(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 104239353 B
(45)授权公告日 2019.12.31

(21)申请号 201310248048.X

(22)申请日 2013.06.20

(65)同一申请的已公布的文献号
申请公布号 CN 104239353 A

(43)申请公布日 2014.12.24

(73)专利权人 上海博达数据通信有限公司
地址 201203 上海市浦东新区张江高科技园区居里路123号

(72)发明人 汪革 彭双庭 郭海涛 陈肖方宇

(74)专利代理机构 上海天翔知识产权代理有限公司 31224
代理人 刘粉宝

(51)Int.Cl.
G06F 16/958(2019.01)

(56)对比文件
CN 102098328 A,2011.06.15,
CN 101958912 A,2011.01.26,
CN 102098229 A,2011.06.15,
CN 103118007 A,2013.05.22,
WO 01/55905 A1,2001.08.02,
石彪.网络环境下的日志监控与安全审计系统研究与实现.《中国优秀硕士学位论文全文数据库》.2006,(第3期),正文第21-30,44-48页.
石彪.网络环境下的日志监控与安全审计系统研究与实现.《中国优秀硕士学位论文全文数据库》.2006,(第3期),正文第21-30,44-48页.
韩巧玲.基于分区的网络行为监控系统数据库设计与优化研究.《中国优秀硕士学位论文全文数据库》.2011,(第4期),正文第26-39页.

审查员 林坚

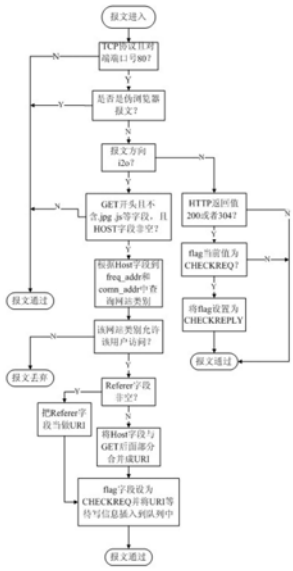
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种WEB分类控制和日志审计的方法

(57)摘要

本发明公开了一种WEB分类控制和日志审计的方法,其首先,识别出网站访问报文,并将所要记录的信息插入队列中;再通过定时器任务,将队列中的网站访问日志写入数据库中。该方法基于模式匹配、DPI识别以及嵌入式数据库技术,有效解决现有技术所存在的问题。



CN 104239353 B

1.一种WEB分类控制和日志审计的方法,其特征在于,所述方法包括如下步骤:

(1)识别出网站访问报文,并将所要记录的信息插入队列中;

在识别报文后,将识别出的HTTP请求报文与网址库中存储的网址信息进行类型匹配,判断HTTP请求报文对应的网站是否属于允许访问的网站类别范围;若HTTP请求报文对应的网站判断为属于允许访问的网站,进一步获取报文中Referer字段的信息,根据Referer字段采取不同的策略,如果该字段非空,就记录该字段内容,否则将Host字段与GET字段后面的内容拼接后记录到内存队列中,以避免记录大量无用信息;

在收到一条HTTP请求报文并在获取访问网站的网址信息及其控制类别的以后,根据报文的五元组信息建立相应的流节点,并将节点加入到内存队列中,每一个新加入队列的流节点将自动被维护一个名为CHECKREQ的初始状态;在此状态基础上如果收到访问网站的回应报文且http状态值是200,即成功或者304,即未修改,就将队列中相应的节点流的状态切换为CHECKREPLY;且在后续的定时器定时写数据库操作中,只有状态为CHECKREPLY的才会真正被移出队列并写入数据库中;由此使得在网站访问失败的情况下就不会留下网站访问的记录,以过滤掉浏览器发出的不可达的网站的访问日志信息,减少无效访问日志的审计和处理;

(2)通过定时器任务,将队列中的网站访问日志写入数据库中。

2.根据权利要求1所述的一种WEB分类控制和日志审计的方法,其特征在于,所述步骤(1)中采用DPI技术识别出真正的HTTP请求报文。

3.根据权利要求2所述的一种WEB分类控制和日志审计的方法,其特征在于,采用DPI技术分析报文中是否含有Accept-Encoding:字段,如果没有含有该字段,则认为是伪装成浏览器报文。

