(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102110132 A
(43)申请公布日 2011.06.29

(21)申请号 201010592296.2

(22)申请日 2010.12.08

(71)申请人 北京星网锐捷网络技术有限公司
地址 100036 北京市海淀区复兴路33号翠微大厦东1106

(72)发明人 魏逢一

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 黄健

(51)Int.Cl.
G06F 17/30(2006.01)
H04L 29/06(2006.01)
H04L 29/08(2006.01)

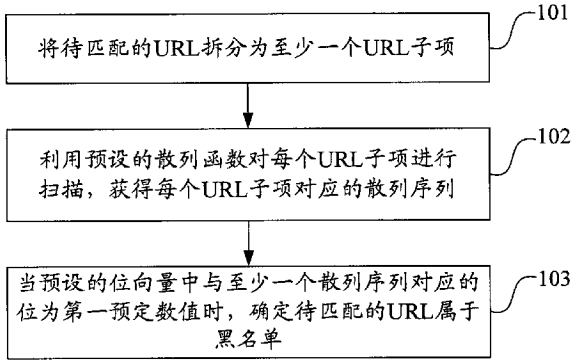
权利要求书 3 页 说明书 11 页 附图 5 页

(54)发明名称

统一资源定位符匹配查找方法、装置和网络侧设备

(57)摘要

本发明实施例提供一种统一资源定位符匹配查找方法、装置和网络侧设备，所述统一资源定位符匹配查找方法包括：将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项；利用预设的散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列；当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定所述待匹配的统一资源定位符属于黑名单。本发明实施例通过使用布隆过滤器存储黑名单中的URL条目，大大压缩了URL的存储空间；并且可以通过一次扫描获得所有URL子项的散列值，大大提升了匹配查找的性能；另外，本发明实施例能够很好的支持URL的前缀匹配和子域名匹配。



CN 102110132 A

1. 一种统一资源定位符匹配查找方法，其特征在于，包括：

将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项；

利用预设的散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列；

当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定所述待匹配的统一资源定位符属于黑名单。

2. 根据权利要求 1 所述的方法，其特征在于，所述将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项包括：

根据统一资源定位符的语法格式将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项，所述至少一个统一资源定位符子项包括所述待匹配的统一资源定位符的域名、各级父域名和前缀。

3. 根据权利要求 1 所述的方法，其特征在于，所述预设的散列函数包括预设的布隆过滤器的查询散列函数，所述利用预设的散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列包括：

利用预设的布隆过滤器的查询散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列。

4. 根据权利要求 3 所述的方法，其特征在于，所述利用预设的布隆过滤器的查询散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列包括：

利用所述查询散列函数按照从尾部至头部的顺序对所述待匹配的统一资源定位符的域名进行扫描，每扫描完一个统一资源定位符子项，则输出所述统一资源定位符子项对应的散列序列；

利用所述查询散列函数按照从头部至尾部的顺序对所述待匹配的统一资源定位符的统一资源标识符进行扫描，每扫描完一个统一资源定位符子项，则输出所述统一资源定位符子项对应的散列序列。

5. 根据权利要求 3 所述的方法，其特征在于，所述预设的位向量为所述布隆过滤器的位向量。

6. 根据权利要求 5 所述的方法，其特征在于，所述将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项之前，还包括：

设置黑名单中需要包含的统一资源定位符条目；

根据所述黑名单中包含的统一资源定位符条目的个数和预先设定的假通过率，确定所述布隆过滤器的位向量长度 L 和表示散列函数的个数 N，L 和 N 为正整数；所述表示散列函数与所述查询散列函数一一对应；

建立长度为 L 的位向量，并将所述位向量的位全部初始化为第二预定数值；

利用设置的 N 个表示散列函数对所述黑名单中的每个统一资源定位符条目进行扫描，获得所述每个统一资源定位符条目对应的散列序列；

将所述位向量中与所述散列序列对应的位设为第一预定数值。

7. 根据权利要求 6 所述的方法，其特征在于，所述利用设置的 N 个表示散列函数对所述黑名单中的每个统一资源定位符条目进行扫描，获得所述每个统一资源定位符条目对应的

散列序列包括：

利用设置的N个表示散列函数按照从尾部至头部的顺序对所述统一资源定位符条目的域名进行扫描；

当所述统一资源定位符条目包括统一资源标识符时，利用所述设置的N个表示散列函数按照从头部至尾部的顺序对所述统一资源标识符进行扫描，获得所述统一资源定位符条目对应的散列序列。

8. 一种统一资源定位符匹配查找装置，其特征在于，包括：

拆分模块，用于将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项；

第一扫描模块，用于利用预设的散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列；

第一确定模块，用于当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定所述待匹配的统一资源定位符属于黑名单。

9. 根据权利要求8所述的装置，其特征在于，所述拆分模块具体用于根据统一资源定位符的语法格式将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项，所述至少一个统一资源定位符子项包括所述待匹配的统一资源定位符的域名、各级父域名和前缀。

10. 根据权利要求8所述的装置，其特征在于，当所述预设的散列函数包括预设的布隆过滤器的查询散列函数时，所述第一扫描模块具体用于利用预设的布隆过滤器的查询散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列。

11. 根据权利要求10所述的装置，其特征在于，所述第一扫描模块包括：

第一子项扫描子模块，用于利用所述查询散列函数按照从尾部至头部的顺序对所述待匹配的统一资源定位符的域名进行扫描，每扫描完一个统一资源定位符子项，则输出所述统一资源定位符子项对应的散列序列；

第二子项扫描子模块，用于在所述第一子项扫描子模块扫描完成之后，利用所述查询散列函数按照从头部至尾部的顺序对所述待匹配的统一资源定位符的统一资源标识符进行扫描，每扫描完一个统一资源定位符子项，则输出所述统一资源定位符子项对应的散列序列。

12. 根据权利要求10所述的装置，其特征在于，还包括：

设置模块，用于设置黑名单中需要包含的统一资源定位符条目；

第二确定模块，用于根据所述黑名单中包含的统一资源定位符条目的个数和预先设定的假通过率，确定所述布隆过滤器的位向量长度L和表示散列函数的个数N，L和N为正整数；所述表示散列函数与所述查询散列函数一一对应；

建立模块，用于建立长度为L的位向量，并将所述位向量的位全部初始化为第二预定数值；

第二扫描模块，用于利用设置的N个表示散列函数对所述黑名单中的每个统一资源定位符条目进行扫描，获得所述每个统一资源定位符条目对应的散列序列；

数值设置模块，用于将所述布隆过滤器的位向量中与所述散列序列对应的位设为第一预定数值。

13. 根据权利要求 12 所述的装置，其特征在于，所述第二扫描模块具体用于利用设置的 N 个表示散列函数按照从尾部至头部的顺序对所述统一资源定位符条目的域名进行扫描；当所述统一资源定位符条目包括统一资源标识符时，再利用所述设置的 N 个表示散列函数按照从头部至尾部的顺序对所述统一资源标识符进行扫描，获得所述统一资源定位符条目对应的散列序列。

