

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 15/16 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200880018566.X

[43] 公开日 2010年3月24日

[11] 公开号 CN 101681340A

[22] 申请日 2008.4.15

[21] 申请号 200880018566.X

[30] 优先权

[32] 2007.4.17 [33] US [31] 60/912,203

[86] 国际申请 PCT/US2008/060378 2008.4.15

[87] 国际公布 WO2008/130946 英 2008.10.30

[85] 进入国家阶段日期 2009.12.2

[71] 申请人 肯尼思·托拉

地址 美国伊利诺斯

共同申请人 厄尔·格兰特劳伦斯

[72] 发明人 肯尼思·托拉 厄尔·格兰特劳伦斯

[74] 专利代理机构 中国国际贸易促进委员会专利
商标事务所
代理人 杜娟

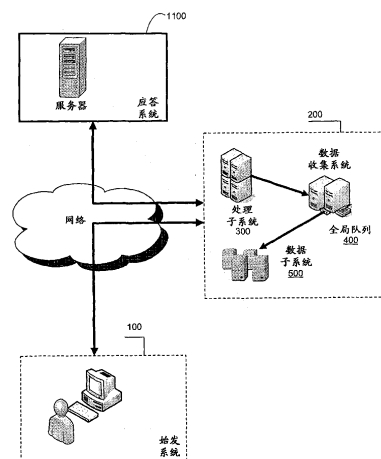
权利要求书 3 页 说明书 19 页 附图 12 页

[54] 发明名称

收集通过网络传输的信息的非介入性方法和系统

[57] 摘要

本发明涉及利用存在于始发系统和应答系统之间的数据收集系统，收集通过网络传输的信息的非介入性方法和系统。在一个实施例中，始发系统可以是 web 浏览器，应答系统可以是 web 服务器。在另一个实施例中，始发系统可以是本地计算机，应答系统可以是网络上的另一台计算机。这些和其它结构都被认为在本发明的范围之内。数据收集系统按照混合的对等/客户机-服务器方式充当应答系统来响应始发系统，同时充当始发系统来响应应答系统。这种结构能够按照完全非介入的方式实现网络流量信息的实时采集和存储，而不需要任何服务器端或客户端代码。



1、一种用于获得和存储通过网络传输的信息的系统，所述系统包括：

a.与所述网络连接的第一网络位置，该第一网络位置能够向所述网络上的第二网络位置传送信息；

b.逻辑上位于所述第一网络位置和所述第二网络位置之间的中间网络位置，所述中间位置能够：

i.接收来自所述第一网络位置的传输；

ii.向所述第二网络位置发送传输，所述传输以从所述第一网络位置接收到的传输的内容为基础；

iii.接收来自所述第二网络位置的传输；

iv.根据从所述第二网络位置接收到的传输的内容，向所述第一网络位置发送传输；

c.其中所述中间网络位置能够把从所述第一网络位置和所述第二网络位置接收到的传输存储在逻辑存储位置中。

2、按照权利要求 1 所述的系统，其中当向所述第二网络位置发送传输时，所述中间网络位置充当所述第一网络位置。

3、按照权利要求 1 所述的系统，其中当向所述第一网络位置发送传输时，所述中间网络位置充当所述第二网络位置。

4、按照权利要求 1 所述的系统，其中所述逻辑存储位置是全局队列。

5、按照权利要求 4 所述的系统，其中所述存储位置中的信息被存放在更永久的存储介质中。

6、按照权利要求 5 所述的系统，其中所述存储介质选自文件系统上的文件或者数据库中的记录。

7、按照权利要求 1 所述的系统，其中至少一个所述传输包括一个不可见的组件，并且追踪值被放入所述不可见的组件中，以便追踪一系列的传输。

8、按照权利要求 1 所述的系统，其中从所述中间网络位置到所述第一网络位置的传输包括活动内容，并且其中该活动内容中的至少某一部分被修改，以便把后续响应引导回所述中间网络位置。

9、按照权利要求 8 所述的系统，其中所述活动内容是一个或多个超链接。

10、按照权利要求 8 所述的系统，其中所述活动内容是嵌入组件。

11、按照权利要求 9 所述的系统，其中所述嵌入组件是 Flash、ActiveX 或 Java 小程序之一。

12、按照权利要求 9 所述的系统，其中所述嵌入组件是客户端脚本。

13、按照权利要求 12 所述的系统，其中所述客户端脚本选自 VBScript 或 JavaScript。

14、按照权利要求 8 所述的系统，还包括能够被所述中间网络位置监视的网络位置列表。

15、按照权利要求 14 所述的系统，其中可选的是，可比较从所

述第一网络位置接收的传输与待监视的所述网络位置列表，从而：

a.如果由来自第一网络位置的传输识别的目标第二网络位置不在所述网络位置列表中，那么来自所述第一网络位置的传输被所述中间网络位置直接转发给所述第二网络位置，所述中间网络位置不采取修改、追踪或存储所述传输的动作。

16、按照权利要求 8 所述的系统，其中通过不同的 URI 值，传输从所述第一网络位置被引导到所述中间网络位置。

17、按照权利要求 16 所述的系统，其中所述 URI 值包含使所述中间网络位置能够确定所述第二网络位置的唯一值。

18、按照权利要求 17 所述的系统，其中所述唯一值以名称-值对的形式被放入所述 URI 中。

19、按照权利要求 1 所述的系统，其中利用 DNS 条目，传输从所述第一网络位置被引导到所述中间网络位置。

20、按照权利要求 1 所述的系统，其中来自所述第一网络位置的唯一信息被用于确定所述第二网络位置，其中来自所述第一网络位置的传输被自动发送给所述中间网络位置。

21、按照权利要求 20 所述的系统，其中来自所述第一网络位置的所述唯一信息是所述第一网络位置的 IP 地址。

22、按照权利要求 20 所述的系统，其中来自所述第一网络位置的所述唯一信息是所述第一网络位置的 MAC 地址。

收集通过网络传输的信息的非介入性方法和系统

技术领域

本发明总体上涉及收集通过网络传输的信息的非介入性方法和系统。

背景技术

数据收集解决方案通常可被分成两种一般性途径。第一种途径(称为服务器端途径)把软件装载到用户的服务器上,例如,分组“探查”软件和日志文件分析软件。这种软件收集许多更常见的使用统计信息,在存储用于传输数据的方法方面非常有益。第二种途径专注于把代码放在客户的计算机上,以捕捉客户与远程站点的交互作用。这些客户端数据收集解决方案采取各种形式。客户端数据收集解决方案的例子包括在页面和文本文件(也被称为“cookies”)上插入的代码,所述页面和文本文件被存储在客户的机器上。

不幸的是,两种途径都存在许多缺陷,使它们不能成为全面的非介入性数据收集的可行选项。这些途径的一个主要缺陷在于代码必须被安装在用户的服务器上(在前一情况下),或者被安装在客户的机器上(在后一情况下)。这种要求加剧了软件兼容性问题、追踪的解决方案的发展约束、和用户/客户时间使用问题。这些途径还限制了追踪的网络化解决方案的有效性或实用性。在服务器端途径中,许多追踪途径使用缓存的组件,它们不能支持复杂的客户端交互作用,所述复杂的客户端交互作用构成许多网络化解决方案的基础。另一方面,客户端途径不能恰当地处理客户机和服务器之间的新的交互作用,因为它们依靠静态使用模式来推断用户活动。最后,跨越关联的服务提供追踪客户机的需求日渐增长,这种能力超出了服务器端解决方案的范围,只有通过使用第三方实用程序依靠客户端解决方案才可行,不过,

多数现代系统默认禁用第三方实用程序。例如，就网站追踪来说，可供这些类型的追踪系统在多个网站上持续存在的唯一手段是使用第三方 cookies。现代 web 浏览器默认拒绝使用这种 cookies 的能力。

语境信息

现有解决途径的其它主要缺陷之一是缺少语境相关数据。为了理解该概念，将分析大脑成像的例子。在较早的正电子发射断层成像 (PET) 扫描方法中，放射性物质被用于追踪人类的大脑功能。这种途径可提供大脑活动的彩色图像，不过没有任何结构，从而医生不能确定大脑的哪个部分负责所观察到的活动。

另一种较早的技术-磁共振成像 (MRI) 非常擅长形成三维组织结构的图像，常被用于寻找诸如肿瘤或凝块之类的密集组织。尽管成像分辨率高，不过 MRI 不提供功能，仍然很难确定什么区域可能受到损伤或者未受到损伤。

在 1991 年，这两种途径被结合成现在称为功能磁共振成像 (fMRI) 的方法。这种技术把功能重叠在结构之上，已导致神经影像诊断学的发展。准确地了解什么结构正在进行什么活动的的能力是恰当确定活动的关键要素。

上面仅仅是从完全无关的技术领域的粗略概念类推，不过仍然特别有益于理解目前的追踪行业。一方面，现代追踪解决方案准确度不断变化地捕捉客户机互动(或功能)。不过，这些追踪解决方案都不能捕捉在这些互动期间目标系统的结构。

另一方面，各种爬虫 (crawler) 每天能够提供成千上万的连网解决方案的详细结构，但是都不能捕捉客户机互动。

没有把网络站点的结构和客户机互动 (这里称为语境信息) 联系起来的能力，那么了解网站功能的能力就会被显著削弱或者降低。

发明人认识到上面提及的缺陷，提供一种克服上面提及的缺陷的收集通过网络传输的信息的系统和方法。

附图说明

结合附图，根据本发明的下述详细说明，本发明的这些和其它目的、特征和优点将变得显而易见，其中：

图 1 是按照本发明的一个实施例的数据收集系统结构的方框图。

图 2 是按照本发明的一个实施例的网络结构的方框图。

图 3 是按照本发明的一个实施例的端口转发的方框图。

图 4 是按照本发明的一个实施例的控制器方框图。

图 5 图解说明通用消息格式。

图 6 图解说明概念性的 URI 查寻表。

图 7 是图解说明按照本发明的一个实施例的关于从始发系统传送给应答系统的消息的事件处理步骤的流程图。

图 8 是图解说明按照本发明的一个实施例的关于从始发系统传送给应答系统的消息的动态内容管理步骤的流程图。

图 9 是图解说明按照本发明的一个实施例的关于从应答系统传送给始发系统的消息的事件处理步骤的流程图。

图 10 是图解说明按照本发明的一个实施例的关于从应答系统传送给始发系统的消息的动态内容管理步骤的流程图。

图 11 是图解说明按照本发明的一个实施例的存储追踪信息的步骤的流程图。

图 12 是提供按照本发明的一个实施例的数据子系统的一种可能结构的方框图。

具体实施方式

本发明的优选实施例提供可被配置成充当起应答系统以与始发系统通信的数据收集系统。从始发系统发送的信息可被存储供以后使用，随后被用于根据始发系统请求的语境产生请求。数据收集系统随后起始发系统的作用，并经网络把请求提交给应答系统。始发消息(请求)包括第一统一资源指示符(URI)，所述第一统一资源指示符可被用于至少部分根据动态 URI 映射，确定应答系统 URI。应答系统随后向数据收集系统返回响应，该响应既可被存储，又可被用于产生返回给

