(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 104657504 A
(43)申请公布日 2015.05.27

(21)申请号 201510109944.7

(22)申请日 2015.03.12

(71)申请人 四川神琥科技有限公司
地址 610041 四川省成都市高新区天府大道中段177号19栋1单元1层5号

(72)发明人 罗阳 陈虹宇 王峻岭

(74)专利代理机构 北京天奇智新知识产权代理有限公司 11340
代理人 郭霞

(51)Int.Cl.
G06F 17/30(2006.01)

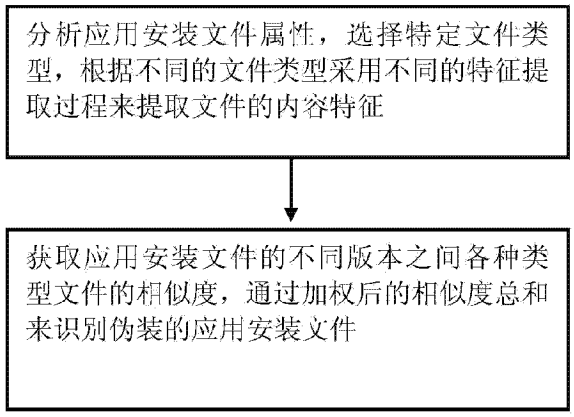
权利要求书2页 说明书7页 附图1页

(54)发明名称

一种文件快速识别方法

(57)摘要

本发明提供了一种文件快速识别方法，该方法包括：分析应用安装文件属性，选择特定文件类型，并根据不同的文件类型采用不同的特征提取过程来提取文件的内容特征，获取应用安装文件的不同版本之间各种类型文件的相似度，通过加权后的相似度总和来识别伪装的应用安装文件。本发明提出了一种文件处理，通过提取应用安装文件内容特征进行识别，且可以有效抵抗文件和目录的伪装和恶意修改带来的干扰，利用特征提取过程缩小文件内容特征规模，提高运算效率。



CN 104657504 A

1. 一种文件快速识别方法，用于识别伪装的应用程序安装文件，其特征在于，包括：

分析应用安装文件属性，选择特定文件类型，并根据不同的文件类型采用不同的特征提取过程来提取文件的内容特征，获取应用安装文件的不同版本之间各种类型文件的相似度，通过加权后的相似度总和来识别伪装的应用安装文件。

2. 根据权利要求 1 所述的方法，其特征在于，所述应用安装文件以压缩文件的形式存在，文件内部以目录的形式组织存放可执行字节码文件、证书文件和资源文件，其中可执行字节码存储在类文件中；证书文件是应用的签名文件；资源文件包括数据库文件、函数库文件、XML 文件、图像文件。

3. 根据权利要求 2 所述的方法，其特征在于，所述特定文件类型文件具备以下条件：文件内容特征具有签名特性，不同应用中提取出的文件内容特征具有差异性，文件内容具有距离特性；并且所述特征提取过程进一步包括，获取安装文件的文件接口，根据压缩文件位置偏移定位特征文件，对应用中的特征文件进行统计，根据统计结果对比不同的算法，对算法进行优化，并且在提取过程中应用多线程方案，重写不支持多线程的部分函数，在特征提取之后，基于文件内容特征进行识别，根据应用的统计特征，采用哈希表计数进行相似度度量。

4. 根据权利要求 3 所述的方法，其特征在于，其中特定类型的特征文件包括界面描述文件、图像文件、音频文件，并将应用程序安装文件描述为 appfile = {image；audio；profile}，每类文件内容特征集合包含了该类所有文件的特征，表示为：

$$image_f = \sum_{i=1}^{n} image_f[i];$$

$$audio_f = \sum_{i=1}^{n} audio_f[i];$$

$$profile_f = \sum_{i=1}^{n} profile_f[i];$$

其中 $image_f$、$audio_f$ 和 $profile_f$ 分别表示图像文件、音频文件、界面描述文件的特征，n 表示每种文件类型包含的文件数量，并且计算文件内容特征相似度函数 com() 的过程表示为：

com(app1，app2) = com(appfile1，appfile2)

