

(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
2016 年 8 月 18 日 (18.08.2016)



(10) 国际公布号
WO 2016/127858 A1

- (51) 国际专利分类号:
G06F 21/56 (2013.01) *H04L 29/06* (2006.01)
- (21) 国际申请号: PCT/CN2016/073148
- (22) 国际申请日: 2016 年 2 月 2 日 (02.02.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510069899.7 2015 年 2 月 10 日 (10.02.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英国开曼群岛大开曼乔治城资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): 周燃 (ZHOU, Ran) [CN/CN]; 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: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.);

中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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, 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)。

[见续页]

(54) Title: METHOD AND DEVICE FOR IDENTIFYING WEBPAGE INTRUSION SCRIPT FEATURES

(54) 发明名称: 网页入侵脚本特征的识别方法及设备

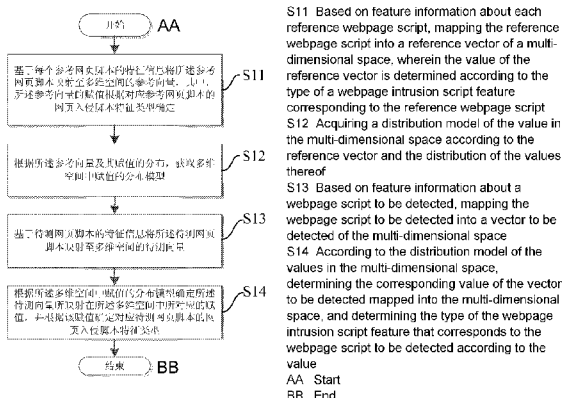


图 5

(57) Abstract: A method and device for identifying webpage intrusion script features. By means of mapping a certain number of acquired reference webpage scripts and feature information thereof into reference vectors of a multi-dimensional space and values thereof, acquiring distribution models of the values in the multi-dimensional space according to the reference vectors and the distribution of the values thereof, and when it is necessary to identify webpage intrusion features and script features, acquiring a webpage script to be detected and using feature information about the webpage script to be detected to map the webpage script to be detected into a vector to be detected of the multi-dimensional space, a webpage intrusion script feature of the webpage script to be detected can simply be obtained only needing to input feature information about the vector to be detected into the distribution models of the values in the multi-dimensional space, thereby reducing the requirement on computer hardware, and greatly shortening the identification time while ensuring the identification accuracy of the webpage intrusion script features.

(57) 摘要:

[见续页]



WO 2016/127858 A1

**本国际公布:**

— 包括国际检索报告(条约第 21 条(3))。

一种网页入侵脚本特征的识别方法及设备，通过将获取的一定数量的参考网页脚本及其特征信息映射为多维空间的参考向量及其赋值，根据所述参考向量及其赋值的分布获取多维空间中赋值的分布模型，当需要进行网页入侵特征脚本特征识别时，获取待测网页脚本并将待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量，只需将待测向量的特征信息输入至所述多维空间中赋值的分布模型即可获得待测网页脚本的网页入侵脚本特征，因此降低了对计算硬件的要求，并在保证对网页入侵脚本特征的识别准确性的同时大大缩小了识别时间。

网页入侵脚本特征的识别方法及设备

技术领域

本申请涉及通信及计算机领域，尤其涉及一种网页入侵脚本特征的识别方法及设备。

5

背景技术

网页入侵脚本（Webshell）是指网页脚本里面具有一定危害性的特定字符串内容的组合，会导致网站被攻击、数据泄露等严重后果。简单的说来，网页入侵脚本就是一个 ASP（动态服务器页面，Active Server Page）或 PHP（超文本预处理器，Hypertext Preprocessor）的木马后门，黑客在入侵了一个网站后，常常在将这些 ASP 或 PHP 木马后门文件放置在网站服务器的 web 目录中，与正常的网页文件混在一起，然后，黑客就可以用 web（网页）的方式，通过 ASP 或 PHP 木马后门控制网站服务器，包括上传下载文件、查看数据库、执行任意程序命令等，再通过 dos 命令或者植入后门木马通过服务器漏洞等达到提权的目的。

15

对于网页入侵脚本来说，网页入侵脚本特征就是那些具有危害性的代码字符串。根据网页入侵脚本特征为分类依据，可以将网页脚本划分为多个特征类型。

目前对网页入侵脚本特征的处理，普遍采取大量采集样本数据放入数据库，然后以正则表达式方式进行特征字符串的查找识别。然而，随着参考网页脚本数量不断增加，参考网页脚本占用磁盘和内存空间高，可以从几兆字节到几百兆字节，因此存在库体积的瓶颈；并且，当体积到一定程度后，整个库在网页入侵脚本特征识别过程中，CPU（内存）占用率越来越高，甚至可达 100% 占用率，无法满足生产场景中对 CPU 使用率的限制要求；此外，识别速度慢，平均单个网页脚本文件识别时间大概在 500 毫秒至 1 秒，进而失去使用价值。

25

发明内容

本申请的目的是提供一种内存占用率低、识别速度快的网页入侵脚本特

征的识别方法。

有鉴于此，本申请提供一种网页入侵脚本特征的识别方法，其中，所述识别方法包括：

5 基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量，其中，所述参考向量的赋值根据对应参考网页脚本的网页入侵脚本特征类型确定；

根据所述参考向量及其赋值的分布，获取多维空间中赋值的分布模型；

基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量；以及

10 根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值，并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

优选的，将所述参考网页脚本映射至多维空间的参考向量以及将所述待测网页脚本映射至多维空间的待测向量基于相同的映射规则。

15 进一步的，基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量包括：

获取若干所述参考网页脚本；

对所述参考网页脚本进行采样和量化，以获取所述参考网页脚本的特征值集合；以及

20 将每一所述参考网页脚本的特征值集合映射至所述多维空间的参考向量。

优选的，对所述参考网页脚本的特征进行采样和量化包括：

从每个所述参考网页脚本中截取相同文本容量的参考网页脚本；以及

25 对所述截取的参考网页脚本的字符进行采样和量化，以获取所述参考网页脚本的特征值集合。

优选的，对从所述参考网页脚本截取的字符依次赋予特征值包括：

根据 ASCII 码规则将所采样的所述参考网页脚本的字符量化成相应特征值。

优选的，所述识别方法还包括：

利用从每个所述参考网页脚本中截取之后剩余的参考网页脚本确定所述待测网页脚本的网页入侵脚本特征类型。

优选的，所述识别方法还包括：

5 采用正则表达式对从每个所述参考网页脚本中截取之后剩余的参考网页脚本进行处理和存储。

进一步的，基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量包括：

获取所述待测网页脚本；

10 对所述待测网页脚本进行采样和量化，以获取所述待测网页脚本的特征值集合；以及

将所述待测网页脚本的特征值集合映射至所述多维空间的待测向量。

优选的，对所述待测网页脚本的特征进行采样和量化包括：

从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本；以及

15 对所述截取的待测网页脚本的字符进行采样和量化，以获取所述待测网页脚本的特征值集合。

优选的，对所述截取的待测网页脚本的字符进行采样和量化包括：

根据 ASCII 码规则将所采样的所述待测网页脚本的字符量化成相应特征值。

20 优选的，将所述多维空间进行划分包括：

利用分类算法获取所述多维空间中赋值的分布模型。

优选的，所述分类算法包括支持向量机算法。

优选的，所述多维空间中赋值的分布模型包括若干多维函数关系组。

25 优选的，根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值包括：

将所述待测向量的赋值代入所述若干多维函数关系组，利用解析几何原理获取所述待测向量所映射在所述多维空间中所对应的赋值。

进一步的，所述网页入侵脚本特征类型包括正常类型、异常类型及待定类型中的一种或任意组合。

本申请还提供一种网页入侵脚本特征的识别设备，其中，所述识别设备包括：

第一装置，用于基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量，其中，所述参考向量的赋值根据对应参考网页脚本的网页入侵脚本特征类型确定；

第二装置，用于根据所述参考向量及其赋值的分布，获取多维空间中赋值的分布模型；

第三装置，用于基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量；以及

第四装置，用于根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值，并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