14. 一种网络侧设备，其特征在于，包括如权利要求 8-13 任意一项所述的统一资源定位符匹配查找装置。

统一资源定位符匹配查找方法、装置和网络侧设备

技术领域

[0001] 本发明涉及网络通信技术领域，尤其涉及一种统一资源定位符匹配查找方法、装置和网络侧设备。

背景技术

[0002] 互联网的迅速普及，不仅带来了诸多便利，也带来了许多负面问题，这些负面问题一般可以分为两个方面：一是娱乐性内容对人们时间的浪费；二是不良信息对人们灵魂的危害。

[0003] 对于前者，互联网上无数的娱乐性内容正在吞噬人们的宝贵时间，这些与工作无关的活动包括在线游戏、网上购物、股票交易、网上电台、流媒体和动态影像专家压缩标准音频层面 3(Moving Picture Experts Group Audio Layer III；以下简称：MP3) 下载等。据一项调查表明，企业员工全部上网活动中，50％以上都是与工作无关的，这意味着这些员工每个月拿到的薪水当中一部分与他们的工作无关。另外，专门研究上网成瘾症状的专家表示，25％到 50％的上网成瘾的人都是在办公室里上网的，如果企业对员工在上班时间的上网情况不闻不问，而且也不对某些不良网站进行禁止，那么很有可能会引发一系列严重的后果。

[0004] 对于后者，黄色网站等不良网站的泛滥，很多青少年因此而荒废学业，成为“网络海洛因”的吸食者。

[0005] 除此之外，病毒、木马网站的泛滥也在侵蚀着网络，访问互联网随时都有可能受到病毒、木马的侵袭。一旦电脑染上病毒或者木马，就可能造成个人账号等信息被盗窃，而且清除电脑病毒和木马的过程中也浪费了大量的宝贵时间。

[0006] 为了解决互联网带来的这些负面问题，维护一个健康、高效的网络环境，统一资源定位符 (Uniform Resource Locator；以下简称：URL) 过滤提供了一种简单而有效的方案，用于防止用户访问与工作无关的、不健康的和恶意的网站。

[0007] 现有技术中，URL 的语法格式如下所示：

[0008] HTTP_URL: ＝" http:" " //" host[:port][abs_path[" ? " query]]

[0009] 其中“http”代表超文本传输协议 (HyperText Transfer Protocol；以下简称：HTTP)，“host[:port]”为 HTTP 请求报文首部主 (host) 域的值，即资源站点的地址，可以是域名，也可以是因特网协议 (Internet Protocol；以下简称：IP)，如果端口号 (port) 为空，则代表端口号为 80。“abs path[" ? " query]”即资源的统一资源标识符 (Uniform Resource Identifier；以下简称：URI)。

[0010] 现有 URL 过滤系统的通常做法是预先定义好 URL 黑名单，其中包括需要屏蔽的各类网站的 URL 集合。接着 URL 过滤系统从用户发送的 HTTP 请求报文中提取出 URL 信息，并查找该 URL 是否属于黑名单中，如果属于，则阻断该 HTTP 请求；否则转发该 HTTP 请求。

[0011] 在 URL 过滤系统中，URL 匹配查找是整个过滤系统的核心，结合 URL 的语法格式，通常情况下，URL 匹配查找实现方式的选择需要考虑如下几个问题：

[0012] (1)URL匹配查找的时间开销：为了保证URL过滤系统有良好的吞吐量，URL匹配查找的时间开销必须越小越好，并且在黑名单中的URL条目数很庞大的情况下，URL过滤系统仍能很好的工作。

[0013] (2)URL匹配查找的空间开销：当黑名单中的URL条目数很庞大时（条目数达到百万级别时），必须能够将整个URL过滤系统的空间需求控制在一个合理的范围。

[0014] (3)URL过滤必须要能够支持前缀匹配：例如URL黑名单中包含URL条目“http://filter.org/path”，则当用户访问“http://filter.org/path”以及“http://filter.org/path/test”时，都能被有效禁止。

[0015] (4)URL过滤必须要能够支持子域名匹配：例如URL黑名单中包含URL条目“filter.org”，则当用户访问“http://filter.org”、“http://test.filter.org”以及“http://one.test.filter.org”时，都能被有效禁止。

[0016] 但是，在实现本发明的过程中，发明人发现：现有技术提供的URL匹配查找的实现方式均未能同时很好地解决URL匹配查找需要考虑的上述问题。

发明内容

[0017] 本发明实施例提供一种统一资源定位符匹配查找方法、装置和网络侧设备，以实现支持统一资源定位符的前缀匹配和子域名匹配，并节省统一资源定位符的存储空间。

[0018] 本发明实施例提供一种统一资源定位符匹配查找方法，包括：

[0019] 将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项；

[0020] 利用预设的散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列；

[0021] 当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定所述待匹配的统一资源定位符属于黑名单。

[0022] 本发明实施例还提供一种统一资源定位符匹配查找装置，包括：

[0023] 拆分模块，用于将待匹配的统一资源定位符拆分为至少一个统一资源定位符子项；

[0024] 第一扫描模块，用于利用预设的散列函数对每个统一资源定位符子项进行扫描，获得所述每个统一资源定位符子项对应的散列序列；

[0025] 第一确定模块，用于当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定所述待匹配的统一资源定位符属于黑名单。

[0026] 本发明实施例还提供一种网络侧设备，包括上述统一资源定位符匹配查找装置。

[0027] 本发明实施例通过布隆过滤器对待匹配的统一资源定位符拆分后的统一资源定位符子项进行匹配，只要有一个统一资源定位符子项属于黑名单，即可确定该待匹配的统一资源定位符属于黑名单；从而可以能够很好地支持统一资源定位符的前缀匹配和子域名匹配。

附图说明

[0028] 为了更清楚地说明本发明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一简单地介绍，显而易见地，下面描述中的附图是本发

明的一些实施例，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。

[0029] 图 1 为本发明统一资源定位符匹配查找方法一个实施例的流程图；

[0030] 图 2 为本发明统一资源定位符匹配查找方法另一个实施例的流程图；

[0031] 图 3 为本发明 URL 条目扫描方向一个实施例的示意图；

[0032] 图 4 为本发明 URL 扫描方向一个实施例的示意图；

[0033] 图 5 为本发明将 URL 条目装入布隆过滤器一个实施例的示意图；

[0034] 图 6 为本发明统一资源定位符匹配查找装置一个实施例的结构示意图；

[0035] 图 7 为本发明统一资源定位符匹配查找装置另一个实施例的结构示意图。

具体实施方式

[0036] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明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发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0037] 图 1 为本发明统一资源定位符匹配查找方法一个实施例的流程图，如图 1 所示，该统一资源定位符匹配查找方法可以包括：

[0038] 步骤 101，将待匹配的 URL 拆分为至少一个 URL 子项。

[0039] 具体地，可以根据 URL 的语法格式将待匹配的 URL 拆分为至少一个 URL 子项，上述至少一个 URL 子项包括待匹配的 URL 的域名、各级父域名和前缀。

[0040] 步骤 102，利用预设的散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列。

[0041] 本实施例中，该预设的散列函数可以为预设的布隆过滤器 (Bloom Filter) 的查询散列函数，则利用预设的散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列可以为：利用预设的布隆过滤器的查询散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列；

