(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 102918533 B
(45)授权公告日 2016.08.03

(21)申请号 201180027177.5
(22)申请日 2011.05.17
(30)优先权数据
12/791,305 2010.06.01 US
(85)PCT国际申请进入国家阶段日
2012.11.30
(86)PCT国际申请的申请数据
PCT/US2011/036733 2011.05.17
(87)PCT国际申请的公布数据
WO2011/152987 EN 2011.12.08
(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州
(72)发明人 R·比索 V·伊斯梅洛夫 L·刘 R·萨科内 M·贝赫
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 蔡悦
(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/00(2006.01)
G06F 15/16(2006.01)
(56)对比文件
US 2005131904 A1,2005.06.16,
US 2004010703 A1,2004.01.15,
US 2006095964 A1,2006.05.04,
US 2008250100 A1,2008.10.09,
US 2008027867 A1,2008.01.31,
US 2009043870 A1,2009.02.12,
审查员 魏晶瑶

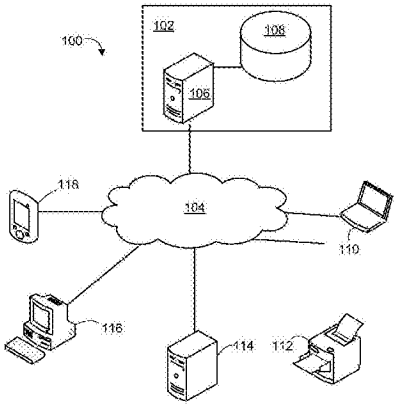
权利要求书2页 说明书10页 附图5页

(54)发明名称

基于声明的内容名誉服务

(57)摘要

在一些实施例中,数据库可存储用于先前评估的数据项的多个内容声明,所述多个内容声明中的每一个都在数据库中与对应存储的先前评估的数据项的数字指纹相关联。一个或多个服务器可被配置为从另一网络节点的客户机设备处接收所确定的数据项的数字指纹,将所确定的数字指纹用作主关键词以向数据库提交查询,并且传输由查询返回至客户机设备的一个或多个内容声明。在一些实施例中,服务器还可被配置为从另一网络节点上的一个或多个计算机处接收内容声明以及与其相关联的数字指纹,以及使得接收到的内容声明以及与其相关联的数字指纹被存储在数据库中。



CN 102918533 B

1.一种用于标识数据项的内容声明的方法,包括:

接收数据项的多个内容声明,所述数据项的多个内容声明包括由针对恶意内容对所述数据项进行了扫描的多个不同发布者所提交的内容声明,其中所述数据项的多个内容声明中的每一个内容声明标识提交所述内容声明的发布者、指示所述发布者所执行的扫描的时间和结果、并且随由所述发布者使用散列函数来计算的数据项的数字指纹一起提交;

将所述数据项的多个内容声明存储在数据库中,所述数据库将所述数据项的多个内容声明中的每一个内容声明与随所述内容声明一起提交的所述数据项的数字签名相关联;

将所述数据项的多个内容声明中的由不同发布者提交的具有相同的数据项的数字指纹的多个内容声明分为一组,其中所述相同的数据项的数字指纹是由所述不同发布者使用散列函数计算的;

从计算设备处接收检索与某一数据相关联的现有内容声明的请求,所述请求包括所确定的该数据的数字指纹,所确定的该数据的数字指纹是由所述计算设备使用所述散列函数计算的;

响应于所述请求,在确定以下的情形时,检索包括由不同发布者提交的所述数据项的多个内容声明中的所述多个内容声明的内容声明集:

所确定的该数据的数字指纹与和所述多个内容声明中的每一个相关联的数据项的相同的数字指纹相匹配,以及

所述多个内容声明中的每一个是在预定时间段内发布的;以及

将所述内容声明集返回给所述计算设备以允许所述计算设备确定是否针对恶意内容来对该数据进行扫描。

2.如权利要求1所述的方法,其特征在于,还包括下列步骤:

接收针对所述数据项的新的内容声明;以及

依据与和所述新的内容声明一起提交的数据项的数字指纹相关联的数据项的任何存储的内容声明来评估所述新的内容声明。

3.如权利要求2所述的方法,其特征在于,还包括:

验证所述数据项的多个内容声明中的每一个内容声明是由受信任的发布者提交的。

4.如权利要求1所述的方法,其特征在于,还包括:

向发送了所述请求的所述计算设备发送用来传送所述内容声明集的消息。

5.如权利要求1所述的方法,其特征在于,还包括:

接收所述数据项的多个内容声明中的每一个内容声明的内容声明的内容证书,其中每个内容证书是由受信任的证书机构签署的并且标识了用于创建所述内容声明的产品。

6.如权利要求1所述的方法,其特征在于,所述内容声明集包括:

指示第一发布者执行的反病毒扫描的结果的内容声明;以及

指示第一发布者执行的恶意软件扫描的结果的内容声明。

7.一种用于标识数据项的内容声明的方法的系统,包括:

用于接收数据项的多个内容声明的装置,所述数据项的多个内容声明包括由针对恶意内容对所述数据项进行了扫描的多个不同发布者所提交的内容声明,其中所述数据项的多个内容声明中的每一个内容声明标识提交所述内容声明的发布者、指示所述发布者所执行的扫描的时间和结果、并且随由所述发布者使用散列函数来计算的数据项的数字指纹一起

提交；

用于将所述数据项的多个内容声明存储在数据库中的装置，所述数据库将所述数据项的多个内容声明中的每一个内容声明与随所述内容声明一起提交的所述数据项的数字签名相关联；

用于将所述数据项的多个内容声明中的由不同发布者提交的具有相同的数据项的数字指纹的多个内容声明分为一组的装置，其中所述相同的数据项的数字指纹是由所述不同发布者使用散列函数计算的；

用于从计算设备处接收检索与某一数据相关联的现有内容声明的请求的装置，所述请求包括所确定的该数据的数字指纹，所确定的该数据的数字指纹是由所述计算设备使用所述散列函数计算的；

用于响应于所述请求，在确定以下的情形时，检索包括由不同发布者提交的所述数据项的多个内容声明中的所述多个内容声明的内容声明集的装置：

所确定的该数据的数字指纹与和所述多个内容声明中的每一个相关联的数据项的相同的数字指纹相匹配，以及

所述多个内容声明中的每一个是在预定时间段内发布的；以及

用于将所述内容声明集返回给所述计算设备以允许所述计算设备确定是否针对恶意内容来对该数据进行扫描的装置。

基于声明的内容名誉服务

背景技术

[0001] 为了确保数字数据遵从商业、安全以及其他策略，近些年的趋势是使这些数据满足不断增长的预访问评估过程的数目。这些过程的示例包括健康扫描、过滤、分类以及数据分析。计算上极其密集的操作可包括，例如病毒/间谍软件扫描、垃圾邮件检测、关键词检测、恶意/不正当/禁止的URL检测、预防数据泄露、数据分类，等等。