优选的，所述第一装置将所述参考网页脚本映射至多维空间的参考向量以及第三装置将所述待测网页脚本映射至多维空间的待测向量基于相同的映射规则。

进一步的，所述第一装置包括：

第一单元，用于获取若干所述参考网页脚本；

第二单元，用于对所述参考网页脚本进行采样和量化，以获取所述参考网页脚本的特征值集合；以及

第三单元，用于将每一所述参考网页脚本的特征值集合映射至所述多维空间的参考向量。

优选的，所述第二单元包括：

第一子单元，用于从每个所述参考网页脚本中截取相同文本容量的参考网页脚本；以及

第二子单元，用于对所述截取的参考网页脚本的字符进行采样和量化，以获取所述参考网页脚本的特征值集合。

优选的，所述第二子单元包括：

第二一子单元，用于根据 ASCII 码规则将所采样的所述参考网页脚本的字符量化成相应特征值。

优选的，所述识别设备还包括：

第五装置，用于利用从每个所述参考网页脚本中截取之后剩余的参考网页脚本确定所述待测网页脚本的网页入侵脚本特征类型。

优选的，所述识别设备还包括：

5 第六装置，用于采用正则表达式对从每个所述参考网页脚本中截取之后剩余的参考网页脚本进行处理和存储。

优选的，所述第三装置包括：

第四单元，用于获取所述待测网页脚本；

10 第五单元，用于对所述待测网页脚本进行采样和量化，以获取所述待测网页脚本的特征值集合；以及

第六单元，用于将所述待测网页脚本的特征值集合映射至所述多维空间的待测向量。

优选的，所述第五单元包括：

15 第三子单元，用于从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本；以及

第四子单元，用于对所述截取的待测网页脚本的字符进行采样和量化，以获取所述待测网页脚本的特征值集合。

优选的，所述第四子单元包括：

20 第四一子单元，用于根据 ASCII 码规则将所采样的所述待测网页脚本的字符量化成相应特征值。

优选的，所述第二装置包括：

第七单元，用于利用分类算法获取所述多维空间中赋值的分布模型。

优选的，所述分类算法包括支持向量机算法。

优选的，所述多维空间中赋值的分布模型包括若干多维函数关系组。

25 优选的，所述第四装置包括：

第八单元，用于将所述待测向量的赋值代入所述若干多维函数关系组，利用解析几何原理获取所述待测向量所映射在所述多维空间中所对应的赋值。

进一步的，所述网页入侵脚本特征类型包括正常类型、异常类型及待定

类型中的一种或任意组合。

与现有技术相比，本申请所述网页入侵脚本特征的识别方法及设备通过
将获取的一定数量的参考网页脚本及其特征信息映射为多维空间的参考向量
及其赋值，根据所述参考向量及其赋值的分布获取多维空间中赋值的分布模
5 型，当需要进行网页入侵特征脚本特征识别时，获取待测网页脚本并将待测
网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量，只需
将待测向量的特征信息输入至所述多维空间中赋值的分布模型即可获得待测
网页脚本的网页入侵脚本特征，因此降低了对计算硬件的要求，并在保证对
网页入侵脚本特征的识别准确性的同时大大缩小了识别时间。

10 进一步的，本申请所述网页入侵脚本特征的识别方法及设备还用从每个
所述参考网页脚本的网页脚本中截取之后剩余的网页脚本进一步确定所述待
测网页脚本的网页入侵脚本特征类型，进一步提高识别的准确性。

附图说明

15 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 示出根据本申请一方面示出的网页入侵脚本特征的识别设备示意
图；

图 2 示出根据本申请一优选的实施例示出的第一装置示意图；

20 图 3 示出根据本申请一优选的实施例示出的第三装置示意图；

图 4 示出根据本申请另一方面示出的网页入侵脚本特征的识别设备示意
图；

图 5 示出根据本申请一方面示出的网页入侵脚本特征的识别方法的流程
图；

25 图 6 示出根据本申请一优选的实施例示出的基于每个参考网页脚本的特
征信息将所述参考网页脚本映射至多维空间的参考向量的流程图；

图 7 示出根据本申请一优选的实施例示出的基于待测网页脚本的特征信
息将所述待测网页脚本映射至多维空间的待测向量；

图 8 示出根据本申请另一方面示出的网页入侵脚本特征的识别方法的流

程示意图。

附图中相同或相似的附图标记代表相同或相似的部件。

具体实施方式

5 在本申请一个典型的配置中，终端、服务网络的设备和可信方均包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和/或非易失性内存等形式，如只读存储器 (ROM) 或闪存 (flash RAM)。内存是计算机可读介质的示例。

10 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只
15 读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带，磁带磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计
算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体 (transitory media)，如调制的数据信号和载波。

20 图 1 示出根据本申请一方面示出的网页入侵脚本特征的识别设备示意图，其中，所述识别设备 1 包括：第一装置 11、第二装置 12、第三装置 13 及第四装置 14。

在较佳的实施例中，所述网页入侵脚本特征类型包括正常类型、异常类型及待定类型中的一种或任意组合。进一步的，正常类型也可包括若干不同的
25 的正常类型，异常类型也可包括若干不同的异常类型，待定类型也可包括若干不同的待定类型，在此，对网页入侵脚本特征类型的进一步划分可根据获得的参考入侵脚本特征确定，不再赘述。

在此，所述识别设备 1 包括但不限于用户设备、网络设备、或用户设备与网络设备通过网络相集成所构成的设备。所述识别设备 1 其包括但不限于

任何一种可与用户通过触摸板进行人机交互的移动电子产品,例如智能手机、PDA等,所述移动电子产品可以采用任意操作系统,如Android操作系统、IOS操作系统等。其中,所述识别设备1包括一种能够按照事先设定或存储的指令,自动进行数值计算和信息处理的电子设备,其硬件包括但不限于微处理器、专用集成电路(ASIC)、可编程门阵列(FPGA)、数字处理器(DSP)、嵌入式设备等。所述识别设备1其包括但不限于计算机、网络主机、单个网络服务器、多个网络服务器集或多个服务器构成的云;在此,云由基于云计算(Cloud Computing)的大量计算机或网络服务器构成,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虚拟超级计算机。

10 所述网络包括但不限于互联网、广域网、城域网、局域网、VPN网络、无线自组织网络(Ad Hoc网络)等。本领域技术人员应能理解,其他的触摸控制设备同样适用于本发明,也应包含在本发明保护范围以内,并在此以引用方式包含于此。

其中,第一装置11用于基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量,其中,所述参考向量的赋值根据对应参考网页脚本的网页入侵脚本特征类型确定;第二装置12用于根据所述参考向量及其赋值的分布,获取多维空间中赋值的分布模型;第三装置13用于基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量;以及第四装置14用于根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值,并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

15

20

在此,上述各装置之间是持续不断工作的,在此,本领域技术人员应理解“持续”是指上述各装置分别实时地或者按照设定的或实时调整的工作模式要求,进行获取所划分区域的分布模型等,直至确定对应待测网页脚本的网页入侵脚本特征类型。

25

图2示出根据本申请一优选的实施例示出的第一装置示意图,其中,第一装置11包括第一单元110、第二单元120以及第三单元130。

其中,第一单元110用于获取若干所述参考网页脚本;第二单元120用于对所述参考网页脚本进行采样和量化,以获取所述参考网页脚本的特征值

集合；以及第三单元 130 用于将每一所述参考网页脚本的特征值集合映射至所述多维空间的参考向量。

在优选的实施例中，所述第二单元 120 包括第一子单元（未标示）和第二子单元（未标示），其中，第一子单元用于从每个所述参考网页脚本中截取
5 相同文本容量的参考网页脚本，第二子单元用于对所述截取的参考网页脚本的字符进行采样和量化，以获取所述参考网页脚本的特征值集合。

在优选的实施例中，第二子单元（未示出）包括第二子单元（未示出），所述第二子单元用于根据 ASCII 码规则将所采样的所述参考网页脚本的字符量化成相应特征值。

10 在此，ASCII 是基于拉丁字母的一套电脑编码系统，在计算机中将所有的数据在存储和运算时都使用二进制数表示。它是现今最通用的单字节编码系统，是一种在计算机中最为通过用的编码系统。因此，通过 ASCII 码规则依次对所述参考网页脚本的网页脚本的字符赋予特征值能够通过计算机直接识别和存储，无需额外增加计算机计算负担，进而提高识别效率。