[0042] 具体地，可以先利用查询散列函数按照从尾部至头部的顺序对待匹配的 URL 的域名进行扫描，每扫描完一个 URL 子项，则输出该 URL 子项对应的散列序列；然后，再利用上述查询散列函数按照从头部至尾部的顺序对待匹配的 URL 的 URI 进行扫描，每扫描完一个 URL 子项，则输出该 URL 子项对应的散列序列。

[0043] 当然以上只是本发明实施例的一个示例，不应构成本发明实施例的限定，本发明实施例对散列函数的具体形式不作限定，只要可以对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列即可。

[0044] 步骤 103，当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定待匹配的 URL 属于黑名单。

[0045] 本实施例中，预设的位向量可以为预设的布隆过滤器的位向量，上述第一预定数值可以为 0 或 1，本实施例对此不作限定，但本实施例以位向量为布隆过滤器的位向量，第一预定数值为 1 为例进行说明。也就是说，本实施例中，当布隆过滤器的位向量中与至少一个散列序列对应的位全为 1 时，即可确定该待匹配的 URL 属于黑名单，需要对该待匹配的

URL 进行过滤处理。

[0046] 另外，本发明实施例对所使用的布隆过滤器的类型不作限定，可以使用现有的传统布隆过滤器，也可以使用计数型布隆过滤器等其他类型的布隆过滤器。

[0047] 上述实施例通过布隆过滤器对待匹配的 URL 拆分后的 URL 子项进行匹配，只要有一个 URL 子项属于黑名单，即可确定该待匹配的 URL 属于黑名单；上述实施例通过使用布隆过滤器存储黑名单中的 URL 条目，大大压缩了 URL 的存储空间，并且能够很好的支持 URL 的前缀匹配和子域名匹配。

[0048] 图 2 为本发明统一资源定位符匹配查找方法另一个实施例的流程图，如图 2 所示，该统一资源定位符匹配查找方法可以包括：

[0049] 步骤 201，设置黑名单中需要包含的 URL 条目。

[0050] 具体地，可以根据实际需要，设定黑名单中需要包含的 URL 条目；具体可以分为如下三种情况：

[0051] (1) 完整的 URL 匹配；

[0052] 举例来说，如果希望过滤“www.test.org/index.html”这个 URL，则可以将 URL 条目“www.test.org/index.html”添加至黑名单中。

[0053] (2) 前缀匹配；

[0054] 举例来说，如果希望过滤 URL 前缀为“www.test.org/path”的所有 URL，例如“www.test.org/path/test.htm”等，则可以将 URL 条目“www.test.org/path”添加至黑名单中。

[0055] (3) 子域名匹配；

[0056] 举例来说，如果希望过滤包含域名“test.org”或该域名下所有子域名的 URL，例如：“www.test.org/index.html”等，则可以将 URL 条目“test.org”添加至黑名单中。

[0057] 步骤 202，根据黑名单中包含的 URL 条目的个数和预先设定的假通过率，确定布隆过滤器的位向量长度 L 和表示散列函数的个数 N，其中，L 和 N 为正整数。

[0058] 本实施例中，假设步骤 201 中设置的黑名单中 URL 条目的个数为 M，M 为正整数，则可以根据 M 和预先设定的假通过率，确定布隆过滤器需要的位向量长度 L 以及所需要的表示散列函数的个数 N。

[0059] 下面介绍确定布隆过滤器需要的位向量长度 L 以及所需要的表示散列函数的个数 N 的两种实现方式。

[0060] （方式一）：一个使用了 N 个表示散列函数的 L 位长的布隆过滤器中装入 M 个元素后，位向量中某一位仍为 0 的概率为

[0061] $(1-1/L)^{NM}$ (1)

[0062] 则假通过率 p 为：

[0063] $p = [1-(1-1/L)^{NM}]^N$ (2)

[0064] 式 (1) 和式 (2) 中，M 为正整数，M 的大小一般是预先设定的，因此可以根据式 (2) 计算出当假通过率在可接受范围内时的 N 和 L。

[0065] 由于 N 为正整数，根据对匹配查找性能的要求，通常设定 N 的值不能大于预定阈值，因此可以采用将 N 的值逐个代入式 (2) 进行计算的方式，比如将 N = 1 代入式 (2) 即得：

[0066] $p = 1-(1-1/L)^M$ (3)

[0067] 式(3)中假通过率p是预先设定的,集合元素个数M是已知的,因此通过解方程即可求得N＝1时的L值了。同理,当N＝2,3,...时都可以计算出一个对应的L值,然后根据实际情况选取合适的N和L即可。

[0068] (方式二):首先将黑名单中的所有URL条目都装入位向量V,然后用一个测试元素集来测试,通过调整位向量的长度L以及表示散列函数的个数N,使得测试元素集的假通过率在可接受的范围内;其中,上述测试元素集中的URL条目都不属于黑名单。

[0069] 可以按照以上两种实现方式中的任意一种确定好布隆过滤器的位向量长度L和表示散列函数的个数N。

[0070] 步骤203,建立长度为L的位向量,并将该位向量的位全部初始化为第二预定数值。

[0071] 其中,该第二预定数值可以为0或1,本实施例对此不作限定,但本实施例以第二预定数值为0为例进行说明。也就是说,本实施例中,建立长度为L的位向量之后,可以先将该位向量的位全部初始化为0。

[0072] 步骤204,利用设置的N个表示散列函数对黑名单中的每个URL条目进行扫描,获得每个URL条目对应的散列序列。

[0073] 图3为本发明URL条目扫描方向一个实施例的示意图。如图3所示,本实施例在进行扫描时,首先判断待扫描的URL条目中是否包括“/”,如果包括,则可以确定该URL条目包括域名(Host)部分和URI部分,并且可以确定该URL条目中从左向右看的第一个“/”的左边为域名部分,第一个“/”的右边为URI部分;然后可以按照从尾部扫到头部的顺序扫描URL条目的域名部分,再按照从头部至尾部的顺序扫描URI部分。如果待扫描的URL条目中不包括“/”,则可以确定待扫描的URL条目仅包括域名部分,这时按照从尾部扫到头部的顺序对待扫描的URL条目的域名部分进行扫描即可。

[0074] 在实际扫描中忽略“http://”部分,因此图3所示URL条目中字符串的实际扫描顺序为:

[0075] “g→r→o→.→t→s→e→t→.→w→w→w→/→p→a→t→h→/→i→n→d→e→x→.→h→t→m→1”。

[0076] 本实施例中,布隆过滤器用到的表示散列函数可以预先设置,举例来说,可以设置布隆过滤器用到的一个表示散列函数的算法为:假设默认散列值为100,每扫描到一个字符,将该字符对应的美国信息互换标准代码(American Standard Code for Information Interchange;以下简称:ASCII)值累加到默认散列值上,在扫描完一个字符串之后,输出该字符串的散列值。其他表示散列函数可以通过类似的方法进行设置,在此不再赘述。

[0077] 当然以上仅是本发明实施例的一个示例,本发明实施例对布隆过滤器用到的表示散列函数的设置方式不作限定,例如:该表示散列函数的算法也可以为:每扫描完N个字符,将这N个字符的ASCII码值累加到预设的默认散列值上,在扫描完一个字符串之后,输出该字符串的散列值;其中,N为正整数。