4.根据权利要求1所述的一种WEB分类控制和日志审计的方法,其特征在于,所述网址库包括存放完整网址库信息的第三级比较表、存放从第三级比较表中挑选出的热门网站的网址信息的第二级比较表、存放用户最近最经常访问网站的网址信息的第一级比较表,在进行类型匹配时,先与第一级比较表中的信息进行匹配,若未找到则继续匹配第二级比较表,仍未找到则匹配第三级比较表,如果还是没有找到,记录网址类型为未分类。

5.根据权利要求4所述的一种WEB分类控制和日志审计的方法,其特征在于,在网站访问日志保存到数据库以后,所述第一级比较表中存储的信息根据数据库中记录的网站访问日志信息不断的动态更新。

6.根据权利要求1所述的一种WEB分类控制和日志审计的方法,其特征在于,所述数据库包括两张数据表,第一张数据表存储最近一段时间内的网站访问日志信息,第二张数据表存储所有的网站访问日志信息,并且规定时间内将第一张数据表内数据移至第二张数据表中。

一种WEB分类控制和日志审计的方法

技术领域

[0001] 本发明涉及计算机网络技术领域，具体涉及网站访问行为管理和监控技术。

背景技术

[0002] 由于在企业、单位中，经常会有一些员工在上班时间浏览一些与工作无关的娱乐网站，导致工作效率下降。因此这些企业、单位就希望能够采取某些方法对员工的网站访问行为进行管理，禁止其在上班时间访问某些类别的网站，以及保存其上网的历史记录。

[0003] 目前有一些厂商已经提供了该功能，但是许多都是简单地记录http报文中的Host字段或者URI字段的内容，这会导致记录的信息过于庞大或者不够准确，因此实用性不强。

[0004] 归结起来主要有以下几个不足之处：

[0005] (1)有的提供专门的设备接在路由器WAN口后面，如果有多个WAN口的话就需要多台，很不方便。

[0006] (2)会把许多p2p下载软件发出的伪浏览器报文误识别为网站访问报文。

[0007] (3)有的为了减少网页访问记录数量，就只记录http请求报文中HOST字段的值，但是该字段的值表示的是网站的名称，并不能准确记录用户究竟访问了哪一个网页，因此意义不大。

[0008] (4)有的记录的是GET后面的字段，但是这容易导致多余记录的问题，例如访问一个www.sina.com.cn的时候，同时发出多个HTTP的GET请求，但我们只需要记录www.sina.com.cn这条原始网页访问记录，其他的并发的衍生出的众多的广告、推送、图片等链接信息不需要记录。

[0009] (5)仅仅根据发出去的http请求报文来记录上网信息，这样的话在网络不通或者网站根本无法访问的情况下仍然会留下上网记录，这是不合理的。

[0010] (6)所有的网站信息统一放在一个数据库中，没有根据使用频率进行分类，导致查询效率低下。

发明内容

[0011] 本发明针对现有网站访问行为管理和监控技术所存在的各项缺陷，而提供一种WEB分类控制和日志审计的方法。该方法基于模式匹配、DPI识别以及嵌入式数据库技术，有效解决现有技术所存在的问题。

[0012] 为了达到上述目的，本发明采用如下的技术方案：

[0013] 一种WEB分类控制和日志审计的方法，所述方法包括如下步骤：

[0014] (1)识别出网站访问报文，并将所要记录的信息插入队列中；

[0015] (2)通过定时器任务，将队列中的网站访问日志写入数据库中。

[0016] 在本发明的优选实例中，所述步骤(1)中采用DPI技术识别出真正的HTTP请求报文。

[0017] 进一步的，采用DPI技术分析报文中是否含有Accept-Encoding:字段，如果没有含

有该字段，则认为是伪装成浏览器报文。

[0018] 进一步的，所述步骤(1)在识别报文后，将识别出的HTTP请求报文与网址库中存储的网址信息进行类型匹配，判断HTTP请求报文对应的网站是否属于允许访问的网站类别范围。所述网址库实际包括存放完整网址库信息的第三级比较表、存放从第三级比较表中挑选出的热门网站的网址信息的第二级比较表、存放用户最近最经常访问网站的网址信息的第一级比较表。在进行类型匹配时，优先与第一级比较表中的信息进行匹配，若未找到则继续匹配第二级比较表，仍未找到则匹配第三级比较表，如果还是没有找到，则记录网址类型为未分类。未分类的暂定为一律允许。

