(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 106295337 B
(45)授权公告日 2018.05.22

(21)申请号 201510377521.3

(22)申请日 2015.06.30

(65)同一申请的已公布的文献号
申请公布号 CN 106295337 A

(43)申请公布日 2017.01.04

(73)专利权人 安一恒通(北京)科技有限公司
地址 100091 北京市海淀区东北旺西路8号,中关村软件园4号楼C座1-03

(72)发明人 张壮 赵长坤 曹亮 董志强

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 马晓亚

(51)Int.Cl.
G06F 21/56(2013.01)
G06F 21/57(2013.01)

(56)对比文件
CN 102708313 A,2012.10.03,
CN 103839006 A,2014.06.04,

审查员 李小青

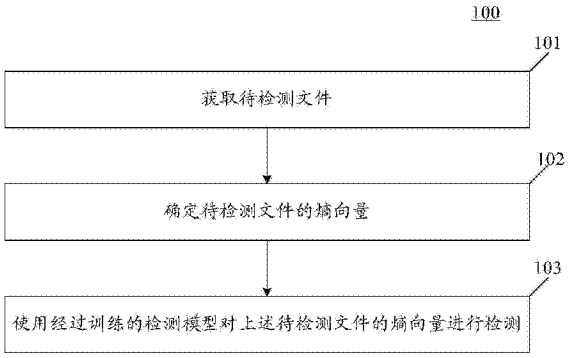
权利要求书3页 说明书7页 附图3页

(54)发明名称

用于检测恶意漏洞文件的方法、装置及终端

(57)摘要

本申请公开了用于检测恶意漏洞文件的方法、装置及终端。所述方法的一具体实施方式包括:获取待检测文件;确定所述待检测文件的熵向量;使用经过训练的检测模型对所述待检测文件的熵向量进行检测,以确定所述待检测文件是否为恶意漏洞文件,其中,所述待检测文件的文件类型和所述检测模型对应的文件类型相同。该实施方式通过提取待检测文件的熵向量,并基于待检测文件的熵向量,确定该待检测文件是否为恶意漏洞文件。解决了现有技术中对恶意漏洞文件查杀速度慢,查杀能力和效率低下的技术问题,提高了对恶意漏洞文件的查杀效率。



CN 106295337 B

1.一种用于检测恶意漏洞文件的方法,其特征在于,所述方法包括:

获取待检测文件;

确定所述待检测文件的熵向量;

使用经过训练的检测模型对所述待检测文件的熵向量进行检测,以确定所述待检测文件是否为恶意漏洞文件,其中,所述待检测文件的文件类型和所述检测模型对应的文件类型相同;

其中,所述检测模型通过如下方式获得:

获取多个文件类型相同且安全类别已知的文件作为训练文件,其中,所述安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别;

按照所述安全类别对所述训练文件进行安全类别的标识;

确定所述训练文件的熵向量;

基于所述训练文件的熵向量和安全类别标识训练并输出检测模型;

其中,所述训练并输出检测模型包括:

基于所述训练文件的熵向量和安全类别标识获得初始检测模型;

测试所述初始检测模型的误判率是否小于预定阈值;

如果否,循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤;

响应于修正后的检测模型的误判率小于预定阈值,停止循环并输出所述修正后的检测模型;

其中,通过如下方式确定文件的熵向量:

将文件切分为预定数量的切片;

获取每个所述切片的熵值;

将所述切片的数量作为熵向量的维度数,每个所述切片对应一个熵向量的方向,基于每个所述切片的熵值确定文件的熵向量。

2.根据权利要求1所述的方法,其特征在于,所述获得初始检测模型,包括:

从所述训练文件中获取部分文件作为第一文件;

对所述第一文件的熵向量进行特征分类;

基于所述特征分类的结果以及所述第一文件的安全类别标识,学习得到初始检测模型。

3.根据权利要求2所述的方法,其特征在于,测试检测模型的误判率是否小于预定阈值,包括:

从所述训练文件中获取部分文件作为第二文件;

使用待测试的检测模型对所述第二文件的熵向量进行检测;

根据检测的结果以及所述第二文件的安全类别标识确定误判率;

将所述误判率与所述预定阈值进行比较,以确定所述误判率是否小于预定阈值;

其中,所述第二文件中不包含所述第一文件;

其中,所述根据检测的结果以及所述第二文件的安全类别标识确定误判率包括:

使用所述初始检测模型对每个第二文件的熵向量进行检测,判断每个第二文件的安全类别;

将判断结果出现错误的次数除以测试的总次数，获得所述初始检测模型的误判率。

4.根据权利要求3所述的方法，其特征在于，所述修正当前检测模型，包括以下至少一项：

增加第一文件的数量并重新学习得到检测模型；以及

调整熵向量的维度数并重新学习得到检测模型。

5.一种用于检测恶意漏洞文件的装置，其特征在于，所述装置包括：

获取单元，用于获取待检测文件；

确定单元，用于确定所述待检测文件的熵向量；

检测单元，用于使用经过训练的检测模型对所述待检测文件的熵向量进行检测，以确定所述待检测文件是否为恶意漏洞文件，其中，所述待检测文件的文件类型和所述检测模型对应的文件类型相同；

其中，所述检测模型通过如下方式获得：

获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，所述安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别；

