(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110462604 B
(45)授权公告日 2023.11.03

(21)申请号 201880020315.9
(22)申请日 2018.01.23
(65)同一申请的已公布的文献号
申请公布号 CN 110462604 A
(43)申请公布日 2019.11.15
(30)优先权数据
15/412,245 2017.01.23 US
(85)PCT国际申请进入国家阶段日
2019.09.23
(86)PCT国际申请的申请数据
PCT/US2018/014807 2018.01.23
(87)PCT国际申请的公布数据
WO2018/136921 EN 2018.07.26
(73)专利权人 萃弈公司
地址 美国加利福尼亚州
(72)发明人 J·阿特拉斯 F·卡洛 J·马
(74)专利代理机构 永新专利商标代理有限公司 72002
专利代理师 胡欣
(51)Int.Cl.
G06F 16/2455(2019.01)
G06F 16/28(2019.01)
G06F 16/901(2019.01)
H04L 67/1396(2022.01)
(56)对比文件
US 9514248 B1,2016.12.06
US 2016/0182657 A1,2016.06.23
US 2014/0095320 A1,2014.04.03
CN 1841380 A,2006.10.04
CN 105144200 A,2015.12.09
审查员 高慧美

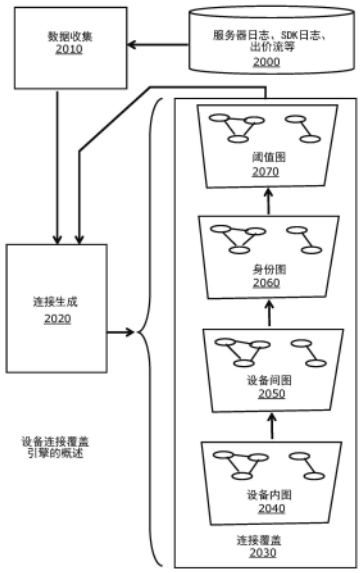
权利要求书3页 说明书26页 附图26页

(54)发明名称

基于设备使用关联互联网设备的数据处理系统和方法

(57)摘要

数据处理系统执行原始或预处理数据的数据处理。数据包括日志文件、比特流数据和包含cookie或设备标识符的其他网络流量。数据处理系统将设备与设备活动历史相关联。数据处理系统包括配对框架，用于基于设备活动历史确定设备对，特征向量生成框架，用于至少部分地基于原始或预处理数据内与确定的候选设备对的设备相关联的特征值，产生与确定的候选设备对相对应的特征向量，评分引擎，用于基于产生的特征向量确定与所确定的设备对相关联的分数，图结构，包括用于表示所确定的设备对的设备的节点(其中，节点对之间的边表示所确定的设备对)，以及聚类引擎，用于识别图结构内表示相应设备组的相应n个或多个节点的聚类。



CN 110462604 B

1.一种用于处理数据的系统,包括:(i)非瞬时性计算机可读储存设备,用于存储将设备标识符与设备活动历史相关联的原始或预处理数据;(ii)配对框架,能够由一个或多个处理器执行,并且被配置用于至少部分地基于所述设备活动历史的至少一部分来确定候选设备对;(iii)连接覆盖引擎包括:数据收集器,被配置成:从数据源接收原始数据或预处理数据;在提取-转换-装载(ETL)模块处处理从所述数据源接收的所述原始数据或预处理数据;在Botnoise过滤器处对从所述ETL模块接收的经处理的数据进行过滤;在归一化和采样模块处对从所述Botnoise过滤器接收的经过滤的数据进行归一化和采样,以产生输出数据;将所述输出数据传输至数据存储器;连接生成器,被配置成生成连接覆盖,所述连接覆盖依次包括设备内图和设备间图;事件访问控制系统,被配置成:从所述数据存储器接收数据;以及生成事件集,其中,所述事件集基于从所述数据存储器接收到的数据和至少一个规则;以及特征向量生成框架,能够由一个或多个处理器执行,并且被配置用于至少部分地基于所述原始或预处理数据内与确定的候选设备对的设备相关联的设备活动历史,产生与确定的候选设备对相对应的多特征值特征向量,其中:所述特征向量生成框架适于至少部分地基于所述原始或预处理数据内与至少一个确定的候选设备对的第一设备标识符相关联的设备活动历史,及所述原始或预处理数据内与至少一个确定的候选设备对的第二设备标识符相关联的设备活动历史,采用至少一个规则来产生在对应于至少一个确定的候选设备对的特征向量内表示的至少一个特征的特征值,其中,对应于所述至少一个确定的候选设备对的所述特征向量包括第一置信度结果和第二置信度结果,所述第一置信度结果被用于产生所述设备内图,并且所述第二置信度结果被用于产生所述设备间图;(iv)评分引擎,能够由一个或多个处理器执行,并且适于至少部分地基于所产生的与确定的候选设备对相关联的特征向量来确定与包括至少一个确定的候选设备对的确定的候选设备对相关联的分数;(v)非瞬时性计算机可读储存设备适于存储包含所述设备内图和所述设备间图中的一个或多个的图结构,其中,图结构内的节点表示所述设备标识符,包括至少一个确定的候选设备对的第一设备标识符和第二设备标识符,并且其中,图结构内的节点对之间的边表示确定的候选设备对;及(vi)聚类引擎,能够由一个或多个处理器执行,并且适于识别所述图结构内表示相应设备组的相应的两个或更多个节点聚类。