[0019] 再进一步的，若HTTP请求报文对应的网站判断为属于允许访问的网站，进一步获取报文中Referer字段的信息，根据Referer字段采取不同的策略，如果该字段非空，就记录该字段内容，否则将Host字段与GET字段后面的内容拼接后记录到内存队列中。

[0020] 再进一步的，所述步骤(1)中的内存队列中的流节点是在收到一条http请求报文并在获取访问网站的网址信息及其控制类别的以后，根据报文的五元组(协议号、源IP地址、源端口、目的IP地址、目的端口)信息来建立并加入的，每一个新加入队列的流节点将自动被维护一个名为CHECKREQ的初始状态。在此状态基础上如果收到访问网站的回应报文且http状态值是200(成功)或者304(未修改)就将队列中相应的流节点的状态切换为CHECKREPLY。且在后续的定时器定时写数据库操作中，只有状态为CHECKREPLY的才会真正被移出队列并写入数据库中。

[0021] 再进一步的，在网站访问日志保存到数据库以后，所述第一级比较表中存储的信息将根据数据库中记录的网站访问日志信息不断的动态更新。

[0022] 进一步的，所述数据库包括两张数据表，第一张数据表存储最近一段时间内的网站访问日志信息，第二张数据表存储所有的网站访问日志信息，并且规定时间内将第一张数据表内数据移至第二张数据表中。

[0023] 本发明的方案在具体实施具有如下优点：

[0024] 1.避免将不是网页访问报文的伪浏览器报文误识别并记录下来。

[0025] 2.准确记录用户访问的确切页面，同时也避免将一些附带发出的http请求报文当作网页访问记录下来。

[0026] 3.能够有效过滤掉浏览器发出的不可达的网站的访问日志信息，减少无效访问日志的审计和处理，节约系统资源。

[0027] 4.网址类型匹配采用多级匹配的策略，提高了效率。

[0028] 5.可用于各企业、单位对员工的网站访问行为进行管理和监控。

附图说明

[0029] 以下结合附图和具体实施方式来进一步说明本发明。

[0030] 图1为本发明中报文处理的流程示意图；

[0031] 图2为本发明中网页访问日志信息写入数据库的流程示意图。

具体实施方式

[0032] 为了使本发明实现的技术手段、创作特征、达成目的与功效易于明白了解，下面结

合具体图示,进一步阐述本发明。

[0033] 本发明基于模式匹配、DPI识别以及嵌入式数据库技术实现WEB分类控制和日志审计,主要包括如下步骤:

[0034] 首先,识别出网站访问报文,并将所要记录的信息插入队列中;

[0035] 再者,通过定时器任务,将队列中的网站访问日志写入数据库中。

[0036] 基于上述原理方案,其具体的实施方案如下:

[0037] (1)在识别报文时,采用DPI技术排除掉常见的伪浏览器报文,从而识别出真正的HTTP请求报文。具体方法是通过DPI技术分析报文中是否含有Accept-Encoding:字段,如果没有含有该字段,则认为是伪装成浏览器报文。这样针对有些下载软件以及网络电视使用伪浏览器报文下载数据,只根据协议号和端口号无法确定该报文究竟是真正的HTTP请求报文还是伪浏览器报文,通过该方案能够有效的别出真正的HTTP请求报文。

[0038] (2)在识别报文后,将识别出的HTTP请求报文与网址库中存储的网址信息进行类型匹配,判断HTTP请求报文对应的网站是否属于允许访问的网站类别范围。

[0039] 其中,网址库用于存储网址类型匹配用的网址信息,如果只用一张大表的话就会导致查询效率低下,这里使用三张表来分级查询:表siteall中存放完整的网址库信息,数据量较大,只需存放在U盘中,无需加载到内存里;表sitecomn中存放的是从表siteall中挑选出来的热门网站的网址信息,数据量不是很大,因此可以加载到内存,作为第二级比较,存放在comn_addr中;表sitefreqt中存放的是用户最近几天最经常访问的网站的网址信息,也需要加载到内存中,作为第一级比较,存放着freq_addr中;在网站类型匹配时,则依次匹配freq_addr和comn_addr,如果查询不到,则该网站类型就属于未分类网站。