按照所述安全类别对所述训练文件进行安全类别的标识；

确定所述训练文件的熵向量；

基于所述训练文件的熵向量和安全类别标识训练并输出检测模型；

其中，所述训练并输出检测模型包括：

基于所述训练文件的熵向量和安全类别标识获得初始检测模型；

测试所述初始检测模型的误判率是否小于预定阈值；

如果否，循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤；

响应于修正后的检测模型的误判率小于预定阈值，停止循环并输出所述修正后的检测模型；

其中，通过如下方式确定文件的熵向量：

将文件切分为预定数量的切片；

获取每个所述切片的熵值；

将所述切片的数量作为熵向量的维度数，每个所述切片对应一个熵向量的方向，基于每个所述切片的熵值确定文件的熵向量。

6.一种终端，其特征在于，所述终端包括处理器，存储器；

其中，所述存储器用于存储经过训练的检测模型，所述处理器用于获取待检测文件，确定所述待检测文件的熵向量，并使用经过训练的检测模型对所述待检测文件的熵向量进行检测，以确定所述待检测文件是否为恶意漏洞文件，其中，所述待检测文件的文件类型和所述检测模型对应的文件类型相同；

其中，所述检测模型通过如下方式获得：

获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，所述安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别；

按照所述安全类别对所述训练文件进行安全类别的标识；

确定所述训练文件的熵向量；

基于所述训练文件的熵向量和安全类别标识训练并输出检测模型；

其中，所述训练并输出检测模型包括：

基于所述训练文件的熵向量和安全类别标识获得初始检测模型；

测试所述初始检测模型的误判率是否小于预定阈值；

如果否，循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤；

响应于修正后的检测模型的误判率小于预定阈值，停止循环并输出所述修正后的检测模型；

其中，通过如下方式确定文件的熵向量：

将文件切分为预定数量的切片；

获取每个所述切片的熵值；

将所述切片的数量作为熵向量的维度数，每个所述切片对应一个熵向量的方向，基于每个所述切片的熵值确定文件的熵向量。

用于检测恶意漏洞文件的方法、装置及终端

技术领域

[0001] 本申请涉及计算机技术领域，具体漏洞检测技术领域，尤其涉及用于检测恶意漏洞文件的方法、装置及终端。

背景技术

[0002] 目前，随着计算机技术的不断发展，计算机已经被广泛地应用于人们的日常生活中，并且功能也越来越多，成为人们生活和工作的重要工具。然而，有一些个人或组织利用先进的攻击手段对特定目标进行长期持续性网络攻击，从而导致恶意代码的执行及敏感信息泄露，威胁了网络的安全。

[0003] 现有的检测恶意漏洞文件的方法有两种：一种为静态检测方法，一种为动态执行的检测方法。静态特征检测方法是比较常用的方法，检测的方式有两种：A、通过文件格式的异常检测异常文档。B、通过检测漏洞利用文件的固定特征检测异常文档。动态执行检测方法是一种启发式检测的方法。在一些较为高级的启发环境中会使用模拟环境执行待执行的文档，检测正常文档不存在的行为。如果文档触发了shellcode(填充数据，属于漏洞代码)，就会具有文档本身不应该有的行为。比如：链接网络、执行程序、注入进程等等。

[0004] 然而，对于静态检测方法来说，可以通过构造文档结构和改变shellcode，很容易的绕过静态检测的方法。因此，静态检测方法启发查杀能力很差，对于新出现的恶意漏洞文件没有什么查杀的能力。对于动态执行检测方法来说，有多种方法可以检测动态执行的虚拟环境，导致不触发相关的病毒代码，从而使得检测失败。因此，动态执行检测方法有一定的启发能力，但是效率低下，速度慢，并且启发能力不是很高。

发明内容

[0005] 本申请提供了一种用于检测恶意漏洞文件的方法、装置及终端。解决了现有技术中对恶意漏洞文件查杀速度慢，查杀能力和效率低下的问题。

[0006] 第一方面，本申请提供了一种用于检测恶意漏洞文件的方法，所述方法包括：获取待检测文件；确定所述待检测文件的熵向量；使用经过训练的检测模型对所述待检测文件的熵向量进行检测，以确定所述待检测文件是否为恶意漏洞文件，其中，所述待检测文件的文件类型和所述检测模型对应的文件类型相同。

[0007] 在某些实施方式中，所述检测模型通过如下方式获得：获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，所述安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别；按照所述安全类别对所述训练文件进行安全类别的标识；确定所述训练文件的熵向量；基于所述训练文件的熵向量和安全类别标识训练并输出检测模型。

[0008] 在某些实施方式中，所述训练并输出检测模型包括：基于所述训练文件的熵向量和安全类别标识获得初始检测模型；

[0009] 测试所述初始检测模型的误判率是否小于预定阈值；如果否，循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤；响应于修

正后的检测模型的误判率小于预定阈值，停止循环并输出所述修正后的检测模型。

[0010] 在某些实施方式中，所述获得初始检测模型，包括：从所述训练文件中获取部分文件作为第一文件；对所述第一文件的熵向量进行特征分类；基于所述特征分类的结果以及所述第一文件的安全类别标识，学习得到初始检测模型。

