(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
2018 年 2 月 15 日 (15.02.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/028430 A1

(51) 国际专利分类号:
H04L 29/06 (2006.01)

(21) 国际申请号: PCT/CN2017/094423

(22) 国际申请日: 2017 年 7 月 26 日 (26.07.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610645819.2 2016年8月8日 (08.08.2016) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 王鹏 (WANG, Peng); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 王中要 (WANG, Zhongyao); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 徐萧萧 (XU, Xiaoxiao); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (CN)。

(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,

(54) **Title:** METHOD, APPARATUS AND SYSTEM FOR IDENTIFICATION AND AUXILIARY IDENTIFICATION OF FAKE TRAFFIC

(54) **发明名称**: 识别及辅助识别虚假流量的方法、装置及系统

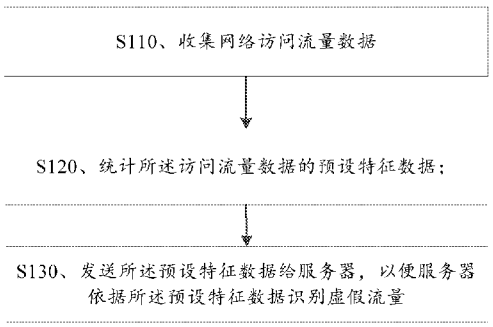


图 2

S110 COLLECTING NETWORK ACCESS TRAFFIC DATA
S120 COUNTING PRE-SET FEATURE DATA OF THE ACCESS TRAFFIC DATA
S130 SENDING THE PRE-SET FEATURE DATA TO A SERVER, SO THAT THE SERVER IDENTIFIES FAKE TRAFFIC ACCORDING TO THE PRE-SET FEATURE DATA

(57) Abstract: Provided are a method, an apparatus and a system for identification and auxiliary identification of fake traffic. The method for auxiliary identification of fake traffic comprises: collecting network access traffic data; counting pre-set feature data of the access traffic data; and sending the pre-set feature data to a server, so that the server identifies the fake traffic according to the pre-set feature data. The method for identification of fake traffic comprises: receiving the pre-set feature data of the access traffic data sent by a client; and inputting the pre-set feature data into a fake traffic identification model acquired by pre-training, so as to obtain an identification result regarding whether the access traffic data is fake traffic. The present application reduces the pressure on network transmission while accurately identifying fake traffic.

(57) 摘要: 本申请提供了一种识别及辅助识别虚假流量的方法、装置及系统，所述辅助识别虚假流量的方法包括：收集网络访问流量数据；统计所述访问流量数据的预设特征数据；发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。所述识别虚假流量的方法包括：接收客户端发送的访问流量数据的预设特征数据；将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。本申请实现了准确识别虚假流量的同时减轻对网络传输造成的压力。

WO 2018/028430 A1

[见续页]

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)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

识别及辅助识别虚假流量的方法、装置及系统

本申请要求2016年08月08日递交的申请号为201610645819.2、发明名称为“识别及辅助识别虚假流量的方法、装置及系统”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，尤其涉及一种识别及辅助识别虚假流量的方法、装置及系统。

背景技术

虚假流量即机器流量，是恶意访问者通过脚本程序自动频繁地向网络服务器发送请求。该虚假流量不仅不产生任何商业和社会价值，还给网络服务器带来可用性和稳定性风险，给正常用户的访问也带来了不便，同时也对网站的流量分析带来了较大的误差。因此，需要识别出虚假流量，以便对虚假流量进行控制，防止对网络产生的上述影响，保证正常用户的访问和更好的用户体验。

目前，识别虚假流量的方法为：客户端收集网络访问流量数据，将收集到的数据发送到服务器端，服务器根据这些访问流量数据来判断该访问流量是否为虚假流量。

其中，一种方案客户端提取所收集的网络访问流量数据的访问内容和手段相关方面的特征数据，包括：网址、访问时间、ip、登陆id等数据，将提取的特征数据发送给服务器端。该方案的缺点在于：由于随着刷流量技术和利益产业的发展，脚本程序在访问内容和手段上的表现已无法和正常流量区分，因此，服务器端基于访问内容和手段很难准确识别出虚假流量。

另一种方案客户端将网络访问流量明细数据发送给服务器端。该方案的缺点在于：由于发送给服务器端的访问流量明显数据存储量的增长导致数据传输的压力越来越大，对带宽也有了更高的要求，甚至成为了处理效率的瓶颈。

可见，已有的虚假流量识别方案中存在着数据量不足无法准确识别虚假流量的问题，或者数据量过大给网络传输造成压力的问题。因此有必要提供一种新的识别虚假流量的方案，以实现准确识别虚假流量的同时减轻对网络传输造成的压力。

发明内容

本申请解决的技术问题之一是提供一种识别及辅助识别虚假流量的方法、装置及系统，实现了准确识别虚假流量的同时减轻对网络传输造成的压力。

根据本申请一方面的一个实施例，提供了一种辅助识别虚假流量的方法，所述方法包括：

收集网络访问流量数据；

统计所述访问流量数据的预设特征数据；

发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。

根据本申请另一方面的一个实施例，提供了一种识别虚假流量的方法，所述方法包括：

接收客户端发送的访问流量数据的预设特征数据；

将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

根据本申请另一方面的一个实施例，提供了一种辅助识别虚假流量的装置，所述装置包括：

收集单元，用于收集网络访问流量数据；

统计单元，用于统计所述访问流量数据的预设特征数据；

发送单元，用于发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。

根据本申请另一方面的一个实施例，提供了一种识别虚假流量的装置，所述装置包括：

预设特征数据接收单元，用于接收客户端发送的访问流量数据的预设特征数据；

识别单元，用将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

根据本申请另一方面的一个实施例，提供了一种识别虚假流量的系统，所述系统包括：终端设备，用于收集网络访问流量数据，统计所述访问流量数据的预设特征数据，发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量；

服务器，用于接收客户端发送的访问流量数据的预设特征数据，将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