[0002] 需要对内容进行的扫描/分类技术的数量已经随着时间持续增长。此外，需要扫描的典型内容的大小持续上升，并且没有显示出趋于平稳的迹象。这两个趋势都造成了需要用来执行扫描/分类的计算机资源(CPU、存储器、网络带宽，等等)的量的持续增长。

[0003] 数据在计算机网络中或者跨计算机网络移动时通常需要被各安全和合规性产品重复地重新分析、重新扫描和/或重新分类的事实造成该问题的进一步恶化。这些产品通常被安装在台式机、笔记本、不同的服务器(例如邮件、文件、协作等)、以及云中的服务。随着数据经过每一个这些路线点，通常会一次又一次地执行相同的计算密集的操作。这导致系统性能和吞吐量的降低，并且需要安装额外的硬件、软件，等等。在服务的情况下，额外的开销将对服务的收益造成直接的影响。

发明内容

[0004] 在本发明的一些实施例中，系统可包括数据库以及一个或多个服务器。例如，数据库可存储用于先前评估的数据项的多个内容声明，所述多个内容声明中的每一个都在数据库中与对应存储的先前评估的数据项的数字指纹相关联。例如，服务器可被配置为从另一网络节点上的客户机设备处接收所确定的数据项的数字指纹，将所确定的数字指纹用作主关键字以向数据库提交查询，以及传输由查询返回至客户机设备的一个或多个内容声明。

[0005] 在一些实施例中，服务器还可被配置为从另一网络节点上的一个或多个计算机处接收内容声明以及与其相关联的数字指纹，以及使得接收到的内容声明以及与其相关联的数字指纹被存储在数据库中。

[0006] 在一些实施例中，客户机设备可包括一个或多个计算机，所述计算机被配置为用散列函数处理数据项以确定该数据项的指纹，将包括所确定的该数据项的数字指纹的第一消息发送给服务器，从服务器接收包括由查询返回的内容声明的第二消息，以及基于第二消息中包括的内容声明来对如何进一步处理该数据项做出决策。

[0007] 在一些实施例中，用于标识数据项的一个或多个内容声明的方法包括将该数据项的数字指纹与存储的与内容声明相关联的数字指纹比较。如果所确定的数字指纹与存储的数字指纹匹配，则确定一个或多个内容声明与所确定的该数据项的数字指纹相关联。

[0008] 在一些实施例中，用指令来编码一个或多个计算机可读存储介质，该指令在由第一网络节点的一个或多个处理器执行时，使得处理器执行用于标识数据项的一个或多个内容声明的方法，该方法包括以下步骤：(a)将所确定的该数据项的数字指纹和与存储的该内容声明相关联的数字指纹比较，以及(b)如果所确定的数字指纹与存储的数字指纹匹配，则确定一个或多个内容声明与所确定的该数据项的数字指纹相关联。

[0009] 在一些实施例中，可从另一网络节点处的一个或多个计算机接收内容声明以及与其相关联的数字指纹，并且可持久地存储所接收的内容声明以及与其相关联的数字指纹。

[0010] 在一些实施例中，可从另一网络节点处的一个或多个计算机接收所确定的数字指纹，并且被确定为与所确定的该数据项的数字指纹相关联的内容声明可被传输到其他网络节点处的一个或多个计算机。

[0011] 在一些实施例中，可从另一网络节点处的一个或多个计算机接收同时包括存储的数字指纹和内容声明的内容证书。

[0012] 除了前面的说明性实施例之外，或者替代前面的说明性实施例，本发明的若干实施例可另外地或另选择呈现或实现一个或多个下列特性、特征和/或功能：

[0013] • 内容健康、过滤、分类和其他内容分析过程的结果可被表示为一组内容声明。

[0014] • 指纹可用作用于将内容声明与经处理的数据相关联的非侵入的和可靠的机制。

[0015] • 内容名誉服务可接受、聚集以及存储由参与的信任实体提交的内容声明。

[0016] • 可向内容名誉服务查询与特定的某一数据相关联的声明。

[0017] • 当发布这些声明时使用的配置和/或安全策略改变时，可以使现有的内容声明无效。

[0018] • 当预定的时间段过去之后，可以使时间敏感的声明无效并且将其从名誉服务数据库中移除。

[0019] • 内容名誉服务可独立于被保护的数据格式。例如，在计算指纹时，数据可被简单地当做字节流，而不是具体格式的数据，从而系统不需要对数据格式的认知。

[0020] • 内容名誉服务可独立于用于传输数据的传输协议。

[0021] • 内容名誉服务可独立于数据存储的存储类型。

[0022] • 可使用不同的网络协议与内容名誉服务通信(为了提交和请求声明的目的)。

[0023] • 使用内容名誉服务可确保为文件(或其他数字内容)创建的声明与相同文件的所有其他副本相关联。

[0024] • 输出内容声明集以及将其串行化至安全和认证的“内容证书”中可使其被传输到因为某些原因不能直接与名誉服务通信(或者完全不知道它)的感兴趣的实体。例如，系统可利用基于已知和被接受的标准的技术来格式化和保护这一内容证书。例如，客户机应用能够将内容证书与关于其发布的数据相关联并且读取与这些数据有关的一个或多个声明。

附图说明

[0025] 附图不旨在按比例绘制。在附图中，各个附图形中示出的每一相同或近乎完全相同的组件由同样的附图标记来表示。出于简明的目的，不是每个组件在每张附图中均被标号。在附图中：

[0026] 图1是示出其中各客户机可访问内容名誉服务的网络的说明性示例的示图；

[0027] 图2示出客户机可执行以向内容名誉服务提交一个或多个声明的例程的说明性示例；

[0028] 图3示出客户机可执行以获取由内容名誉服务维护的一个或多个现有声明的例程的说明性示例；

[0029] 图4是对包括用于图像的实际声明数据的声明集的概念表示；

[0030] 图5示出可由内容名誉服务的服务器执行以处理客户机提交声明的请求的例程的说明性示例；

[0031] 图6示出可由内容名誉服务的服务器执行以处理客户机检索和返回内容声明的请求的例程的说明性示例；以及

[0032] 图7示出图4所示的声明集如何在被输出为内容证书时出现的说明性示例。

具体实施方式

[0033] 我们意识到由于在一个计算机上运行的应用或服务不能利用由其他应用(在相同计算机或一个或多个不同计算机上运行)在相同内容上产生的结果，因此会发生由现有系统执行的冗余扫描。反病毒应用是一个好例子。在现有系统中，在组织中移动的文件(以及可能是其相同的副本)通常需要像其在不同的计算机和服务器(电子邮件、文件、协作，等等)之间移动一样被重复地扫描。我们还意识到，例如通过提供与相同应用或服务的所有实例以及可以利用这些结果的其他感兴趣的实体共享以前的扫描和分类结果的安全方法，可避免这样的重复扫描和分类。

