(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 105763633 B
(45)授权公告日 2019.05.21

(21)申请号 201610230263.0

(22)申请日 2016.04.14

(65)同一申请的已公布的文献号
申请公布号 CN 105763633 A

(43)申请公布日 2016.07.13

(73)专利权人 上海牙木通讯技术有限公司
地址 200030 上海市徐汇区番禺路1028号305室

(72)发明人 张大顺

(74)专利代理机构 上海立群专利代理事务所（普通合伙）31291
代理人 毛立群 杨楷

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 29/12(2006.01)
G06F 16/951(2019.01)

(56)对比文件
CN 105005600 A,2015.10.28,说明书第40-116段.
CN 104065532 A,2014.09.24,说明书第68-80段.
CN 105357054 A,2016.02.24,全文.
US 2010250742 A1,2010.09.30,全文.

审查员 王亭

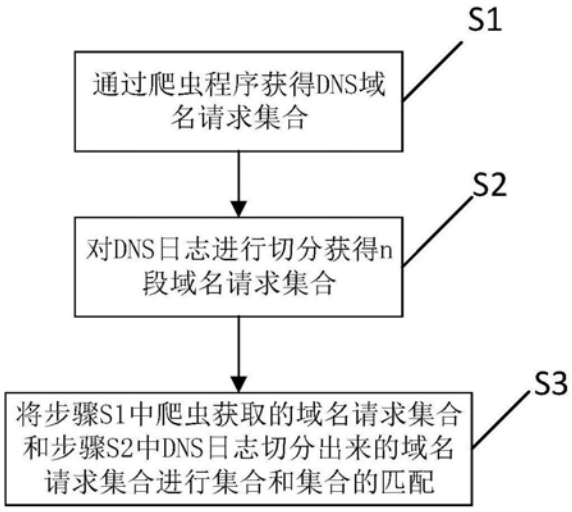
权利要求书1页 说明书3页 附图2页

(54)发明名称

一种域名和网站访问行为的关联方法

(57)摘要

本发明提出了一种域名和网站访问行为的关联方法，包括以下步骤：步骤S1：通过爬虫程序模拟用户访问网站的行为，获得当次HTTP请求中所有的DNS域名请求，即抓取的DNS域名请求集合；步骤S2：对DNS日志进行切分获得n段域名请求集合，n为大于等于1的整数；以及步骤S3：将步骤S1中抓取的DNS域名请求集合和步骤S2中DNS日志切分出来的域名请求集合进行集合和集合的匹配，如果DNS日志切分出来的一段域名请求集合等于或被包含于所述抓取的DNS域名请求集合，则认为所述DNS日志表示用户点击了所述爬虫程序抓取时请求的URL的域名。通过本发明的域名和网站访问行为的关联方法，通过DNS日志也能实现分析用户的互联网浏览行为。



CN 105763633 B

1.一种域名和网站访问行为的关联方法,其特征在于,包括如下步骤:

步骤S1:通过爬虫程序模拟用户访问网站的行为,获得当次HTTP请求中所有的DNS域名请求,即抓取的DNS域名请求集合;

步骤S2:对DNS日志进行切分获得n段域名请求集合,n为大于等于1的整数;以及

步骤S3:将步骤S1中抓取的DNS域名请求集合和步骤S2中DNS日志切分出来的n段域名请求集合进行集合和集合的匹配,如果DNS日志切分出来的一段域名请求集合等于或被包含于所述抓取的DNS域名请求集合,则认为所述DNS日志表示用户点击了所述爬虫程序抓取时请求的URL的域名;