本申请实施例通过预先确定识别虚假流量所需的特征数据作为预设特征数据，由客

户端收集网络访问流量数据并统计网络访问流量的预设特征数据发送给服务器端，该预设特征数据需要客户端基于所收集的访问流量数据进行统计计算获得，而并非访问流量的明细数据，既能满足准确识别虚假流量，同时有效降低了客户端传输给服务器端的数据量，减轻了存储及传输的压力。

本领域普通技术人员将了解，虽然下面的详细说明将参考图示实施例、附图进行，但本申请并不仅限于这些实施例。而是，本申请的范围是广泛的，且意在仅通过后附的权利要求限定本申请的范围。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

图 1 是根据本申请一个实施例的识别虚假流量的实现框架图；

图 2 是根据本申请一个实施例的辅助识别虚假流量的方法的流程图；

图 3 是根据本申请一个实施例的识别虚假流量的方法的流程图；

图 4 是根据本申请一个实施例的辅助识别虚假流量的装置的结构示意图；

图 5 是根据本申请一个实施例的识别虚假流量的装置的结构示意图；

图 6 是根据本申请一个实施例的识别虚假流量的系统的架构示意图。

本领域普通技术人员将了解，虽然下面的详细说明将参考图示实施例、附图进行，但本申请并不仅限于这些实施例。而是，本申请的范围是广泛的，且意在仅通过后附的权利要求限定本申请的范围。

具体实施方式

在更加详细地讨论示例性实施例之前应当提到的是，一些示例性实施例被描述成作为流程图描绘的处理或方法。虽然流程图将各项操作描述成顺序的处理，但是其中的许多操作可以被并行地、并发地或者同时实施。此外，各项操作的顺序可以被重新安排。当其操作完成时所述处理可以被终止，但是还可以具有未包括在附图中的附加步骤。所述处理可以对应于方法、函数、规程、子例程、子程序等等。

所述计算机设备包括用户设备与网络设备。其中，所述用户设备包括但不限于电脑、智能手机、PDA 等；所述网络设备包括但不限于单个网络服务器、多个网络服务器组成的服务器组或基于云计算（Cloud Computing）的由大量计算机或网络服务器构成的云，

其中，云计算是分布式计算的一种，由一群松散耦合的计算机集组成的一个超级虚拟计算机。其中，所述计算机设备可单独运行来实现本申请，也可接入网络并通过与网络中的其他计算机设备的交互操作来实现本申请。其中，所述计算机设备所处的网络包括但不限于互联网、广域网、城域网、局域网、VPN 网络等。

需要说明的是，所述用户设备、网络设备和网络等仅为举例，其他现有的或今后可能出现的计算机设备或网络如可适用于本申请，也应包含在本申请保护范围以内，并以引用方式包含于此。

后面所讨论的方法（其中一些通过流程图示出）可以通过硬件、软件、固件、中间件、微代码、硬件描述语言或者其任意组合来实施。当用软件、固件、中间件或微代码来实施时，用以实施必要任务的程序代码或代码段可以被存储在机器或计算机可读介质（比如存储介质）中。（一个或多个）处理器可以实施必要的任务。

这里所公开的具体结构和功能细节仅仅是代表性的，并且是用于描述本申请的示例性实施例的目的。但是本申请可以通过许多替换形式来具体实现，并且不应当被解释成仅仅受限于这里所阐述的实施例。

应当理解的是，虽然在这里可能使用了术语“第一”、“第二”等等来描述各个单元，但是这些单元不应当受这些术语限制。使用这些术语仅仅是为了将一个单元与另一个单元进行区分。举例来说，在不背离示例性实施例的范围的情况下，第一单元可以被称为第二单元，并且类似地第二单元可以被称为第一单元。这里所使用的术语“和/或”包括其中一个或更多所列出的相关联项目的任意和所有组合。

应当理解的是，当一个单元被称为“连接”或“耦合”到另一单元时，其可以直接连接或耦合到所述另一单元，或者可以存在中间单元。与此相对，当一个单元被称为“直接连接”或“直接耦合”到另一单元时，则不存在中间单元。应当按照类似的方式来解释被用于描述单元之间的关系的其他词语（例如“处于...之间”相比于“直接处于...之间”，“与...邻近”相比于“与...直接邻近”等等）。

这里所使用的术语仅仅是为了描述具体实施例而不意图限制示例性实施例。除非上下文明确地另有所指，否则这里所使用的单数形式“一个”、“一项”还意图包括复数。还应当理解的是，这里所使用的术语“包括”和/或“包含”规定所陈述的特征、整数、步骤、操作、单元和/或组件的存在，而不排除存在或添加一个或更多其他特征、整数、步骤、操作、单元、组件和/或其组合。

还应当提到的是，在一些替换实现方式中，所提到的功能/动作可以按照不同于附图

中标示的顺序发生。举例来说，取决于所涉及的功能/动作，相继示出的两幅图实际上可以基本上同时执行或者有时可以按照相反的顺序来执行。

本申请发明人在对已有识别虚假流量的技术方案进行分析过程中发现，已有技术在识别虚假流量时，由于无法准确确定服务器端识别虚假流量所需的网络访问流量数据的特征信息，导致发送给服务器端的数据量不足或者数据量过大，从而造成服务器端无法准确识别虚假流量，或者对网络传输造成压力，造成处理效率的性能瓶颈。本申请实施例为克服该问题通过离线大数据的机器学习过程确定服务器端识别虚假流量所需的访问流量的特征数据，该机器学习过程所确定的特征数据既保证了识别虚假流量的准确性，又能有效减少客户端传送给服务器端的数据量，提升了数据传输以及虚假流量判断的效率。

下面结合附图对本申请的技术方案作进一步详细描述。

图 1 是本申请实施例识别虚假流量的实现框架图，由图 1 可以看出，客户端收集访问流量数据，所述客户端包括但不限于：PC、PAD、移动终端等。之后，针对收集的访问流量数据进行特征提取和分析，也就是统计获得所述访问流量数据的预设特征数据，再将统计获得的预设特征数据通过网络传输给服务器，以便服务器依据所述预设特征数据识别虚假流量。