对两个应用的每种特征进行对比，获得文件特征相似度计算如下，表示安装文件内文件相似度等价于两个应用安装文件内所有该类型的相似度：

$$com\_image(app1，app2) = \sum_{i=1}^{n} \sum_{j=1}^{m} com(image_{f1}[i]，image_{f2}[j]);$$

$$com\_audio(app1，app2) = \sum_{i=1}^{n} \sum_{j=1}^{m} com(audio_{f1}[i]，audio_{f2}[j]);$$

$$com\_profile(app1，app2) = \sum_{i=1}^{n} \sum_{j=1}^{m} com(profile_{f1}[i]，profile_{f2}[j]);$$

其中 m 表示每种文件类型包含的文件数量；

对三种文件内容相似度赋予权值，通过三种文件内容特征的加权相似度表示应用安装

文件相似度，加权相似度公式表示如下：

com(app1，app2) ＝ com(appfile1，appfile2) ＝

com_image×α+com_audio×β+com_profile×γ ；

其中 α，β，γ 的值根据 com_image，com_audio，com_profile 内容的不同而动态变化，即根据 com_image，com_audio，com_profile 三个值的大小赋予权值，通过学习确定三个最优权值。

一种文件快速识别方法

技术领域

[0001] 本发明涉及文件处理，特别涉及一种应用安装文件的处理方法。

背景技术

[0002] 在移动应用领域中，开发者将应用程序提交给应用市场，用户通过应用市场下载应用。但是官方市场内依然无法避免恶意软件的存在；安全保障机制不够完善，导致恶意软件的比例居高不下。其中，嵌入已知代码和伪装应用安装文件是主要威胁。现有的技术方案采用反编译工具或者动态行为分析工具得到应用行为序列，对行为序列进行预处理得到行为序列特征，通过比较行为序列特征的距离得到应用是否被伪装的量化数据。该方法可以识别应用代码的改变，但是行为序列特征的提取容易受到代码混淆技术的影响，因而在针对实际问题进行分析时具有一定的局限性。

[0003] 因此，针对相关技术中所存在的上述问题，目前尚未提出有效的解决方案。

发明内容

[0004] 为解决上述现有技术所存在的问题，本发明提出了一种文件快速识别方法，包括：

[0005] 分析应用安装文件属性，选择特定文件类型，并根据不同的文件类型采用不同的特征提取过程来提取文件的内容特征，获取应用安装文件的不同版本之间各种类型文件的相似度，通过加权后的相似度总和来识别伪装的应用安装文件。

[0006] 优选地，所述应用安装文件以压缩文件的形式存在，文件内部以目录的形式组织存放可执行字节码文件、证书文件和资源文件，其中可执行字节码存储在类文件中；证书文件是应用的签名文件；资源文件包括数据库文件、函数库文件、XML 文件、图像文件。

[0007] 优选地，所述特定文件类型文件具备以下条件：文件内容特征具有签名特性，不同应用中提取出的文件内容特征具有差异性，文件内容具有距离特性；并且所述特征提取过程进一步包括，获取安装文件的文件接口，根据压缩文件位置偏移定位特征文件，对应用中的特征文件进行统计，根据统计结果对比不同的算法，对算法进行优化，并且在提取过程中应用多线程方案，重写不支持多线程的部分函数，在特征提取之后，基于文件内容特征进行识别，根据应用的统计特征，采用哈希表计数进行相似度度量。

[0008] 本发明相比现有技术，具有以下优点：

[0009] 本发明提出了一种文件处理，通过提取应用安装文件内容特征进行识别，且可以有效抵抗文件和目录的伪装和恶意修改带来的干扰，利用特征提取过程缩小文件内容特征规模，提高运算效率。

附图说明

[0010] 图 1 是根据本发明实施例的文件快速识别方法的流程图。

具体实施方式

[0011] 下文与图示本发明原理的附图一起提供对本发明一个或者多个实施例的详细描述。结合这样的实施例描述本发明，但是本发明不限于任何实施例。本发明的范围仅由权利要求书限定，并且本发明涵盖诸多替代、修改和等同物。在下文描述中阐述诸多具体细节以便提供对本发明的透彻理解。出于示例的目的而提供这些细节，并且无这些具体细节中的一些或者所有细节也可以根据权利要求书实现本发明。

