(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112019449 A
(43)申请公布日 2020.12.01

(21)申请号 202010818919.7

(22)申请日 2020.08.14

(71)申请人 四川电科网安科技有限公司
地址 610000 四川省成都市中国(四川)自由贸易试验区成都高新区吉泰路666号1栋22层6号

(72)发明人 李勋 庄阿刚

(74)专利代理机构 北京细软智谷知识产权代理有限责任公司 11471
代理人 牛晴

(51)Int.Cl.
H04L 12/851(2013.01)

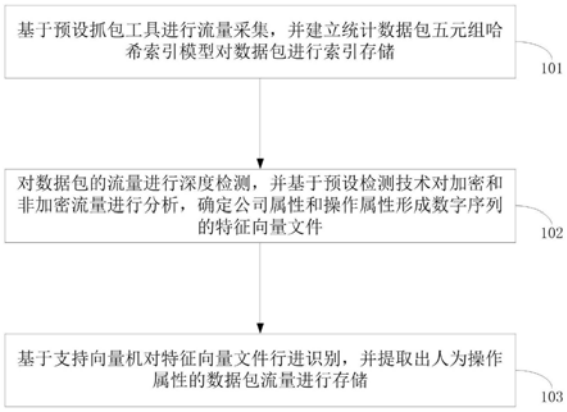
权利要求书2页 说明书6页 附图3页

(54)发明名称

流量识别抓包方法和装置

(57)摘要

本发明涉及一种流量识别抓包方法和装置，包括：基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储；对数据包的流量进行深度检测，并基于预设检测技术对加密和非加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件；基于支持向量机对特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储。本发明的有益效果为：基于特征向量文件所构成的样本和支持向量机进行自动完成对流量的分类，并且显著标明人在操作上网时候的特征，非人为上网如软件升级、后台自动运行等垃圾流量则过滤掉，只有识别为人为上网的通信才保存，不仅节省了存储空间而且便于进行查看等操作。



CN 112019449 A

1.一种流量识别抓包方法,其特征在于,包括:

基于预设抓包工具进行流量采集,并建立统计数据包五元组哈希索引模型对数据包进行索引存储;

对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件;

基于支持向量机对所述特征向量文件行进识别,并提取出人为操作属性的数据包流量进行存储。

2.根据权利要求1所述的流量识别抓包方法,其特征在于,所述基于预设抓包工具进行流量采集,并建立统计数据包五元组哈希索引模型对数据包进行索引存储包括:

基于抓包网卡采集流量并对数据包的源IP地址、目的IP地址、源端口、目的端口和协议类型五元组进行流量统计,并基于端口号建立统计数据包五元组哈希索引模型。

3.根据权利要求2所述的流量识别抓包方法,其特征在于,所述哈希索引模型中的每个匹配流至少包括:上行流量、下行流量、数据包个数、源和目标IP个数、平均流量以及数据包平均间隔。

4.根据权利要求3所述的流量识别抓包方法,其特征在于,所述对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件包括:

基于DNS检测技术对非加密流量进行分析,确定流量对应的公司属性。

5.根据权利要求4所述的流量识别抓包方法,其特征在于,所述对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件还包括:

基于深度检测技术对加密流量进行分析,确定流量对应的公司属性。

6.根据权利要求5所述的流量识别抓包方法,其特征在于,所述对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件还包括:

对数据流量包进行关键特征提取确定对应IP地址是否为人为操作属性。

7.根据权利要求6所述的流量识别抓包方法,其特征在于,所述关键特征至少包括:

长度、首32字节、尾部32字节、是否HTTP协议和是否含有域名信息。

8.根据权利要求7所述的流量识别抓包方法,其特征在于,所述基于支持向量机对所述特征向量文件行进识别,并提取出人为操作属性的数据包流量进行存储包括:

基于所述特征向量文件构建高维的样本特征向量数据为X=[x1,x2,…,xl];

将每个网络流量样本标记为D(X,y_i),其中y_i为该类流量数据样本标示的类别标签,$y_i \in \{+1,-1\}$;

区别不同类别的最优分类面表示为w·X+b=0,w为权重,b为常数项;

最优分类面满足:

$$\min \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

$y_i(w*X+b)-1 \geq 0$,i=1,...n;

其中i=1,…,n,中n表示样本数;

基于支持向量机分类判别函数进行判别：

$$f(x)=\mathrm{sgn}\{w*X+b\}=\mathrm{sgn}\{\sum_{i=1}^{l}\alpha l*y_i*xl+b\}\ ;$$