图 2 是根据本申请一个实施例的辅助识别虚假流量的方法的流程图，该方法用于客户端收集服务器端识别虚假流量所需的特征数据发送给服务器端，从而辅助服务器端识别虚假流量，所述方法主要包括如下步骤：

S110、收集网络访问流量数据；

S120、统计所述访问流量数据的预设特征数据；

S130、发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。

下面对上述各步骤做进一步详细介绍。

步骤 S110 中所述的网络访问流量数据包括但不限于如下几类数据：

（1）所访问的网站内容，包括但不限于：所访问网站的各级域名，页面所包含的业务内容，页面的跳转上下游关系等数据。

（2）访问网站所使用的手段，包括但不限于：PC or 无线，ip 地址，mac 地址，wifi 名称，wifimac 地址，sessionid，登陆 id 等数据。

（3）用户访问习惯数据，包括但不限于：访问每个页面的具体时刻及停留时长，文

本框输入的字符内容、顺序、具体时刻及敲击按键时长，访问时鼠标划动轨迹坐标及具体时刻，访问时滚轮滚动具体时刻及幅度，用户使用触屏设备时手指滑动和拖动的幅度及具体时刻等数据。

上述用户访问习惯数据能够从用户的操作行为习惯上来识别虚假流量，更加符合实际业务场景的逻辑。

步骤 S120 所述预设特征数据为通过机器学习过程确定；具体确定方式为：服务器在离线时基于大数据量的机器学习过程所确定。由客户端完成该特征的提取。该预设特征数据需要客户端基于所收集的访问流量数据进行统计计算获得，而并非访问流量的明细数据。所需统计的预设特征数据包括但不限于如下至少一项：

a. 统计用户访问的页面数，页面跳转次数，用户通过客户端发起访问请求的网络 ip，mac，wifimac，sessionid，登陆 id；其中，所统计的用户访问的页面数及页面跳转次数可以为预设时长范围内，例如，1 秒、10 秒、20 秒或 1 分钟内用户访问的页面数及页面跳转次数。

b. 计算用户访问不同类型页面的时间段分布序列，页面访问时长平均值，页面访问时长方差。

c. 计算用户文本框输入字符串时连续字符的输入时间间隔均值和方差，输入不同字符的敲击按键的时长均值和方差，用户输入特定字符组合的按键顺序和时间间隔、键盘敲击的时长（敲击一个键的持续时长）。

d. 计算用户鼠标划动轨迹的 X 轴和 Y 轴两个方向的坐标最大和最小值，鼠标滑动的速度，计算用户最频繁出现的屏幕相对位置区域的频率。

e. 计算用户鼠标滚轮滚动时长的均值和方差，两次滚轮时间间隔的均值和方差，滚轮滚动幅度（屏数）的均值和方差。

f. 计算用户操作触屏设备时手指滑动和拖动的幅度（屏数）的均值和方差，页面跳转时用户手指滑动和拖动的时长均值和方差，页面跳转时间间隔均值和方差。

可见，本申请实施例将存储量较大的网络访问流量明细数据转换为存储量较少的访问流量特征数据，有效降低了存储及传输压力，且该预设特征数据为基于机器学习所确定，因此依据该数据足以准确识别该访问流量是否为虚假流量。

可以理解的是，上述预设特征数据仅为发明人所列举的几种实例，而本申请实施例并不局限于此，具体的特征数量更多，而且具体使用哪些是由机器学习算法来决定。

本申请实施例在机器学习确定预设特征数据时，可进一步确定上述预设特征数据中

的预判特征数据及预判特征数据对应的预判规则。所述预判特征数据为所述预设特征数据中权重较大的特征数据，也就是该预判特征数据对虚假流量识别结果影响较大。所述预判特征数据对应的预判规则可以为对应的阈值范围。例如，所述预判特征数据包括：用户访问的页面数、键盘敲击的时长及鼠标滑动的速度。用户访问的页面数对应的预判规则可以为用户单位时间内访问的页面数的阈值；键盘敲击的时长对应的预判规则可以为一键盘敲击时长阈值；鼠标滑动的速度对应的预判规则可以鼠标滑动速度阈值；为防止误判，会将各预判规则对应的阈值设置的较高，避免误杀正常用户的访问行为。从而根据该预判特征数据即可初步识别该流量是否为虚假流量。客户端可以保存所述预判特征数据及各预判特征数据对应的预判规则。从而在统计所述访问流量数据的预设特征数据后，可根据所述预设特征数据中的预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量的初步识别，得到初步识别结果。若所述初步识别结果为是虚假流量，则可对所述访问流量进行预控制。所述预控制包括但不限于：人机验证、限流控制等。例如，在初步识别结果为是虚假流量的情况下，将该识别结果发送给服务器，并接收服务器返回的用于进行人机验证的验证码以实现人机验证，或者客户端本地产生用于进行人机验证的验证码。

步骤 S130 为发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。其中，对于客户端利用预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量的初步识别，且识别结果为是虚假流量的情况，可以将该识别结果发送给服务器，同时发送所述预设特征数据给服务器，或者为减轻网络传输压力，可以仅发送该初步识别结果，而不发送该被初步识别为虚假流量的网络访问流量的预设特征数据。

本申请实施例在将预设特征数据发送给服务器后，服务器可依据该预设特征数据识别该网络访问流量是否为虚假流量，并将识别结果发送给客户端。客户端接收该服务器返回的所述访问流量是否为虚假流量的识别结果后，针对初步识别为虚假流量的网络访问流量，判断初步识别结果与服务器的识别结果是否一致；也就是判断初步识别为是虚假流量的访问流量所对应的所述服务器返回的识别结果是否为是虚假流量；若服务器的识别结果不是虚假流量，则解除对所述访问流量的预控制。也就是服务器与客户端的初步识别结果不一致，则解除对上述访问流量的预控制，若一致，则保持对该访问流量的预控制，并可进一步采取强制控制措施，例如，屏蔽该访问流量对应设备的后续访问流量。

