(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 101512522 B
(45)授权公告日 2011.11.09

(21)申请号 200780025882.5
(22)申请日 2007.07.09
(30)优先权数据
11/484,335 2006.07.10 US
(85)PCT申请进入国家阶段日
2009.01.08
(86)PCT申请的申请数据
PCT/US2007/015730 2007.07.09
(87)PCT申请的公布数据
WO2008/008339 EN 2008.01.17
(73)专利权人 网圣公司
地址 美国加利福尼亚州
(72)发明人 维克托·L·巴杜尔
斯蒂芬·切尼特 丹·哈伯德
尼古拉斯·J·维雷尼尼
阿里·A·梅斯达克
(74)专利代理机构 北京律盟知识产权代理有限责任公司 11287
代理人 刘国伟
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
WO 0155873 A1,2001.08.02,
EP 1318468 A2,2003.06.11,
EP 1318468 A2,2003.06.11,
WO 0155873 A1,2001.08.02,
CN 1527207 A,2004.09.08,
审查员 刘琳

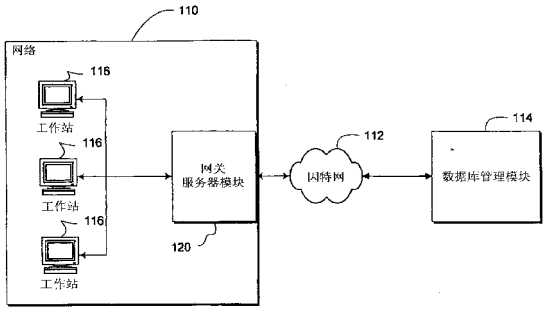
权利要求书 2 页 说明书 17 页 附图 22 页

(54)发明名称

分析网络内容的系统和方法

(57)摘要

本发明提供一种用于识别网络上的网站中的活动内容的系统和方法。一个实施例包含一种将网络内容分类的方法。在一个实施例中，分类指示活动和/或恶意内容。所述方法包含至少部分地基于所述网页的内容识别与所述网页相关联的属性，以及将所述属性存储在网页属性数据库中。所述方法进一步包含对至少一个定义与存储在所述网页属性数据库中的属性进行比较，以及基于对至少一个定义与所述存储的属性进行比较来识别具有所述定义的网页。所述方法进一步包含识别具有与所述至少一个定义相关联的至少一个类别的网页，其中所述类别指示与所述网页相关联的活动内容。其它实施例包含经配置以执行此类方法的系统。



CN 101512522 B

1. 一种将网络内容分类的方法，所述方法包括：

接收一个网页的内容；

确定所述网页包括活动内容和静态内容；

基于所述网页的所述静态内容识别与所述网页相关联的属性；

基于通过执行所述网页的所述活动内容而产生的所述网页的所述活动内容的行为识别与所述网页相关联的属性；

将所述属性存储在网页属性数据库中；

对至少一个定义与存储在所述网页属性数据库中的属性进行比较；

基于对至少一个定义与所述存储的属性进行比较而将所述网页与至少一个定义相关联；以及

基于所关联的定义对所述网页进行分类，其中分类类别指示所述网页中的所述活动内容。

2. 根据权利要求 1 所述的方法，其中对所述网页与所述定义进行比较包括执行与至少一个定义相关联的至少一个数据库查询，其中所述查询从所述网页属性数据库中选择所述网页，所述选择是至少部分地基于所述选择的至少一个网页的所述属性。

3. 根据权利要求 1 所述的方法，其中所述执行所述网页的所述活动内容包括：

将所述网页的所述活动内容传送到一个沙盒环境；

在所述沙盒环境中执行所述网页的所述活动内容；以及

监视计算机的状态以识别所述活动内容的行为。

4. 根据权利要求 1 所述的方法，其中所述将所述网页与至少一个定义相关联包括存储将所述网页中的统一资源定位符与所述类别相关联的数据。

5. 根据权利要求 1 所述的方法，其中所述类别将所述至少一个网页识别为具有经计算以利用客户端计算机的弱点的内容。

6. 根据权利要求 1 所述的方法，其进一步包括从定义数据库接收所述至少一个定义。

7. 根据权利要求 1 所述的方法，其中所述定义中的至少一者包括逻辑表达式。

8. 根据权利要求 7 所述的方法，其中所述逻辑表达式包括至少一个项，所述项包括至少一个网页属性与至少一个其它值的关系。

9. 根据权利要求 8 所述的方法，其中所述至少一个其它值包括常数值。

10. 根据权利要求 8 所述的方法，其中所述至少一个其它值包括至少一个其它网页属性。

11. 根据权利要求 1 所述的方法，其中所述属性中的至少一者与字符串相关联。

12. 根据权利要求 1 所述的方法，其中所述属性中的至少一者与常规表达式相关联。

13. 根据权利要求 11 所述的方法，其中所述属性中的所述至少一者包括指示所述网页的所述内容内的发生率的数字。

14. 根据权利要求 11 所述的方法，其进一步包括确定与所述网页的 URL 相关联的记分，其中识别具有至少一个类别的所述网页是至少部分地基于所述记分。

15. 一种用于将网络内容分类的系统，所述系统包括：

存储与网页相关联的属性的装置；

接收一个网页的内容的装置；

确定所述网页包括活动内容和静态内容的装置；

基于所述网页的所述静态内容来识别与所述网页相关联的属性的装置；

基于通过执行所述网页的所述活动内容而产生的所述网页的所述活动内容的行为来识别与所述网页相关联的属性的装置；

将所述属性存储在所述网页属性数据库中的装置；

对至少一个定义与存储在所述网页属性数据库中的属性进行比较的装置；

基于对至少一个定义与所述存储的属性进行比较而将所述网页与至少一个定义相关联的装置；以及

基于所关联的定义对所述网页进行分类，其中分类类别指示与所述网页中的所述活动内容的装置。

16. 根据权利要求 15 所述的系统，其进一步包括至少部分地通过执行与至少一个定义相关联的至少一个数据库查询来对所述网页与所述定义进行比较的装置，其中所述查询从所述网页属性数据库选择所述网页，所述选择是至少部分地基于所述选择的至少一个网页的所述属性。

17. 根据权利要求 15 所述的系统，其中所述活动内容通过以下装置执行：

接收来自所述第一处理器的所述网页的所述活动内容的装置；

将所述活动内容存储在一个沙盒环境的装置；

在所述沙盒环境中执行所述网页的所述活动内容的装置；以及

监视计算机的状态以识别所述活动内容的行为的装置。

18. 根据权利要求 15 所述的系统，其进一步包括至少部分地通过存储将所述网页中的所述至少一者的统一资源定位符与所述类别相关联的数据来关联所述网页和一个类别的装置。

19. 根据权利要求 15 所述的系统，其中所述类别将所述网页中的所述至少一者识别为具有经计算以利用客户端计算机的弱点的内容。

20. 根据权利要求 15 所述的系统，其进一步包括存储所述网页的所述属性的装置。

21. 根据权利要求 15 所述的系统，其进一步包括存储所述至少一个定义的装置。

22. 根据权利要求 15 所述的系统，其中所述至少一个定义包括逻辑表达式。

23. 根据权利要求 22 所述的系统，其中所述逻辑表达式包括至少一个项，所述项包括至少一个网页属性与至少一个其它值的关系。

24. 根据权利要求 23 所述的系统，其中所述至少一个其它值包括常数值。

25. 根据权利要求 23 所述的系统，其中所述至少一个其它值包括至少一个其它网页属性。

26. 根据权利要求 15 所述的系统，其中所述属性中的至少一者与字符串相关联。

27. 根据权利要求 15 所述的系统，其中所述属性中的至少一者与常规表达式相关联。

28. 根据权利要求 15 所述的系统，其中所述属性中的至少一者包括指示所述网页的所述内容内的发生率的数字。

分析网络内容的系统和方法

[0001] 相关申请案

[0002] 本申请案涉及 2006 年 7 月 10 日申请的第 11/484,240 号美国专利申请案（代理人案号 WEBSEN.083A），所述申请案的全文以引用的方式并入本文。

技术领域

[0003] 本申请案涉及数据和应用程序安全性。确切地说，本申请案揭示收集和挖掘数据以确定数据是否与恶意内容相关联的系统方法。

背景技术

[0004] 传统上，计算机病毒和其它恶意内容最经常通过将受感染的磁盘或某种其它物理媒体插入计算机而被提供到客户端计算机。随着电子邮件和因特网的使用增加，电子邮件附件攻击变为用于向计算机分布病毒代码的普遍方法。为了用这些类型的具有恶意内容的病毒感染计算机，通常需要用户的某种同意动作，例如打开受感染的文件附件或从网站下载受感染的文件并在用户的计算机上启动所述文件。随着时间的过去，反病毒软件制作者开发出日益有效的程序，所述程序经设计以扫描文件并在其有机会感染客户端计算机之前将其消毒。因此，计算机黑客不得不创造更聪明且创新的方法来用其恶意代码感染计算机。

[0005] 在当今的日渐联网的数字世界，正开发分布式应用程序以在开放的、合作的联网环境中向用户提供越来越多的功能性。尽管这些应用程序较有力且复杂，但其增加的功能性要求网络服务器以更集成的方式与客户端计算机交互。举例来说，在先前网络应用程序主要向客户端浏览器提供 HTML 内容并经由 HTTP 邮递命令从客户端接收回数据的情况下，许多新的网络应用程序经配置以向客户端计算机发送各种形式的目标内容（例如活动内容），其引起在较新的网络浏览器的增强特征内启动应用程序。举例来说，许多基于网络的应用程序现在利用活动 -X(Active-X) 控件，其必须下载到客户端计算机以使得其可被有效地利用。在特定例子中 Java 小程序 (Java applet)、Java 脚本 (JavaScript) 和 VB 脚本 (VBScript) 命令也有能力修改客户端计算机文件。

[0006] 这些功能性增加所带来的便利也有代价。较新的网络应用程序和内容显然比先前的应用程序环境更有力。因此，其还为将恶意代码下载到客户端计算机提供了机会。另外，随着操作系统和网络浏览应用程序的复杂性增加，更难以识别可能会允许黑客将恶意代码传送到客户端计算机的安全性弱点。尽管浏览器和操作系统厂商通常会发布软件更新以补救这些弱点，但许多用户尚未配置其计算机以下载这些更新。因此，黑客已开始编写利用这些弱点以将其本身下载到用户的机器而不用依赖于用户的任何特定活动（例如启动受感染的文件）的恶意代码和应用程序。此攻击的一个实例是使用嵌入在网站上的活动内容对象的恶意代码。如果恶意代码已经配置以利用网络浏览器中的弱点，那么用户可能仅仅因为访问过所述页面就会被恶意代码感染或损害，因为所述页面中的目标内容将在用户的计算机上执行。

[0007] 解决嵌入活动内容中的恶意代码的问题的一种尝试是利用网络浏览器上的升高

的安全性设定。然而在许多公司环境中，内部网或外部网应用程序经配置而向客户端计算机发送可执行内容。将浏览器设定设定为高安全性等级往往会妨碍或阻碍对这些类型的“安全”应用程序的有效使用。解决问题的另一尝试是使用网络防火墙应用程序来阻止所有可执行内容。此强力方法在许多环境中也是低效的，因为为了让软件正确地起作用，有必要对特定类型内容进行选择性接入。

[0008] 需要一种允许检测恶意网络内容而不会损害用户功能性的系统和方法。此外，需要一种可检测例如活动内容的目标内容并快速识别和归类其行为，且以最小延迟向大量客户端计算机提供针对恶意内容的保护的系统。

发明内容

[0009] 本发明的系统、方法和装置每一者均具有若干方面，其中任何单个一个方面均不唯一负责其所需的属性。现在将简要论述本发明的若干特征，但并不限制本发明的范围。

[0010] 一个实施例包含将网络内容分类的方法。所述方法包含接收至少一个网页的内容。所述方法进一步包含至少部分地基于所述网页的所述内容识别与所述网页相关联的属性。所述方法进一步包含将属性存储在网页属性数据库中。所述方法进一步包含对至少一个定义与存储在所述网页属性数据库中的属性进行比较。所述方法进一步包含基于对至少一个定义与存储的属性进行比较来识别具有所述定义的网页。所述方法进一步包含识别具有与所述至少一个定义相关联的至少一个类别的所述网页，其中所述类别指示与所述网页相关联的活动内容。

[0011] 一个实施例包含用于将网络内容分类的系统。所述系统包含数据库，其经配置以与网页相关联的属性。所述系统进一步包含至少一个处理器，其经配置以至少部分地基于网页的内容识别与网页相关联的属性，以及将属性存储在网页属性数据库中。所述处理器进一步经配置以对至少一个定义与存储在网页属性数据库中的属性进行比较，基于对至少一个定义与存储的属性进行比较来识别具有所述定义的网页，以及识别具有与所述至少一个定义相关联的至少一个类别的网页，其中所述类别指示与网页相关联的活动内容。

附图说明

[0012] 在本描述内容中参看附图，其中始终以相同标号指代相同部分。

[0013] 图 1 是根据本发明方面的系统的各种组件的方框图。

[0014] 图 2 是来自图 1 的工作站模块的方框图。

[0015] 图 3 是来自图 1 的网关服务器模块的方框图。

[0016] 图 4 是记录数据库的实例。

[0017] 图 5 是 URL 接入策略数据库表的实例。

[0018] 图 6A 和 6B 分别是经归类和未经归类的 URL 的实例。

[0019] 图 7 是来自图 1 的数据库管理模块的方框图。

[0020] 图 8 是来自图 7 的收集系统的方框图。

[0021] 图 9 是来自图 8 的收集模块的方框图。

[0022] 图 10 展示根据本发明某些方面的蜜罐客户端系统。

[0023] 图 11 是由来自图 9 的收集模块收集的 URL 相关数据的实例。

[0024] 图 12 是来自图 7 的记分和归类模块的方框图。

[0025] 图 13A 是属性表的实例。

[0026] 图 13B 是经处理的网页属性表的实例。

[0027] 图 13C 是定义表的实例。

[0028] 图 14 是说明来自图 7 的训练模块的一个实施例的方框图。

[0029] 图 15 是说明来自图 12 的活动分析系统的一个实施例的方框图。

[0030] 图 16 是描述在一个实施例中可如何在网关服务器模块中处理 URL 的流程图。

[0031] 图 17 是描述根据某些实施例可如何结合策略模块通过网关服务器模块处理 URL 的流程图。

[0032] 图 18 是描述收集系统可如何在网关服务器模块内处理 URL 的流程图。

[0033] 图 19 是描述收集系统可如何在数据库管理模块内处理 URL 的流程图。

[0034] 图 20 是数据挖掘系统的方框图。

[0035] 图 21 是说明在数据库管理模块内将 URL 归类的方法的一个实施例的流程图。

[0036] 图 22 是说明在图 21 的方法中识别 URL 的属性的方法的一个实施例的流程图。

[0037] 图 23 是说明在图 21 的方法中基于 URL 属性将 URL 归类的方法的一个实施例的流程图。

[0038] 图 24 是说明识别在图 22 和 23 的方法中将 URL 归类时使用的属性的方法的一个实施例的流程图。

具体实施方式

[0039] 以下详细描述是针对本发明的某些具体实施例。然而，本发明可以权利要求书定义和涵盖的许多不同方式来实施。在本描述内容中参看附图，其中始终以相同标号指代相同部分。

[0040] 特定实施例提供识别和归类在通过统一资源定位符 (URL) 识别的位置发现的网络内容的系统和方法，所述内容包含可能可执行的网络内容和恶意内容。如本文使用，可能可执行的网络内容通常指包含由网络浏览器或网络客户端计算机执行的指令的任何类型的内容。可能可执行的网络内容可包含例如小程序、嵌入 HTML 或其它超文本文档（包含例如 Java 脚本或 VB 脚本的脚本语言）的可执行代码、嵌入其它文档（例如微软 Word 宏或样式表）中的可执行代码。可能可执行的网络内容也可指执行位于另一位置（例如另一网页、另一计算机或网络浏览器计算机本身上）中的代码的文档。举例来说，通常可认为包含“对象”元素且因此可引起活动 X 或其它可执行组件的执行的 HTML 网页是可能可执行的网络内容，无论所述可执行组件的位置如何。恶意内容可指不可执行但可经计算以利用客户端计算机的弱点的内容。然而，可能可执行的网络内容也可能是恶意内容。举例来说，已使用图像文件来在所述图像经处理用于显示时利用某些操作系统中的弱点。而且，恶意网络内容也可指例如“网络钓鱼 (phishing)”方案的交互内容，在所述方案中，HTML 表格或其它网络内容经设计以表现为由例如银行的另一（通常是受到信任的）网站提供，以便欺骗用户向未经授权方提供证书或其它敏感信息。

[0041] 系统的描述

[0042] 图 1 提供示范性系统的最高级说明。系统包含网络 110。网络 110 可以是局域网、

广域网或某种其它类型的网络。网络 110 可包含一个或一个以上工作站 116。工作站 116 可以是附接到网络的各种类型的客户端计算机。客户端计算机 116 可以是桌上型计算机、笔记型计算机、手持式计算机或类似计算机。客户端计算机也可装载有操作系统，所述操作系统允许客户端计算机通过例如网络浏览器、电子邮件程序等各种软件模块利用网络。

