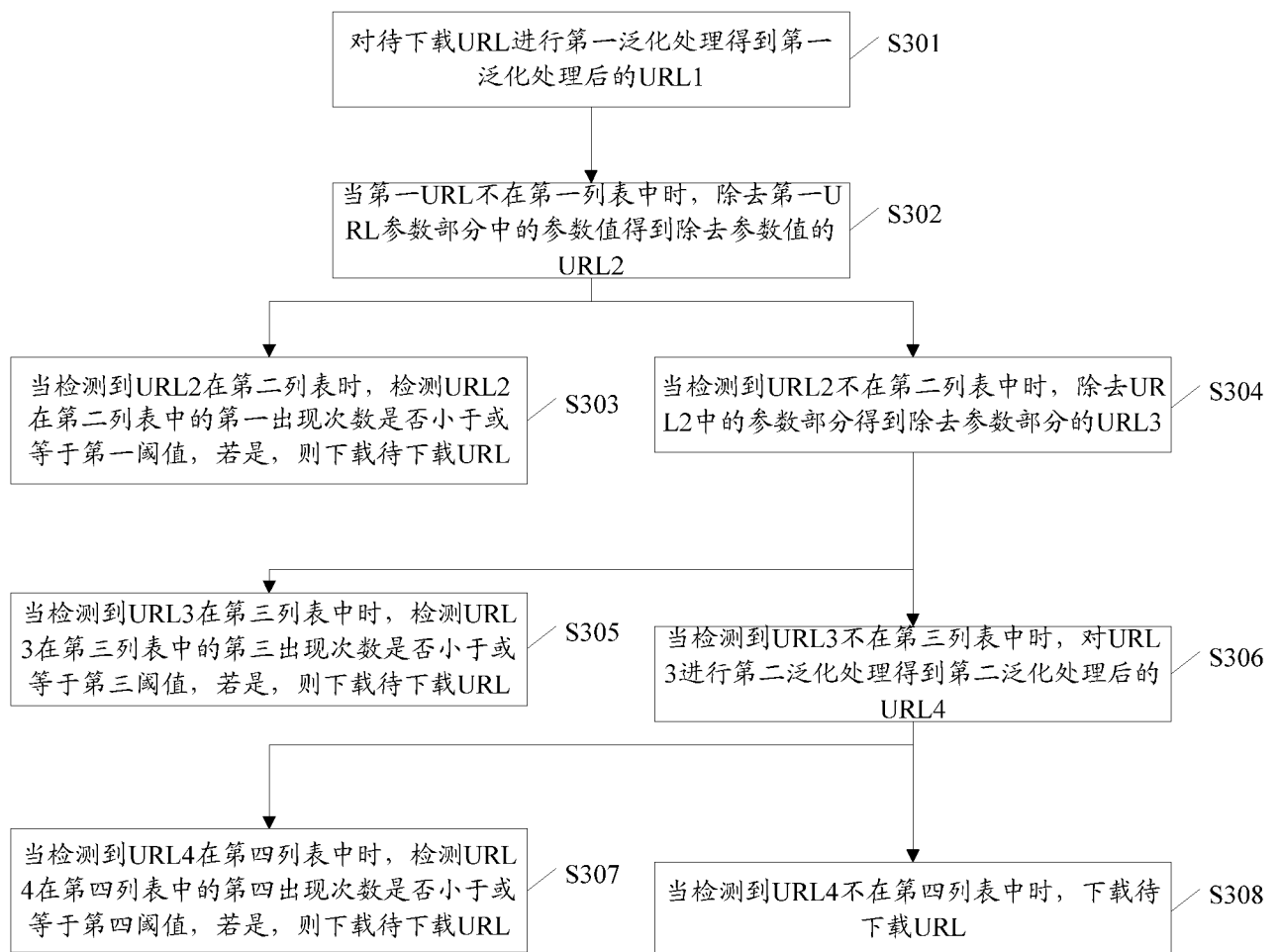
图 1

第一列表	出现次数
URL1	$1 \rightarrow 1+1=2$
URL4	3
URL7	5

图 2a

第二列表	出现次数
02	$2 \rightarrow 2+1=3$
06	2

图 2b



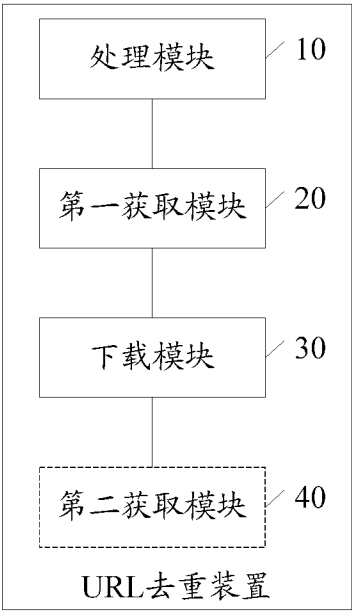


图 4

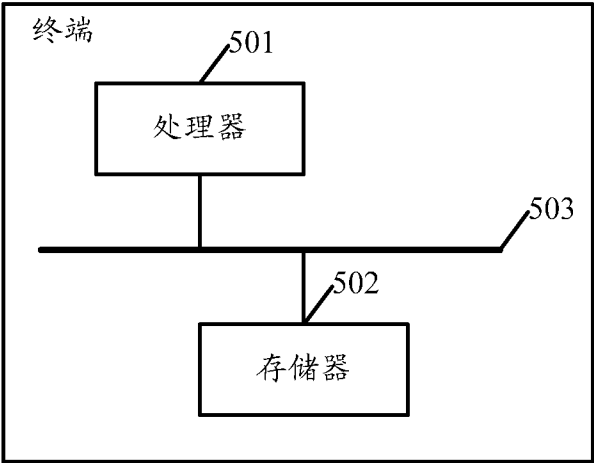


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/108715

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 去重 统一资源定位符 URL 泛化 参数 下载 deduplication uniform parameter download repetition

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106919570 A (ACADEMY OF BROADCASTING SCIENCE, SAPPRFT) 04 July 2017 (2017-07-04) description, paragraphs [0083]-[0131]	1-20
X	CN 107885820 A (BEIJING XIAODUHUYU TECHNOLOGY CO., LTD.) 06 April 2018 (2018-04-06) description, paragraphs [0056]-[0068]	1-20
X	CN 104933056 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 23 September 2015 (2015-09-23) claims 1-28	1-20
A	CN 106547764 A (BEIJING GRIDSUM TECHNOLOGY CO., LTD.) 29 March 2017 (2017-03-29) entire document	1-20
A	WO 2013143363 A1 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) 03 October 2013 (2013-10-03) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

18 March 2019

Date of mailing of the international search report

29 March 2019

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (ISA/
CN)**
**No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
 100088**
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/108715

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106919570	A	04 July 2017	None			
CN	107885820	A	06 April 2018	None			
CN	104933056	A	23 September 2015	None			
CN	106547764	A	29 March 2017	None			
WO	2013143363	A1	03 October 2013	CN	103365865	A	23 October 2013
				US	2014149457	A1	29 May 2014

国际检索报告

国际申请号

PCT/CN2018/108715

A. 主题的分类 G06F 17/00 (2019.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																				
