(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116015772 A
(43)申请公布日 2023.04.25

(21)申请号 202211590350.9

(22)申请日 2022.12.12

(71)申请人 深圳安巽科技有限公司
地址 518000 广东省深圳市南山区粤海街道滨海社区海天一路6号百度国际大厦东塔楼27层

(72)发明人 王晓伟 马庆贺 高磊 杨真

(74)专利代理机构 深圳市恒程创新知识产权代理有限公司 44542
专利代理师 孔德丞

(51)Int.Cl.
H04L 9/40(2022.01)
G06V 30/10(2022.01)

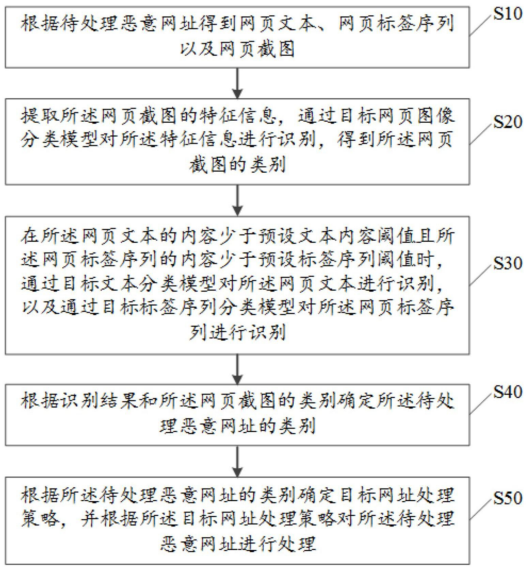
权利要求书3页 说明书9页 附图4页

(54)发明名称

恶意网址的处理方法、装置、设备及存储介质

(57)摘要

本发明涉及数据安全技术领域，公开了一种恶意网址的处理方法、装置、设备及存储介质，所述方法包括：根据待处理恶意网址得到网页文本、网页标签序列以及网页截图；通过目标网页图像分类模型对特征信息进行识别；通过目标文本分类模型对网页文本进行识别，以及通过目标标签序列分类模型对网页标签序列进行识别；根据识别结果和网页截图的类别确定待处理恶意网址的类别；根据待处理恶意网址的类别确定目标网址处理策略，并根据目标网址处理策略对待处理恶意网址进行处理；通过上述方式，根据类别确定的目标网址处理策略对待处理恶意网址进行处理，从而能够有效提高处理恶意网址的效率和准确率。



CN 116015772 A

1.一种恶意网址的处理方法,其特征在于,所述恶意网址的处理方法包括以下步骤:

根据待处理恶意网址得到网页文本、网页标签序列以及网页截图;

提取所述网页截图的特征信息,通过目标网页图像分类模型对所述特征信息进行识别,得到所述网页截图的类别;

在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时,通过目标文本分类模型对所述网页文本进行识别,以及通过目标标签序列分类模型对所述网页标签序列进行识别;

根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别;

根据所述待处理恶意网址的类别确定目标网址处理策略,并根据所述目标网址处理策略对所述待处理恶意网址进行处理。

2.如权利要求1所述的恶意网址的处理方法,其特征在于,所述根据待处理恶意网址得到网页文本、网页标签序列以及网页截图,包括:

获取待处理恶意网址,在虚拟机上通过目标HTTP get命名对所述待处理恶意网址进行访问,得到恶意网址源码和恶意网址内容;

对所述恶意网址源码进行解析,得到网页文本和网页标签数据;

根据所述网页标签数据得到对应的网页标签序列;

通过目标操作浏览器对所述恶意网址内容进行截图,得到网页截图。

3.如权利要求1所述的恶意网址的处理方法,其特征在于,所述提取所述网页截图的特征信息,通过目标网页图像分类模型对所述特征信息进行识别,得到所述网页截图的类别,包括:

对所述网页截图进行检测,得到网页截图形状;

根据预设固定图像形状对所述网页截图形状进行调整;

根据调整形状后的网页截图得到对应的截图像素值;

在截图像素值位于预设像素区间时,对所述截图像素值进行均值计算,得到当前截图像素均值,以及对所述截图像素值进行方差计算,得到当前截图像素方差;

在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时,提取所述网页截图的特征信息,通过目标网页图像分类模型对所述特征信息进行识别,得到所述网页截图的类别。

4.如权利要求3所述的恶意网址的处理方法,其特征在于,所述在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时,提取所述网页截图的特征信息,通过目标网页图像分类模型对所述特征信息进行识别,得到所述网页截图的类别,包括:

在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时,通过RestNet网络对所述网页截图进行特征提取,得到各个尺度的特征信息;

通过MaxPooling网络层对所述各个尺度的特征信息进行融合,得到多尺度特征信息;

通过目标网页图像分类模型的全连接层对所述多尺度特征信息进行识别,得到所述网页截图所属各个类别的概率值;

提取所述各个类别的概率值中的最大概率值,将所述最大概率值对应类别作为所述网页截图的类别。

5.如权利要求1所述的恶意网址的处理方法，其特征在于，所述在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别，包括：