[0078] 采用上述方式,利用设置的N个表示散列函数对黑名单中的每个URL条目进行扫描,即可获得每个URL条目对应的散列序列。

[0079] 步骤205,将位向量中与上述散列序列对应的位设为第一预定数值。

[0080] 其中,该第一预定数值可以为0或1,本实施例对此不作限定,但本实施例以第一

预定数值为 1 为例进行说明。

[0081] 至此，一个针对黑名单中所有 URL 条目的布隆过滤器就完成了。

[0082] 当需要查询一个 URL 是否属于黑名单时，可以执行如下步骤：

[0083] 步骤 206，将待匹配的 URL 拆分为至少一个 URL 子项。

[0084] 具体地，可以根据 URL 的语法格式将待匹配的 URL 拆分为至少一个 URL 子项，该至少一个 URL 子项包括待匹配的 URL 的域名、各级父域名和前缀。

[0085] 举例来说，对于 URL :http://www.test.org/path/index.html，其包含的子项有：

[0086] 1、一级父域名 :org

[0087] 2、二级父域名 :test.org

[0088] 3、域名 :www.test.org

[0089] 4、第一个 URL 前缀 :www.test.org/

[0090] 5、第二个 URL 前缀 :www.test.org/path

[0091] 6、完整的 URL :www.test.org/path/index.html

[0092] 步骤 207，利用预设的布隆过滤器的查询散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列。

[0093] 现有技术中，计算字符串的散列值都是将字符串从头扫到尾，然后得出一个散列值。但是在 URL 所包含的 URL 子项较多的情况下，采用这种方法进行 URL 匹配查找的性能较低。因此，本发明实施例提供一种散列值计算方法，改变字符串的扫描方向，使得通过一次扫描即可获得所有 URL 子项的散列值，从而可以有效地提高散列计算效率。

[0094] 具体地，可以先确定“http://”之后的第一个“/”为域名部分与 URI 部分的分界点，然后计算域名部分的散列值，利用查询散列函数从域名的尾部向头部扫描，每当扫到一个点号（“.”）时，表示已经扫描完一个父域名，此时输出该父域名的散列值；以此类推，直至扫描完域名部分。接下来，可以利用查询散列函数从头部向尾部扫描 URI 部分，每扫描到一个斜线（“/”），表示已经扫描完一个 URL 前缀，此时输出该 URL 前缀的散列值；以此类推，直至扫描完 URI 部分。

[0095] 仍以 URL :http://www.test.org/path/index.html 为例，其扫描过程如图 4 所示，图 4 为本发明 URL 扫描方向一个实施例的示意图。

[0096] 图 4 中“1”对应第 1 步扫描的第 1 个 URL 子项，“2”对应第 2 步扫描的第 2 个 URL 子项，以此类推。

[0097] URL :http://www.test.org/path/index.html 的扫描顺序如下所示：

[0098] g → r → o →输出第 1 个 URL 子项的散列值；

[0099] → . → t → s → e → t →输出第 2 个 URL 子项的散列值；

[0100] → . → w → w → w →输出第 3 个 URL 子项的散列值；

[0101] → / →输出第 4 个 URL 子项的散列值；

[0102] → p → a → t → h → / →输出第 5 个 URL 子项的散列值；

[0103] → i → n → d → e → x → . → h → t → m → 1 →输出第 6 个 URL 子项的散列值。

[0104] 本实施例中，布隆过滤器的查询散列函数与表示散列函数一一对应，查询散列函数与表示散列函数对每个字符采用的散列值计算方式是相同的。因此，利用查询散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列的具体实现方式可参照步骤

204 中的描述，在此不再赘述。

[0105] 步骤 208，当布隆过滤器的位向量中与至少一个散列序列对应的位为第一预定数值时，确定待匹配的 URL 属于黑名单。

[0106] 本实施例中，第一预定数值为 1，当布隆过滤器的位向量中与一个散列序列对应的位全为 1 时，即可确定该散列序列对应的 URL 子项属于黑名单，只要有一个 URL 子项属于黑名单，即可确定该待匹配的 URL 属于黑名单，需要进行过滤处理。

[0107] 反之，当布隆过滤器的位向量中与每个散列序列对应的位都不全为 1 时，可以确定待匹配的 URL 的所有 URL 子项都不属于黑名单，因此该待匹配的 URL 也不属于黑名单。

[0108] 上述实施例通过布隆过滤器对待匹配的 URL 拆分后的 URL 子项进行匹配，只要有一个 URL 子项属于黑名单，即可确定该待匹配的 URL 属于黑名单；上述实施例通过使用布隆过滤器存储黑名单中的 URL 条目，大大压缩了 URL 的存储空间，并且本发明实施例提出的散列值计算方法，可以通过一次扫描获得所有 URL 子项的散列值，大大提升了匹配查找的性能，同时实现了匹配查找性能与黑名单中的 URL 条目数无关；并且本实施例能够很好的支持 URL 的前缀匹配和子域名匹配。

[0109] 下面结合具体实例对本发明实施例的具体实施方式进行介绍。

[0110] （一）假设希望过滤域名“test.org”下所有 URL 的访问，并且，希望过滤 URL 前缀匹配“www.test2.org/sport”或“www.test3.org/news/sport”的所有 URL 的访问。同时，需要精确过滤如下几个 URL：“www.test3.org/file1.html”、“www.test3.org/file2.html”。

[0111] 假设假通过率为万分之一，即访问一万个正常的 URL，最多只能有一个 URL 被误判为属于黑名单。

[0112] 步骤一：设置黑名单中需要包含的 URL 条目。

[0113] 根据上文的场景假设，设置黑名单中需要包含的 URL 条目为：

[0114] test.org

[0115] www.test2.org/sport

[0116] www.test3.org/news/sport

[0117] www.test3.org/file1.html

[0118] www.test3.org/file2.html

[0119] 步骤二：设计布隆过滤器

[0120] 黑名单中包含 5 个 URL 条目，在具体实现时，可以根据 URL 过滤系统对性能的要求，使用 4 个表示散列函数和 4 个查询散列函数，此时可以根据本发明图 2 所示实施例步骤 202 中提供的方式二确定布隆过滤器的位向量长度 L，本例中 L 为 400 比特（即 50 字节）。

[0121] 根据本发明图 2 所示实施例步骤 204 中提供的散列计算方式，设计 4 个不同的表示散列函数 (F_1, F_2, F_3, F_4)，并对应地设计 4 个查询散列函数 (F_1', F_2', F_3', F_4')，用于 URL 匹配查找时使用。

[0122] 其中 F1 算法如下：初始化当前散列值 h 为 5381，其中该当前散列值 h 的大小可以为任意数值，本发明实施例对此不作限定，只要保证整个实施过程都采用同一个值即可。按照本发明图 2 所示实施例步骤 204 中介绍的扫描方向，对于扫描到的每个字符 c，执行 $h_1 += (h_1 << 5)+(c)$，当扫描完一个 URL 条目的所有字符时，获得的 h_1 值即为该 URL 条目的