15 在实际应用场景中，网页入侵脚本特征通常会出现现在在网页脚本文件的 256KB 字节之前（其中，256KB 字节仅为举例并不被限制）。

基于上述，在本实施例中，所述网页入侵脚本特征识别方法基于建立对网页脚本中前 256KB 字节字符的特征识别模型，即首先在一个以 256×1024 维的多维度空间进行建模，将每一参考网页脚本的特征信息都会映射为以
20 256×1024 维的多维度空间中的一个参考向量，每一向量具有 256×1024 维度的向量坐标，且每个参考向量具有一个根据对应参考网页脚本的网页入侵脚本特征类型确定的赋值。

具体地，第一装置 11 的第一单元 110 获取若干所述参考网页脚本之后，第二单元 120 对所述参考网页脚本进行采样，其中对于小于 256KB（字节）
25 的参考网页脚本则获取该参考网页脚本的全部网页脚本，对于大于 256KB 的参考网页脚本即截取该参考网页脚本的前 256KB 字节的部分参考网页脚本。

然后，第二单元 120 再对采样获得参考网页脚本中的字符进行量化，以使各字符形成特征值。经第二单元 120 量化后的每一参考网页脚本可获得具

有 256*1024 数量的特征值组成的特征值集合。然后，由第三单元 130 将每一待测网页脚本的特征值集合中的每一特征值依次作为向量的向量坐标，从而将每一待测网页脚本映射为一个具有 256*1024 维度的向量。

在具体的应用场景中，目前网页入侵脚本特征通常会出现现在在网页脚本

5 文件的 256KB 字节之前，故以此为例：

更具体地，例如某参考网页脚本内容如下：

?PHP namespace

...

Xxbbxxbb

10

其中，参考网页脚本中“Xxbbxxbb”代表一个 Webshell 特征。

将上述参考网页脚本的每个字符依据计算机 ASCII 码转换为 16 进制数值，则量化后形成的向量如下：

(0x50,0x68,0x70,0x6e,0x61,0x6d,0x65,0x73,0x70,0x61,0x63,0x65,.....,0x58,
15 0x78,0x62,0x62,0x78,0xx78,0x62,0x62.....)。

则上述参考网页脚本就形成了一个多维空间的向量。

接着，多维空间的向量的赋值可以采用但不限于十进制数值表示。

例如一参考网页脚本特征类型为异常类型，则根据异常类型标记类型值，
例如用十进制的两位数标示：10，表示异常及异常的类型，将该类型值赋予
20 其参考网页脚本的特征所映射的赋值。

从而，将参考网页脚本的特征形成为如下内容：

10: (0x50,0x68,0x70,0x6e,0x61,0x6d,0x65,0x73,0x70,0x61,0x63,0x65,
.....,0x58,0x78,0x62,0x62,0x78,0xx78,0x62,0x62.....)

上述采样和量化的方法大大简化的参考网页脚本的存储数量，且便于后
25 续进行特征提取和计算。

当然，本领域技术人员应当能够理解，所述的对多维空间向量的赋值方式以及对所述参考网页脚本的网页脚本的字符赋予特征值仅为举例，其他获得的赋值和特征值的方式，如可适用于本申请，也应包含在本申请保护范围以内，并在此以引用方式包含于此。

从而,通过第一装置 11 持续获得参考网页脚本,并基于获取的参考网页脚本的特征信息将参考网页脚本映射至多维空间的参考向量,从而形成若干参考向量及其对应的参考向量的赋值。

接着,第二装置 12 根据所述赋值分布,将所述多维空间进行划分,并获取所划分区域的分布模型根据所述参考向量及其赋值的分布,获取多维空间中赋值的分布模型。在优选的实施例中,所划分的区域的分布模型包括若干所述区域所分布的所述多维空间中赋值的分布模型包括若干多维函数关系组。

进一步的,所述第二装置 12 包括第七单元(未标示),所述第七单元用于利用分类算法获取所述多维空间中赋值的分布模型。较佳的,所述分类算法包括采用支持向量机(SVM, Support Vector Machine)算法。支持向量机算法能够将巨量的样本数据压缩成一个多元方程组。

在具体的实施例中,每个参考网页脚本经过处理后得到的是一个特征值集合组成的特征向量,每个特征向量可以等价为一个多维函数关系式(方程式)的系数。多个参考向量经过处理后就可以得到若干多维函数关系组,若干多维函数关系组即可描述所述多维空间。使用支持向量机算法能够根据若干多维函数关系组最终计算得出这个多维空间的向量表达法。具体地,例如在一参考网页脚本的采样到三个特征字符(1:66, 2:97, 3:99),参考网页脚本的类型值为 10,则可等价为多维函数关系式: $10=66x+97y+99z$,例如在另一参考网页脚本采样到三个特征字符(1:66, 2:97, 3:99),且该参考网页脚本的类型值为 20, $20=74x+101y+104z+80q$ 。

在此,多维关系式中“+”里仅仅是举例,并不限于此,其他运算方式,例如加减乘除等运算符号适用于本申请的,亦可以引用的方式包含在本申请之中。

在此,支持向量机(Support Vector Machine, SVM)是一种基于统计学习理论的模式识别方法,主要应用于模式识别领域。支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中,以建立识别模型。

具体地，第二装置 12 将第一装置 11 创建好的参考向量及其对应赋值使用支持向量机的多分类算法予以运算，根据支持向量机的特性，耗时可能从几分钟到一个月不等，且因为支持向量机的特性原因，中途不可中断。根据目前软硬件情况以及支持向量机本身的特性，运行时间在几百的数量级，该
5 这个计算所花时间成本是可以接受的。

在一具体实施例中，最终支持向量机会算出一个文本方式的展现结果，类似如下：

0 # kernel type

3 # kernel parameter -d

10 1 # kernel parameter -g

1 # kernel parameter -s

1 # kernel parameter -r

empty# kernel parameter -u

50 # highest feature index

15 500 # number of training documents

101 # number of support vectors plus 1

9.9308382 # threshold b, each following line is a SV (starting with alpha*y)

-3.2864346412986876e-007 1:0 2:247 3:211 4:204 5:202 6:223 7:206 8:241

9:208 10:222

20 该文本方式的展现结果实质是关于若干所述多维函数关系组的相关方程系数，所述第二装置 12 将若干甚至是巨量的参考网页脚本文件压缩为一个多维函数关系组，该函数关系组仅包括方程、方程系数等函数关系的相关数据，从而实现了巨量的数据压缩和保存。

25 因此，将巨量样本数据根据一定的映射算法，映射到一个多维空间里，并且利用支持向量机的多分类算法，找出多个类之间的分割面，最后持久化保存这些分割面的方程组表现，达到将巨量样本数据压缩成一个多元函数关系的结果。

图 3 示出根据本申请一优选的实施例示出的第三装置示意图，其中，第三装置 13 包括：第四单元 310、第五单元 320 以及第六单元 330。

其中，第四单元 310 用于获取所述待测网页脚本；第五单元 320 用于对所述待测网页脚本进行采样和量化，以获取所述待测网页脚本的特征值集合；第六单元 330 用于将所述待测网页脚本的特征值集合映射至所述多维空间的待测向量。

5 在优选的实施例中，所述第一装置 11 将所述参考网页脚本映射至多维空间的参考向量以及第三装置 13 将所述待测网页脚本映射至多维空间的待测向量基于相同的映射规则。其中，相同的映射规则包括基于相同的规则对参考网页脚本和待测网页脚本进行采样和量化，以使参考网页脚本和待测网页脚本能够映射在同一多维空间。

10 在优选的实施例中，所述第五单元 320 包括第三子单元（未标示）和第四子单元（未标示）。其中，第三子单元用于从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本；第四子单元用于对所述截取的待测网页脚本的字符进行采样和量化，以获取所述待测网页脚本的特征值集合。

15 在优选的实施例中，所述第四子单元包括第四一子单元（未标示），所述第四一子单元用于根据 ASCII 码规则将所采样的所述待测网页脚本的字符量化成相应特征值。

具体地，继续以所述网页入侵脚本特征识别方法基于建立对页脚本中前 256KB 字节字符的特征识别模型，即在一个以 256*1024 维的多维度空间进行
20 建模为例，将待测网页脚本的特征信息都映射为以 256*1024 维的多维度空间中的待测向量，待测向量具有 256*1024 维度的向量坐标。