[0012] 图 1 是根据本发明实施例的文件快速识别方法流程图。提出了一种应用程序安装文件的伪装识别方法。通过分析应用安装文件属性，选择文件类型，提取内容特征，并根据文件类型采用不同的内容特征提取算法，对其相似度赋予权值，从而提高应用程序伪装识别的准确性和运算效率。

[0013] 应用安装文件以压缩文件的形式存在，内部以目录的形式组织存放可执行字节码文件、证书文件和资源文件，其中可执行字节码存储在类文件中；证书文件是应用的签名文件；资源文件包括数据库文件、函数库文件、XML 文件、图像文件等。

[0014] 在一个实施例中，将应用安装文件描述为集合 app = {exe；lib；profile；image；audio；etc}，其中 exe 表示安装文件中的可执行字节码，lib 表示程序中的原生代码库，profile 表示用于程序数据存储和布局描述的 XML 文档，image 表示程序中的图像文件，etc 表示程序中的其他文件。根据集合 app 的描述可知：本发明的目标是根据 exe，lib，profile，image 等相关文件的内容特征，执行应用程序安装文件的伪装识别。

[0015] 为了准确、有效地通过文件内容分析安装文件是否被伪装，并符合实际的识别需求，本发明提出的方法着力达到以下三个目标：1) 适应大数据运算，应用市场内的数量大、增长快，能快速处理大量数据的系统框架是适应大数据运算的基础；2) 选择合适的特征文件，安装文件内有上千种文件类型，提取哪些文件的内容直接影响伪装识别的效率和准确性；3) 高效的特征提取和准确的特征算法，提取文件内容特征的速度决定了系统效率，同时准确的特征算法是保证系统能够正确给出判定结果的基本保证。

[0016] 本发明在提取文件内容特征、计算文件相似度的过程中保证提高效率的同时不失运算结果的准确性。

[0017] 首先要求算法针对的目标不能过于复杂，如果针对目标过于复杂，那么需要对这个目标进行缩减，选出其中关键的要素进行对比；其次算法效率高；最后，在构建算法过程的时候，要尽可能对算法的运行环境进行优化，减少算法的中间步骤，削减算法中可能引起大量时间和空间消耗的内容。

[0018] 首先需要选择合适的特征文件，一个应用安装文件中的文件从几百个到几千个不等，如对全部文件的内容进行特征提取，容易造成目标过于复杂、分析效率低下的结果，且容易受到插入无用文件的干扰。因此本发明根据普遍性、代表性和可度量性原则，选择部分合适的文件类型作为特征文件，在最大程度保证特征文件有效表示应用安装文件的情况下缩小特征规模，从而减小运算量。

[0019] 接下来，从安装文件中提取已选定文件的特征，获取安装文件的文件接口，根据压缩文件位置偏移定位特征文件，省去对其他无关文件进行解压的步骤以提高运算效率。首先对应用中的特征文件进行统计，根据统计规律对比不同的算法实现，对算法进行最合适的优化，在保证准确性的前提下采用效率最高的算法，并在提取过程中应用多线程方案，重

写不支持多线程的部分函数，保证所有运算的线程安全性，进一步提高运算效率。

[0020] 最后，基于文件内容特征进行伪装识别，在相似度度量算法设计时，根据应用的统计特征，采用哈希表计数，用空间消耗换取时间优化。

[0021] 通过文件内容特征计算文件相似度，首先要从复杂的文件类型中选择合适的特征文件。合适的特征文件需要具有以下三个特点。大多数安装文件内包含该类型的文件，如果某个文件类型仅在少数应用内存在，则无法通过该类文件内容特征进行相似度比较；文件内容特征具有“签名”特性，可以代表该应用，不同应用中提取出的文件内容特征具有差异性；文件内容具有距离特性，相似文件中的文件内容距离近，反之不同文件中的文件内容距离远。在一个实施例中，选择界面描述文件、图像文件、音频文件作为特征文件，可描述为 appfile = {image ;audio ;profile}，主要思路是计算文件内容特征相似度，以此分析相似度，可用以下公式表示：

