(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 102902693 B
(45)授权公告日 2018.01.12

(21)申请号 201110215012.2

(22)申请日 2011.07.29

(65)同一申请的已公布的文献号
申请公布号 CN 102902693 A

(43)申请公布日 2013.01.30

(73)专利权人 慧与发展有限责任合伙企业
地址 美国德克萨斯州

(72)发明人 H-M.侯 J-M.金 L-M.焦 S.H.麟

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 刘春元 王洪斌

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

(56)对比文件
US 2004249979 A1,2004.12.09,
US 2005038635 A1,2005.02.17,
CN 101443751 A,2009.05.27,

审查员 高丹丹

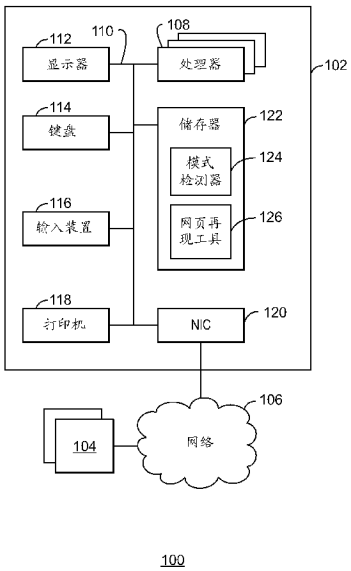
权利要求书2页 说明书5页 附图8页

(54)发明名称

检测在网页上的重复模式

(57)摘要

本发明涉及检测在网页上的重复模式。本技术的示例性实施例可以生成DOM-树以及基于DOM-树和节点列表生成信号。可以分析所述信号以及可以在所述信号中选择节点以形成周期波。可以使用所述周期波和所述节点来检测重复模式。



CN 102902693 B

1.一种用于检测网页上的重复模式的系统,所述系统包括:

处理器,所述处理器适于执行所存储的指令;以及

存储指令的存储器装置,所述存储器装置包括处理器可执行代码,所述处理器可执行代码当由所述处理器执行时,适于:

生成DOM-树;

基于所述DOM-树和节点列表来生成信号,其中所述节点列表是所述DOM-树中的节点以它们在所述DOM-树中被遍历的次序的列单;

分析所述信号;

在所述信号中选择节点以形成周期波;以及

使用所述周期波和所述节点来检测重复模式。

2.根据权利要求1所述的系统,其中,所述节点列表包括DOM-树中的每个节点的节点深度。

3.根据权利要求1所述的系统,其中,所述存储器存储处理器可执行代码,所述处理器可执行代码适于通过如下步骤来基于DOM-树和节点列表生成所述信号:

从DOM-树获得节点列表,其中,所述节点列表包括DOM-树中的每个节点的节点深度;以及

基于所述节点列表中的叶节点以及所述节点深度来生成1D信号,其中,每个节点对应于所述信号上的点,其中该点的x坐标对应于该节点的遍历次序以及y坐标对应于所述节点深度。

4.根据权利要求1所述的系统,其中,所述存储器存储处理器可执行代码,所述处理器可执行代码适于通过如下步骤基于DOM-树和节点列表生成信号:

从DOM-树获得节点列表,其中,所述节点列表包括DOM-树中的每个节点的节点深度;以及

基于节点列表中的叶节点以及节点深度生成1D信号,其中,每个节点对应于所述信号上的点,其中该点的x坐标对应于该节点的遍历次序,以及y坐标对应于节点深度乘以节点特性得分。

5.根据权利要求1所述的系统,其中,所述存储器存储处理器可执行代码,所述处理器可执行代码适于通过如下步骤基于DOM-树和节点列表生成信号:

从DOM-树获得节点列表,其中,所述节点列表包括DOM-树中的每个节点的节点深度;以及

基于节点列表中的叶节点以及节点深度生成2D信号,其中,每个节点对应于该信号上的点,其中该点的x坐标对应于该节点的遍历次序,y坐标对应于节点深度,以及z坐标基于节点特性得分。

6.根据权利要求1所述的系统,其中,所述存储器存储处理器可执行代码,所述处理器可执行代码适于使用时间-频率分析技术来分析信号,所述时间-频率分析技术诸如快速傅立叶变换、数字小波变换或自相关。

7.根据权利要求1所述的系统,其中,通过过滤掉不传达任何信息或不满足特定阈值的节点,来形成周期波。

8.一种用于检测网页上的重复模式的方法,所述方法包括:

从由web浏览器或网页再现工具再现的网页，生成DOM-树；

基于DOM-树和节点列表生成信号，其中所述节点列表是DOM-树中的节点以它们在DOM-树中被遍历的次序的列单；

分析所述信号；

在所述信号中选择节点以形成周期波；以及

使用所述周期波和所述节点来检测重复模式。

9.根据权利要求8所述的方法，其中，所述节点列表包括DOM-树中的每个节点的节点深度。

10.根据权利要求8所述的方法，其中，基于DOM-树和节点列表生成信号包括：

从DOM-树获得节点列表，其中，所述节点列表包括DOM-树中的每个节点的节点深度；以及

基于节点列表中的叶节点以及节点深度生成1D信号，其中，每个节点对应于所述信号上的点，其中该点的x坐标对应于该节点的遍历次序以及y坐标对应于节点深度。

11.根据权利要求8所述的方法，其中，基于DOM-树和节点列表生成信号包括：

从DOM-树获得节点列表，其中，所述节点列表包括DOM-树中的每个节点的节点深度；以及

基于节点列表中的叶节点以及节点深度生成1D信号，其中，每个节点对应于所述信号上的点，其中该点的x坐标对应于该节点的遍历次序以及y坐标对应于节点深度乘以节点特性得分。

12.根据权利要求8所述的方法，其中，基于DOM-树和节点列表生成信号包括：

从DOM-树获得节点列表，其中，所述节点列表包括DOM-树中的每个节点的节点深度；以及

基于节点列表中的叶节点以及节点深度生成2D信号，其中，每个节点对应于所述信号上的点，其中该点的x坐标对应于该节点的遍历次序，y坐标对应于节点深度，以及z坐标基于节点特性得分。