另外，若所接收的服务器的识别结果与客户端的初步识别结果不一致，则说明所应用的预判特征数据对应的预判规则不够准确，在此情况下，可调整所述预判特征数据对应的预判规则。例如，假设客户端依据用户访问的页面数及对应的用户单位时间内访问的页面数的阈值对访问流量进行初步识别，初步识别结果为是虚假流量。而接收的服务器返回的识别结果为不是虚假流量，则说明该用户单位时间访问的页面数的阈值较宽松，正常用户的访问可以达到该阈值，需要提升该用户单位时间访问的页面数的阈值。针对每一预判特征数据的预判规则可预先设置一调整幅度，则可按照该预先设置的调整幅度调整所述预判规则，或者调整预判特征参数，也就是调整参与预判的预判特征数据，例如，根据一预判特征数据进行预判不够准确，不再利用该预判特征参数作为预判参数。

可以理解的是，针对初步识别结果为正常流量的访问流量，在接收到服务器返回的识别结果后，可根据该服务器的识别结果确定是否为虚假流量，进而确定是否对该访问流量进行控制。

图 3 是根据本申请一个实施例的识别虚假流量的方法的流程图，该方法用于服务器接收到客户端发送的预设特征数据后，依据该预设特征数据执行虚假流量的识别操作，所述方法主要包括如下步骤：

S210、接收客户端发送的访问流量数据的预设特征数据；

S220、将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

下面对上述各步骤做进一步详细介绍。

步骤 S210 所述预设特征数据为通过机器学习过程确定；所述的机器学习的过程可以在服务器离线执行。所述机器学习过程包括：

首先，收集样本数据；可将已有的虚假流量标记数据以及正常访问数据作为样本数据；可以理解的是，机器学习过程所收集的样本数据数量越大，则所训练的模型的准确性越高。本申请实施例可离线进行基于大量的样本数据进行训练，因此，可准确确定出识别虚假流量所需的特征数据，以及各特征数据的权重，并可进一步确定权重较大的特征数据为预判特征数据，例如，将权重超过预设权重阈值的特征数据确定为预判特征数据。同时可确定各预判特征数据对应的预判规则，所述预判规则如前面实施例中所述，此处不再赘述。所确定的预设特征数据、预判特征数据以及各预判特征数据对应的预判规则可发送给客户端，并保存在客户端。

之后，提取样本数据的预设特征数据输入到待训练的虚假流量识别模型中；

也就是，在训练虚假流量识别模型时，预先确定所需要的特征数据，从而通过训练过程确定各特征数据的权重大小。可以理解的是，可根据线上实际情况，不断更新所述预设特征数据，并基于更新的预设特征数据训练所述虚假流量识别模型。

最后，所述虚假流量识别模型输出虚假流量识别结果，同时确定了所述虚假流量识别模型所需要的预设特征数据以及各特征数据的权重。其中，该虚假流量识别结果可以为访问流量是虚假流量的概率。当该概率到达预设概率阈值时，可以认为该访问流量为虚假流量。也就是预置一个概率阈值，判断虚假流量识别模型输出的该访问流量为虚假流量的概率是否达到该概率阈值，若达到该概率阈值，则确定该访问流量为虚假流量，否则不是虚假流量。

本申请一种实施例离线通过机器学习过程所确定的预设特征数据包括但不限于：

用户所访问的网站内容，包括：用户访问的页面数，页面跳转次数；

用户访问网站所使用的手段，包括：用户通过客户端发起访问请求的网络 ip，mac，wifimac，sessionid，登陆 id；其中，所统计的用户访问的页面数及页面跳转次数可以为预设时长范围内，例如，1 秒、10 秒或 20 秒内用户访问的页面数及页面跳转次数。

用户访问习惯数据，包括：

用户访问不同类型页面的时间段分布序列，页面访问时长平均值，页面访问时长方差。

用户文本框输入字符串时连续字符的输入时间间隔均值和方差，输入不同字符的敲击按键的时长均值和方差，用户输入特定字符组合的按键顺序和时间间隔、键盘敲击的时长。

用户鼠标划动轨迹的 X 轴和 Y 轴两个方向的坐标最大和最小值，鼠标滑动的速度、计算用户最频繁出现的屏幕相对位置区域的概率。

用户鼠标滚轮滚动的时长均值和方差，滚轮时间间隔均值和方差，滚轮滚动幅度（屏数）的均值和方差。

用户手指滑动和拖动的幅度（屏数）的均值和方差，页面跳转时用户手指滑动和拖动的时长均值和方差，页面跳转时间间隔均值和方差。

上述基于机器学习所确定的特征数据即包含访问内容及访问手段，又包含用户的操作行为习惯，因此更加符合实际业务场景的逻辑，该数据量足以满足服务器端识别虚假流量所需，又有效减小了对网络传输造成压力。

步骤S220为线上基于从客户端接收的预设特征数据识别虚假流量的过程，将所述预设特征数据输入离线训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

服务器端本地可保存识别结果为虚假流量的预设特征数据，以便根据对虚假流量对应的设备的后续访问流量进行相应控制。所述控制包括但不限于：多重校验、限流控制甚至屏蔽控制等。或者服务器端将识别结果发送给客户端，以便客户端根据该识别结果对访问流量进行控制。

本申请一种实施例针对客户端将所有访问流量的预设特征数据均发送给服务器端的场景，也就是服务器端接收到所有访问流量的预设特征数据（包括初步识别结果为是虚假流量的访问流量的预设特征数据），可发送所述识别结果给所述客户端，以便所述客户端根据所述识别结果对所述访问流量进行控制或解除预控制。也就是将所有识别结果（包括识别结果为正常流量和虚假流量）均发送给客户端。