其中α1为优化的Lagrange算子。

9.根据权利要求1至8任一项所述的流量识别抓包方法，其特征在于，还包括：将识别为人为操作的数据包流量以IP地址为标记建立文件进行保存，并在存储容量超过50M时进行分割保存。

10.一种流量识别抓包装置，其特征在于，包括：

数据抓取模块，用于基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储；

深度检测模块，用于对数据包的流量进行深度检测，并基于预设检测技术对加密和非加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件；以及

识别模块，用于基于支持向量机对所述特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储。

流量识别抓包方法和装置

技术领域

[0001] 本发明属于数据采集技术领域,具体涉及一种流量识别抓包方法和装置。

背景技术

[0002] 作为一种常见的数据分析手段,现在的电子取证技术,都是从网口抓取数据包并保存到硬盘,然后分析这些数据包。抓包设备是对多个硬盘存储空间进行判断和管理,对每个抓取到的数据包信息采用连续存储的方式存储到硬盘上。

[0003] 而随着网络带宽和传输速度的不断发展,一般小区的带宽入口已经升级到了10Gb以上,家庭带宽1000M也已经在逐步的普及,现有的传统的抓包方法因无法有效识别人为上网,致使硬盘要么很快就被占满,要么就是记录了海量的垃圾数据,并且后期分析查看特别困难。

发明内容

[0004] 为了解决现有技术存在的无法有效识别人的上网行为导致存储垃圾文件占用硬盘存储空间的问题,本发明提供了一种流量识别抓包方法和装置,其具有有效识别人的上网行为,节省存储空间、便于查看等特点。

[0005] 根据本发明具体实施方式提供的一种流量识别抓包方法,包括:

[0006] 基于预设抓包工具进行流量采集,并建立统计数据包五元组哈希索引模型对数据包进行索引存储;

[0007] 对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件;

[0008] 基于支持向量机对所述特征向量文件行进识别,并提取出人为操作属性的数据包流量进行存储。

[0009] 进一步地,所述基于预设抓包工具进行流量采集,并建立统计数据包五元组哈希索引模型对数据包进行索引存储包括:

[0010] 基于抓包网卡采集流量并对数据包的源IP地址、目的IP地址、源端口、目的端口和协议类型五元组进行流量统计,并基于端口号建立统计数据包五元组哈希索引模型。

[0011] 进一步地,所述哈希索引模型中的每个匹配流至少包括:上行流量、下行流量、数据包个数、源和目标IP个数、平均流量以及数据包平均间隔。

[0012] 进一步地,所述对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件包括:

[0013] 基于DNS检测技术对非加密流量进行分析,确定流量对应的公司属性。

[0014] 进一步地,所述对数据包的流量进行深度检测,并基于预设检测技术对加密和非加密流量进行分析,确定公司属性和操作属性形成数字序列的特征向量文件还包括:

[0015] 基于深度检测技术对加密流量进行分析,确定流量对应的公司属性。

[0016] 进一步地,所述对数据包的流量进行深度检测,并基于预设检测技术对加密和非

加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件还包括：

[0017] 对数据流量包进行关键特征提取确定对应IP地址是否为人为操作属性。

[0018] 进一步地，所述关键特征至少包括：

[0019] 长度、首32字节、尾部32字节、是否HTTP协议和是否含有域名信息。

[0020] 进一步地，所述基于支持向量机对所述特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储包括：

[0021] 基于所述特征向量文件构建高维的样本特征向量数据为X＝[x1,x2,…,xl]；

[0022] 将每个网络流量样本标记为D(X,y_i)，其中y_i为该类流量数据样本标示的类别标签，y_i∈{+1,-1}；

[0023] 区别不同类别的最优分类面表示为w·X+b＝0，w为权重，b为常数项；

[0024] 最优分类面满足：

[0025] $$\min \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

[0026] $y_i(w*X+b)-1\geqslant 0$，i＝1,...n；

[0027] 其中i＝1,…,n，中n表示样本数；

[0028] 基于支持向量机分类判别函数进行判别：

[0029] $$f(x)=\mathrm{sgn}\{w*X+b\}=\mathrm{sgn}\{\sum_{i=1}^{t}\alpha l*y_i*xl+b\}\ ;$$