[0034] 在本发明的一些实施例中，以将声明与生成该声明所针对的数据相关联的方式，针对数字内容的扫描、分类或任何其他操作的结果可作为一组内容声明被保存在可被感兴趣的实体访问的集中式储存库中。例如，在一些实施例中，这可通过使用可通过网络访问的集中式内容名誉服务来实现。这一机制可允许完全或者至少部分地避免未来对未修改的(或复制的)的重新扫描和/或重新分类。在一些实施例中，可将结果(声明集)与和声明有关的数据分开存储，并且该过程不需要对数据本身的任何修改。这一方案可因而确保任何已发布的声明的完整性和真实性。

[0035] 在一些实施例中，在基于内容的分析/检查/扫描期间执行的基于各种类型内容的健康和/或过滤技术的结果可用作一组内容声明。例如，信任的服务和应用可将它们的操作结果提交给内容名誉服务以用于和可用作稍后访问这些结果的标识符一起存储。例如，在一些实施例中，经评估的内容的数字指纹可用作这一标识符。任何参与的实体随后可通过计算主题数据的数字指纹(或以其它方式确定标识符)并且将其作为请求的一部分发送给内容名誉服务来请求给定的数字内容的现有声明。内容名誉服务随后可向请求者返回与该数据相关联的声明(如果存在)。例如，内容名誉服务可在把标识符用作关键词的关系数据库中存储声明。

[0036] 这一技术可因而允许在将来为了各种目的而重复使用内容声明。一个此类目的可以是当数据经过网络上的计算机而没有改变时避免对数据的重复分析/检查/扫描。另一个此类目的可以是使数据的客户能够使用内容声明集来做出可验证的信任决策。另一个目的可以是减少用于跨网络的健康、过滤、分类和其他内容检查的计算机资源的消耗。此外，在一些实施例中，驻留在内容名誉服务的数据库中的对内容声明的分析可提供与数据使用、数据的地理位置迁移、感染的源头等有关的可行的统计信息。

[0037] 图1示出其中内容名誉服务102可通过网络104访问的系统100的说明性实施例。如图1所示，内容名誉服务102可例如包括一个或多个服务器106(或者其他计算设备)，这些服务器具有足够的计算能力来处理来自系统100中的需要访问服务102的所有其他设备的请

求。取决于内容名誉服务102部署在其中的系统100的操作需求和规模，内容名誉服务102的计算设备可采用众多形式中的任何一种，并且本发明不需要使用任何特定类型或配置的计算设备。对于大规模实现，可例如采用传统的缩放和/或负载平衡技术在服务102内部署的各服务器和/或其他计算资源之间分发对请求的处理。

[0038] 数据库108可采用众多形式和配置中的任一种，并且本发明不限于使用任何特定类型的数据库。例如，在一些实施例中，数据库108可以是存储和访问与使用关键词的特定数字内容相关联的一个或多个内容声明的关系数据库。在一些实施例中，这些关键词可例如包括数字内容的数字指纹，服务为该数字内容维护一个或多个内容声明。取决于性能和数据库大小的考虑，可使用各种表格和外部关键词以提高性能。例如，在一些实施例中，表格可将指纹映射到内部的身份关键词，并且每个这种内部身份关键词可用于访问与其相关联的一个或多个内容声明。但是，应当明白在其他实施例中，数据库108可包括能够将标识符(例如，指纹)与对应的存储的内容声明相关联的任何其他数据库架构或存储机制。在一些实施例中，甚至可在例如硬盘驱动器或RAM的地址索引的存储设备中存储内容声明，并且表格可用来将指纹(或其他标识符)映射到对应的内容声明的存储器地址。如此处所使用的，术语“数据库”旨在包括所有这些存储架构。

[0039] 网络104可以是众多网络、或网络组中的任一个，并且本发明不限于使用任何特定类型的配置的网络。例如，包括104可包括诸如在公司环境内使用的局域网和/或诸如因特网之类的广域网。可在各实施例中使用任何网络架构和/或通信协议。但是合适的网络和协议的一些示例包括以太网、令牌环、TCP/IP、HTTP、SOAP、REST、RPC、XML-PRC，等等。

[0040] 如图1所示，除了内容名誉服务102自己之外，系统100可另外地包括能够通过网络104与服务102通信的各种客户机。示出了可能是服务的客户机的计算设备的若干示例(例如，由于这些客户机能够向服务102提交内容声明和/或能够访问由服务维护的内容声明)。但是，应当明白可使用那些示出的设备类型之外或者替代它们的设备类型，并且本发明不旨在限于使用任何特定类型的设备。在图1的示例中，示出的客户机包括膝上型计算机110、打印机/传真机/扫描仪112、服务器114、台式计算机116、以及手持式计算设备(例如，PDA、智能电话，等等)118。

[0041] 如此处所使用的，“网络节点”指的是在网络上具有或共享唯一的地址或地址组件的设备或设备组。在一些情况下，给定的网络节点可包括一个或多个子节点。在这一情况下，网络地址的一个组件可唯一标识网络上的节点，并且该地址的另一组件可唯一标识每个子节点。在图1的示例中，内容名誉服务102、膝上型计算机110、打印机/传真机/扫描仪112、服务器114、台式计算机116、以及手持式计算设备118中的每一个都位于网络104的不同节点。

[0042] 例如，在一些实施例中，内容名誉服务102可接受、聚集、存储、以及在请求时提供与数字内容(文件或任何其他类型的数据)有关的声明。此外，在一些实施例中，可采取步骤以确保只有受信任的实体被允许向服务102提交声明。在这些实施例中，由未知或不信任的客户机提交的声明将不被接受。在一些实施例中，对哪些用户被允许查找现有声明没有限制。

[0043] 如以上所述，在一些实施例中，声明可通过数字指纹与数据相关联。可用众多方式中的任何一种来完成对数字指纹的计算，并且本发明不限于任何特定的指纹技术。例如，在

一些实施例中，可使用加密的散列函数来计算指纹这一实现可为生成的指纹提供好的统一性，并且取决于所使用的散列函数，可显著地最小化碰撞（意外的或有意的）的概率。由于加密的散列是单向函数，由散列值推断出原始的内容（或者甚至其特性）是不可能的。合适的散列函数的示例有SHA1、SHA-256和SHA-512。但是，应当明白，可在某些实施例中另外地或另选地使用其他指纹技术。例如，对于数据安全不是问题的应用，可另外地或另选地使用非加密的散列函数。

[0044] 通常可用最小的计算代价来可靠地确定数字指纹，并且相同的数据段将总是产生相同的数字指纹。因此，在使用数字指纹作为声明标识符的实施例中，对数据的任何修改都将产生不同的指纹并且将自动地打破所有现有声明与该文件修改的副本之间的关联。