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, CNKI, WPI, EPDOC: 去重 统一资源定位符 URL 泛化 参数 下载 deduplication uniform parameter download repetition																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 106919570 A (国家新闻出版广电总局广播科学研究院) 2017年 7月 4日 (2017 - 07 - 04) 说明书第83-131段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 107885820 A (北京小度互娱科技有限公司) 2018年 4月 6日 (2018 - 04 - 06) 说明书第56-68段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 104933056 A (腾讯科技深圳有限公司) 2015年 9月 23日 (2015 - 09 - 23) 权利要求1-28</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 106547764 A (北京国双科技有限公司) 2017年 3月 29日 (2017 - 03 - 29) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>WO 2013143363 A1 (TENCENT TECH SHENZHEN CO., LTD.) 2013年 10月 3日 (2013 - 10 - 03) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 106919570 A (国家新闻出版广电总局广播科学研究院) 2017年 7月 4日 (2017 - 07 - 04) 说明书第83-131段	1-20	X	CN 107885820 A (北京小度互娱科技有限公司) 2018年 4月 6日 (2018 - 04 - 06) 说明书第56-68段	1-20	X	CN 104933056 A (腾讯科技深圳有限公司) 2015年 9月 23日 (2015 - 09 - 23) 权利要求1-28	1-20	A	CN 106547764 A (北京国双科技有限公司) 2017年 3月 29日 (2017 - 03 - 29) 全文	1-20	A	WO 2013143363 A1 (TENCENT TECH SHENZHEN CO., LTD.) 2013年 10月 3日 (2013 - 10 - 03) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 106919570 A (国家新闻出版广电总局广播科学研究院) 2017年 7月 4日 (2017 - 07 - 04) 说明书第83-131段	1-20																		
X	CN 107885820 A (北京小度互娱科技有限公司) 2018年 4月 6日 (2018 - 04 - 06) 说明书第56-68段	1-20																		
X	CN 104933056 A (腾讯科技深圳有限公司) 2015年 9月 23日 (2015 - 09 - 23) 权利要求1-28	1-20																		
A	CN 106547764 A (北京国双科技有限公司) 2017年 3月 29日 (2017 - 03 - 29) 全文	1-20																		
A	WO 2013143363 A1 (TENCENT TECH SHENZHEN CO., LTD.) 2013年 10月 3日 (2013 - 10 - 03) 全文	1-20																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期 2019年 3月 18日		国际检索报告邮寄日期 2019年 3月 29日																		
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 方婷 电话号码 86-010-53961654																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/108715

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106919570	A	2017年 7月 4日	无	
CN	107885820	A	2018年 4月 6日	无	
CN	104933056	A	2015年 9月 23日	无	
CN	106547764	A	2017年 3月 29日	无	
WO	2013143363	A1	2013年 10月 3日	CN 103365865 A	2013年 10月 23日
				US 2014149457 A1	2014年 5月 29日