[0022] com(app1, app2) = com(appfile1, appfile2)。

[0023] 本发明用这三类文件的内容特征表示安装文件的特征。每类文件内容特征集合包含了此类所有文件的特征，用如下公式表示：

[0024] $$image_f = \sum_{i=1}^{n} image_f[i];$$

[0025] $$audio_f = \sum_{i=1}^{n} audio_f[i];$$

[0026] $$profile_f = \sum_{i=1}^{n} profile_f[i]。$$

[0027] n 表示每种文件类型包含的文件数量，计算图像、音频、界面描述文件的内容相似度，对两个应用的每种特征进行对比，可以推导出文件特征相似度计算公式如下，表示安装文件内文件相似度等价于两个应用安装文件内所有该类型的相似度：

[0028] $$com\_image(app1, app2) = \sum_{i=1}^{n} \sum_{j=1}^{m} com(image_{f1}[i], image_{f2}[j]);$$

[0029] $$com\_audio(app1, app2) = \sum_{i=1}^{n} \sum_{j=1}^{m} com(audio_{f1}[i], audio_{f2}[j]);$$

[0030] $$com\_profile(app1, app2) = \sum_{i=1}^{n} \sum_{j=1}^{m} com(profile_{f1}[i], profile_{f2}[j])。$$

[0031] m 表示每种文件类型包含的文件数量。

[0032] 单独使用图像、音频或者界面描述文件内容特征相似度代表应用安装文件相似度，结果不够理想，如果阈值设置较高会导致漏报；如果阈值设置过低则会导致误报。因此，本发明对图像、音频和界面描述文件内容相似度赋予权值，通过三种文件内容特征的加权相似度表示应用安装文件相似度，加权相似度公式表示如下：

[0033] com(app1, app2) = com(appfile1, appfile2) =

[0034] com_image × α +com_audio × β +com_profile × γ 。

[0035] 上式表示应用 app1 和 app2 的相似度等价于 app1 和 app2 内部文件的相似度，等价于两个安装文件内图像、声音、界面描述文件相似度的加权值。此处 α，β，γ 的值根据

com_image，com_audio，com_profile 的不同而动态变化。

[0036] 图像、音频和界面描述文件在安装文件中的数量不一，而且部分应用不包含音频文件，所以固定的 α，β，γ 无法有效地计算文件相似度。本发明的实施例利用动态权值的方法：即根据 com_image，com_audio，com_profile 三个值的大小赋予权值，通过学习确定三个最合适的权值，分别为 0.6，0.3，0.1，com_image，com_audio，com_profile 中值最大的权值为 0.6，其次权值为 0.3，最小的权值为 0.1。

[0037] 经过以上过程可以获得两个文件的相似度，通过比较相似度和阈值 T 的大小可以判断两个文件是否属于相似应用，即是否为伪装文件。

[0038] 本发明采用文件内容特征代表应用特征，针对不同文件的特点提出具体的特征提取方法以及相似度算法。

[0039] 目前，现有的图像相似度匹配算法需要较大的空间和时间开销，无法应用在大规模计算环境中。而伪装应用安装文件通常采用两种方式影响图像：1) 在原有图像基础上进行修改；2) 改变原图像分辨率。基于这样的考虑，在图像内容特征提取过程中，需要选择一种算法，能够降低修改图像和消除分辨率降低带来的干扰。因此，本发明首先缩小安装文件中的图像尺寸，并将彩色图像转换为灰度图像，计算平均灰度级，根据相似度哈希算法提取图像内容特征，根据图像的亮度和构图为每张图像生成一个字符串作为图像的“指纹”，图像的指纹越相似则表示 2 张图像越相似。提高了准确性的同时降低了运算复杂度。

[0040] 其中缩小图像尺寸是将图像缩小到 K×K 像素，该过程主要用于消除图像分辨率对相似度比较的干扰、去除图像尺寸和图像比例的差异，只保留结构、亮度等基本信息，这里的 K 值一般设为 128。40×40 分辨率的图像在移动应用中出现比例最高。图像内容相似度比较需要计算指纹的汉明距离，即两个指纹字符串对应位置不一样的字符个数，K ＝ 40，则字符串长度为 K×K/8 ＝ 200。本发明对该步骤进行简化，采用字符串是否相等代替汉明距离，代价是相似度结果只能显示两个图像指纹是否一致，无法通过汉明距离识别图像指纹是否相似。