[0011] 在某些实施方式中，测试检测模型的误判率是否小于预定阈值，包括：从所述训练文件中获取部分文件作为第二文件；使用待测试的检测模型对所述第二文件的熵向量进行检测；根据检测的结果以及所述第二文件的安全类别标识确定误判率；将所述误判率与所述预定阈值进行比较，以确定所述误判率是否小于预定阈值；其中，所述第二文件中不包含所述第一文件。

[0012] 在某些实施方式中，所述修正当前检测模型，包括以下至少一项：增加第一文件的数量并重新学习得到检测模型；以及调整熵向量的维度数并重新学习得到检测模型。

[0013] 在某些实施方式中，通过如下方式确定文件的熵向量：将文件切分为预定数量的切片；获取每个所述切片的熵值；将所述切片的数量作为熵向量的维度数，每个所述切片对应一个熵向量的方向，基于每个所述切片的熵值确定文件的熵向量。

[0014] 第二方面，本申请提供了一种用于检测恶意漏洞文件的装置，获取单元，用于获取待检测文件；确定单元，用于确定所述待检测文件的熵向量；检测单元，用于使用经过训练的检测模型对所述待检测文件的熵向量进行检测，以确定所述待检测文件是否为恶意漏洞文件，其中，所述待检测文件的文件类型和所述检测模型对应的文件类型相同。

[0015] 在某些实施方式中，所述检测模型通过如下方式获得：获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，所述安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别；按照所述安全类别对所述训练文件进行安全类别的标识；确定所述训练文件的熵向量；基于所述训练文件的熵向量和安全类别标识训练并输出检测模型。

[0016] 在某些实施方式中，所述训练并输出检测模型包括：基于所述训练文件的熵向量和安全类别标识获得初始检测模型；测试所述初始检测模型的误判率是否小于预定阈值；如果否，循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤；响应于修正后的检测模型的误判率小于预定阈值，停止循环并输出所述修正后的检测模型。

[0017] 在某些实施方式中，所述获得初始检测模型，包括：从所述训练文件中获取部分文件作为第一文件；对所述第一文件的熵向量进行特征分类；基于所述特征分类的结果以及所述第一文件的安全类别标识，学习得到初始检测模型。

[0018] 在某些实施方式中，测试检测模型的误判率是否小于预定阈值，包括：从所述训练文件中获取部分文件作为第二文件；使用待测试的检测模型对所述第二文件的熵向量进行检测；根据检测的结果以及所述第二文件的安全类别标识确定误判率；将所述误判率与所述预定阈值进行比较，以确定所述误判率是否小于预定阈值；其中，所述第二文件中不包含所述第一文件。

[0019] 在某些实施方式中，所述修正当前检测模型，包括以下至少一项：增加第一文件的数量并重新学习得到检测模型；以及调整熵向量的维度数并重新学习得到检测模型。

[0020] 在某些实施方式中，所述确定单元配置用于：将文件切分为预定数量的切片；获取每个所述切片的熵值；将所述切片的数量作为熵向量的维度数，每个所述切片对应一个熵

向量的方向，基于每个所述切片的熵值确定文件的熵向量。

[0021] 第三方面，本申请提供了一种终端，所述终端包括处理器，存储器；其中，所述存储器用于存储经过训练的检测模型，所述处理器用于获取待检测文件，确定所述待检测文件的熵向量，并使用经过训练的检测模型对所述待检测文件的熵向量进行检测，以确定所述待检测文件是否为恶意漏洞文件，其中，所述待检测文件的文件类型和所述检测模型对应的文件类型相同。

[0022] 在某些实施方式中，所述检测模型通过如下方式获得：获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，所述安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别；按照所述安全类别对所述训练文件进行安全类别的标识；确定所述训练文件的熵向量；基于所述训练文件的熵向量和安全类别标识训练并输出检测模型。

[0023] 本申请提供的用于识别手势的方法、装置及终端，通过提取待检测文件的熵向量，并基于待检测文件的熵向量，确定该待检测文件是否为恶意漏洞文件。解决了现有技术中对恶意漏洞文件查杀速度慢，查杀能力和效率低下的技术问题，提高了对恶意漏洞文件的查杀效率。

附图说明

[0024] 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：

[0025] 图1是本申请实施例提供的用于检测恶意漏洞文件的方法的一个实施例的流程图；

[0026] 图2是本申请实施例提供的恶意漏洞文件内容的熵值曲线变化示意图；

[0027] 图3是本申请实施例提供的包含shellcode的文件的内容熵值曲线变化示意图；

[0028] 图4是本申请实施例提供的获得检测模型的方法的一个实施例的流程图；

[0029] 图5是本申请实施例提供的用于检测恶意漏洞文件的装置的一个实施例的结构示意图；

[0030] 图6是本申请实施例提供的终端的一个实施例的结构示意图。

具体实施方式

[0031] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0032] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0033] 本申请所涉及的终端可以包括但不限于智能手机、平板电脑、个人数字助理、膝上型便携计算机以及台式电脑等等。出于示例描述目的以及为了简洁起见，在接下来的讨论中，结合台式电脑来描述本申请的示例性实施例。