另一种实施例，针对客户端将所有访问流量的预设特征数据均发送给服务器端的场景，也就是服务器端接收到所有访问流量的预设特征数据（包括初步识别结果为是虚假流量的访问流量的预设特征数据），且同时接收到初步识别结果，则在得到识别结果后，可将该识别结果与从客户端接收的初步识别结果进行比对是否一致，若不一致，则发送该识别结果给客户端。由前面实施例的介绍可知，客户端仅将初步识别结果为是虚假流量的初步识别结果发送给服务器，则服务器端在得到该初步识别结果为是虚假流量的识别结果后，判断服务器的识别结果是否也是虚假流量，若是虚假流量，由于此时客户端已经对该虚假流量进行了预控制，则可以不发送给识别结果，若识别结果为不是虚假流量，说明客户端的初步识别出现误判，则将识别结果发送给客户端，以解除所述预控制。这样既实现了虚假流量的及时控制，又有效避免了误判现象。也就是本实施例中在接收客户端发送的根据所述预设特征数据中的预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量初步识别的初步识别结果后，比对所述虚假流量识别模型的识别结果与所述初步识别结果是否一致，若不一致，则发送识别结果给客户端。

可以理解的是，对于未接收到初步识别结果的访问流量，可以仅发送是虚假流量的识别结果，也可以所有识别结果给客户端。

另外，本申请实施例在虚假流量识别模型的识别结果与初步识别结果不一致情况下，可以调整预判特征数据对应的预判规则，并发送调整后的所述预判特征数据对应的预判

规则给所述客户端。可以预置各预判规则的调整幅度值及预判特征参数，根据该幅度值及预判特征参数调整预判规则。

又一种实施例，针对客户端仅将初步识别结果为正常流量对应的预设特征数据发送给服务器端的情况，在得到识别结果后，可以仅将虚假流量的识别结果发送给客户端。而对于接收的是虚假流量的初步识别结果，可以记录该识别结果，并对该虚假流量对应设备的后续流量进行控制。也就是针对此场景，直接接受客户端的初步识别结果。

本申请实施例还提供一种与上述辅助识别虚假流量的方法对应的辅助识别虚假流量的装置，所述装置结构示意图如图 4 所示，所述装置主要包括如下单元：

收集单元 310，用于收集网络访问流量数据；

统计单元 320，用于统计所述访问流量数据的预设特征数据；所述预设特征数据为通过机器学习过程确定。

发送单元 330，用于发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。

所述装置还包括：

初步识别单元 340，用于根据所述预设特征数据中的预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量的初步识别，得到初步识别结果；

预控制单元 350，用于在所述初步识别结果为是虚假流量情况下，对所述访问流量进行预控制。

所述装置还包括：

接收单元 360，用于接收所述服务器返回的所述访问流量是否为虚假流量的识别结果；

判断单元 370，用于判断初步识别结果为是虚假流量的访问流量所对应的所述服务器返回的识别结果是否为是虚假流量；

解除预控制单元 380，用于在所述判断单元判断所述识别结果不是虚假流量情况下，解除对所述访问流量的预控制。

所述装置还包括：

调整单元 390，用于在所述判断单元判断所述识别结果不是虚假流量情况下，调整所述预判特征数据对应的预判规则。

所述装置还包括：

初步识别结果发送单元3110，用于发送初步识别结果为是虚假流量的初步识别结果给所述服务器。

本申请实施例还提供一种与上述识别虚假流量的方法对应的识别虚假流量的装置，所述装置结构示意图如图5中所示，所述装置主要包括如下单元：

预设特征数据接收单元410，用于接收客户端发送的访问流量数据的预设特征数据；所述预设特征数据为通过机器学习过程确定。

识别单元 420，用将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

所述识别结果包括：

识别所述访问流量数据为虚假流量的概率。

所述装置还包括：

发送单元430，用于发送所述识别结果给所述客户端。

所述装置还包括：

初步识别结果接收单元 440，用于接收客户端发送的根据所述预设特征数据中的预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量初步识别的初步识别结果；

比对单元 450，用于比对所述虚假流量识别模型的识别结果与所述初步识别结果是否一致；

调整单元 460，用于在所述虚假流量识别模型的识别结果与所述初步识别结果不一致情况下，调整所述预判特征数据对应的预判规则；

预判规则发送单元 470，用于发送调整后的所述预判特征数据对应的预判规则给所述客户端。

如图6所示为本申请实施例的识别虚假流量的系统的架构示意图，该系统包括：终端设备 610，用于收集网络访问流量数据，统计所述访问流量数据的预设特征数据，发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量；

服务器 620，用于接收客户端发送的访问流量数据的预设特征数据，将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

综上所述，本申请实施例通过预先确定识别虚假流量所需的特征数据作为预设特征

数据，由客户端收集网络访问流量数据并统计网络访问流量的预设特征数据发送给服务器端，该预设特征数据需要客户端基于所收集的访问流量数据进行统计计算获得，而并非访问流量的明细数据，既能满足准确识别虚假流量，同时有效降低了客户端传输给服务器端的数据量，减轻了存储及传输的压力。

需要注意的是，本申请可在软件和/或软件与硬件的组合体中被实施，例如，可采用专用集成电路（ASIC）、通用目的计算机或任何其他类似硬件设备来实现。在一个实施例中，本申请的软件程序可以通过处理器执行以实现上文所述步骤或功能。同样地，本申请的软件程序（包括相关的数据结构）可以被存储到计算机可读记录介质中，例如，RAM 存储器，磁或光驱动器或软磁盘及类似设备。另外，本申请的一些步骤或功能可采用硬件来实现，例如，作为与处理器配合从而执行各个步骤或功能的电路。