具体地，第三装置 13 的第四单元 310 获取若干所述待测网页脚本之后，第五单元 320 对所述待测网页脚本进行采样，其中对于小于 256KB（字节）的待测网页脚本即获取该待测网页脚本的全部网页脚本，对于大于 256KB 的
25 的待测网页脚本即截取该待测网页脚本的前 256KB 字节的部分网页脚本。然后，第五单元 320 再对采样获得网页脚本中的各字符进行量化，以形成特征值，对于小于 256KB 的待测网页脚本，其形成的特征值的数量小于 256*1024 个，则不足数量的特征值可以用特殊的特征值标示，例如特征值为“0”标示。则经第五单元 320 量化后的每一待测网页脚本可获得具有 256*1024 数量的特

征值组成的特征值集合，由第三单元 130 将每一所述待测网页脚本的特征值集合中的每一特征值依次作为向量的向量坐标，从而将每一待测网页脚本映射为一个具有 256×1024 维度的向量。因而在后续可以根据参考向量的赋值的分布规律获取待测向量的赋值信息。

- 5 进一步的，第四装置 14 根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值，并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

在优选的实施例中，所述第四装置 14 进一步包括第八单元（未示出），所述第八单元用于将所述待测向量的赋值代入所述若干多维函数关系组，利用解析几何原理获取所述待测向量所映射在所述多维空间中所对应的赋值。

继续以所述网页入侵脚本特征识别方法基于建立对页脚本中前 256KB 字节字符的特征识别模型，即在一个以 256×1024 维的多维度空间进行建模为例，将待测向量的 256×1024 维度的向量坐标代入前述的所述第二装置 12 获得的多维函数关系组中，则能够获得待测脚本所映射的待测向量所在区域，进而通过该区域的赋值，确定所述待测脚本的类型值。

因此，采用支持向量机获得关于多维空间中赋值的多维函数关系组，将待测网页脚本的特征通过支持向量机的多分类识别接口，即可得出分类值，从而获得待测网页脚本的网页入侵脚本特征类型。通过使用分类算法对参考网页脚本进行特征提取和空间映射，将参考网页脚本的特征及其网页入侵脚本特征类型映射为多维空间的向量及其赋值，并根据赋值对多维空间进行区域分布，则只需获得待测网页脚本所映射的区域的赋值即可获得待测网页脚本的网页入侵脚本特征类型。

图 4 示出根据本申请另一方面示出的网页入侵脚本特征的识别设备示意图，在另一优选的实施例中，本申请所述网页入侵脚本特征的识别设备 1 包括第一装置 11'、第二装置 12'、第三装置 13'、第四装置 14' 以及第五装置 15'。

在此，识别设备 1 的第一装置 11'、第二装置 12'、第三装置 13' 以及第四装置 14' 与图 1 中第一装置 11、第二装置 12、第三装置 13 以及第四装置 14 的内容对应相同或基本相同，为简明起见，不再赘述，仅以引用的方式包含在此。

其中，第五装置 15' 利用从每个所述参考网页脚本中截取之后剩余的参考网页脚本确定所述待测网页脚本的网页入侵脚本特征类型。所述第五装置 15' 获得的剩余的网页脚本可以用于进一步确定所述待测网页脚本的网页入侵脚本特征类型，例如，当第四装置 14' 获得的网页入侵脚本特征类型为待确定类型时，则可以利用所述第五装置 15' 将待测网页脚本与从参考网页脚本中获得的剩余的网页脚本进行进一步比对，进一步检测其网页入侵脚本特征类型。

继续以所述网页入侵脚本特征识别方法基于建立对页脚本中前 256KB 字节字符的特征识别模型，即在一个以 256*1024 维的多维度空间进行建模为例，第五装置 15' 对于大于 256KB 的参考网页脚本的文件，将网页脚本中第 256KB 以后的字符按照正则表达式进行存储，以便于后续利用数据库中的正则表达式存储的参考网页脚本的内容对待测网页脚本进行进一步匹配，对分类算法获得的网页入侵脚本特征类型进行进一步匹配，以提高识别的准确性。

图 5 示出根据本申请一方面示出的网页入侵脚本特征的识别方法的流程图，所述识别设备方法 1 包括：第一装置步骤 S11、第二装置步骤 S12、第三装置步骤 S13 及第四装置步骤 S14。

在较佳的实施例中，所述网页入侵脚本特征类型包括正常类型、异常类型及待定类型中的一种或任意组合。进一步的，正常类型也可包括若干不同的正常类型，异常类型也可包括若干不同的异常类型，待定类型也可包括若干不同的待定类型，在此，对特征类型的进一步划分可根据获得的参考入侵脚本特征确定，不再赘述。

其中，在步骤 S11 中，基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量，其中，所述参考向量的赋值根据对应参考网页脚本的网页入侵脚本特征类型确定；在步骤 S12 中，根据所述参考向量及其赋值的分布，获取多维空间中赋值的分布模型；在步骤 S13 中，基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量；以及在步骤 S14 中，根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值，并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

图 6 示出根据本申请的一个优选的实施例示出的基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量的流程图, 其中, 步骤 S11 包括步骤 S110、步骤 S120 以及步骤 S130。

其中, 在步骤 S110 用于中, 获取若干所述参考网页脚本; 在步骤 S120 5 中, 用于对所述参考网页脚本进行采样和量化, 以获取所述参考网页脚本的特征值集合; 以及在步骤 S130 中, 用于将每一所述参考网页脚本的特征值集合映射至所述多维空间的参考向量。

在优选的实施例中, 所述步骤 S120 进一步包括以下步骤:

从每个所述参考网页脚本中截取相同文本容量的参考网页脚本; 以及对 10 所述截取的参考网页脚本的字符进行采样和量化, 以获取所述参考网页脚本的特征值集合。

在优选的实施例中, 从每个所述参考网页脚本中截取相同文本容量的参考网页脚本进一步包括:

根据 ASCII 码规则将所采样的所述参考网页脚本的字符量化成相应特征 15 值。

在此, ASCII 是基于拉丁字母的一套电脑编码系统, 在计算机中将所有的数据在存储和运算时都使用二进制数表示。它是现今最通用的单字节编码系统, 是一种在计算机中最为通过用的编码系统。因此, 通过 ASCII 码规则依次对所述参考网页脚本的网页脚本的字符赋予特征值能够通过计算机直接 20 识别和存储, 无需额外增加计算机计算负担。

在实际应用场景中, 网页入侵脚本特征通常会出现现在在网页脚本文件的 256KB 字节之前 (其中, 256KB 字节仅为举例并不被限制)。

基于上述, 在本实施例中, 所述网页入侵脚本特征识别方法基于建立对 页脚本中前 256KB 字节字符的特征识别模型, 即首先在一个以 256×1024 维 25 的多维度空间进行建模, 将每一参考网页脚本的特征信息都会映射为以 256×1024 维的多维度空间中的一个参考向量, 每一向量具有 256×1024 维度的向量坐标, 且每个参考向量具有一个根据对应参考网页脚本的网页入侵脚本特征类型确定的赋值。

具体地, 步骤 S110 获取若干所述参考网页脚本之后, 步骤 S120 对所述

参考网页脚本进行采样，其中对于小于 256KB（字节）的参考网页脚本则获取该参考网页脚本的全部网页脚本，对于大于 256KB 的参考网页脚本即截取该参考网页脚本的前 256KB 字节的部分网页脚本。

5 然后，步骤 S120 再对采样获得参考网页脚本中的各字符进行量化，以使各字符形成特征值。经第二单元 120 量化后的每一参考网页脚本可获得具有 256*1024 数量的特征值组成的特征值集合。

然后，在步骤 S130 中将每一所述待测网页脚本的特征值集合中的每一特征值依次作为向量的向量坐标，从而将每一待测网页脚本映射为一个具有 256*1024 维度的向量。

10 在具体的应用场景中，目前 Webshell 特征通常会出现现在在网页脚本文件的 256KB 字节之前，故以此为例：

更具体地，例如某参考网页脚本内容如下：

?PHP namespace

...

15 Xxbbxxbb

....