[0034] 请参考图1，其示出了根据本申请的用于检测恶意漏洞文件的方法的一个实施例的流程100。

[0035] 如图1所示，在步骤101中，获取待检测文件。

[0036] 接着，在步骤102中，确定待检测文件的熵向量。

[0037] 一般来说，恶意漏洞文件在文档中构造了大量的重复字串，然后构造ROP(Return-oriented programming，返回导向编程)，执行其他模块中的代码，从而绕过DEP(Data Execution Prevention，数据执行保护)以释放病毒体。

[0038] 在本实施例中，通过对一些恶意漏洞文件进行深入地分析发现，构造好的可疑文件会将病毒文件加密放在文本后面，此部分内容的熵值必然非常大，而且会用大量的重复数据填充，所以文件的熵值曲线应该在末尾有一个突增。

[0039] 例如，图2示出了恶意漏洞文件内容的熵值曲线变化示意图，如图2所示，横坐标表示文件内容的片段的位置，横坐标的原点表示文件头的位置，横坐标的值越大表示文件的内容片段越靠后。纵坐标表示文件中对应于横坐标位置的内容片段的熵值。从图2可以看出恶意漏洞文件内容在末尾的片段的熵值有一个突增。

[0040] 又例如，图3示出了包含shellcode的文件的内容熵值曲线变化示意图，如图3所示，横坐标表示文件内容的片段的位置，横坐标的原点表示文件头的位置，横坐标的值越大表示文件的内容片段越靠后。纵坐标表示文件中对应于横坐标位置的内容片段的熵值。从图3可以看出，包含shellcode的文件的内容熵值包含了大量的连续的4左右的数据。

[0041] 因此，在本实施例中，可以根据待检测文件的熵向量的特征判断待检测文件是否为恶意漏洞文件。

[0042] 需要说明的是，文件片段的熵值表征了该文件片段的混乱程度，文字、图片、代码、压缩包、应用程序等，由于组织方式不同，熵值也不相同。例如，图片经过压缩，压缩包也经过压缩，其熵值就会很高，而且有一定的规律。可以用数据编码的信息熵来表示编码的状态。

[0043] 在本实施例中，可以通过如下方式确定文件的熵向量：首先，将文件等值切分为预定数量的切片，计算每一片切片的信息熵，以表示整个文件编码的变化情况。其中，预定数量可以是用户预先设定的一个值，可以理解，本申请对预定数量的具体数值不限定。将上述切片的数量作为熵向量的维度数，每个切片对应一个熵向量的方向，基于每个切片的熵值确定文件的熵向量。例如，假设将文件等值切分为3个切片，分别为切片i，切片j，切片k，计算这3个切片对应的熵值，分别为a，b，c，则该文件的熵向量为3维向量，熵向量可以表示为 $\vec{S}=a\vec{i}+b\vec{j}+c\vec{k}$ 。

[0044] 最后，在步骤103中，使用经过训练的检测模型对上述待检测文件的熵向量进行检测，以确定该待检测文件是否为恶意漏洞文件。

[0045] 在本实施例中，可以使用经过训练的检测模型对上述待检测文件的熵向量进行检测，分析待检测文件的熵向量的特征，根据待检测文件的熵向量的特征确定该待检测文件是否为恶意漏洞文件。

[0046] 需要说明的是，恶意漏洞文件的文件类型不同，其熵向量的特征就不相同。因此，每种文件类型对应一种检测模型，在检测待检测文件时，所选取的检测模型对应的文件类型与待检测文件的文件类型相同。

[0047] 本申请的上述实施例提供的方法，通过提取待检测文件的熵向量，并基于待检测文件的熵向量，确定该待检测文件是否为恶意漏洞文件。解决了现有技术中对恶意漏洞文件查杀速度慢，查杀能力和效率低下的技术问题，提高了对恶意漏洞文件的查杀效率。

[0048] 进一步参考图4，其示出了获得检测模型的方法的一个实施例的流程400。

[0049] 如图4所示，在步骤401中，获取多个文件类型相同且安全类别已知的文件作为训练文件。

[0050] 在本实施例中，任意获取多个文件类型相同的文件作为训练文件，并且这些文件的安全类别已知，其中，安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别。需要说明的是，上述训练文件的安全类别可以是通过其它方法确定的，可以理解，本申请对确定上述训练文件的安全类别的具体方法不限定。

[0051] 接着，在步骤402中，按照安全类别对上述训练文件进行安全类别的标识。

[0052] 在本实施例中，按照训练文件的安全类别对上述训练文件进行标识，在本实施例的一种实现中，可以采用特殊颜色对训练文件进行安全类别的标识，不同的颜色表示不同的安全类别。在本实施例的另一种实现中，可以采用特殊符号对训练文件进行安全类别的标识，不同的符号表示不同的安全类别。可以理解，还可以通过其它的方式对上述训练文件进行安全类别的标识，本申请对此方面不限定。

[0053] 继而，在步骤403中，确定上述训练文件的熵向量。

[0054] 最后，在步骤404中，基于上述训练文件的熵向量和安全类别标识训练并输出检测模型。

[0055] 在本实施例中，首先，基于上述训练文件的熵向量和安全类别标识获得初始检测模型。具体地，先从训练文件中获取部分文件作为第一文件，对第一文件的熵向量进行特征分类。可以采用SVM(Support Vector Machine，支持向量机)算法对第一文件的熵向量进行特征分类。可以理解，还可以采用其它的方式对第一文件的熵向量进行特征分类，本申请对进行特征分类所采用的具体方式方面不限定。然后，基于特征分类的结果以及第一文件的安全类别标识，学习得到初始检测模型。