13.根据权利要求8所述的方法，使用时间-频率分析技术来分析所述信号，所述时间-频率分析技术诸如快速傅立叶变换、数字小波变换或自相关。

14.根据权利要求8所述的方法，其中，通过过滤掉不传达任何信息或者不满足特定阈值的节点，来形成周期波。

检测在网页上的重复模式

技术领域

[0001] 本公开涉及检测在网页上的重复模式的系统和方法。

背景技术

[0002] 通常使用固定的模板或模式来再现网页上的信息。模式可能在网页上重复地出现，并经常被称为重复模式。可以基于在网页上找到的模式，对网页进行分割。例如，片段可以是导航条、头部、尾部、广告、相关链接、版权信息或实际网页内容自身。识别在网页中的模式在很多应用中是有用的，所述应用诸如在小屏幕装置上显示网页、数据挖掘、搜索引擎以及打印装置。进一步地，识别重复模式可以提供关于网页设计、网页结构以及网页上包含的内容的信息。

[0003] 为了从网页识别并检索内容，网页分割算法可以对相似元素进行聚类。在这些算法中，可能不对重复的元素组进行聚类，因为重复的元素可能根本不相似。因此，重复模式可能无法在聚类的元素中检测到，以及由重复模式传达的信息可能丢失。

发明内容

[0004] 根据本公开的第一方面，提供了一种用于检测网页上的重复模式的系统，所述系统包括：处理器，所述处理器适于执行所存储的指令；以及存储指令的存储器装置，所述存储器装置包括处理器可执行代码，所述处理器可执行代码当由所述处理器执行时，适于：生成DOM-树；基于所述DOM-树和节点列表来生成信号，其中所述节点列表是所述DOM-树中的节点以它们在所述DOM-树中被遍历的次序的列单；分析所述信号；在所述信号中选择节点以形成周期波；以及使用所述周期波和所述节点来检测重复模式。

[0005] 根据本公开的第二方面，提供了一种用于检测网页上的重复模式的方法，所述方法包括：从由web浏览器或网页再现工具再现的网页，生成DOM-树；基于DOM-树和节点列表生成信号，其中所述节点列表是DOM-树中的节点以它们在DOM-树中被遍历的次序的列单；分析所述信号；在所述信号中选择节点以形成周期波；以及使用所述周期波和所述节点来检测重复模式。

附图说明

[0006] 参考附图在以下的详细说明中对某些示例性实施例进行说明，在这些附图中：

[0007] 图1是根据本技术的实施例的、可以检测在网页上的重复模式的系统的框图；

[0008] 图2是根据本技术的实施例的、用于检测在网页中的重复模式的方法的过程流程图；

[0009] 图3是示出根据本技术的实施例的、具有节点的网页的一部分的图示；

[0010] 图4是示出根据本技术的实施例的、网页的DOM-树的一部分的图示；

[0011] 图5是示出根据本技术的实施例的、用于网页的1D信号的图示；

[0012] 图6是示出根据本技术的实施例的、对信号进行信号分析的结果的框图；

[0013] 图7是示出根据本技术的实施例的、对于网页的所检测的重复模式的标记选择结果的图示；

[0014] 图8是示出根据本技术的实施例的、存储用于检测在网页上的重复模式的代码的非暂时的(non-transitory)、计算机可读介质的框图。

具体实施方式

[0015] 检测网页上的重复元素使得重复元素能够被分组(group)为重复模式。一个实施例包括系统，所述系统能够使用信号分析方法来检测在网页上的重复模式，包括使用树数据结构的网页文档对象模型(DOM)生成信号。DOM是用于表示各种标记语言文档中的对象以及与所述对象交互的跨平台且与语言无关的协定。DOM的各方面(诸如其元素)可以被寻址以及操纵。元素是所使用的特定标记语言的单独组件。DOM-树将这些元素再现为树中的节点。节点也可以对应于驻留在网页上的小的数据单元。

[0016] 各种用于网页分割的技术能够使用树匹配算法来识别重复模式，以及然后使用对齐信息来过滤掉不想要的数据。可以通过在DOM-树中使用自下至上的次序遍历每个节点，从局部最优解获得全局最优解。但是，自下至上遍历是递归的，以及这种递归计算可能是耗时的。进一步地，如果重复模式没有被完全显示则它们可能不被检测到，使得一个子树不包含模式的一些节点，但是实际上是网页的模式。

[0017] 用于网页分割的其他技术可以使用哑元树(dummy tree)匹配算法，以通过检查在DOM-树的所有层中的独特标签以及然后比较独特标签的总数，来查验DOM-树中的数据记录的相似性。但是，当一个子树不包含模式的所有节点时，这种技术也可能出现问题。类似地，如果数据记录具有不同的属性，则使用视觉一致性来定位和提取模式或数据区域可能效果不好。

[0018] 在实施例中，可以以鲁棒方式来检测重复模式，而不管重复模式中的节点数量如何或者数据记录是否具有不同的属性。另外，即便没有在网页上完全显示重复模式也可能检测到重复模式。进一步，信号分析技术，诸如快速傅立叶变换(FFT)、数字小波变换(DWT)、自相关或任何其他时间-频率分析技术可以用于分析该信号。通过本技术，web重复模式检测问题可以被作为信号分析问题进行求解，其中，信号分析技术被用于获得准确且鲁棒的结果。因为重复模式可以被用于分割网页，所述结果在网页打印以及web内容提取中可以是有用的。

[0019] 图1是根据本技术的一个实施例的、可以检测在网页上的重复模式的系统的框图。该系统通常用参考数字100来表示。本领域普通技术人员将认识到，图1中所示的功能框和装置可以包括：硬件元素，包括电路；软件元素，包括在有形的、机器可读介质上存储的计算机代码；或者硬件元素与软件元素的组合。另外，系统100的功能框和装置不过是可以在一个实施例中实现的功能框和装置的一个示例。本领域普通技术人员将容易地能够根据特定电子装置的设计考虑而定义具体的功能框。

[0020] 系统100可以包括经由网络106通信的服务器102以及一个或多个客户端计算机104。如图1中所示，服务器102可以包括一个或多个处理器108，其可以通过总线110连接到显示器112、键盘114、一个或多个输入装置116以及输出装置，诸如打印机118。输入装置116可以包括诸如鼠标或触摸屏的装置。处理器108可以包括单核、多核或者在云计算体系架构