始发系统的响应。该信息随后可被用于支持关于监视的有网络功能的站点的高级用户交互作用分析。

按照一个优选实施例(下面有时称为 **DataTrendz™**), 提供追踪通过网络传输的消息的方法和系统。**DataTrendz™**的直接把处理插入请求-响应流中的能力首次允许用户存储和/或分析结构和功能。收集这种语境相关数据将提供超出简单的追踪和报告的重大新见识。对具有宽范围的不同网络位置的网络, 比如因特网来说, **DataTrendz™**提供的效用和功能性是可实现的。在本例中, 网络位置可包括网络服务器、网站服务器、个人计算机、诸如能够接入因特网的电话机之类的移动设备、和其它具有网络能力的设备的主机。不过, **DataTrendz™**也向其它网络, 比如企业内部网提供优选的功能性和效用, 在所述其它网络中, 与因特网相比, 网络位置的范围可能更单一。于是, **DataTrendz™**的具体实现可以包括连接实际上任意类型的网络位置与实际上任意其它类型的网络位置的实际上任意类型的网络。

DataTrendz™解决了限制当前的追踪方法的众多挑战, 同时把通信追踪和分析的概念扩展到对基于网络的通信的限制之外。

网站特有的益处

在网站域(website-domain)内, **DataTrendz™**提供许多益处, 比如(但不限于):

代码密集的。问题: 许多数据收集解决方案要求在客户机或用户机器上的大量代码。解决方案: **DataTrendz™**的系统和方法不要求在客户机或用户机器上的代码。

陈旧的推断方法。问题: 在已知处理始发请求的方法的新技术途径的情况下, 许多追踪解决方案用于确定线索的典型服务器处理使用模式不再有效。解决方案: 作为其语境数据收集过程的一部分, **DataTrendz™**捕捉实际的线索信息, 使通过推断, 或者至少完全或者主要通过推断确定功能的思想变得陈旧。

跨域问题。问题: 在不借助于第三方 cookies 的情况下, 典型的数据收集解决方案无法跨网站追踪用户。解决方案: 由于

DataTrendz™起非介入性追踪系统的作用，因此能够在不使用 cookies，或者任何其它用户/客户端代码的情况下，跨数目无限的网站进行追踪。

高速缓存。问题：一些数据收集解决方案响应始发请求，发送高速缓存的各种版本的用户网站。这种途径不能支持具有高级的客户端功能性的复杂网站。解决方案：当利用非介入性追踪系统时，不需要任何高速缓存。另外，通过在套接字层工作，动态请求、解析和 HTML 包创建与请求链中的任何其它网络跳程(network hop)一样快。

浏览器不可知论。问题：利用客户端 JavaScript 或者服务器端帧-目前的数据收集途径中就是这样-会导致浏览器依赖性问题的。解决方案：**DataTrendz™**不要求在客户机的浏览器上放置会影响用户界面的任何东西，于是不存在和该追踪途径相关的任何浏览器问题。

图 1 图解说明包括数据收集系统 200 的网络。数据收集系统 200 管理往来于始发系统 100 和应答系统 1100 发送的消息。在本发明的一个实施例中，网络包括采用有线、无线蜂窝或其它媒介的因特网。在本发明的另一个实施例中，网络选自：局域网(LAN)和广域网(WAN)。本发明并不局限于任何特定网络结构方面的实现方式。相反，本发明可应用于任何类型的包含配置成利用以电子方式传送的消息相互通信的互连计算机的系统。

在本发明的一个实施例中，处理子系统 300、全局队列子系统 400 和数据子系统 500 可存在于分开的物理设备或一组设备之中。在另一个实施例中，这些子系统可存在于同一设备中，或者可存在于多个设备的任意组合中。

在一个实施例中，设备驱动器层的网络通信（或流量(traffic)）可根据存储器中规则被重新路由到所得到的 URI 地址。通过利用这些基于软件的与 DNS 相关的路由系统，**DataTrendz™**有能力在外部使用任何域名，并把该通信路由给希望的内部位置，而不需要独立的 URI 值。本实施例可被用于通过利用跨相同服务器、局域网、广域网或者它们的任意组合的处理位置，以对称或者固定的方式把通信量

(或流量、通信流量)均衡到已知的处理位置。

在本发明的一个优选实施例中,始发系统利用域名源(DNS)统一资源标识符(URI)发送请求。在用实线表示的从始发系统 100 到地理负载均衡器 201 的主路径上,该 URI 把消息传给地理负载均衡器 201。URI 是作为定位外部资源的当前例子提供的,并不意图限制本发明。

可以存在和为确保完全可用性所需一样多的地理负载均衡器 201,甚至不存在任何地理负载均衡器 201,出于举例说明的目的,示出了两个地理负载均衡器 201。在本实施例中,地理负载均衡器 201 相互通信,以确保每个站点 203 正在恰当地运行,并跨区域均衡负载。如果主要的地理负载均衡器 201 未能应答用户请求,那么 DNS 协议会自动失效转移(或故障转移(failover))到辅助地理负载均衡器 201,如图 2 中用虚线所示。域名系统(DNS)是作为目前实现的识别外部资源的手段的例子提供的,并不意图作为对本发明的限制。这种失效转移过程将继续进行和给出的实施例的实现中提供的站点 203 一样多的次数。

在站点 203 内,利用站点负载均衡器来保持一个或多个处理子系统 300 之间的功能性。如果指定的处理子系统 300 发生故障(或失效),那么所有的通信量将被转移给其余的处理子系统。如果站点 203 内的所有处理子系统都不在进行处理,那么站点负载均衡器 202 将向地理负载均衡器 201 返回该消息,以便在另一个站点 203 进行处理。

在本发明的一个优选实施例中,数据收集系统 200 包含配置成与始发系统 100 和应答系统 1100 通信的服务器。数据收集系统 200 动态监视从始发系统 100 传送的要给应答系统 1100 的消息,和从应答系统 1100 传送的要给始发系统 100 的消息。为此,数据收集系统 200 包括在处理子系统 300 内的端口监视器 301,如图 3 中所示。

如图 3 中所示,当端口监视器 301 收到请求时,该请求被端口转发给多个端口处理器 600 之一。在一个实施例中,只有一个端口处理器 600,在另一个实施例中,有多个端口处理器 600。在一个实施例

中，一个或多个端口处理器 600 存在于一个物理设备中，在另一个实施例中，一个或多个端口处理器存在于多个独立设备的任意组合中。所有实施例都被认为在本发明的范围之内。

端口处理器 600 包括数据表示 601，数据表示 601 包含子域(SD)值 102 和它们对应的应答统一资源标识符(URI)103 之间的映射，如图 5 中所示。在本发明的一个优选实施例中，映射表包含存储器中 XML 文件 608，如图 6 中所示。图 6 中图解说明的存储器中 XML 文件 608 是概念表示。因而，它不指示特定数目的条目，也不指示条目的所有细节。数据表示 601 的确切实现方式可变化。所有这些变化都在本发明的范围之内。

数据表示 601 包含子域值 102。每个子域条目包括表示对应的应答域的值和目标 URI。对应的 URI 示于图 6 中。URI 是用于识别应答系统 1100 的一种手段。按照本发明的一个优选实施例，网络的多个应答系统 1100 被映射到图 6 中的对应的唯一子域值 102。这里，把应答系统 1100 的映射子域值 102 称为“应答域”。

在本发明的一个优选实施例中，映射表包含存储器中 XML 文件 608，存储器中 XML 文件 608 包含 URI 103。在本发明的另一个实施例中，映射表包含 XML 文件，该 XML 文件包含应答系统统一资源定位符。在本发明的一个优选实施例中，映射表被存储在数据收集系统 200 的存储器中。在本发明的另一个实施例中，映射表被存储在端口处理器 600 的存储器中。

图 5 图解说明表示通常用于经因特网传送的消息的通用消息结构。消息 107 包含报头部分 101、URI 部分 103 和页面内容部分 104。URI 部分 103 包含子域部分 102、基本域部分 105 和查询串部分 106。该消息是作为一种参考而提供的，该消息的所有组合或衍生物都在本发明的范围之内，在优选实施例中，这些消息(及其组合和衍生物)能够包含报头信息和内容。

处理子系统 300

图 4 图解说明按照本发明的一个优选实施例，在图 1 中图解说明

的数据收集系统 200 的处理子系统 600。处理子系统 300 包含把通信转发给一个或多个端口处理器 600 的端口监视器(PM)301。端口处理器 600 包括动态内容管理单元(DCMU)900、事件汇点发生器(ESG)700、全局队列接口 408、事件处理器单元(EHU)1000、数据表示 601、和用户代理(UA)800。处理子系统 600 还通过全局队列接口 408 与全局队列 400 通信。

端口监视器 301

端口监视器 301 被配置成检测包含网络上的通信的数据流。端口监视器监视数据收集系统 200 的一个或多个端口(例如, 端口 80、81 等), 以检测网络通信流量。网络通信流量的一个例子是从始发系统 100(图 1 中图解所示)传送的针对由应答系统 1100 提供的信息(例如网页)的消息。该通信流量可以是安全的或不安全的; 任意类型的网络上的两个设备之间的有线、无线通信, 或蜂窝通信, 或任何其它形式的通信。