散列值。

[0123] 对应地，F_1' 的算法如下：初始化当前散列值 h_1' 为 5381，同样 h_1' 的大小可以为任意数值，本发明实施例对此不作限定，只要保证整个实施过程都采用同一个值即可。按照本发明图 2 所示实施例步骤 207 介绍的扫描方向，对于扫描到的每个字符 c'，同样执行 h_1' += (h_1' << 5)+(c')，当扫描完一个 URL 子项的所有字符时，获得的 h_1' 值即为该 URL 子项的散列值。需要说明的是，上述公式中的 (c) 表示字符 c 的 ASCII 码值，(c') 表示字符 c' 的 ASCII 码值。

[0124] 本例中，F_2 的算法思路与 F_1 一致，其对每个扫描到的字符执行如下处理：h_2 = 31×h_2+(c)；对应地，F_2' 的算法思路与 F_1' 一致，其对每个扫描到的字符执行如下处理：h_2' = 31×h_2' +(c')；

[0125] F_3 的算法思路与 F_1 一致，其对每个扫描到的字符执行如下处理：h_3^ = (h_3 << 5)+(c)+(h_3 >> 2)；对应地，F_3' 的算法思路与 F_1' 一致，其对每个扫描到的字符执行如下处理：h_3' ^ = (h_3' << 5)+(c')+(h_3' >> 2)

[0126] F_4 的算法思路与 F_1 一致，其对每个扫描到的字符执行如下处理：h_4 = (c)+(h_4 << 6)+(h_4 << 16)-h_4；对应地，F_4' 的算法思路与 F_1' 一致，其对每个扫描到的字符执行如下处理：h_4' = (c')+(h_4' << 6)+(h_4' << 16)-h_4'。

[0127] 步骤三：将黑名单中的 URL 条目逐一装入布隆过滤器。

[0128] 首先，建立一个长度为 400 比特的位向量，然后将该位向量中的 400 个二进制位全部初始化为 0。

[0129] 然后，利用表示散列函数 (F_1, F_2, F_3, F_4) 对黑名单中的每个 URL 条目进行扫描，获得每个 URL 条目对应的散列序列 (f_1, f_2, f_3, f_4)，其中 f_1 为 h_1 对 400 求余后获得的值，f_2 为 h_2 对 400 求余后获得的值，f_3 为 h_3 对 400 求余后获得的值，f_4 为 h_4 对 400 求余后获得的值，因此 f_1、f_2、f_3 和 f_4 的取值均为 1 到 400 之间的一个值。当然本发明实施例并不仅限于此，本发明实施例对 f_n 与 h_n(n = 1,2,3,4) 之间的关系不作限定，只要可以通过预定的映射关系，使得 f_n 与 h_n(n = 1,2,3,4) 一一对应，并且 f_n 的取值在 1 到 400 之间即可。

[0130] 最后，将位向量中与每个散列序列对应的二进制位设为 1。

[0131] 在对黑名单中的每个 URL 条目都进行上述处理后，就将黑名单中的 URL 条目都装入了布隆过滤器，一个针对上述黑名单中的 URL 条目的布隆过滤器就完成了。

[0132] 以 URL 条目“test.org”为例，将该 URL 条目装入布隆过滤器的过程可以如图 5 所示，图 5 为本发明将 URL 条目装入布隆过滤器一个实施例的示意图。

[0133] 步骤四：查询一个 URL 是否属于黑名单。

[0134] 假设现有如下 URL 访问：“www.good.com/index.html”，可以先根据本发明图 2 所示实施例步骤 206 中介绍的方法将该 URL 拆分为至少一个 URL 子项，该 URL 的 URL 子项包括：

[0135] (1)com

[0136] (2)good.com

[0137] (3)www.good.com

[0138] (4)www.good.com/

[0139] (5)www.good.com/index.html

[0140] 然后，可以采用查询散列函数 (F_1', F_2', F_3', F_4')，为上述 URL 子项计算对应的散列序列 (t_{s1}, t_{s2}, t_{s3}, t_{s4})，$1 \leqslant s \leqslant 5$，s 为正整数；其中，$t_{s1}$ 为 h_1' 对 400 求余后获得的值，t_{s2} 为 h_2' 对 400 求余后获得的值，t_{s3} 为 h_3' 对 400 求余后获得的值，t_{s4} 为 h_4' 对 400 求余后获得的值，因此 t_{s1}，t_{s2}，t_{s3} 和 t_{s4} 的取值均为 1 到 400 之间的一个值。当然本发明实施例并不仅限于此，本发明实施例对 t_{sN} 与 h_N' (N = 1,2,3,4) 之间的关系不作限定，只要可以通过预定的映射关系，使得 t_{sN} 与 h_N' (N = 1,2,3,4) 一一对应，并且 t_{sN} 的取值在 1 到 400 之间即可。具体来说：

[0141] 1、URL 子项“com”对应的散列序列可以表示为 (t_{11}, t_{12}, t_{13}, t_{14})，布隆过滤器的位向量中与该散列序列对应的位不全为 1，所以该 URL 子项不属于黑名单；

[0142] 2、URL 子项“good.com”对应的散列序列可以表示为 (t_{21}, t_{22}, t_{23}, t_{24})，布隆过滤器的位向量中与该散列序列对应的位不全为 1，所以该 URL 子项不属于黑名单；

[0143] 3、URL 子项“www.good.com”对应的散列序列可以表示为 (t_{31}, t_{32}, t_{33}, t_{34})，布隆过滤器的位向量中与该散列序列对应的位不全为 1，所以该 URL 子项不属于黑名单；

[0144] 4、URL 子项“www.good.com/”对应的散列序列可以表示为 (t_{41}, t_{42}, t_{43}, t_{44})，布隆过滤器的位向量中与该散列序列对应的位不全为 1，所以该 URL 子项不属于黑名单；

[0145] 5、URL 子项“www.good.com/index.html”对应的散列序列可以表示为 (t_{51}, t_{52}, t_{53}, t_{54})，布隆过滤器的位向量中与该散列序列对应的位不全为 1，所以该 URL 子项不属于黑名单。

[0146] 由于所有的 URL 子项都不属于黑名单，因此 URL “www.good.com/index.html”不属于黑名单。

[0147] 再举一个例子，假设现有如下 URL 访问：“news.test.org/file1.html”，同样，可以先根据本发明图 2 所示实施例步骤 206 中介绍的方法将该 URL 拆分为至少一个 URL 子项，该 URL 的 URL 子项包括：

[0148] (1)org

[0149] (2)test.org

[0150] (3)news.test.org

[0151] (4)news.test.org/

[0152] (5)news.test.org/file 1.html

[0153] 然后，可以采用查询散列函数 (F_1', F_2', F_3', F_4')，为上述 URL 子项计算对应的散列序列，计算方式如上所述，在此不再赘述。

[0154] 本例中，URL 子项“test.org”对应的散列序列在布隆过滤器的位向量中的对应位全为 1，所以该 URL 子项“test.org”属于黑名单，因此该 URL“news.test.org/file 1.html”属于黑名单。

[0155] 本发明实施例提供的统一资源定位符匹配查找方法，主要是匹配的时候，将待匹配的 URL 拆分为至少一个 URL 子项，将每个 URL 子项放进布隆过滤器中看是否匹配。性能方面，布隆过滤器主要的时间开销是在散列值计算上，本发明实施例提出了一种散列值计算方式，通过一次扫描即可获得所有 URL 子项的散列值，因此大大提升了匹配查找的性能。