[0056] 接着，测试上述初始检测模型的误判率是否小于预定阈值。具体地，从训练文件中不包含第一文件的部分文件中获取多个文件作为第二文件(第二文件中不包含第一文件)，使用初始检测模型(待测试的检测模型)对每个第二文件的熵向量进行检测，判断每个第二文件的安全类别。然后将初始检测模型判断的结果与每个第二文件的安全类别标识进行比较。如果初始检测模型判断的结果与某个第二文件的安全类别标识所对应的安全类别一致，则该判断结果正确。如果初始检测模型判断的结果与某个第二文件的安全类别标识所对应的安全类别不一致，则该判断结果错误。用判断结果出现错误的次数除以测试的总次数，从而获得该初始检测模型的误判率。将该误判率与预定阈值进行比较，以确定误判率是否小于预定阈值。

[0057] 继而，如果初始检测模型的误判率大于等于预定阈值，说明该模型的准确率不够高，因此，循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤。具体地，修正当前检测模型，可以包括以下至少一项：增加第一文件的数量并重新学习得到检测模型以及调整熵向量的维度数并重新学习得到检测模型。

[0058] 最后，响应于修正后的检测模型的误判率小于预定阈值，说明该模型的准确率已经满足条件，停止循环并输出修正后的检测模型。

[0059] 应当注意，尽管在附图中以特定顺序描述了本发明方法的操作，但是，这并非要求或者暗示必须按照该特定顺序来执行这些操作，或是必须执行全部所示的操作才能实现期

望的结果。相反，流程图中描绘的步骤可以改变执行顺序。例如，在图4的流程400中，可以先执行步骤403，确定上述训练文件的熵向量，然后再执行步骤402，按照安全类别对上述训练文件进行安全类别的标识。附加地或备选地，可以省略某些步骤，将多个步骤合并为一个步骤执行，和/或将一个步骤分解为多个步骤执行。

[0060] 进一步参考图5，其示出了根据本申请的用于检测恶意漏洞文件的装置的一个实施例的结构示意图。

[0061] 如图5所示，本实施例的装置500包括：获取单元501，确定单元502和检测单元503。其中，获取单元501用于获取待检测文件。确定单元502用于确定待检测文件的熵向量。检测单元503用于使用经过训练的检测模型对待检测文件的熵向量进行检测，以确定待检测文件是否为恶意漏洞文件，其中，待检测文件的文件类型和检测模型对应的文件类型相同。

[0062] 在一些可选实施方式中，检测模型通过如下方式获得：获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别。按照安全类别对训练文件进行安全类别的标识。确定训练文件的熵向量。基于训练文件的熵向量和安全类别标识训练并输出检测模型。

[0063] 在一些可选实施方式中，训练并输出检测模型包括：基于训练文件的熵向量和安全类别标识获得初始检测模型。测试初始检测模型的误判率是否小于预定阈值。如果否，循环执行修正当前检测模型的步骤以及测试修正后的检测模型的误判率是否小于预定阈值的步骤。响应于修正后的检测模型的误判率小于预定阈值，停止循环并输出所述修正后的检测模型。

[0064] 在一些可选实施方式中，获得初始检测模型，包括：从训练文件中获取部分文件作为第一文件。对第一文件的熵向量进行特征分类。基于特征分类的结果以及第一文件的安全类别标识，学习得到初始检测模型。

[0065] 在一些可选实施方式中，测试检测模型的误判率是否小于预定阈值，包括：从训练文件中获取部分文件作为第二文件。使用待测试的检测模型对第二文件的熵向量进行检测。根据检测的结果以及第二文件的安全类别标识确定误判率。将误判率与预定阈值进行比较，以确定误判率是否小于预定阈值。其中，第二文件中不包含第一文件。

[0066] 在一些可选实施方式中，修正当前检测模型，包括以下至少一项：增加第一文件的数量并重新学习得到检测模型，以及调整熵向量的维度数并重新学习得到检测模型。

[0067] 在一些可选实施方式中，确定单元配置用于：将文件切分为预定数量的切片。获取每个切片的熵值。将切片的数量作为熵向量的维度数，每个切片对应一个熵向量的方向，基于每个切片的熵值确定文件的熵向量。

[0068] 应当理解，装置500中记载的诸单元或模块与参考图1-4描述的方法中的各个步骤相对应。由此，上文针对方法描述的操作和特征同样适用于装置500及其中包含的单元，在此不再赘述。装置500可以预先设置在终端中，也可以通过下载等方式而加载到终端中。装置500中的相应单元可以与终端中的单元相互配合以实现用于识别手势的方案。

[0069] 进一步参考图6，其示出了根据本申请的终端的一个实施例的结构示意图。

[0070] 如图6所示，本实施例的终端600包括：至少一个处理器601，例如CPU(Central Processing Unit，中央处理器)，至少一个通信接口602，至少一个用户接口603，存储器604，至少一个通信总线605。通信总线605用于实现上述组件之间的连接通信。终端600可选