[0041] 安装文件内的界面描述文件以 XML 文件格式存储，因此，界面描述文件内容特征提取等同于 XML 文件内容特征提取。XML 文件相似度比较包括结构相似度和内容相似度 2 方面，将 XML 文件转换为树结构，通过比较树的差异得到 XML 结构差异，通过比较树的节点差异得到 XML 内容差异。

[0042] 在应用中界面描述文件是按照预定规则存储的，在已知规则的情况下，本发明采用了一种简单的结构和内容特征提取方法：首先，根据界面描述文件说明，得到结构名列表；然后，根据结构名列表提取结构特征，过滤界面描述文件内的结构特征和符号信息，得到内容信息；最终对结构和内容信息计算哈希值，得到结构特征值和内容特征。界面描述文件经过处理后得到一个哈希数组，从而将界面描述文件的内容相似度转化为比较哈希数组的相似度。

[0043] 经过对安装文件内的音频文件进行分析发现，伪装应用安装文件包不对音频文件进行大的修改，因此本发明采用文件哈希值作为音频文件特征。计算音频文件哈希值。对于大规模运算来说其哈希空间较小，哈希结果易发生碰撞。因此，本发明提出以下哈希方法，在保证运算速度的情况下大大减小哈希碰撞。输入音频文件流 S，并预设常量字符串 M，计算输入音频文件流 S 的 MD5 哈希值 H1，然后将输入音频文件流 S 与预设常量字符串 M 相

加，并计算相加结果的 MD5 哈希值 H2，对 H1 和 H2 进行求和，得到最终哈希值。通过以上算法得到音频文件的二次哈希值作为音频文件的内容特征。

[0044] 应用安装文件内容特征包含图像内容特征、界面描述文件内容特征和音频文件内容特征。图像内容特征为图像“指纹”集合；一个界面描述文件内容特征为一个哈希集合，应用安装文件内的所有界面描述文件内容特征由多个哈希集合组成；音频文件内容特征为哈希集合。三种文件内容特征集合均可视为字符串集合。本发明选择内容相似度作为标准，其计算方法是：集合 A 和 B 的交集元素在 A 和 B 中较小的集合所占的比例。这种方法可以有效地衡量不同长度的集合之间的相似度。内容相似度用 L(A,B) 表示如下：

[0045] $L(A,B) = |A \cap B|/\min(|A|,|B|)$。

[0046] 由此，特征文件集合相似度计算公式由文件特征相似度公式和内容相似度公式推导，表示文件集合相似度等价于文件集合的内容相似度；文件相似度计算由加权相似度公式推导，表示文件相似度等价于文件集合相似度加权值，即三种特征文件内容相似度的加权值。

[0047] 通过计算特征文件内容相似度得到文件相似度，不受文件目录结构改变的干扰；而选择的相似度计算方法采用两个集合中较小的集合长度作为标准，因此可以有效抵抗插入垃圾文件的干扰。

[0048] 根据本发明的另一方面，还提出了一种移动应用防伪装系统，首先采用报文摘要算法对服务器各文件进行初始化指纹采样，存储于远程安全数据库和本地安全文件中。创建防伪装组件，处理客户端提交的访问请求。分析访问请求，提取访问路径，将应用安装文件指纹与库中指纹作对比后给出应答方案；直接追溯到网页文件，适用于动态和静态页面的站点。采用调用本地页面快照和文件校验对比的方式恢复被伪装的页面文件。