按照本发明的一个优选实施例, 始发系统 100 包含用户计算机。来自用户计算机的消息的一个例子是用户经始发系统 100 对应答系统 1100 提供的网页的请求。用户的请求可被送往包含数据收集系统 200 的服务器。注意用户的请求最好终止于数据收集系统 200, 尽管用户请求的信息存在于应答系统 1100 上。端口监视器 301 能够检测网络通信, 并按照负载均衡的方式把该信息传送给一个或多个端口处理器 600。

端口处理器 600

响应端口监视器 301 检测到的用户请求, 端口处理器 600 产生请求消息。端口处理器 600 请求可被传送给最好由在数据表示 601 中找到的映射确定的目标应答系统 1100。应答系统 1100 同步地应答来自端口处理器 600 的请求。应答系统 1100 把其响应送往数据收集系统 200, 所述响应被端口监视器 301 捕捉, 并被转发给同一端口处理器 600。

事件处理器(event handler)单元(EHU)1000

在端口处理器 600 内, EHU 1000 被配置成与消息输入单元 609、DCMU 900、URI 映射的数据表示 601 和全局队列接口 408 通信。EHU 1000 进行这里称为事件消息处理的过程。第一步是根据输入的 URI 解析子域, 并根据数据表示 601 进行查找查询。如果查找得到应答域, 那么 EHU 1000 把输入的请求和应答域传递给 DCMU 900 和全局队列接口 400。如果查找未得到应答域, 那么请求被直接递传给应答系统 1100, 从而绕过数据收集系统 200 的数据收集和存储机构。

对于来自始发系统 100 的关于出自应答系统 1100 的信息的请求, EHU 1000 被配置成执行图 7 中图解说明的方法。这种情况下, 始发系统 100 发送请求, 如在图 7 的 108 所示, 端口监视器 300 接收该请求, 如在图 7 的 302 所示。对于来自应答系统 1100 的提供端口处理器 600 所请求信息的响应, EHU 1000 被配置成执行图 9 中图解说明的方法。这种情况下, 应答系统 1100 发送响应, 如在图 9 的 1101 所示。端口监视器 301 接收该响应, 如在图 9 的 303 所示。

参见图 7, 消息输入单元 609 从端口监视器 301 接收消息 107, 如步骤 605 中所示。EHU 1000 从端口处理器 600 接收表示来自始发系统 100 的请求的消息。在本发明的一个优选实施例中, 请求包含图 5 中在 107 图示的一般种类的消息。通过解析消息 107 的 URI 103, 从而识别子域 102 值, EHU 1000 能够评估接收的请求(图 7 的步骤 1001-1003)。EHU 1000 确定消息 107 的子域 102 是否对应于被监视的子域 102 值。被监视的子域 102 值是由数据收集系统 200 为应答系统 1100 分配的值。

如果 EHU 1100 确定 URI 103 中的子域 102 值是被监视的子域 102(图 7 的步骤 1005), 那么 EHU 1100 把消息 107 发送给动态内容管理单元(DCMU)900(图 7 的步骤 1007)。另外, EHU 1100 把消息 107 提供给全局高速缓存 400(图 7 的步骤 1008)。EHU 1100 根据应答域的值做出确定。如果应答系统 1100 URI 103 不在数据表示 601 中, 那么 EHU 1100 把消息传递给应答系统 1100。

按照对应的方式, 消息输入单元 609 能够从端口监视器 301 接收

消息 107, 该消息 107 表示应答系统 1100 响应来自同一端口处理器 600 的请求而传送的响应, 如图 9 中所示。该端口处理器把消息传递给 EHU 1000, 如图 9 中的步骤 607 中所示。这种情况下, EHU 1000 最好进行图 9 中图解说明的步骤。EHU 1100 确定消息 107 中的子域 102 值是否在数据表示 601 中, 如图 9 的步骤 1009、1002 和 1003 中所示。EHU 1100 随后在步骤 1007 中把消息 107 提供给全局高速缓存 400, 在步骤 1008 中把消息 107 提供给 DCMU 900。

DCMU 400

DCMU 900 实现下面说明的如图 10 中所示的一般功能。

内容检索。DCMU 900 使用输入消息 107 的内容, 以及输入 URI 103 的值动态生成请求。在 DCMU 900 模仿始发系统 100 的情况下, 该请求被发送给应答系统 1100。来自应答域 1100 的响应被捕捉, 并被暂时存储为存储器中消息 107。来自应答系统 1100 的响应的内容被用于生成将被回送给始发系统 100 的消息 107。如在图 10 的步骤 802 中所示, 插入自定义报头 101, 以便在后续传输中识别该消息。关于要追踪的所有可动作组件(例如, JavaScript, Form Post Addresses, 超链接等)的基本 URI 103 被修改, 以指回数据收集系统 200 和被端口监视器 301 监视的端口。动态响应被回送给 EHU 1000。

图 8 图解说明按照本发明的一个优选实施例, 由 DCMU 900 执行的方法的步骤。如在图 8 的 901 所示, DCMU 900 从 EHU 1000 接收消息 107 和应答系统 1100 URI 103。DCMU 900 把消息 107 解析成报头 101 部分和页面内容 104 部分(在步骤 902 所示)。报头 101 部分和页面内容 104 部分被提供给 ESG 700。ESG 700 用 EHU 1000 提供的应答系统 1100 URI 103 替换页面内容 104 中的子域(SD)102 值(在图 8 的步骤 701 所示)。消息 107 被提供给用户代理 800, 如图 6 中在步骤 801 所示。用户代理 800 从消息 107 的报头部分中除去自定义报头 101, 并把消息 107 回送给 ESG 70, 以便进一步处理。ESG 700 用 EHU 1000 提供的应答系统 1100 URI 103 替换报头 101 集合中的子域 102 值。

DCMU 900 创建新消息 107 信封,如在图 8 的 903 所示。DCMU 900 把 ESG 700(在步骤 702)提供的页面内容 104 转移到新消息 107 信封(在步骤 904)。在步骤 905, DCMU 900 把 ESG 700(在步骤 704)提供的报头 101 集合转移到新消息 107 信封中。在步骤 900 中,在 DCMU 900 提供的信封中,该消息被传送给应答系统 1100。

图 8 图解说明处理来自应答系统 1100 的响应的 DCMU 900 过程。对于响应来说, DCMU 900 充当应答系统 1100 的客户机。如图 8 中所示,当如步骤 907 中所示, DCMU 900 从 EHU 100 收到消息 107 和始发系统 100 URI 103 时,开始 DCMU 900 过程。DCMU 900 把响应解析成报头 101 集合部分和页面内容 104 部分(步骤 902)。最好,页面内容 104 部分的所有可动作组件被 DCMU 900 修改,以致基本 URI 103 往回指向数据收集系统 200(步骤 7-1)。在步骤 703 中,自定义报头 101 被添加到报头 101 集合中, DCMU 创建这里称为“动态响应消息”的新消息 107。在步骤 702 和 704 中提供的页面内容 104 和报头 101 集合信息被转移到新消息 107 中, DCMU 900 把所得到的动态响应消息提供给 EHU 1000(在步骤 903-906 示出)。

事件汇点发生器(ESG)700

ESG 700 与 DCMU 900 耦接。ESG 700 准备在响应来自用户的情况下,将由系统恰当处理的动态响应。在本发明的一个实施例中, ESG 700 实现下述功能。

会话创建。如果尚不存在关于该动态响应的会话,那么生成一个新的全局唯一标识符(GUID),并将其添加到报头 101 集合中。从消息 107 的报头 101 集合查询会话。会话 GUID 被输入到消息 107 的报头 101 集合中。包含 DataTrendz™会话报头值的消息组件集合被称为“监控响应”。监控响应随后被回送给 EHU 1000。

全局队列 400

全局队列 400 把关于指定请求的信息存储到通过全局队列管理器 409 管理和存留的存储器中位置中,如图 11 中所示。全局队列可包含单独地或者在群集环境中进行处理的一个或多个服务器。独立于

全局队列管理器的物理实现方式，全局队列管理器能够管理一个或多个全局队列 400，不论这些队列是驻留在同一的物理机器上还是驻留在独立的物理机器上。

全局队列接口 408 为 EHU 1000 过程提供按照即发即弃 (fire-and-forget) 方式把新消息 107 放置在队列中的手段。在一个实施例中，每个 EHU 1000 过程可存在一个全局队列 400，在另一个实施例中，全局队列 400 和 EHU 1000 过程可共有多对多关系。

在本发明的一个实施例中，全局高速缓存 400 是由两个或者更多的过程访问的共享系统资源。在一个优选实施例中，全局高速缓存 400 是用于传递数据的异步排队/缓存机构。在本节中描述的和根据该评论推测的所有实施例都被认为落入本发明的范围中。

全局管理器 409 负责监视指定的全局队列 400 内的各种队列存储过程。如果任意存储过程变慢或者没有响应，那么全局队列管理器负责发起新的队列存储过程，同时温和地终止有问题的存储过程。这种思想被称为旋转加速 (spinning up) 和旋转减速 (spinning down) 过程。

如图 11 中所示，全局队列利用下述步骤处理输入的消息：会话确定。查询报头 101 集合，以确定会话存在。如果对于该消息来说，会话尚不存在，那么生成新会话 GUID。该会话 GUID 被输入消息 07 的报头 101 集合中。页面存储。随后查询存储器中配置文件，以确定是否存储该页面的所有内容。如果该页面需要被存储，那么语境相关信息 (报头集合、页面内容、表格内容等) 连同会话 ID 一起被输入数据库中。动作存储。实际的动作 (例如，关于 HTTP 的 GET 或 POST 命令) 也和会话 ID 一起被存储在数据库中。关于主内容捕捉的所有事件此时也被记录到数据库中 - 包括所有有关的追踪信息。

图 11 图解说明按照本发明的一个实施例的全局队列 400 的操作。在 401，从 EHU 900 接收消息 107。消息 107 被解析成子分量 (步骤 402)。在步骤 403，解析的子分量或组件 (component) 被发送给全局高速缓存 400，在步骤 404 和 405 中，针对存储的解析消息检查全局高速缓存 400。当在全局高速缓存 400 中找到解析的消息时，从全局