[0043] 每一工作站 116 均可与网关服务器模块 120 电连通。网关服务器模块可驻存在网络 110 的边缘，使得从因特网 112 和向因特网 112 发送的业务可在进入或离开网络 110 的途中经过网关服务器模块。网关服务器模块 112 可采用安装在服务器上的软件模块的形式，所述服务器作为向比工作站 116 直接附接到的网络 110 广的区域网络 112 的网关而起作用。数据库管理模块 114 也连接到因特网 112。数据库管理模块也可以是驻存在一个或一个以上计算装置上的软件模块（或一个或一个以上硬件器件）。数据库管理模块 114 可驻存在包含某类网络连接硬件（例如网络接口卡）的机器上，所述网络连接硬件允许数据库管理模块 114 向因特网 112 发送数据和信息以及从因特网 112 接收数据和信息。

[0044] 现在参看图 2，呈现工作站 116 的更详细视图。工作站 116 可包含工作站模块 130。工作站模块 130 可采用经安装以在工作站 116 的操作系统上运行的软件的形式。或者，工作站模块 130 可以是在另一机器上运行的由工作站 116 远程启动的应用程序。

[0045] 下作站模块 130 可包含各种组件。工作站模块可包含本地活动内容模块 132 的清单 (inventory)，其记录存储在工作站 116 上的所有网络内容。举例来说，本地内容清单模块 132 可周期性列出所有本地内容的清单。清单中列出的数据可上载到网关服务器模块 120 以与经归类的 URL/ 内容数据库 146 进行比较。本地内容清单模块 132 可通过与清单中列出的本地内容 132 进行比较来确定是否有新内容正在被引入到工作站 116。

[0046] 工作站模块还可包含上载 / 下载模块 134 和 URL 请求模块 136。上载 / 下载模块 134 可用于通过网关服务器模块 120 从网络 110 向因特网 112 发送和接收数据。URL 请求模块 136 从用户或某个系统过程接收 URL 输入，且可经由网关服务器模块 120 发送请求以检索与所述 URL 相关联的文件和 / 或内容。通常，上载 / 下载模块 134 和 URL 请求模块 136 中的每一者的功能可由例如网络浏览器的软件应用程序执行，其中因特网探测器® (Internet Explorer®)、谋智火狐 (Mozilla Firefox)、奥普拉 (Opera)、远征 (Safari) 是此项技术中众所周知的浏览软件的实例。或者，模块的功能可在不同的软件应用程序之间划分。举例来说，FTP 应用程序可执行上载 / 下载模块 134 的功能，而网络浏览器可执行 URL 请求。其它类型的软件也可执行上载 / 下载模块 134 的功能。尽管工作站上通常不需要这些类型的软件，但例如间谍软件 (Spyware) 或特洛伊木马 (Trojan Horses) 的软件可能做出从因特网发送和接收数据的请求。

[0047] 工作站模块 130 可与网关服务器模块 120 通信。网关服务器模块 120 可用于分析传入和传出的网络业务并做出关于所述业务对工作站 116 可能造成的影响的各种确定。现在参看图 3，提供网关服务器模块 120 的实例。网关服务器模块 120 与工作站 116 双向通信。其可从工作站模块 130 接收文件上载和下载以及 URL 请求。网关服务器模块 120 还与因特网 112 双向通信。因此，源自网络 110 的工作站 116 内的请求可能需要在其前进到因特网时通过网关服务器模块 120。在一些实施例中，网关服务器模块 120 可与保护网络 110 免受来自因特网 112 的未经授权的入侵的某个防火墙硬件或软件集成。在其它实施例中，网关服务器模块 120 可以是独立的硬件器件乃至是安装在驻存于到因特网 112 的网络网关

处的单独网关服务器上的软件模块。

[0048] 如上论述，网关服务器模块 120 可借助于工作站模块 130 而从工作站 116 接收 URL 请求和上载 / 下载数据。网关服务器模块 120 可包含基于所接收数据执行各种功能的各种组件。

[0049] 网关服务器模块 120 中包含的一个特征是经归类 URL 数据库 146。URL 数据库 146 可用于存储包含与 URL 相关联的数据的关于 URL 的信息。经归类 URL 数据库 146 可以是关系数据库，或其可以例如平面文件、面向对象的数据库的某种其它形式存储，且可经由应用程序编程接口 (API) 或某个数据库管理软件 (DBMS) 存取。URL 数据库 146 通常可用于帮助确定由 URL 请求模块 136 发送的 URL 请求是否将被许可完成。在一个实施例中，将存储在 URL 数据库 146 中的 URL 归类。

[0050] 网关服务器模块 120 还可包含策略模块 142。策略模块 142 可用于实施关于特定内容将如何由网关服务器模块 120 或由安装在网络 110 内的防火墙或某种其它安全性软件处理的网络策略。在一个实施例中，策略模块 142 可经配置以提供关于如何处理针对经归类 URL 的 URL 请求的系统引导。举例来说，网关服务器模块 120 可经配置以不允许归类为“恶意”或“间谍软件”的 URL 请求。在其它实施例中，策略模块 142 可用于确定如何处理未经归类的 URL 请求。在一个实施例中，系统可经配置以阻止针对不在经归类 URL 数据库 146 中的 URL 的所有请求。策略模块 142 还可经配置以基于做出请求的用户或做出请求的时间而允许某些对未经归类 URL 的请求。这允许系统在通用型 (one-size-fits-all) 配置将不满足运行网关服务器模块 120 的组织的业务需要时避免具有所述配置。

[0051] 网关服务器模块 120 可包含收集模块 140。收集模块 140 可以是用于收集关于 URL 的数据的软件程序、例行程序或过程。在一个实施例中，当从 URL 请求模块 136 接收到针对特定 URL 的请求时，收集模块 140 可经配置以访问所述 URL 并下载页面数据到网关服务器模块 120 以供网关服务器模块 120 的组件进行分析。下载的数据还可经由因特网 112 发送以传递到数据库管理模块 114(如下文将进一步论述的)。

[0052] 在一些实施例中，网关服务器模块 120 还可包含记录数据库 144。记录数据库 144 可执行各种功能。举例来说，其可存储网络 110 内的特定类型发生情况的记录。在一个实施例中，记录数据库 144 可经配置以记录工作站 116 请求未经授权 URL 的每一事件。在一些实施例中，记录数据库 144 还可经配置以记录特定未经归类 URL 被请求的频率。此信息可用于确定未经归类 URL 是否应具有特定重要性或优先权且应先于较早的接收到的数据而由数据库管理模块 114 归类。在一些实施例中，未经归类 URL 可单独存储在未经归类 URL 数据库 147 中。

[0053] 举例来说，可编写某个间谍软件以从特定 URL 请求数据。如果网络 110 内的许多工作站 116 被所述间谍软件感染，则对特定 URL 的重复请求可提供网络内存在某种异常的指示。记录数据库也可经配置以记录对经归类 URL 数据的请求。在一些实施例中，对经归类 URL 的请求归类可有助于确定特定 URL 是否被错误地特征化。

[0054] 现在参看图 4，论述记录数据库 144 的实例。记录数据库 144 包含四列数据。第一列“页面请求次数”152 指示网络 110 内的用户请求特定 URL 的次数。第二列“URL”154 记录正在记录数据库 144 中记录的特定 URL 串。因此，当将 URL 发送到记录数据库 144 时，可首先搜索数据库以确定所述 URL 串是否已在其中。如果不是，那么可将 URL 串添加到数据

库。在一些实施例中，收集模块 140 可经配置以访问所请求的 URL 并收集关于所述 URL 的数据。收集模块 140 可检索所请求 URL 的页面来源并对其进行扫描以查找可能指示内容类型的特定关键词。举例来说，如果页面来源包含“javascript://”，那么所述页可被识别为具有 Java 脚本。尽管此内容并非固有危险的，但具有 Java 脚本的网页包含恶意内容的可能性可能更大，所述恶意内容经设计以利用浏览器应用程序处理 Java 脚本函数调用的方式。在一些实施例中，此数据可存储在记录数据库 144 中在 Java 脚本列 155 中。记录数据库也可从包含活动 -X 内容的页面接收类似的信息并将所述内容存储在活动 X 列 156 内。在其它实施例中，可针对 Java 小程序、VB 脚本等检测和存储其它类型的活动内容。

[0055] 再次参看图 3，网关服务器模块 120 可进一步包含管理界面模块 148 或“管理模块”。管理模块 148 可用于允许网络管理员或组织内的其它技术人员配置网关服务器模块 120 的各种特征。在某些实施例中，管理模块 148 允许网络管理员或某种其它网络管理类型来配置策略模块 142。

[0056] 现在参看图 5，提供 URL 接入策略数据库 158 的实例。URL 接入策略数据库 158 可由策略模块 142 用于实施用于网络 110 内的工作站 116 接入基于网络的内容的策略。在所示的实施例中，URL 接入策略数据库 158 包含具有四列的表。第一列是用户列 160。“用户”列 160 包含关于服从于在表的给定行中定义的策略的用户的数据。下一列“类别”162 列出所述行定义的策略所适用的内容的类别。第三列“总是阻止”164 表示当所请求内容的用户和类别 166 匹配于所述特定行中定义的用户和类别时系统实施的行为或策略。在一个实施例中，“总是阻止”字段可以是其中数据可设定为真或假的布尔型字段。因此，在数据表所示的第一行中，策略模块 142 经配置以“总是阻止”用户“asmith”对“恶意内容”的请求。

[0057] 如上所述，策略模块还可经配置以基于不同时间实施策略。在图 5 提供的实施例中，第四列“允许的时间”166 提供此功能性。第二行数据提供如何实施时间策略的实例。用户 164 设定为“bnguyen”且类别 162 是“赌博”。正如保留为空白的字段所指示的，策略未经配置以针对“bnguyen”“总是阻止”赌博内容。然而，这些 URL 请求被许可的时间限于从 6PM 到 8AM。因此，采用这些类型的策略允许网络管理员向工作站和用户提供某一程度的灵活性，但此灵活性的提供是以在典型工作时间期间网络业务不受损害的方式进行的。

[0058] 图 6A 和 6B 提供对经归类 URL 数据库 146 可如何存储经归类数据的说明。在一个实施例中，经归类 URL 可存储在例如图 6A 所示的两列数据库表中。在一个实施例中，所述表可包含 URL 列 172，其可仅存储已经特征化的 URL 串。类别列 174 可存储关于所述 URL 已如何由数据库模块 114 特征化的数据（如下文将详细描述）。在一个实施例中，可对 URL 字段编索引以使得其可被实时地更快速地搜索。因为经归类 URL 的列表可能涉及到数百万个 URL，所以快速接入例行程序是有益的。

[0059] 现在参看图 6B，提供未经归类 URL 的表 147（上文结合图 3 描述）。此表中可填充有来自工作站 116 的 URL 请求，所述 URL 请求是请求在经归类 URL 表 146 中不存在的 URL。如下文将更详细描述，网关服务器模块 120 可经配置以查询经归类 URL 数据库 146 以确定是否应阻止所请求的 URL。如果所请求 URL 在经归类数据库 146 中，则策略模块可确定是否允许所述请求前进到因特网 112。然而如果在经归类 URL 数据库中没有发现所请求 URL，则可将其添加到未经归类 URL 列表 176，使得其可经由因特网 112 发送到数据库管理模块 114 并稍后经分析和归类且下载到经归类 URL 数据库 146 中。

[0060] 图 7 是对数据库管理模块 114 中可包含的各种组件的说明。如上文论述，数据库管理模块 114 可位于网络 110 及其相关联工作站 116 的远端（可经由因特网 112 接入）。数据库管理模块可采用一个或许多不同硬件和软件组件的形式，例如同时运行数百个服务器以实现改善性能的服务器库。

[0061] 在一个实施例中，数据库管理模块 114 可包含上载 / 下载模块 178。上载 / 下载模块 178 可以是软件或硬件组件，其允许数据库管理模块 114 从因特网 112 向任意数目的位置发送和接收数据。在一个实施例中，上载 / 下载模块经配置以向因特网 112 上的网关服务器模块 120 发送新归类的 URL 以添加到其本地 URL 数据库 146。

[0062] 数据库管理模块 114 还可包含 URL/ 内容数据库 180。URL/ 内容数据库 180 可采用数据仓库的形式，其存储 URL 串和关于已由收集系统 182 收集的 URL 的信息。URL/ 内容数据库 180 可以是经编索引以提供快速且有效的数据搜索的关系数据库。在某些实施例中，URL 数据库可以是数据入库应用程序，其跨越许多物理硬件组件和存储媒体。URL 数据库可包含例如以下数据 :URL 串、与这些串相关联的内容、关于如何收集到内容（例如，通过蜜罐客户端、通过客户提交等）的信息，且可能包含 URL 被写入到 URL/ 内容数据库 180 内的日期。

[0063] 数据库管理模块 114 可进一步包含训练系统 184。训练系统 184 可以是软件 / 硬件模块，其用于定义可用于归类基于网络的内容的属性和定义。数据库管理模块 114 可进一步提供记分 / 分类系统 186，其利用由训练系统 184 创建的定义和属性来向网络内容提供记分或分类（例如，归类），使得所述归类可经由上载 / 下载模块 178 传递到网关服务器模块 120。

[0064] 现在参看图 8，提供收集系统 182 的更详细视图。收集系统 182 可包含收集模块 190，其（直接或间接）耦合到数据挖掘模块 192。收集模块 190 可由数据库管理模块 114 用于为 URL 数据库 180 收集关于未经归类的 URL 的数据。除了 URL 之外，URL 数据库 180 还可存储与 URL 相关联的内容。收集模块还可用于收集 URL 供其它系统组件进行额外分析。收集模块 190 可与其可从其收集关于 URL 的数据的一个或一个以上收集源 194 相关联。收集源可采用各种形式。在一些实施例中，收集源 194 可包含主动与被动蜜罐和蜜罐客户端、存储在网关服务器模块 120 上的记录数据库 144 的用以识别应用程序的数据分析、用于收集的 URL 和协议。收集源也可以是网络爬行 (webcrawling) 应用程序，其针对特定关键词搜索因特网 112，或在页面内容内搜索短语。收集源 194 还可包含从 DNS 数据库挖掘的 URL 和 IP 地址数据以识别与已知恶意 IP 地址相关联的域。在一些实施例中，可通过从共享恶意代码和恶意 URL 样本的其它组织接收此信息以收集用于归类的 URL。在又一些实施例中，可经由电子邮件模块收集 URL，所述模块经配置以从整个公众接收举报 (tip)，近似于通过罪犯举报热线来识别罪犯的方式。

[0065] 现在参看图 9，提供收集模块 190 的更详细视图。收集模块 190 可包含允许其有效利用上述收集源中每一者的各种子组件。收集模块 190 可包含搜索短语数据模块 197 和表达式数据模块 198。搜索短语数据模块 197 收集并提供可能与识别不适当内容相关的搜索短语。表达式数据模块可包含各种类型的表达式，例如常规表达式、操作数或某种其它表达式。搜索短语数据模块 197 和表达式数据模块 198 每一者可包含可更新的记录组，其可用于定义用于网络爬行收集源 194 的搜索参数。收集模块 190 还可包含优先权模块 200。

优先权模块 200 可采用在收集系统 182 内运行的软件过程的形式，或者其可作为单独过程运行。优先权模块可用于对收集模块收集的数据区分优先次序，以便使较可能危险或可疑的 URL（或数据）在较可能无害的 URL 之前受到严格的检查。在一个实施例中，优先权模块 200 可基于接收的 URL 来自的收集源 194 而指派优先权。举例来说，如果从客户报告接收到 URL，则可为其指定较高的优先权。类似地，如果从接入在过去主机恶意内容已知的域或 IP 地址或子网的网络爬行器接收到 URL，则所述 URL 可得到高优先权。类似地，由蜜罐客户端（下文更详细论述）识别的可能危险的网站也可得到高优先权。收集模块 190 还可包含数据选择模块 202，其可与优先权模块 200 一起工作以确定所识别 URL 是否应被标记为用于归类的候选 URL。在一个实施例中，数据选择 URL 可提供用于接收搜索参数的用户界面以通过基于优先权和内容搜索数据来进一步细化经区分优先次序的数据。

[0066] 如上文指示，收集模块还可包含数据下载模块 204。数据下载模块 204 可经配置以识别 URL 以进行访问以及从所访问 URL 下载数据和内容。数据下载模块可结合收集模块中的各种子系统一起工作，以检索用于 URL 数据库 180 的数据。一个此子系统是网络爬行器模块 206。网络爬行器模块 206 可以是软件应用程序，其经配置以通过接入网页并跟随包含在所述页面中的超链接来接入因特网 112 上的网站。网络爬行器模块 206 可配置有若干同时的过程，所述过程允许模块同时爬行许多网站并将所访问 URL 报告回 URL 数据库 180，如下文将更详细论述。收集模块 190 还可包含蜜罐客户端模块 208。蜜罐客户端模块 208 是软件过程，其经配置而以吸引存储在所访问页面内的恶意代码的方式模仿网络浏览者访问网站的行为。蜜罐客户端模块 208 可访问网站并跟踪网站的行为，且将内容下载回到 URL 数据库 180 供进一步分析。

[0067] 下载模块 204 还可包含第三方供应者模块 212，其经配置以从第三方接收 URL 和相关联的内容。举例来说，第三方模块 212 可经配置以提供可由一般公众接入的网站。所述模块可经配置以接收输入 URL 串，所述串随后可被输入到 URL 数据库 180 中。在一些实施例中，第三方模块还可经配置以接收来自专有或公共邮寄列表的电子邮件，且识别所述电子邮件内嵌入的任何 URL 数据以存储在 URL 数据库 180 中。