其中,步骤S2中,对所述DNS日志进行切分包括2次切分,即先根据源IP进行切分,再根据时间戳之差进行切分。

2.根据权利要求1所述的域名和网站访问行为的关联方法,其特征在于,步骤S2中,所述DNS日志是访问行为当天的DNS日志。

3.根据权利要求1所述的关联方法,其特征在于,根据源IP对DNS日志进行切分是获得一段时间内相同源IP的连续的DNS日志。

4.根据权利要求3所述的关联方法,其特征在于,所述根据时间戳之差对日志进行切分是对根据源IP切分后的日志再根据DNS日志的时间戳之间的差进行切分,如果两个DNS日志的时间戳之间的差大于规定时间长度,则切开所述两个DNS日志。

5.根据权利要求4所述的关联方法,其特征在于,所述规定时间长度为3秒。

一种域名和网站访问行为的关联方法

技术领域

[0001] 本发明涉及互联网DNS域名解析领域以及网络爬虫技术，尤其涉及一种域名和网站访问行为的关联方法。

背景技术

[0002] DNS(Domain Name System，域名系统)，是因特网上作为域名和IP地址相互映射的一个分布式数据库，能够使用户更方便的访问互联网，而不用去记住能够被机器直接读取的IP数串。“DNS域名解析技术”是指：当用户需要访问一个网站时，他需要在浏览器中输入这个网站的域名。敲击回车后浏览器会先发起一个DNS请求，通过DNS技术，浏览器可以获取这个域名对应的服务器IP地址，然后再对这个IP地址发起HTTP请求。

[0003] 网络爬虫技术，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。其模拟用户对网站发起HTTP请求并记录该过程中产生的DNS请求。

[0004] DNS的数据的价值一直没有的到相应的重视，仅仅被认为是一种IP和域名的对应关系，所以目前市场上并没有人通过DNS数据去进行相应的关联。

发明内容

[0005] 本发明提出了一种域名和网站访问行为的关联方法，通过DNS日志采集和网络爬虫技术的结合，使得通过DNS日志也能分析用户的互联网浏览行为。

[0006] 本发明的一种域名和网站访问行为的关联方法，包括如下步骤：步骤S1：通过爬虫程序模拟用户访问网站的行为，获得当次HTTP请求中所有的DNS域名请求，即抓取的DNS域名请求集合；步骤S2：对DNS日志进行切分获得n段域名请求集合，n为大于等于1的整数；以及步骤S3：将步骤S1中抓取的DNS域名请求集合和步骤S2中DNS日志切分出来的域名请求集合进行集合和集合的匹配，如果DNS日志切分出来的一段域名请求集合等于或被包含于所述抓取的DNS域名请求集合，则认为所述DNS日志表示用户点击了所述爬虫程序抓取时请求的URL的域名。

[0007] 优选地，步骤S2中，所述DNS日志是访问行为当天的DNS日志。

[0008] 优选地，步骤S2中，对所述DNS日志进行切分包括2次切分，即先根据源IP进行切分，再根据时间戳之差进行切分。

[0009] 优选地，根据源IP对DNS日志进行切分是获得一段时间内相同源IP的连续的DNS日志。

[0010] 优选地，所述根据时间戳之差对日志进行切分是对根据源IP切分后的日志再根据DNS日志的时间戳之间的差进行切分，如果两个DNS日志的时间戳之间的差大于规定时间长度，则切开所述两个DNS日志。

[0011] 优选地，所述规定时间长度为3秒。

[0012] 通过本发明的域名和网站访问行为的关联方法，通过DNS日志也能实现分析用户的互联网浏览行为。

附图说明

[0013] 图1是爬虫程序抓取的DNS域名请求集合的示意图。

[0014] 图2是本发明的域名和网站访问行为的关联方法的流程图。

具体实施方式

[0015] 以下，将结合附图和实施例对发明进行详细说明。以下实施例并不是对本发明的限制。在不背离发明构思的精神和范围下，本领域技术人员能够想到的变化和优点都被包括在本发明中。

