(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 110417841 B
(45)授权公告日 2022.01.18

(21)申请号 201810403070.X
(22)申请日 2018.04.28
(65)同一申请的已公布的文献号
申请公布号 CN 110417841 A
(43)申请公布日 2019.11.05
(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱
(72)发明人 徐道晨
(74)专利代理机构 北京博浩百睿知识产权代理有限责任公司 11134
代理人 褚敏 宋子良
(51)Int.Cl.
H04L 67/02(2022.01)
H04L 61/30(2022.01)
G06F 16/955(2019.01)
G06F 16/958(2019.01)
审查员 程茹

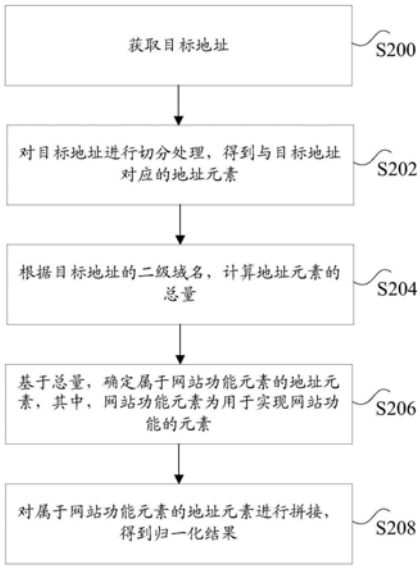
权利要求书4页 说明书16页 附图7页

(54)发明名称

地址归一化处理方法、装置和系统、数据处理方法

(57)摘要

本申请公开了一种地址归一化处理方法、装置和系统、数据处理方法。其中，该方法包括：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。本申请解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。



CN 110417841 B

1.一种地址归一化处理方法,其特征在于,包括:

获取目标地址;

对所述目标地址进行切分处理,得到与所述目标地址对应的地址元素;

根据所述目标地址的二级域名,计算所述地址元素的总量;

基于所述总量,确定属于网站功能元素的地址元素,其中,所述网站功能元素为用于实现网站功能的元素;

对所述属于网站功能元素的地址元素进行拼接,得到归一化结果;

其中,所述对所述目标地址进行切分处理,得到与所述目标地址对应的地址元素包括:

将所述目标地址切分为一级域名、路径及参数,并从所述一级域名中解析出所述二级域名;对所述一级域名按照第一规则进行拆解,得到所述一级域名对应的第一地址元素,对所述路径按照第二规则进行拆解,得到所述路径对应的第二地址元素,对所述参数按照第三规则进行拆解,得到所述参数对应的第三地址元素;

所述根据所述目标地址的二级域名,计算所述地址元素的总量包括:按照所述二级域名,对所述第一地址元素、所述第二地址元素以及所述第三地址元素进行分组;计算每个分组中的地址元素总个数;

所述基于所述总量,确定属于网站功能元素的地址元素包括:判断所述地址元素总个数是否满足预设条件,所述预设条件为根据元素属性及网站状态得到的,所述网站状态用于指示元素的计数分布情况;若所述地址元素总个数满足所述预设条件,确定对应分组中的地址元素属于所述网站功能元素;若所述地址元素总个数不满足所述预设条件,确定对应分组中的地址元素不属于所述网站功能元素。

2.根据权利要求1所述的方法,其特征在于,所述获取目标地址包括:

从网页日志中提取状态码为预设值的第一地址;

将所述第一地址中的字符转换为预设字符,得到所述目标地址。

3.据权利要求1所述的方法,其特征在于,所述对所述属于网站功能元素的地址元素进行拼接,得到归一化结果包括:

将属于所述网站功能元素的地址元素保留,将不属于所述网站功能元素的地址元素替换为预设标识;

对所述属于所述网站功能元素的地址元素以及所述预设标识进行拼接,得到所述归一化结果。

4.一种地址归一化处理装置,其特征在于,包括:

获取模块,用于获取目标地址;

切分模块,用于对所述目标地址进行切分处理,得到与所述目标地址对应的地址元素;

计算模块,用于根据所述目标地址的二级域名,计算所述地址元素的总量;

确定模块,用于基于所述总量,确定属于网站功能元素的地址元素,其中,所述网站功能元素为用于实现网站功能的元素;

拼接模块,用于对所述属于网站功能元素的地址元素进行拼接,得到归一化结果;

其中,所述切分模块还用于将目标地址切分为一级域名、路径及参数,并从一级域名中解析出二级域名;对一级域名按照第一规则进行拆解,得到一级域名对应的第一地址元素,对路径按照第二规则进行拆解,得到路径对应的第二地址元素,对参数按照第三规则进行

拆解，得到参数对应的第三地址元素；

所述计算模块还用于按照二级域名，对第一地址元素、第二地址元素以及第三地址元素进行分组；计算每个分组中的地址元素总个数；

所述确定模块还用于判断地址元素总个数是否满足预设条件，预设条件为根据元素属性及网站状态得到的，网站状态用于指示元素的计数分布情况；若地址元素总个数满足预设条件，确定对应分组中的地址元素属于网站功能元素；若地址元素总个数不满足预设条件，确定对应分组中的地址元素不属于网站功能元素。

5.一种存储介质，其特征在于，所述存储介质包括存储的程序，其中，在所述程序运行时控制所述存储介质所在设备执行如下步骤：获取目标地址；对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素；根据所述目标地址的二级域名，计算所述地址元素的总量；基于所述总量，确定属于网站功能元素的地址元素，其中，所述网站功能元素为用于实现网站功能的元素；对所述属于网站功能元素的地址元素进行拼接，得到归一化结果；其中，所述对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素包括：将所述目标地址切分为一级域名、路径及参数，并从所述一级域名中解析出所述二级域名；对所述一级域名按照第一规则进行拆解，得到所述一级域名对应的第一地址元素，对所述路径按照第二规则进行拆解，得到所述路径对应的第二地址元素，对所述参数按照第三规则进行拆解，得到所述参数对应的第三地址元素；所述根据所述目标地址的二级域名，计算所述地址元素的总量包括：按照所述二级域名，对所述第一地址元素、所述第二地址元素以及所述第三地址元素进行分组；计算每个分组中的地址元素总个数；所述基于所述总量，确定属于网站功能元素的地址元素包括：判断所述地址元素总个数是否满足预设条件，所述预设条件为根据元素属性及网站状态得到的，所述网站状态用于指示元素的计数分布情况；若所述地址元素总个数满足所述预设条件，确定对应分组中的地址元素属于所述网站功能元素；若所述地址元素总个数不满足所述预设条件，确定对应分组中的地址元素不属于所述网站功能元素。

6.一种处理器，其特征在于，所述处理器用于运行程序，其中，所述程序运行时执行如下步骤：获取目标地址；对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素；根据所述目标地址的二级域名，计算所述地址元素的总量；基于所述总量，确定属于网站功能元素的地址元素，其中，所述网站功能元素为用于实现网站功能的元素；对所述属于网站功能元素的地址元素进行拼接，得到归一化结果；其中，所述对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素包括：将所述目标地址切分为一级域名、路径及参数，并从所述一级域名中解析出所述二级域名；对所述一级域名按照第一规则进行拆解，得到所述一级域名对应的第一地址元素，对所述路径按照第二规则进行拆解，得到所述路径对应的第二地址元素，对所述参数按照第三规则进行拆解，得到所述参数对应的第三地址元素；所述根据所述目标地址的二级域名，计算所述地址元素的总量包括：按照所述二级域名，对所述第一地址元素、所述第二地址元素以及所述第三地址元素进行分组；计算每个分组中的地址元素总个数；所述基于所述总量，确定属于网站功能元素的地址元素包括：判断所述地址元素总个数是否满足预设条件，所述预设条件为根据元素属性及网站状态得到的，所述网站状态用于指示元素的计数分布情况；若所述地址元素总个数满足所述预设条件，确定对应分组中的地址元素属于所述网站功能元素；若所述地址元素总个数不满足所

述预设条件，确定对应分组中的地址元素不属于所述网站功能元素。

7.一种地址归一化处理系统，其特征在于，包括：

处理器；以及

存储器，与所述处理器连接，用于为所述处理器提供处理以下处理步骤的指令：获取目标地址；对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素；根据所述目标地址的二级域名，计算所述地址元素的总量；基于所述总量，确定属于网站功能元素的地址元素，其中，所述网站功能元素为用于实现网站功能的元素；对所述属于网站功能元素的地址元素进行拼接，得到归一化结果；其中，所述对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素包括：将所述目标地址切分为一级域名、路径及参数，并从所述一级域名中解析出所述二级域名；对所述一级域名按照第一规则进行拆解，得到所述一级域名对应的第一地址元素，对所述路径按照第二规则进行拆解，得到所述路径对应的第二地址元素，对所述参数按照第三规则进行拆解，得到所述参数对应的第三地址元素；所述根据所述目标地址的二级域名，计算所述地址元素的总量包括：按照所述二级域名，对所述第一地址元素、所述第二地址元素以及所述第三地址元素进行分组；计算每个分组中的地址元素总个数；所述基于所述总量，确定属于网站功能元素的地址元素包括：判断所述地址元素总个数是否满足预设条件，所述预设条件为根据元素属性及网站状态得到的，所述网站状态用于指示元素的计数分布情况；若所述地址元素总个数满足所述预设条件，确定对应分组中的地址元素属于所述网站功能元素；若所述地址元素总个数不满足所述预设条件，确定对应分组中的地址元素不属于所述网站功能元素。