[0030] 其中α1为优化的Lagrange算子。

[0031] 进一步地，所述流量识别抓包方法还包括：将识别为人为操作的数据包流量以IP地址为标记建立文件进行保存，并在存储容量超过50M时进行分割保存。

[0032] 根据本发明具体实施方式提供的一种流量识别抓包装置，包括：

[0033] 数据抓取模块，用于基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储；

[0034] 深度检测模块，用于对数据包的流量进行深度检测，并基于预设检测技术对加密和非加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件；以及

[0035] 识别模块，用于基于支持向量机对所述特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储。

[0036] 本发明的有益效果为：通过抓取端口的数据包流量，进行哈希索引存储后，对数据包流量进行深度分析得到数据包的公司属性和操作属性形成数字序列的特征向量文件，然后基于特征向量文件所构成的样本和支持向量机进行自动完成对流量的分类，并且显著标明人在操作上网时候的特征，非人为上网如软件升级、后台自动运行等垃圾流量则过滤掉，只有识别为人为上网的通信才保存，这样保存的数据报文流量减少了50％-95％以上，不仅节省了存储空间而且便于进行查看等操作。

附图说明

[0037] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以

根据这些附图获得其他的附图。

[0038] 图1是根据一示例性实施例提供的流量识别抓包方法的流程图；

[0039] 图2是根据一示例性实施例提供的数据抓取的流程图；

[0040] 图3是根据一示例性实施例提供的支持向量机的分类图；

[0041] 图4是根据一示例性实施例提供的流量识别抓包装置的原理图。

具体实施方式

[0042] 为使本发明的目的、技术方案和优点更加清楚，下面将对本发明的技术方案进行详细的描述。显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所得到的所有其它实施方式，都属于本发明所保护的范围。

[0043] 参照图1所示本发明的实施例提供了一种流量识别抓包方法，包括以下步骤：

[0044] 101、基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储；

[0045] 102、对数据包的流量进行深度检测，并基于预设检测技术对加密和非加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件；

[0046] 103、基于支持向量机对特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储。

[0047] 具体的，通过采用支持向量机建立流量特征值库，样本的训练通过对抓包工具抓取的数据包流量进行深度分析确定公司属性和操作属性形成数值序列的特征向量文件，再利用该库对收集回来的流量进行分类或者分析并应用到设备的策略上实现对指定应用的数据传输的记录，最终获得应用层的分类识别检测结果，从而判断IP地址五元组对应的APP应用情况，判断是人为上网还是设备(如路由器/手机/电脑)后台自行发出的流量，如果是人为上网，则写入日志文件。这样保存的数据报文流量减少了50％-95％以上，同时准确率高达99％。

[0048] 作为上述实施例可行的实现方式，参照图2所示，基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储具体的可包括：

[0049] 首先用pfring或者DPDK高性能抓包网卡采集流量，获取到源IP地址、目的IP地址、源端口、目的端口和协议类型五元组进行流量统计，人为上网中超过99.99％是TCP和UDP协议，其他协议完全可以忽略，根据端口号1-65535建立统计数据包五元组hash_index模型，每个匹配流包括：上行流量、下行流量、数据包个数、源和目标IP个数、平均流量、数据包平均间隔等。

[0050] 而对一些常见的应用如DNS、SSH、HTTP、HTTPS、MAIL等协议端口基本是固定的，端口分析主要是针对常见流量，但是随着网络的发展，目前很多应用，尤其是P2P、视频语音协议类型的应用，会使用动态随机端口或者伪装端口的方法，使得端口识别这种方法识别精度变低，甚至失效。针对这种识别方法的缺点，主要使用深度包检测技术和DNS检测技术，尤其是DNS检测技术可以大幅增强流量公司属性的识别的精度。

[0051] 其中可先进行分析DNS和域名组成，如*.weixin.qq.com，首先截获所有的DNS请求报文和对应的DNS应答报文进行DNS解码，获得站点域名对应的IP地址；然后把后续的网络

流量分流出源地址或目的地址根据IP打上标签,比如凡是还有weixin.qq.com域名对应的IP,我们可以先标记为微信相关应用,流量对应的公司属性则为腾讯公司。

[0052] 而人为上网正常情况下TCP端口的80和443占90%左右的流量,而且随着HTTPS的发展,443数据加密越来越多,所以需要深度检测数据包的流量来识别加密流量。具体的识别方法为:

[0053] 根据SSL协议版本、SSL返回的数字证书(内有公司域名、公司名称、签名机构),这样和DNS一样,可以把绝大部分的IP地址打上公司标签,如腾讯公司,由于一个公司可能使用*.qq.com这种域名,光靠数字证书还不能准确识别到某种应用。这时结合关键数据提取,在每个IP在特定的位置提取n份特征,包括包长度、首32字节、尾部32字节、是否HTTP协议、是否含有域名信息等然后通过关键特征提取,用以区分IP对应的属性如微信后台数据或者微信人为操作属性,再把属性翻译为数字序列的特征向量文件,最后通交给引擎程序识别为二进制序列。

[0054] 主要机器学习算法采用决策树、随机森林和支持向量机。随机森林用随机方式建立一个森林,里面也是很多决策树组成,而每一颗决策树之间是没有关联的原因是决策树训练时间复杂度低、预测过程非常快捷,非常符合网络流量的模型特征。

[0055] 电子取证家庭环境的单IP环境决策树一般就满足一般包括:

[0056] 样本采集:从不同的终端如手机、电脑、平板等,安装N个常见应用软件,这N个软件的流量基本覆盖了互联网99.9%的人为流量,人工操作这些APP,建立数据流量模型。

[0057] 决策树:将N个样本用来训练一个决策树,作为决策树根节点处的样本;

[0058] 每个样本中,可以进行分析的属性有X个,比如域名、所属公司、平均报文长度、流持续时间、最大报文长度、最小报文长度、平均短报文长度、平均长报文长度、平均数量等X个属性,从这X个属性中选取x个(x<X)属性。然后从这x个属性中选取一个属性,并采用信息增益策略作为该节点分裂属性在随机森林形成过程中,对上述决策树分裂到不能分裂为止,并建立大量决策树,即可得到数据特征模型。

[0059] 特殊大流量有不同维度的统计特征,难以通过决策树等直观规则实现映射。SVM方法建立在统计学习理论的基础上,具备很强的认知能力,尤其是对于小样本学习问题,可以通过统计学习掌握潜在非规则描述性规律,实现多维特征联合映射,具体的:

[0060] 经过流量特征选择后,单位时间内获得网络流量基本特征及统计特征共计1个维度,构建高维的样本特征向量数据为X=[x1,x2,…,x1],每个网络流量样本可以标记为D(X,y_i),其中yi为该类流量数据样本标示的类别标签,$y_i \in \{+1,-1\}$。区别不同类别的最优分类面可以表示为w·X+b=0,w为权重,b为常数项。最优分类面可以使得不同类别分类间隔最大。获取最优分类面等价为:

[0061] $$\min \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

[0062] $y_i(w*X+b)-1 \geq 0, i=1,...n$;

[0063] 其中i=1,…,n,中n表示样本数;w不仅与样本的位置有关,还与样本的类别有关。对于本发明所涉及的两分类问题,SVM分类判别函数可表示为:

[0064] $$f(x)=\mathrm{sgn}\{w*X+b\}=\mathrm{sgn}\{\sum_{i=1}^{t} \alpha l * y_i * xl + b\}$$;

[0065] 参照图3所示的流量分类实现过程图，其中α1为优化的Lagrange算子，w和b确定分类面方程w·X+b=0。对于多分类问题，通过在任意两个类别间设计SVM辨别模型，对于k个类别，就需要设计Ck2个SVM分类，对于未知类别的待分类样本，获得票数最多的类别即为该样本的类别如微信聊天或者微信自动升级。

[0066] 最后基于已经学习好的流量特征规则向量，获得基于SVM的网络流量检测分类模型。网络流量数据经过特征提取后，利用分类模型进行分类，最终获得应用层的分类识别检测结果。从而判断IP地址五元组对应的APP应用情况，判断是人为上网还是设备(如路由器/手机/电脑)后台自行发出的流量，如果是人为上网，则写入日志文件。

[0067] 在本发明的一些具体实施例中，可把最终存储的报文用自定义的格式存为文件，一天每个IP地址一个文件，超过50M则分割保存，只有识别为人为上网的通信才保存，这样保存的数据报文流量减少了50％-95％以上。

[0068] 参照图4所示，基于同样的设计思路本发明的另一些实施例还提供了一种流量识别抓包装置，包括：

[0069] 数据抓取模块，用于基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储；

[0070] 深度检测模块，用于对数据包的流量进行深度检测，并基于预设检测技术对加密和非加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件；以及