[0016] 如前所提到的，DNS(Domain Name System，域名系统)，是因特网上作为域名和I P地址相互映射的一个分布式数据库，能够使用户更方便地访问互联网，而不用去记住能够被机器直接读取的IP数串。当用户访问一个网站时，先在浏览器中输入这个网站的域名，敲击回车后浏览器会先发起一个DNS请求，通过DNS技术，浏览器可以获取这个域名对应的服务器IP地址，然后再对这个IP地址发起HTTP请求。这就是DNS域名解析技术。

[0017] 在上述域名解析的过程中，会产生DNS日志。DNS日志会记录每次DNS请求的应答内容，几乎能记录所有用户请求的域名信息。DNS日志的格式如下所示：

[0018] 14.***.***.10|www.baidu.com|20141211035932|180.***.***.107;180.***.***.108|0源IP|域名|时间戳|解析IP|状态码

[0019] 即DNS日志包括“源IP”，“域名”，“时间戳”，“解析IP”和“状态码”五部分内容。下面结合图1详细说明本发明的域名和网站访问行为的关联方法。

[0020] 首先，通过爬虫程序模拟用户访问网站的行为，获得当次HTTP请求中所有的DNS域名请求，即抓取的DNS域名请求集合(步骤S1)。例如，打开某个页面或点击某个URL(链接)，爬虫程序会抓取当次HTTP请求中所有的DNS域名请求。由于当一个用户点击一个URL时，除了请求当前URL的域名外还会请求一些其他的域名，通过爬虫技术可以获取点击该URL后产生的所有DNS域名请求。这里，统一资源定位符(URL)是对可以从互联网上得到的资源的位置和访问方法的一种简洁的表示，是互联网上标准资源的地址。互联网上的每个文件都有一个唯一的URL，它包含的信息指出文件的位置以及浏览器应该怎么处理它。

[0021] 例如，用户点击一个具体的URL(链接)，如下所示：

[0022] “http://baike.baidu.com/link?url＝Lm-TkKUzV687IRoPCDVUAG5qslgMyZtNa6e6A3nPnWXorcXEAI15OO6XHZWpTJat”。

[0023] 爬虫程序会抓取点击该URL后产生的所有DNS域名请求，即DNS域名请求集合，具体如图1所示。

[0024] 接下来，对DNS日志进行切分获得n段域名请求集合，n为大于等于1的整数(步骤S2)。这里，DNS日志一般为访问行为当天的日志。所述切分包括2次切分，即先根据源IP进行切分，再根据时间戳之差进行切分。

[0025] 1)对DNS日志根据源IP进行切分，即日志的源IP不同，则将连续的日志切分开。根据源IP切分是获得一段时间内相同源IP的连续的DNS日志。如下所示：

[0026] 1.1.1.1|www.baidu.com|20141211035932|180.***.***.107;180.***.***.108|0

[0027] 1.1.1.1|www.qq.com|20141211035932|180.***.***.107;180.***.***.108|0

[0028] ----------日志切割线----------

[0029] 2.2.2.2|www.baidu.com|20141211035932|180.***.***.107;180.***.***.108|0

[0030] 2.2.2.2|www.qq.com|20141211035932|180.***.***.107;180.***.***.108|0

[0031] 2)根据时间戳之差切分是指对根据源IP切分后的日志再根据DNS日志的时间戳之间的差值进行切分。如果两个连续日志之间的时间戳之差大于规定时间长度,则被切分开(切分的原因是日志的时间间隔过久则被认为是两个不同的行为)。该规定时间长度可以根据需要调整。本实施例中,所述规定时间长度为3秒,即时间戳相隔大于3秒会被切分开。

[0032] 例如,对源IP2.2.2.2的DNS日志进一步根据其时间戳之差值进行切分,如下所示。(时间戳20141211035932表示2014年12月11日3点59分32秒)