高速缓存 400 取回解析的消息 107, 并将其写入归档服务器 (archiver server) 501。

用户代理单元 600

通过产生指向数据收集系统 200 的命令, 人工创建用户代理 800。最好该命令中的 URI 103 包含基本域部分中的有效的应答系统 1100 子域 102 值。除了该规则之外, 用户代理单元 800 是灵活的。用户代理单元 800 具有各种各样的实现方式。例如, 可在 SEM 和横幅广告、网站上的超链接、电子邮件和各个站点上的提交物(submission)中实现用户代理 800, 仅仅列举了一些可能的实现方式。此外, 用户代理 800 能够采取二进制、TCP、通信协议、甚至由所实现的网络保证的无线/蜂窝传输地址的形式。

数据子系统 500

数据子系统 500 被用于捕捉、存储、聚集和分析数据收集系统 200 捕捉的数据。数据子系统利用从属数据收集模型, 其中一个或多个归档服务器 501 被用于快速把消息 107 从全局队列 400 转移给更永久的存储机构, 如图 12 中所示。

在一个优选实施例中, 归档服务器 501 利用关系数据库存储信息。在另一个实施例中, 信息被写成二进制文件格式, 并被存留在磁盘上。归档服务器 501 的主要用途是把存储器中的全局队列 400 消息转移到更有复原力的存储介质中。

依据系统定义的时间间隔, 中间数据库服务器 502 从一个或多个归档服务器 501 引出信息, 以便把该数据装载到站点数据仓库或数据集市(datamart)中。在一个实施例中, 归档服务器 501 采用与中间数据库服务器 502 的多对一关系。在一个优选实施例中, 归档服务器 501 采用与中间数据库服务器 502 的直接一对一关系, 在又一个实施例中, 归档服务器 501 采用与中间数据库服务器 502 的一对多关系。

此外, 在指定实施例中, 归档服务器 501 和中间数据库服务器 502 能够驻留在利用厂家软件平台的同一物理设备上。在另一个实施例中, 归档服务器 501 和中间数据库服务器 502 能够驻留在利用同一

厂家软件的分开的物理设备上。在又一个实施例中，归档服务器 501 和中间数据库服务器 502 能够采用不同的厂家软件平台，而不管它们的物理位置如何。在本节中描述的和根据该评论推测的所有实施例都被认为落入本发明的范围内。

类似地，站点数据仓库 503 能够驻留在同一或者分开的物理设备上，它可以采用与归档服务器 501 和中间数据库服务器 502 相同或不同的厂家软件平台。站点数据仓库 503 按照有利的方式存储信息，以便以各种方式分析通信量。

可选的是，就多站点操作来说，全局数据仓库 504 可被用于跨各个站点合并数据。类似地，全局数据仓库 503 可驻留在同一或分开的物理设备上，它可以采用与归档服务器 501、中间数据库服务器 502 和站点数据仓库服务器 503 相同或不同的厂家软件平台。

从而，数据收集系统 200 实现收集通过网络传输的信息的系统。数据收集系统 200 通过网络与始发系统 100 通信，从起端点服务器作用的始发系统 100 接收具有 URI 103 的消息 107。数据收集系统 200 根据输入的始发系统 100 URI 107，确定消息 107 的应答系统 1100 URI 102。数据收集系统 200 被配置成分析消息 107 的内容，并根据对初始消息 107 的分析结果，产生后续消息 107。数据收集系统 200 在利用全局队列 400 的过程中，存储始发系统 100 消息 107 的语境相关组件或分量（component），同时把后续消息 107 传送给起始发系统作用的应答系统 1100 URI 103。

语境数据

语境数据有三个主要分量（或组件）：结构、交互作用和时间。

结构与在指定网络位置发现的组件内和组件间定义相关。组件可包括(但不限于)网页、web 服务、远程访问的软件资源和公开可用的多组数据。结构包括(但不限于)如同会在网站图或系统图上发现的那样组件是如何被链接在一起的。结构还包括指定组件是如何构成的(例如，如在网页的结构或一组 API 调用的结构中那样)，以及指定组件的内容是如何呈现给用户的。本质上，结构包括从指定服务器发给用

户的一切东西。

交互作用通常被表示成从通过 DataTrendz™发明直接或间接追踪的客户机动作得到的一切。在一个实施例中，用户可向服务器发送请求或响应，这种情况下，所有信息通过 DataTrendz™体系结构，随后如同所述那样被捕捉。在另一个实施例中，异步调回机制、诸如 AJAX 或 JavaScript 之类的客户端脚本、诸如 ActiveX 控件或 Java 小程序之类的构件、或者甚至诸如工具栏和插件之类的下载组件都可被用于把关于用户交互作用的信息发送给 DataTrendz™系统。上述列举并不包括所有可能的选项，相反它只打算表示一些可能备选方案的例子。

时间指的是 DataTrendz™发明随着时间追踪结构和交互作用的能力。这能够实现用户活动的移动观察，并且能够实现获得用户行为模式和网站响应模式的能力。

通过实现这种数据的捕捉、存储和分析，DataTrendz™提供或者关于服务器的响应，或者关于各种时间相关措施观察数据的能力。

工业应用

通过其各种实施例，DataTrendz™发明可在各种行业中得到应用。本节将深入研究这些行业中的一些行业，突出通过本发明获得的提高。下面的列举不应被认为是全面的，相反只是提供本发明的典型应用例子。

DataTrendz™消除了妨碍目前的许多追踪解决方案的一些更重大的障碍。DataTrendz™提供追踪用户交互作用的能力，而不需要在应答系统上的代码。DataTrendz™也捕捉以前从未获得的数据，比如与站点结构相关的语境数据和实际表单提交值。最后，DataTrendz™能够跨域追踪用户，而不需要在始发系统上的特殊 cookies。从基于因特网/外联网的网站追踪到基于内部网的企业内容资源追踪，DataTrendz™提供显著增强的追踪用户交互作用的能力。

点击欺诈(click fraud)宽松地定义一种致力于分析活动模式，力图确定欺诈性活动的行业。点击欺诈的例子包括(但不限于)自动化

(BOT)程序、脚本化(scripted)点击模式活动和黑客服务攻击。点击欺诈分析具有获得具有网络能力的站点的静态结构数据的内容爬虫和捕捉用户动作的当前追踪解决方案之间存在差距的缺点。**DataTrendz™**提供把用户交互作用重叠在具有网络能力的站点结构之上能力,使新的数据算法能够确定欺诈性活动。在下面的章节中将更详细地涉及数据挖掘。

行为定向是应用于试图当用户遍历一组被监视站点内的不同网站时,向这些用户提供针对性商业内容的那些解决方案提供者的名称。例如,如果用户遍历汽车特许经销商网站的指定网络时,这种途径最终会确定用户对某种车辆感兴趣,从而会提供显示可供选择的汽车的广告。关于行为定向的主要挑战在于它要求系统跨网站追踪用户。在**DataTrendz™**之前,这意味或者使用多数浏览器默认禁用的第三方 cookies,或者厂家不得不设法根据分开地收集的数据使用户信息相关联。**DataTrendz™**的跨网站实际跟随用户的能力能够实现在目前的市场中未实现的实时行为定向。

搜索引擎优化(SEO)公司试图确定利用比如外部链接、字典排列之类的事物,尽可能高地移动客户机的自然搜索结果的各种手段。这全部都是确定哪些搜索引擎在任何时候都最有价值的努力。这些努力的主要缺点是确定原因和效果的手段不直接。这些解决方案能够获得用户交互作用,但是它们不能同时获得站点结构。例如,指定的解决方案可能能够确定用户访问过指定页面,不过它不能确定该页面上的确切内容。由于**DataTrendz™**获得语境数据,因此在连续的迭代中探索不同的可能途径的条件下,能够实时进行SEO。

搜索引擎营销(SEM)描述致力于利用关键字层面的自然搜索结果的相关付费广告的排列的行业。SEM公司的目标之一是增大目标网站的销售或线索。多数SEM提供存在众多的限制,包括不能直接报告用户内容(即,表单提交数据),和不能直接把搜索引擎内容和所得到的访问者动作联系起来。**DataTrendz™**位于搜索引擎和目标网站之间,能够把用户交互作用和搜索引擎活动联系起来。本发明的基于因

特网的实施例是搜索引擎的有用配合，因为 DataTrendz™为 SEM 公司提供相当数量的语境信息。

这样大量的进行中的语境信息的收集还提供单一的市场信息仓库。通过利用创新的数据挖掘算法，DataTrendz™将能够提供以前不能获得的市场分析和预测能力。

联盟营销描述商家的使其它在线商家能够代表该商家做广告的实践。联盟营销建立在跨许多商家网站追踪用户活动，以便核实购买和其它用户动作的能力之上。过去这一直是极其困难的过程，要求联盟网络和商家的漫长努力。通过消除放在每个商家的站点上放置代码的需要，DataTrendz™消除这些障碍中的许多障碍。此外，由于多数联盟营销网络通过一系列的 HTTP 重定向过程传递通信量，因此，通过利用冗余网络解决方案消除这种重定向，DataTrendz™实际上将降低网络可视性，同时提高稳定性和追踪能力。

数据挖掘

一旦系统收集了语境数据，就进行有目的的分析，以便获得额外的商业和战略见识。这种分析通常被称为分布式数据挖掘。分布式数据挖掘技术目前应用于各种数据类型。尽管本领域的技术人员可选择使用他们自己偏爱的实现方法，不过一种优选的途径是首先把语境数据的功能组件重叠在结构要素之上，以便逐步形成、可视化和更好地理解数据能够支持的语境和潜在商业目的或其它目的。该过程一旦完成，就利用时间分片算法沿时间轴存储结构化的功能数据。

在恰当地结合和存储这组新颖的数据的情况下，数据挖掘方面被证实的公知理论方法可被用于定义使用模式、序列模式、活动的模式、确定新的或增长的影响点，以及得到关于上面提及的一些或全部的市场可变性，最终得到未来预测。

通过利用这些新的模式，通过对照模拟结果和加权的影响点，测量活动的周期性波动，二次分析分揭示另外的关心点。这些周期性波动可由任意时间段构成，包括(但不限于)与时间相关的周期性、区域特性、网络位置信息和/或用户属性。交互作用可包括这些波动与归因

于该数据的任意单一或多重的数据属性的任意组合。例如，关心的可能组合可以是美国东北部妇女购买家庭用品的习惯的月度波动。

利用这些二次分析，显现成为诸如搜索引擎优化(例如，内容的什么特性使网站更有效)或者企业内容管理(例如，当利用指定的分类学上层管理组织内容时，发现内容更有效或者不太有效)之类工作的基础的其它活动模式。此外，背离标准活动模式能够实现影响分析的发展，影响分析能够在诸如点击欺诈分析之类工作中达到极至。