中的核集群。服务器102还可以通过总线110连接到网络接口卡(NIC)120。NIC 120可以将服务器102连接到网络106。

[0021] 网络106可以是局域网(LAN)、广域网(WAN)或另一网络配置。网络106可以包括路由器、交换机、调制解调器或用于互连的任何其他类型的接口装置。网络106可以连接到数个客户端计算机104。通过网络106,数个客户端计算机104可以连接到服务器102。客户端计算机104可以与服务器102的结构类似。

[0022] 服务器102可以具有通过总线110可操作地耦合到处理器108的其他单元。这些单元可以包括有形的、机器可读存储介质,诸如储存器122。储存器122可以包括硬盘驱动器、只读存储器(ROM)、随机存取存储器(RAM)、RAM驱动器、闪速驱动器、光学驱动器、高速缓冲存储器等的任意组合。储存器122可以包括模式检测器124和网页再现工具126。模式检测器124可以从网页生成DOM-树。可以使用网络106对网页进行访问。网页可以使用web浏览器或网页再现工具126被再现在显示器112上。网页再现工具126可以允许web设计师来验证网站设计的方面。

[0023] 模式检测器124也可以基于DOM-树和节点列表来生成信号。节点列表是DOM-树中的节点以它们在DOM-树中被访问或遍历的次序的列单。模式检测器124也可以分析该信号并选择信号中的节点来形成周期波。根据周期波,模式检测器124可以使用周期波和节点来检测重复模式。

[0024] 图2是根据本技术一个实施例的、检测在网页中的重复模式的方法200的过程流程图。在框202处,可以生成DOM-树。诸如网页再现工具126(图1)的网页再现工具可以被用于生成网页的DOM-树。

[0025] 在框204处,可以生成信号。该信号可以通过使用任何树遍历方法(诸如前序遍历)遍历DOM-树,而基于该DOM-树。通常,树遍历指代以有秩序的(methodical)方式访问树数据结构中的每个节点的过程。遍历过程可以根据对每个节点进行访问或遍历的次序进行改变。当前序遍历树数据结构时,首先访问根节点,接着是左子树然后是右子树。

[0026] 通过遍历DOM-树,获得DOM-树的节点列表。如上所述,节点列表是按照在DOM-树中对节点进行遍历的次序的、所述节点的列单。在DOM-树中,叶节点可以对应于实际的网页内容信息,诸如文本、图像以及视频。其他子树节点,诸如具有孩子的节点,包含网页的结构以及风格信息。节点列表可以包括节点深度,所述节点深度表示在DOM-树中的每个节点的深度。节点列表中的叶节点可以与从DOM-树获得的节点深度一起使用,以形成1D信号。对于1D信号来说,信号中的点的x坐标可以对应于DOM-树中的节点的遍历次序,而y坐标可以对应于在DOM-树中的相同节点的节点深度。这样的1D信号在图5中示出并在本文进行进一步说明。

[0027] 节点特性得分可以用于优化1D信号。可以通过为节点的特性设置得分来计算节点的特性得分,所述特性包括但不局限于标签信息、文本字体以及位置坐标。标签可以对应于在标记语言文档中嵌入的编码指令。Web浏览器可以读取标签以在诸如显示器112(图1)的显示器上再现网页。可以将不同特性的得分加到一起来计算节点特性得分。也可以通过生成标签串的哈希值作为其得分,来计算节点特性得分。为了优化1D信号,可以将节点深度乘以节点特性得分,以形成1D信号的y坐标。在实施例中,可以使用节点特性得分来形成z轴以生成2D信号。因此,2D信号的z坐标可以基于节点特性得分。

[0028] 在框206处，可以分析该信号。该信号可以是1D或2D信号，以及可以使用包括但不限于FFT、DWT或自相关的技术进行分析。信号分析可以将信号变换到时频域，其中，重复的频率值可以被用于提取周期波。信号分析的结果可以按信号的位置、波长以及周期进行记录。所记录的结果可以用于形成周期波。

[0029] 在框208处，可以选择子树节点。对于所提取的周期波中的每个“波”，找到DOM树中包括对应于该特定波的所有叶节点的最小子树。为了选择每个子树，可以将提取的周期波从时频域变换回到1D信号以及与原始的DOM-树进行比较。

[0030] 不传达任何信息的节点或者不满足特定阈值的节点可以被过滤掉或忽略。例如，通过使用网页再现工具来为每个节点生成包围盒，可以使用阈值来忽略在高度或宽度上小于10个像素的节点。通常，这样的小节点具有很少有用的网页内容。

[0031] 发现的包含周期波的特定波的叶节点的、每个子树的父或根节点能够被用于构造在重复模式中找到的叶节点。如上所述，叶节点典型地包含网页的内容，而具有孩子的节点可能包含结构和风格信息。在父节点以及其他子树节点中找到的结构和风格信息可以被用于构造在叶节点中找到的内容。

[0032] 在框210处，检测重复模式。在找到的每个子树中，孩子节点可以形成重复模式。通过检测重复模式，即便一些模式不完全匹配实际的重复模式，也可以对网页进行鲁棒分割。以此方式，分割后的网页能够被用于在其中诸如在小显示装置或打印装置上再现所有网页片段可能是不希望的场景下，再现网页的实际内容。

[0033] 图3是示出根据本技术一个实施例的、具有节点的网页300的一部分的图示。图3中的节点被虚线矩形所包围，所述虚线矩形表示每个节点的边界矩形。如由在本文中所述技术确定的，重复模式位于节点302和304处。用于每个节点的数字在该节点的边界矩形中被圈出，诸如在参考数字306处的数字1626，以及在参考数字308处的数字1554。每个节点的数字是在DOM-树中的该节点的遍历次序，以及每个节点在图3和图4中具有相同的数字。为了便于说明，仅仅示出少许节点。但是，网页300可以具有任何数量的节点。

[0034] 用于再现参考数字310、312和314处的节点的标记语言是类似的，赋予节点相似的节点特性。例如，在节点316、318和320处的图像具有相同的尺寸。同样，节点322、324和326的文本字体是相同的。因此，显然在标记语言中使用固定的模板来在参考数字310、312和314处再现内容，以及参考数字310、312和314形成重复模式。