[0071] 识别模块，用于基于支持向量机对特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储。

[0072] 上述流量识别抓包装置的具体实现方式可参见流量识别抓包方法的具体实施例，本发明在此不再赘述。

[0073] 本发明实施例所提供的流量识别抓包方法和装置，通过采用半监督学习识别流量，即样本训练由在识别前事先完成，设备实际部署后就不需要学习了，也就是我们把事先学习好的一些流量特征规则植入设备类，找出每一个应用如微信/QQ/Facebook等关键的数据包，由机器自动完成对流量的分类，并且显著标明人在操作上网时候的特征。非人为上网如软件升级、后台自动运行等垃圾流量则过滤掉，只对人为上网的通信进行保存，从而使数据报文流量减少了50％-95％以上，节约了硬盘空间提高了利用效率。

[0074] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0075] 此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

[0076] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何

的一个或多个实施例或示例中以合适的方式结合。

[0077] 上文的描述包括一个或多个实施例的举例。当然，为了描述上述实施例而描述部件或方法的所有可能的结合是不可能的，但是本领域普通技术人员应该认识到，各个实施例可以做进一步的组合和排列。因此，本文中描述的实施例旨在涵盖落入所附权利要求书的保护范围内的所有这样的改变、修改和变型。此外，就说明书或权利要求书中使用的术语“包含”，该词的涵盖方式类似于术语“包括”，就如同“包括”在权利要求中用作衔接词所解释的那样。此外，使用在权利要求书的说明书中的任何一个术语“或者”是要表示“非排它性的或者”。

[0078] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

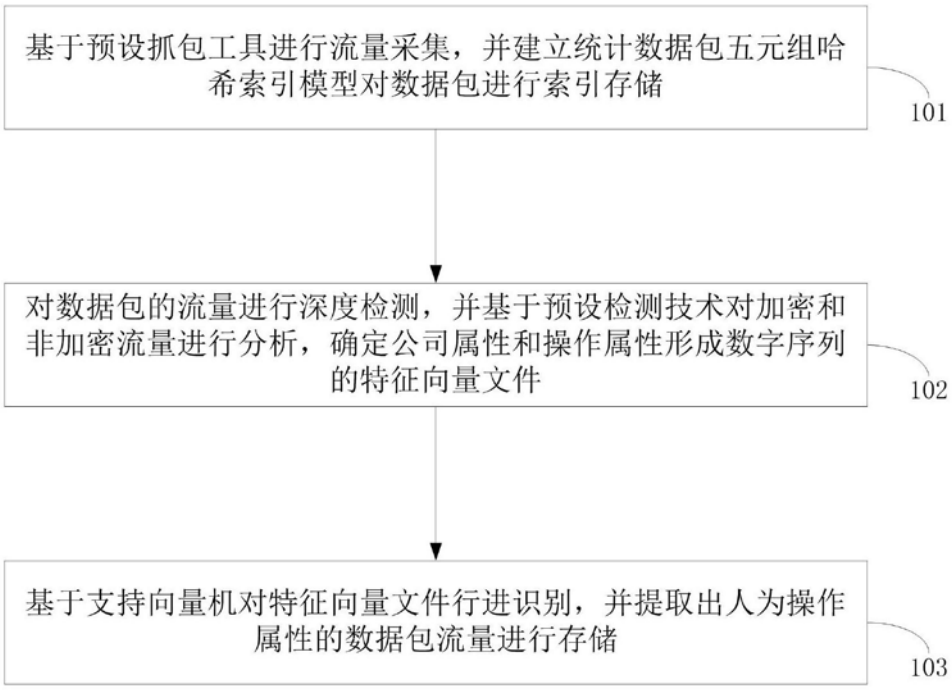

基于预设抓包工具进行流量采集，并建立统计数据包五元组哈希索引模型对数据包进行索引存储
101
对数据包的流量进行深度检测，并基于预设检测技术对加密和非加密流量进行分析，确定公司属性和操作属性形成数字序列的特征向量文件
102
基于支持向量机对特征向量文件行进识别，并提取出人为操作属性的数据包流量进行存储
103