的包含用户接口603，如显示组件，键盘或者点击设备(例如，鼠标，轨迹球(trackball)，触感板或者触感显示屏)等等。存储器604可能包含高速RAM(Random Access Memory，随机存取存储器)，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。存储器604可选的可以包含至少一个位于远离前述处理器601的存储装置。

[0071] 在一些实施方式中，存储器604存储了如下的元素，可执行模块或者数据结构，或者他们的子集，或者他们的扩展集：

[0072] 操作系统614，包含各种系统程序，用于实现各种基础业务以及处理基于硬件的任务。

[0073] 应用程序624，包含各种应用程序，用于实现各种应用业务。

[0074] 在本实施例中，存储器604用于存储经过训练的检测模型。处理器601用于获取待检测文件，确定待检测文件的熵向量，并使用经过训练的检测模型对待检测文件的熵向量进行检测，以确定待检测文件是否为恶意漏洞文件，其中，待检测文件的文件类型和所述检测模型对应的文件类型相同。

[0075] 进一步地，检测模型通过如下方式获得：获取多个文件类型相同且安全类别已知的文件作为训练文件，其中，安全类别包括恶意漏洞文件类别以及非恶意漏洞文件类别。按照安全类别对训练文件进行安全类别的标识。确定训练文件的熵向量。基于训练文件的熵向量和安全类别标识训练并输出检测模型。

[0076] 描述于本申请实施例中所涉及到的单元模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元模块也可以设置在处理器中，例如，可以描述为：一种处理器包括获取单元，确定单元，检测单元。其中，这些单元模块的名称在某种情况下并不构成对该单元模块本身的限定，例如，获取单元还可以被描述为“用于获取待检测文件的单元”。

[0077] 作为另一方面，本申请还提供了一种计算机可读存储介质，该计算机可读存储介质可以是上述实施例中所述装置中所包含的计算机可读存储介质；也可以是单独存在，未装配入终端中的计算机可读存储介质。所述计算机可读存储介质存储有一个或者一个以上程序，所述程序被一个或者一个以上的处理器用来执行描述于本申请的用于检测恶意漏洞文件的方法。

[0078] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离所述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