2.根据权利要求1所述的系统,其中,所述系统还包括能够由一个或多个处理器执行的目标引擎,用于使用所述设备标识符的组使信息通过互联网以与所述设备标识符的组相关联的用户为目标。

3.根据权利要求1或2所述的系统,其中,所述原始或预处理数据用于将所述设备标识符与一个或多个网络源/目的地地址标识符相关联;并且其中,所述配对框架适于至少部分地基于所述设备标识符和相关联的网络源/目的地地址标识符来识别设备对。

4.根据权利要求1所述的系统,其中,所述原始或预处理数据用于将所述设备标识符与一个或多个网络源/目的地地址标识符和一个或多个时间戳相关联;并且其中,所述配对框架被配置为至少部分地基于所述设备标识符和关联的一个或多个网络源/目的地地址标识符和一个或多个时间戳来识别设备对。

5.根据权利要求1所述的系统,其中,所述特征向量生成框架适于至少部分地基于所述原始或预处理数据内与各个确定的设备对的设备的所述设备标识符相关联的设备活动历史来产生包括多个特征值的特征向量,以与各个确定的设备对相关联。

6.根据权利要求1所述的系统，其中，所述特征向量生成框架适于生成包括多个特征值的特征向量，以与各个确定的设备对相关联，并且其中，所述评分引擎被配置为评估与不同的各个确定的设备对相关联的特征向量之间的相关性。

7.根据权利要求6所述的系统，其中，所述评分引擎适于使用评分模型来为所确定的设备对的至少一部分产生分数。

8.根据权利要求1所述的系统，其中，所述特征向量生成框架适于产生包括多个特征值的特征向量，以与各个确定的设备对相关联，其中，所述评分引擎适于至少部分地基于与确定的设备对相关联的特征向量之间的亲和度的评估来产生确定的设备对的多个重映射聚类；其中，所述评分引擎适于评估与相同重映射聚类内的不同的各个确定的设备对相关联的特征向量之间的相关性；其中，所述评分引擎适于基于对与不同的各个确定的设备对相关联的特征向量之间的相关性的评估来采用所述非瞬时性计算机可读储存设备中的评分模型；并且其中，所述评分引擎适于使用评分模型来为确定的设备对的至少一部分产生分数。

9.根据权利要求1所述的系统，其中，所述特征向量生成框架适于产生包括多个特征值的特征向量，以与各个确定的设备对相关联，其中，所述评分引擎适于评估与不同的各个确定的设备对相关联的特征向量之间的相关性；其中，所述评分引擎适于基于对产生的聚类内的有标签设备对的向量与无标签设备对的向量之间的相关性的评估来产生存储在所述非瞬时性计算机可读储存设备中的评分模型；并且其中，所述评分引擎适于使用所述评分模型来为确定的设备对的至少一部分产生分数。

10.根据权利要求1所述的系统，其中，所述评分引擎适于产生确定的设备对的多个重映射聚类，其中评估有标签设备对的向量与无标签设备对的向量之间的相关性，其中，所述评分引擎适于基于对产生的重映射聚类内的有标签设备对的向量与无标签设备对的向量之间的相关性的评估来产生评分模型；并且其中，所述评分引擎适于使用所述评分模型来为确定的设备对的至少一部分产生分数。