其中，参考网页脚本中“Xxbbxxbb”代表一 Webshell 特征。

将上述参考网页脚本的每个字符依据计算机 ASCII 码转换为 16 进制数值，则量化后形成的向量如下：

20 (0x50,0x68,0x70,0x6e,0x61,0x6d,0x65,0x73,0x70,0x61,0x63,0x65,.....,0x58,0x78,0x62,0x62,0x78,0xx78,0x62,0x62.....)。

则上述参考网页脚本就形成了一个多维空间的向量。

接着，多维空间的向量的赋值可以采用但不限于十进制数值表示。

25 例如一参考网页脚本特征类型为异常类型，则根据异常类型标记类型值，例如用十进制的两位数标示：10，表示异常及异常的类型，将该类型值赋予其参考网页脚本的特征所映射的赋值。

从而，将参考网页脚本的特征形成为如下内容：

10: (0x50,0x68,0x70,0x6e,0x61,0x6d,0x65,0x73,0x70,0x61,0x63,0x65,.....,0x58,0x78,0x62,0x62,0x78,0xx78,0x62,0x62.....)

上述采样和量化的方法大大简化的参考网页脚本的存储数量，且便于后续进行特征提取和计算。

当然，本领域技术人员应当能够理解，所述的对多维空间向量的赋值方式以及对所述参考网页脚本的网页脚本的字符赋予特征值仅为举例，其他获得的赋值和特征值的方式，如可适用于本申请，也应包含在本申请保护范围以内，并在此以引用方式包含于此。

从而，通过步骤 S11 持续获得参考网页脚本，并基于获取的参考网页脚本的特征信息将参考网页脚本映射至多维空间的参考向量，从而形成若干参考向量及其对应的参考向量的赋值。

接着，步骤 S12 根据所述参考向量及其赋值的分布，获取多维空间中赋值的分布模型。在优选的实施例中，所述多维空间中赋值的分布模型包括若干多维函数关系组。

进一步的，所述步骤 S12 包括：利用分类算法获取所述多维空间中赋值的分布模型。

较佳的，所述分类算法采用支持向量机算法。支持向量机算法能够将巨量的样本数据压缩成一个多元方程组。

在具体的实施例中，每个参考网页脚本经过处理后得到的是一个特征值集合组成的特征向量，每个特征向量可以等价为一个多维函数关系式（方程式）的系数。多个参考向量经过处理后就可以得到若干多维函数关系组，若干多维函数关系组即可描述所述多维空间。使用支持向量机算法能够根据若干多维函数关系组最终计算得出这个多维空间的向量表达法。具体地，例如在一参考网页脚本的采样到三个特征字符（1:66, 2:97, 3:99），参考网页脚本的类型值为 10，则可等价为多维函数关系式： $10=66x+97y+99z$ ，例如在另一参考网页脚本采样到三个特征字符（1:66, 2:97, 3:99），且该参考网页脚本的类型值为 20， $20=74x+101y+104z+80q$ 。

在此，多维关系式中“+”里仅仅是举例，并不限于此，其他运算方式，例如加减乘除等运算符号适用于本申请的，亦可以引用的方式包含在本申请之中。

在此，支持向量机(Support Vector Machine, SVM)是一种基于统计学习理

论的模式识别方法，主要应用于模式识别领域。支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以建立识别模型。

- 5 具体地，步骤 S12 将步骤 S11 创建好的参考向量及其对应赋值使用支持向量机的多分类算法予以运算，根据支持向量机的特性，耗时时间可能从几分钟到一个月不等。根据目前软硬件情况以及支持向量机本身的特性，运行时间在几百的数量级，为可承受的计算时间成本。

在一具体实施例中，最终支持向量机会算出一个文本方式的展现结果，
10 类似如下：

```
0 # kernel type
3 # kernel parameter -d
1 # kernel parameter -g
1 # kernel parameter -s
15 1 # kernel parameter -r
empty# kernel parameter -u
50 # highest feature index
500 # number of training documents
101 # number of support vectors plus 1
20 9.9308382 # threshold b, each following line is a SV (starting with alpha*y)
-3.2864346412986876e-007 1:0 2:247 3:211 4:204 5:202 6:223 7:206 8:241
9:208 10:222
```

该文本方式的展现结果实质是关于若干所述多维函数关系组的相关方程系数，所述步骤 S12 将若干甚至是巨量的参考网页脚本文件压缩为一个多维
25 函数关系组，该函数关系组仅包括方程、方程系数等函数关系的相关数据，从而实现了巨量的数据压缩和保存。

因此，将巨量样本数据根据一定的映射算法，映射到一个多维空间里，并且利用支持向量机的多分类算法，找出多个类之间的分割面，最后持久化保存这些分割面的方程组表现，达到将巨量样本数据压缩成一个多元函数关

系的结果。

图 7 示出根据本申请一优选的实施例示出的基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量，其中，所述步骤 S13 包括步骤 S310、步骤 S 320 以及步骤 S330。

- 5 其中，在步骤 S310 中，获取所述待测网页脚本；在步骤 S320 中，对所述待测网页脚本进行采样和量化，以获取所述待测网页脚本的特征值集合；在步骤 S330 中，将所述待测网页脚本的特征值集合映射至所述多维空间的待测向量。

- 10 在优选的实施例中，所述步骤 S11 将所述参考网页脚本映射至多维空间的参考向量以及步骤 S13 将所述待测网页脚本映射至多维空间的待测向量基于相同的映射规则。其中，相同的映射规则包括基于相同的规则对参考网页脚本和待测网页脚本进行采样和量化，以使参考网页脚本和待测网页脚本能够映射在同一多维空间。

在优选的实施例中，所述步骤 S320 包括：

- 15 从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本；以及

对所述截取的待测网页脚本的字符进行采样和量化，以获取所述待测网页脚本的特征值集合。

- 20 在优选的实施例中，从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本包括：根据 ASCII 码规则将所采样的所述待测网页脚本的字符量化成相应特征值。

- 具体地，继续以所述网页入侵脚本特征识别方法基于建立对页脚本中前 256KB 字节字符的特征识别模型，即在一个以 256*1024 维的多维度空间进行建模为例，将待测网页脚本的特征信息都映射为以 256*1024 维的多维度空间
25 中的待测向量，待测向量具有 256*1024 维度的向量坐标。

具体地，步骤 S310 获取若干所述待测网页脚本之后，步骤 S320 对所述待测网页脚本进行采样，其中对于小于 256KB（字节）的待测网页脚本即获取该待测网页脚本的全部网页脚本，对于大于 256KB 的的待测网页脚本即截取该待测网页脚本的前 256KB 字节的部分网页脚本。然后，步骤 S320 再对

采样获得网页脚本中的各字符进行量化,以形成特征值,对于小于 256KB 的待测网页脚本,其形成的特征值的数量小于 256×1024 个,则不足数量的特征值可以用特殊的特征值标示,例如特征值为“0”标示。则经步骤 S320 量化后的每一待测网页脚本可获得具有 256×1024 数量的特征值组成的特征值集合,由步骤 S130 将每一所述待测网页脚本的特征值集合中的每一特征值依次作为向量的向量坐标,从而将每一待测网页脚本映射为一个具有 256×1024 维度的向量。因而在后续可以根据参考向量的赋值的分布规律获取待测向量的赋值信息。

进一步的,在步骤 S14 中,根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值,并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

在优选的实施例中,步骤 S14 进一步包括:将所述待测向量的赋值代入所述若干多维函数关系组,利用解析几何原理获取所述待测向量所映射在所述多维空间中所对应的赋值。

继续以所述网页入侵脚本特征识别方法基于建立对页脚本中前 256KB 字节字符的特征识别模型,即在一个以 256×1024 维的多维度空间进行建模为例,将待测向量的 256×1024 维度的向量坐标代入前述的所述第二装置 12 获得的多维函数关系组中,则能够获得待测脚本所映射的待测向量所在区域,进而通过该区域的赋值,确定所述待测脚本的类型值。

因此,采用支持向量机获得关于多维空间中赋值的多维函数关系组,将待测网页脚本的特征通过支持向量机的多分类识别接口,即可得出分类值,从而获得待测网页脚本的网页入侵脚本特征类型。通过使用分类算法对参考网页脚本进行特征提取和空间映射,将参考网页脚本的特征及其网页入侵脚本特征类型映射为多维空间的向量及其赋值,并根据赋值对多维空间进行区域分布,则只需获得待测网页脚本所映射的区域的赋值即可获得待测网页脚本的网页入侵脚本特征类型。

图 8 示出根据本申请另一方面示出的网页入侵脚本特征的识别方法的流程示意图,在另一优选的实施例中,本申请所述网页入侵脚本特征的识别方法包括步骤 S11'、步骤 S12'、步骤 S13'、步骤 S14'以及步骤 S15'。