另外，本申请的一部分可被应用为计算机程序产品，例如计算机程序指令，当其被计算机执行时，通过该计算机的操作，可以调用或提供根据本申请的方法和/或技术方案。而调用本申请的方法的程序指令，可能被存储在固定的或可移动的记录介质中，和/或通过广播或其他信号承载媒体中的数据流而被传输，和/或被存储在根据所述程序指令运行的计算机设备的工作存储器中。在此，根据本申请的一个实施例包括一个装置，该装置包括用于存储计算机程序指令的存储器和用于执行程序指令的处理器，其中，当该计算机程序指令被该处理器执行时，触发该装置运行基于前述根据本申请的多个实施例的方法和/或技术方案。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请的精神或基本特征的情况下，能够以其他的具体形式实现本申请。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

权 利 要 求 书

1. 一种识别虚假流量的系统，其特征在于，包括：

终端设备，用于收集网络访问流量数据，统计所述访问流量数据的预设特征数据，发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量；

服务器，用于接收客户端发送的访问流量数据的预设特征数据，将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

2. 一种辅助识别虚假流量的方法，其特征在于，所述方法包括：

收集网络访问流量数据；

统计所述访问流量数据的预设特征数据；

发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。

3. 如权利要求 2 所述的方法，其特征在于，统计所述访问流量数据的预设特征数据后，所述方法还包括：

根据所述预设特征数据中的预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量的初步识别，得到初步识别结果；