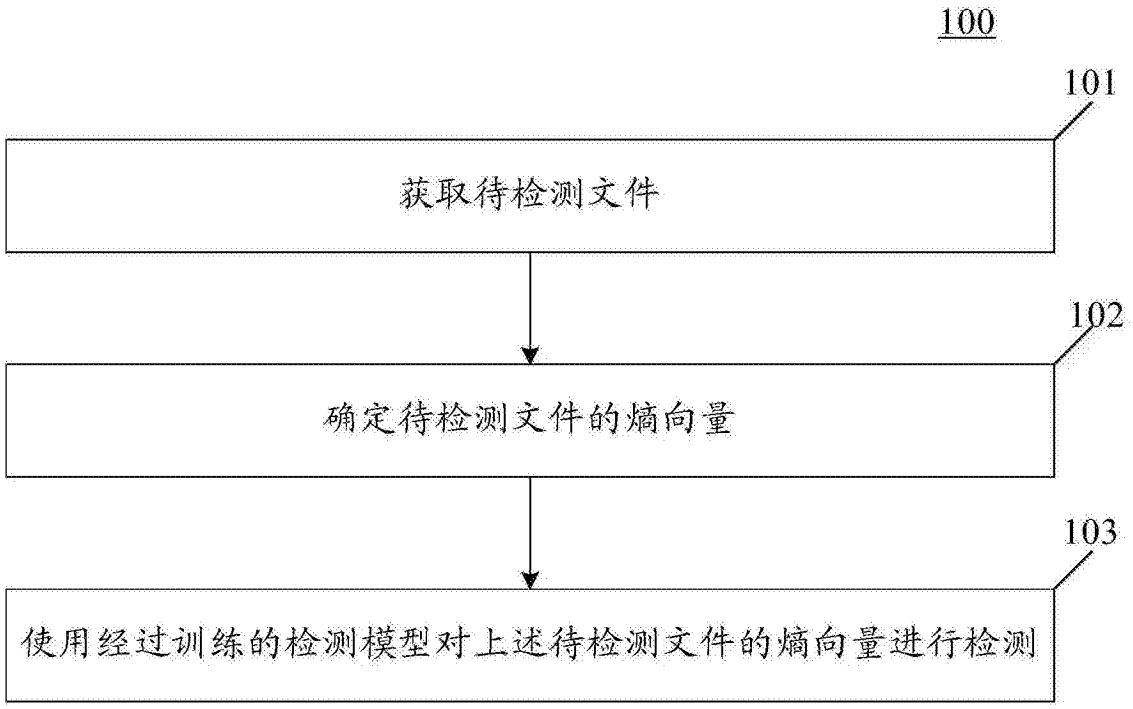

100
101
获取待检测文件
102
确定待检测文件的熵向量
103
使用经过训练的检测模型对上述待检测文件的熵向量进行检测

图1

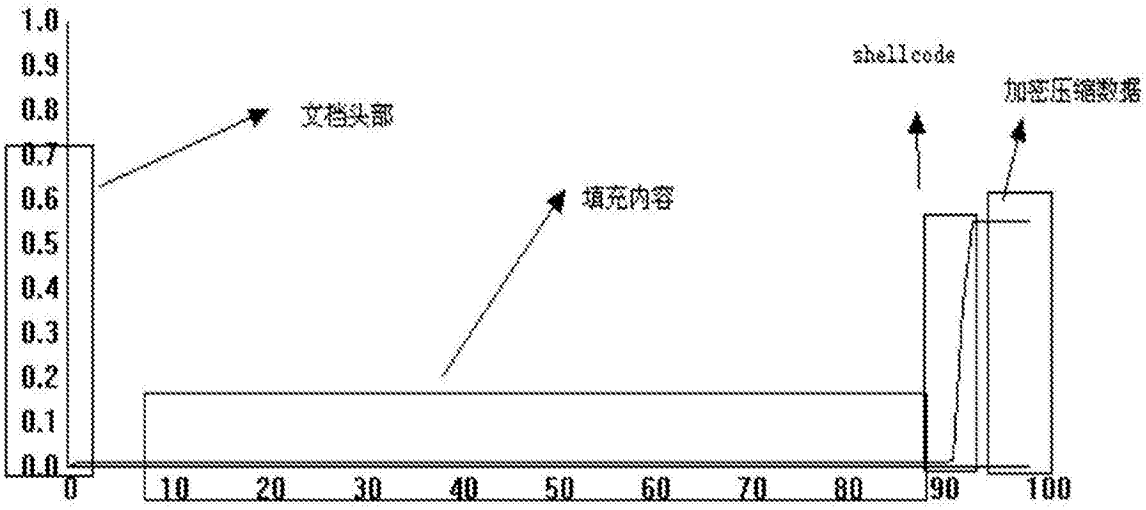

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
文档头部
填充内容
shellcode
加密压缩数据
0
10
20
30
40
50
60
70
80
90
100