8.一种数据处理方法，其特征在于，包括：

获取待处理元素；

从所述待处理元素中确定属于网站功能元素的待处理元素，所述网站功能元素为用于实现网站功能的元素；

对属于网站功能元素的待处理元素进行拼接，得到归一化结果；

其中，在获取待处理元素之前，所述方法还包括：

获取目标地址；将所述目标地址切分为一级域名、路径及参数，并从所述一级域名中解析出二级域名；对所述一级域名按照第一规则进行拆解，得到所述一级域名对应的第一地址元素，对所述路径按照第二规则进行拆解，得到所述路径对应的第二地址元素，对所述参数按照第三规则进行拆解，得到所述参数对应的第三地址元素；按照所述二级域名，对所述第一地址元素、所述第二地址元素以及所述第三地址元素进行分组；计算每个分组中的地址元素总个数；将所述地址元素总个数作为所述待处理元素；判断所述地址元素总个数是否满足预设条件，所述预设条件为根据元素属性及网站状态得到的，所述网站状态用于指示元素的计数分布情况；若所述地址元素总个数满足所述预设条件，确定对应分组中的地址元素属于所述网站功能元素；若所述地址元素总个数不满足所述预设条件，确定对应分组中的地址元素不属于所述网站功能元素。

9.一种数据处理方法，其特征在于，包括：

获取目标地址；

对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素；

根据所述目标地址的二级域名，计算所述地址元素的总量；

确定用于实现网站功能的地址元素；

对所述用于实现网站功能的地址元素进行拼接；

其中，所述对所述目标地址进行切分处理，得到与所述目标地址对应的地址元素包括：将所述目标地址切分为一级域名、路径及参数，并从所述一级域名中解析出所述二级域名；对所述一级域名按照第一规则进行拆解，得到所述一级域名对应的第一地址元素，对所述路径按照第二规则进行拆解，得到所述路径对应的第二地址元素，对所述参数按照第三规则进行拆解，得到所述参数对应的第三地址元素；

所述根据所述目标地址的二级域名，计算所述地址元素的总量包括：按照所述二级域名，对所述第一地址元素、所述第二地址元素以及所述第三地址元素进行分组；计算每个分组中的地址元素总个数；

所述确定用于实现网站功能的地址元素包括：判断所述地址元素总个数是否满足预设条件，所述预设条件为根据元素属性及网站状态得到的，所述网站状态用于指示元素的计数分布情况；若所述地址元素总个数满足所述预设条件，确定对应分组中的地址元素属于所述网站功能元素；若所述地址元素总个数不满足所述预设条件，确定对应分组中的地址元素不属于所述网站功能元素。

10.根据权利要求9所述的方法，其特征在于，所述对所述用于实现网站功能的地址元素进行拼接包括：

将所述用于实现网站功能的地址元素保留，将不用于实现网站功能的地址元素替换为预设标识；

对所述用于实现网站功能的地址元素以及所述预设标识进行拼接。

地址归一化处理方法、装置和系统、数据处理方法

技术领域

[0001] 本申请涉及互联网领域，具体而言，涉及一种地址归一化处理方法、装置和系统、数据处理方法。

背景技术

[0002] URL(Uniform Resource Location，统一资源定位符)中常常携带有参数，例如，URL地址为/friend/zhangsan/index.php，其中，zhangsan是作为参数传输至网站后台的。黑客常常利用这一点，通过不断变化填充的参数，让请求的地址两两不同，以规避网站的处置。为了解决上述问题，可以将功能相近的地址压缩成一类的地址归一化，以上述地址为例，可以压缩为/friend/{参数}/index.php。

[0003] 现有技术中的地址归一化方案，为自底向上的地址聚合，即通过地址间的互相比较，判断是否需要合并，但是，该归一化方法的阈值不容易确定，而且不符合网站功能。具体举例如下：日志中存在如下九个地址，假设同一位置大于两个的变参，确定需要合并：

[0004] 1)/friend/photo/1.png

[0005] 2)/friend/photo/2.png

[0006] 3)/friend/photo/3.png

[0007] 4)/friend/photo/4.png

[0008] 5)/friend/a.js

[0009] 6)/friend/m.css

[0010] 7)/friend/index.html

[0011] 8)/friend/index.html

[0012] 9)/friend/index.html

[0013] 从功能而言，1-4可以归一化成/friend/photo/{参数}，7-9可以归一化成/friend/index.html，5-6可以归成/friend/{参数}，符合网站功能分类，但是，按照自底向上的做法，1-4将统一归成/friend/{参数}/{参数}，5-9将归为/friend/{参数}，不但处理效率较低而且损失了URL功能信息。

[0014] 针对现有的URL归一化处理方法处理效率低，且不符合网站功能的问题，目前尚未提出有效的解决方案。

发明内容

[0015] 本申请实施例提供了一种地址归一化处理方法、装置和系统、数据处理方法，以至少解决现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

[0016] 根据本申请实施例的一个方面，提供了一种地址归一化处理方法，包括：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结

果。

[0017] 根据本申请实施例的另一方面，还提供了一种地址归一化处理装置，包括：获取模块，用于获取目标地址；切分模块，用于对目标地址进行切分处理，得到与目标地址对应的地址元素；计算模块，用于根据目标地址的二级域名，计算地址元素的总量；确定模块，用于基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；拼接模块，用于对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0018] 根据本申请实施例的另一方面，还提供了一种存储介质，存储介质包括存储的程序，其中，在程序运行时控制存储介质所在设备执行如下步骤：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0019] 根据本申请实施例的另一方面，还提供了一种处理器，处理器用于运行程序，其中，程序运行时执行如下步骤：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0020] 根据本申请实施例的另一方面，还提供了一种地址归一化处理系统，包括：处理器；以及存储器，与处理器连接，用于为处理器提供处理以下处理步骤的指令：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0021] 根据本申请实施例的另一方面，还提供了一种数据处理方法，包括：获取待处理元素；从待处理元素中确定属于网站功能元素的待处理元素，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的待处理元素进行拼接，得到归一化结果。

[0022] 根据本申请实施例的另一方面，还提供了一种数据处理方法，包括：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；确定用于实现网站功能的地址元素；对用于实现网站功能的地址元素进行拼接。

[0023] 在本申请实施例中，在获取到目标地址之后，首先可以对目标地址进行切分处理，得到与目标地址对应的地址元素，然后根据目标地址的二级域名，计算每个地址元素的总量，进一步基于总量确定属于网站功能元素的地址元素，最后将属于网站功能元素的地址元素进行拼接，从而得到目标地址的归一化结果。

[0024] 通过本申请上述实施例所提供的方案，基于目标地址中切分得到的地址元素的总量，确定属于网站功能元素的地址元素，并根据属于网站功能元素的地址元素，得到归一化结果，与现有技术相比，仅仅依靠网站的web日志，即可自动化得到网站核心功能的归一化结果，从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果，达到了提高处理效率，同时符合网站功能，并且能够规避黑客对网站的处置的技术效果，进而解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

附图说明

[0025] 此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

[0026] 图1是根据本申请实施例一种用于实现地址归一化处理方法的计算机终端(或移动设备)的硬件结构框图；

[0027] 图2是根据本申请实施例1的一种地址归一化处理方法的流程图；

[0028] 图3是根据本申请实施例1的一种可选的地址归一化处理方法的示意图；

[0029] 图4是根据本申请实施例1的一种地址归一化处理装置的示意图；

[0030] 图5是根据本申请实施例4的一种数据处理方法的流程图；

[0031] 图6是根据本申请实施例5的一种数据处理装置的示意图；

[0032] 图7是根据本申请实施例6的一种数据处理方法的流程图；以及

[0033] 图8是根据本申请实施例的一种计算机终端的结构框图。

具体实施方式

[0034] 为了使本技术领域的人员更好地理解本申请方案，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，所描述的实施例仅仅是本申请一部分的实施例，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，都应当属于本申请保护的范围。

[0035] 需要说明的是，本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0036] 首先，在对本申请实施例进行描述的过程中出现的部分名词或术语适用于如下解释：

[0037] URL归一化(URL Normalization)：可以将功能相近的URL合并为同一类；

[0038] 一级域名：可以是由一串用点‘.’分隔的字符组成的互联网上某一台计算机或计算机组的名称，例如，一级域名可以是www.tmall.com。

[0039] 二级域名：可以是指一级域名中顶级域名之下的域名，是域名的倒数第二个部分，可以是一级域名中，最后一个点‘.’的左边的字符，例如，对于一级域名www.tmall.com，其中，顶级域名为.com，二级域名为tmall.com。