[0040] (3)对于HTTP请求报文对应的网站判断为属于允许访问的网站,在记录请求报文对应的网站访问日志的时候,进一步获取报文中Referer字段的信息,根据Referer字段采取不同的策略,如果该字段非空,就记录该字段内容,否则将Host字段与GET字段后面的内容拼接后记录下来。通过该方案能够避免记录大量无用信息。因为,当访问一个网页,如www.sina.com.cn,的时候,会接连发出多个HTTP页面请求,若仅仅依据GET字段或者HOST字段的内容就会导致记录大量无用或者重复的信息,由于这些附带的HTTP请求它们都有相同的Referer字段,如www.sina.com.cn,这样如果选择记录Referer字段并在数据库中进行相同替换就可以避免记录大量无用信息。

[0041] (4)在收到一条HTTP请求报文并在获取访问网站的网址信息及其控制类别的以后,根据报文的五元组(协议号、源IP地址、源端口、目的IP地址、目的端口)信息建立相应的流节点,并将节点加入到内存队列中,每一个新加入队列的流节点将自动被维护一个名为CHECKREQ的初始状态。在此状态基础上如果收到访问网站的回应报文且http状态值是200(成功)或者304(未修改)就将队列中相应的节点流的状态切换为CHECKREPLY。且在后续的定时器定时写数据库操作中,只有状态为CHECKREPLY的才会真正被移出队列并写入数据库中。这样的话在网站访问失败的情况下就不会留下网站访问的记录。就能够有效过滤掉浏览器发出的不可达的网站的访问日志信息,减少无效访问日志的审计和处理,节约系统资源。

[0042] (5)由于前面在报文检查的时候,只是将要写入数据库的信息插入到队列当中,对于真正写入数据库的操作,本发明通过一个单独的定时器任务完成的。

[0043] 由于往数据库中写数据的时候是进行替换操作，如果数据库太大必然造成耗时过大。为此，本发明使用两张表，表1-表surfingCurHour存放的是最近1小时内的网站访问日志信息；表2-表surfingDaily则是所有的网站访问日志信息；且每隔1小时就将表1中的数据移到表2中。这样在定时器超时往数据库中写数据时，只需操作较小的表1，从而节省了时间。

[0044] 本发明通过定时器任务每隔10s就往数据库中写入300条记录。另外这个定时任务还在比较空闲，例如晚上的时候，统计表surfingDaily中访问次数最高的前1000个网站作为表sitefreqt的内容。

[0045] 根据上述方案，以下通过一具体实例来对本发明中WEB分类控制和日志审计方案进一步说明。

[0046] 该实例中无需使用专门的设备，只要基于宽带路由器进行功能扩展，并将网址库，网址日志信息等存放着外接U盘中，这样方便用户备份数据。

[0047] 该实例实施时，分为两个部分：partA：识别出网站访问报文，并将所要记录的信息插入队列中；partB：定时器任务，将队列中的网站访问日志写入数据库中，定时调整用户访问网站的一级、二级和三级比较表中的网址信息，以及在用户设定的时间点自动将数据库中名为surfingCurHour的网站访问日志表中的访问日志信息归并到名为surfingDaily的网站访问日志表中。

[0048] 下面具体阐述：

[0049] partA：报文处理任务，该任务具体工作如下(参见图1)：

[0050] 1.当一个报文进入路由器时，通过DPI识别技术识别出该报文所属的类型，如果不是伪浏览器报文，并且是TCP报文，对端端口号80，此时认为是http报文，进入下一步处理。