对所述网页文本进行检测，根据网页文本检测结果得到对应的文本内容；

对所述网页标签序列进行检测，根据标签序列检测结果得到对应的标签序列内容；

在所述网页文本的内容少于预设文本内容阈值时，判断所述网页标签序列的内容是否少于预设标签序列阈值；

在所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别。

6.如权利要求5所述的恶意网址的处理方法，其特征在于，所述在所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别，包括：

在所述网页标签序列的内容少于预设标签序列阈值时，对所述网页文本进行词义分析，得到各个词汇；

统计所述各个词汇出现的频率，并将所述频率大于预设频率阈值的词汇从所述各个词汇筛选；

根据筛选得到的频率构建对应的词汇表，并根据所述词汇表构建词嵌入矩阵；

根据所述网页标签序列和所述词嵌入矩阵得到词向量列表；

根据所述词嵌入矩阵查询出与所述网页文本对应的词向量，并通过目标文本分类模型对所述词向量进行识别；

通过全局池化层和权连接层对所述词向量和所述词向量列表进行汇聚，得到目标词向量特征；

通过目标标签序列分类模型对所述目标词向量特征进行识别。

7.如权利要求1至6中任一项所述的恶意网址的处理方法，其特征在于，所述根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理，包括：

根据所述待处理恶意网址的类别在目标恶意网址处理策略集合中选取目标网址处理策略；

对所述待处理恶意网址进行拦截，并获取所述待处理恶意网址的统一资源定位符；

根据所述统一资源定位符的域名信息得到对应的统一资源定位符段；

根据目标网址处理策略在所述统一资源定位符段的预设位置插入阻隔字符，并计算所述统一资源定位符段的哈希值；

将所述统一资源定位符段的哈希值存储至恶意网址区块链。

8.一种恶意网址的处理装置，其特征在于，所述恶意网址的处理装置包括：

获取模块，用于根据待处理恶意网址得到网页文本、网页标签序列以及网页截图；

提取模块，用于提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别；

识别模块，用于在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的

内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别；

确定模块，用于根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别；

处理模块，用于根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理。

9.一种恶意网址的处理设备，其特征在于，所述恶意网址的处理设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的恶意网址的处理程序，所述恶意网址的处理程序配置有实现如权利要求1至7中任一项所述的恶意网址的处理方法。

10.一种存储介质，其特征在于，所述存储介质上存储有恶意网址的处理程序，所述恶意网址的处理程序被处理器执行时实现如权利要求1至7中任一项所述的恶意网址的处理方法。

恶意网址的处理方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及数据安全技术领域，尤其涉及恶意网址的处理方法、装置、设备及存储介质。

背景技术

[0002] 互联网为人们带来便捷的同时也带来了危害，例如，诈骗，且诈骗方式也跟随着的互联网技术不断变化，使得越来越多行接触网络的老年用户或者长期使用网络的青少年因诈骗受到了财产损失和精神损失，而诈骗的途径之一就是通过恶意网址对用户实施诈骗，例如，贷款、刷单、杀猪盘以及公检法诈骗等，仅仅识别出恶意网址是远远不够的，如何处理恶意网址才是重中之重，目前，常用的相关技术是防火墙，具体是通过防火墙拦截恶意网址，禁止对恶意网址的访问，但是恶意网址的制造者会根据防火墙的工作原理制造出新型的恶意网址，使得防火墙的抵御能力急速下降，造成处理恶意网址的效率和准确率较低。

[0003] 上述内容仅用于辅助理解本发明的技术方案，并不代表承认上述内容是现有技术。

发明内容

[0004] 本发明的主要目的在于提供一种恶意网址的处理方法、装置、设备及存储介质，旨在解决现有技术无法处理恶意网址的效率和准确率较低的技术问题。

[0005] 为实现上述目的，本发明提供了一种恶意网址的处理方法，所述恶意网址的处理方法包括以下步骤：

[0006] 根据待处理恶意网址得到网页文本、网页标签序列以及网页截图；

[0007] 提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别；

[0008] 在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别；

[0009] 根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别；

[0010] 根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理。

[0011] 可选地，所述根据待处理恶意网址得到网页文本、网页标签序列以及网页截图，包括：

[0012] 获取待处理恶意网址，在虚拟机上通过目标HTTP get命名对所述待处理恶意网址进行访问，得到恶意网址源码和恶意网址内容；

[0013] 对所述恶意网址源码进行解析，得到网页文本和网页标签数据；

[0014] 根据所述网页标签数据得到对应的网页标签序列；

[0015] 通过目标操作浏览器对所述恶意网址内容进行截图，得到网页截图。

[0016] 可选地，所述提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别，包括：

[0017] 对所述网页截图进行检测，得到网页截图形状；

[0018] 根据预设固定图像形状对所述网页截图形状进行调整；

[0019] 根据调整形状后的网页截图得到对应的截图像素值；

[0020] 在截图像素值位于预设像素区间时，对所述截图像素值进行均值计算，得到当前截图像素均值，以及对所述截图像素值进行方差计算，得到当前截图像素方差；

[0021] 在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时，提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别。

[0022] 可选地，所述在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时，提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别，包括：

[0023] 在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时，通过RestNet网络对所述网页截图进行特征提取，得到各个尺度的特征信息；