其中，识别方法中步骤 S11'、步骤 S12'、步骤 S13'以及步骤 S14'与图 1 中步骤 S11、步骤 S12、步骤 S13 以及步骤 S14 的内容对应相同或基本相同，为简明起见，不再赘述，仅以引用的方式包含在此。

其中，步骤 S15' 利用从每个所述参考网页脚本中截取之后剩余的参考网页脚本确定所述待测网页脚本的网页入侵脚本特征类型。所述第五装置 15' 获得的剩余的网页脚本可以用于进一步确定所述待测网页脚本的网页入侵脚本特征类型，例如，当步骤 S14'获得的网页入侵脚本特征类型为待确定类型时，则可以利用所述步骤 S15'将待测网页脚本与从参考网页脚本中获得的剩余的网页脚本进行进一步比对，进一步检测其网页入侵脚本特征类型。

继续以所述网页入侵脚本特征识别方法基于建立对页脚本中前 256KB 字节字符的特征识别模型，即在一个以 256*1024 维的多维度空间进行建模为例，步骤 S15'对于大于 256KB 的参考网页脚本的文件，将网页脚本中第 256KB 以后的字符按照正则表达式进行存储，以便于后续利用数据库中的正则表达式存储的参考网页脚本的内容对待测网页脚本进行进一步匹配，对分类算法获得的网页入侵脚本特征类型进行进一步匹配，以提高识别的准确性。

与现有技术相比，本申请所述网页入侵脚本特征的识别方法及设备通过将获取的一定数量的参考网页脚本及其特征信息映射为多维空间的参考向量及其赋值，根据所述参考向量及其赋值的分布获取多维空间中赋值的分布模型，当需要进行网页入侵特征脚本特征识别时，获取待测网页脚本并将待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量，只需将待测向量的特征信息输入至所述多维空间中赋值的分布模型即可获得待测网页脚本的网页入侵脚本特征，因此降低了对计算硬件的要求，并在保证对网页入侵脚本特征的识别准确性的同时大大缩小了识别时间。

进一步的，本申请所述网页入侵脚本特征的识别方法及设备还用从每个所述参考网页脚本的网页脚本中截取之后剩余的网页脚本进一步确定所述待测网页脚本的网页入侵脚本特征类型，进一步提高识别的准确性。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

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）

5 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。

10

15

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。装置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，

20

25 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、一种网页入侵脚本特征的识别方法，其中，所述识别方法包括：

基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量，其中，所述参考向量的赋值根据对应参考网页脚本的网页入侵脚本特征类型确定；

根据所述参考向量及其赋值的分布，获取多维空间中赋值的分布模型；

基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量；以及

根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值，并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

2、根据权利要求 1 所述的识别方法，其中，将所述参考网页脚本映射至多维空间的参考向量以及将所述待测网页脚本映射至多维空间的待测向量基于相同的映射规则。

3、根据权利要求 1 或 2 所述的识别方法，其中，基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量包括：

获取若干所述参考网页脚本；

对所述参考网页脚本进行采样和量化，以获取所述参考网页脚本的特征值集合；以及

将每一所述参考网页脚本的特征值集合映射至所述多维空间的参考向量。

4、根据权利要求 3 所述的识别方法，其中，对所述参考网页脚本的特征进行采样和量化包括：

从每个所述参考网页脚本中截取相同文本容量的参考网页脚本；以及

对所述截取的参考网页脚本的字符进行采样和量化，以获取所述参考网页脚本的特征值集合。

5、根据权利要求 4 所述的识别方法，其中，对从所述参考网页脚本截取的字符依次赋予特征值包括：

根据 ASCII 码规则将所采样的所述参考网页脚本的字符量化成相应特征

值。

6、根据权利要求4或5所述的特征识别方法，其中，所述识别方法还包括：
利用从每个所述参考网页脚本中截取之后剩余的参考网页脚本确定所述待测网页脚本的网页入侵脚本特征类型。

5 7、根据权利要求6所述的识别方法，其中，所述识别方法还包括：

采用正则表达式对从每个所述参考网页脚本中截取之后剩余的参考网页脚本进行处理和存储。

8、根据权利要求1或7中任一项所述的识别方法，其中，基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量包括：

10 获取所述待测网页脚本；

对所述待测网页脚本进行采样和量化，以获取所述待测网页脚本的特征值集合；以及

将所述待测网页脚本的特征值集合映射至所述多维空间的待测向量。

9、根据权利要求8所述的识别方法，其中，对所述待测网页脚本的特征进行采样和量化包括：

从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本；以及

对所述截取的待测网页脚本的字符进行采样和量化，以获取所述待测网页脚本的特征值集合。

20 10、根据权利要求9所述的识别方法，其中，对所述截取的待测网页脚本的字符进行量化包括：

根据ASCII码规则将所采样的所述待测网页脚本的字符量化成相应特征值。

25 11、根据权利要求1至10中任一项所述的识别方法，其中，将所述多维空间进行划分包括：

利用分类算法获取所述多维空间中赋值的分布模型。

12、根据权利要求11所述的识别方法，其中，所述分类算法包括支持向量机算法。

13、根据权利要求1至12中任一项所述的识别方法，其中，所述多维空间

中赋值的分布模型包括若干多维函数关系组。

14、根据权利要求 13 所述的识别方法，其中，根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值包括：

将所述待测向量的赋值代入所述若干多维函数关系组，利用解析几何原理

5 获取所述待测向量所映射在所述多维空间中所对应的赋值。

15、根据权利要求 1 至 14 中任一项所述的识别方法，其中，所述网页入侵脚本特征类型包括正常类型、异常类型及待定类型中的一种或任意组合。

16、一种网页入侵脚本特征的识别设备，其中，所述识别设备包括：

第一装置，用于基于每个参考网页脚本的特征信息将所述参考网页脚本映射至多维空间的参考向量，其中，所述参考向量的赋值根据对应参考网页脚本的网页入侵脚本特征类型确定；

第二装置，用于根据所述参考向量及其赋值的分布，获取多维空间中赋值的分布模型；

第三装置，用于基于待测网页脚本的特征信息将所述待测网页脚本映射至多维空间的待测向量；以及

第四装置，用于根据所述多维空间中赋值的分布模型确定所述待测向量所映射在所述多维空间中所对应的赋值，并根据该赋值确定对应待测网页脚本的网页入侵脚本特征类型。

17、根据权利要求 16 所述的识别设备，其中，所述第一装置将所述参考网页脚本映射至多维空间的参考向量以及第三装置将所述待测网页脚本映射至多维空间的待测向量基于相同的映射规则。