11.根据权利要求1所述的系统，其中，所述原始或预处理数据用于将各个设备标识符与标签信息相关联，所述标签信息指示与所述各个设备标识符相关联的各个设备是有标签的还是无标签的；其中，所述评分引擎适于评估有标签设备对的向量与无标签设备对的向量之间的相关性；其中，所述评分引擎适于基于对所产生的聚类内的有标签设备对的向量与无标签设备对的向量之间的相关性的评估来产生存储在所述非瞬时性计算机可读储存设备中的评分模型；并且其中，所述评分引擎适于使用所述评分模型来为所确定的设备对的至少一部分产生分数。

12.根据权利要求1所述的系统，其中，所述聚类引擎适于使用模拟退火过程来识别所述图结构内的节点的聚类。

13.根据权利要求1所述的系统，其中，所述特征向量生成框架适于生成包括多个特征值的特征向量，以与各个确定的设备对相关联；其中，所述评分引擎适于评估与不同的各个确定的设备对相关联的特征向量之间的相关性；其中，所述评分引擎适于基于对所产生的聚类内的有标签设备对的向量与无标签设备对的向量之间的相关性的评估来产生存储在所述非瞬时性计算机可读储存设备中的评分模型；并且其中，所述评分引擎适于使用所述评分模型来为所确定的设备对的至少一部分产生分数；并且其中，所述聚类引擎适于至少

部分地基于所产生的分数来识别所述图结构内的节点的聚类。

14.一种用于处理数据的方法,其中,所述方法包括:(i)使用非瞬时性计算机可读储存设备来存储将设备标识符与设备活动历史相关联的原始或预处理数据;(ii)使用配对框架,其被配置为至少部分地基于设备活动历史的至少一部分来确定候选设备对;(iii)使用连接覆盖引擎,其包括:数据收集器,被配置成:从数据源接收原始数据或预处理数据;在提取-转换-装载(ETL)模块处处理从所述数据源接收的所述原始数据或预处理数据;在Botnoise过滤器处对从所述ETL模块接收的经处理的数据进行过滤;在归一化和采样模块处对从所述Botnoise过滤器接收的经过滤的数据进行归一化和采样,以产生输出数据;将所述输出数据传输至数据存储器;连接生成器,被配置成生成连接覆盖,所述连接覆盖依次包括设备内图和设备间图;事件访问控制系统,被配置成:从所述数据存储器接收数据;以及生成事件集,其中,所述事件集基于从所述数据存储器接收到的数据和至少一个规则;以及特征向量生成框架以至少部分地基于所述原始或预处理数据内与确定的候选设备对的设备相关联的设备活动历史,来产生与确定的候选设备对相对应的多特征值特征向量,其中,所述方法还包括:(iv)布置所述特征向量生成框架以至少部分地基于所述原始或预处理数据内与至少一个确定的候选设备对的第一设备标识符相关联的设备活动历史,及所述原始或预处理数据内与至少一个确定的候选设备对的第二设备标识符相关联的设备活动历史,采用至少一个规则来产生在对应于至少一个确定的候选设备对的特征向量内表示的至少一个特征的特征值,其中,对应于所述至少一个确定的候选设备对的所述特征向量包括第一置信度结果或第二置信度结果,所述第一置信度结果被用于产生所述设备内图,并且所述第二置信度结果被用于产生所述设备间图,并且,所述第一置信度结果指示设备内候选对,所述第二置信度结果指示设备间候选对;(v)使用评分引擎以至少部分地基于所产生的与确定的候选设备对相关联的特征向量来确定与包括至少一个确定的候选设备对的确定的候选设备对相关联的分数;(vi)使用非瞬时性计算机可读储存设备来存储包括所述设备内图和所述设备间图中的一个或多个的图结构,其中,所述图结构内的节点表示所述设备标识符,包括所述至少一个确定的候选设备对的第一设备标识符和第二设备标识符,并且其中,所述图结构内的节点对之间的边表示确定的候选设备对;及(vii)使用聚类引擎来识别所述图结构内表示相应设备组的相应的两个或更多个节点聚类。

15.如权利要求14所述的方法,其中,所述图结构包括身份图和家庭图中的一个或多个。

基于设备使用关联互联网设备的数据处理系统和方法

技术领域

[0001] 本公开内容涉及用于执行用于基于设备使用来关联互联网设备的数据处理的数据处理系统。此外，本公开内容涉及使用基于设备使用来关联互联网设备的上述数据处理系统的方法。此外，本公开内容涉及记录在机器可读数据储存介质上的软件产品，其特征在于，软件产品可在计算硬件上执行以执行上述方法。