[0024] 通过MaxPooling网络层对所述各个尺度的特征信息进行融合，得到多尺度特征信息；

[0025] 通过目标网页图像分类模型的全连接层对所述多尺度特征信息进行识别，得到所述网页截图所属各个类别的概率值；

[0026] 提取所述各个类别的概率值中的最大概率值，将所述最大概率值对应类别作为所述网页截图的类别。

[0027] 可选地，所述在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别，包括：

[0028] 对所述网页文本进行检测，根据网页文本检测结果得到对应的文本内容；

[0029] 对所述网页标签序列进行检测，根据标签序列检测结果得到对应的标签序列内容；

[0030] 在所述网页文本的内容少于预设文本内容阈值时，判断所述网页标签序列的内容是否少于预设标签序列阈值；

[0031] 在所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别。

[0032] 可选地，所述在所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别，包括：

[0033] 在所述网页标签序列的内容少于预设标签序列阈值时，对所述网页文本进行词义分析，得到各个词汇；

[0034] 统计所述各个词汇出现的频率，并将所述频率大于预设频率阈值的词汇从所述各个词汇筛选；

[0035] 根据筛选得到的频率构建对应的词汇表，并根据所述词汇表构建词嵌入矩阵；

[0036] 根据所述网页标签序列和所述词嵌入矩阵得到词向量列表；

[0037] 根据所述词嵌入矩阵查询出与所述网页文本对应的词向量，并通过目标文本分类模型对所述词向量进行识别；

[0038] 通过全局池化层和权连接层对所述词向量和所述词向量列表进行汇聚，得到目标词向量特征；

[0039] 通过目标标签序列分类模型对所述目标词向量特征进行识别。

[0040] 可选地，所述根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理，包括：

[0041] 根据所述待处理恶意网址的类别在目标恶意网址处理策略集合中选取目标网址处理策略；

[0042] 对所述待处理恶意网址进行拦截，并获取所述待处理恶意网址的统一资源定位符；

[0043] 根据所述统一资源定位符的域名信息得到对应的统一资源定位符段；

[0044] 根据目标网址处理策略在所述统一资源定位符段的预设位置插入阻隔字符，并计算所述统一资源定位符段的哈希值；

[0045] 将所述统一资源定位符段的哈希值存储至恶意网址区块链。

[0046] 此外，为实现上述目的，本发明还提出一种恶意网址的处理装置，所述恶意网址的处理装置包括：

[0047] 获取模块，用于根据待处理恶意网址得到网页文本、网页标签序列以及网页截图；

[0048] 提取模块，用于提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别；

[0049] 识别模块，用于在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别；

[0050] 确定模块，用于根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别；

[0051] 处理模块，用于根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理。

[0052] 此外，为实现上述目的，本发明还提出一种恶意网址的处理设备，所述恶意网址的处理设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的恶意网址的处理程序，所述恶意网址的处理程序配置为实现如上文所述的恶意网址的处理方法。

[0053] 此外，为实现上述目的，本发明还提出一种存储介质，所述存储介质上存储有恶意网址的处理程序，所述恶意网址的处理程序被处理器执行时实现如上文所述的恶意网址的处理方法。

[0054] 本发明提出的恶意网址的处理方法，根据待处理恶意网址得到网页文本、网页标签序列以及网页截图；提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别；在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别；根据

识别结果和所述网页截图的类别确定所述待处理恶意网址的类别;根据所述待处理恶意网址的类别确定目标网址处理策略,并根据所述目标网址处理策略对所述待处理恶意网址进行处理;通过上述方式,根据类别确定的目标网址处理策略对待处理恶意网址进行处理,从而能够有效提高处理恶意网址的效率和准确率。

附图说明

[0055] 图1是本发明实施例方案涉及的硬件运行环境的恶意网址的处理设备的结构示意图;

[0056] 图2为本发明恶意网址的处理方法第一实施例的流程示意图;

[0057] 图3为本发明恶意网址的处理方法第二实施例的流程示意图;

[0058] 图4为本发明恶意网址的处理装置第一实施例的功能模块示意图。

[0059] 本发明目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0060] 应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。

[0061] 参照图1,图1为本发明实施例方案涉及的硬件运行环境的恶意网址的处理设备结构示意图。

[0062] 如图1所示,该恶意网址的处理设备可以包括:处理器1001,例如中央处理器(Central Processing Unit,CPU),通信总线1002、用户接口1003,网络接口1004,存储器1005。其中,通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏(Display)、输入单元比如键盘(Keyboard),可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如无线保真(Wireless-Fidelity,Wi-Fi)接口)。存储器1005可以是高速的随机存取存储器(Random Access Memory,RAM)存储器,也可以是稳定的非易失性存储器(Non-Volatile Memory,NVM),例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

[0063] 本领域技术人员可以理解,图1中示出的结构并不构成对恶意网址的处理设备的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0064] 如图1所示,作为一种存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及恶意网址的处理程序。

[0065] 在图1所示的恶意网址的处理设备中,网络接口1004主要用于与网络一体化平台工作站进行数据通信;用户接口1003主要用于与用户进行数据交互;本发明恶意网址的处理设备中的处理器1001、存储器1005可以设置在恶意网址的处理设备中,所述恶意网址的处理设备通过处理器1001调用存储器1005中存储的恶意网址的处理程序,并执行本发明实施例提供的恶意网址的处理方法。