[0049] 进一步地，系统主要用于协调访问请求、伪装识别、站点文件更新和事件告警 4 个动作之间的关系。当系统接收到 Web 访问请求时，调用伪装识别模块对每个 HTTP 请求进行分析，追踪被调用文件及访问路径；采用挂载在安全组件内的防伪装组件计算应用安装文件的数字指纹，将其与安全区域中原始指纹进行对比，判断应用安装文件是否被伪装；若未被伪装，Web 服务器以正常的 HTTP 请求响应用户访问。否则，立即启用应急恢复模块，调用本地页面快照响应用户，之后启用恢复模块调用本地备份替换伪装文件，完成修复。在启用快照技术的情况下，即使页面文件被黑客伪装或重置，服务器也不会将伪装后的页面误传给浏览者，避免造成不良的后果。系统记录伪装日志，以手机短信或电子邮件方式通知管理人员。服务器中各服务器文件在启用防伪装后将被锁定，未经授权将无法更新；经身份认证解锁后可采用 FTP 或 SSL 方式进行更新。本地指纹库、备份文件及快照与远程库文件进行适时同步，以保证数据一致。

[0050] 系统通过提供保护站点文件的完整性、监控及处理 HTTP 访问请求、快速恢复伪装文件、告警和可信发布五大功能来实现防伪装机制。由此，将系统服务端、客户端和发布端 3 个部分。

[0051] (1) 服务端。以数据库为枢纽完成与多个客户端之间的通信。为各客户端提供文件备份、快照存储、站点文件初始化数字指纹的处理、各类日志及伪装告警信息的保存。服务端在工作过程中仅开放与客户端及数据库通信的端口，以最大限度提供系统的安全性；此结构将为系统移植打下基础。

[0052] (2) 客户端。在不改变原有网络拓扑结构的情况下安装于受保护的服务器中，与服务端和发布端建立可信通信。客户端包括初始化、被访文件监控与追踪、站点目录锁、伪装识别、伪装恢复和本地资源备份六大功能，是整个防伪装系统的核心部分。

[0053] 首次启用时，将对服务器受保护站点文件进行初始化，采集各文件的数字指纹，存储于服务端的安全数据库中，并将其备份存储于本地指纹文件中；为保证本地文件的安全，对本地数字指纹和备份文件及快照采用对称密钥加密法 AES 进行加密处理。当收到发布端的更新命令后，解锁其保护目录，对被更新文件的数字指纹进行更新。防伪装模块在处理客户页面访问请求时，根据被访页面文件名及访问路径计算其指纹并与安全区域中的指纹对比，如果一致则响应；否则启用伪装恢复和事件告警模块，执行事后处理过程，并记录访问请求的源 IP、源端口及目的端口号、伪装进程 ID、修改内容，构建警告信息通知管理人员。为快速响应客户端请求，恢复模块先读取页面快照缓解用户访问；再取本地备份文件解密后替换伪装文件，当本地备份文件被破坏时，将从服务端下发备份文件进行恢复，以处理灾难性事件。

[0054] (3) 发布端。主要完成新服务器的发布及原有服务器文件的更新。发布端通过客户端认证后，客户端根据请求创建新的站点或解锁被请求站点，完成发布指令；结束时客户端重新锁定站点。

[0055] 系统初始化过程包括各客户端提交服务器特征信息后，在服务端服务器指定区域中建立以特征信息命名的目录。为了保证其唯一性，特征信息采用客户机的 IP 地址、CPU ID、硬盘 ID 所组成字符串的哈希值来表示。服务端建立统一的数据库，存储各站点文件指纹、日志和告警信息。客户端完成与服务端连接信息验证后配置本地运行环境，指定当前服务器中需要保护的站点及不同站点的文件类型，以站点名称和创建时间的密文为文件名创建本地安全目录，用来存储备份及快照文件，存储指纹数据、日志和告警数据的 XML 文档。

[0056] 文件预处理组件首先调用加密引擎，采用公钥加密算法 RSA 生成一对密钥；公钥采用 AES 加密后存于备份服务器中，再将加密后的公钥副本和私钥一起保存到本地安全目录，并及时与备份服务器交换公钥，为数据同步及站点恢复提供工作环境。然后读取需保护的站点目录及文件类型后调用遍历引擎，遍历规定后缀的服务器文件，采用 MD5 算法计算其唯一的、不可逆的数字指纹，将指纹结果按一定的数据结构存入数据库中，再按站点名称生成 XML 文档存入本地安全文件中，供防伪装组件对比使用。最后将遍历完的站点数据采用客户端的公钥加密，存于本地安全目录中等待与备份服务器进行同步。整个