[0156] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序

在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0157] 图 6 为本发明统一资源定位符匹配查找装置一个实施例的结构示意图，本实施例中的统一资源定位符匹配查找装置可以作为网络侧设备，或网络侧设备的一部分，实现本发明图 1 所示实施例的流程。

[0158] 如图 6 所示，该统一资源定位符匹配查找装置可以包括：拆分模块 61、第一扫描模块 62 和第一确定模块 63。

[0159] 其中，拆分模块 61，用于将待匹配的 URL 拆分为至少一个 URL 子项；具体地，拆分模块 61 可以根据 URL 的语法格式将待匹配的 URL 拆分为至少一个 URL 子项，其中，该至少一个 URL 子项包括待匹配的 URL 的域名、各级父域名和前缀。

[0160] 第一扫描模块 62，用于利用预设的散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列。

[0161] 第一确定模块 63，用于当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定待匹配的 URL 属于黑名单；其中，该第一预定数值可以为 0 或 1，本实施例对此不作限定，但本实施例以第一预定数值为 1 为例进行说明。也就是说，本实施例中，当预设的位向量中与至少一个散列序列对应的位全为 1 时，第一确定模块 63 即可确定该待匹配的 URL 属于黑名单，需要对该待匹配的 URL 进行过滤处理。

[0162] 本实施例中的网络侧设备可以为路由器、交换机或网关设备等可以对网络访问进行管理和控制的设备。

[0163] 上述统一资源定位符匹配查找装置能够很好的支持 URL 的前缀匹配和子域名匹配。

[0164] 图 7 为本发明统一资源定位符匹配查找装置另一个实施例的结构示意图，本实施例中的统一资源定位符匹配查找装置可以作为网络侧设备，或网络侧设备的一部分，实现本发明图 2 所示实施例的流程。

[0165] 与图 6 所示的统一资源定位符匹配查找装置相比，不同之处在于，图 7 所示的统一资源定位符匹配查找装置中，当预设的散列函数包括预设的布隆过滤器的查询散列函数时，第一扫描模块 62 具体可以利用预设的布隆过滤器的查询散列函数对每个 URL 子项进行扫描，获得每个 URL 子项对应的散列序列。

[0166] 本实施例中，第一扫描模块 62 可以包括：第一子项扫描子模块 621 和第二子项扫描子模块 622；

[0167] 其中，第一子项扫描子模块 621，用于利用查询散列函数按照从尾部至头部的顺序对待匹配的 URL 的域名进行扫描，每扫描完一个 URL 子项，则输出该 URL 子项对应的散列序列；

[0168] 第二子项扫描子模块 622，用于在第一子项扫描子模块 621 扫描完成之后，利用上述查询散列函数按照从头部至尾部的顺序对待匹配的 URL 的 URI 进行扫描，每扫描完一个 URL 子项，则输出该 URL 子项对应的散列序列。

[0169] 本实施例中，预设的位向量为上述布隆过滤器的位向量，进一步地，该统一资源定位符匹配查找装置还可以包括：设置模块 64、第二确定模块 65、建立模块 66、第二扫描模块 67 和数值设置模块 68；

[0170] 其中，设置模块 64，用于设置黑名单中需要包含的 URL 条目；

[0171] 第二确定模块 65，用于根据黑名单中包含的 URL 条目的个数和预先设定的假通过率，确定上述布隆过滤器的位向量长度 L 和表示散列函数的个数 N；其中，L 和 N 为正整数，并且表示散列函数与查询散列函数一一对应；

[0172] 建立模块 66，用于建立长度为 L 的位向量，并将该位向量的位全部初始化为第二预定数值；其中，该第二预定数值可以为 0 或 1，本实施例对此不作限定，但本实施例以第二预定数值为 0 为例进行说明。也就是说，本实施例中，建立模块 66 建立长度为 L 的位向量之后，可以先将该位向量的位全部初始化为 0；

[0173] 第二扫描模块 67，用于利用设置的 N 个表示散列函数对黑名单中的每个 URL 条目进行扫描，获得每个 URL 条目对应的散列序列；具体地，第二扫描模块 67 可以利用设置的 N 个表示散列函数按照从尾部至头部的顺序对 URL 条目的域名进行扫描；当上述 URL 条目包括 URI 时，再利用设置的 N 个表示散列函数按照从头部至尾部的顺序对该 URI 进行扫描，获得 URL 条目对应的散列序列；

[0174] 数值设置模块 68，用于将布隆过滤器的位向量中与上述散列序列对应的位设为第一预定数值；其中，该第一预定数值可以为 0 或 1，本实施例对此不作限定，但本实施例以第一预定数值为 1 为例进行说明。

[0175] 本实施例中，在数值设置模块 68 将布隆过滤器的位向量中与上述散列序列对应的位设为 1 之后，一个针对黑名单中所有 URL 条目的布隆过滤器就完成了。

[0176] 本实施例中的网络侧设备可以为路由器、交换机或网关设备等可以对网络访问进行管理和控制的设备。

[0177] 上述实施例通过布隆过滤器对待匹配的 URL 拆分后的 URL 子项进行匹配，只要有一个 URL 子项属于黑名单，即可确定该待匹配的 URL 属于黑名单；上述实施例通过使用布隆过滤器存储黑名单中的 URL 条目，大大压缩了 URL 的存储空间，并且本发明实施例提出的散列值计算方法，可以通过一次扫描获得所有 URL 子项的散列值，大大提升了匹配查找的性能，同时实现了匹配查找性能与黑名单中的 URL 条目数无关；并且本实施例能够很好的支持 URL 的前缀匹配和子域名匹配。

[0178] 本领域技术人员可以理解附图只是一个优选实施例的示意图，附图中的模块或流程并不一定是实施本发明所必须的。

[0179] 本领域技术人员可以理解实施例中的装置中的模块可以按照实施例描述进行分布于实施例的装置中，也可以进行相应变化位于不同于本实施例的一个或多个装置中。上述实施例的模块可以合并为一个模块，也可以进一步拆分成多个子模块。

[0180] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

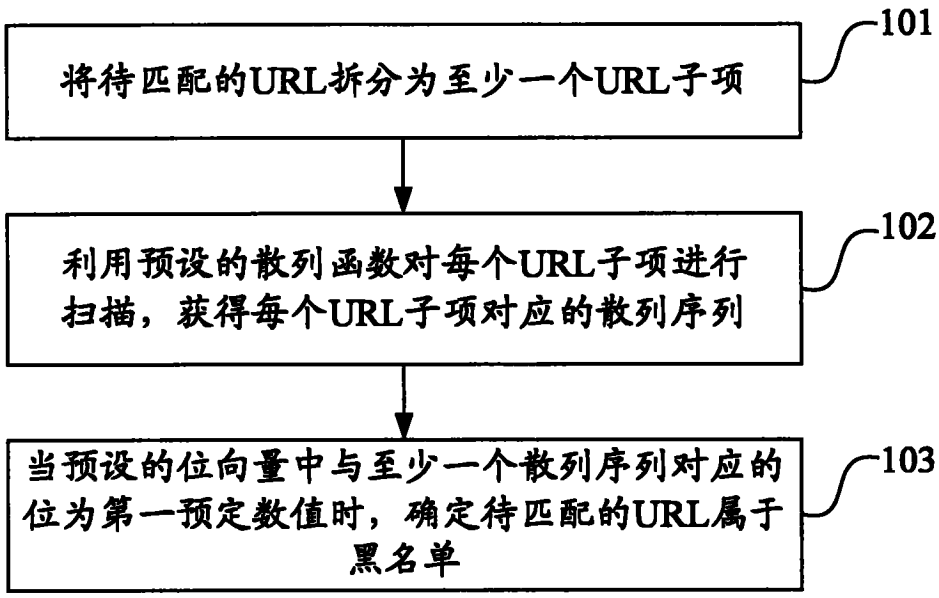

将待匹配的URL拆分为至少一个URL子项
101
利用预设的散列函数对每个URL子项进行扫描，获得每个URL子项对应的散列序列
102
当预设的位向量中与至少一个散列序列对应的位为第一预定数值时，确定待匹配的URL属于黑名单
103