[0068] 下载模块还可包含网关服务器接入模块 210。网关服务器接入模块是软件组件或程序，其可经配置以有规律地接入网关服务器模块 120 上的记录数据库 144 以下载 / 上载由记录数据库 144 识别的所有新未经归类的网络内容。

[0069] 返回参看图 8，收集系统还可包含数据挖掘模块 192。数据挖掘模块 192 可用于获得关于存储在 URL 数据库 180 中的 URL 的额外数据。在许多例子中，由收集源 194 供应到收集模块 190 和 URL 数据库 180 的信息仅限于 URL 串。因此，为了使系统有效地归类所述 URL 内的内容，可能必须有更多数据。举例来说，可能需要检查实际的页面内容以确定是否存在嵌入 URL 内的危险内容。数据挖掘模块 192 用于收集关于 URL 的此额外必要数据，且下文将更详细论述。

[0070] 图 10 提供蜜罐客户端系统 208 的更详细视图。蜜罐客户端系统 208 包含控制服务器 220。控制服务器 220 用于控制多个蜜罐挖掘器 (honey miner)222，其经配置以访问网站并模仿人类浏览者的行为以尝试检测网站上的恶意代码。蜜罐挖掘器 222 可以是被动蜜罐挖掘器或主动蜜罐挖掘器。被动蜜罐挖掘器类似于上述的网络爬行器。然而，不同于仅访问网站并报告从所述站点可获得的 URL 链接的上述网络爬行器，被动蜜罐挖掘器可经

配置以下载页面内容并将其传回控制服务器 220 以用于插入到 URL 数据库 180 中。蜜罐挖掘器 222 可以是单一机器上的软件模块，或者其每一者可实施在单独计算装置上。

[0071] 在一个实施例中，每一控制服务器可控制 17 个被动蜜罐挖掘器 222。控制服务器 220 可从 URL 数据库 180 提取或接收需要额外信息以便完全分析或归类的 URL。控制服务器 220 将所述 URL 提供到挖掘器，挖掘器又检阅 URL 并存储收集的数据。当被动挖掘器 222 完成特定 URL 时，其可从其控制服务器 222 请求另一 URL。在一些实施例中，挖掘器 222 可经配置以跟随 URL 内容上的链接，使得除了访问由控制服务器 220 指定的 URL 之外，挖掘器还可访问其链接到所述 URL 的内容。在一些实施例中，挖掘器 222 可经配置以相对于每一原始 URL 挖掘到指定深度。举例来说，挖掘器 222 可经配置以向下挖掘穿过四层网络内容，然后从控制服务器 220 请求新的 URL 数据。

[0072] 在其它实施例中，控制服务器 220 可经配置以控制主动蜜罐挖掘器 222。与仅访问网站并存储站点上呈现的内容的被动蜜罐挖掘器相比，主动蜜罐挖掘器 222 可经配置以访问 URL 并运行或执行在站点上识别的内容。在一些实施例中，主动蜜罐挖掘器 222 包含实际的网络浏览软件，其经配置以访问网站并经由浏览器软件接入网站上的内容。控制服务器 220（或蜜罐挖掘器本身 222）可经配置以在其执行其访问的网站上的内容时监视蜜罐挖掘器 222 的特性。在一个实施例中，控制服务器 220 将记录由于执行所访问网站上的应用程序或内容而由蜜罐挖掘器访问的 URL。因此，主动蜜罐挖掘器 222 可提供更准确地跟踪系统行为并发现先前未识别出的利用（exploit）的方式。因为主动蜜罐挖掘器将其本身暴露于可执行内容的危险，所以在一些实施例中主动蜜罐挖掘器 222 可位于沙盒（sandbox）环境中，其提供一组受到紧密控制的资源用于客人程序（guest program）在其中运行，以便保护其它计算机免于可能由恶意内容造成的危险。在一些实施例中，沙盒可采用模拟操作系统的虚拟机的形式。在其它实施例中，沙盒可采用与网络隔离的实际系统的形式。可通过实时跟踪对沙盒机器上的文件系统做出的改变来检测反常行为。在一些实施例中，由主动蜜罐挖掘器 222 执行的代码可能会引起运行所述挖掘器的机器由于嵌入在网页内容中的恶意代码而变为不可操作。为了解决此问题，控制服务器可控制替代挖掘器，其可插手帮助完成在挖掘过程期间损坏的蜜罐挖掘器 222 的工作。

[0073] 现在参看图 11，提供已由收集系统收集的一组 URL 相关数据的实例。尽管提供所收集数据的特定实例，但所属领域的技术人员将了解，除了此实例中提供的数据之外还可收集其它数据。所收集数据中包含针对 URL 的 IP 地址 230。IP 地址 230 可用于识别正代管同一 IP 地址下或同一服务器上的可疑内容的多个域的网站。因此，如果具有恶意内容的 URL 被识别为来自特定 IP 地址，那么可针对具有相同 IP 地址的其它 URL 挖掘 URL/ 内容数据库 180 中的数据的其余部分，以便对其进行选择和对其进行更仔细的分析。所收集 URL 数据还可包含 URL 232，如图 11 中的第二列指示。在使用例如上述蜜罐客户端过程的挖掘过程收集数据的例子中，URL 232 可常包含来自相同网域的各种页面，因为挖掘器可能经配置以爬行通过网站中的所有链接。所收集数据还可包含针对特定 URL 的页面内容 234。因为 URL 的内容可呈图形、文本、应用程序和 / 或其它内容的形式，所以在一些实施例中，存储此 URL 数据的数据库可经配置以将页面内容存储为数据记录中的二进制大对象（blob）或应用程序对象。然而，由于某些网页只含有文本，因此页面内容 234 也可存储为文本。在一些实施例中，收集例行程序可经配置以确定 URL 是否含有可执行内容。在这些例子中，所收

集数据的所得数据集可包含 URL 在其页面代码内是否具有可执行内容 236 的指示。此信息可稍后用于从具有候选数据的 URL/ 内容数据库 180 中选择数据以供分析。

[0074] 图 12 是说明来自图 7 的记分和归类模块 186 的方框图。在一个实施例中，记分和归类模块 168 包含属性数据库 320、经处理网页属性数据库 324、定义数据库 326、静态内容分类模块 328 以及内容记分模块 330。在一个实施例中，记分和归类模块 186 包含活动分析模块 332。内容分析模块 322 接收来自 URL 数据库 180 的一个或一个以上候选 URL 并从属性数据库 320 中识别其发现的与每一候选 URL 相关联的属性。每一 URL 的属性的值和 / 或计数存储在经处理网页属性数据库 324 中。静态内容分类模块 328 基于来自定义数据库 326 的定义查询经处理网页属性数据库 324 以将类别与候选 URL 相关联。内容记分模块 330 可进一步将记分与每一 URL 相关联，所述记分可用于进一步归类或改变由静态内容分类模块 328 识别的类别。在一个实施例中，内容记分模块 330 可识别候选 URL 以供活动分析模块 332 进行处理。活动分析模块 332 下载和执行任何活动内容以识别与 URL 相关联的行为属性。接着可将这些属性提供到内容记分模块以进一步归类候选 URL，例如改变其类别或添加额外类别。

[0075] 举例来说，由内容分析模块 322 处理的 URL 可得到“恶意”类别。内容记分模块 330 接着可将记分（例如，低分）与 URL 相关联，所述记分指示 URL 不是恶意的。为了解决，内容记分模块 330 可将 URL 作为候选 URL 提供到活动分析模块 332 以识别更多属性或行为记分，其可由内容记分模块 330 使用以确定“恶意”类别是否适当。

[0076] 属性数据库 320 包含可用于归类网页的关键词、常规表达式以及其它网页属性。属性也可以是与网页相关联的值，例如 HTTP 请求标头数据或与网页相关联的其它元数据。举例来说，属性可包含将在文档中识别的例如“<java 脚本 >”“< 对象 >”的关键词、例如“数据 = .*\.txt”的常规表达式（例如，关键词“数据 =”之后是任意长度的字符串，之后是“.txt”），或来自 HTTP 标头的数据的内容类型。图 13A 是属性数据库的实例，其包含属性和识别属性类型的额外字段，例如关键词或常规表达式。在说明性数据库中，属性 ID 字段用于提供用于每一属性的唯一（在数据库内）识别符。在其它实施例中，可使用其它合适类型的关键词。

[0077] 在一个实施例中，内容分析模块 322 接收来自 URL 数据库的已经由收集系统 182 识别的候选 URL。内容分析模块接收内容和与 URL 相关联的其它数据（例如 HTTP 标头），并识别属性数据库 320 中与候选网页相关联的一个或一个以上属性，且将与那些属性相关的数据存储在经处理网页属性数据库 324 中。内容分析模块 322 可接收来自 URL 数据库的候选网页的内容或其本身可下载数据。在一个实施例中，蜜罐客户端模块 208 获得并存储 URL 数掘库中每一候选网页的内容。在另一实施例中，作为针对属性处理网页的一部分，内容分析模块 322 下载候选网页的内容。

[0078] 大体上，属性数据库 320 存储属性和充足信息以识别与网页相关联的属性。举例来说，针对关键词或常规表达式属性，属性数据库 320 可存储关键词或常规表达式。相比之下，经处理网页属性数据库 324 可存储由内容分析模块 322 发现与每一网页相关联的关键词或常规表达式的计数。对于常规表达式，取决于实施例，可将匹配表达式的计数或匹配表达式本身或所述两者存储在经处理网页属性数据库 324 中。举例来说，对于特定网页，经处理网页属性数据库 324 可能存储值 3，其指属性“<java 脚本 >”在页面中出现的次数，值 0，

指属性“< 对象 t>”出现的次数，以及“data = http://www.example.url/example.txt.”，指常规表达式属性“数据 = .*\.txt.”。

[0079] 图 13B 说明经处理网页属性数据库 324 中的表的一个实施例，其中图 13A 的实例属性已经相对于若干网页经处理。在说明的实施例中，数据库包含两个表，一个将 URL 与唯一（在数据库内）识别符相关，第二个将 URL 识别符与同所述 URL 相关联的属性相关。在说明的实施例中，表包含针对与 URL 相关联的网络内容数据的每一属性的条目或行。在一个实施例中，数据库还包含针对对应于关键词属性的每一属性 /URL 的数字值，其指示在网页中发现特定属性的次数。数据库，例如在 URL/ 属性表中，还可包含匹配于 URL 的常规表达式属性的实际表达式。在一个实施例中，可在页面主体中和标头或其它元数据中搜索关键词属性。在一个实施例中，仅搜索页面主体。在又一实施例中，属性可与例如属性数据库 320 中的数据相关联，其指示在识别网页中的属性的过程中应处理哪些数据。

[0080] 在一个实施例中，静态内容分类模块 328 存取网页属性数据库 324 并对一个或一个以上网页的属性与来自定义数据库 326 的定义进行比较。当网页匹配于特定定义时，以与所述定义相关联的一个或一个以上类别识别所述网页。在一个实施例中，这些类别存储在与 URL 相关联的 URL 数据库中。在一个实施例中，根据网页的一个或一个以上属性表达每一定义。在一个实施例中，定义表达为与一个或一个以上所述属性相关的一阶逻辑运算。在一个实施例中，定义的项包含网页属性之间或属性与值（包含常数值）之间的比较。举例来说，定义可能包含表达式，例如“属性 _1”=“属性 2”AND “属性 _3”的发生 > 5。除了比较之外，项可包含对网页属性的其它运算，例如算术、字符串或任何其它合适的计算表达式。举例来说，简单的定义可以是“data =，*\.txt”=“data = xyx333.txt”，其匹配于具有字符串“data = xyx333.txt”（匹配于常规表达式属性“data =，*\.txt”）作为其内容的一部分的任何网页。更复杂的定义可包括对所述项的逻辑运算。此类逻辑运算可包含 AND、OR、NOT、XOR、IF-THEN-ELSE，或对属性的常规表达式匹配。在一个实施例中，定义还可包含或对应于数据库查询表达式，例如标准 SQL 数据库比较函数和逻辑运算。在一个实施例中，定义可包含可执行代码，例如可执行程序的脚本或引用或至少部分地确定 URL 的分类的脚本。图 13C 说明根据一个实施例的定义数据库 326 的示范性部分。如本文使用，类别可指任何类型的分类。举例来说，类别可仅仅是指示针对 URL 应执行进一步处理或分析以识别 URL 的类别的分类。

[0081] 在一个实施例中，内容记分模块 330 进一步分析网页并向网页指派与一个或一个以上类别相关联的记分。在一个实施例中，记分可基于在网页中发现关键词的次数的加权组合。在一个实施例中，权数存储在与对应属性相关联的属性数据库中。

[0082] 在另一实施例中，记分可基于关于网页的 URL 的信息来确定。举例来说，可基于因特网地址和 / 或域名向特定者指派记分。数据库可向整个子网络指派记分（例如，匹配于 128.2.*.* 的所有地址可具有特定记分）。此类网络或子网络帮助将网站识别为位于特定国家或具有特定服务提供商。已发现这对记分有用，因为由于不同的法律或执法不严，某些国家和服务提供商已经关联于特定类型的网络内容。网络或子网络的记分系统可基于具有特定类别的特定网络或域中的 URL 的相对数目。举例来说，如果 URL 数据库 180 中针对特定网络的 URL 的 95％被分类为恶意的，那么可给予新 URL 高分。在一个实施例中，具有高于阈值的记分的 URL 被识别为具有一类别，例如恶意的，而无论通过对网页的内容分析识别

的类别如何或除了所述类别以外。在一个实施例中，向每一 URL 指派与不同类别相关联的多个记分，且用 URL 识别对应于高于给定阈值的每一记分的类别。在一个实施例中，采用多个阈值。举例来说，基于记分自动分类具有高于一个阈值的记分的 URL。在一个实施例中，将具有低于第一阈值但高于第二阈值的记分的 URL 传送给人类分析员以用于分类。在一个实施例中，内容记分模块 330 将此类 URL 传送到活动分析模块 332 以用于额外分析。

[0083] 一个实施例可包含记分和归类系统，例如标题为“用于控制对因特网站点的接入的系统和方法”(" System and method for controlling access to internet sites,") 的第 6,606,659 号美国专利中说明，所述文档的全文以引用的方式并入。

[0084] 在一个实施例中，活动分析模块 332 执行网页的活动内容以识别其行为属性。这些属性可接着用于为网页记分和分类。在一个实施例中，静态内容分类模块 328 和内容记分模块 330 中的一者或一者以上识别 URL 以用于由活动分析模块 332 处理。在接收到候选 URL 之后，活动分析模块 332 可将与一个或一个以上行为属性（例如，比如“写入到注册表”的属性）相关联的行为记分或数据提供到内容记分模块以用于进一步归类。

[0085] 图 14 是说明来自图 7 的训练模块 184 的一个实施例的方框图。在一个实施例中，训练模块包含分析任务分配模块 352，其识别针对其需要额外类别的具有例如活动内容等内容的网页或 URL。在一个实施例中，收集模块 190 识别具有活动内容的 URL。在另一实施例中，例如安全性研究员等外部源识别具有已经识别出具有一个或一个以上类别（例如，键盘记录程序、病毒、恶意内容、蠕虫等）的活动内容的特定 URL。在一个实施例中，这些可存储在 URL 数据库 180 中。在一个实施例中，任务分配模块 352 维持此类 URL 的数据库（未图示）。在一个实施例中，任务分配模块 352 数据库维持针对这些 URL 的优先权，并基于优先权将其呈现给分析员。

[0086] 属性识别模块 354 识别网页的属性和基于所述属性的定义，所述属性和定义对网页进行归类。在一个实施例中，属性识别模块 354 为人类分析员提供使用记分和分类模块 186 向 URL 应用特定规则或定义的界面。另外在一个实施例中，属性识别模块 354 可提供一界面，供分析员将 URL 识别为供图 10 的活动分析模块 332 执行 URL 的行为分析的候选，以便从活动分析模块 332 接收回用于将 URL 分类的额外数据。属性识别模块 354 接着可将此数据提供给分析员。在一个实施例中，分析员分析来自记分和分类模块 186（包含活动分析模块 332）的 URL 数据以帮助识别将 URL 以及（在可能时）涉及类似分类的内容的其它 URL 适当分类的属性和定义。在一个实施例中，属性识别模块 354 将这些新识别的属性和定义提供到数据库更新模块 356，数据库更新模块 356 将新定义和属性存储到属性数据库 320 和定义数据库 326。

[0087] 图 15 是说明来自图 12 的活动分析模块 332 的一个实施例的方框图。在一个实施例中，活动分析模块 332 包含沙盒模块 370，在沙盒模块 370 中如将在典型工作站 116 上所发生的那样下载 URL 和执行任何活动内容。沙盒模块 370 以透明方式监视计算机的状态以识别网络内容的行为，所述行为影响例如新产生进程、网络接入、处理器使用、存储器使用、系统资源使用、文件系统存取或修改以及注册表存取或修改中的一者或一者以上。

[0088] 行为分析模块 372 将来自沙盒模块的所监视动作与特征化所监视动作的列表、数据库或规则进行比较。在一个实施例中，这些特征化定义 URL 的属性，所述属性随后由图 12 的静态内容分类模块 328 分析。在另一实施例中，活动记分分类模块 374 可使用与行为属