[0040] 实施例1

[0041] 根据本申请实施例，还提供了一种URL归一化处理方法的实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0042] 本申请实施例一所提供的方法实施例可以在移动终端、计算机终端或者类似的运算装置中执行。图1示出了一种用于实现地址归一化处理方法的计算机终端(或移动设备)的硬件结构框图。如图1所示,计算机终端10(或移动设备10)可以包括一个或多个(图中采用102a、102b,……,102n来示出)处理器102(处理器102可以包括但不限于微处理器MCU或可编程逻辑器件FPGA等的处理装置)、用于存储数据的存储器104、以及用于通信功能的传输装置106。除此以外,还可以包括:显示器、输入/输出接口(I/O接口)、通用串行总线(USB)端口(可以作为I/O接口的端口中的一个端口被包括)、网络接口、电源和/或相机。本领域普通技术人员可以理解,图1所示的结构仅为示意,其并不对上述电子装置的结构造成限定。例如,计算机终端10还可包括比图1中所示更多或者更少的组件,或者具有与图1所示不同的配置。

[0043] 应当注意到的是上述一个或多个处理器102和/或其他数据处理电路在本文中通常可以被称为“数据处理电路”。该数据处理电路可以全部或部分的体现为软件、硬件、固件或其他任意组合。此外,数据处理电路可为单个独立的处理模块,或全部或部分的结合到计算机终端10(或移动设备)中的其他元件中的任意一个内。如本申请实施例中所涉及到的,该数据处理电路作为一种处理器控制(例如与接口连接的可变电阻终端路径的选择)。

[0044] 存储器104可用于存储应用软件的软件程序以及模块,如本申请实施例中的地址归一化处理方法对应的程序指令/数据存储装置,处理器102通过运行存储在存储器104内的软件程序以及模块,从而执行各种功能应用以及数据处理,即实现上述的地址归一化处理方法。存储器104可包括高速随机存储器,还可包括非易失性存储器,如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中,存储器104可进一步包括相对于处理器102远程设置的存储器,这些远程存储器可以通过网络连接至计算机终端10。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0045] 传输装置106用于经由一个网络接收或者发送数据。上述的网络具体实例可包括计算机终端10的通信供应商提供的无线网络。在一个实例中,传输装置106包括一个网络适配器(Network Interface Control ler,NIC),其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中,传输装置106可以为射频(Radio Frequency,RF)模块,其用于通过无线方式与互联网进行通讯。

[0046] 显示器可以例如触摸屏式的液晶显示器(LCD),该液晶显示器可使得用户能够与计算机终端10(或移动设备)的用户界面进行交互。

[0047] 此处需要说明的是,在一些可选实施例中,上述图1所示的计算机设备(或移动设备)可以包括硬件元件(包括电路)、软件元件(包括存储在计算机可读介质上的计算机代码)、或硬件元件和软件元件两者的结合。应当指出的是,图1仅为特定具体实例的一个实例,并且旨在示出可存在于上述计算机设备(或移动设备)中的部件的类型。

[0048] 在上述运行环境下,本申请提供了如图2所示的地址归一化处理方法。图2是根据本申请实施例1的一种地址归一化处理方法的流程图。如图2所示,该方法包括如下步骤:

[0049] 步骤S200,获取目标地址。

[0050] 上述步骤S200中的目标地址可以是从web日志中获取到的需要进行URL归一化处理的多个URL。

[0051] 步骤S202,对目标地址进行切分处理,得到与目标地址对应的地址元素。

[0052] 上述步骤S202中可以是对目标地址进行切分，得到的多个数组，例如，当目标地址为/friend/zhangsan/index.php时，对目标地址进行切分处理，得到的地址元素可以是friend、zhangsan和index.php。

[0053] 需要说明的是，由于地址是由域名(host)、路径(path)及参数(query)三部分组成的，因此可以按照上述三部分对地址进行切分，得到每部分的数组。

[0054] 步骤S204，根据目标地址的二级域名，计算地址元素的总量。

[0055] 需要说明的是，为了能够按照功能对地址进行归一化处理，可以根据目标地址的二级域名进行归一化处理。

[0056] 上述步骤S204的总量可以是目标地址中每一个地址元素的总数量。

[0057] 步骤S206，基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素。

[0058] 上述步骤S206中，可以参照元素本身(包括但不限于：元素长度、是否包含数字、是否包含特殊字符、元素语义信息、属于域名或路径或参数)和网站整体情况(包括但不限于：网站元素计数分布情况、按元素分类的计数分布情况)，选出总量满足阈值的地址元素，从而确定属于网站功能元素的地址元素。

[0059] 步骤S208，对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0060] 需要说明的是，在地址中，属于网站功能元素的地址元素保持不变，而不属于网站功能元素的地址元素可以认为是不断变化的参数，在生成归一化结果的过程中，可以按照地址原来的顺序，将属于网站功能元素的地址元素和参数进行拼接，从而得到归一化结果。

[0061] 下面结合图3对本申请实施例中一种优选的地址归一化处理方法进行详细说明。如图3所示，该方法可以包括如下步骤：

[0062] 步骤S32，预处理。

[0063] 可选地，可以对web日志中存储的URL进行预处理，筛选出正常的多个URL。具体可以从web日志中获取状态码为预设值，且字符转换为预设字符的URL，例如，预处理后得到多个URL可以包括九个URL，具体如下：

[0064] 1)/friend/photo/1.png，该URL仅包含路径部分，表示路径“/friend/photo”下，格式为png，文件名为1的文件；

[0065] 2)/friend/photo/2.png，该URL仅包含路径部分，表示路径“/friend/photo”下，格式为png，文件名为2的文件；

[0066] 3)/friend/photo/3.png，该URL仅包含路径部分，表示路径“/friend/photo”下，格式为png，文件名为3的文件；

[0067] 4)/friend/photo/4.png，该URL仅包含路径部分，表示路径“/friend/photo”下，格式为png，文件名为4的文件；

[0068] 5)/friend/a.js，该UR仅包含路径部分，L表示路径“/friend”下，格式为js，文件名为a的文件；

[0069] 6)/friend/m.css，该URL仅包含路径部分，表示路径“/friend”下，格式为css，文件名为m的文件；

[0070] 7)/friend/index.html，该URL仅包含路径部分，表示路径“/friend”下，格式为html，文件名为index的文件；

[0071] 8)/friend/index.html；

[0072] 9)/friend/index.html。

[0073] 由上可知，第1个至第4个URL为相同位置“/friend/photo”下的四个变参，第5个至第9个URL为相同位置“/friend”下的三个变参。

[0074] 步骤S34，切分。

[0075] 可选地，将预处理后得到的URL进行切分，切分成域名、路径以及参数，并从域名中解析出二级域名。进一步地对域名、路径以及参数部分进行拆解，得到每个部分的地址元素。例如，对上述的九个URL进行切分，切分结果如下所示：

[0076] 1)[friend,photo,1.png]，表示该URL包含路径对应的三个地址元素，分别为地址元素friend、地址元素photo和地址元素1.png；

[0077] 2)[friend,photo,2.png]，表示该URL包含路径对应的三个地址元素，分别为地址元素friend、地址元素photo和地址元素2.png；

[0078] 3)[friend,photo,3.png]，表示该URL包含路径对应的三个地址元素，分别为地址元素friend、地址元素photo和地址元素3.png；

[0079] 4)[friend,photo,4.png]，表示该URL包含路径对应的三个地址元素，分别为地址元素friend、地址元素photo和地址元素4.png；

[0080] 5)[friend,a.js]，表示该URL包含路径对应的两个地址元素，分别为地址元素friend和地址元素a.js；

[0081] 6)[friend,m.css]，表示该URL包含路径对应的两个地址元素，分别为地址元素friend和地址元素m.css；

[0082] 7)[friend,index.html]，表示该URL包含路径对应的两个地址元素，分别为地址元素friend和地址元素index.html；

[0083] 8)[friend,index.html]；

[0084] 9)[friend,index.html]。

[0085] 步骤S36，判断日志是否累计到一定量。

[0086] 需要说明的是，如果web日志未累计到一定量，则切分后的地址元素的总量无法满足条件，也即，无法确定出属于网站功能元素的地址元素。

[0087] 可选地，在对多个URL进行拆分之后，可以判断web日志是否累计到一定量，当判断出web日志累计到一定量时，可以进入步骤S38。

[0088] 步骤S38，统计。

[0089] 可选地，在判断出web日志累计到一定量之后，可以按照二级域名划分统计范围，计算出网站拆解出URL各地址元素总量，每个URL中的地址元素去重后计入统计。例如，假设上述九个URL划分为同一组，对上述九个URL拆分后的结果进行统计，从而得到每个地址元素的总量如下：地址元素friend的总量为9，地址元素photo的总量为4，地址元素index.html的总量为3，地址元素1.png的总量为1，地址元素2.png的总量为1，地址元素3.png的总量为1，地址元素4.png的总量为1，地址元素a.js的总量为1，地址元素m.css的总量为1。

[0090] 步骤S310，确定网站功能元素。

[0091] 可选地，可以根据地址元素本身以及网站整体情况，通过综合考虑上述两部分内

容,可以根据网站需求,确定流入网站功能元素池的标准。根据确定后的标准从统计后的地址元素中确定网站功能元素。例如,根据上述的统计结果,确定九个URL中的所有网站功能元素如下:friend、photo和index.html。