图2

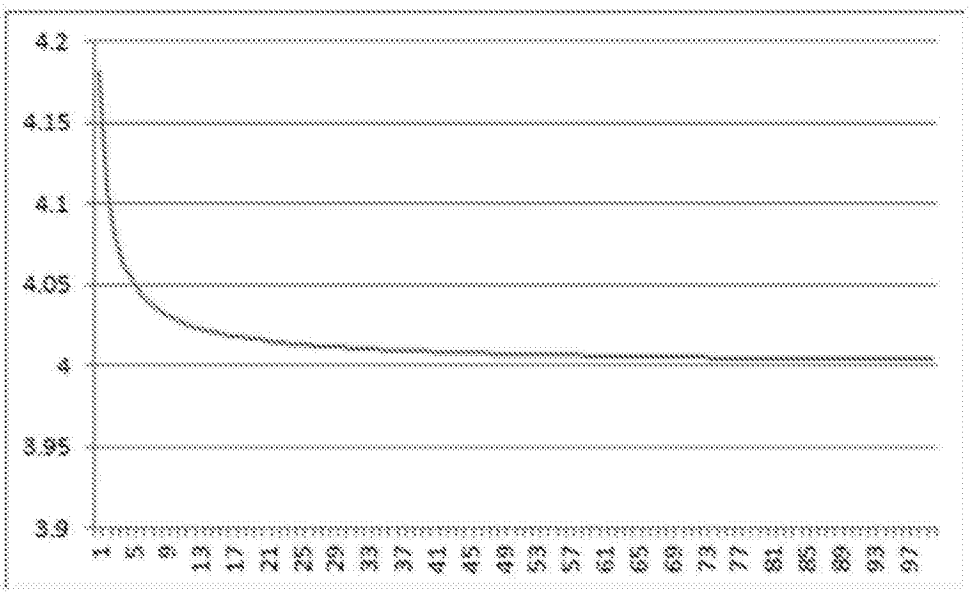

4.2
4.15
4.1
4.05
4
3.95
3.9

图3

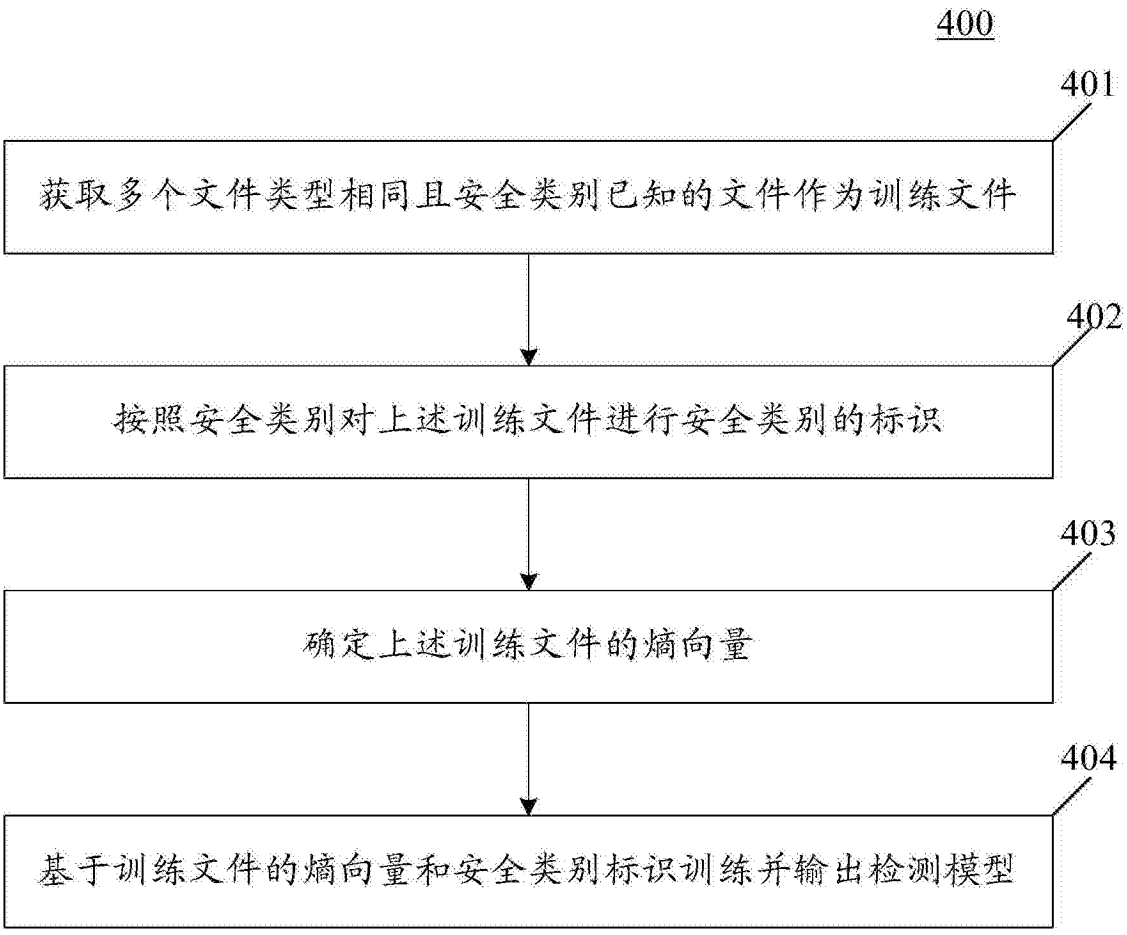

400
401
获取多个文件类型相同且安全类别已知的文件作为训练文件
402
按照安全类别对上述训练文件进行安全类别的标识
403
确定上述训练文件的熵向量
404
基于训练文件的熵向量和安全类别标识训练并输出检测模型

图4

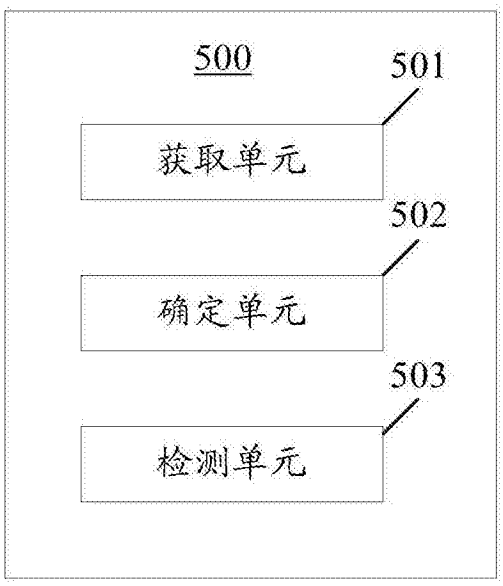

500
501
获取单元
502
确定单元
503
检测单元

图5

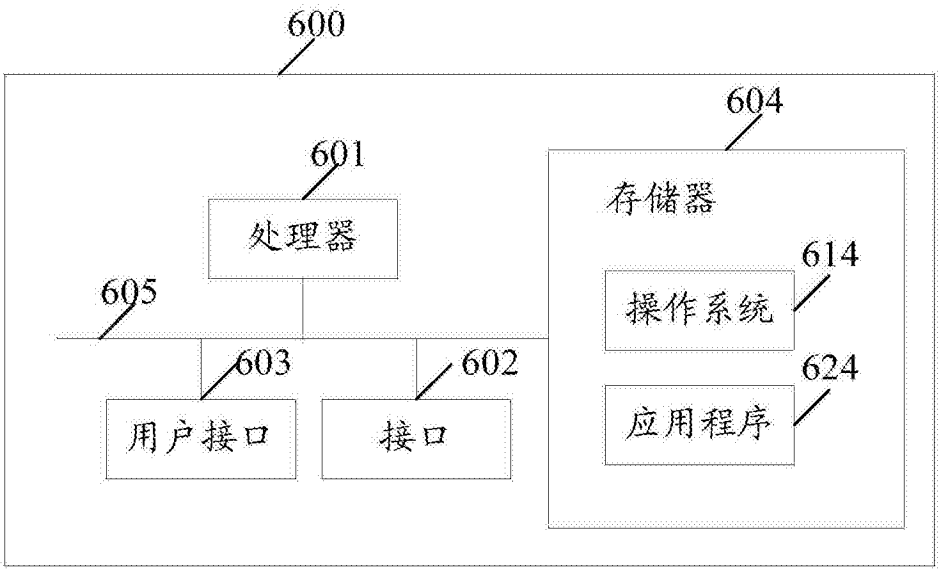

600
604
601
处理器
存储器
614
操作系统
605
603
602
624
用户接口
接口
应用程序

图6