性相关联的记分来确定 URL 的记分。在一个实施例中，记分是这些属性的加权记分。此记分可用于将 URL 分类或将其传送到内容记分模块以用于分类。在另一实施例中，将例如来自定义数据库 332 的规则或定义应用于 URL 的行为属性（且在一个实施例中，经处理网页属性 324）以识别与 URL 相关联的一个或一个以上类别。

[0089] 使用和操作的方法描述

[0090] 取决于实施例，本文描述的方法的动作或事件可以不同顺序执行、可合并，或可完全省略（例如，并非所有动作或事件对于实践所述方法都是必要的），除非正文中另有具体且清楚的陈述。另外，本文描述的方法可包含额外的动作或事件，除非正文中另有具体且清楚的陈述。而且，除非另有清楚陈述，否则可例如通过中断处理或多个处理器同时执行而不是顺序执行动作或事件。

[0091] 如上文结合图 3 论述，在一些实施例中，网关服务器模块 120 可经配置以基于经归类 URL 数据库 146 中存储的数据来控制对特定 URL 的接入。图 16 是描述网关服务器模块处理来自工作站 116 的请求的实施例的流程图。

[0092] 在方框 1200，工作站 116 从因特网 112 请求 URL。在方框 1202，此请求在因特网网关处被拦截并被转发到网关服务器模块 120。在方框 1204，查询经归类 URL 数据库 146 以确定所请求 URL 是否存储在数据库 146 中。如果发现所请求 URL 是数据库中的一份记录，那么过程继续移动到方框 1206，其中所述过程分析 URL 记录以确定 URL 的类别是否是应针对工作站用户阻止的类别。如果所述类别被阻止，则过程跳转到方框 1212 且请求被阻止。然而如果所述类别未被阻止，则在方框 1208 处允许所述请求。

[0093] 如果在方框 1204 处并未发现所请求 URL 是经归类 URL 数据库 146 中的记录，则系统前进到方框 1210。在方框 1210 处，系统确定如何处理未经归类内容。在一些实施例中，系统可利用策略模块 142 来做出此确定。如果网关服务器模块 120 经配置以阻止针对未经归类内容的请求，则过程移动到方框 1212，且阻止请求。另一方面，如果模块经配置以允许这些类型的未经归类请求，则过程移动到方框 1208，其中允许所述请求前进到因特网 112。

[0094] 在一些实施例中，对 URL 数据的请求可导致新记录添加到记录数据库 144。这些记录可稍后传送到数据库管理模块 114 供进一步分析。现在参看图 17，提供描述网关服务器模块可借以处理 URL 请求的过程的另一流程图。在方框 1300，网关服务器模块 120 接收针对 URL 的请求。如上所述，此请求可来自工作站 116。在方框 1302，接着将 URL 与经归类 URL 数据库 146 进行比较，且系统在方框 1304 确定所请求 URL 是否在经归类 URL 数据库中。

[0095] 如果 URL 已经在经归类 URL 数据库 146 中，则过程跳转到方框 1308。然而如果在经归类 URL 数据库 146 中没有发现所请求 URL，则过程移动到方框 1306，其中将 URL 插入到未经归类 URL 数据库 147 中。（在一些实施例中，记录数据库 144 和未经归类 URL 数据库 147 可以是同一数据库。）在将 URL 插入到数据库中之后，方法前进到方框 1308。在方框 1308，检查策略数据库以获得关于如何处理所接收 URL 的指令。一旦策略模块 142 已经被检查，就在方框 1310 更新记录数据库 144 以记录 URL 已经被请求。在更新记录数据库 144 之后，如果策略数据库许可工作站 116 接入 URL，则过程移动到方框 1314，且将 URL 请求发送到因特网 112。然而如果策略数据库不允许所述请求，则过程跳转到方框 1316 且阻止请求。

[0096] 在一些实施例中，网关服务器模块 120 可执行收集活动以减少数据库管理模块

114 的收集系统 182 的负担。图 18 提供网关服务器收集模块 140 用于收集关于未经归类 URL 的数据的系统的实例。在方框 1400，网关服务器模块接收针对 URL 的请求。接着，在方框 1402，将所请求 URL 与经归类 URL 数据库进行比较。如果在方框 1404 系统确定所请求 URL 在 URL 数据库中，则过程移动到方框 1410，其中依据 URL 如何被归类而将请求转发到因特网 112 或阻止请求。

[0097] 如果所请求 URL 不在经归类 URL 数据库 146 中，则过程移动到方框 1406，其中将 URL 发送到网关收集模块 140。接着在方框 1408，收集模块 140 收集关于所请求 URL 的 URL 数据。在一些实施例中，此数据可存储在未经归类 URL 数据库 147 中。或者，此数据可简单地经由因特网 112 转发到数据库管理模块 114。一旦数据已被收集并存储，则过程移动到方框 1410，其中基于策略模块 142 中指示的策略而允许或阻止 URL 请求。

[0098] 如先前论述，未经归类 URL 数据可从网关服务器模块 120 发送到数据库管理模块 114 供进一步分析，使得 URL 可经归类并添加到经归类 URL 数据库 146。然而，因为未经归类数据的量有时很大，以至于或许不可能在无损于准确性或速度的情况下将所有接收的数据归类。因此，在一些例子中，可能需要识别未经归类数据内的最有可能对工作站 116 和网络 110 引起威胁的候选 URL。

[0099] 图 19 提供用于识别候选 URL 供进一步分析的方法的实例。所述方法以将 .URL 接收到数据库模块 114 的收集系统 182 中开始。在方框 1502，预处理 URL 或应用程序以确定其是否携带已知的恶意数据元素或数据签名。接着在方框 1504，如果系统确定 URL 包含已知的恶意元素，则过程跳转到方框 1514，其中将 URL 标记为候选 URL 并将其发送到训练系统 184 供进一步分析。如果在方框 1504 中对 URL 的初始分析没有显示恶意元素，则过程移动到方框 1506，其中将 URL 添加到可能的候选 URL 的数据库。接着在方框 1508，数据挖掘模块 192 经配置以基于预先配置的条件（例如，攻击串、病毒签名等）从源 194（可能的候选 URL 的数据库是其中之一）选择 URL。接着在方框 1510 将包含所有数据源 194 的数据集发送到数据挖掘模块 192，其中在方框 1512 通过数据挖掘模块 192 分析每一 URL。如果 URL 满足所定义的预先配置的条件，则过程移动到方框 1514，其中将 URL 标记为候选 URL 并将其转送到记分 / 分类系统 186 供额外分析。然而如果 URL 不满足为将其转换为候选 URL 而指定的条件，则方法前进到方框 1516 且不将 URL 标记为候选。尽管在 URL 候选分类的上下文中描述此实施例，但所属领域的技术人员将容易了解，可使用上述过程类似地分析应用程序并将其标记为候选。

[0100] 如上论述，收集并分析因特网数据以确定其是否包含有害的活动内容的难点之一就是必须收集和分析的数据的量。在又一实施例中，数据挖掘模块 192 可用于通过收集大量相关数据来解决这些问题以有效且高效地利用系统资源。现在参看图 20，提供数据挖掘系统 192 的更详细的方框图。数据挖掘系统 192 可采用软件模块的形式，其运行多个异步过程以实现最大效率和输出。数据挖掘系统 192 可包含插入模块 242，其接收提供关于应如何处理输入数据的指令的配置参数。在一个实施例中，由插件模块接收的指令可采用 HTTP 协议插件的形式，其为数据挖掘系统 192 接收 URL 数据并基于由数据挖掘系统对 URL 数据实施的各种 HTTP 相关指令分析和补充数据提供参数。在另一实施例中，可朝挖掘例如 FTP、NNTP 或某种其它数据形式的某种其它协议的方向来调整插件。

[0101] 也可用于实施被动蜜罐客户端的数据挖掘系统 192 还包含调度程序 248 的库 246。

调度程序 248 是单个单个的异步处理实体，其基于输入到数据挖掘系统中的数据（用于分析）和由插件模块 242 接收的配置数据来接收任务指派。库 246 是由驱动程序 244 控制的调度程序的集合。驱动程序 244 是用于库的管理机制。驱动程序 244 可经配置以监视库 246 中的调度程序 248 的活动以确定何时将额外数据发送到库 246 中用于挖掘和分析。在一个实施例中，驱动程序可经配置以每当任何调度程序 248 空闲便将新数据单元发送到库 246 中。在一个实施例中，驱动程序 244 可用作控制服务器以管理如上文结合图 10 描述的蜜罐客户端挖掘器 222。库 246 可将数据单元传递到空闲的调度程序 248。调度程序 248 读取插件配置并根据插件 242 执行动作。

[0102] 在一个实施例中，插件模块可接收 HTTP 插件。HTTP 插件可经配置以接收呈 URL 串形式的输入数据，关于所述数据，数据挖掘系统 192 将获得额外信息，例如 URL 的页面内容、在接入 URL 时由 URL 返回的 HTTP 消息（例如，“4xx- 文件未找到”或“5xx- 服务器错误”）。插件可进一步指定网络爬行模式，其中调度程序除了收集页面内容以外还将 URL 内容内的 URL 链接添加到待分析的 URL 数据集。

[0103] 图 21 是说明在数据库管理模块 114 内将 URL 归类的方法 2000 的一个实施例的流程图。方法 2000 开始于方框 2002，其中开发可用于将网页归类的属性。在一个实施例中，训练模块 184 用于开发属性数据库 320 中的属性。在一个实施例中，开发属性包含开发定义（例如与一个或一个以上属性相关的表达式），并将定义存储在定义数据库 326 中。接着在方框 2004 处，识别网页以用于内容分析。在一个实施例中，收集模块 190 识别网页以用于内容分析。在一个实施例中，识别具有活动内容的属性或其它指示的网页以用于内容分析。

[0104] 移动到方框 2006，内容分析模块 322 识别与每一所识别网页相关联的一个或一个以上属性。下文参看图 22 更详细描述方框 2006 的功能。前进到方框 2010，静态内容分类模块 328 至少部分地基于属性识别具有一个或一个以上类别的网页。在一个实施例中，静态内容分类模块 328 将来自定义数据库 326 的定义与每一网页的属性进行比较以识别其属性。在一个实施例中，类别包含指示网页是否与活动内容相关联的那些类别。在一个实施例中，类别包含指示与网页相关联或由网页引用的活动内容的类型（例如，恶意、网络钓鱼站点、键盘记录程序、病毒、蠕虫等）的那些类别。在一个实施例中，活动内容包含在网页的主体中。在一个实施例中，在网页的链接或活动 X 对象元素中引用活动内容。在一个实施例中，活动内容包含交互式“网络钓鱼”站点，其包含往往误导用户提供证书或其它敏感、私人或个人信息的内容。在一个实施例中，记分模块 330 进一步为网页记分和分类。移动到方框 2012，将与网页相关联的类别存储在 URL 数据库中。在一个实施例中，图 7 的上载下载模块 178 将新 URL 类别分布到一个或一个以上网关服务器模块 120 或工作站 116（两者均见图 1）。在一个实施例中，方法 2000 的一个或一个以上方框（例如，方框 2006-2012）也可在收集模块 190 接收到新 URL 时连续执行。在一个实施例中，方法 2000 的一个或一个以上方框（例如，方框 2006-2012）可周期性执行。

[0105] 图 22 是说明执行图 21 的方框 2006 的功能的方法的一个实施例的流程图。方法开始于方框 2020，其中内容分析模块 322 接收 URL 数据库 180 中的网页 URL 的列表。在一个实施例中，收集模块 190 提供候选 URL 的列表。接着在方框 2022，针对每一 URL，内容分析模块 322 接收下载的网页内容。在一个实施例中，收集模块 190 下载内容并将其存储在

URL 数据库 180 中，内容分析模块 322 从 URL 数据库 180 中存取所述内容。在另一实施例中，内容分析模块 322 下载并处理内容。移动到方框 2024，内容分析模块 322 从属性数据库 320 存取属性。接着在方框 2026，内容分析模块 322 至少部分地基于每一网页的内容而识别与每一网页相关联的属性。在一个实施例中，内容分析模块 322 扫描内容以识别来自属性数据库 320 的字符串、关键词和常规表达式属性。在一个实施例中，内容分析模块 322 还可在扫描属性之前和 / 或之后解码内容。举例来说，内容分析模块 322 可在扫描之前解码例如 URL 的 URL 编码部分或十六进制编码网络地址的网络内容，以帮助防止通过编码或部分编码关键词而将关键词隐藏。前进到方框 2028，内容分析模块 322 将与每一网页相关联的所识别属性存储在经处理网页属性数据库 324 中。

[0106] 图 23 是说明执行图 21 的方框 2010 的功能的方法的一个实施例的流程图。方法开始于方框 2042，其中静态内容分类模块 328 从定义数据库 326 存取指示网页类别的定义。接着在方框 2044，针对每一定义，静态内容分类模块 328 对照经处理网页属性数据库 324 识别与每一定义相关联的一个或一个以上查询。在一个实施例中，查询包括 SQL 查询。

[0107] 移动到方框 2046，静态内容分类模块 328 将网页属性数据库中的 URL 的属性与查询进行比较以识别匹配于查询的 URL。在一个实施例中，静态内容分类模块 328 通过执行所述一个或一个以上所识别数据库查询，而对照经处理网页属性数据库 324 执行比较。接着在方框 2050，静态内容分类模块 328 将任何所识别 URL 与定义进行比较以识别匹配于定义的所识别 URL 中的任一者。在一个实施例中，此比较包含使用额外可执行指令（例如 Perl 脚本）比较数据库查询的结果以识别匹配的 URL。前进到方框 2052，静态内容分类模块 328 基于定义将所识别的 URL 归类。在一个实施例中，每一定义与单一类别相关联。在另一实施例中，每一定义与每一者用 URL 来识别的若干类别相关联。在又一实施例中，定义可包含逻辑表达式，其识别一个或一个以上待用 URL 识别的类别。举例来说，if-then-else 表达式可依据 if 表达式的结果识别不同的类别。在一个实施例中，内容记分模块进一步对 URL 记分。基于记分，可用 URL 识别相同、不同或额外的类别。接着在方框 2054，静态内容分类模块 328 将每一 URL 的类别存储到经归类网页数据库。在一个实施例中，URL 数据库 180 包含经归类网页数据库。

[0108] 图 24 是说明作为识别在图 22 和 23 的方法中将 URL 归类时使用的属性的一部分来执行图 21 的方框 2002 的功能的方法的一个实施例的流程图。方法开始于方框 2062，其中图 14 的分析任务分配模块 352 接收与活动内容相关联的活动内容数据或 URL。接着在方框 2064，属性识别模块 254 识别区分与活动内容数据相关的目标 URL 与其它 URL 且识别与目标 URL 相关联的一个或一个以上类别的属性。在一个实施例中，记分和分类系统 186 用于帮助识别这些属性。另外，可识别包括一个或一个以上所述属性的定义，所述一个或一个以上属性区分与特定类别相关联的目标 URL 与不应与所述类别相关联的其它 URL。移动到方框 2068，数据库更新模块 356 将属性、定义和类别存储在属性数据库 320 和定义数据库 326 中。因此使这些经更新的属性和定义可用于使用例如图 21 说明的方法来处理 URL。

[0109] 如本文中所使用，“数据库”指存储在可由计算机存取的媒体上的所存储数据的任何集合。举例来说，数据库可指平面数据文件或结构化数据文件。而且，将认识到结合本文中所揭示的实施例描述的各种说明性数据库可实施为组合各种说明性数据库的方面的数据库，或者可将所述说明性数据库划分为多个数据库。举例来说，各种说明性数据库中的一

者或一者以上可实施为一个或一个以上关系数据库中的表。实施例可以关系数据库实施，所述关系数据库包含例如 mySQL 的 SQL 数据库、面向对象的数据库、对象关系数据库、平面义件或任何其它合适的数据存储系统。

[0110] 所属领域的技术人员将认识到，结合本文中所揭示的实施例描述的各种说明性逻辑区块、模块、电路和算法步骤可实施为电子硬件、计算机软件或两者的组合。为了清楚地说明硬件与软件的此可互换性，上文已大体在功能性方面描述各种说明性组件、区块、模块、电路和步骤。此功能性实施为硬件还是软件取决于特定应用和强加于总体系统的设计约束。所属领域的技术人员可针对每一特定应用以各种方式实施所述功能性，但此类实施方案决策不应被解释为导致偏离本发明的范围。

[0111] 结合本文揭示的实施例描述的各种说明性逻辑区块、模块和电路可用如下装置实施或执行：通用处理器、数字信号处理器 (DSP)、专用集成电路 (ASIC)、现场可编程门阵列 (FPGA) 或其它可编程逻辑装置、离散门或晶体管逻辑、离散硬件组件或其经设计以执行本文所述功能的任意组合。通用处理器可以是微处理器，但在替代方案中，处理器可以是任何常规处理器、控制器、微控制器或状态机。处理器也可实施为计算装置的组合，例如 DSP 与微处理器的组合、多个微处理器、一个或一个以上微处理器结合 DSP 核心，或任何其它此配置。

[0112] 结合本文揭示的实施例描述的方法或算法的步骤可直接以硬件实施、以由处理器执行的软件模块实施，或以两者的组合实施。软件模块可驻存在 RAM 存储器、快闪存储器、ROM 存储器、EPROM 存储器、EEPROM 存储器、寄存器、硬盘、可移除盘、CD-ROM 或此项技术中已知的任何其它形式的存储媒体中。示范性存储媒体耦合到处理器，使得处理器可从存储媒体读取信息和向存储媒体写入信息。在替代方案中，存储媒体可与处理器成为一体。处理器和存储媒体可驻存在 ASIC 中。ASIC 可驻存在用户终端中。在替代方案中，处理器和存储媒体可作为离散组件驻存在用户终端中。