动态 cookies

成为本系统的基础的更有趣的创新之一专注于解决跨越对目标网站的多次访问追踪访问者的问题。为了实现跨天、跨周、跨月和跨年追踪个人的能力，本发明的一个实施例利用动态 cookies 的构思来处理对指定网站的后续访问。

动态 cookies 是能够手动下载、自动安装或按它们的某种组合下载安装到用户计算机上的小型实用程序。该实用程序对要追踪的用户访问网站的内部列表施加影响，同时监视用户的网络活动。

每当用户重新访问指定的网站时，该实用程序自动把用户重定向到 DataTrendz™系统，在 DataTrendz™系统中，追踪被重新初始化。在一个优选实施例中，除这种自动重定向功能之外，动态 cookies 不与用户的计算机交互作用，也不能够实现任何其它动作。

在一个实施例中，该实用程序可采取监视关于指定 web 浏览器的所有网络流量的浏览器插件、ActiveX 或 Java 小程序的形式。这些对象被认为是例子，而不是对本发明的限制。在另一个实施例中，作为响应的一部分，DataTrendz™会向始发系统发送一个可执行文件。可以图像或者避免用户安全问题的其它某种文件格式的形式嵌入该可执行文件。该可执行文件随后会按照与当前的 cookie 技术类似的方式，把它自己装在用户计算机上，从而监视通信量。这些是可如何实现动态 cookies 的两个例子，只是作为例子提供的。它们不应被视为可能的备选实现方案的穷举，所有其它备选方案都在本发明的范围之内。

尽管举例说明了本发明的优选实施例，不过如上所述，在不脱离本发明的精神和范围的情况下，可做出许多变化。因此，本发明的范围不受优选实施例的公开内容限制。相反，本发明应完全由下面的权利要求限定。

本申请要求早先于 2007 年 4 月 17 日提交的美国临时专利申请 No. 60/912,203 的优先权。通过引用将其全部内容并入于此，就好像在此进行了全面阐述一样。