[0066] 基于上述硬件结构,提出本发明恶意网址的处理方法实施例。

[0067] 参照图2,图2为本发明恶意网址的处理方法第一实施例的流程示意图。

[0068] 在第一实施例中,所述恶意网址的处理方法包括以下步骤:

[0069] 步骤S10,根据待处理恶意网址得到网页文本、网页标签序列以及网页截图。

[0070] 需要说明的是,本实施例的执行主体为恶意网址的处理设备,还可为其他可实现

相同或相似功能的设备，例如网址处理器等，本实施例对此不作限制，在本实施例中，以网址处理器为例进行说明。

[0071] 应当理解的是，网页文本指的是在虚拟机上访问待处理恶意网址生成的网页的文本内容，网页标签序列指的是所生成的网页的标签序列，该网页标签可以为网页HTML标签，网页截图指的是网页内容的截图，该网页内容包括但不限于网页文本和网页图片。

[0072] 进一步地，步骤S10，包括：获取待处理恶意网址，在虚拟机上通过目标HTTP get命名对所述待处理恶意网址进行访问，得到恶意网址源码和恶意网址内容；对所述恶意网址源码进行解析，得到网页文本和网页标签数据；根据所述网页标签数据得到对应的网页标签序列；通过目标操作浏览器对所述恶意网址内容进行截图，得到网页截图。

[0073] 可以理解的是，在得到待处理恶意网址后，为了避免恶意网址对设备的攻击和侵入，本实施例在虚拟机通过目标HTTP get命名对待处理恶意网址进行访问，得到恶意网址源码和恶意网址内容，恶意网址源码指的是待处理恶意网址对应的网页的源代码，网页标签数据指的是待处理恶意网址对应的网页的标签数据，该网页标签数据位于源代码的两端，然后根据网页标签数据得到对应的网页标签序列，然后通过目标操作浏览器对恶意网址内容进行截图，得到网页截图，该目标操作浏览器可以为Selenium操作浏览器。

[0074] 步骤S20，提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别。

[0075] 可以理解的是，特征信息指的是能够唯一识别网页截图的信息，该特征信息可以为网页截图标识字段，目标网页图像分类模型指的是对网页图像进行分类的模型，该目标网页图像分类模型是采用ImageNet数据集预训练的模型进行迁移学习到网页截图数据集微调得到的，相较于一般的图像分类模型来说，该目标网页图像分类模型的深度进行了增加并使用跳跃方式将内部的残差块进行连接，从而可以缓解增加深度带来的梯度消失的困扰。

[0076] 进一步地，步骤S20，包括：对所述网页截图进行检测，得到网页截图形状；根据预设固定图像形状对所述网页截图形状进行调整；根据调整形状后的网页截图得到对应的截图像素值；在截图像素值位于预设像素区间时，对所述截图像素值进行均值计算，得到当前截图像素均值，以及对所述截图像素值进行方差计算，得到当前截图像素方差；在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时，提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别。

[0077] 应当理解的是，在得到网页截图形状后，需要将网页截图形状调整成预设固定图像形状，然后判断调整形状后的网页截图的截图像素值是否位于预设像素区间，若是，则需要按照比例将截图像素值缩减至预设像素区间，该预设像素区间为[0,1]，然后分别计算出截图像素值的当前截图像素均值和当前截图像素方差，再判断是否满足当前截图像素均值为预设均值阈值且当前截图像素方差为预设方差阈值的条件，若否，则需要将当前截图像素均值规范化为预设均值阈值，以及将当前截图像素方差规范化为预设方差阈值，该预设均值阈值为0，该预设方差阈值为1，然后继续通过目标网页图像分类模型识别出网页截图的类别。

[0078] 进一步地，所述在所述当前截图像素均值为预设均值阈值且所述当前截图像素方

差为预设方差阈值时，提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别，包括：在所述当前截图像素均值为预设均值阈值且所述当前截图像素方差为预设方差阈值时，通过RestNet网络对所述网页截图进行特征提取，得到各个尺度的特征信息；通过MaxPooling网络层对所述各个尺度的特征信息进行融合，得到多尺度特征信息；通过目标网页图像分类模型的全连接层对所述多尺度特征信息进行识别，得到所述网页截图所属各个类别的概率值；提取所述各个类别的概率值中的最大概率值，将所述最大概率值对应类别作为所述网页截图的类别。

[0079] 可以理解的是，在得到满足条件的网页截图后，通过RestNet网络进行特征提取，该RestNet网络包括不同尺度的网络层，参考图3，该RestNet网络包括但不限于(7×7conv,64,/2)、(3×3conv,64)、(3×3conv,128,/2)、(3×3conv,128)、(3×3conv,256,/2)、(3×3conv,256)、(3×3conv,512,/2)、(3×3conv,512)，因此，特征提取到各个尺度下的特征信息，然后通过MaxPooling网络层将各个尺度的特征信息融合成多尺度特征信息，然后通过目标网页图像分类模型的全连接层对多尺度特征信息进行识别，并输出网页截图所属各个类别的概率值，然后将各个类别的概率值中的最大概率值对应的类别作为网页截图的类别，例如，类别1的概率值为60％，类别2的概率值为80％，类别3的概率值为95％，则将类别3作为网页截图的类别。