[0113] 鉴于上文内容，将了解本发明的实施例通过提供处理因特网上可用的大量 URL 以识别 URL 的类别（尤其是具有活动内容的 URL）的高效方式来克服此项技术中的许多长期存在的问题。具有许多类型的活动内容的 URL 甚至对于人类分析员来说可能也难以归类，因为相关属性可埋入于可执行代码（包含脚本）中，或埋入于活动 X 组件的参数中。可经高效处理的属性和定义的使用允许通过自动过程来有效地识别活动 X 内容。此外，通过将网页的属性存储在数据库中用于稍后查询，可在识别出活动内容的新定义时基于这些存储的属性来立即将大量 URL 归类。

[0114] 尽管上述详细描述已展示、描述和指出应用于各种实施例的本发明的新颖特征，但将了解，在不脱离本发明精神的情况下，所属领域的技术人员可对所说明的装置或过程做出形式和细节上的各种省略、替代和改变。将认识到，本发明可以并不提供本文陈述的所有特征和益处的形式实施，因为有些特征可与其它特征分开地使用或实践。本发明的范围由所附权利要求书指示而不是由上述描述内容指示。在权利要求书的等效物的意义和范围内的所有改变应包含在权利要求书的范围内。

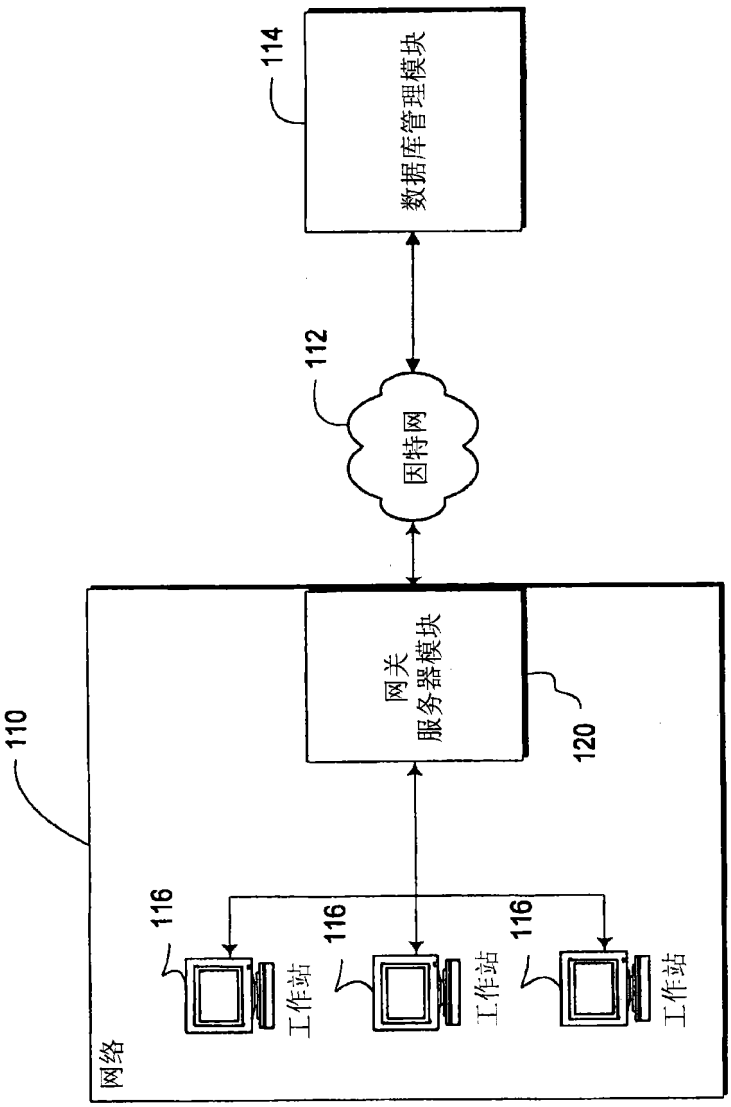


图1

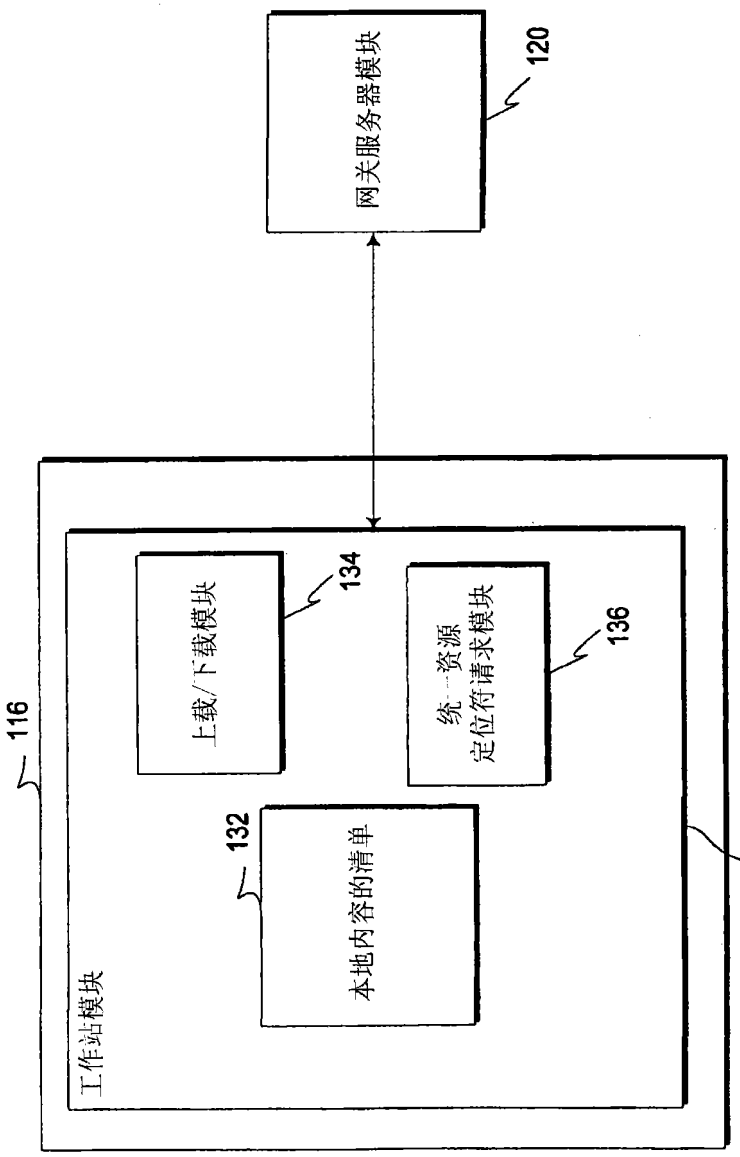

网关服务器模块
120
116
上载/下载模块
134
统一资源
定位符请求模块
136
132
本地内容的清单
130
工作站模块

图2

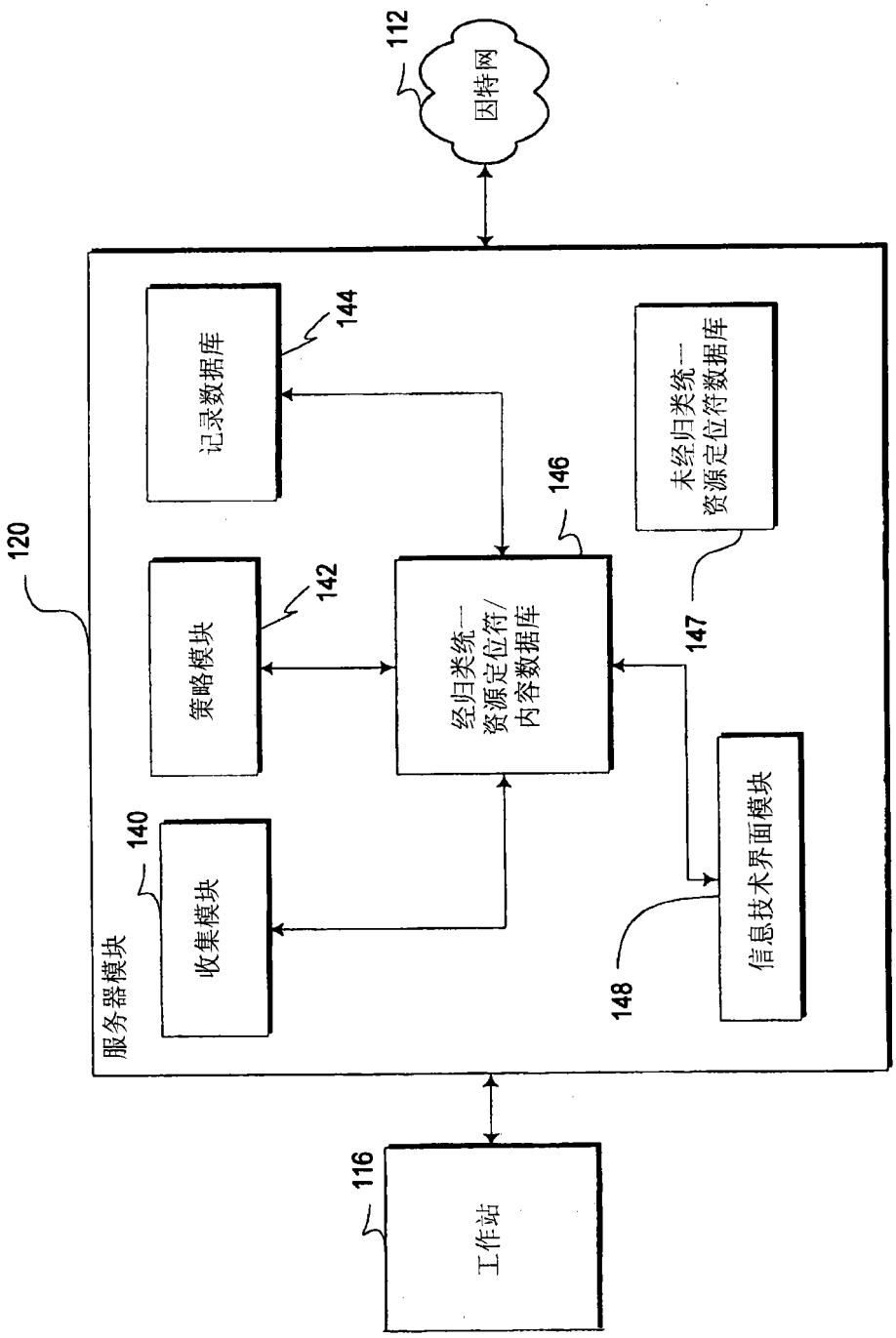

112
因特网
120
记录数据库
144
策略模块
142
经归类统一资源定位符/内容数据库
146
未经归类统一资源定位符数据库
147
收集模块
140
信息技术界面模块
148
服务器模块
工作站
116

图3

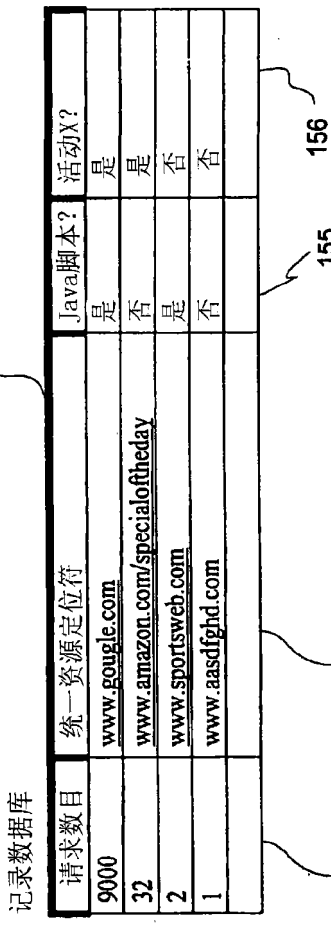

144
记录数据库
请求数目
统一资源定位符
Java脚本?
活动X?
9000
www.gougle.com
是
是
32
www.amazon.com/specialoftheday
否
是
2
www.sportsweb.com
是
否
1
www.aasdfghd.com
否
否
152
154
155
156

图4

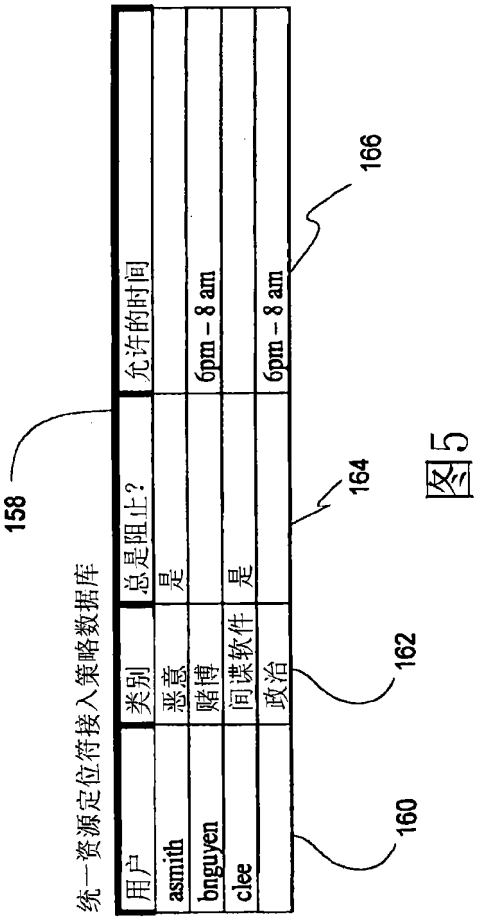

158
统一资源定位符接入策略数据库
用户
类别
总是阻止?
允许的时间
asmith
恶意
是
bnguyen
赌博
6pm – 8 am
clee
间谍软件
是
6pm – 8 am
政治
160
162
164
166

图5

经归类统一资源定位符 146

统一资源定位符	类别
http://example1.com/abc	恶意
http://example2.biz/abc	赌博
http://example4.com/abc	间谍软件

172 174

图6A

未经归类统一资源定位符 147

统一资源定位符
http://example4.com/abc
http://example6.biz/abc
http://example1.com/abc

图6B

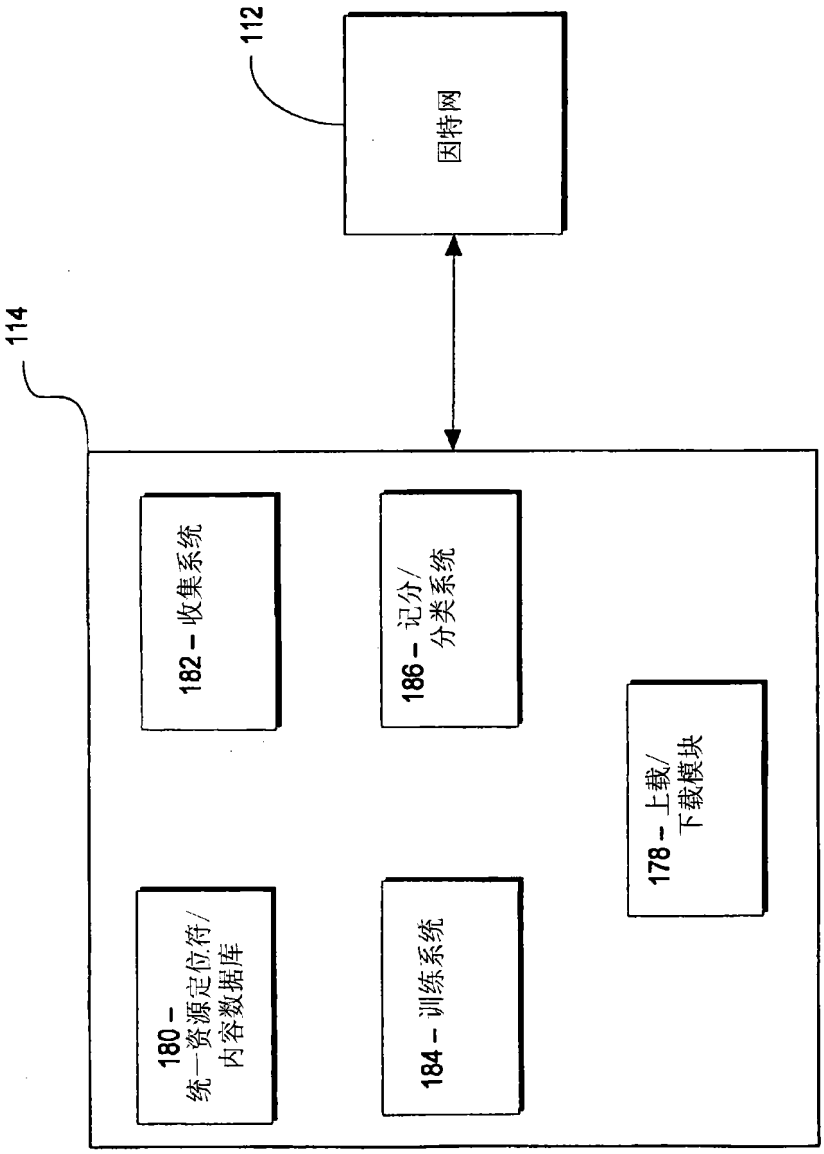

112
因特网
114
182 – 收集系统
186 – 记分/
分类系统
180 –
统一资源定位符/
内容数据库
184 – 训练系统
178 – 上载/
下载模块