图 1

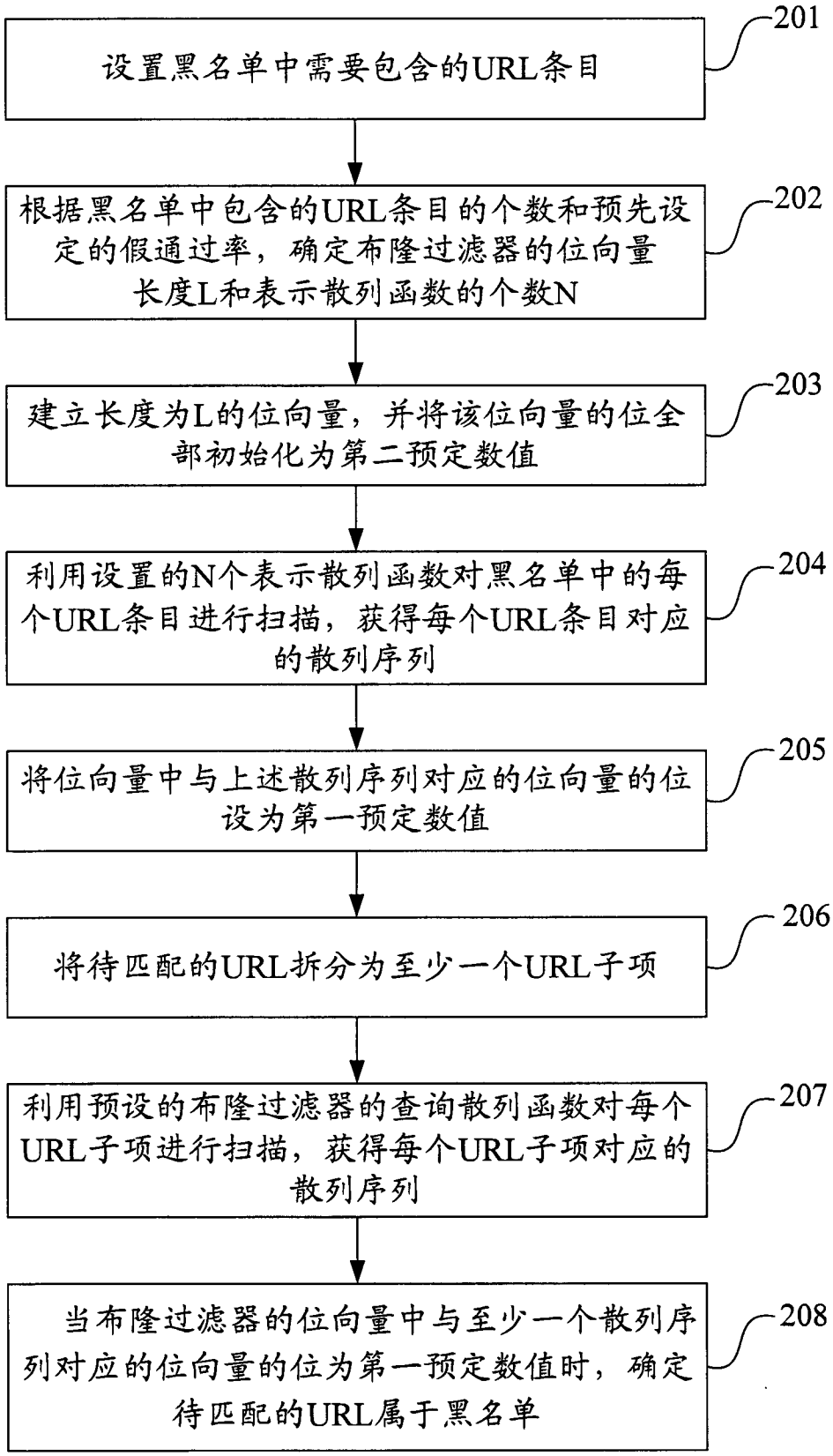

设置黑名单中需要包含的URL条目
201
根据黑名单中包含的URL条目的个数和预先设定的假通过率，确定布隆过滤器的位向量长度L和表示散列函数的个数N
202
建立长度为L的位向量，并将该位向量的位全部初始化为第二预定数值
203
利用设置的N个表示散列函数对黑名单中的每个URL条目进行扫描，获得每个URL条目对应的散列序列
204
将位向量中与上述散列序列对应的位向量的位设为第一预定数值
205
将待匹配的URL拆分为至少一个URL子项
206
利用预设的布隆过滤器的查询散列函数对每个URL子项进行扫描，获得每个URL子项对应的散列序列
207
当布隆过滤器的位向量中与至少一个散列序列对应的位向量的位为第一预定数值时，确定待匹配的URL属于黑名单
208