[0057] 伪装识别过程利用安全组件开发防伪装组件，对 HTTP 请求进行分析，通过 HTTP 客户端提交的数据，提取访问路径及文件名，对其完整性进行实时监控，文件发生变化的合法性进行验证。采用核心内嵌技术开发防伪装组件，并借助安全组件接口在映射表中建立映射关系，被服务器加载到进程空间中，完成对每个被访页面指纹的计算和原始指纹对比工作。

[0058] 服务器收到 HTTP 请求时，首先对请求应用安装文件进行追踪，再计算应用安装文件的哈希值，最后调用指纹对比组件；读取本地安全区域中应用安装文件的原始指纹解密后与当前计算指纹进行对比，如果匹配则应答 HTTP 请求，否则进入恢复处理及应急响应流程。应急响应组件在接收到对比失败指令后，系统调用本地同名快照文件解密后生成 HTML 格式文本响应 HTTP 请求，保证服务器 HTTP 请求应答的效率和质量；响应完后以最快的速度

调用本地安全区域中原始文件经 AES 解密后对伪装文件进行替换恢复，从最大程度上为文件提供安全的保障；如果恢复失败则禁用当前文件，将请求重定向到指定页面。在文件恢复的同时，系统记录伪装日志，通过手机短信或电子邮件的方式为管理人员发送告警信息，为日后的数据分析及管理提供依据。快照调用与访问重定向过程所需时间在数毫秒内发生，请求者无法收到被伪装页面的响应内容。客户端按一定的周期时长调用文件底层过滤驱动模块对当前服务器受保护的站点及规定类型的文件的数字指纹进行计算、对比、识别，以确保各处数字指纹的相似度。

[0059] 综上所述，本发明提出了一种文件处理，通过提取应用安装文件内容特征进行识别，且可以有效抵抗文件和目录的伪装和恶意修改带来的干扰，利用特征提取过程缩小文件内容特征规模，提高运算效率。

[0060] 显然，本领域的技术人员应该理解，上述的本发明的各模块或各步骤可以用通用的计算系统来实现，它们可以集中在单个的计算系统上，或者分布在多个计算系统所组成的网络上，可选地，它们可以用计算系统可执行的程序代码来实现，从而，可以将它们存储在存储系统中由计算系统来执行。这样，本发明不限制于任何特定的硬件和软件结合。

[0061] 应当理解的是，本发明的上述具体实施方式仅仅用于示例性说明或解释本发明的原理，而不构成对本发明的限制。因此，在不偏离本发明的精神和范围的情况下所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。此外，本发明所附权利要求旨在涵盖落入所附权利要求范围和边界、或者这种范围和边界的等同形式内的全部变化和修改例。

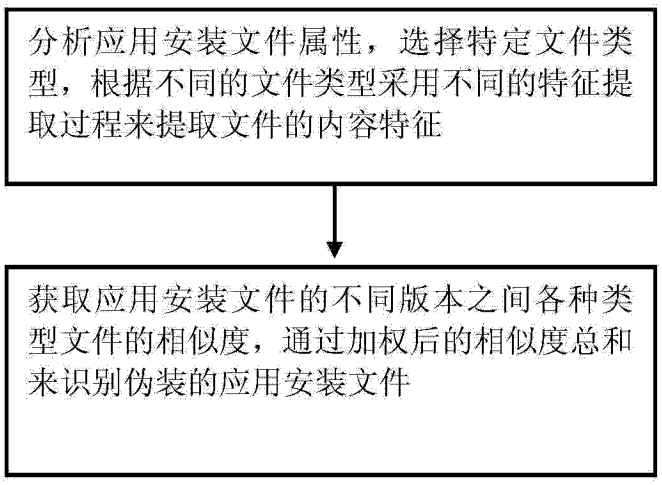

分析应用安装文件属性，选择特定文件类型，根据不同的文件类型采用不同的特征提取过程来提取文件的内容特征
获取应用安装文件的不同版本之间各种类型文件的相似度，通过加权后的相似度总和来识别伪装的应用安装文件

图 1