18、根据权利要求 16 或 17 所述的识别设备，其中，所述第一装置包括：

第一单元，用于获取若干所述参考网页脚本；

第二单元，用于对所述参考网页脚本进行采样和量化，以获取所述参考网页脚本的特征值集合；以及

第三单元，用于将每一所述参考网页脚本的特征值集合映射至所述多维空间的参考向量。

19、根据权利要求 18 所述的识别设备，其中，所述第二单元包括：

第一子单元，用于从每个所述参考网页脚本中截取相同文本容量的参考网

页脚本；以及

第二子单元，用于对所述截取的参考网页脚本的字符进行采样和量化，以获取所述参考网页脚本的特征值集合。

20、根据权利要求 19 所述的识别设备，其中，所述第二子单元包括：

- 5 第二一子单元，用于根据 ASCII 码规则将所采样的所述参考网页脚本的字符量化成相应特征值。

21、根据权利要求 19 或 20 所述的识别设备，其中，所述识别设备还包括：

第五装置，用于利用从每个所述参考网页脚本中截取之后剩余的参考网页脚本确定所述待测网页脚本的网页入侵脚本特征类型。

- 10 22、根据权利要求 21 所述的识别设备，其中，所述识别设备还包括：

第六装置，用于采用正则表达式对从每个所述参考网页脚本中截取之后剩余的参考网页脚本进行处理和存储。

23、根据权利要求 16 至 22 中任一项所述的识别设备，其中，所述第三装置包括：

- 15 第四单元，用于获取所述待测网页脚本；

第五单元，用于对所述待测网页脚本进行采样和量化，以获取所述待测网页脚本的特征值集合；以及

第六单元，用于将所述待测网页脚本的特征值集合映射至所述多维空间的待测向量。

- 20 24、根据权利要求 23 所述的识别设备，其中，所述第五单元包括：

第三子单元，用于从所述待测网页脚本中截取与所述参考网页脚本所截取的参考网页脚本相同文本容量的待测网页脚本；以及

第四子单元，用于对所述截取的待测网页脚本的字符进行采样和量化，以获取所述待测网页脚本的特征值集合。

- 25 25、根据权利要求 24 所述的识别设备，其中，所述第四子单元包括：

第四一子单元，用于根据 ASCII 码规则将所采样的所述待测网页脚本的字符量化成相应特征值。

26、根据权利要求 16 至 25 中任一项所述的识别设备，其中，所述第二装置包括：

第七单元，用于利用分类算法获取所述多维空间中赋值的分布模型。

27、根据权利要求 26 所述的识别设备，其中，所述分类算法包括支持向量机算法。

28、根据权利要求 16 至 27 中任一项所述的识别设备，其中，所述多维空间
5 中赋值的分布模型包括若干多维函数关系组。

29、根据权利要求 28 所述的识别设备，其中，所述第四装置包括：

第八单元，用于将所述待测向量的赋值代入所述若干多维函数关系组，利用解析几何原理获取所述待测向量所映射在所述多维空间中所对应的赋值。

30、根据权利要求 15 至 29 中任一项所述的识别设备，其中，所述网页入
10 侵脚本特征类型包括正常类型、异常类型及待定类型中的一种或任意组合。

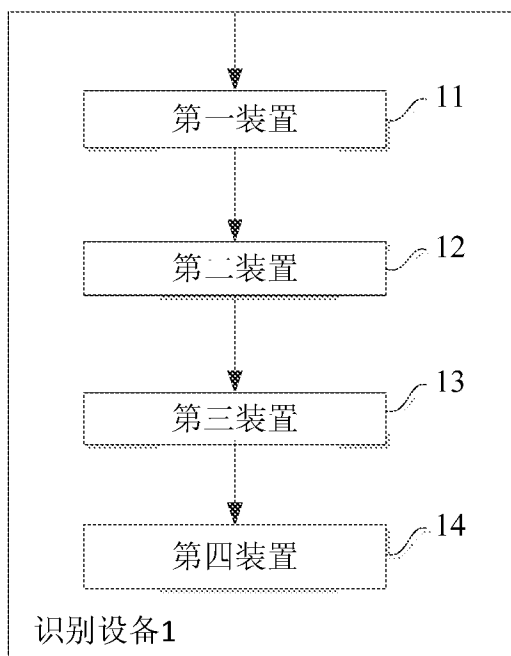


图 1

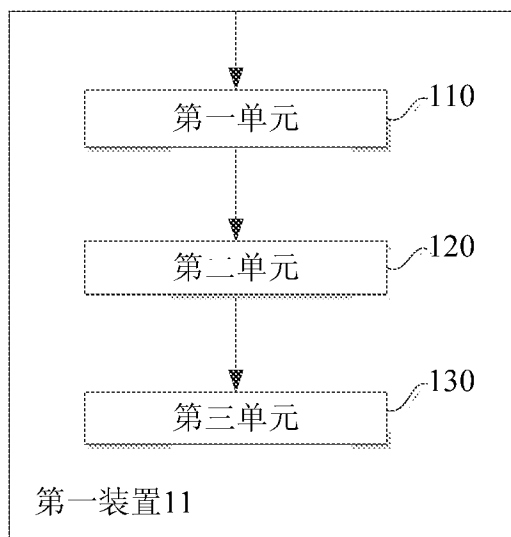


图 2

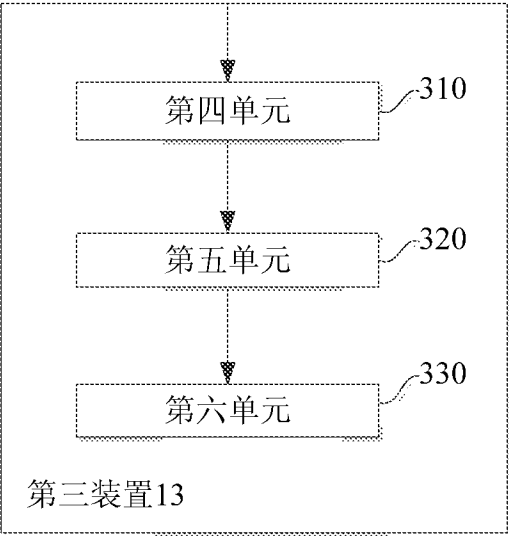


图 3

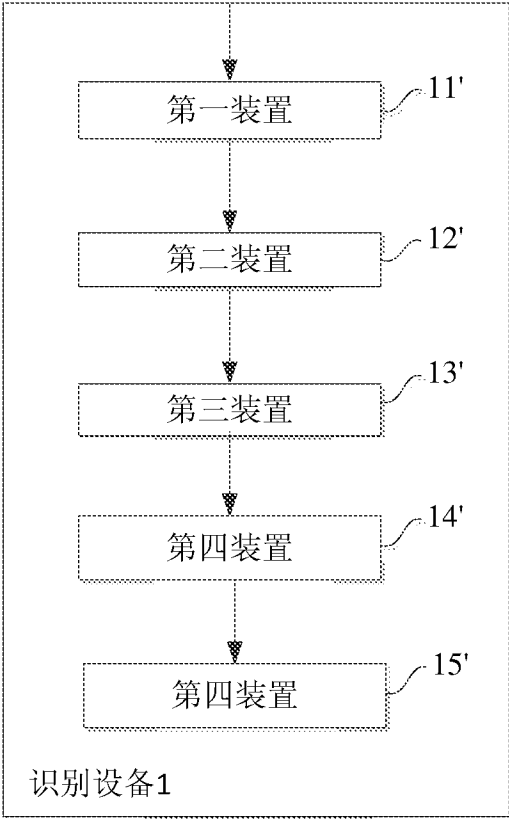


图 4

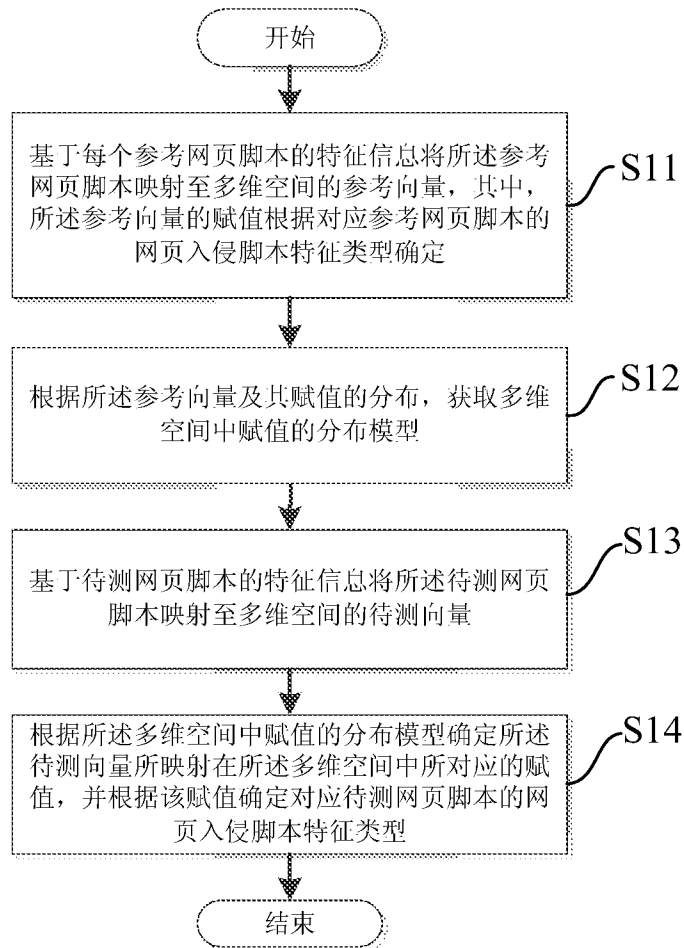


图 5

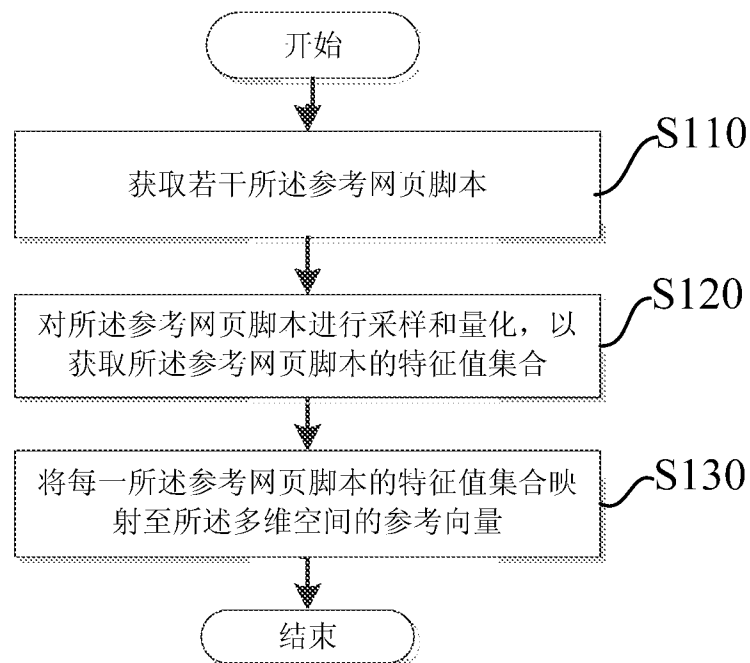


图 6

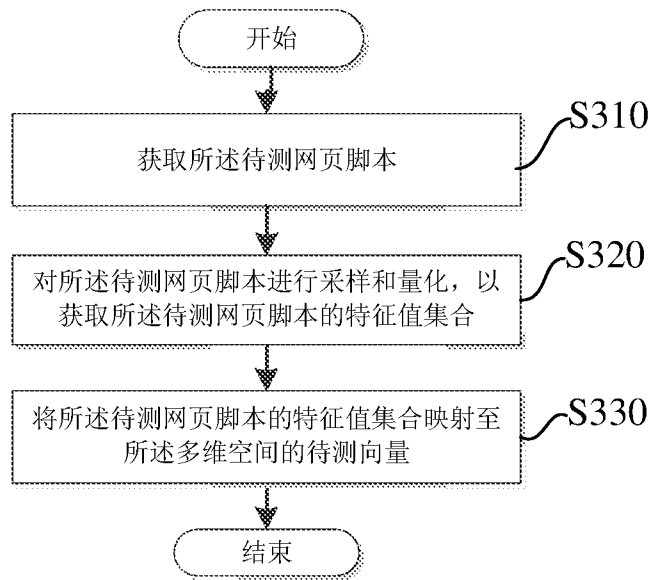


图 7

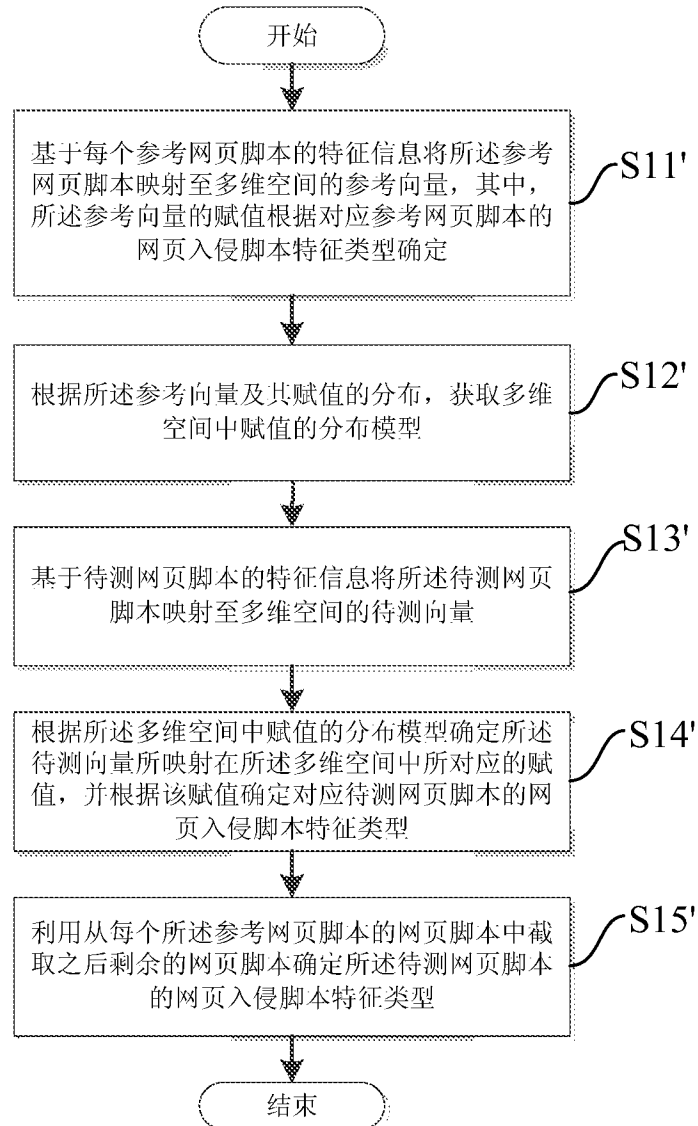


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/073148

A. CLASSIFICATION OF SUBJECT MATTER

G06F 21/56 (2013.01) i; 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)

G06F; H04L

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)

CNKI, CNPAT, EPODOC, WPI: web page, invade, script, information interchange code, support vector machine, web, shell, trojan, backdoor, characteristic, feature, vector, model, type, detect, monitor, ASCII, SVM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103973664 A (INFOSEC CO., LTD.), 06 August 2014 (06.08.2014), description, paragraphs [0028]-[0048]	1-30