图 2

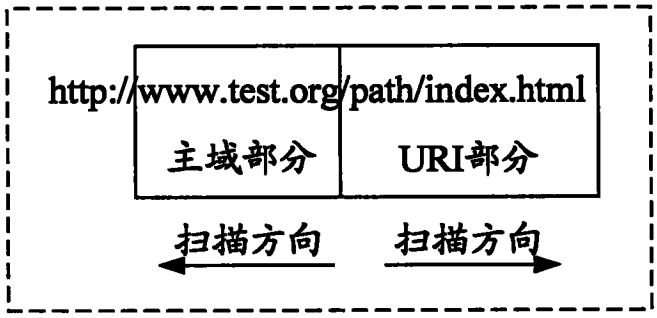

http://www.test.org/path/index.html
主域部分
URI部分
扫描方向
扫描方向

图 3

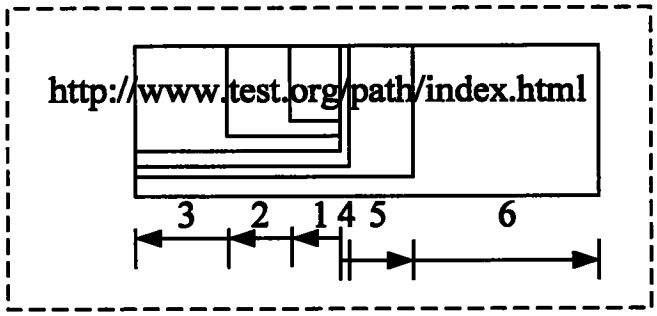

http://www.test.org/path/index.html
3
2
1
4
5
6

图 4

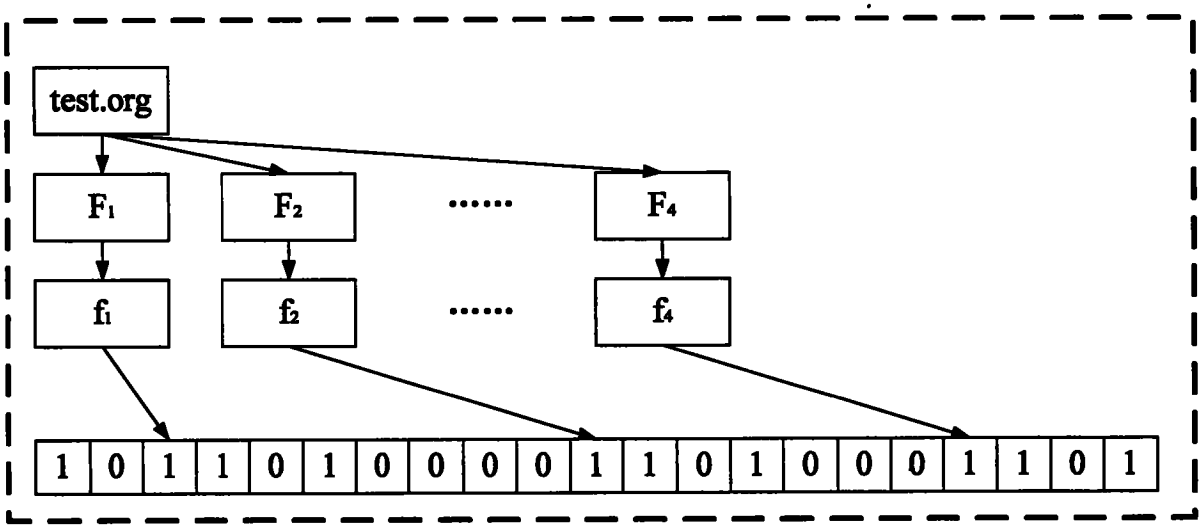

test.org
F1
F2
.....
F4
f1
f2
.....
f4
1 0 1 1 0 1 0 0 0 0 1 1 0 1 0 0 0 1 1 0 1

图 5

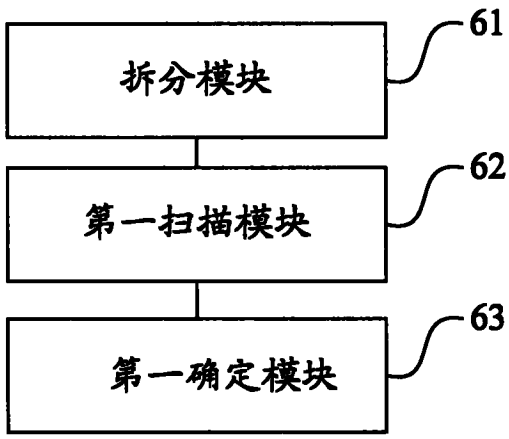


图 6

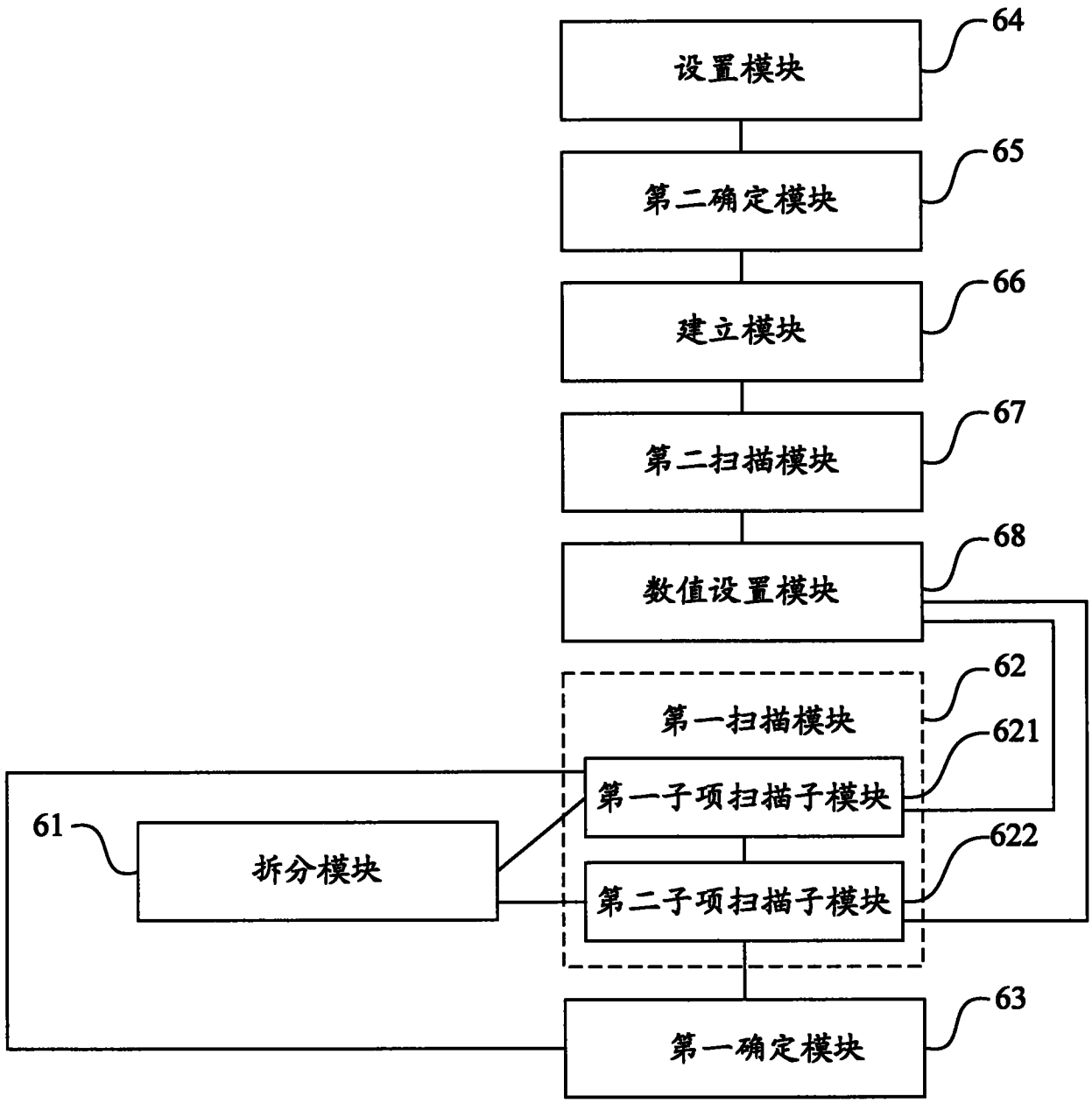

设置模块
64
第二确定模块
65
建立模块
66
第二扫描模块
67
数值设置模块
68
62
第一扫描模块
621
第一子项扫描子模块
61
拆分模块
622
第二子项扫描子模块
63
第一确定模块

图 7