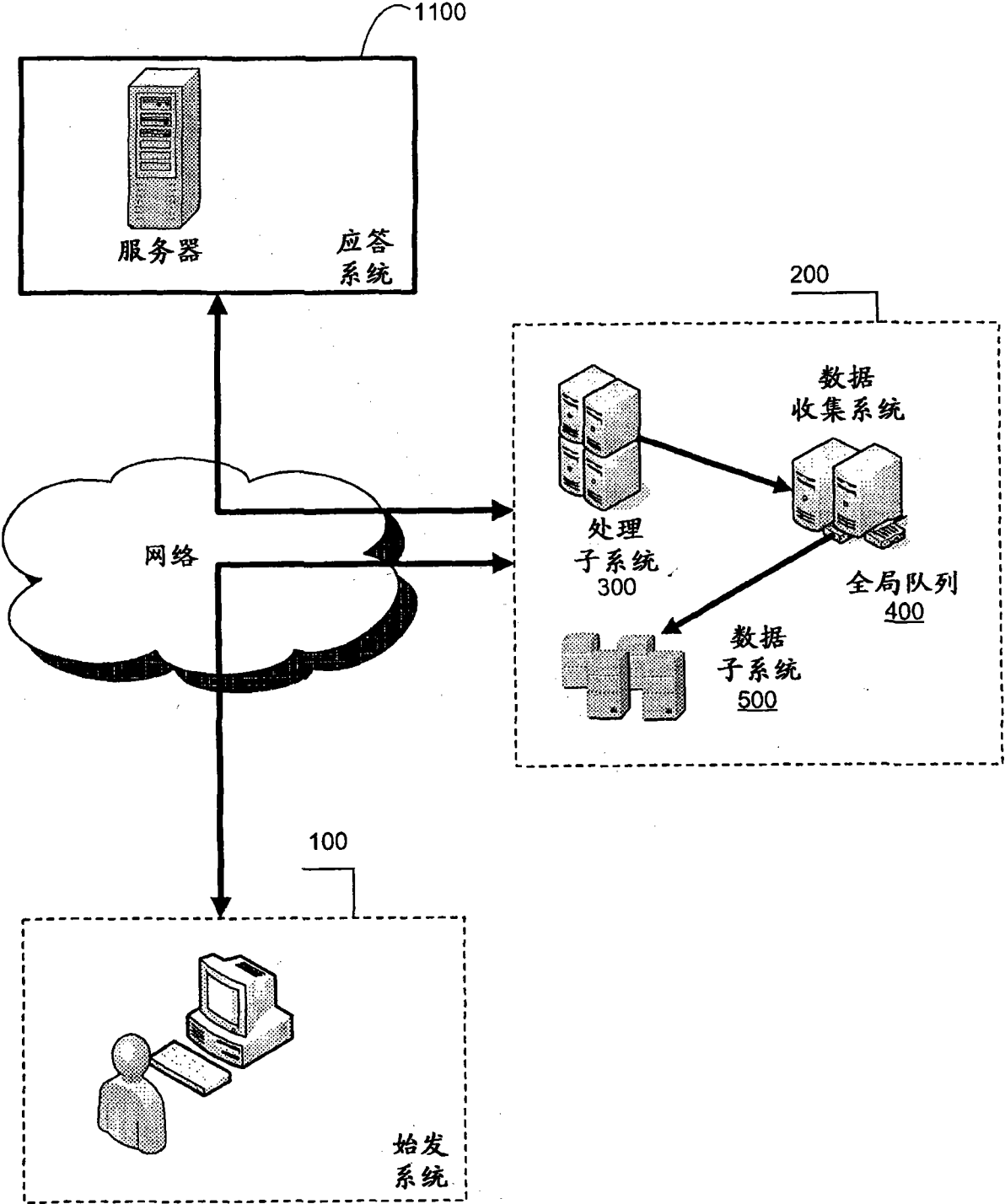


图1

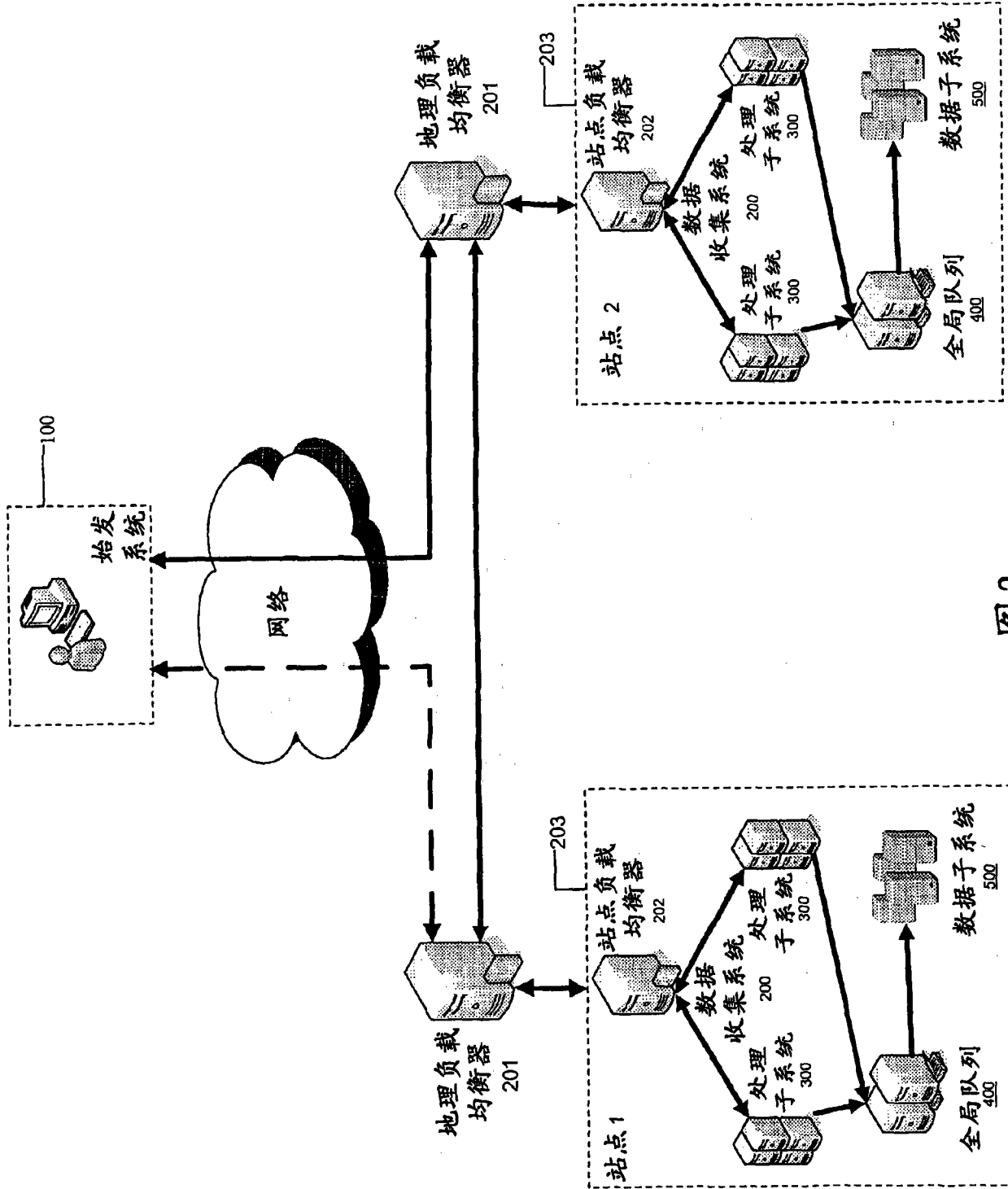


图 2

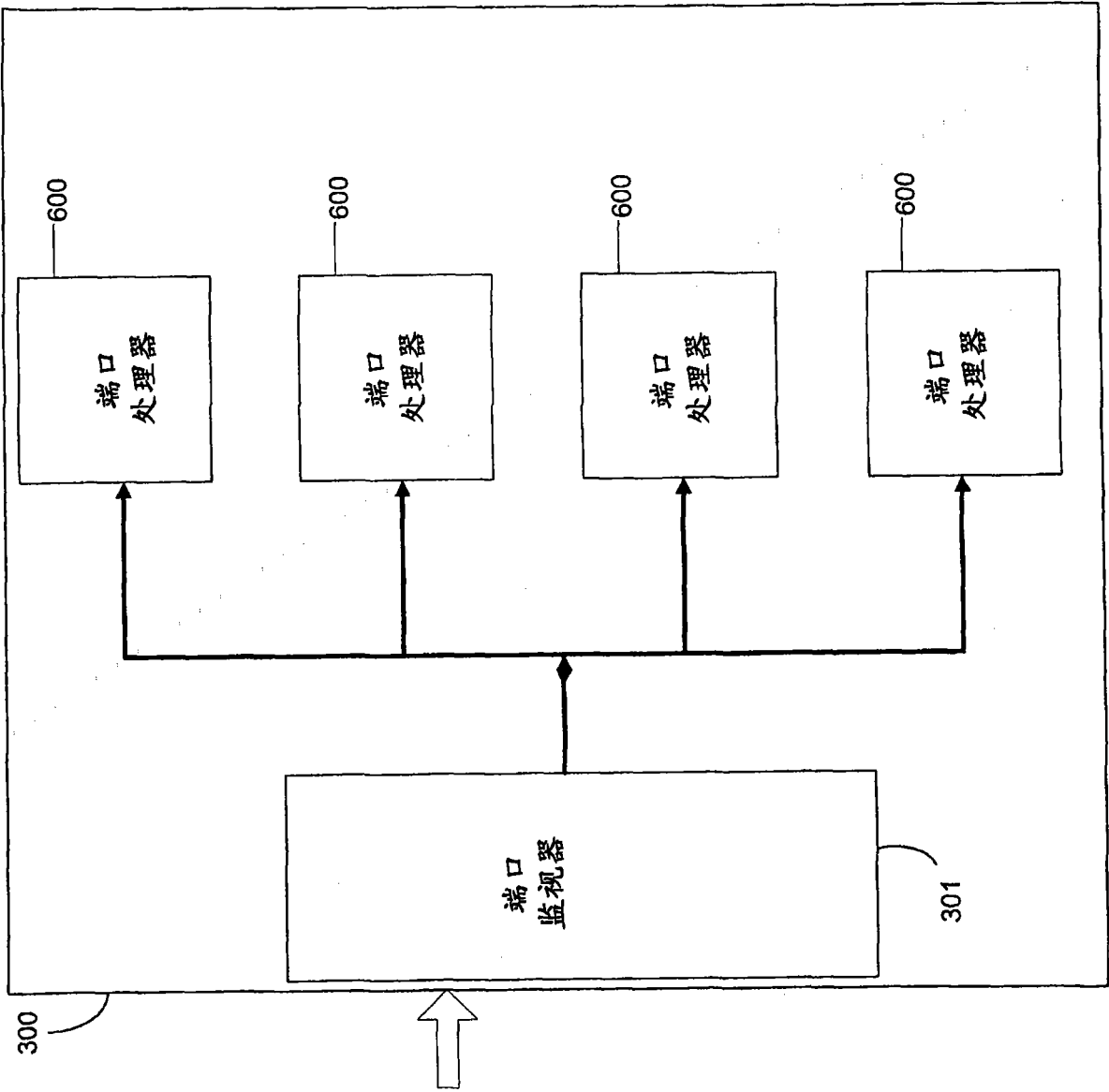


图3

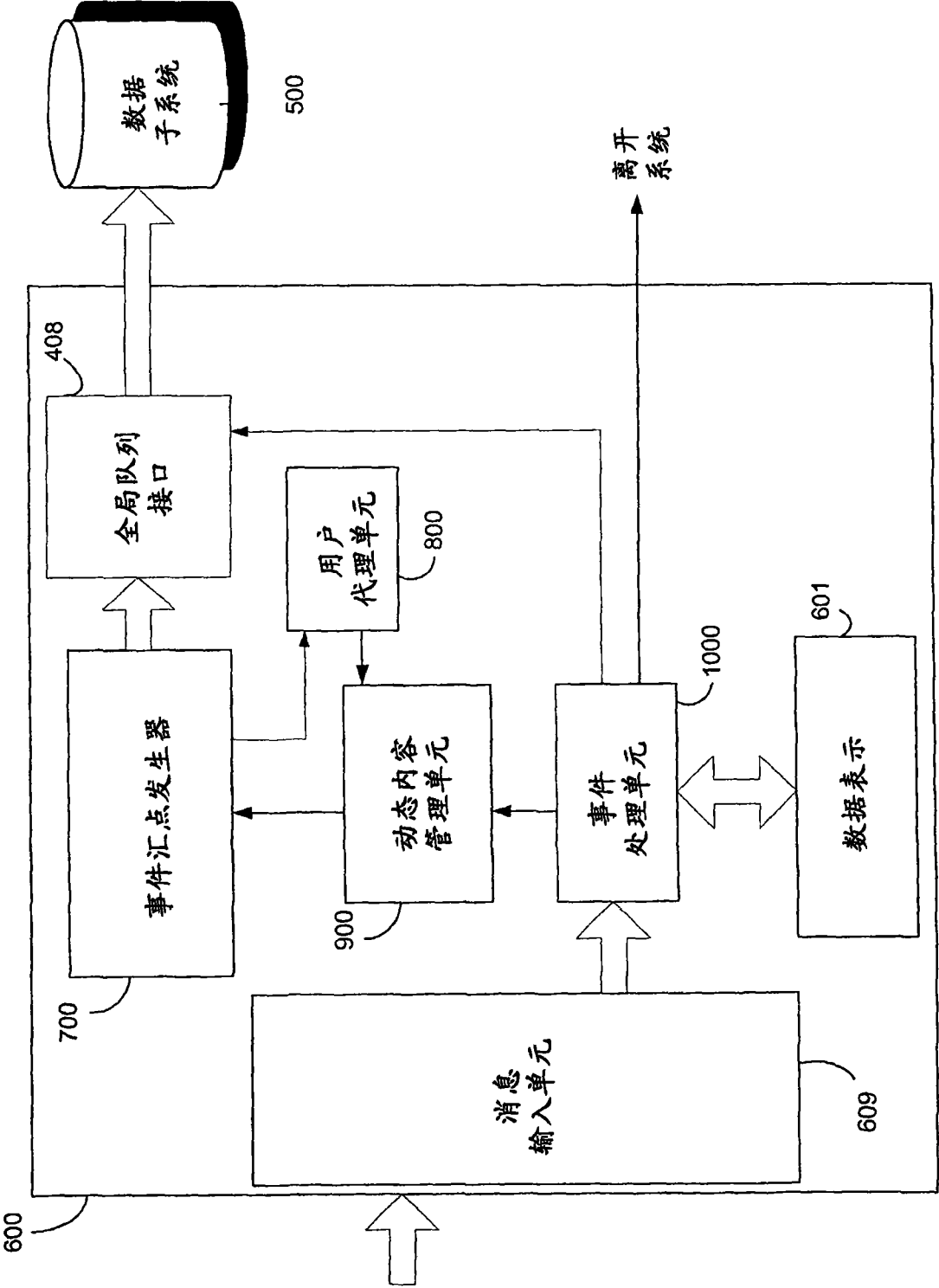


图4