[0092] 步骤S312,拼接,得到URL归一化结果。

[0093] 可选地,可以按照顺序拼接元素得到URL归一化结果,对于域名部分的地址元素,可以按照‘.’拼接,对于路径部分的地址元素,可以按照‘/’拼接,对于参数部分的地址元素,可以按照‘&’拼接,如果路径部分不为空,则可以加前缀‘/’拼接至域名部分拼接结果之后,如果参数部分不为空,则可以加前缀‘?’拼接至域名部分和路径部分的拼接结果之后。例如,上述九个URL的URL归一化结果如下:

[0094] 1)/friend/photo/{参数},该URL归一化结果表示第1个至第4个URL功能相近,路径“/friend/photo”下携带有参数;

[0095] 2)/friend/photo/{参数};

[0096] 3)/friend/photo/{参数};

[0097] 4)/friend/photo/{参数};

[0098] 5)/friend/{参数},该URL归一化结果表示第5个至第6个URL功能相近,路径“/friend”下携带有参数;

[0099] 6)/friend/{参数};

[0100] 7)/friend/index.html,该URL归一化结果表示第7个至第9个URL功能相近,未携带参数;

[0101] 8)/friend/index.html;

[0102] 9)/friend/index.html。

[0103] 需要说明的是,对于非网站功能元素的其他地址元素,可以用{参数}进行代替。

[0104] 基于上述实施例所限定的方案可以获知,在获取到目标URL之后,首先可以对目标URL进行切分处理,得到与目标URL对应的地址元素,然后根据URL的二级域名,计算每个地址元素的总量,进一步基于总量确定属于网站功能元素的地址元素,最后将属于网站功能元素的地址元素进行拼接,从而得到目标URL的归一化结果。

[0105] 通过本申请上述实施例所提供的方案,基于目标地址中切分得到的地址元素的总量,确定属于网站功能元素的地址元素,并根据属于网站功能元素的地址元素,得到归一化结果,与现有技术相比,仅仅依靠网站的web日志,即可自动化得到网站核心功能的归一化结果,从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果,达到了提高处理效率,同时符合网站功能,并且能够规避黑客对网站的处置的技术效果,进而解决了现有的URL归一化处理方法处理效率低,且不符合网站功能的技术问题。

[0106] 本申请提供的一种可选实施例中,获取目标地址包括:从网页日志中提取状态码为预设值的第一地址;将第一地址中的字符转换为预设字符,得到目标地址。

[0107] 为了避免异常地址的干扰,上述的预设值可以是200;为了方便对地址元素进行统计,上述的预设字符可以是小写字符。

[0108] 在一种可选的实施例中,可以从web日志中筛选出状态码为200的地址,并将地址中的所有字符转换为小写字符,从而得到目标地址,例如,从web日志中筛选出状态码为200的地址,并将地址中的所有字符转换为小写字符,也即,得到上述的九个地址。

[0109] 本申请提供的一种可选实施例中，对目标地址进行切分处理，得到与目标地址对应的地址元素包括：将目标地址切分为一级域名、路径及参数，并从一级域名中解析出二级域名；对一级域名按照第一规则进行拆解，得到一级域名对应的第一地址元素，对路径按照第二规则进行拆解，得到路径对应的第二地址元素，对参数按照第三规则进行拆解，得到参数对应的第三地址元素。

[0110] 在一个地址中，对于host和path部分，元素可以是拆解得到的数组中每一个字符串；对于query部分，元素可以是拆解得到数组中每一个键值对的键(key)。

[0111] 需要说明的是，对于一个地址，均可以切分成一级域名、路径和参数三个部分，但是，路径或参数可以为空。

[0112] 在一个地址中，一级域名可以是地址中第一个斜杠(/)之前的部分，参数可以是地址中第一个问号(?)之后的部分，路径可以是地址中第一个斜杠之后，第一个问号之间的部分，例如，地址为www.tmall.com/index.php?a＝XXX，则host部分可以是www.tmall.com，pach部分可以是index.php，query部分可以是a＝XXX。对于host部分，可以根据域名规范拆解出二级域名，顶级域名前一个‘.’可以是二级域名，例如，对于host部分www.tmall.com，拆解出的二级域名可以是tmall.com，对于host部分www.aaa.com.cn，拆解出的二级域名可以是aaa.com.cn。

[0113] 上述的第一规则可以是去除二级域名后缀，按照‘.’进行拆分，第二规则可以是按照‘/’进行拆分，并去除空字符，第三规则可以是按照‘&’进行拆分，去除空字符，并按第一个‘＝’切分为键值对。例如，对于host部分www.tmall.com进行拆解，拆解后得到的第一地址元素可以是www；对于path部分index.php，由于不包含‘/’，因此，拆解得到的第二地址元素可以是index.php；对于query部分a＝XXX进行拆解，拆解后得到的第三地址元素可以是a:XXX。

[0114] 本申请提供的一种可选实施例中，根据目标地址的二级域名，计算地址元素的总量包括：按照二级域名，对第一地址元素、第二地址元素以及第三地址元素进行分组；计算每个分组中的地址元素总个数。

[0115] 在一种可选的实施例中，可以以二级域名划分统计范围，并计算网站拆解出的地址中各地址元素总量，每个地址中的元素去重后计入统计。

[0116] 本申请提供的一种可选实施例中，基于总量，确定属于网站功能元素的地址元素包括：判断地址元素总个数是否满足预设条件，预设条件为根据元素属性及网站状态得到的，网站状态用于指示元素的计数分布情况；若地址元素总个数满足预设条件，确定对应分组中的地址元素属于网站功能元素；若地址元素总个数不满足预设条件，确定对应分组中的地址元素不属于网站功能元素。

[0117] 上述的预设条件可以根据网站整体情况进行确定，网站整体情况包括网站元素技术分布情况，按元素分类的技术分布情况等，例如，预设条件是网站最高频元素的0.5次方得到的阈值，例如，对于上述的九个URL，最高频元素可以是friend，频次为9，则得到的预设条件为3。

[0118] 在一种可选的实施例中，在获取到每个地址元素的总个数之后，可以将每个地址元素的总个数与阈值进行比较，如果地址元素的总个数大于等于预设条件，则确定该地址元素属于网站功能元素，如果地址元素的总个数小于预设条件，则确定该地址元素不属于

网站功能元素。例如,对于上述九个URL,计算得到的每个地址元素的总个数如下:地址元素friend的总个数为9,地址元素photo的总个数为4,地址元素index.html的总个数为3,地址元素1.png的总个数为1,地址元素2.png的总个数为1,地址元素3.png的总个数为1,地址元素4.png的总个数为1,地址元素a.js的总个数为1,地址元素m.css的总个数为1。其中最高频元素是friend,根据该总数量的0.5次方得到的阈值为3,将每个地址元素的总个数与该阈值进行比较,由于地址元素friend的总个数为9>3,则确定地址元素friend属于网站功能元素,由于地址元素photo的总个数为4>3,则确定地址元素photo属于网站功能元素,由于地址元素index.html的总个数为3=3,则确定地址元素index.html属于网站功能元素,由于地址元素1.png的总个数为1<3,则确定地址元素1.png不属于网站功能元素,由于地址元素2.png的总个数为1<3,则确定地址元素2.png不属于网站功能元素,由于地址元素3.png的总个数为1<3,则确定地址元素3.png不属于网站功能元素,由于地址元素4.png的总个数为1<3,则确定地址元素4.png不属于网站功能元素,由于地址元素a.js的总个数为1<3,则确定地址元素a.js不属于网站功能元素,由于地址元素m.css的总个数为1<3,则确定地址元素m.css的不属于网站功能元素。从而得到确定九个URL中的所有网站功能元素如下:friend、photo和index.html。

[0119] 本申请提供的一种可选实施例中,对属于网站功能元素的地址元素进行拼接,得到归一化结果包括:将属于网站功能元素的地址元素保留,将不属于网站功能元素的地址元素替换为预设标识;对属于网站功能元素的地址元素以及预设标识进行拼接,得到归一化结果。

[0120] 上述的预设标识可以是{参数},也即,不属于网站功能元素的地址元素可以根据需要进行修改设置,可以作为参数传递至网站后台。

[0121] 在一种可选的方案中,为了确保处理后的处理结果满足网站功能,可以保留属于网站功能元素的地址元素,并且不属于网站功能元素的地址元素替换为{参数},进一步按照URL顺序,将属于网站功能元素的地址元素和{参数}进行拼接,从而得到归一化结果。例如,由于1.png、2.png、3.png、4.png、a.js和m.css不属于功能元素,因此可以将上述几个地址元素可以是替换为{参数},并且后的归一化结果如上述步骤S312中示例所示。在形式上相当于直接将URL中不属于网站功能元素的地址元素替换为{参数}。

[0122] 需要说明的是,对于前述的各方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本申请并不受所描述的动作顺序的限制,因为依据本申请,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本申请所必须的。