[0045] 图2和图3示出客户机在与内容名誉服务102交互时可以如何使用指纹的示例。具体地，图2示出客户机（例如图1所示的计算设备110、112、114、116和118中的一个）可执行以向内容名誉服务102提交一个或多个声明的例程的说明性示例，而图3示出客户机可执行以获取由服务102维护的一个或多个现有声明的例程的示例。例如，所示例程可使用存储在计算机可读介质中的指令来实现，该计算机可读介质可由客户机机器的处理器访问和执行。

[0046] 如图2所示，在处理数据之后（例如，执行病毒或恶意软件扫描、分类数据，等等）（步骤202），客户机可创建表达被执行的处理的结果的一个或多个声明（步骤204）。在创建一个或多个声明之后，客户机可计算该数据的指纹（步骤206），然后例如通过网络104向服务102发送消息，将所述一个或多个声明与计算的数据指纹一起提交给内容名誉服务102（步骤208）。如以下进一步描述的，在一些实施例中，这一消息可包括由受信任的认证机构签署的客户机端的数字证书，以标识出用于创建一个或多个声明的产品。例如，这一技术可帮助防止有意危害数据库108的可能的攻击。

[0047] 如图3所示，在提交资源以评估具体数据的内容（例如，通过执行病毒或恶意软件扫描、分类数据，等等）之前，客户机可计算该数据的指纹（步骤302），并且向内容名誉服务102提交指纹以作为对现有声明的一部分请求（步骤304）。如果名誉服务102标识出与所提交的指纹相关联的任何声明，则可检索和返回那些声明给客户机（步骤306）。然后客户机可对是否和/或如何基于任何返回的声明中所包含的信息来评估数据做出进一步的决策（308）。在一些实施例中，由内容名誉服务102接收的消息可携带有效的、可验证的证书，从而允许客户机确认他们正在与受信任的源通信。此外，在一些实施例中，还可数字地签署返回的声明从而进一步提高系统的安全和可靠性。

[0048] 在一些实施例中，无限数量的内容声明可与给定的数据相关联。尽管这些声明可由不同的受信任的发布者创建，但是当它们被发布至相同的数据段（相同的数据段产生相同的指纹值）时，它们都可被数据指纹组合。

[0049] 在某些实现中，当客户机向内容名誉服务102提出返回关于数字内容的声明的请求时，客户机可请求所有现有的声明，或者通过指定其感兴趣的声明的类型（例如，发布者、声明发布的时间、内容断言，等等）来缩小返回的组的范围。

[0050] 如以上所述，在一些实现中，任何对数据的修改将导致不同的计算的指纹。因此，任何对文件的修改将自动地使该文件与所有先前发布的声明分离。

[0051] 表1（下表）示出单个内容声明中可包含的性质/属性的说明性示例。

[0052]

属性	*描述*
声明类型	声明的类型。可以有多个预定义的内容声明类型。内容名誉服务的消费者还可定义自己的声明类型。
声明时间	发布声明的日期和时间。
断言	关于内容的断言。可根据声明类型来解释断言的含义。 例如，如果声明类型是“病毒”并且断言是“不存在的属性”，则可被解释为表示内容不含病毒。同样地，声明类型“恶意的 URL”结合断言“存在的属性”则可被解释为表示内容含有恶意的 URL。 除了预定义的断言以外，声明发布者可定义自己的断言。
发布者	发布这一声明的实体。例如，该实体可标识具体的应用、服务、用户，等等。
数据	发布者可附加在声明后的可选的自定义数据。名誉服务不需要尝试解释这一数据，它可简单地存储这一数据并且在被请求时将数据与声明一起返回。例如，反病毒应用可提交病毒引擎或签名版本以及它做出的每个病毒声明。采用这一方式，当它接收返回的声明时，由于刚接收的声明已被发布，可确定病毒引擎或签名被更新。它从而可以决定仍必须要扫描文件，尽管现有的声明指出文件是不含病毒的，并且然后提交新的声明。

[0053]

[0054] 表1

[0055] 如以上所指出的，在一些实施例中，由不同实体发布的多个声明可存在于内容名誉服务102的数据库108中。当被请求时，这些声明可作为“声明集”被返回。

[0056] 表2(下表)示出如何安排这一声明集的格式的说明性示例。

[0057]

声明集

指纹算法
指纹值
声明1
声明2
声明…
声明N

[0058] 表2

[0059] 图4是对包括实际声明数据(在这一例子中,用于图像)的声明集的概念表示。

[0060] 在一些实施例中,取决于用来与内容名誉服务102通信的协议,可用不同的格式返回不同的声明集。例如,客户机可使用SOAP消息(web服务)或者支持基于连接或分组/消息的安全的其他网络协议(例如,REST、HTTP、RPC、XML-PRC等)以与服务通信。在一些实施例中,服务的实现还可支持多个绑定和/或能够在同一时间使用不同的协议通信。内容名誉服务可被安装在屋内、云中,或两者。

[0061] 如以上所述,在一些实施例中,为了防止数据库中毒以及遭到其他类型的攻击,仅允许受信任的应用提交内容声明。例如,出于此目的,受信任的机制可使用广泛使用的工业标准,诸如WS-信任以及服务器和客户机端证书等。

[0062] 在一些实施例中,不管用来与内容名誉服务102通信的协议如何,所有内容声明可与用作主关键词的指纹一起被存储在集中式的关系数据库108中。

[0063] 图5和图6示出内容名誉服务102如何可以处理从客户机处接收的提交和检索内容声明的请求。具体地,图5示出内容名誉服务102的一个或多个服务器106可与数据库108一起执行以处理提交一个或多个声明的客户机请求的例程的说明性示例,而图6示出这些服务器可与数据库一起执行以处理检索和返回一个或多个内容声明的客户机请求的例程的示例。例如,可使用在一种或多种计算机可读介质中存储的指令来实现所示的例程,该计算机可读介质可由一个或多个服务器106的处理器和/或数据库108的控制器访问和执行。

[0064] 如图5所示,在接收到提交一个或多个声明给服务102的客户机请求后(步骤502),如果相同的声明(例如,相同的声明类型和断言)还不在数据库108中(见步骤503),则服务器106可对接收到的声明格式化并且使用接收到的作为用于查询数据库条目的主关键词的指纹来将声明写入数据库108(步骤504)。如以上所述,在一些实施例中,服务器106可拒绝接受不是由受信任的证书机构签名或者无法标识用于创建声明的产品的任何新声明提交。

[0065] 在一些实施例中,如果对提交的数字指纹有相同的声明被发现已经在数据库108中(见步骤503),则内容名誉服务102可依据现有声明的内容评估新声明的内容并且基于该评估来更新声明中的部分或者全部信息(步骤505)。