[0051] 2.如果是内网主机发出去的http请求报文，则只处理以GET开头且不含有.gif、.jpg、.js等信息，还必须Host:字段非空的报文。然后获取GET字段后面跟的URI信息，以及Host字段的信息。根据Host字段的信息匹配freq_addr，如果未找到则继续匹配comn_addr，如果还是没有找到，则网址类型为0(未分类)。然后根据网址类别信息，判断该网站是否属于允许访问的网站类别范围(未分类的暂定为一律允许)，如果不属于就把该报文丢掉，从而禁止了该用户访问这个类别的网站。

[0052] 如果网站是允许访问的，则进一步获取报文中的Referer字段的信息。如果Referer信息是非空的，就将Referer信息作为URI信息记录下来，否则就将Host信息与GET后面的信息合并作为URI信息。得到URI信息后，将其与内网主机号等信息存放在一个动态分配的节点中，并将flag字段设置为CHECKREQ，然后插入队列中，待定时器超时的时候根据flag字段值决定是否写入数据库中。

[0053] 3.如果是服务器端发来的http回复报文，则检查报文中的HTTP返回值是否是200(成功)或者304(内容未修改)。如果是，并且flag字段值是CHECKREQ，就把flag字段改为CHECKREPLY，表示成功地访问了该网页。

[0054] partB：定时器任务，超时时间为10s，主要是完成将网页访问日志信息写入数据库以及其他的数据库处理操作。该任务具体工作如下(参见图2)：

[0055] 1.如果相关数据库表没有创建成功，就先创建。表surfingCurHour存放最近1小时的网站访问记录，表surfingDaily存放所有的网站访问记录。每隔1小时，就把

surfingCurHour中的记录移到surfingDaily中。

[0056] 表siteall，需要手动编辑然后放入U盘中，并非由路由器自动创建。这个表中存放的是最完整的网站类别信息，但是由于这个表太大，查找起来比较费时，因此只在晚上设备清闲的时候(如凌晨1点)进行一些查找工作。

[0057] 表sitecomn中存放的是从表siteall中筛选出来的常用网站，也是手动存放如U盘中。

[0058] 表sitefreqt中存放的是根据surfingDaily统计出的访问次数最多的前1000记录。如果不存在该表需要创建，一开始该表内容与表sitecomn相同。

[0059] 2.每次从队列中取下300条记录，如果flag字段为CHECKREPLY就将该URI以及主机、时间等信息写入数据库的表surfingCurHour中。写数据库时，根据当前时间、内网主机ip、URI信息生成一个historyid。其中时间部分需要处理，如果秒数小于9，则取9，如果小于19，则取19，以此类推。URI信息用于计算的部分不超过64字节，如果URI信息超过64字节，则取最后的64字节部分。由于将historyid定义为主键，因此在写数据库的时候，historyid相同的会被替换。这样由于时间部分的特殊处理，同样的记录在10秒内只会被记录一次。

[0060] 3.每隔1小时就把表surfingCurHour中的记录移到surfingDaily中。如果当前是凌晨1点，就统计表surfingDaily，根据主站Host字段统计出访问次数最多的1000个网站，如果这1000个高频度访问的网站中有网站的类型是‘未分类’，就到siteall这张表中查询，若查不到就把域名开头的部分改为www再到siteall表中查找，如果还是查不到就把该网站归为‘未分类’。然后再把这些1000条网站信息更新到表sitefreqt中，并更新到内存中的fqt_addr变量。

[0061] 以上显示和描述了本发明的基本原理、主要特征和本发明的优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明的原理，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化和改进都落入要求保护的本发明范围内。本发明要求保护范围由所附的权利要求书及其等效物界定。