若所述初步识别结果为是虚假流量，则对所述访问流量进行预控制。

4. 如权利要求 3 所述的方法，其特征在于，所述方法还包括：

接收所述服务器返回的所述访问流量是否为虚假流量的识别结果；

判断初步识别结果为是虚假流量的访问流量所对应的所述服务器返回的识别结果是否为是虚假流量；

若不是虚假流量，则解除对所述访问流量的预控制。

5. 如权利要求 4 所述的方法，其特征在于，若不是虚假流量，所述方法还包括：

调整所述预判特征数据对应的预判规则。

6. 如权利要求 3 所述的方法，其特征在于，所述预判规则包括：预判特征数据对应的阈值范围及预判特征参数。

7. 如权利要求 3 所述的方法，其特征在于，若所述初步识别结果为是虚假流量，所述方法还包括：

发送所述初步识别结果给所述服务器。

8. 如权利要求 2 所述的方法，其特征在于，收集网络访问流量数据前，所述方法还

包括：

在离线计算框架下，利用已有的虚假流量标记数据，使用预先收集到的流量的预设特征数据，训练机器学习模型，得到模型的各项参数。

9.如权利要求 2 所述的方法，其特征在于，收集的所述网络访问流量数据包括但不限于如下至少一类数据：

所访问的网站内容、访问网站所使用的手段、用户访问习惯数据。

10.如权利要求 2 所述的方法，其特征在于，所述预设特征数据包括如下至少一项：

用户所访问的网站内容，包括：用户访问的页面数，页面跳转次数；

用户访问网站所使用的手段，包括：用户通过客户端发起访问请求的网络 ip，mac，wifimac，sessionid，登陆 id；

用户访问习惯数据，包括：用户访问不同类型页面的时间段分布序列，页面访问时长平均值，页面访问时长方差；

用户文本框输入字符串时连续字符的输入时间间隔均值和方差，输入不同字符的敲击按键的时长均值和方差，用户输入特定字符组合的按键顺序和时间间隔、键盘敲击的时长；

用户鼠标划动轨迹的 X 轴和 Y 轴两个方向的坐标最大和最小值，鼠标滑动的速度，用户最频繁出现的屏幕相对位置区域的频率；

用户鼠标滚轮滚动时长的均值和方差，两次滚轮时间间隔的均值和方差，滚轮滚动幅度的均值和方差；

用户操作触屏设备时手指滑动和拖动的幅度的均值和方差，页面跳转时用户手指滑动和拖动的时长均值和方差，页面跳转时间间隔均值和方差。

11.一种识别虚假流量的方法，其特征在于，所述方法包括：

接收客户端发送的访问流量数据的预设特征数据；

将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

12.如权利要求 11 所述的方法，其特征在于，所述识别结果包括：

识别所述访问流量数据为虚假流量的概率。

13.如权利要求 11 所述的方法，其特征在于，所述方法还包括：

发送所述识别结果给所述客户端。

14.如权利要求 11 所述的方法，其特征在于，所述方法还包括：

接收客户端发送的根据所述预设特征数据中的预判特征数据及所述预判特征数据对应的预判规则对所述访问流量数据进行虚假流量初步识别的初步识别结果；

比对所述虚假流量识别模型的识别结果与所述初步识别结果是否一致；

若不一致，则调整所述预判特征数据对应的预判规则；

发送调整后的所述预判特征数据对应的预判规则给所述客户端。

15. 如权利要求 11 所述的方法，其特征在于，所述方法还包括：

在离线计算框架下，利用已有的虚假流量标记数据，使用预先收集到的流量的预设特征数据，训练所述虚假流量识别模型，得到模型的各项参数。

16. 如权利要求 15 所述的方法，其特征在于，所述方法还包括：

将所述识别结果作为离线计算框架下训练所述虚假流量识别模型的迭代输入，以实现对所述虚假流量识别模型进行迭代训练。

17. 一种辅助识别虚假流量的装置，其特征在于，所述装置包括：

收集单元，用于收集网络访问流量数据；

统计单元，用于统计所述访问流量数据的预设特征数据；

发送单元，用于发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量。

18. 一种识别虚假流量的装置，其特征在于，所述装置包括：

预设特征数据接收单元，用于接收客户端发送的访问流量数据的预设特征数据；

识别单元，用将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果。

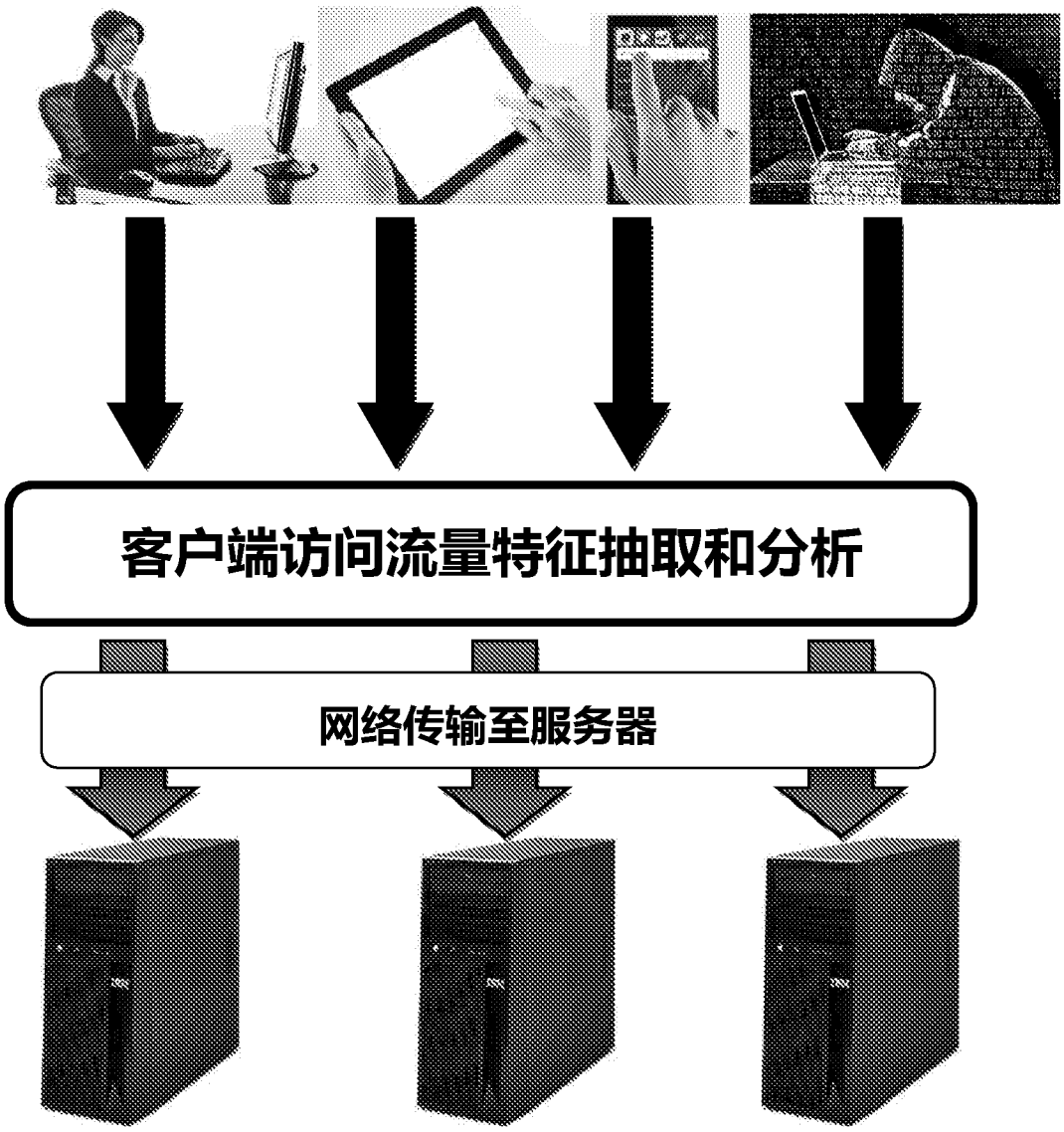

客户端访问流量特征抽取和分析
网络传输至服务器

图 1

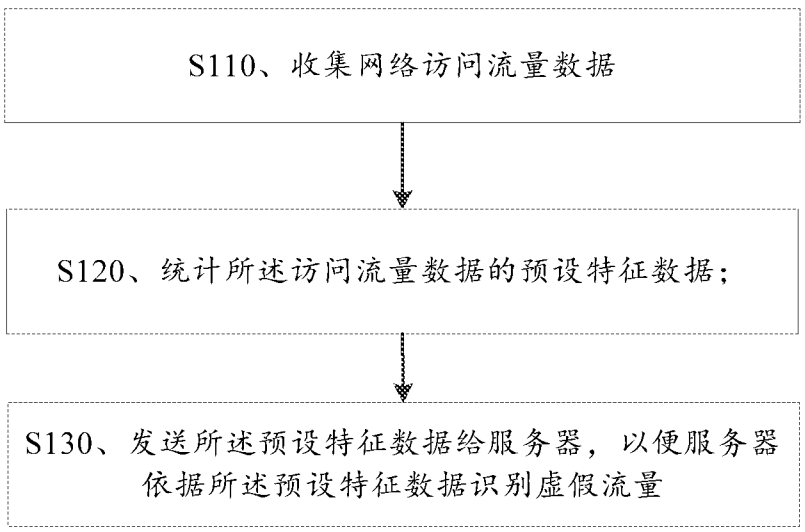

S110、收集网络访问流量数据
S120、统计所述访问流量数据的预设特征数据；
S130、发送所述预设特征数据给服务器，以便服务器依据所述预设特征数据识别虚假流量

图 2

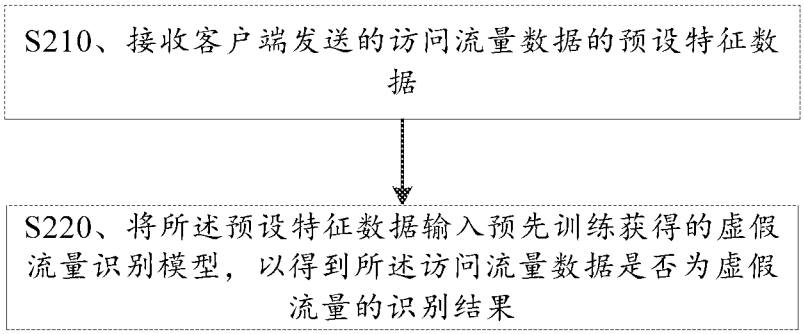

S210、接收客户端发送的访问流量数据的预设特征数据
S220、将所述预设特征数据输入预先训练获得的虚假流量识别模型，以得到所述访问流量数据是否为虚假流量的识别结果