[0123] 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根据上述实施例的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,本申请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)执行本申请各个实施例所述的方法。

[0124] 实施例2

[0125] 根据本申请实施例，还提供了一种用于实施上述地址归一化处理方法的地址归一化处理装置，如图4所示，该装置400包括：

[0126] 获取模块402，用于获取目标地址。

[0127] 切分模块404，用于对目标地址进行切分处理，得到与目标地址对应的地址元素。

[0128] 计算模块406，用于根据目标地址的二级域名，计算地址元素的总量。

[0129] 确定模块408，用于基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素。

[0130] 拼接模块410，用于对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0131] 此处需要说明的是，上述获取模块402、切分模块404、计算模块406、确定模块408和拼接模块410对应于实施例1中的步骤S200至步骤S208，五个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例1所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例1提供的计算机终端10中。

[0132] 基于上述实施例所限定的方案可以获知，在获取到目标地址之后，首先可以对目标地址进行切分处理，得到与目标地址对应的地址元素，然后根据目标地址的二级域名，计算每个地址元素的总量，进一步基于总量确定属于网站功能元素的地址元素，最后将属于网站功能元素的地址元素进行拼接，从而得到目标地址的归一化结果。

[0133] 通过本申请上述实施例所提供的方案，基于目标地址中切分得到的地址元素的总量，确定属于网站功能元素的地址元素，并根据属于网站功能元素的地址元素，得到归一化结果，与现有技术相比，仅仅依靠网站的web日志，即可自动化得到网站核心功能的归一化结果，从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果，达到了提高处理效率，同时符合网站功能，并且能够规避黑客对网站的处置的技术效果，进而解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

[0134] 本申请提供的一种可选实施例中，获取模块402还用于从网页日志中提取状态码为预设值的第一地址；将第一地址中的字符转换为预设字符，得到目标地址。

[0135] 本申请提供的一种可选实施例中，切分模块404还用于将目标地址切分为一级域名、路径及参数，并从一级域名中解析出二级域名；对一级域名按照第一规则进行拆解，得到一级域名对应的第一地址元素，对路径按照第二规则进行拆解，得到路径对应的第二地址元素，对参数按照第三规则进行拆解，得到参数对应的第三地址元素。

[0136] 本申请提供的一种可选实施例中，计算模块406还用于按照二级域名，对第一地址元素、第二地址元素以及第三地址元素进行分组；计算每个分组中的地址元素总个数。

[0137] 本申请提供的一种可选实施例中，确定模块408还用于判断地址元素总个数是否满足预设条件，预设条件为根据元素属性及网站状态得到的，网站状态用于指示元素的计数分布情况；若地址元素总个数满足预设条件，确定对应分组中的地址元素属于网站功能元素；若地址元素总个数不满足预设条件，确定对应分组中的地址元素不属于网站功能元素。

[0138] 本申请提供的一种可选实施例中，拼接模块410还用于将属于网站功能元素的地址元素保留，将不属于网站功能元素的地址元素替换为预设标识；对属于网站功能元素的地址元素以及预设标识进行拼接，得到归一化结果。

[0139] 需要说明的是，本实施例的可选或优选实施方式可以参见实施例1中的相关描述，

在此不在赘述。

[0140] 实施例3

[0141] 根据本申请实施例，还提供了一种用于实施上述地址归一化处理方法的地址归一化处理系统，包括：

[0142] 处理器；以及

[0143] 存储器，与处理器连接，用于为处理器提供处理以下处理步骤的指令：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0144] 需要说明的是，本实施例的可选或优选实施方式可以参见实施例1中的相关描述，在此不在赘述。

[0145] 实施例4

[0146] 根据本申请实施例，还提供了一种数据处理方法的实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0147] 图5是根据本申请实施例4的一种数据处理方法的流程图。如图5所示，该方法包括如下步骤：

[0148] 步骤S502，获取待处理元素。

[0149] 上述步骤S502中，可以从web日志中获取到需要进行地址归一化处理的多个地址，也即，获取到目标地址，并对目标地址进行切分，得到与目标地址对应的地址元素，也即，得到上述的待处理元素，也即，待处理元素可以是多个目标地址中的多个数组。例如，对于当目标地址为/friend/zhangsan/index.php时，得到的待处理元素可以是friend、zhangsan和index.php。

[0150] 需要说明的是，由于地址是由域名(host)、路径(path)及参数(query)三部分组成的，因此可以按照上述三部分对地址进行切分，得到每部分的数组。

[0151] 步骤S504，从待处理元素中确定属于网站功能元素的待处理元素，网站功能元素为用于实现网站功能的元素。

[0152] 上述步骤S504中，可以参照元素本身(包括：元素长度、是否包含数字、是否包含特殊字符、元素语义信息、属于域名或路径或参数等)和网站整体情况(包括网站元素计数分布情况、按元素分类的计数分布情况等)，选出总量满足阈值的待处理元素，从而确定属于网站功能元素的待处理元素。

[0153] 需要说明的是，为了能够按照功能对地址进行归一化处理，可以根据目标地址的二级域名进行归一化处理。

[0154] 步骤S506，对属于网站功能元素的待处理元素进行拼接，得到归一化结果。

[0155] 需要说明的是，在地址中，属于网站功能元素的待处理元素保持不变，而不属于网站功能元素的待处理元素可以认为是不断变化的参数，在生成归一化结果的过程中，可以按照地址原来的顺序，将属于网站功能元素的待处理元素和参数进行拼接，从而得到归一

化结果。

[0156] 基于上述实施例所限定的方案可以获知，在获取到待处理元素之后，可以从待处理元素中确定属于网站功能元素的待处理元素，进一步对属于网站功能元素的待处理元素进行拼接，从而得到归一化结果。

[0157] 通过本申请上述实施例所提供的方案，从待处理元素中确定属于网站功能元素的待处理元素，并根据属于网站功能元素的待处理元素，得到归一化结果，与现有技术相比，仅仅依靠网站的web日志，即可自动化得到网站核心功能的归一化结果，从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果，达到了提高处理效率，同时符合网站功能，并且能够规避黑客对网站的处置的技术效果，进而解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

[0158] 实施例5

[0159] 根据本申请实施例，还提供了一种用于实施上述数据处理方法的数据处理装置，如图6所示，该装置600包括：

[0160] 获取模块602，用于获取待处理元素。

[0161] 确定模块604，用于从待处理元素中确定属于网站功能元素的待处理元素，网站功能元素为用于实现网站功能的元素。

[0162] 拼接模块606，用于对属于网站功能元素的待处理元素进行拼接，得到归一化结果。

[0163] 此处需要说明的是，上述获取模块602、确定模块604和拼接模块606对应于实施例4中的步骤S502至步骤S506，三个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例4所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例1提供的计算机终端10中。

[0164] 需要说明的是，本实施例的可选或优选实时方式可以参见实施例4中的相关描述，在此不在赘述。

[0165] 实施例6

[0166] 根据本申请实施例，还提供了一种数据处理方法的实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0167] 图7是根据本申请实施例6的一种数据处理方法的流程图。如图7所示，该方法包括如下步骤：

[0168] 步骤S700，获取目标地址。

[0169] 步骤S702，对目标地址进行切分处理，得到与目标地址对应的地址元素。

[0170] 步骤S704，根据目标地址的二级域名，计算地址元素的总量。

[0171] 步骤S706，确定用于实现网站功能的地址元素。

[0172] 步骤S708，对用于实现网站功能的地址元素进行拼接。

[0173] 基于上述实施例所限定的方案可以获知，在获取到目标地址之后，首先可以对目标地址进行切分处理，得到与目标地址对应的地址元素，然后根据目标地址的二级域名，计算每个地址元素的总量，进一步确定属于网站功能元素的地址元素，最后将属于网站功能

元素的地址元素进行拼接，从而得到目标地址的归一化结果。

[0174] 通过本申请上述实施例所提供的方案，基于目标地址中切分得到的地址元素的总量，确定属于网站功能元素的地址元素，并根据属于网站功能元素的地址元素，得到归一化结果，与现有技术相比，仅仅依靠网站的web日志，即可自动化得到网站核心功能的归一化结果，从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果，达到了提高处理效率，同时符合网站功能，并且能够规避黑客对网站的处置的技术效果，进而解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

[0175] 本申请提供的一种可选实施例中，对用于实现网站功能的地址元素进行拼接包括：将用于实现网站功能的地址元素保留，将不用于实现网站功能的地址元素替换为预设标识；对用于实现网站功能的地址元素以及预设标识进行拼接。

[0176] 需要说明的是，本实施例的可选或优选实施方式可以参见实施例1中的相关描述，在此不在赘述。

[0177] 实施例7

[0178] 根据本申请实施例，还提供了一种用于实施上述数据处理方法的数据处理装置，如图4所示，该装置400包括：

[0179] 获取模块402，用于获取目标地址。

[0180] 切分模块404，用于对目标地址进行切分处理，得到与目标地址对应的地址元素。

[0181] 计算模块406，用于根据目标地址的二级域名，计算地址元素的总量。

[0182] 确定模块408，用于确定属于网站功能元素的地址元素。

[0183] 拼接模块410，用于对属于网站功能元素的地址元素进行拼接。

[0184] 此处需要说明的是，上述获取模块402、切分模块404、计算模块406、确定模块408和拼接模块410对应于实施例6中的步骤S700至步骤S708，五个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例1所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例1提供的计算机终端10中。