[0035] 在参考数字310处的模式包含四个节点，而在参考数字312和314处的模式每个包含三个节点。即使在参考数字310处的模式可能具有比在参考数字312和314处的模式少的节点，本技术也能够识别在模式之间的相似性且检测出重复模式。

[0036] 图4是示出根据本技术一个实施例的、网页的DOM-树400的一部分的图示。所述DOM-树可以根据在框202(图2)处描述的技术来生成。DOM-树400基于网页300(图3)。能够通过分析根据DOM-树400生成的信号来检测网页300(图3)的重复模式。

[0037] DOM-树400包括来自网页300(图3)的每个节点，如由每个节点的圈出数字指示的。DOM-树的叶节点(如在参考数字402、404和406处的叶节点)可以用于生成信号，如在框204(图2)处描述的。在通过在框206(图2)处描述的信号分析找到了周期波之后，该周期波可以用于选择子树节点，如在框208(图2)处所述的。可以找到包括对应于每个特定波的所有叶节点的、DOM-树的最小子树。例如，周期波中的 “波”可以对应于叶节点402、404和406。子树

节点410是包括叶节点402、404和406的最小子树节点。相似地，包括对应于周期波的波的其他叶节点的、最小子树节点可以是来自DOM-树的节点，诸如节点410、412、414和416。

[0038] 图5是示出根据本技术一个实施例的、用于网页的1D信号500的图示。该1D信号500通过使用在DOM-树中的叶节点以及节点列表从DOM-树400(图4)生成。x-坐标，如由x-轴502指示的，对应于在DOM-树中的节点的遍历次序。y-坐标，如由y-轴504指示的，对应于在DOM-树中的相同节点的节点深度。

[0039] 图6是示出根据本技术一个实施例的、对信号600进行信号分析的结果的图示。信号600示出根据在框206(图2)处描述的技术对信号500(图5)进行信号分析的结果。该信号600具有x-轴602和y-轴604。通过信号分析，可以在参考数字606和608处找到两个周期波。

[0040] 在参考数字606处的周期波内，可以在参考数字610处找到两个重复模式。在参考数字610处的这两个重复模式可以被从时间频率域变换回到1D信号。在周期波中找到的节点可以用于在DOM-树中找到在参考数字408和412(图4)处的节点，其对应于在参考数字302和304(图3)处的重复模式。相似地，可以在参考数字608处的周期波中找到三个重复模式612。该三个重复模式612可以用于找到在节点410、414和416(图4)处的子树，其对应于在参考数字316、318和320(图3)处的重复模式。

[0041] 图7是示出根据本技术一个实施例的、网页700的所检测到的重复模式的标记选择结果的图示。本技术能够发现在网页700上的数个重复模式。例如，在参考数字702处的部分包含6个超链接，其对应于关于我们、求职、投资者、查看全部、运输以及返回链接，它们被分组为三个部分。在参考数字702处的模式重复三次，如参考数字704、706和708所指示的。参考数字704、706和708示出网页的重复模式。同样地，在参考数字710处的虚线内，存在对应于各种广告的五个重复模式712、714、716、718和720。

[0042] 图8是示出根据本技术一个实施例的、存储用于检测在网页上的重复模式的代码的非暂时的、计算机可读介质的框图。所述非暂时的、计算机可读介质通常用参考数字800表示。

[0043] 所述非暂时的、计算机可读介质800可以对应于存储计算机实现指令的任何典型的存储装置，所述计算机实现指令诸如编程代码等。例如，所述非暂时的、计算机可读介质800可以包括非易失性存储器、易失性存储器和/或一个或多个存储装置中的一种或多种。

[0044] 非易失性存储器的示例包括但不限于电可擦除可编程只读存储器(EEPROM)以及只读存储器(ROM)。易失性存储器的示例包括但不局限于静态随机存取存储器(SRAM)以及动态随机存取存储器(DRAM)。存储装置的示例包括但不限于硬盘、致密盘驱动器、数字多用途盘驱动器以及闪速存储器装置。

[0045] 处理器802通常取回在非暂时的计算机可读介质800中存储的、用于检测在网页上的重复模式的计算机实现指令，并执行所述计算机实现指令。在框804处，再现模块可以生成DOM-树，以及根据所述DOM-树和节点列表生成信号。所述再现模块可以分析该信号，以及选择在所述信号中的节点来形成周期波。在框806处，检测模块可以使用来自再现模块的周期波以及节点来检测重复模式。