图 3


辅助识别虚假流量的装置
收集单元310
统计单元320
发送单元330

图 4

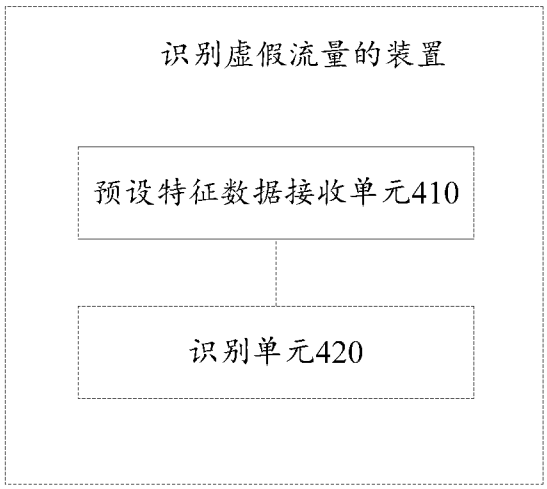

识别虚假流量的装置
预设特征数据接收单元410
识别单元420

图 5

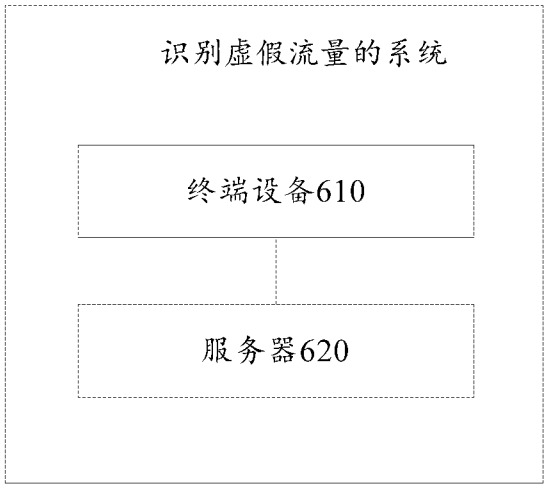


图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/094423

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/06 (2006.01) 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)

H04L; H04W; G06F; H04Q

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: 虚假, 机器, 脚本, 程序, 爬虫, 广告, 恶意, 流量, 数据量, 访问量, 点击量, 识别, 检测, 判断, 分类, 统计, 整理, 初步, 预先, 预判, 首次, 一次, 减少, 降低, 减轻, 缓解, 传输, 带宽, 压力, 需求, 服务器, 客户端, 终端, 发送, 接收, machine, fake, false, robot, script, malicious, crawler, spider, advertis+, flow, traffic, detect, distinguish, identify, recognize, character+, data, parameter, server, client, terminal, first, initial, eliminate, decrease, bandwidth, press+, burden, load, require+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105808639 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.), 27 July 2016 (27.07.2016), description, paragraphs [0021]-[0062] and [0078]-[0090]	1-18
A	CN 105187396 A (XIAOMI TECHNOLOGY CO., LTD.), 23 December 2015 (23.12.2015), entire document	1-18
A	CN 105119735 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 02 December 2015 (02.12.2015), entire document	1-18
A	US 2003227917 A1 (NETRAKE CORPORATION), 11 December 2003 (11.12.2003), entire document	1-18

☐ 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	Date of mailing of the international search report
13 October 2017	26 October 2017
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer LIU, Cheng Telephone No.: (86-10) 010-61648536

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/094423

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105808639 A	27 July 2016	None	
CN 105187396 A	23 December 2015	None	
CN 105119735 A	02 December 2015	None	
US 2003227917 A1	11 December 2003	EP 1512302 A1	09 March 2005
		WO 03105506 A1	18 December 2003
		AU 2003248677 A1	22 December 2003
		US 2004215770 A1	28 October 2004
		JP 2006513590 A	20 April 2006

国际检索报告

国际申请号
PCT/CN2017/094423

A. 主题的分类

H04L 29/06(2006.01)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)

H04L; H04W; G06F; H04Q

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,EPODOC: 虚假，机器，脚本，程序，爬虫，广告，恶意，流量，数据量，访问量，点击量，识别，检测，判断，分类，统计，整理，初步，预先，预判，首次，一次，减少，降低，减轻，缓解，传输，带宽，压力，需求，服务器，客户端，终端，发送，接收，machine, fake, false, robot, script, malicious, crawler, spider, advertis+, flow, traffic, detect, distinguish, identify, recognize, character+, data, parameter, server, client, terminal,first, initial, eliminate, decrease, bandwidth, press+, burden, load, require+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 105808639 A (平安科技深圳有限公司) 2016年 7月 27日 (2016-07-27) 说明书第 [0021]-[0062]、[0078]-[0090]段	1-18
A	CN 105187396 A (小米科技有限责任公司) 2015年 12月 23日 (2015-12-23) 全文	1-18
A	CN 105119735 A (百度在线网络技术北京有限公司) 2015年 12月 2日 (2015-12-02) 全文	1-18
A	US 2003227917 A1 (NETRAKE CORPORATION) 2003年 12月 11日 (2003-12-11) 全文	1-18

☐ 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	国际检索报告邮寄日期
2017年 10月 13日	2017年 10月 26日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 刘成 电话号码 (86-10)010-61648536

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/094423

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	105808639	A	2016年 7月 27日	无			
CN	105187396	A	2015年 12月 23日	无			
CN	105119735	A	2015年 12月 2日	无			
US	2003227917	A1	2003年 12月 11日	EP	1512302	A1	2005年 3月 9日
				WO	03105506	A1	2003年 12月 18日
				AU	2003248677	A1	2003年 12月 22日
				US	2004215770	A1	2004年 10月 28日
				JP	2006513590	A	2006年 4月 20日