[0080] 步骤S30，在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别。

[0081] 应当理解的是，在得到网页文本后和网页标签序列后，需要判断是否满足网页文本的内容少于预设文本内容阈值且网页标签序列的内容少于预设标签序列阈值，若是，则表明网页文本的内容和网页标签序列的内容过少，此时通过目标文本分类模型对网页文本进行识别，通过目标标签序列分类模型对网页标签序列进行识别，目标文本分类模型和目标标签序列分类模型均是通过TextCNN深度学习算法训练得到的，该目标文本分类模型训练采用的文本类别有色情、博彩、贷款、刷单、ETC诈骗、仿冒公检法以及正常合法。

[0082] 步骤S40，根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别。

[0083] 可以理解的是，在得到网页文本和网页标签序列的识别结果后，结合网页截图的类别综合考虑并确定待处理恶意网址的类别。

[0084] 步骤S50，根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理。

[0085] 应当理解的是，目标网址处理策略指的是处理恶意网址的策略，由于不同类别的恶意网址的处理策略均不相同，因此，在得到待处理恶意网址的类别，根据待处理恶意网址的类别确定最契合的网址处理策略，然后通过目标网址处理策略对待处理恶意网址进行处理。

[0086] 进一步地，步骤S50，包括：根据所述待处理恶意网址的类别在目标恶意网址处理策略集合中选取目标网址处理策略；对所述待处理恶意网址进行拦截，并获取所述待处理恶意网址的统一资源定位符；根据所述统一资源定位符的域名信息得到对应的统一资源定位符段；根据目标网址处理策略在所述统一资源定位符段的预设位置插入阻隔字符，并计算所述统一资源定位符段的哈希值；将所述统一资源定位符段的哈希值存储至恶意网址区

块链。

[0087] 可以理解的是，在选取最契合待处理恶意网址的类别的目标网址处理策略后，然后对待处理恶意网址进行拦截，即不会访问待处理恶意网址，然后根据待处理恶意网址的统一资源定位符的域名信息得到对应的统一资源定位符段，然后在统一资源定位符段的预设位置插入阻隔字符，使得整个待处理恶意网址处理无效状态，然后将统一资源定位符段的哈希值存储至恶意网址区块链，在其他用户遇到该待处理恶意网址时，就会自动弹出恶意标签，以避免其他用户的设备受到待处理恶意网址的危害。

[0088] 本实施例根据待处理恶意网址得到网页文本、网页标签序列以及网页截图；提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别；在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别；根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别；根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理；通过上述方式，根据类别确定的目标网址处理策略对待处理恶意网址进行处理，从而能够有效提高处理恶意网址的效率和准确率。

[0089] 在一实施例中，如图3所述，基于第一实施例提出本发明恶意网址的处理方法第二实施例，所述步骤S30，包括：

[0090] 步骤S301，对所述网页文本进行检测，根据网页文本检测结果得到对应的文本内容。

[0091] 应当理解的是，文本内容指的是网页文本的内容，该文本内容包括但不限于网页文本和网页图片，具体是在得到网页文本后，对网页文本进行检测，以得到对应的文本内容。

[0092] 步骤S302，对所述网页标签序列进行检测，根据标签序列检测结果得到对应的标签序列内容。

[0093] 可以理解的是，标签序列内容指的是网页标签序列的内容，具体是在得到网页标签序列后，对网页标签序列进行检测，以得到对应的标签序列内容。

[0094] 步骤S303，在所述网页文本的内容少于预设文本内容阈值时，判断所述网页标签序列的内容是否少于预设标签序列阈值。

[0095] 应当理解的是，在得到网页文本的内容后，需要判断网页文本的内容是否少于预设文本内容阈值，若是，则需要继续判断网页标签序列的内容是否少于预设标签序列阈值。

[0096] 步骤S304，在所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别。

[0097] 可以理解的是，在判定网页标签序列的内容少于预设标签序列阈值时，表明网页标签序列和网页文本的内容过少，此时通过目标文本分类模型对网页文本进行识别，通过目标标签序列分类模型对网页标签序列进行识别。

[0098] 进一步地，步骤S304，包括：在所述网页标签序列的内容少于预设标签序列阈值时，对所述网页文本进行词义分析，得到各个词汇；统计所述各个词汇出现的频率，并将所

述频率大于预设频率阈值的词汇从所述各个词汇筛选;根据筛选得到的频率构建对应的词汇表,并根据所述词汇表构建词嵌入矩阵;根据所述网页标签序列和所述词嵌入矩阵得到词向量列表;根据所述词嵌入矩阵查询出与所述网页文本对应的词向量,并通过目标文本分类模型对所述词向量进行识别;通过全局池化层和权连接层对所述词向量和所述词向量列表进行汇聚,得到目标词向量特征;通过目标标签序列分类模型对所述目标词向量特征进行识别。