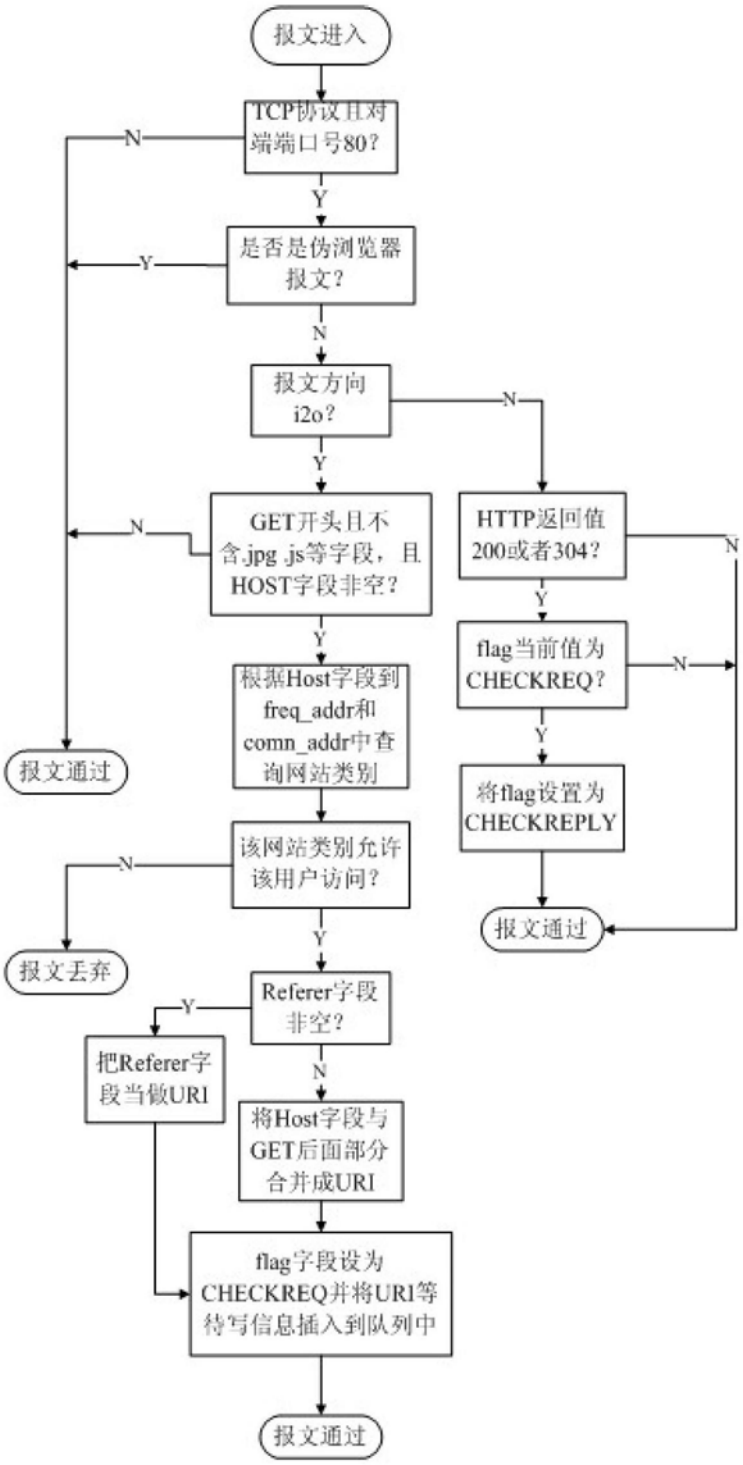

报文进入
TCP协议且对端端口号80?
N
Y
是否是伪浏览器报文?
Y
N
报文方向i2o?
N
Y
GET开头且不含.jpg .js等字段，且HOST字段非空?
N
Y
根据Host字段到freq_addr和comn_addr中查询网站类别
该网站类别允许该用户访问?
N
Y
Referer字段非空?
Y
N
把Referer字段当做URI
将Host字段与GET后面部分合并成URI
flag字段设为CHECKREQ并将URI等待写信息插入到队列中
报文通过
报文丢弃
报文通过
HTTP返回值200或者304?
N
Y
flag当前值为CHECKREQ?
N
Y
将flag设置为CHECKREPLY
报文通过