[0185] 基于上述实施例所限定的方案可以获知，在获取到目标地址之后，首先可以对目标地址进行切分处理，得到与目标地址对应的地址元素，然后根据目标地址的二级域名，计算每个地址元素的总量，进一步确定属于网站功能元素的地址元素，最后将属于网站功能元素的地址元素进行拼接，从而得到目标地址的归一化结果。

[0186] 通过本申请上述实施例所提供的方案，基于目标地址中切分得到的地址元素的总量，确定属于网站功能元素的地址元素，并根据属于网站功能元素的地址元素，得到归一化结果，与现有技术相比，仅仅依靠网站的web日志，即可自动化得到网站核心功能的归一化结果，从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果，达到了提高处理效率，同时符合网站功能，并且能够规避黑客对网站的处置的技术效果，进而解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

[0187] 本申请提供的一种可选实施例中，拼接模块410还用于将用于实现网站功能的地址元素保留，将不用于实现网站功能的地址元素替换为预设标识；对用于实现网站功能的地址元素以及预设标识进行拼接。

[0188] 需要说明的是，本实施例的可选或优选实施方式可以参见实施例1中的相关描述，在此不在赘述。

[0189] 实施例8

[0190] 本申请的实施例可以提供一种计算机终端，该计算机终端可以是计算机终端群中的任意一个计算机终端设备。可选地，在本实施例中，上述计算机终端也可以替换为移动终端等终端设备。

[0191] 可选地，在本实施例中，上述计算机终端可以位于计算机网络的多个网络设备中的至少一个网络设备。

[0192] 在本实施例中，上述计算机终端可以执行应用程序的地址归一化处理方法中以下步骤的程序代码：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0193] 可选地，图8是根据本申请实施例的一种计算机终端的结构框图。如图8所示，该计算机终端A可以包括：一个或多个(图中仅示出一个)处理器802和存储器804。

[0194] 其中，存储器可用于存储软件程序以及模块，如本申请实施例中的地址归一化处理方法和装置对应的程序指令/模块，处理器通过运行存储在存储器内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的地址归一化处理方法。存储器可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器可进一步包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至终端A。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0195] 处理器可以通过传输装置调用存储器存储的信息及应用程序，以执行下述步骤：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0196] 可选的，上述处理器还可以执行如下步骤的程序代码：从网页日志中提取状态码为预设值的第一地址；将第一地址中的字符转换为预设字符，得到目标地址。

[0197] 可选的，上述处理器还可以执行如下步骤的程序代码：将目标地址切分为一级域名、路径及参数，并从一级域名中解析出二级域名；对一级域名按照第一规则进行拆解，得到一级域名对应的第一地址元素，对路径按照第二规则进行拆解，得到路径对应的第二地址元素，对参数按照第三规则进行拆解，得到参数对应的第三地址元素。

[0198] 可选的，上述处理器还可以执行如下步骤的程序代码：按照二级域名，对第一地址元素、第二地址元素以及第三地址元素进行分组；计算每个分组中的地址元素总个数。

[0199] 可选的，上述处理器还可以执行如下步骤的程序代码：判断地址元素总个数是否满足预设条件，预设条件为根据元素属性及网站状态得到的，网站状态用于指示元素的计数分布情况；若地址元素总个数满足预设条件，确定对应分组中的地址元素属于网站功能元素；若地址元素总个数不满足预设条件，确定对应分组中的地址元素不属于网站功能元素。

[0200] 可选的，上述处理器还可以执行如下步骤的程序代码：将属于网站功能元素的地

址元素保留，将不属于网站功能元素的地址元素替换为预设标识；对属于网站功能元素的地址元素以及预设标识进行拼接，得到归一化结果。

[0201] 处理器可以通过传输装置调用存储器存储的信息及应用程序，以执行下述步骤：获取待处理元素；从待处理元素中确定属于网站功能元素的待处理元素，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的待处理元素进行拼接，得到归一化结果。

[0202] 采用本申请实施例，提供了一种地址归一化处理的方案。在获取到目标地址之后，首先可以对目标地址进行切分处理，得到与目标地址对应的地址元素，然后根据目标地址的二级域名，计算每个地址元素的总量，进一步基于总量确定属于网站功能元素的地址元素，最后将属于网站功能元素的地址元素进行拼接，从而得到目标地址的归一化结果。

[0203] 通过本申请上述实施例所提供的方案，基于目标地址中切分得到的地址元素的总量，确定属于网站功能元素的地址元素，并根据属于网站功能元素的地址元素，得到归一化结果，与现有技术相比，仅仅依靠网站的web日志，即可自动化得到网站核心功能的归一化结果，从而实现将功能相近的地址自动、高效地压缩成一类的地址归一化结果，达到了提高处理效率，同时符合网站功能，并且能够规避黑客对网站的处置的技术效果，进而解决了现有的URL归一化处理方法处理效率低，且不符合网站功能的技术问题。

[0204] 本领域普通技术人员可以理解，图8所示的结构仅为示意，计算机终端也可以是智能手机(如Android手机、iOS手机等)、平板电脑、掌声电脑以及移动互联网设备(Mobile Internet Devices，MID)、PAD等终端设备。图8其并不对上述电子装置的结构造成限定。例如，计算机终端A还可包括比图8中所示更多或者更少的组件(如网络接口、显示装置等)，或者具有与图8所示不同的配置。

[0205] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令终端设备相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：闪存盘、只读存储器(Read-Only Memory，ROM)、随机存取器(Random Access Memory，RAM)、磁盘或光盘等。

[0206] 实施例9

[0207] 本申请的实施例还提供了一种存储介质。可选地，在本实施例中，上述存储介质可以用于保存上述实施例一所提供的地址归一化处理方法所执行的程序代码。

[0208] 可选地，在本实施例中，上述存储介质可以位于计算机网络中计算机终端群中的任意一个计算机终端中，或者位于移动终端群中的任意一个移动终端中。

[0209] 可选地，在本实施例中，存储介质被设置为存储用于执行以下步骤的程序代码：获取目标地址；对目标地址进行切分处理，得到与目标地址对应的地址元素；根据目标地址的二级域名，计算地址元素的总量；基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素；对属于网站功能元素的地址元素进行拼接，得到归一化结果。

[0210] 上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

[0211] 在本申请的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

[0212] 在本申请所提供的几个实施例中，应该理解到，所揭露的技术内容，可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为

一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元或模块的间接耦合或通信连接，可以是电性或其它的形式。

[0213] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0214] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0215] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或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中，包括若干指令用以使得一台计算机设备(可为个人计算机、服务器或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、只读存储器(ROM,Read-OnlyMemory)、随机存取存储器(RAM,Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0216] 以上所述仅是本申请的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