A	CN 104331663 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 04 February 2015 (04.02.2015), the whole document	1-30
A	KR 20140042478 A (EZ SECURE INC.), 07 April 2014 (07.04.2014), the whole document	1-30
A	KR 20130077184 A (SECUI CORP.), 09 July 2013 (09.07.2013), the whole document	1-30
A	CN 103294952 A (BEIJING ANTIY ELECTRONIC EQUIPMENT CO., LTD.), 11 September 2013 (11.09.2013), the whole document	1-30

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 31 March 2016 (31.03.2016)	Date of mailing of the international search report 05 May 2016 (05.05.2016)
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 SONG, Chao Telephone No.: (86-10) 82245412

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2016/073148

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103973664 A	06 August 2014	KR 101291782 B1	31 July 2013
		US 2014215619 A1	31 July 2014
		JP 2014146307 A	14 August 2014
CN 104331663 A	04 February 2015	None	
KR 20140042478 A	07 April 2014	None	
KR 20130077184 A	09 July 2013	None	
CN 103294952 A	11 September 2013	None	

国际检索报告

国际申请号

PCT/CN2016/073148

A. 主题的分类

G06F 21/56(2013.01)i; 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)

G06F; H04L

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))

CNKI, CNPAT, EPDOC, WPI; 网页, 入侵, 脚本, 木马, 后门, 特征, 向量, 模型, 类型, 检测, 探测, 信息互换代码, 信息交换代码, 支持向量机, web, shell, trojan, backdoor, characteristic, feature, vector, model, type, detect, monitor, ASCII, SVM

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 103973664 A (信息安全有限公司) 2014年 8月 6日 (2014 - 08 - 06) 说明书[0028]-[0048]段	1-30
A	CN 104331663 A (北京奇虎科技有限公司等) 2015年 2月 4日 (2015 - 02 - 04) 全文	1-30
A	KR 20140042478 A (EZ SECURE INC.) 2014年 4月 7日 (2014 - 04 - 07) 全文	1-30
A	KR 20130077184 A (SECUI CORP.) 2013年 7月 9日 (2013 - 07 - 09) 全文	1-30
A	CN 103294952 A (北京安天电子设备有限公司) 2013年 9月 11日 (2013 - 09 - 11) 全文	1-30

☐ 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

2016年 3月 31日

国际检索报告邮寄日期

2016年 5月 5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

授权官员

宋朝

电话号码 (86-10)82245412

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/073148

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	103973664	A	2014年 8月 6日	KR	101291782	B1	2013年 7月 31日
				US	2014215619	A1	2014年 7月 31日
				JP	2014146307	A	2014年 8月 14日
CN	104331663	A	2015年 2月 4日	无			
KR	20140042478	A	2014年 4月 7日	无			
KR	20130077184	A	2013年 7月 9日	无			
CN	103294952	A	2013年 9月 11日	无			