图1

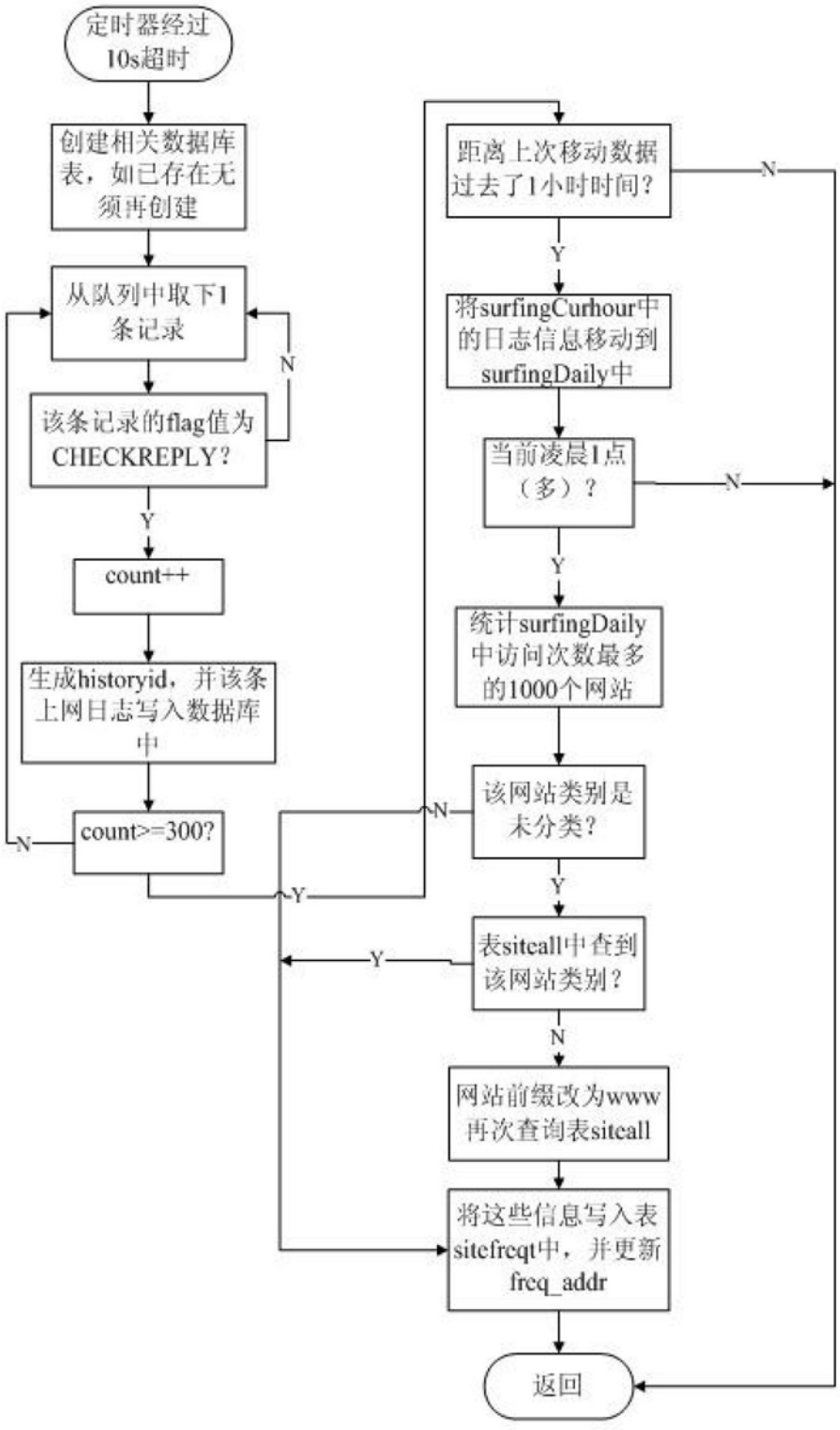

定时器经过10s超时
创建相关数据库表，如已存在无须再创建
从队列中取下1条记录
该条记录的flag值为CHECKREPLY?
N
Y
count++
生成historyid，并该条上网日志写入数据库中
count>=300?
N
Y
距离上次移动数据过去了1小时时间?
N
Y
将surfingCurhour中的日志信息移动到surfingDaily中
当前凌晨1点（多）?
N
Y
统计surfingDaily中访问次数最多的1000个网站
该网站类别是未分类?
N
Y
表siteall中查到该网站类别?
Y
N
网站前缀改为www再次查询表siteall
将这些信息写入表sitefreqt中，并更新freq_addr
返回

图2