[0066] 例如,内容名誉服务102可更新现有声明的信息的场景的一个示例是,新提交的声明包含比与相同的数字指纹相关联的现有声明的病毒签名版本更新的病毒签名版本。例如,这一场景可发生在由于客户机具有比现有声明所反映的病毒签名版本更新的病毒签名版本,客户机决定尽管存在文件的现有声明但还是扫描文件的时候。例如,在使用更新的病毒签名版本执行扫描之后,这一场景中的客户机可向内容名誉服务102提交病毒声明(反映曾用于扫描的病毒签名版本)。

[0067] 例如,当内容名誉服务102接收到这一声明时,它可确定具有相同断言的相同类型

(并且甚至可能来自相同的发布者)的声明已经存在。例如,内容名誉服务102然后可比较病毒签名版本以及各个声明的创建日期和时间,并且用其确定为声明的最新和最可靠的信息来更新数据库108中的条目(例如,数据库列)。例如,在更新病毒签名版本的示例中,现有声明的更新的条目可包括病毒扫描的日期和时间以及用于扫描的病毒签名版本。

[0068] 如图6所示,在接收到检索现有声明的客户机请求后(步骤602),服务器106可使用作为主关键词被包括在客户机请求中的指纹来制定数据库查询。如果数据库108中存在匹配的指纹(步骤606),则服务器106可检索与指纹相关联的声明(步骤608),根据客户机指定的任何准则来过滤那些声明(步骤609),以及返回经过滤的声明数据(如果不使用过滤或者客户机没有指定过滤准则,则可以为检索的声明数据的子集或整个声明集)至请求的客户机(步骤610)。如果匹配的指纹不在数据库中(步骤606),则服务器106可通知客户机没有发现匹配的声明(步骤612)。内容名誉服务102从而可通过将数据的经计算的指纹(从客户机处接收)与存储在数据库中的与被查询的数据的声明相关联的指纹比较以标识与给定数据相关联的任何声明。

[0069] 实际上,可在系统100中的数据第一次满足特定类型的分析时创建内容声明。此后,当数据在系统内传输并且需要被访问时,可使用以前发布的声明以获取关于数据的必要信息,而不要再次重新分析该数据。在一些实施例中,可当在数据上执行新类型的扫描时添加关于数据的另外的声明,从而用新信息扩展了声明集。

[0070] 以下实际示例示出内容名誉服务102可如何用来最小化需要用来检查具体数据的内容的资源。首先,考虑文档文件(例如,微软WORD®文档)被附加至发送到组织中某人的电子邮件的常见情况。在接收到具有附件的电子邮件后,该组织的边缘服务器可扫描该附件,确定它没有病毒、间谍软件以及恶意URL,并且用内容名誉服务102来创建三个声明。例如,接收者然后可接收文件并且将其上传至组织内部的SHAREPOINT ®站点。(假设在这一SHAREPOINT ®站点上执行的安全策略需要为所有文件扫描病毒、间谍软件、恶意URL、以及DLP。)在上传期间,该SHAREPOINT ®站点的安全扫描器可计算该文件的指纹并且向内容名誉服务102发送请求。然后可返回包括三个先前创建的声明的声明集。因此,安全扫描器可确定仅需在文件上执行DLP扫描,并且在执行扫描后可向内容名誉服务发布另外的DLP声明。例如,此后如果相同的文件被上传到组织内的另一个SHAREPOINT®站点,则该SHAREPOINT®站点的安全扫描器可确定不需执行其他扫描,因为该扫描器向内容名誉服务的请求将返回所有必要的声明。

[0071] 重要的是,在大多数情况下,由与内容名誉服务102交互造成的开销要显著低于实际扫描或其他数据评估过程的开销。还需注意,在一些实施例中,包括另外的数据检查过程(该检查过程增加扫描/评估时间)将不会对声明提交和检查时间造成不良影响。

[0072] 在某些实施例中,内容名誉服务102可将内容声明或声明集输出为经过数字签名的文件,例如XML文件。例如,这一“内容证书”可在具有或不具有(如果接收者已经具有该数据)对应数据的情况下被传递至因为某些原因无法访问内容名誉服务以及不能直接与内容名誉服务直接通信的实体。尽管有这一事实,但是接收者可以可靠地验证内容证书的有效性,并且如果有效,可决定信任部分或者全部所包括的内容声明

[0073] 图7示出图4所示的声明集如何在被输出为内容证书时出现的示例。例如,可使用包围的XML数字签名来对这一XML文件签名。这一文件的接收者可首先验证数字签名以确保

XML的完整性未被破坏。例如，如果签名有效，则接收者可使用在XML中指定的算法来计算数据的指纹。如果生成值与XML中的指纹匹配，则接收者可确定这一XML中的所有声明与该数据有关并且是可信任的。因此，在一些实施例中，通过将数据经计算的指纹与同样包含声明的内容证书中所包括的指纹比较，这一内容证书的接收者能够标识与给定数据相关联的一个或多个声明。在一些实施例中，可提供对这些内容证书的自动验证的验证工具。

[0074] 尽管在至少某些情况下可能是不需要的，但是在一些实施例中，内容证书可另外地或另选地被直接附加至或嵌入它所关于的文件中。其中可能有这一实现的文件类型的一个示例是电子邮件。例如，内容证书可被放置在电子邮件的头部空间，而不影响该邮件其余的内容。某些文件格式，例如微软OFFICE ®文件，也允许可存储另外的有效负载的可扩展性。另外地，在一些实施例中，可使用同时存储源文件以及内容证书的通用文件包装封套。例如，可出于此目的使用微软的通用文件保护(GFP)文件包装器。

[0075] 在数据项自身已经被修改时，使用数据项的内容证书(无论作为分别的文件还是作为被附加至或者嵌入这些项的信息)还可提供一些额外的灵活性。例如，在一些实施例中，可通过使用分类技术避免至少某些对内容的重新分类(例如，PII、HBI，等等)，该分类技术生成可用于确定文档与原始文档有多相近的“软散列”。在这些实施例中，如果结果在容限内，则可避免整个重新分类过程。

[0076] 至此描述了本发明的至少一个实施例的若干方面，可以理解，本领域的技术人员可容易地想到各种更改、修改和改进。

[0077] 这样的更改、修改和改进旨在是本发明的一部分，并且旨在处于本发明的精神和范围内。从而，上述描述和附图仅用作示例。

[0078] 可以用多种方式中的任一种来实现本发明的上述实施例。例如，可使用硬件、软件或其组合来实现各实施例。当使用软件实现时，该软件代码可在无论是在单个计算机中提供的还是在多个计算机之间分布的任何合适的处理器或处理器的集合上执行。该处理器可被实现为集成电路，集成电路组件中具有一个或多个处理器。然而，可按照任何适合的方式使用电路实现处理器。

[0079] 此外，应当理解，计算机可以用多种形式中的任一种来具体化，诸如机架式计算机、台式计算机、膝上型计算机、或图形输入板计算机。另外，计算机可以具体化在通常不被认为是计算机但具有合适的处理能力的设备中，包括个人数字助理(PDA)、智能电话、或任何其他适合的便携式或固定电子设备。