图1

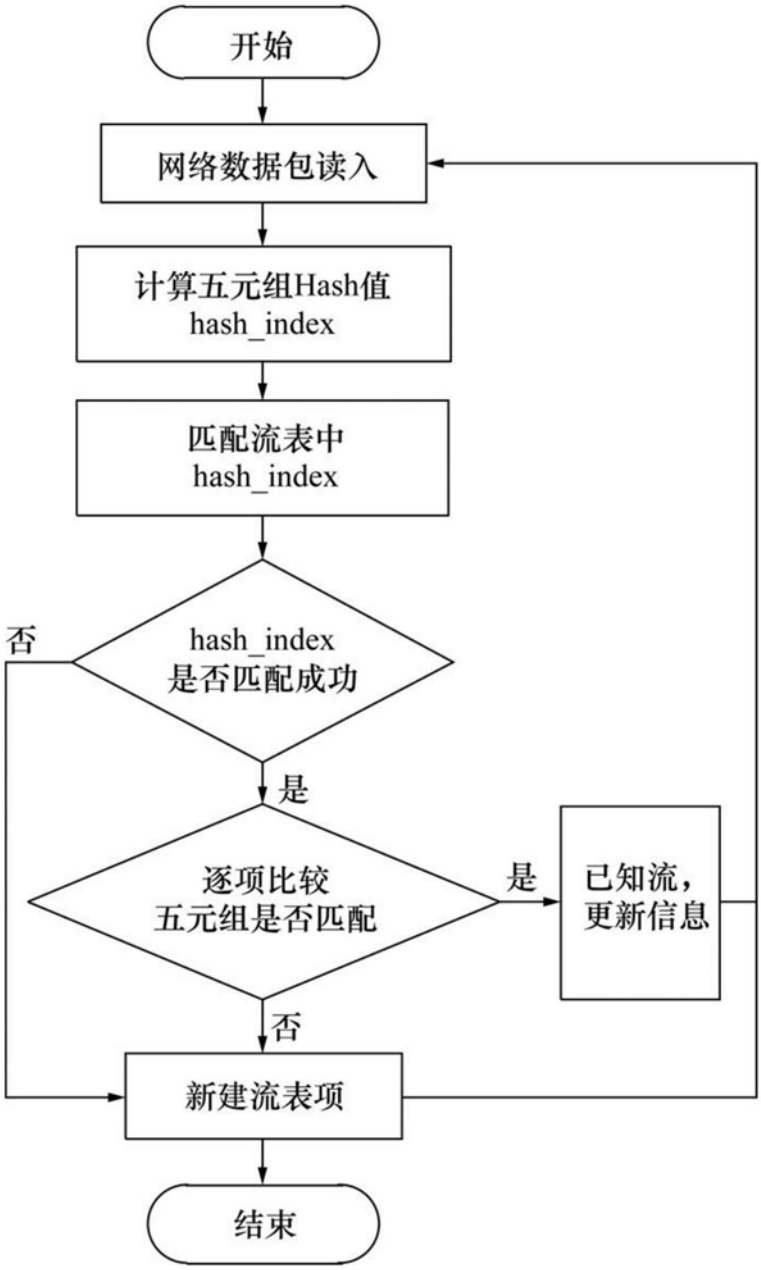

开始
网络数据包读入
计算五元组Hash值
hash_index
匹配流表中
hash_index
hash_index
是否匹配成功
否
是
逐项比较
五元组是否匹配
是
已知流，
更新信息
否
新建流表项
结束

图2

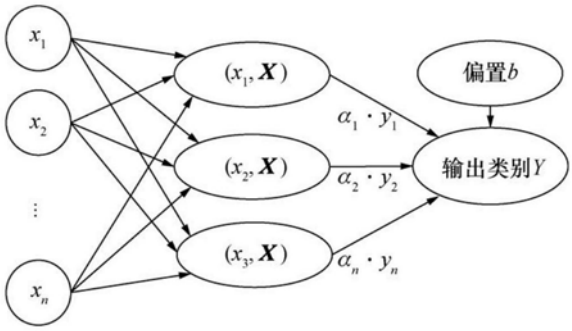

x_1
x_2
x_n
$(x_1, \boldsymbol{X})$
$(x_2, \boldsymbol{X})$
$(x_3, \boldsymbol{X})$
偏置b
$\alpha_1 \cdot y_1$
$\alpha_2 \cdot y_2$
$\alpha_n \cdot y_n$
输出类别Y

图3

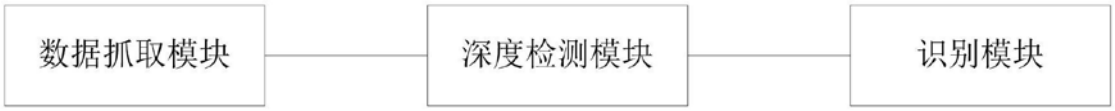

数据抓取模块
深度检测模块
识别模块

图4