图7

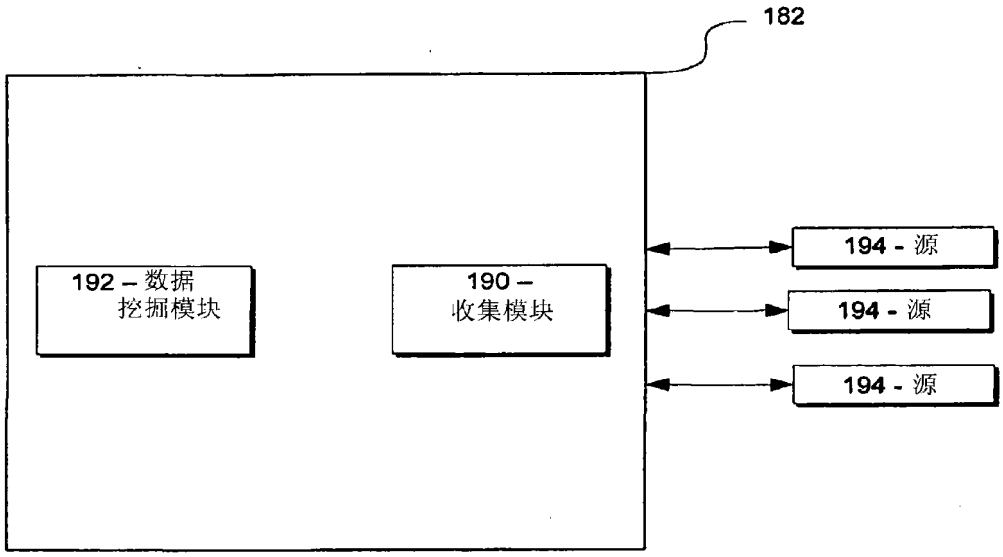

182
192 – 数据
挖掘模块
190 –
收集模块
194 - 源
194 - 源
194 - 源

图8

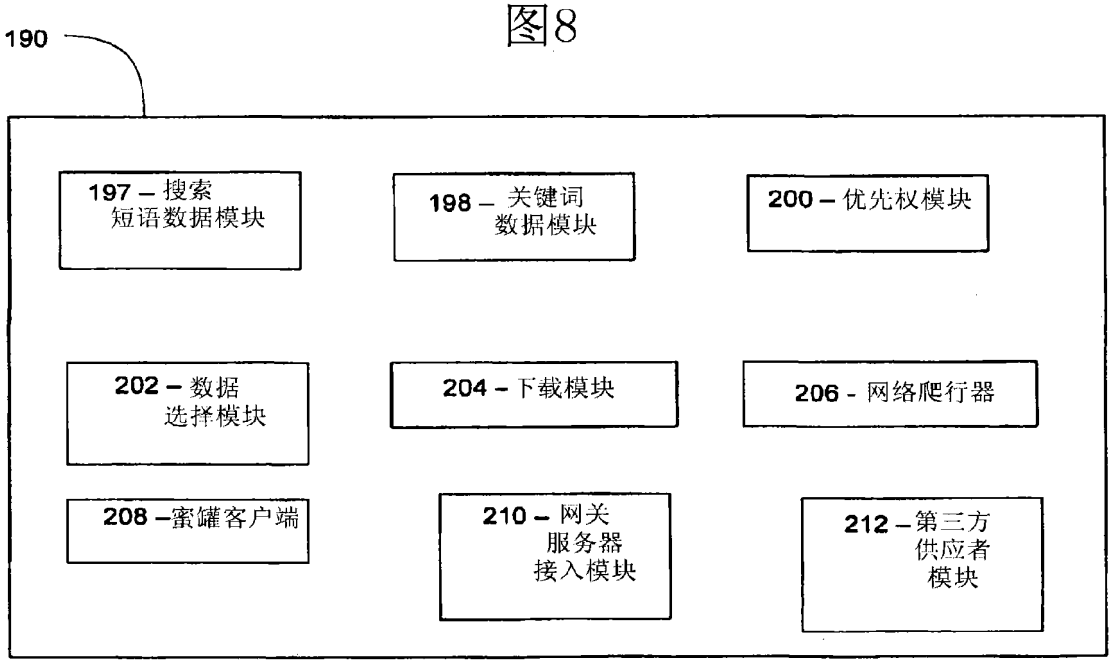

190
197 – 搜索
短语数据模块
198 – 关键词
数据模块
200 – 优先权模块
202 – 数据
选择模块
204 – 下载模块
206 - 网络爬行器
208 –蜜罐客户端
210 – 网关
服务器
接入模块
212 – 第三方
供应者
模块

图9

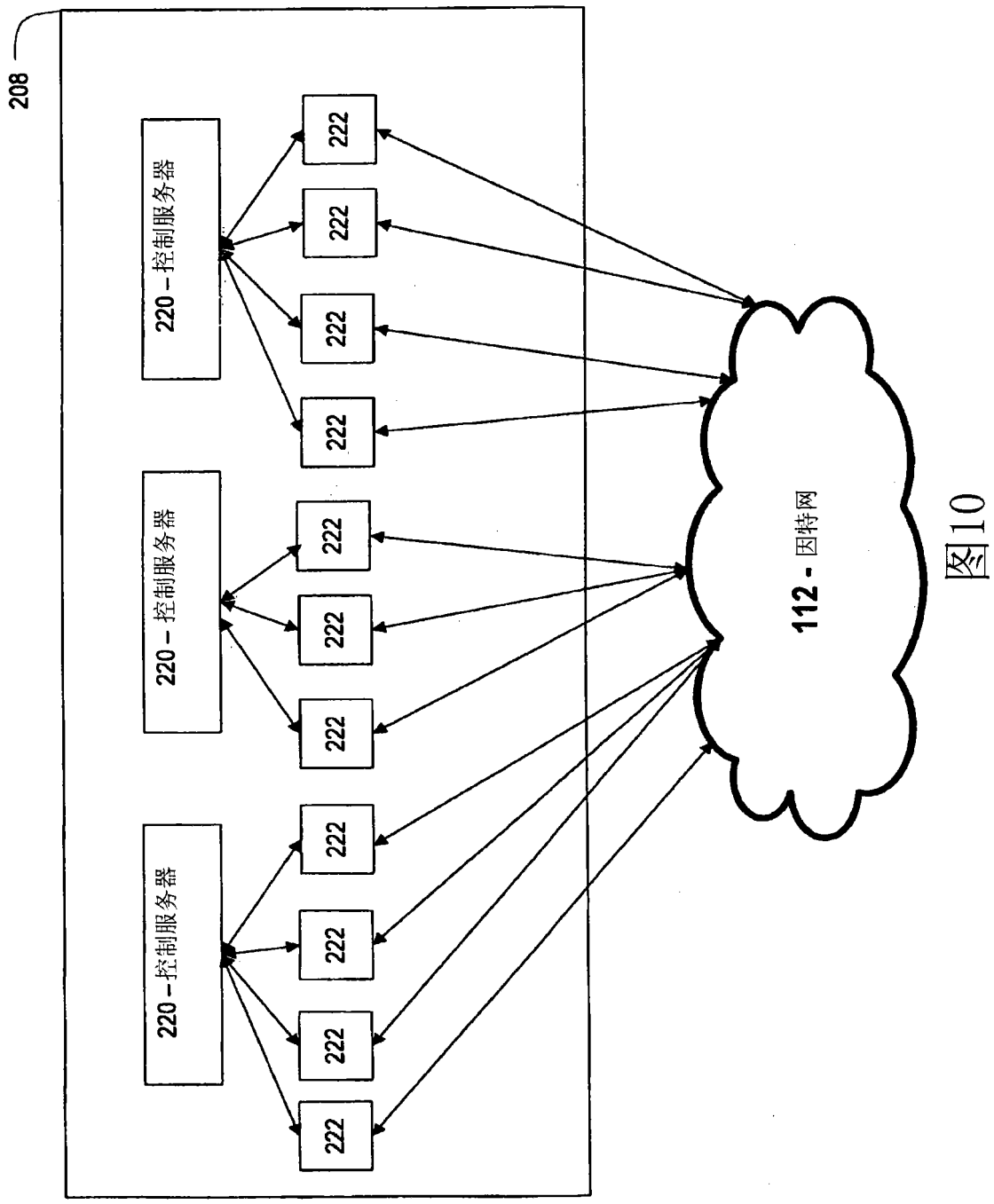


图10

由收集系统收集的统一资源定位符数据

因特网协议地址	统一资源定位符	页面内容	活动内容?
134.34.54.158	www.gougle.com	二进制大对象	是
152.68.94.129	www.amazon.com/specialoftheday	二进制大对象	是
10.42.228.233	www.sportsweb.com	二进制大对象	否
152.36.242.21	www.aasdfghd.com	文本	否
134.34.54.158	www.gougle.com/page2.html	二进制大对象	是

230 232 234 236

图11

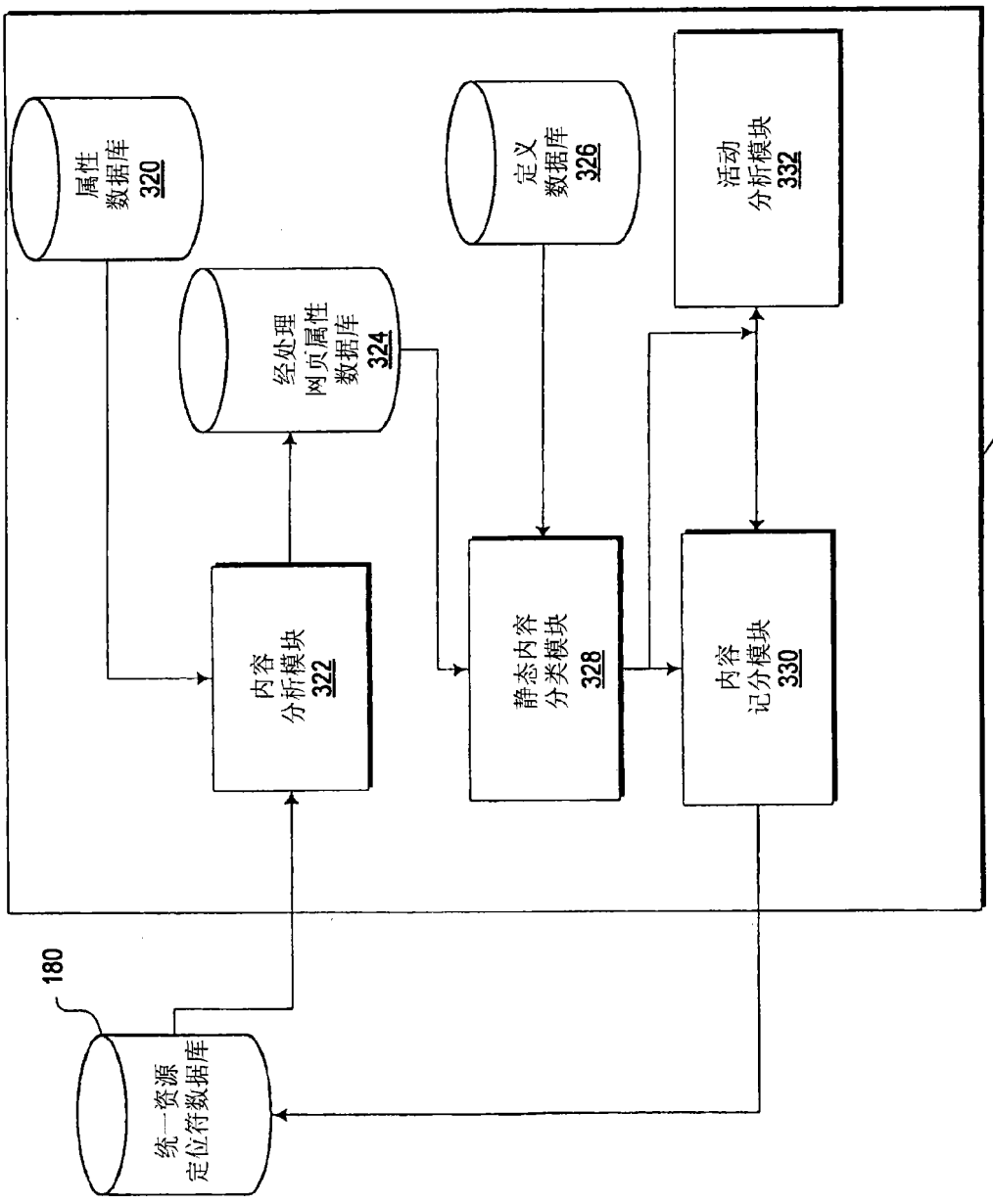

属性
数据库
320
经处理
网页属性
数据库
324
定义
数据库
326
活动
分析模块
332
内容
分析模块
322
静态内容
分类模块
328
内容
记分模块
330
186
180
统一资源
定位符数据库

图12

属性数据库

属性_识别号	属性	类型
1	〈java脚本〉	关键词
2	〈对象〉	关键词
3	数据=*\.txt	常规表达式
4	内容-类型	超文本传输协议数据

图13A

经处理网页属性数据库

统一资源定位符_识别号	统一资源定位符
1	http://www.a.com
2	http://www.b.com
3	http://www.c.com
4	http://www.d.com

统一资源定位符_识别号	属性_识别号
1	1
1	2
2	3
2	4
3	1

图13B

定义数据库

表达式	类别
内容－类型＝超文本标记语言且“〈java脚本〉”〉2	恶意
〈对象〉＝6且统一资源定位符匹配于“a. com”	赌博
“数据＝.*\.txt” ＝“数据＝xvx333. txt”	间谍软件

图13C

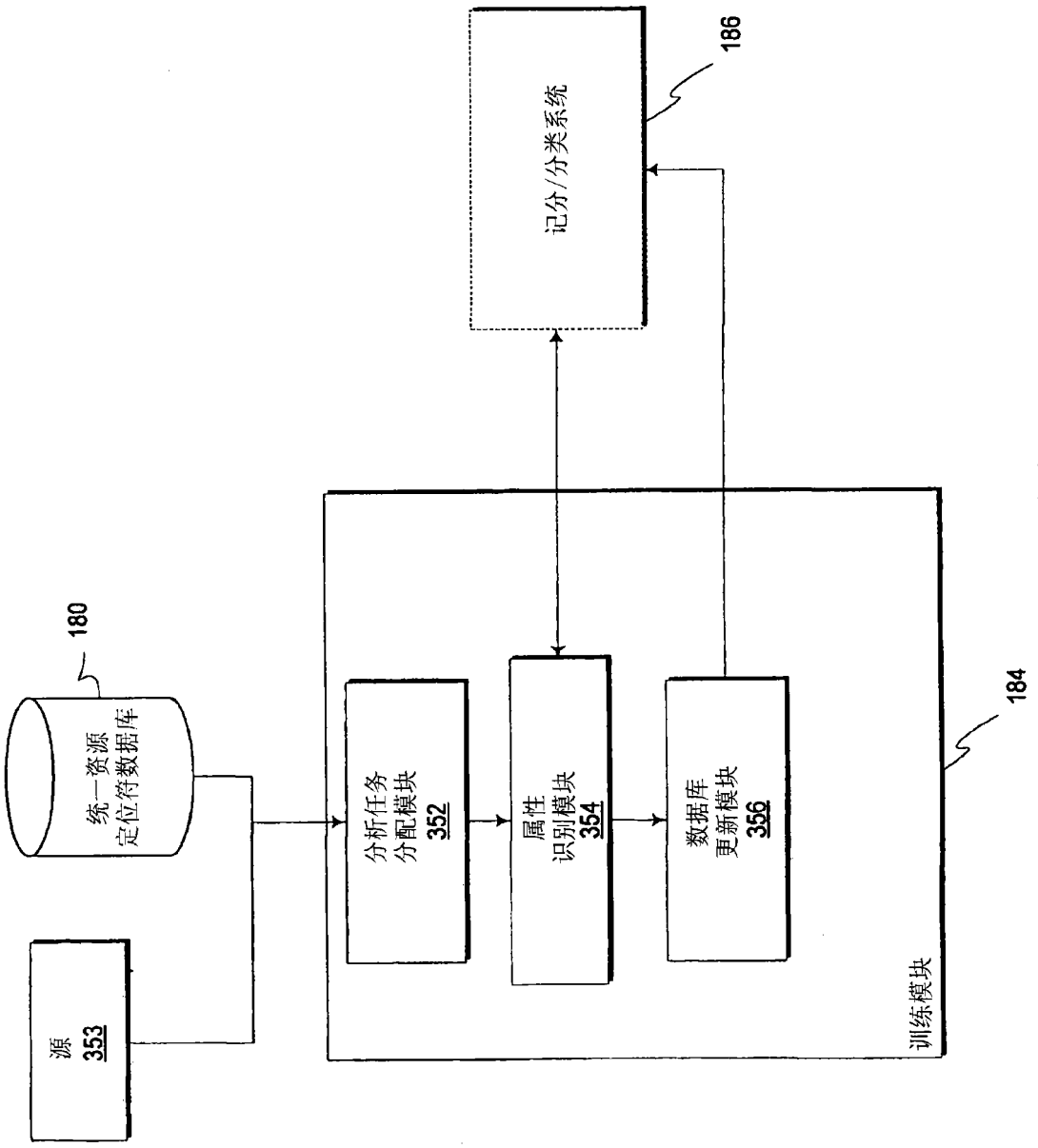

记分/分类系统
186
180
统一资源
定位符数据库
分析任务
分配模块
352
属性
识别模块
354
数据库
更新模块
356
184
源
353
训练模块