[0080] 同样，计算机可以具有一个或多个输入和输出设备。这些设备主要可被用来呈现用户界面。可被用来提供用户界面的输出设备的示例包括用于可视地呈现输出的打印机或显示屏和用于可听地呈现输出的扬声器或其他声音生成设备。可被用于用户界面的输入设备的示例包括键盘和诸如鼠标、触摸板和数字化输入板等定点设备。作为另一示例，计算机可以通过语音识别或以其他可听格式来接收输入信息。

[0081] 这些计算机可以通过任何合适形式的一个或多个网络来互连，包括作为局域网或广域网，如企业网络或因特网。这些网络可以基于任何合适的技术并可以根据任何合适的协议来操作，并且可以包括无线网络、有线网络或光纤网络。

[0082] 而且，此处略述的各种方法或过程可被编码为可在采用各种操作系统或平台中任何一种的一个或多个处理器上执行的软件。另外，这样的软件可使用多种合适的程序设计

语言和/或程序设计或脚本工具中的任何一种来编写，而且还可被编译为可执行机器语言代码或在框架或虚拟机上执行的中间代码。

[0083] 就此，本发明可被具体化为用一个或多个程序编码的一种计算机可读介质(或多种计算机可读介质)(例如，计算机存储器、一个或多个软盘、紧致盘(CD)、光盘、数字视频盘(DVD)、磁带、闪存、现场可编程门阵列或其他半导体器件中的电路配置、或其他非瞬态的有形计算机存储介质)，当这些程序在一个或多个计算机或其他处理器上执行时，它们执行实现本发明的上述各个实施例的方法。这一种或多种计算机可读介质可以是可移植的，使得其上存储的一个或多个程序可被加载到一个或多个不同的计算机或其他处理器上以便实现本发明上述的各个方面。如此处所使用的，术语“非瞬态计算机可读存储介质”只包含可被认为是制品或机器的计算机可读介质。

[0084] 此处以一般的意义使用术语“程序”或“软件”来指可被用来对计算机或其他处理器编程以实现本发明上述的各个方面的任何类型的计算机代码或计算机可执行指令集。另外，应当理解，根据本实施例的一个方面，当被执行时实现本发明的方法的一个或多个计算机程序不必驻留在单个计算机或处理器上，而是可以按模块化的方式分布在多个不同的计算机或处理器之间以实现本发明的各方面。

[0085] 计算机可执行指令可以具有可由一个或多个计算机或其他设备执行的各种形式，诸如程序模块。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。通常，程序模块的功能可以按需在各个实施例中进行组合或分布。

[0086] 而且，数据结构能以任何适合的形式存储在计算机可读介质上。为简化说明，数据结构可被示为具有通过该数据结构中的位置而相关的字段。这些关系同样可以通过对各字段的存储分配传达各字段之间的关系的计算机可读介质中的位置来得到。然而，可以使用任何适合的机制来在数据结构的各字段中的信息之间建立关系，例如通过使用指针、标签、或在数据元素之间建立关系的其他机制。

[0087] 本发明的各个方面可单独、组合或以未在前述实施例中特别讨论的各种安排来使用，从而并不将其应用限于前述描述中所述或附图中所示的组件的细节和安排。例如，可使用任何方式将一个实施例中描述的各方面与其他实施例中描述的各方面组合。

[0088] 同样，本发明可被具体化为方法，其示例已经提供。作为该方法的一部分所执行的动作可以按任何适合的方式来排序。因此，可以构建各个实施例，其中各动作以与所示的次序所不同的次序执行，不同的次序可包括同时执行某些动作，即使这些动作在各说明性实施例中被示为顺序动作。

[0089] 在权利要求书中使用诸如“第一”、“第二”、“第三”等序数词来修饰权利要求元素本身并不意味着一个权利要求元素较之另一个权利要求元素的优先级、先后次序或顺序、或者方法的各动作执行的时间顺序，而仅用作将具有某一名字的一个权利要求元素与(若不是使用序数词则)具有同一名字的另一元素区分开的标签以区分各权利要求元素。

[0090] 同样，此处所使用的短语和术语是出于描述的目的而不应被认为是限制。此处对“包括”、“包含”、或“具有”、“含有”、“涉及”及其变型的使用旨在包括其后所列的项目及其等效物以及其他项目。