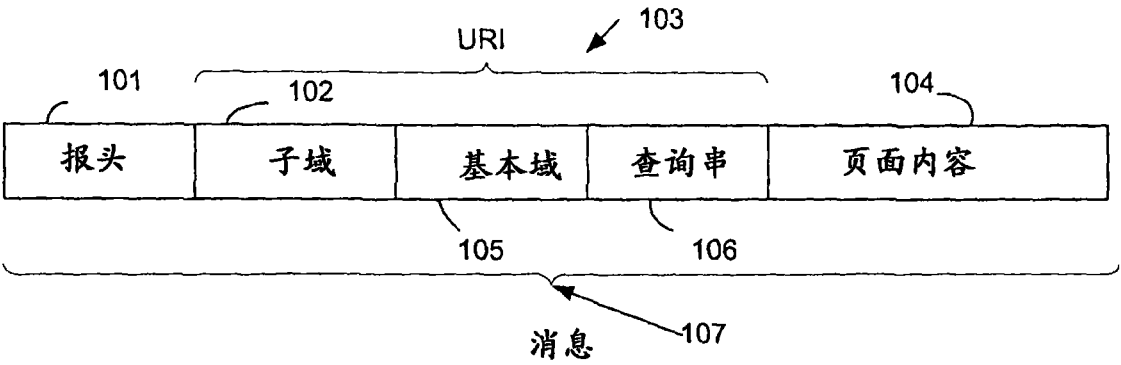


图5

存储器中XML文件	
子域	目标 URI
SD值 (目标域a)	目标 URI a
SD值 (目标域b)	目标 URI b

图 6

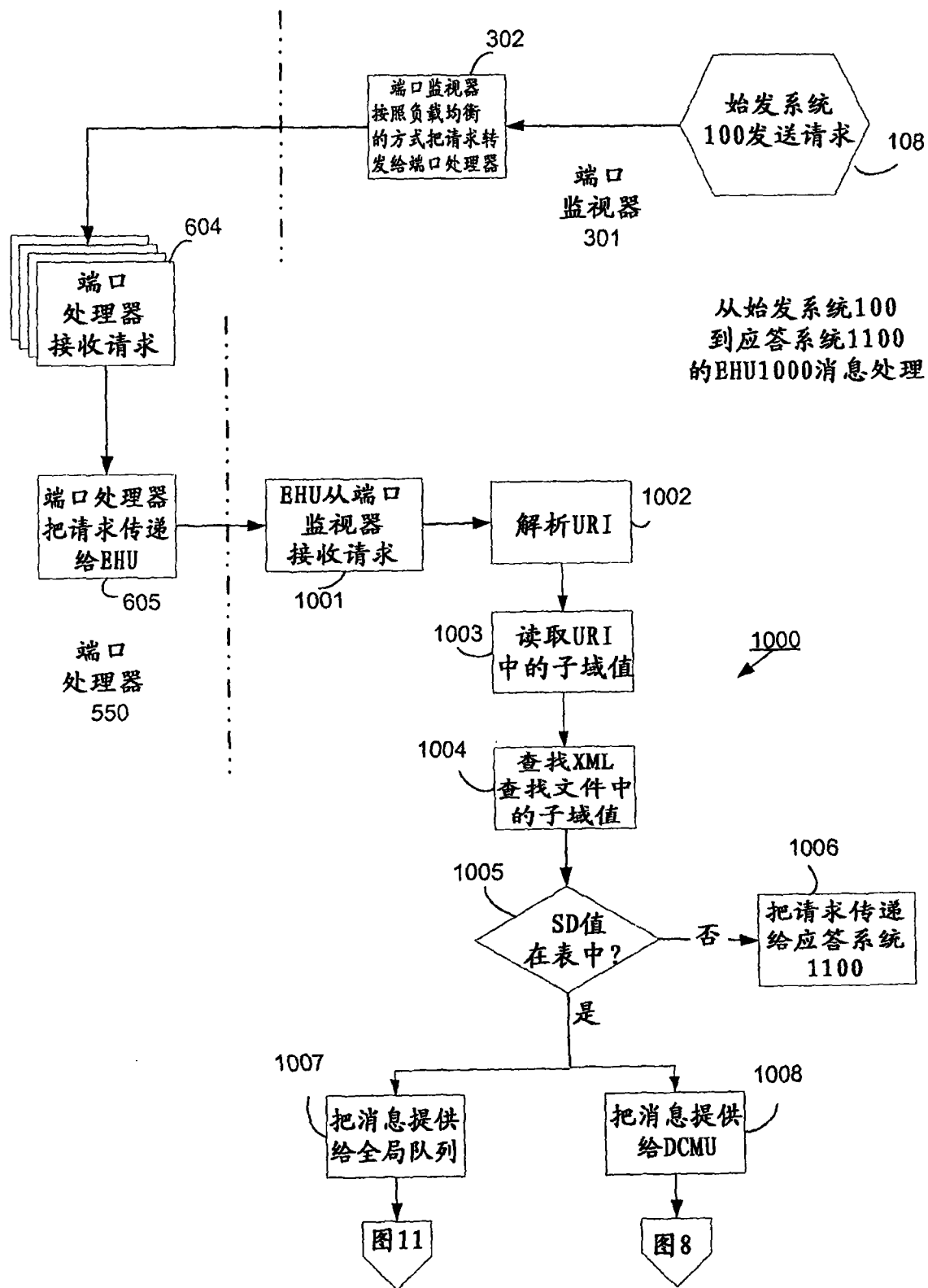


图7

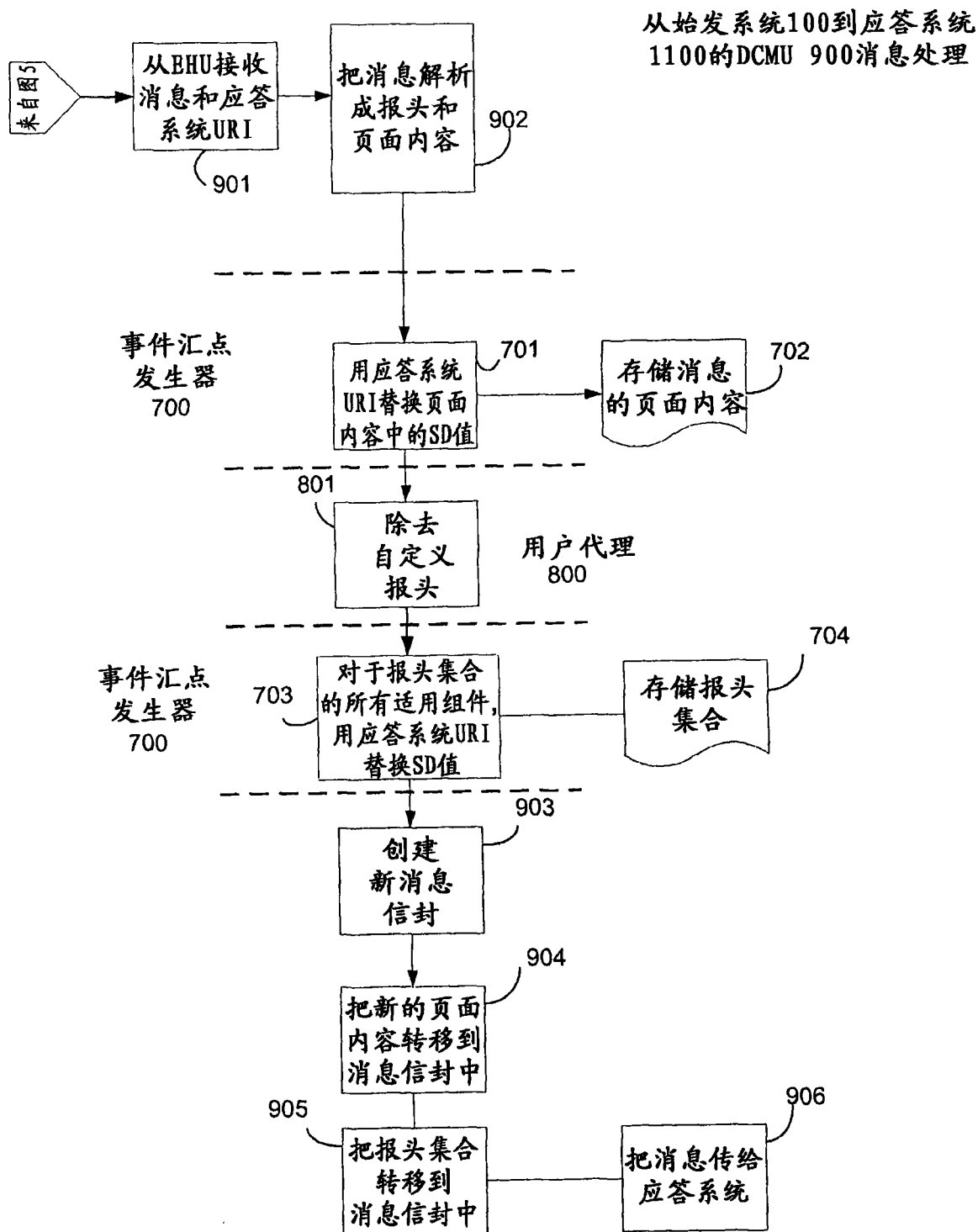