[0099] 应当理解的是,在判定网页标签序列的内容少于预设标签序列阈值时,将网页文本分为词粒度,然后根据词粒度得到各个词汇,然后统计各个词汇出现的频率,再判断统计出的频率是否大于预设频率阈值,若是,则将频率对应的词汇构建对应的词汇表,然后根据词汇表构建词嵌入矩阵,此时的词嵌入矩阵可以通过任一词查询到该词对应的词向量,通过该词向量可以表征该词各个维度的特征,然后通过目标文本分类模型对词向量进行识别,并通过全局池化层和权连接层对词向量和词向量列表进行汇聚,然后通过目标标签序列分类模型对汇聚得到的目标词向量特征进行识别。

[0100] 本实施例通过对所述网页文本进行检测,根据网页文本检测结果得到对应的文本内容;对所述网页标签序列进行检测,根据标签序列检测结果得到对应的标签序列内容;在所述网页文本的内容少于预设文本内容阈值时,判断所述网页标签序列的内容是否少于预设标签序列阈值;在所述网页标签序列的内容少于预设标签序列阈值时,通过目标文本分类模型对所述网页文本进行识别,以及通过目标标签序列分类模型对所述网页标签序列进行识别;通过上述方式,分别对网页文本和网页标签序列进行检测,然后判断是否满足网页文本的内容少于预设文本内容阈值且网页标签序列的内容少于预设标签序列阈值的条件,若是,则通过目标文本分类模型对网页文本进行识别,以及通过目标标签序列分类模型对网页标签序列进行识别,从而能够有效提高识别网页文本和网页标签序列的准确性。

[0101] 此外,本发明实施例还提出一种存储介质,所述存储介质上存储有恶意网址的处理程序,所述恶意网址的处理程序被处理器执行时实现如上文所述的恶意网址的处理方法的步骤。

[0102] 由于本存储介质采用了上述所有实施例的全部技术方案,因此至少具有上述实施例的技术方案所带来的所有有益效果,在此不再一一赘述。

[0103] 此外,参照图4,本发明实施例还提出一种恶意网址的处理装置,所述恶意网址的处理装置包括:

[0104] 获取模块10,用于根据待处理恶意网址得到网页文本、网页标签序列以及网页截图。

[0105] 提取模块20,用于提取所述网页截图的特征信息,通过目标网页图像分类模型对所述特征信息进行识别,得到所述网页截图的类别。

[0106] 识别模块30,用于在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时,通过目标文本分类模型对所述网页文本进行识别,以及通过目标标签序列分类模型对所述网页标签序列进行识别。

[0107] 确定模块40,用于根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别。

[0108] 处理模块50,用于根据所述待处理恶意网址的类别确定目标网址处理策略,并根

据所述目标网址处理策略对所述待处理恶意网址进行处理。

[0109] 本实施例根据待处理恶意网址得到网页文本、网页标签序列以及网页截图;提取所述网页截图的特征信息,通过目标网页图像分类模型对所述特征信息进行识别,得到所述网页截图的类别;在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时,通过目标文本分类模型对所述网页文本进行识别,以及通过目标标签序列分类模型对所述网页标签序列进行识别;根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别;根据所述待处理恶意网址的类别确定目标网址处理策略,并根据所述目标网址处理策略对所述待处理恶意网址进行处理;通过上述方式,根据类别确定的目标网址处理策略对待处理恶意网址进行处理,从而能够有效提高处理恶意网址的效率和准确率。

[0110] 需要说明的是,以上所描述的工作流程仅仅是示意性的,并不对本发明的保护范围构成限定,在实际应用中,本领域的技术人员可以根据实际的需要选择其中的部分或者全部来实现本实施例方案的目的,此处不做限制。

[0111] 另外,未在本实施例中详尽描述的技术细节,可参见本发明任意实施例所提供的恶意网址的处理方法,此处不再赘述。

[0112] 本发明所述恶意网址的处理装置的其他实施例或具有实现方法可参照上述各方法实施例,此处不再赘余。

[0113] 此外,需要说明的是,在本文中,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

[0114] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0115] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质(如只读存储器(Read Only Memory,ROM)/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,一体化平台工作站,或者网络设备等)执行本发明各个实施例所述的方法。

[0116] 以上仅为本发明的优选实施例,并非因此限制本发明的专利范围,凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本发明的专利保护范围内。