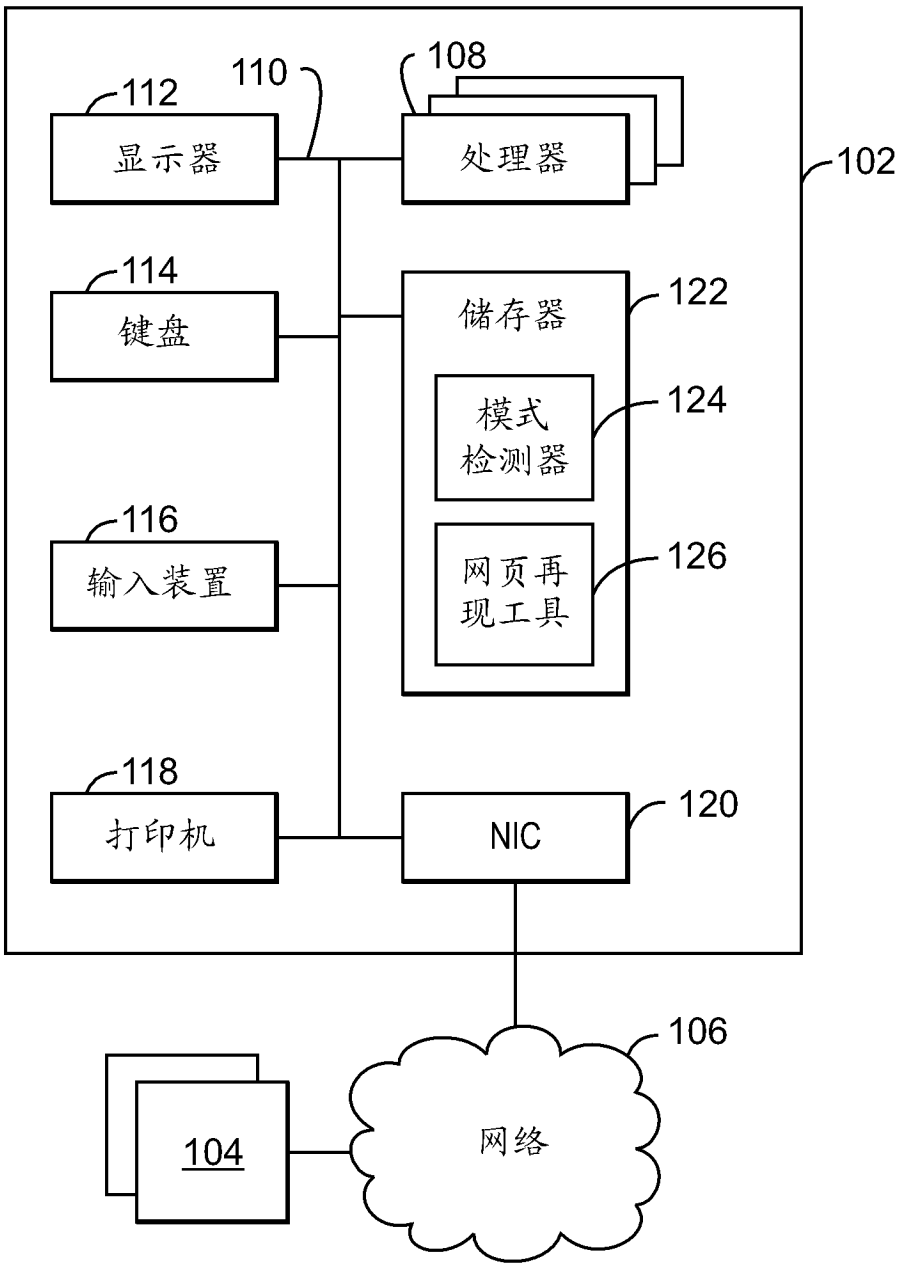

112
110
108
显示器
处理器
102
114
键盘
储存器
122
模式
检测器
124
116
输入装置
网页再
现工具
126
118
打印机
NIC
120
106
104
网络
100