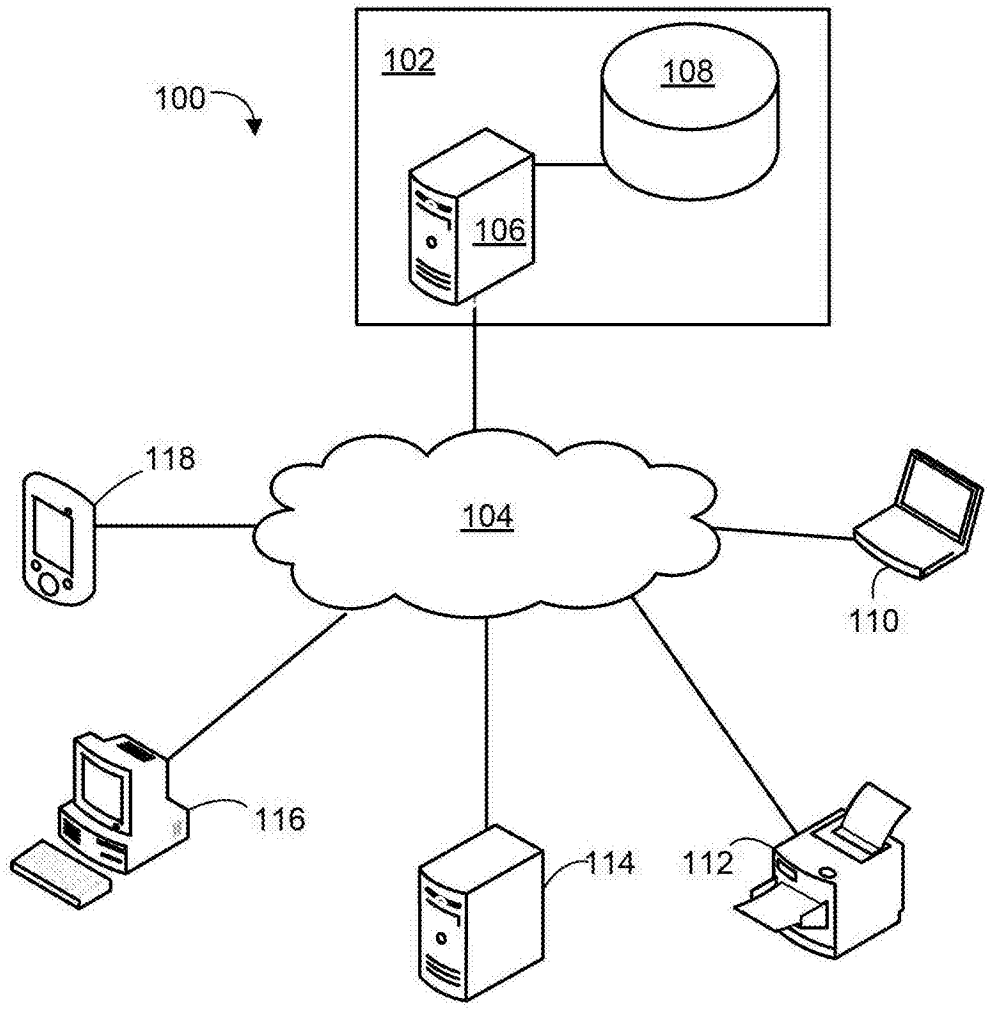

100
102
108
106
104
118
110
116
114
112

图1

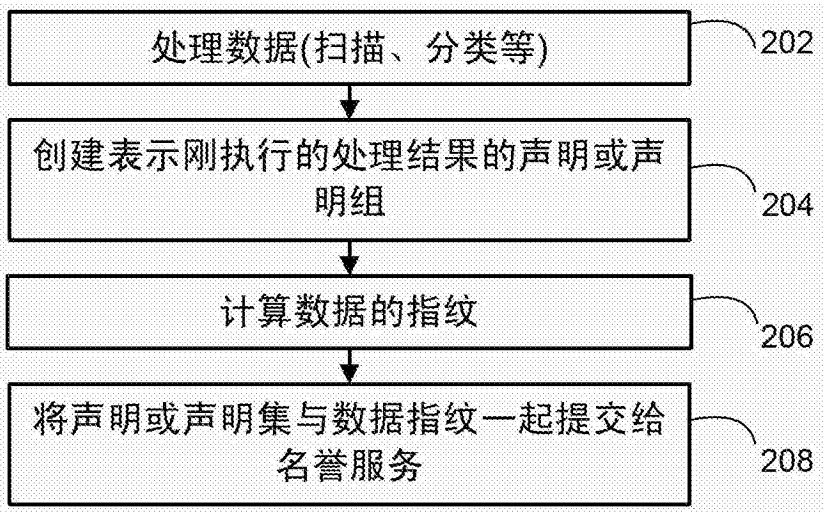

处理数据(扫描、分类等)
202
创建表示刚执行的处理结果的声明或声明组
204
计算数据的指纹
206
将声明或声明集与数据指纹一起提交给名誉服务
208

图2

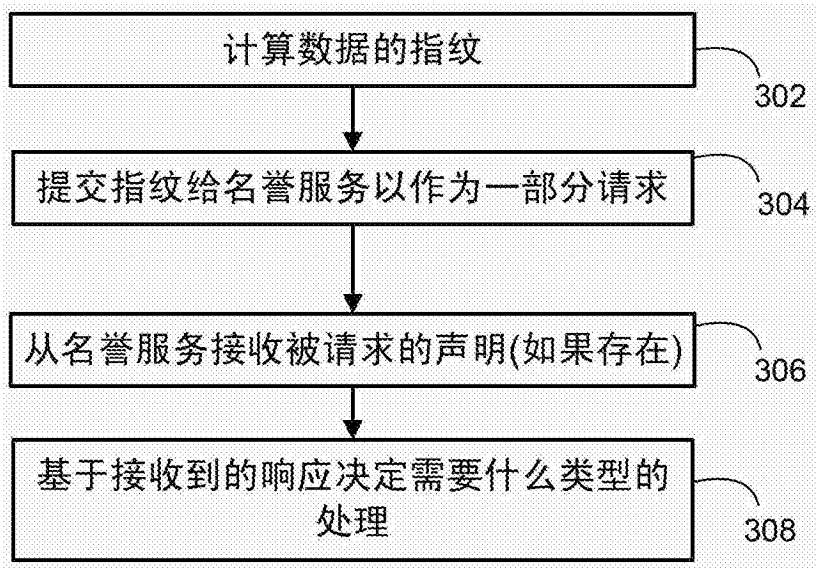

计算数据的指纹
302
提交指纹给名誉服务以作为一部分请求
304
从名誉服务接收被请求的声明(如果存在)
306
基于接收到的响应决定需要什么类型的处理
308

图3

[声明集]
[算法ID: SHA1]
[指纹:
AD6578FF098873EE2DE1BFF4450F2A1BFF436785]
[声明]
类型=位图
时间=05/05/2010 23:11:17
断言=文件类型
发布者=FCI
数据=空
[声明]
类型=明确的内容
时间=05/06/2010 14:24:12
断言=属性不存在
发布者=图像分类器
数据=空
[声明]
类型=版权所有的材料
时间=05/06/2010 14:24:12
断言=属性存在
发布者=图像分类器
数据=空
[声明]
类型=病毒
时间=05/06/2010 14:22:55
断言=属性不存在
发布者=微软
数据=0xfe657dff
[声明]
类型=间谍软件
时间=05/06/2010 11:14:01
断言=属性不存在
发布者=卡巴斯基
数据=<卡巴斯基指定的数据>

图4

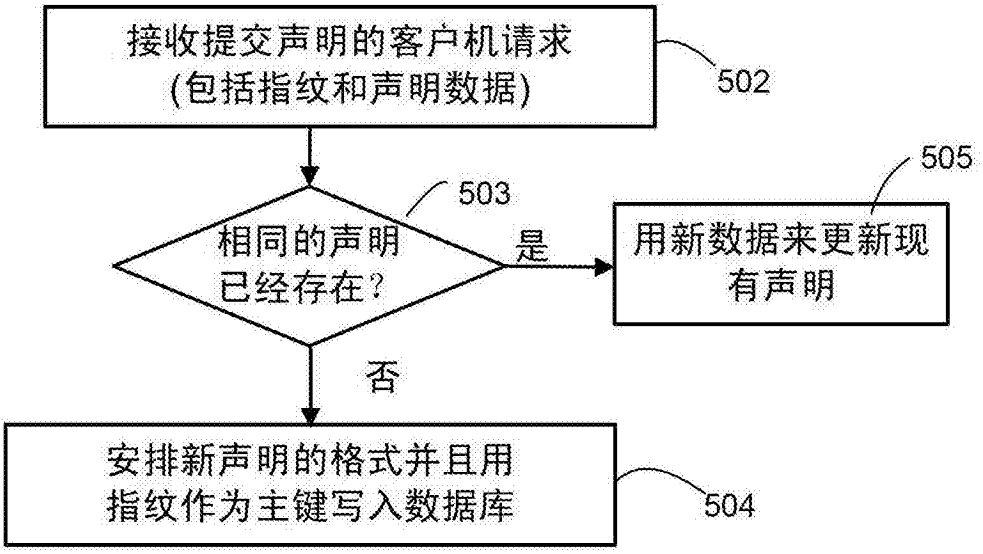

接收提交声明的客户机请求
(包括指纹和声明数据)
502
503
相同的声明
已经存在？
是
505
用新数据来更新现
有声明
否
安排新声明的格式并且用
指纹作为主键写入数据库
504