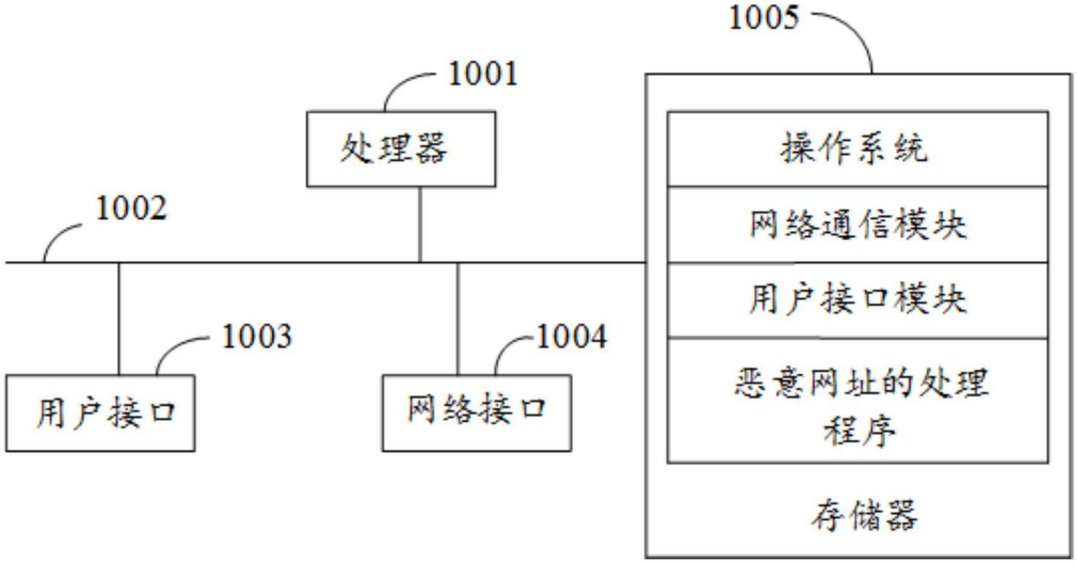

1005
1001
处理器
操作系统
网络通信模块
1002
用户接口模块
1003
1004
恶意网址的处理程序
用户接口
网络接口
存储器

图1

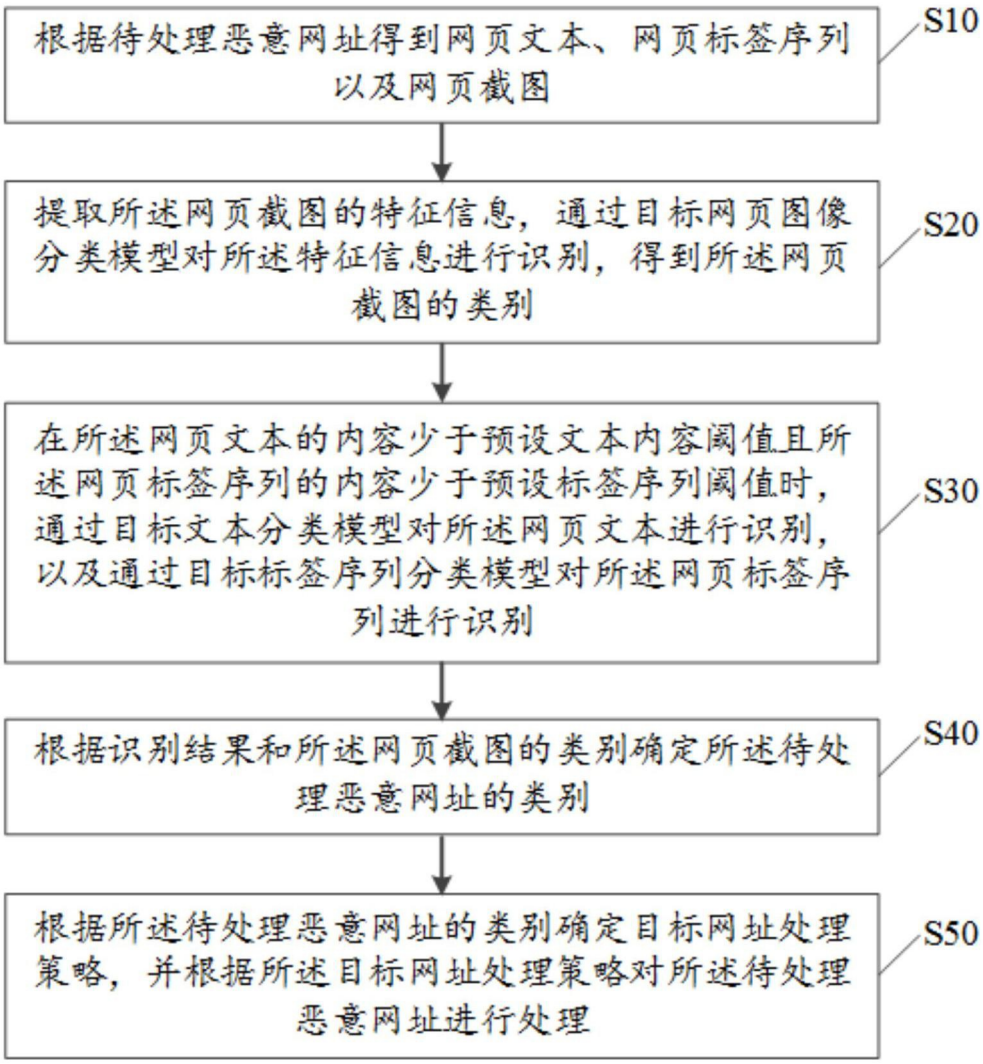

根据待处理恶意网址得到网页文本、网页标签序列以及网页截图
S10
提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别
S20
在所述网页文本的内容少于预设文本内容阈值且所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别
S30
根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别
S40
根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理
S50