图 1

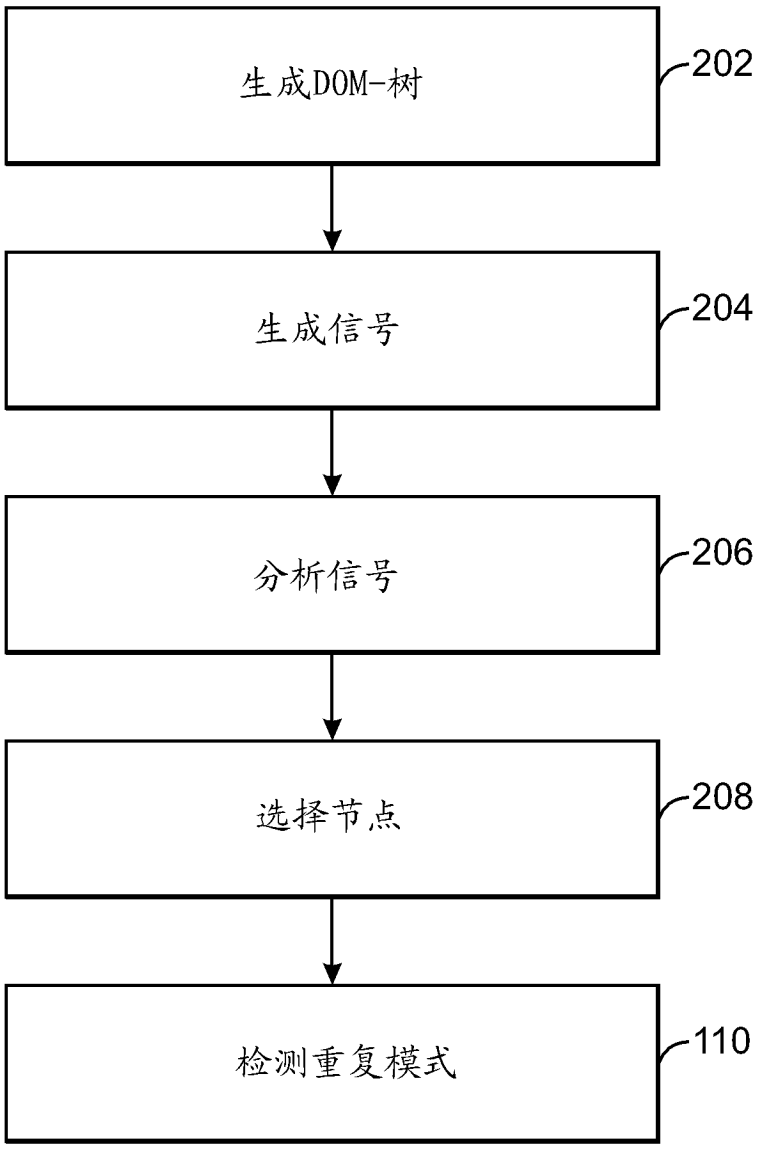
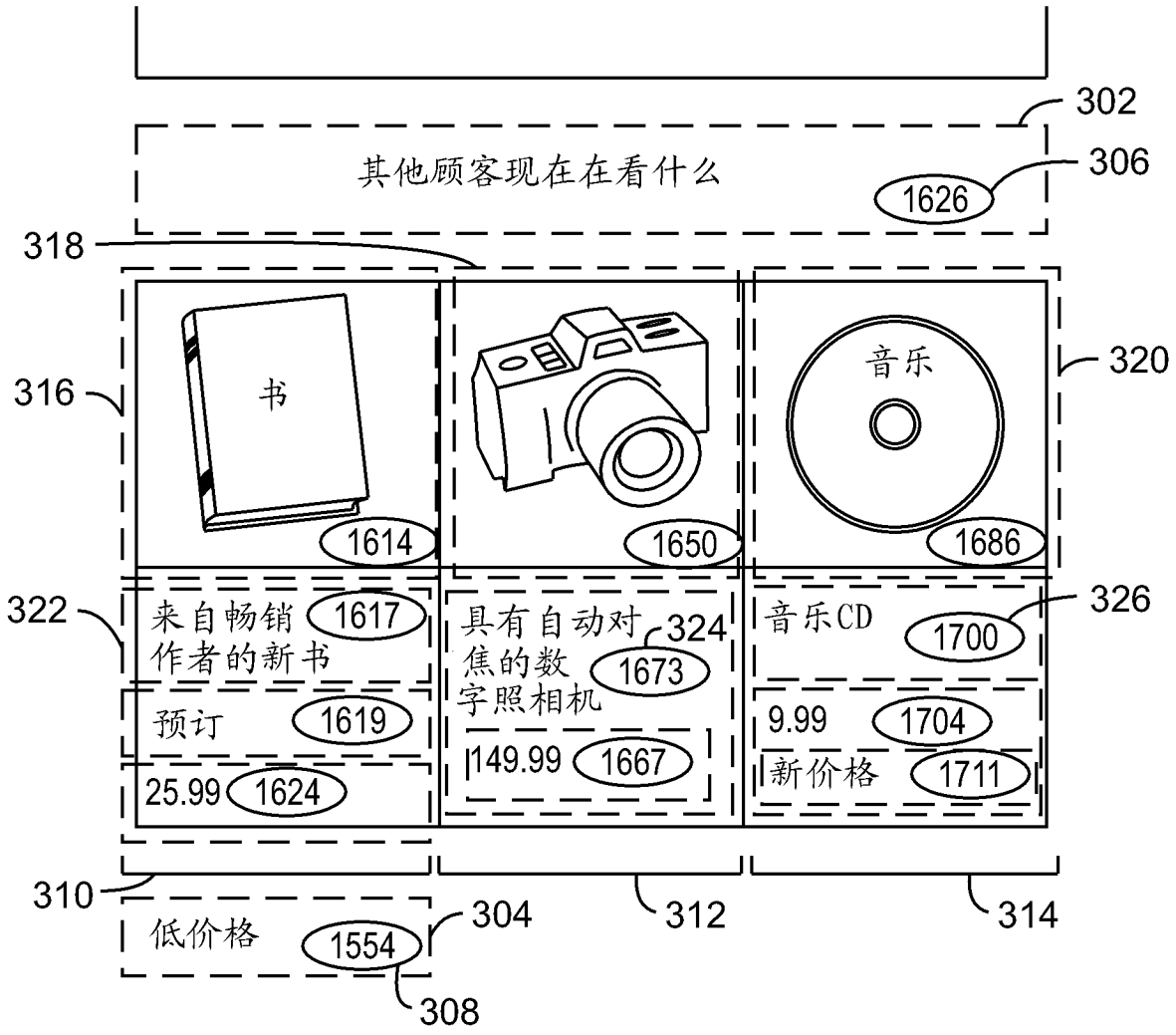
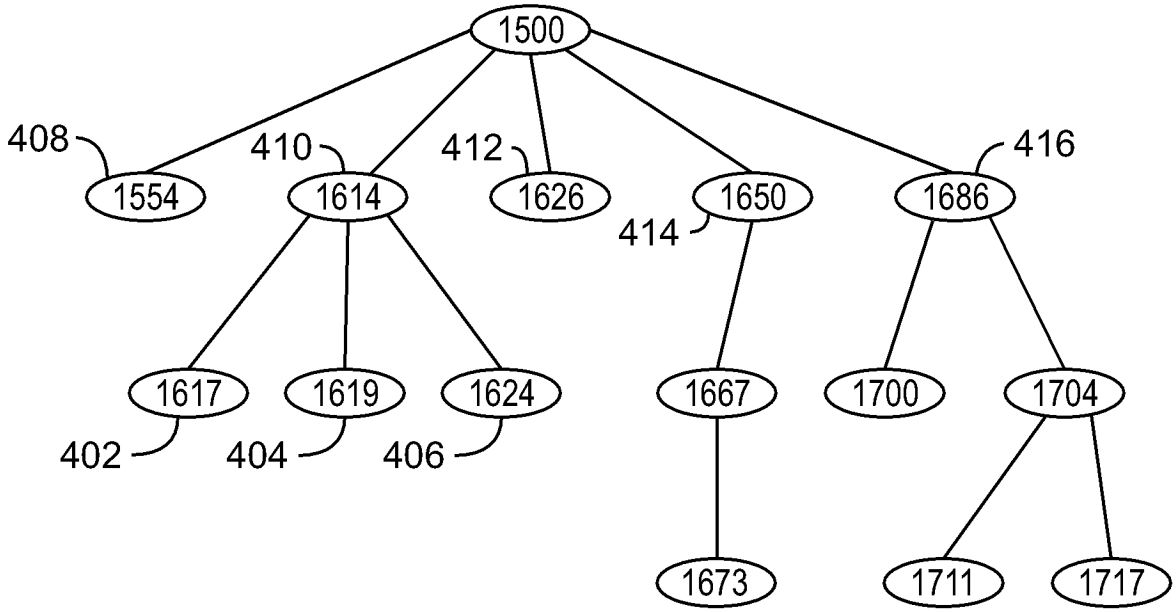


图 2


302
其他顾客现在在看什么
306
1626
318
书
316
1614
1650
音乐
320
1686
322
来自畅销作者的新书
1617
预订
1619
25.99
1624
具有自动对焦的数字照相机
324
1673
149.99
1667
音乐CD
1700
326
9.99
1704
新价格
1711
310
低价格
1554
304
308
312
314

300

图 3


1500
408
410
412
416
1554
1614
1626
1650
1686
414
1617
1619
1624
1667
1700
1704
402
404
406
1673
1711
1717