图14

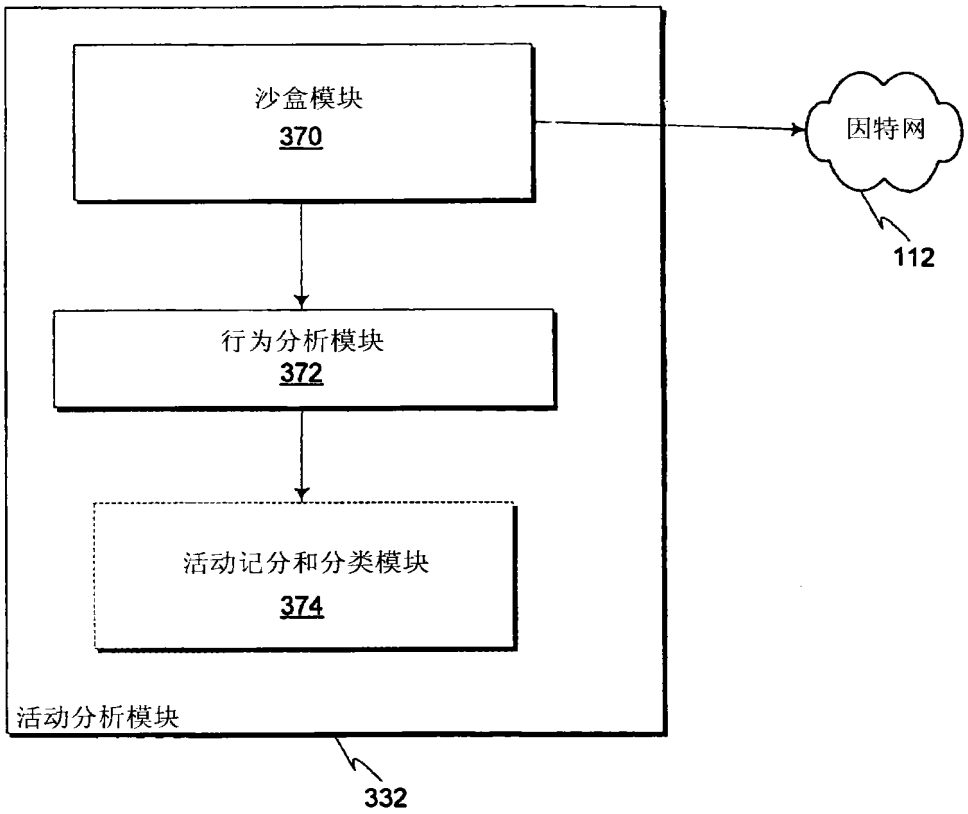


图15

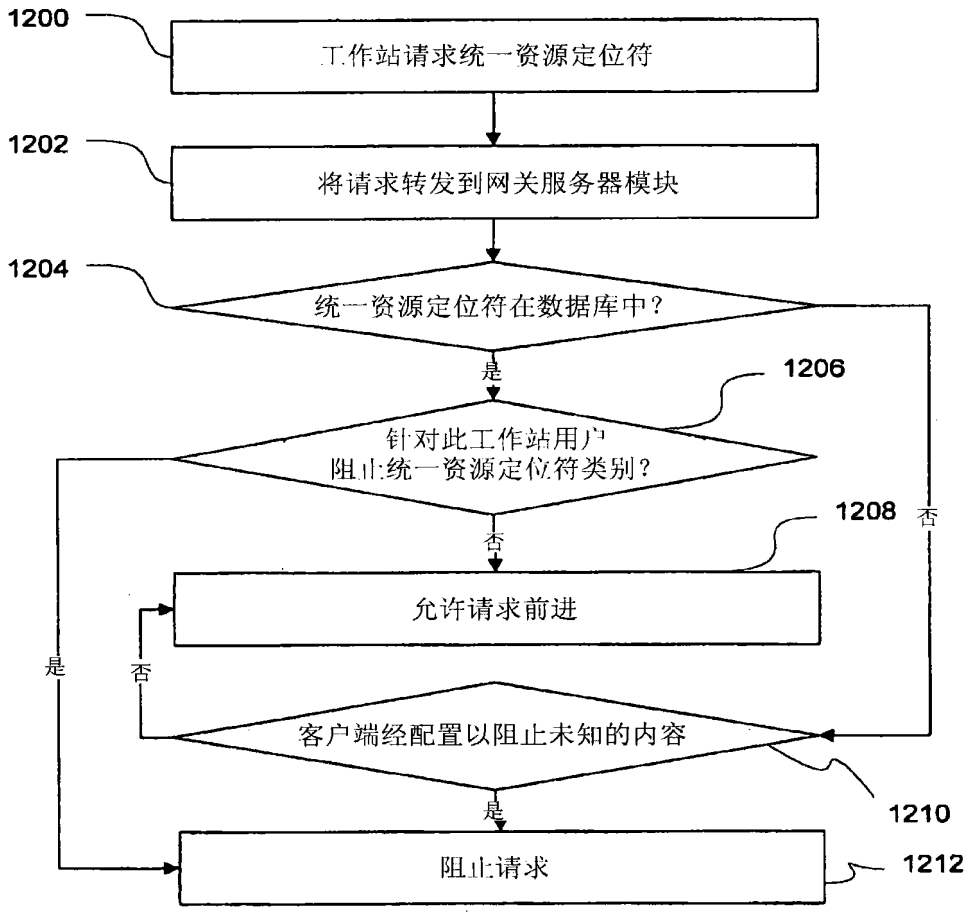

1200
工作站请求统一资源定位符
1202
将请求转发到网关服务器模块
1204
统一资源定位符在数据库中？
是
1206
针对此工作站用户
阻止统一资源定位符类别？
否
1208
否
允许请求前进
是
否
客户端经配置以阻止未知的内容
1210
是
阻止请求
1212

图 16

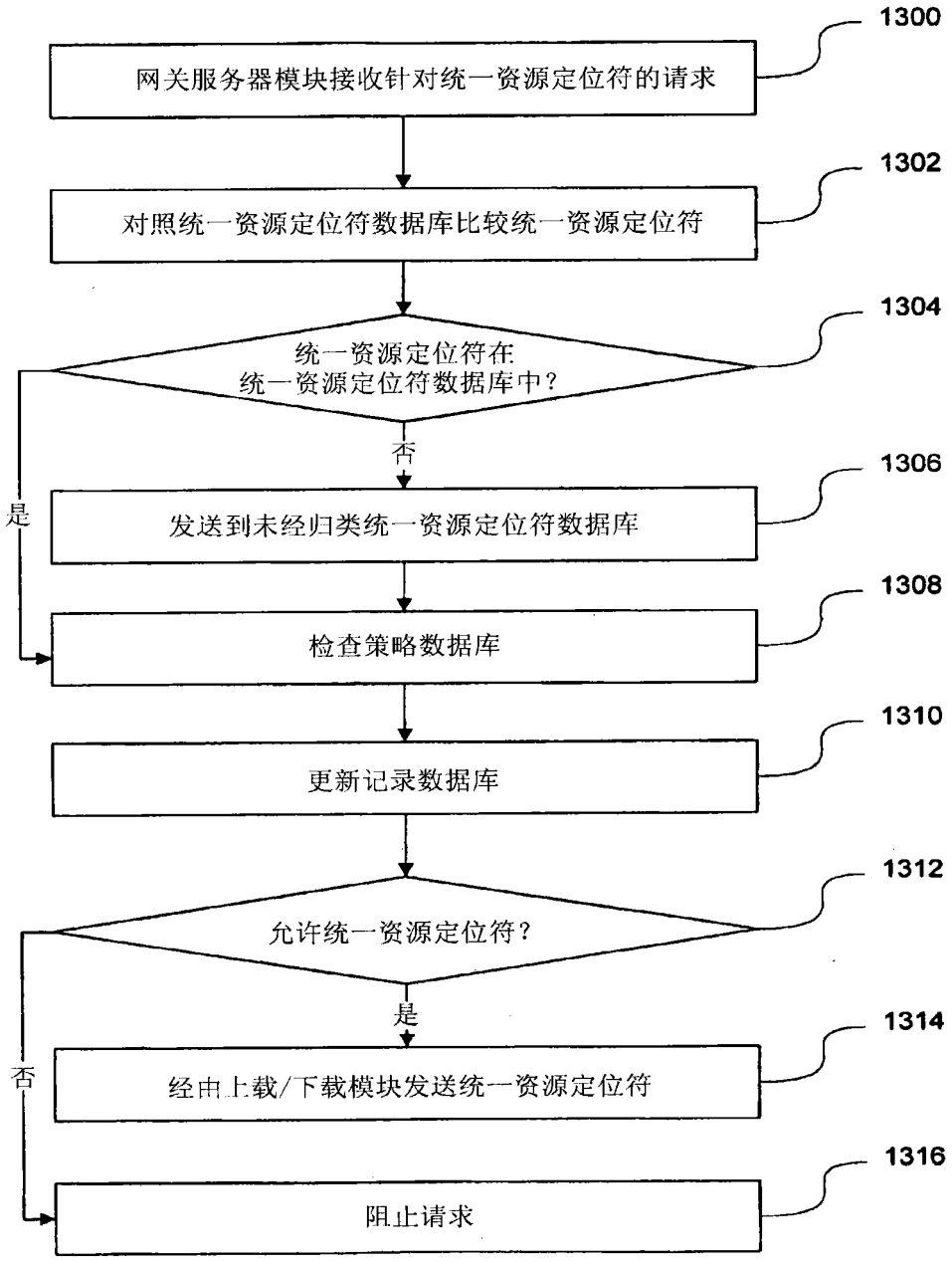

1300
网关服务器模块接收针对统一资源定位符的请求
1302
对照统一资源定位符数据库比较统一资源定位符
1304
统一资源定位符在
统一资源定位符数据库中？
否
是
1306
发送到未经归类统一资源定位符数据库
1308
检查策略数据库
1310
更新记录数据库
1312
允许统一资源定位符？
是
否
1314
经由上载/下载模块发送统一资源定位符
1316
阻止请求

图 17

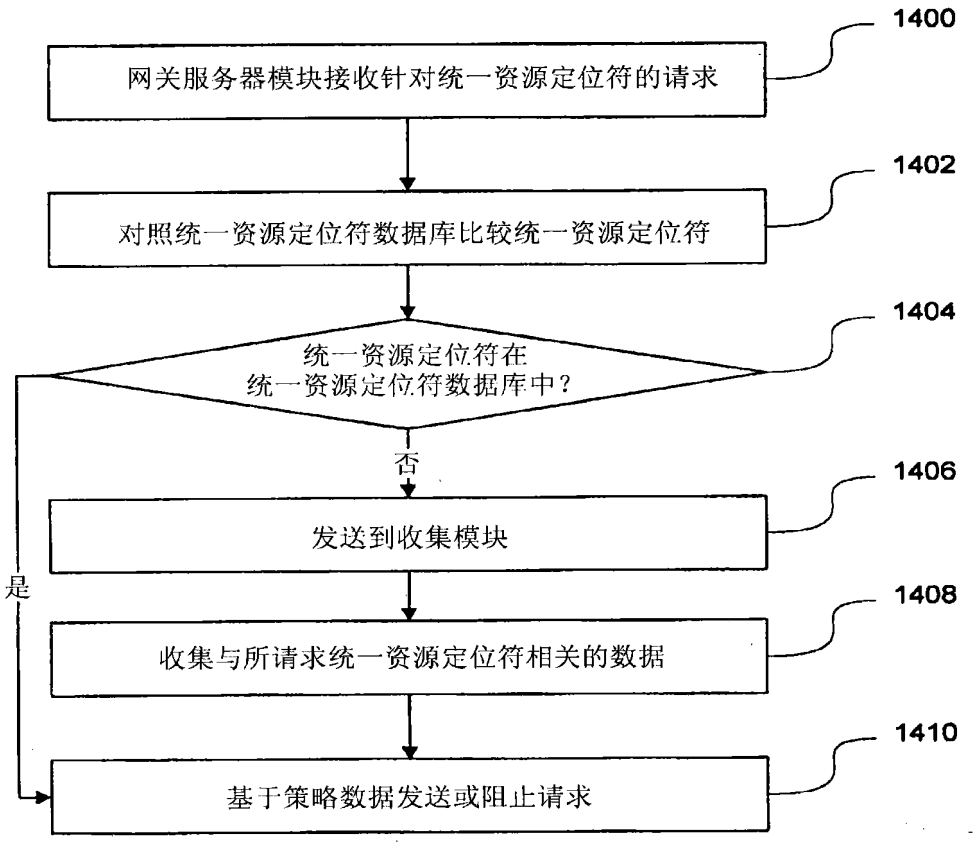

1400
网关服务器模块接收针对统一资源定位符的请求
1402
对照统一资源定位符数据库比较统一资源定位符
1404
统一资源定位符在
统一资源定位符数据库中？
否
是
1406
发送到收集模块
1408
收集与所请求统一资源定位符相关的数据
1410
基于策略数据发送或阻止请求

图 18

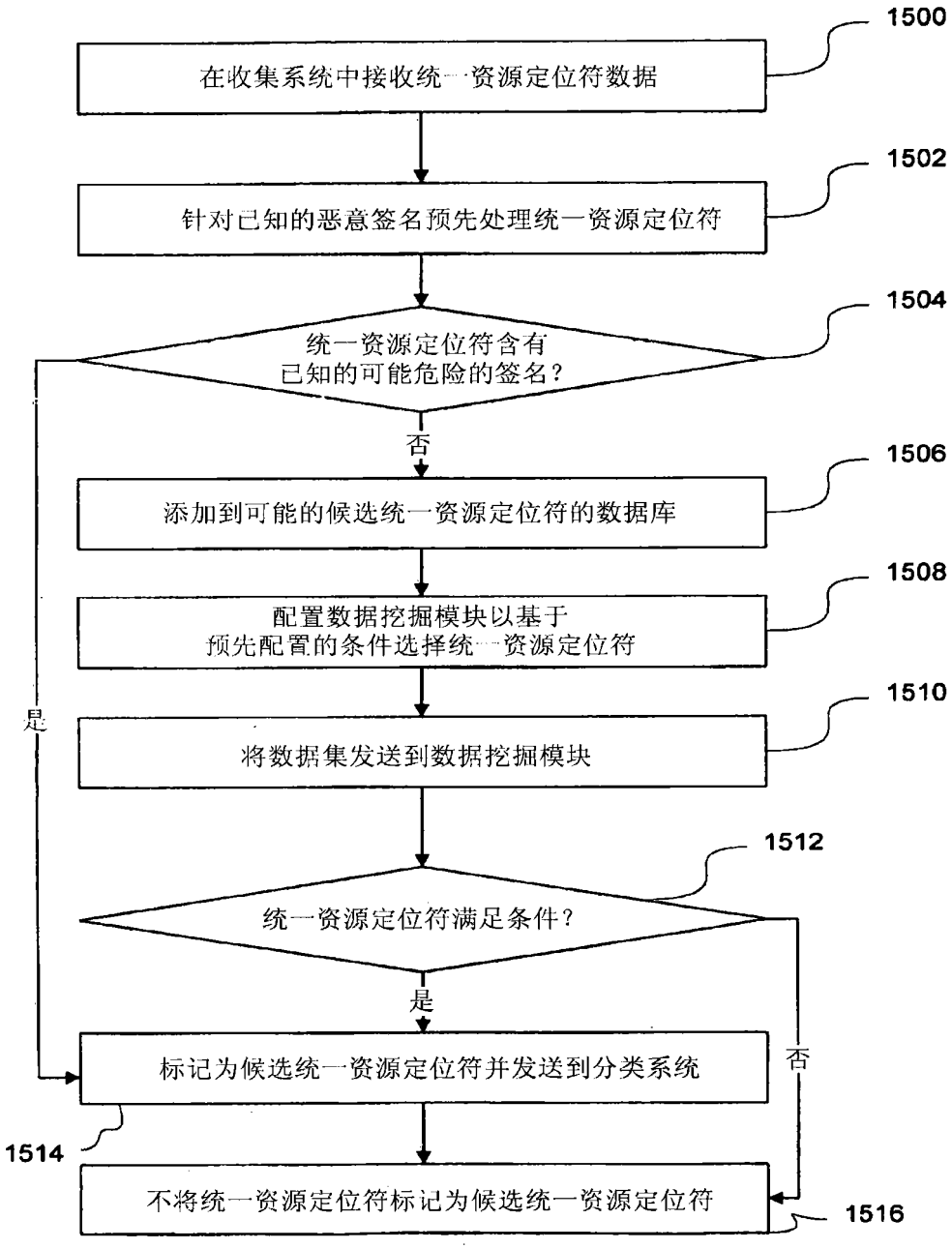

1500
在收集系统中接收统一资源定位符数据
1502
针对已知的恶意签名预先处理统一资源定位符
1504
统一资源定位符含有
已知的可能危险的签名？
否
1506
添加到可能的候选统一资源定位符的数据库
1508
配置数据挖掘模块以基于
预先配置的条件选择统一资源定位符
1510
将数据集发送到数据挖掘模块
1512
统一资源定位符满足条件？
是
否
是
标记为候选统一资源定位符并发送到分类系统
1514
不将统一资源定位符标记为候选统一资源定位符
1516