图8

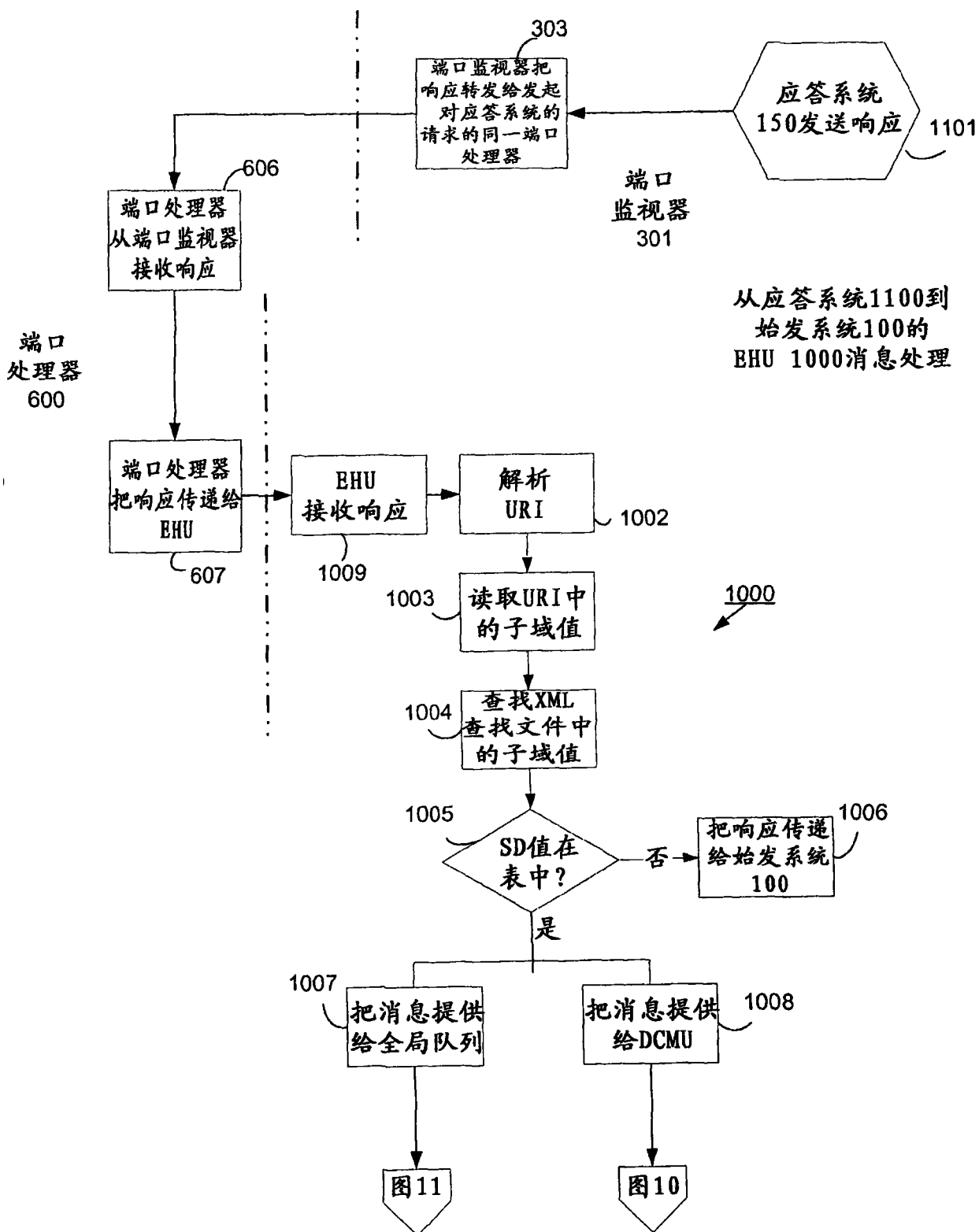


图 9

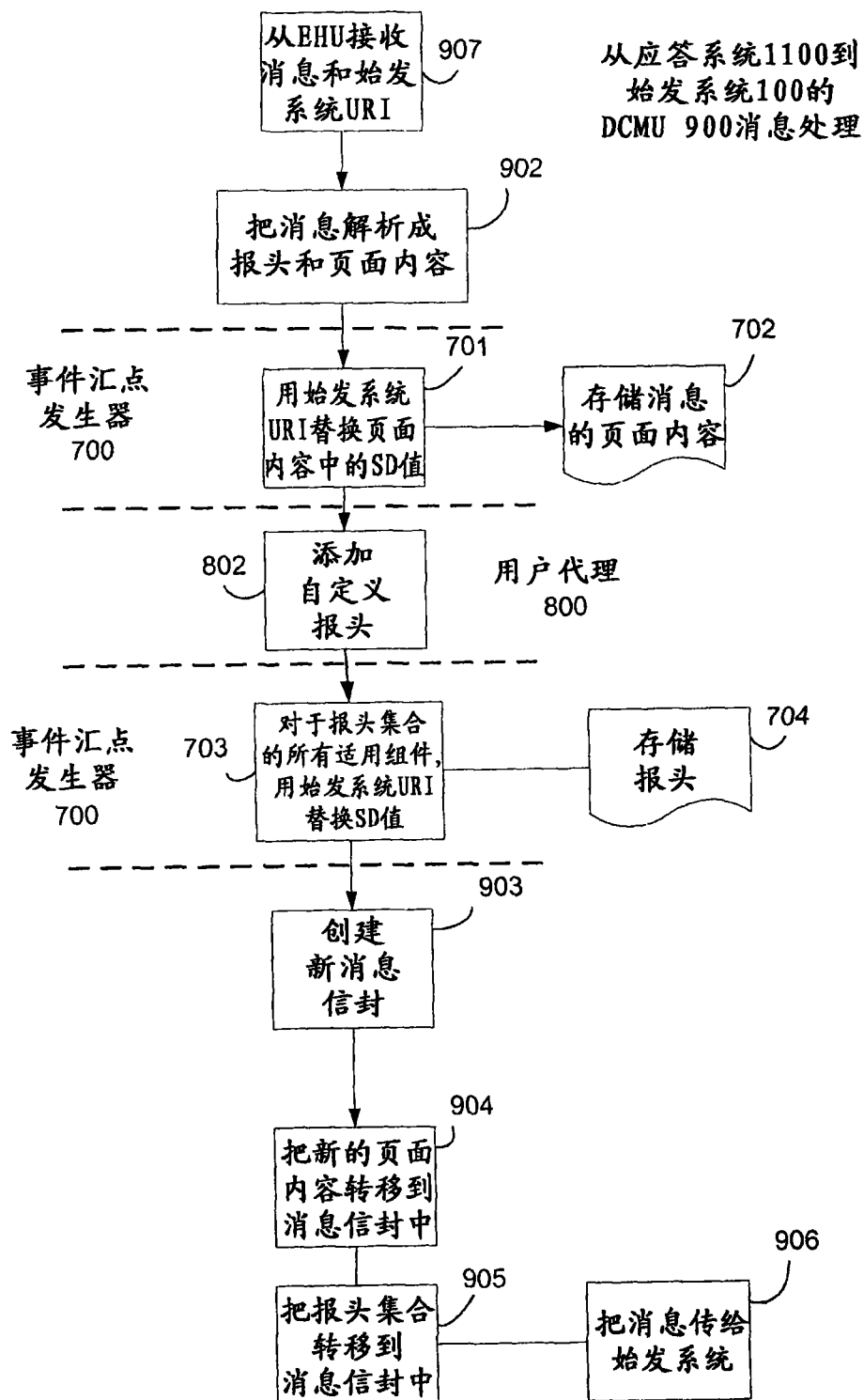


图10

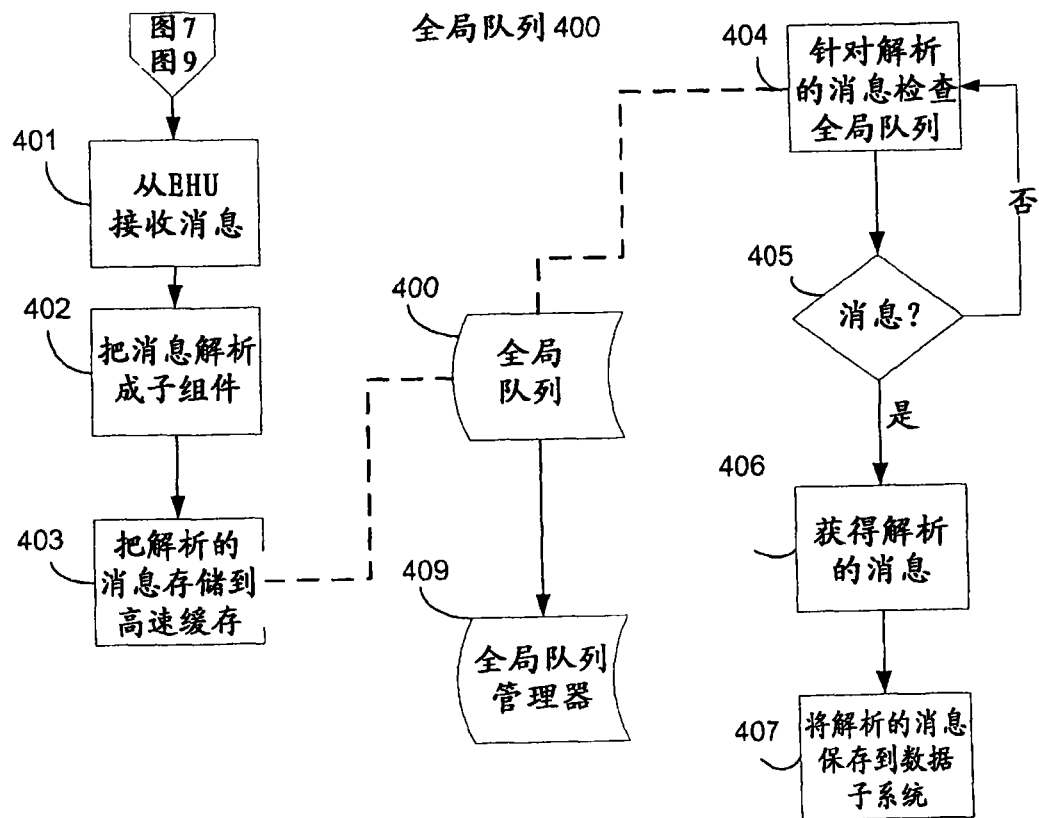


图11

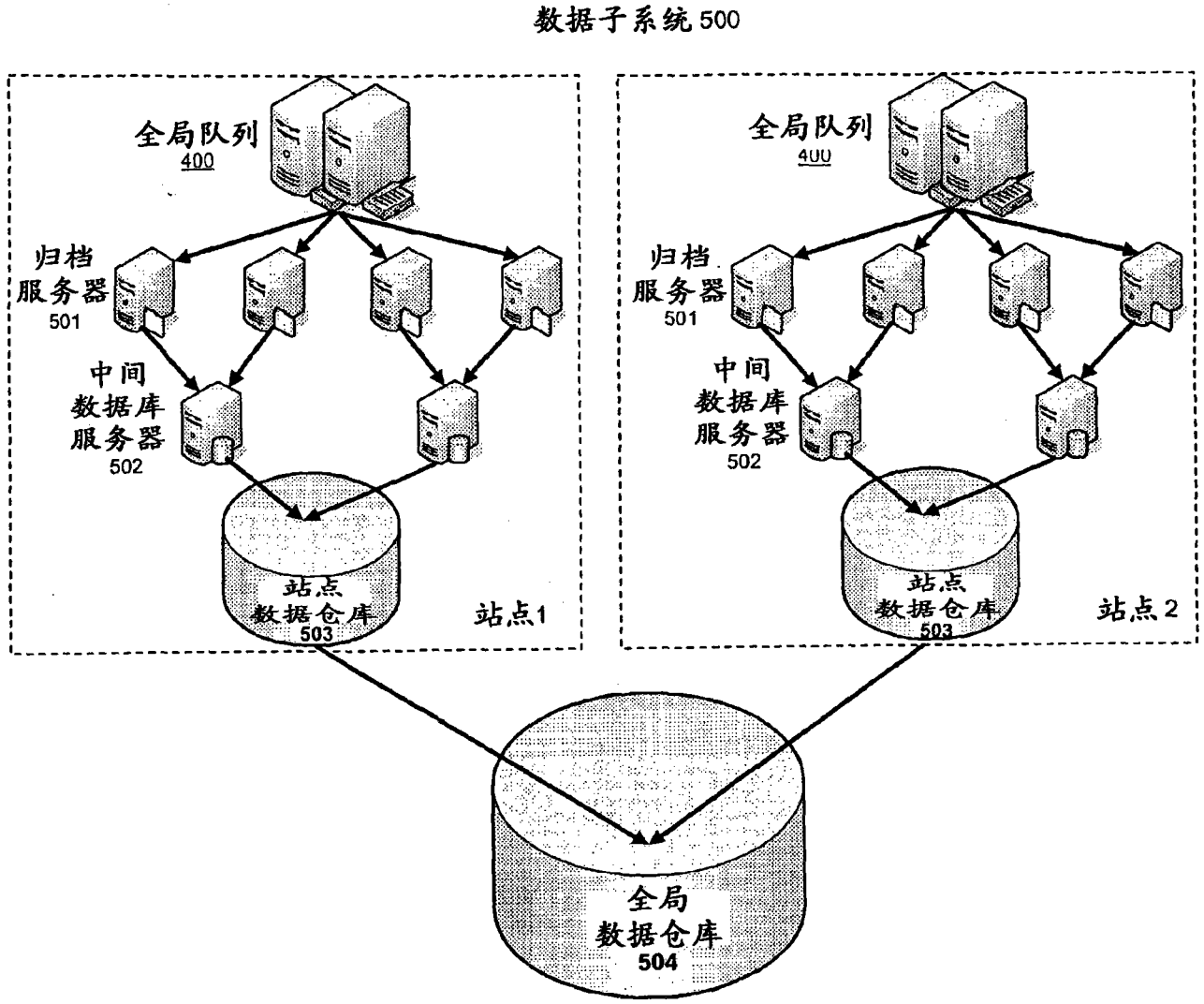


图12