400

图 4

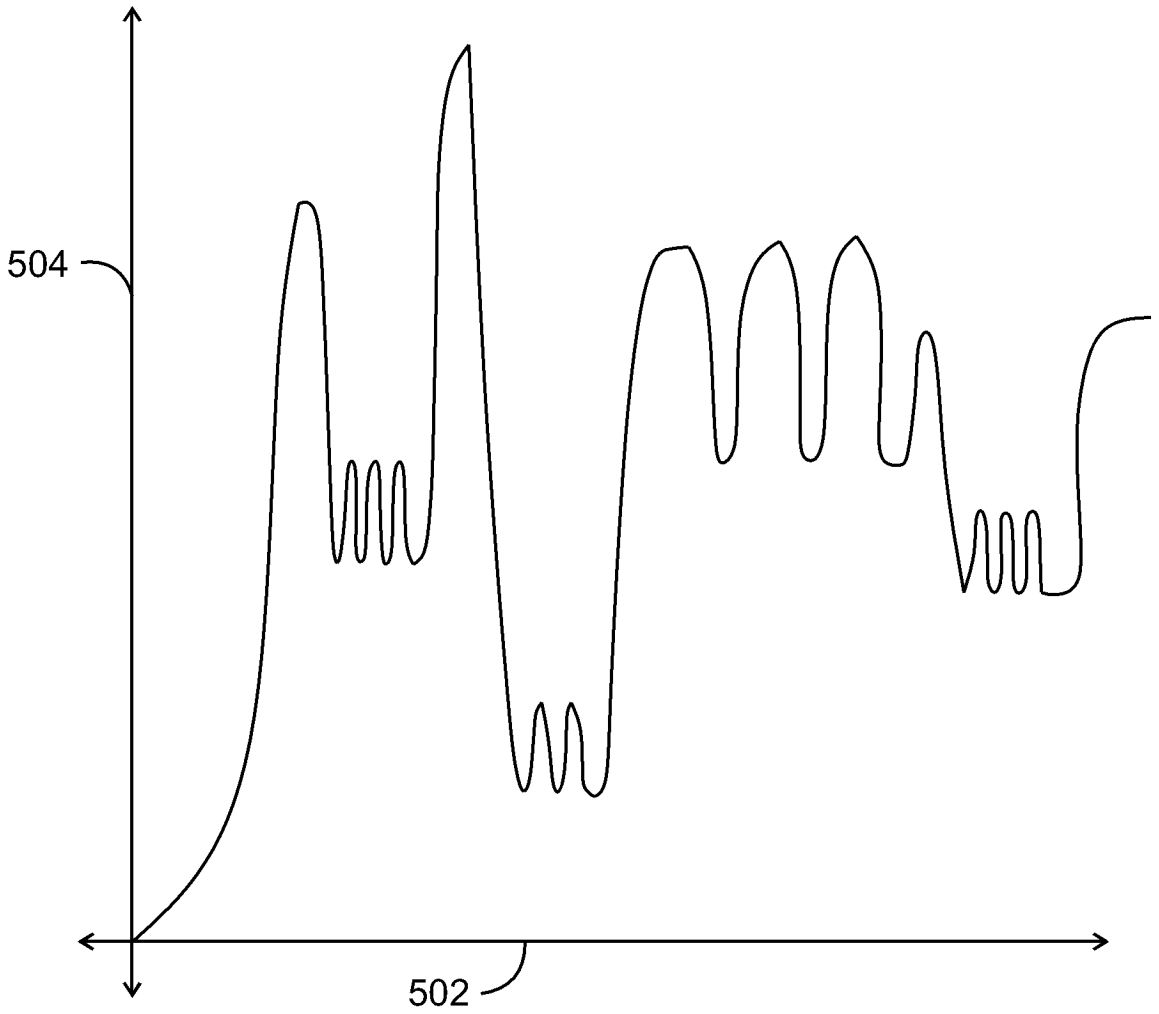


图 5

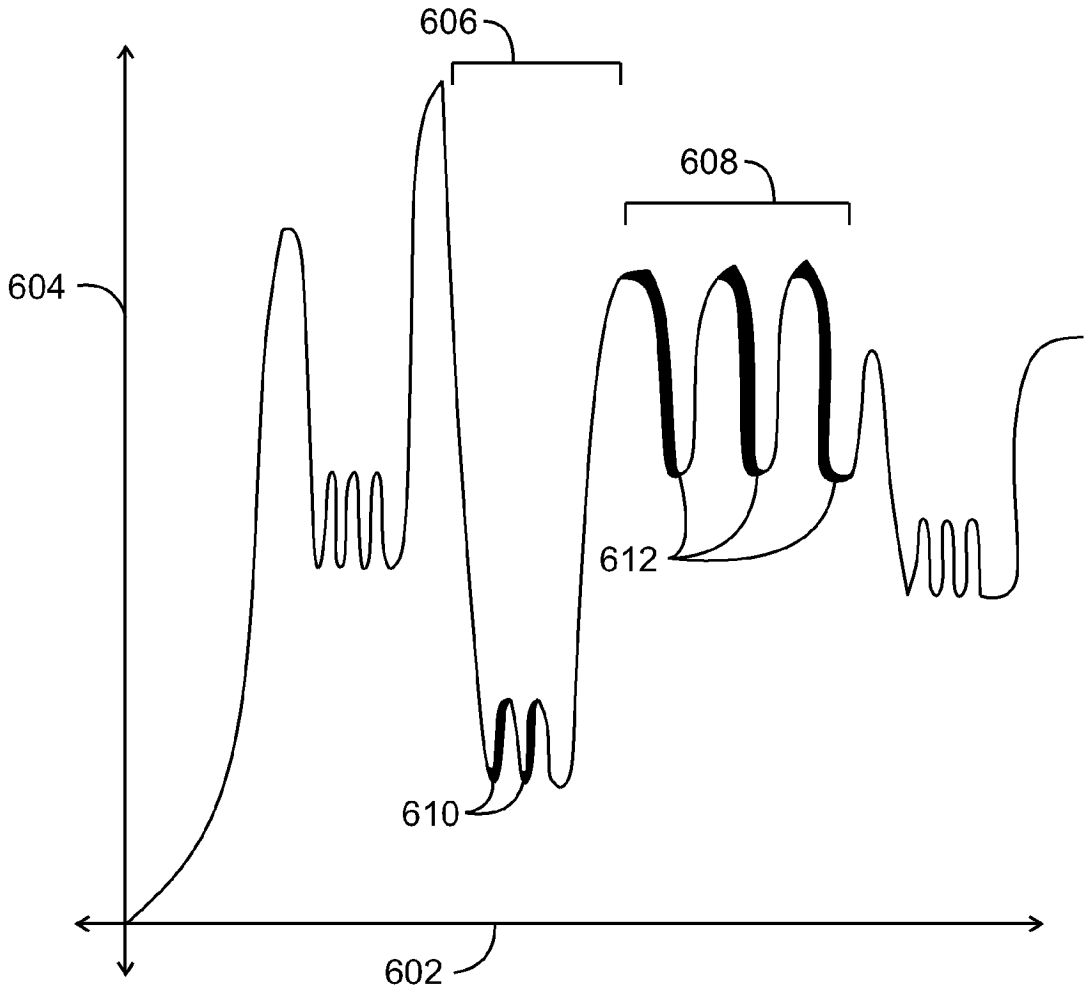


图 6

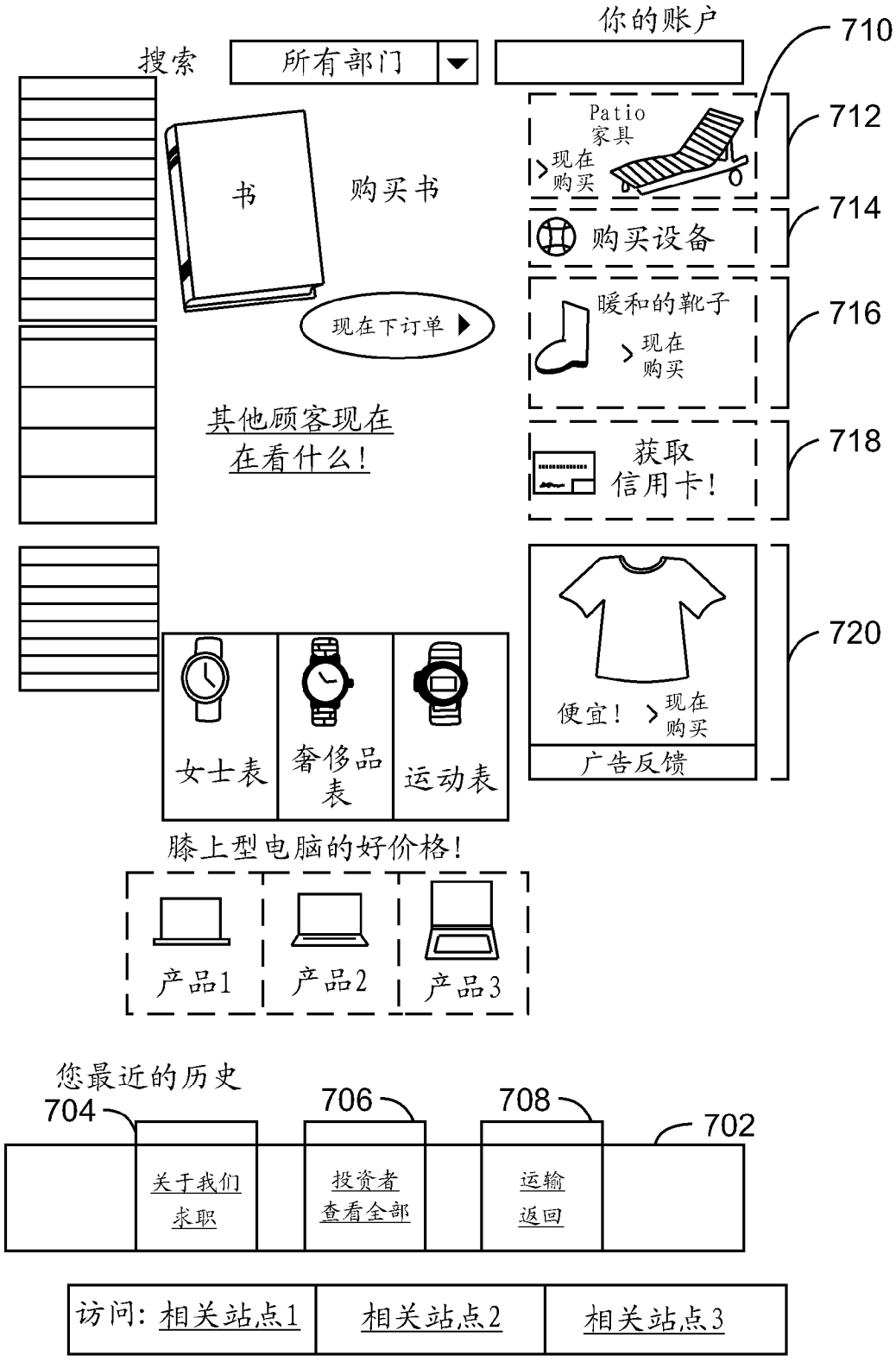

你的账户
搜索
所有部门
710
712
Patio
家具
现在
购买
书
购买书
714
购买设备
现在下订单
暖和的靴子