图 19

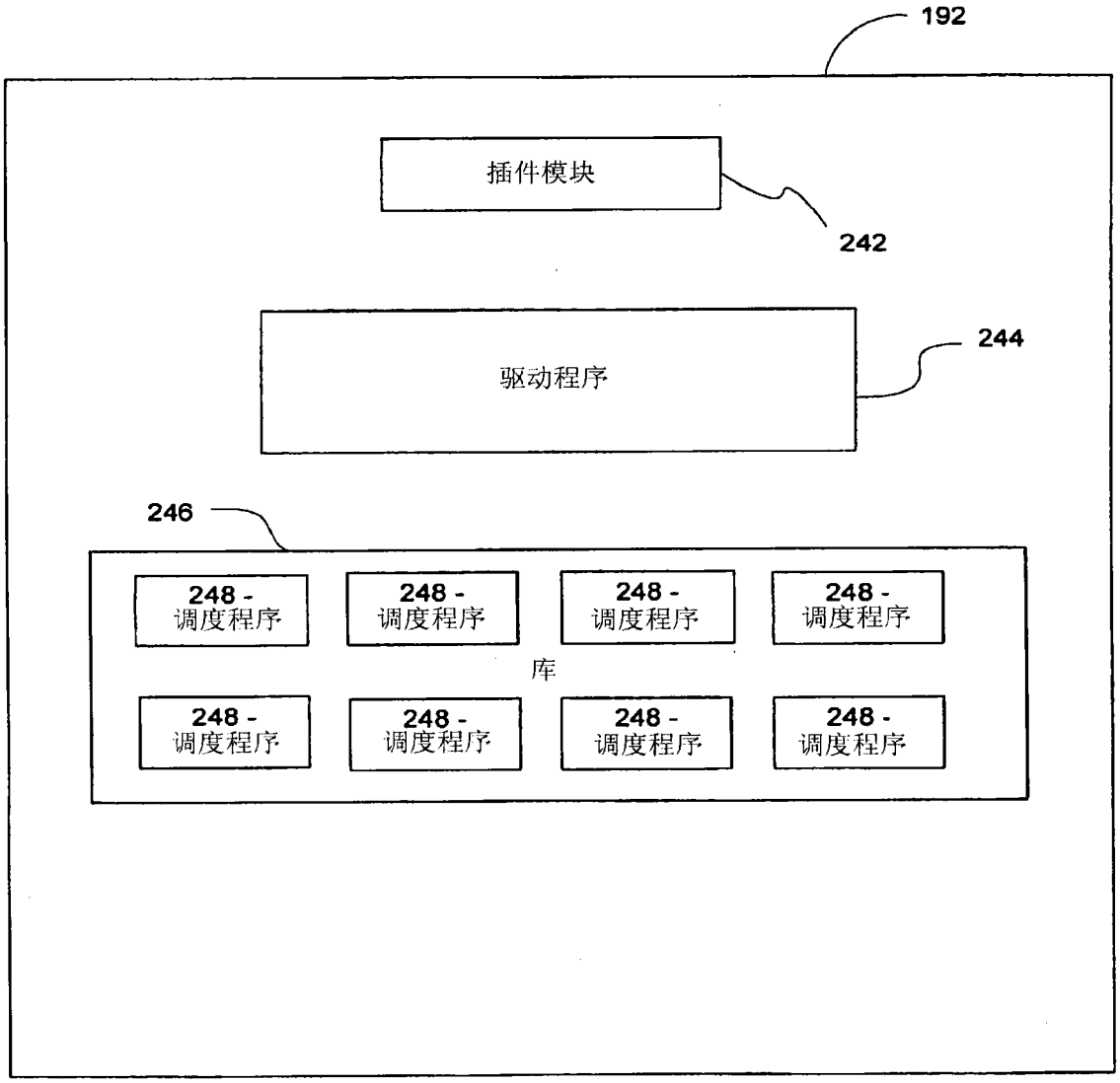

192
插件模块
242
驱动程序
244
246
248 - 调度程序
248 - 调度程序
248 - 调度程序
248 - 调度程序
库
248 - 调度程序
248 - 调度程序
248 - 调度程序
248 - 调度程序

图 20

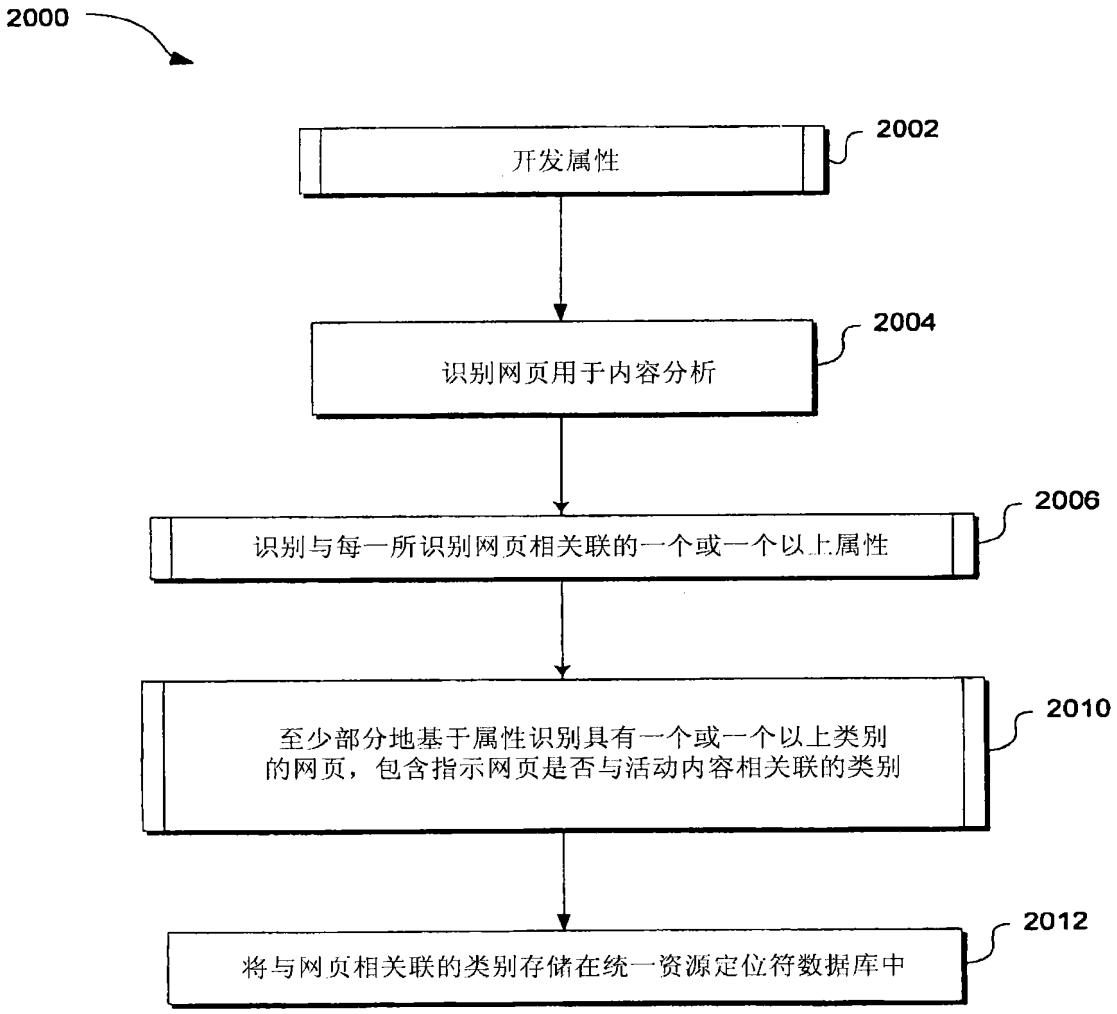

2000
开发属性
2002
识别网页用于内容分析
2004
识别与每一所识别网页相关联的一个或一个以上属性
2006
至少部分地基于属性识别具有一个或一个以上类别的网页，包含指示网页是否与活动内容相关联的类别
2010
将与网页相关联的类别存储在统一资源定位符数据库中
2012

图 21

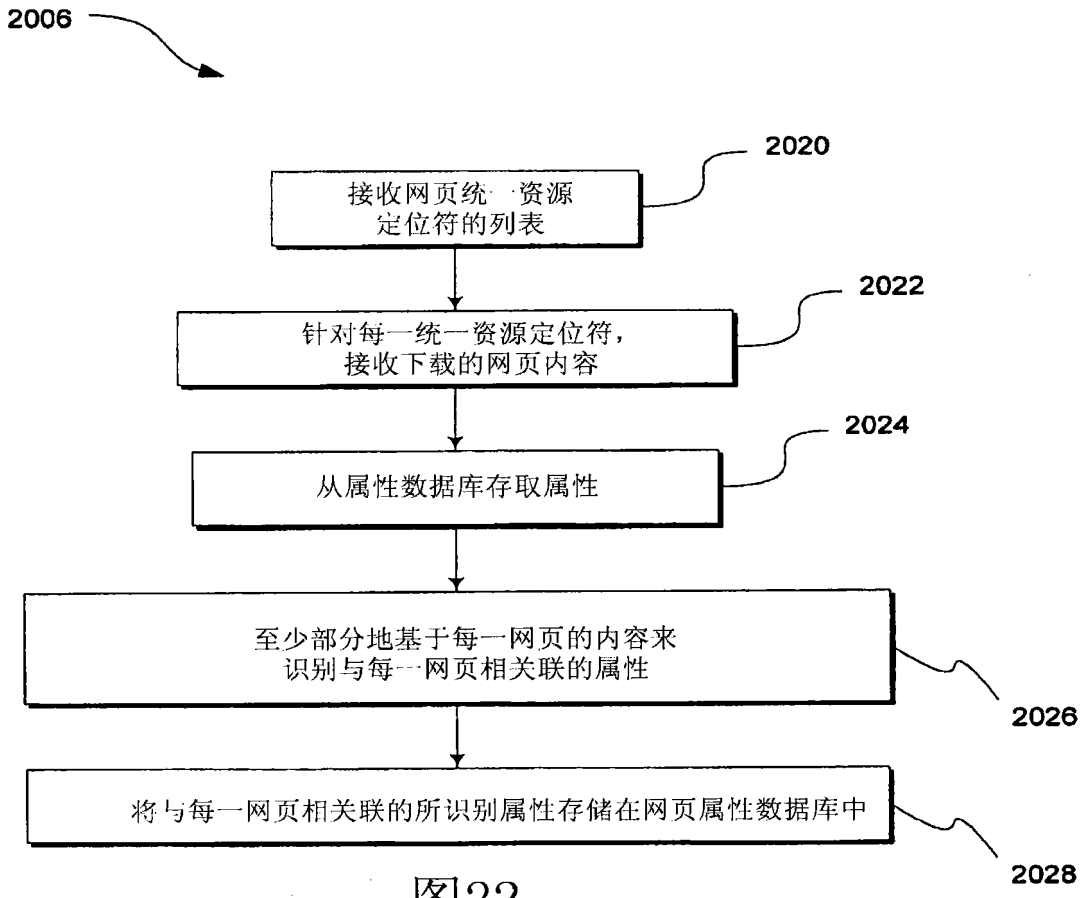

2006
2020
接收网页统一资源定位符的列表
2022
针对每一统一资源定位符，接收下载的网页内容
2024
从属性数据库存取属性
至少部分地基于每一网页的内容来识别与每一网页相关联的属性
2026
将与每一网页相关联的所识别属性存储在网页属性数据库中
2028

图22

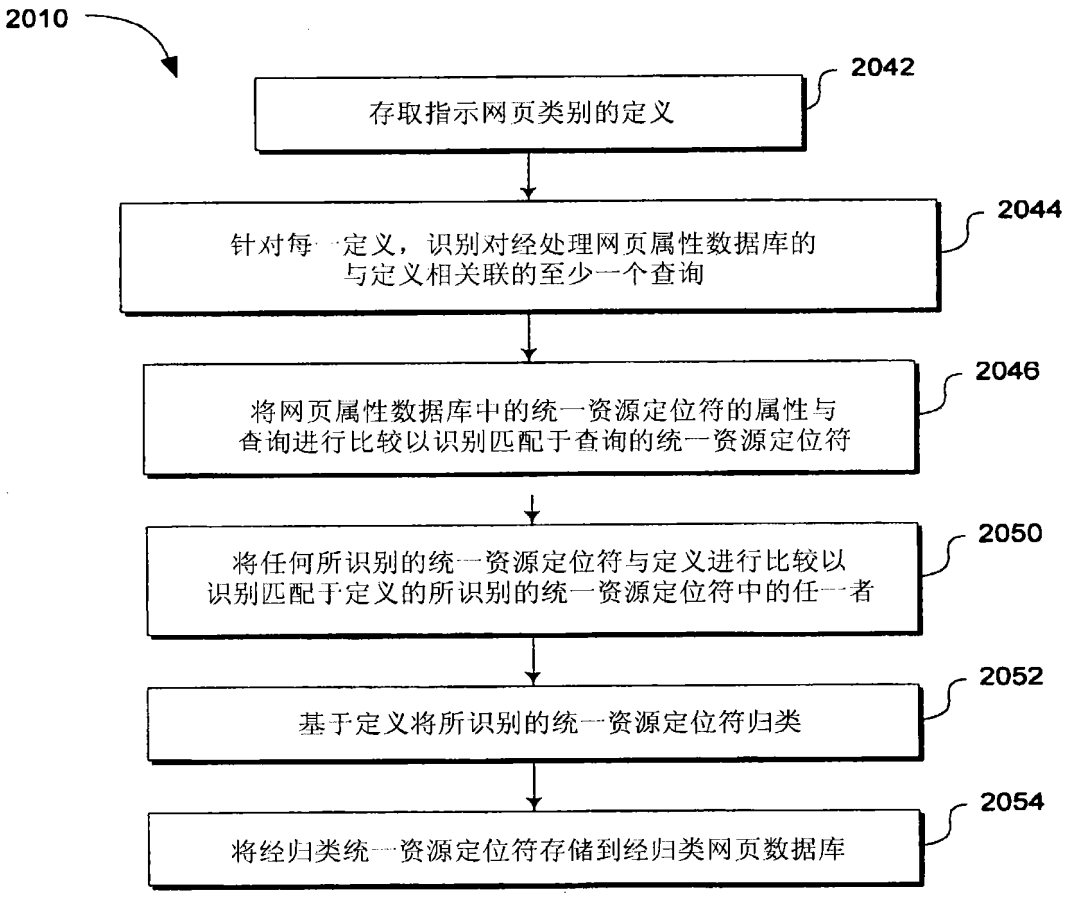

2010
2042
存取指示网页类别的定义
2044
针对每一定义，识别对经处理网页属性数据库的与定义相关联的至少一个查询
2046
将网页属性数据库中的统一资源定位符的属性与查询进行比较以识别匹配于查询的统一资源定位符
2050
将任何所识别的统一资源定位符与定义进行比较以识别匹配于定义的所识别的统一资源定位符中的任一者
2052
基于定义将所识别的统一资源定位符归类
2054
将经归类统一资源定位符存储到经归类网页数据库

图 23

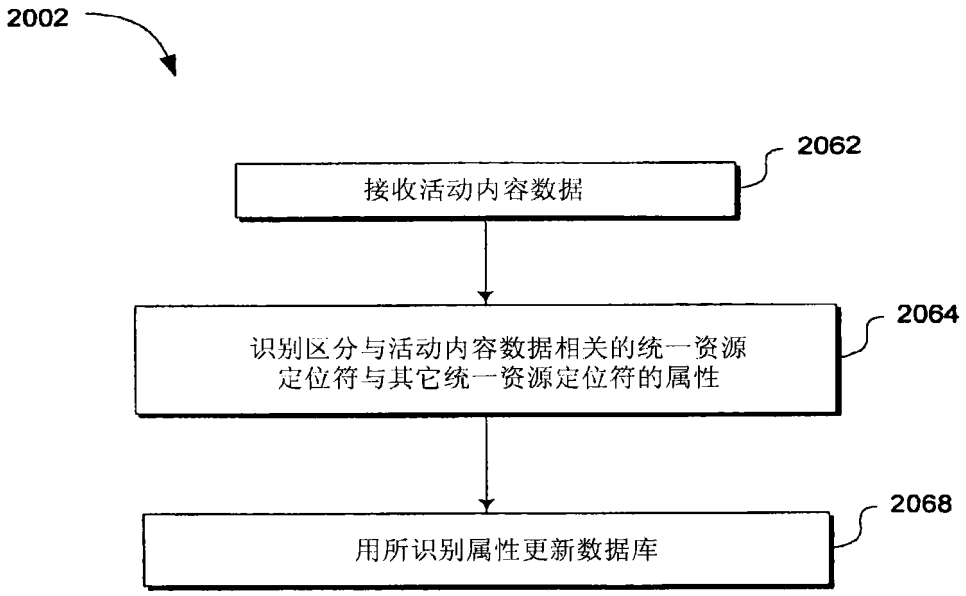

2002
2062
接收活动内容数据
2064
识别区分与活动内容数据相关的统一资源定位符与其它统一资源定位符的属性
2068
用所识别属性更新数据库

图 24