图2

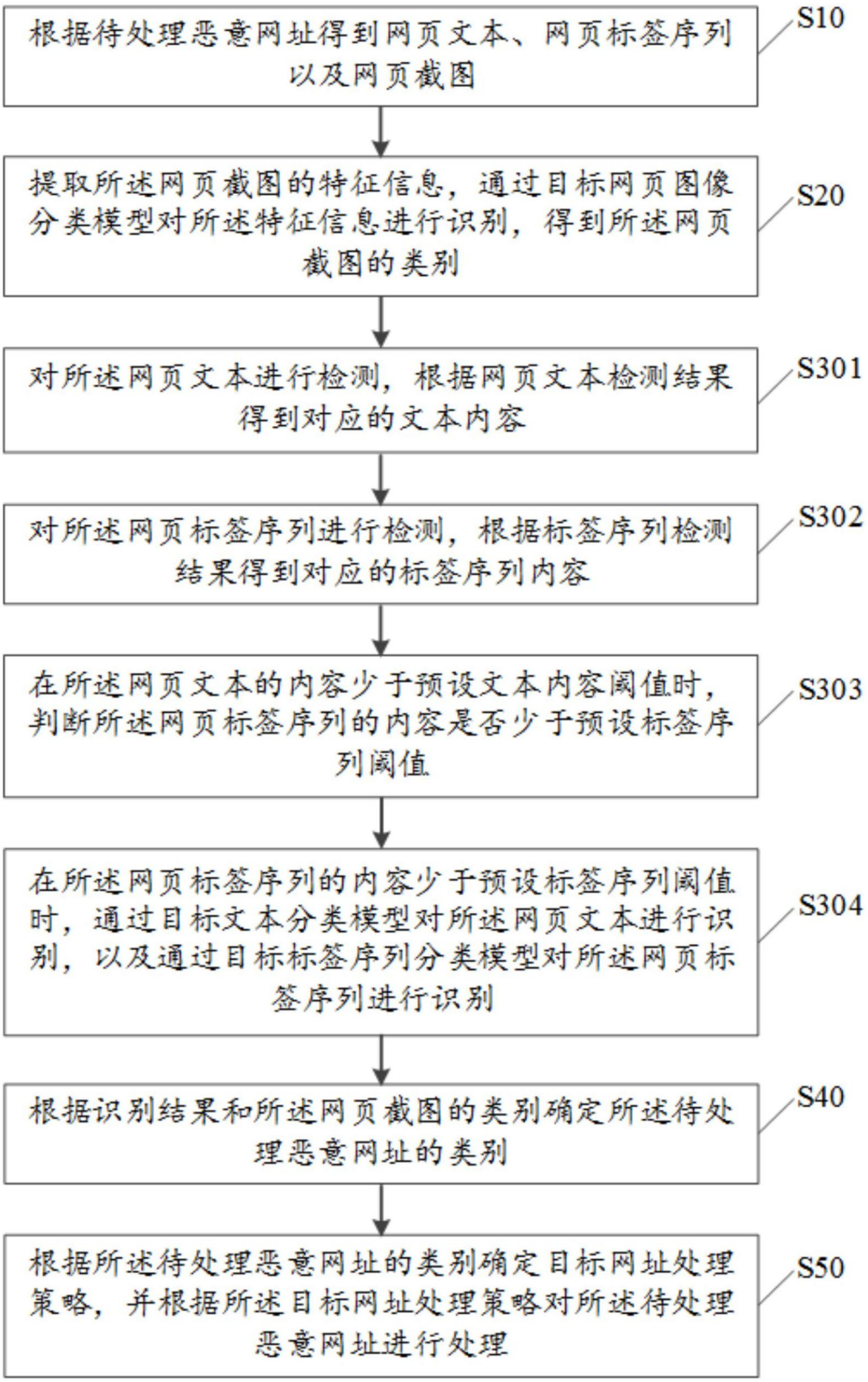

根据待处理恶意网址得到网页文本、网页标签序列以及网页截图
S10
提取所述网页截图的特征信息，通过目标网页图像分类模型对所述特征信息进行识别，得到所述网页截图的类别
S20
对所述网页文本进行检测，根据网页文本检测结果得到对应的文本内容
S301
对所述网页标签序列进行检测，根据标签序列检测结果得到对应的标签序列内容
S302
在所述网页文本的内容少于预设文本内容阈值时，判断所述网页标签序列的内容是否少于预设标签序列阈值
S303
在所述网页标签序列的内容少于预设标签序列阈值时，通过目标文本分类模型对所述网页文本进行识别，以及通过目标标签序列分类模型对所述网页标签序列进行识别
S304
根据识别结果和所述网页截图的类别确定所述待处理恶意网址的类别
S40
根据所述待处理恶意网址的类别确定目标网址处理策略，并根据所述目标网址处理策略对所述待处理恶意网址进行处理
S50

图3

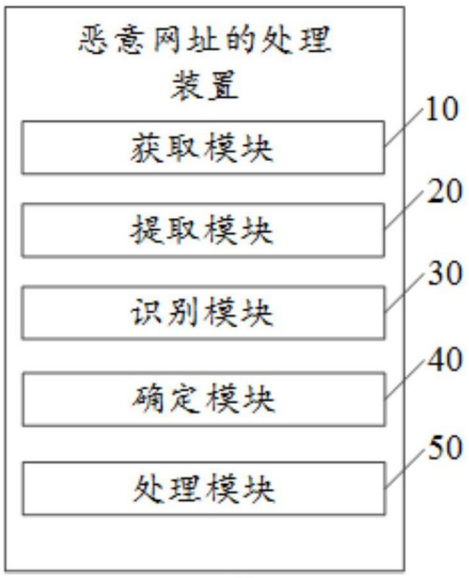


图4