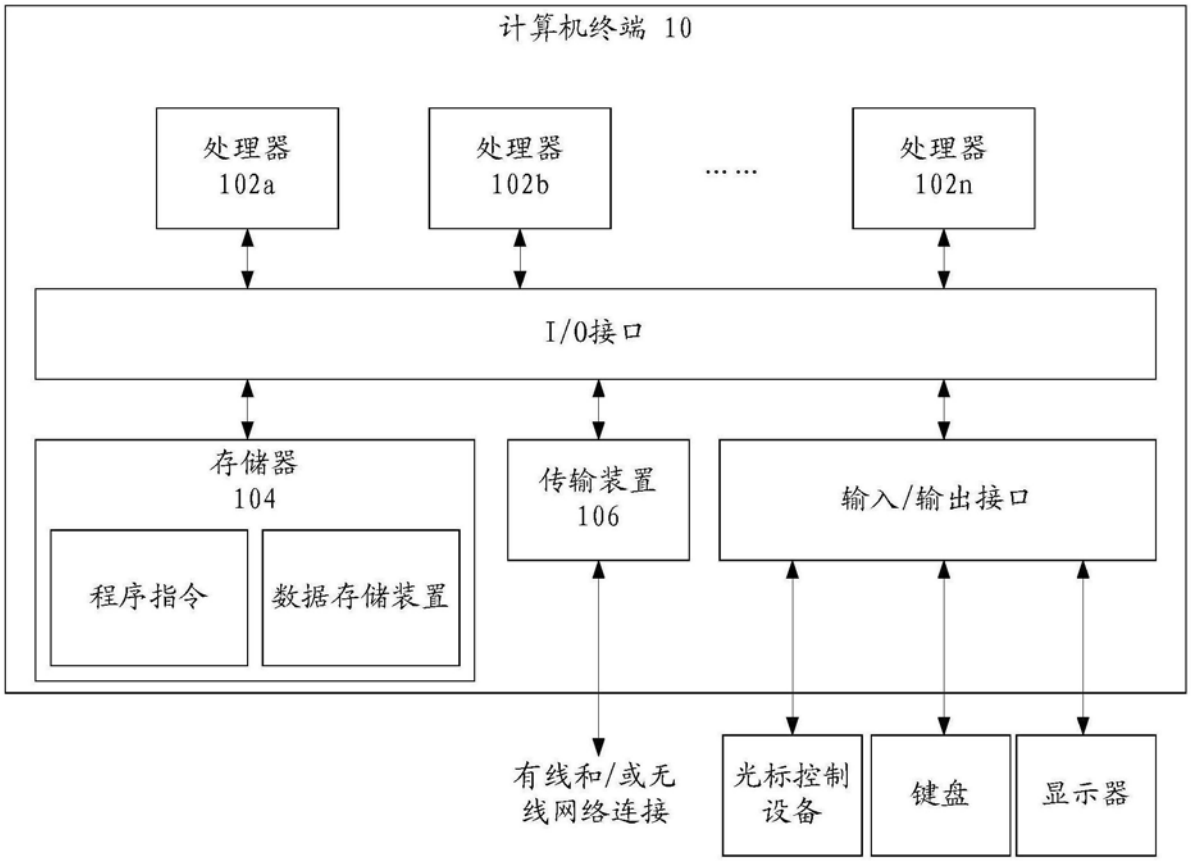

计算机终端 10
处理器
102a
处理器
102b
……
处理器
102n
I/O接口
存储器
104
程序指令
数据存储装置
传输装置
106
输入/输出接口
有线和/或无
线网络连接
光标控制
设备
键盘
显示器

图1

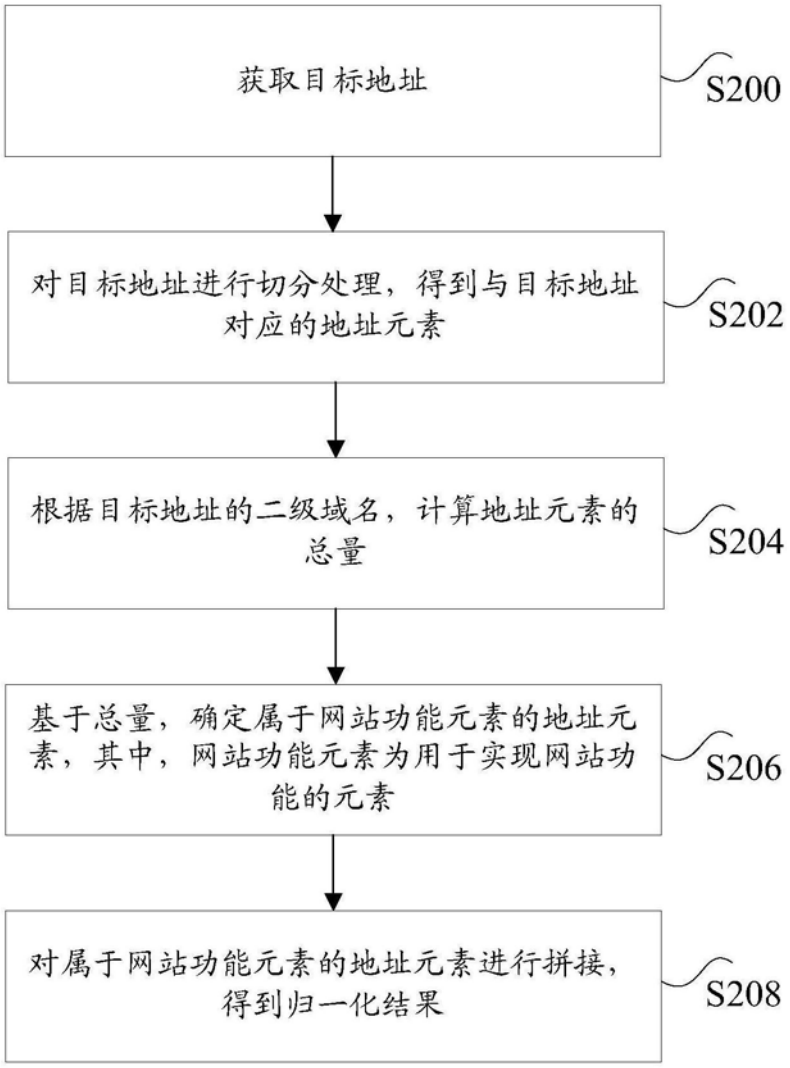

获取目标地址
S200
对目标地址进行切分处理，得到与目标地址对应的地址元素
S202
根据目标地址的二级域名，计算地址元素的总量
S204
基于总量，确定属于网站功能元素的地址元素，其中，网站功能元素为用于实现网站功能的元素
S206
对属于网站功能元素的地址元素进行拼接，得到归一化结果
S208

图2

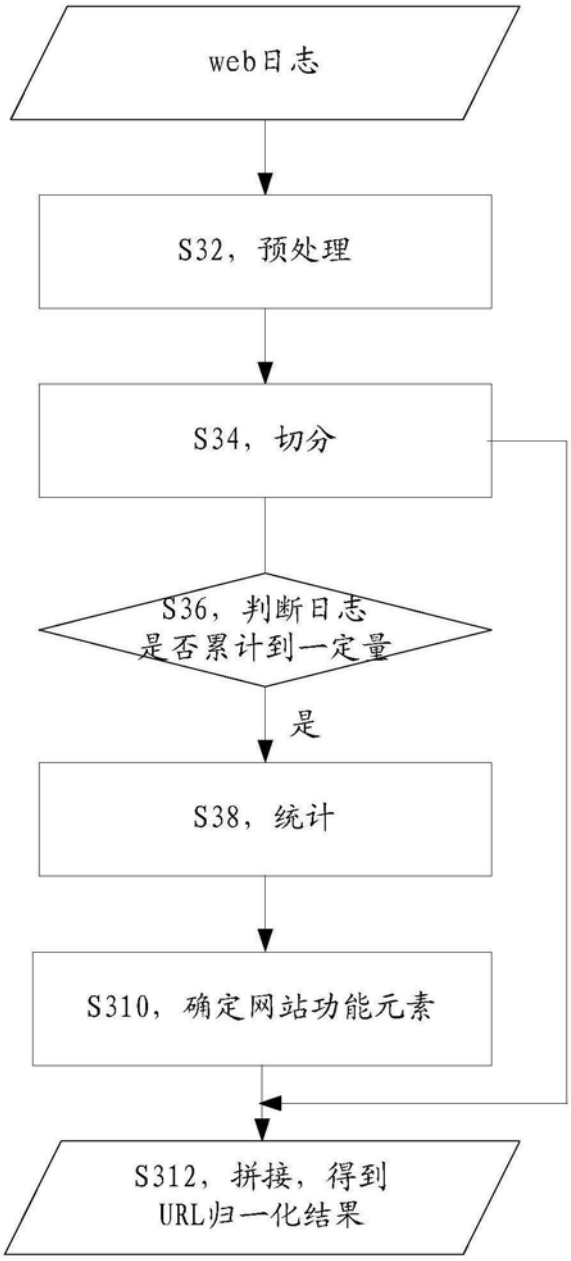

web日志
S32，预处理
S34，切分
S36，判断日志
是否累计到一定量
是
S38，统计
S310，确定网站功能元素
S312，拼接，得到
URL归一化结果