716
现在
购买
其他顾客现在
在看什么!
718
获取
信用卡!
720
便宜!
现在
购买
广告反馈
女士表
奢侈品
表
运动表
膝上型电脑的好价格!
产品1
产品2
产品3
您最近的历史
704
706
708
702
关于我们
求职
投资者
查看全部
运输
返回
访问: 相关站点1
相关站点2
相关站点3
700

图 7

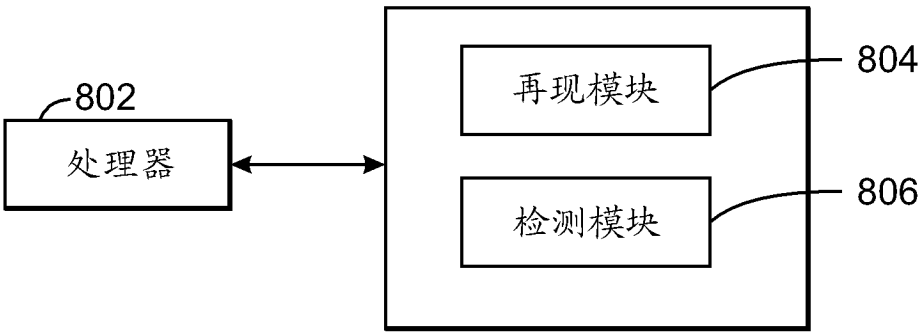


图 8