[0033] 源IP|域名|时间戳|解析IP|状态码

[0034] 2.2.2.2|www.baidu.com|20141211000001|180.***.***.107;180.***.***.108|0

[0035] 2.2.2.2|a.qq.com|20141211000002|180.***.***.107;180.***.***.108|0

[0036] 2.2.2.2|b.baidu.com|20141211000003|180.***.***.107;180.***.***.108|0

[0037] 2.2.2.2|c.tanx.com|20141211000004|180.***.***.107;180.***.***.108|0

[0038] 2.2.2.2|c.allyes.com|20141211000005|180.***.***.107;180.***.***.108|0

[0039] ----------日志切割线----------

[0040] 2.2.2.2|www.sina.com|20141211000009|180.***.***.107;180.***.***.108|0

[0041] 如上所示,由于时间戳20141211000005的05秒和20141211000009的09秒之间相差4秒(大于3秒),所以日志被切开。

[0042] www.baidu.com,a.qq.com,b.baidu.com,c.tanx.com,c.tanx.com即为DNS日志中的一段域名请求集合。

[0043] 接着将步骤S1中爬虫获取的域名请求集合和步骤S2中DNS日志切分出来的域名请求集合进行集合和集合的匹配(步骤S3)。匹配的规则是【(a,b,c)＝(b,c,a)＝(a,c,b)】。

[0044] 匹配日志后,如果DNS日志的一段域名请求集合包含在爬虫抓取的域名请求集合内,或两个集合相同,即认为该DNS日志表示用户点击了该域名(即爬虫抓取时请求的URL的域名)。例如:

[0045] 爬虫抓取的URL是www.a.com/doc/1234(该URL为一个用户的点击行为)。抓取的所有域名请求集合A为“www.a.com、www.b.com、www.c.com、www.d.com、www.e.com”。

[0046] DNS日志切分后有一段的域名请求集合B为“www.a.com、www.b.com、www.e.com、www.d.com”

[0047] 如上,B集合包含在A集合内,则认为域名请求集合B反映了域名集合A映射的www.a.com/doc/1234这一用户访问行为。这样,通过DNS日志也能实现分析用户的互联网浏览行为。