图3

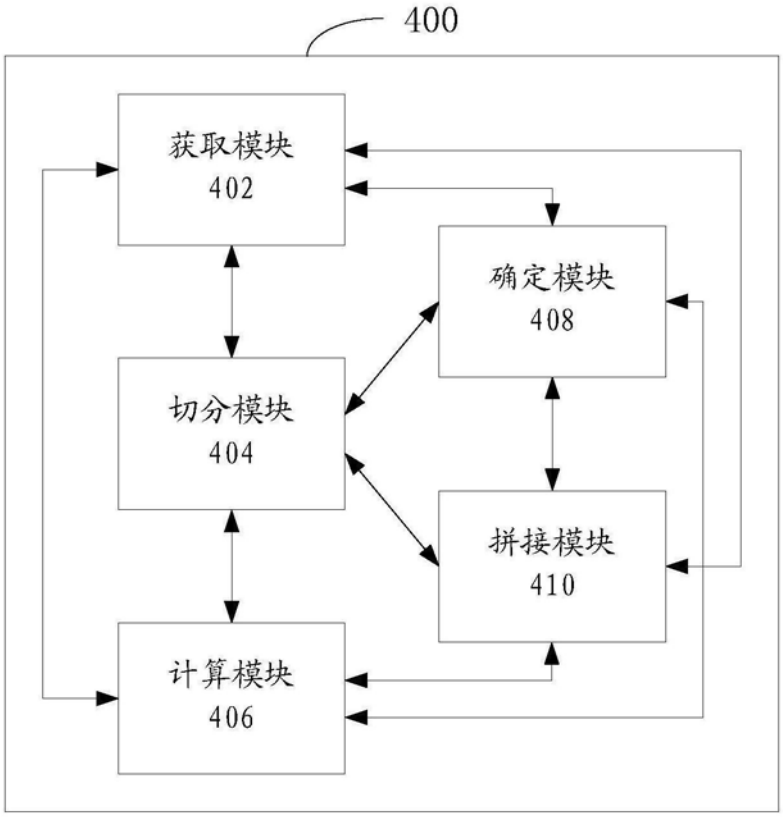

400
获取模块
402
确定模块
408
切分模块
404
拼接模块
410
计算模块
406

图4

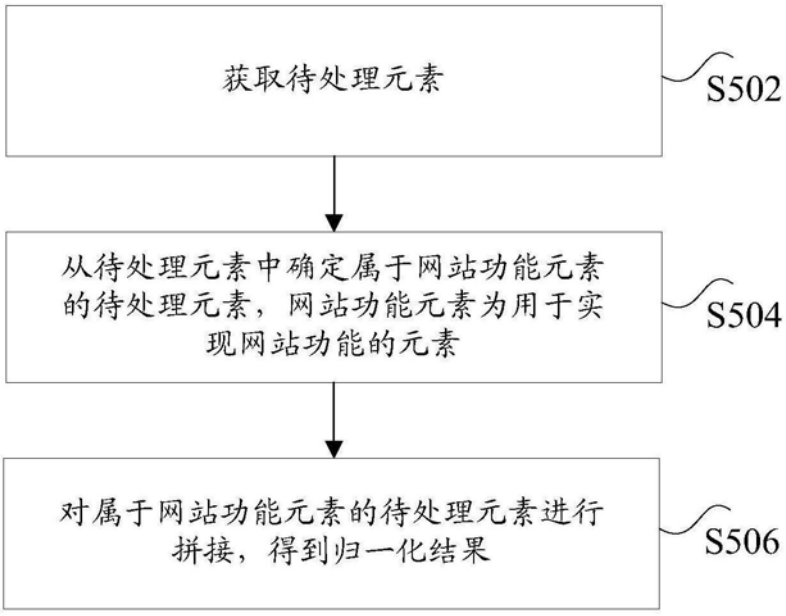

获取待处理元素
S502
从待处理元素中确定属于网站功能元素的待处理元素，网站功能元素为用于实现网站功能的元素
S504
对属于网站功能元素的待处理元素进行拼接，得到归一化结果
S506

图5

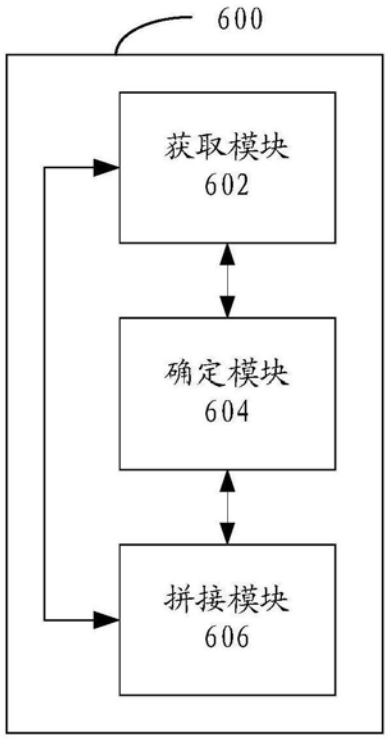


图6

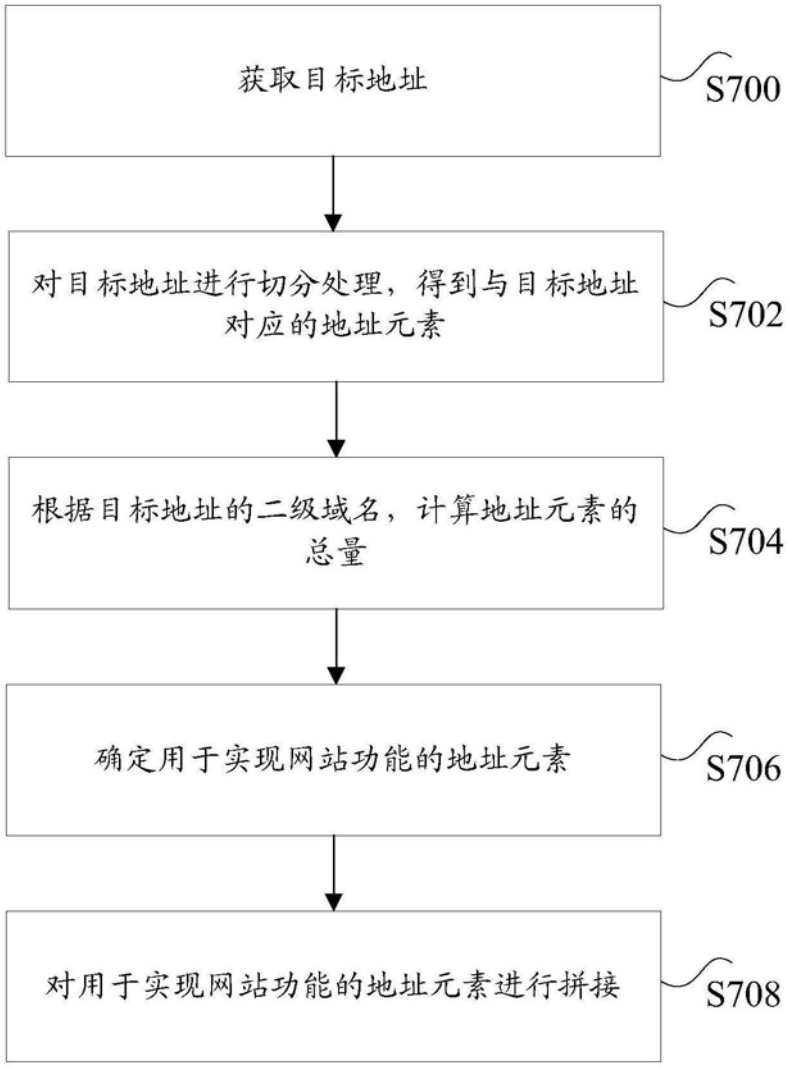

获取目标地址
S700
对目标地址进行切分处理，得到与目标地址对应的地址元素
S702
根据目标地址的二级域名，计算地址元素的总量
S704
确定用于实现网站功能的地址元素
S706
对用于实现网站功能的地址元素进行拼接
S708

图7

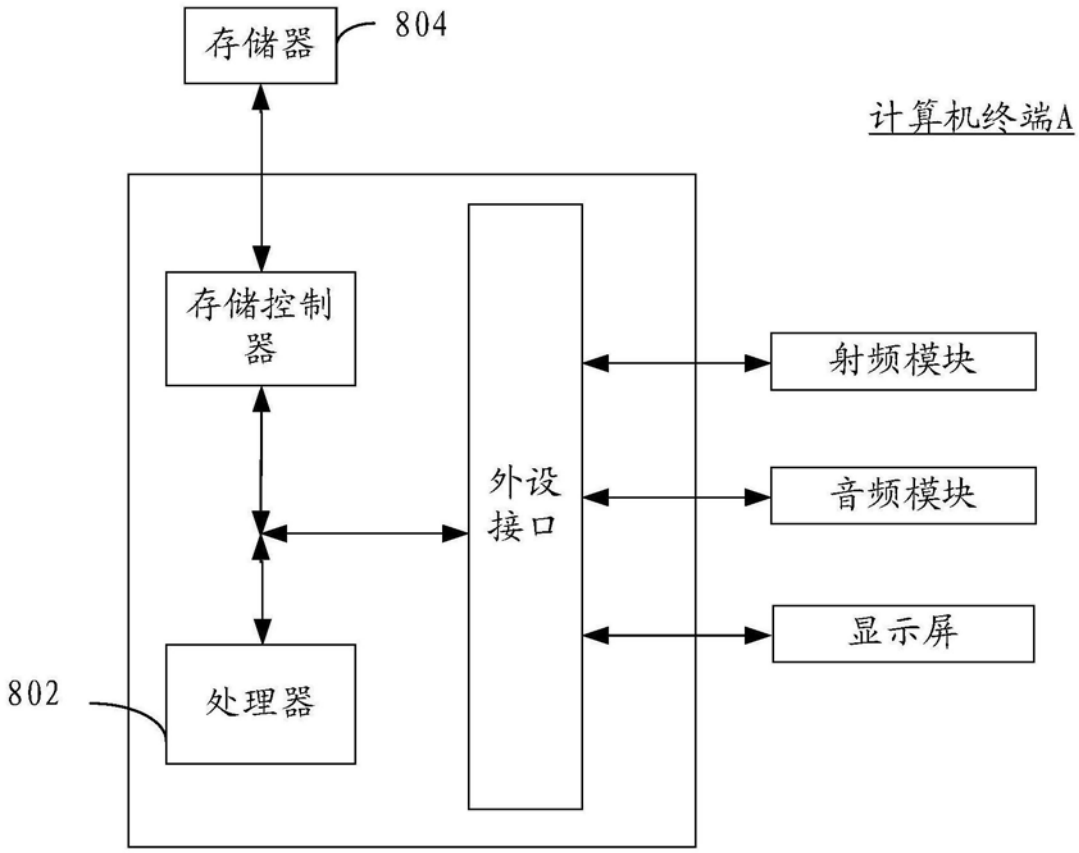


图8