图5

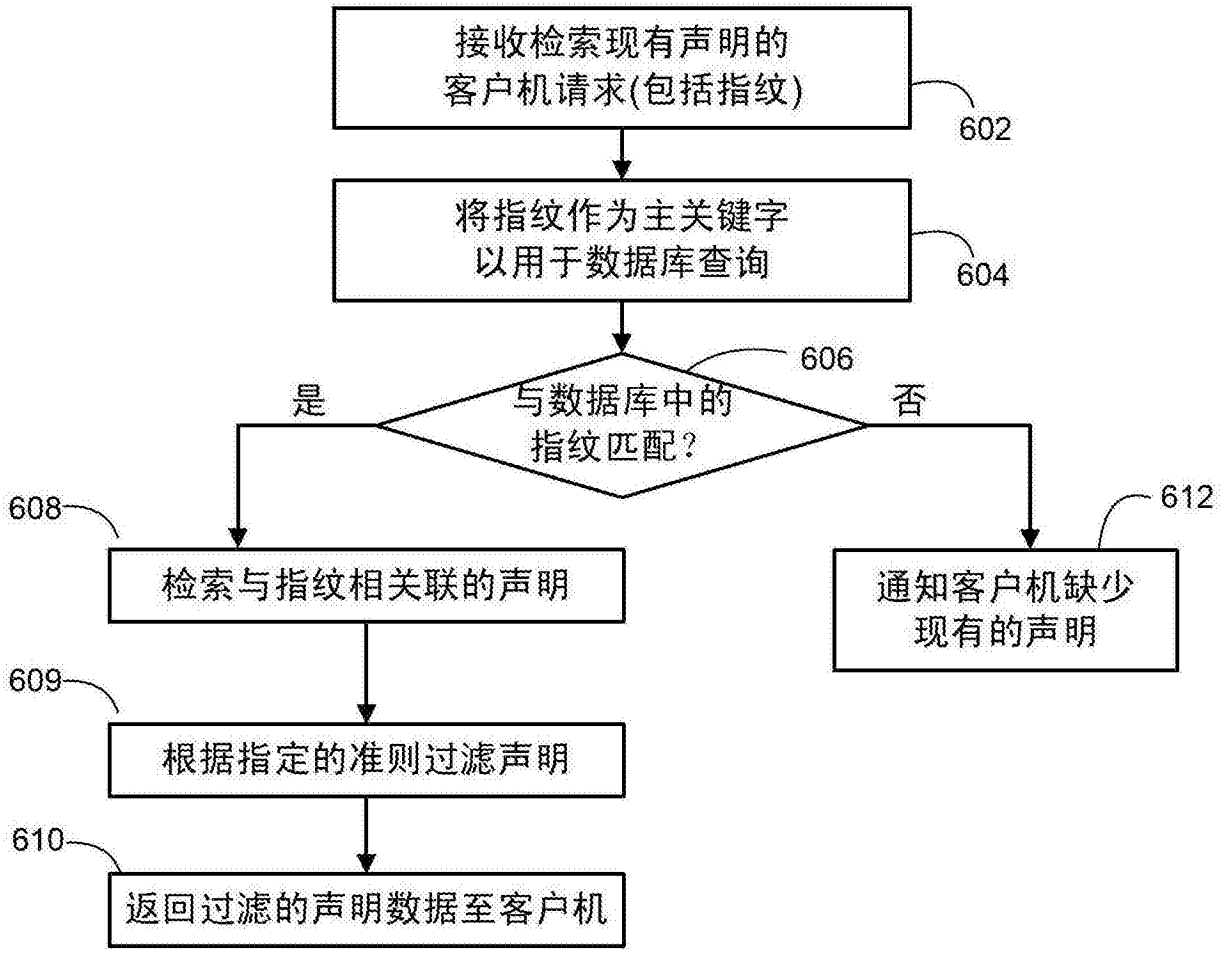

接收检索现有声明的
客户机请求(包括指纹)
602
将指纹作为主关键字
以用于数据库查询
604
606
是
与数据库中的
指纹匹配？
否
608
检索与指纹相关联的声明
612
通知客户机缺少
现有的声明
609
根据指定的准则过滤声明
610
返回过滤的声明数据至客户机

图6

```
      <ContentClaimSet>
        <Fingerprint algorithm="SHA1"
value="AD6578FF098873EE2DE1BFF4450F2A1BFF436785" />
        <ContentClaims>
          <ContentClaim type="Bitmap" time="2010-05-05T23:11:17Z"
assertion="FileType" Issuer="FCI" />
          <ContentClaim type="ExplicitContent" time="2010-05-
06T14:24:12Z" assertion="PropertyAbsent" Issuer="ImageCategorizer" /
>
          <ContentClaim type="CopyrightedMaterial" time="2010-05-
06T14:24:12Z" assertion="PropertyPresent"
Issuer="ImageCategorizer" />
          <ContentClaim type="Virus" time="2010-05-06T14:22:55Z"
assertion="PropertyAbsent" Issuer="Microsoft">
            <Data>0xfe657dff</Data>
          </ContentClaim>
          <ContentClaim type="Spyware" time="2010-05-06T11:14:01Z"
assertion="PropertyAbsent" Issuer="Kaspersky">
            <Data>lH675j44IB7l0pYybHR2QSXsXQYS</Data>
          </ContentClaim>
        </ContentClaims>
        <Signature xmlns="http://www.w3.org/2000/09/xmldsig#">
          <SignedInfo>
            <CanonicalizationMethod Algorithm="http://www.w3.org/TR/
2001/REC-xml-c14n-20010315" />
            <SignatureMethod Algorithm="http://www.w3.org/2000/09/
xmldsig#rsa-sha1" />
            <Reference URI="">
              <Transforms>
                <Transform Algorithm="http://www.w3.org/2000/09/
xmldsig#enveloped-signature" />
              </Transforms>
              <DigestMethod Algorithm="http://www.w3.org/2000/09/
xmldsig#sha1" />
              <DigestValue>hapronkY7kBvxFggzkiQMp/Qpz8=</DigestValue>
            </Reference>
          </SignedInfo>
          <SignatureValue>ImrxJa7gx/
53lH675j44IBQSXsXQYS+e4N5j3ZDBQxbkKJGk...</SignatureValue>
          <KeyInfo>
            <KeyValue>
              <RSAKeyValue>

<Modulus>lgT3XgoYCDGLmQB00Sh8VcLLD65FLpF7c+1MHlrw6mX671snc...</
Modulus>
                <Exponent>AQAB</Exponent>
              </RSAKeyValue>
            </KeyValue>
            <KeyName>MS_CONTENT_REPUTATION</KeyName>
          </KeyInfo>
        </Signature>
      </ContentClaimSet>
```

图7