[0048] 综上所述仅为本发明的较佳实施例,并非用来限定本发明的实施范围。即凡依本发明申请专利范围的内容所作的等效变化与修饰,都应属于本发明的技术范畴。

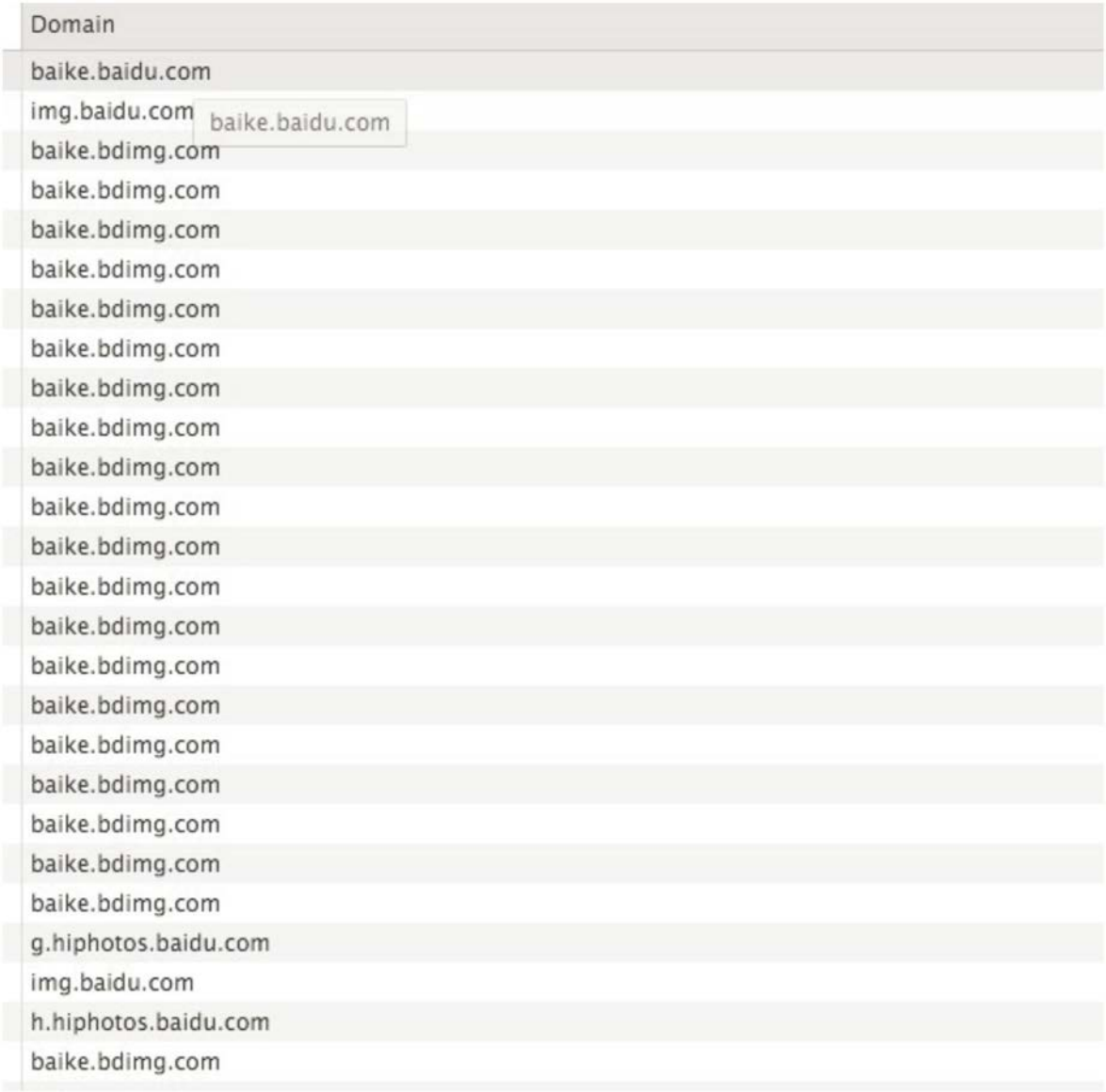

Domain
baike.baidu.com
img.baidu.com
baike.baidu.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
baike.bdimg.com
g.hiphotos.baidu.com
img.baidu.com
h.hiphotos.baidu.com
baike.bdimg.com

图1

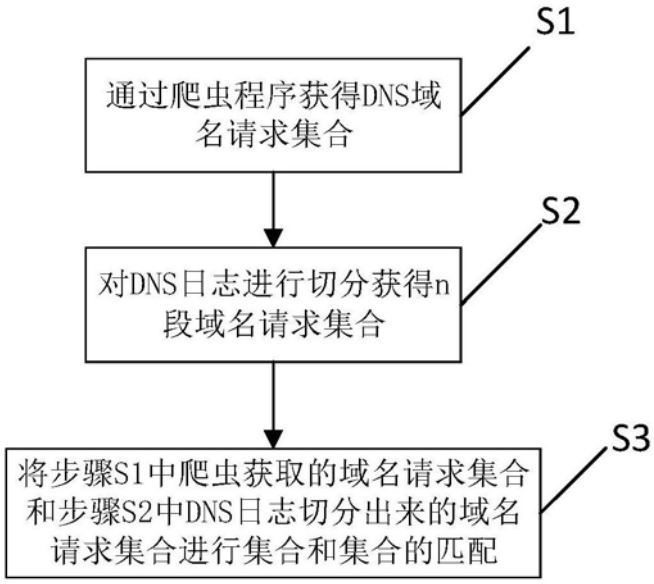

S1
通过爬虫程序获得DNS域名请求集合
S2
对DNS日志进行切分获得n段域名请求集合
S3
将步骤S1中爬虫获取的域名请求集合和步骤S2中DNS日志切分出来的域名请求集合进行集合和集合的匹配

图2