背景技术

[0002] 传统上，互联网提供数据通信网络，通过该网络，人们能够使用各种不同类型的设备来交换信息。例如，给定用户拥有智能电话、移动平板电脑、笔记本电脑和连接的电视。随着用户在多个不同的可连接互联网的设备上工作、社交、研究和购买产品，商业公司正在继续将注意力转移到通过其各自的多个设备更有效地到达用户。例如，虽然给定人员拥有并使用不同的设备通过互联网进行通信，但不同设备和不同设备的用户之间的关系对于外界来说并不明显，例如寻求通过给定人员的多个设备到达给定人员的商业公司。

[0003] 这种上述关系可能涉及给定人员使用具有不同设备标识符的不同设备来通过互联网进行通信。例如，给定人员通过互联网匿名通信而不公开给定人员的个人标识符。用户设备与互联网的连接通常是暂时的和动态的。设备通常通过与网际协议(IP)地址关联的连接点连接到互联网。但是，用户设备可能在不同时间使用不同的网络地址。在通过互联网的通信期间，可能交换用户设备识别信息，例如设备标识符或用户标识符。然而，使用一个设备的给定用户在互联网通信期间使用的识别信息可能与同一用户使用不同设备在互联网通信期间使用的识别信息不同。另外，相同的给定设备可以在不同的互联网通信期间使用不同的识别信息。因此，当通过互联网与不同设备通信时，人可能在不同时间使用不同的IP地址。例如，用户具有多个不同的电子邮件帐户并且可能以不同假名参与使用社交媒体。因此，目前还没有现有的识别使用不同设备访问互联网的用户的现成可用的可靠确定性方法。

发明内容

[0004] 所公开的实施例的各方面试图提供用于基于描述设备的数据网络活动的原始或未处理的数据，例如互联网浏览活动，确定设备之间的不同类型的关联的改进的系统(例如，设备内关联，包括所有权的设备间关联，使用份额中的关联，家庭关联等)；当跨越平台、设备、登录等的边界时，这种改进的系统基于给定个人的身份是恒定的，并且基于收集的与给定个人有关的使用信息，可以推断给定个人的身份。

[0005] 根据第一方面，提供了一种系统，包括：

[0006] (i)计算机可读储存设备，用于存储将设备标识符与设备活动历史相关联的原始或预处理数据；

[0007] (ii)配对框架，用于至少部分地基于设备活动历史的至少一部分来确定不同类型的候选设备对；

[0008] (iii)特征向量生成框架,用于根据一种或多种类型的对并且至少部分地基于原始或预处理数据内与确定的候选设备对的设备相关联的设备活动历史,产生与确定的候选设备对相对应的多特征值特征向量,

[0009] 其特征在于:

[0010] (iv)特征向量生成框架可操作以至少部分地基于原始或预处理数据内与至少一个确定的候选设备对的第一设备标识符相关联的设备活动历史,及原始或预处理数据内与至少一个确定的候选设备对的第二设备标识符相关联的设备活动历史,采用至少一个规则来产生在对应于至少一个确定的候选设备对的特征向量内表示的至少一个特征的特征值;

[0011] (v)评分引擎可操作以至少部分地基于所产生的与确定的候选设备对相关联的特征向量来确定与包括至少一个确定的候选设备对的确定的候选设备对相关联的分数;

[0012] (vi)计算机可读储存设备可操作以存储图结构,其中,图结构内的节点表示设备标识符,包括至少一个确定的候选设备对的第一设备标识符和第二设备标识符,并且其中,图结构内的节点对之间的边表示确定的候选设备对;及

[0013] (vii)聚类引擎可操作以识别图结构内表示相应设备组的相应两个或更多个节点的聚类。

[0014] 所公开实施例的各方面提供的优点在于,系统能够依据原始或未处理数据提供更有效地识别用户设备的关联(即,避免生成任何形式的索引结构的需要,从而节省计算工作量和/或更快地生成关联结果)。

[0015] 可以理解,就“活动历史”而言,例如,来自设备的每个互联网通信可以被捕获为具有传输信息的事件,并且这些事件的序列可以被视为设备在互联网上的活动的历史。

[0016] 可以理解,例如,就“框架”而言,对于不同类型的对可能存在不同的引擎,例如,稍后在本公开内容中描述至少两个这样的引擎(例如,用于设备内对和设备间对)。

[0017] 可选地,在系统中,原始或未处理的数据包括例如日志文件、比特流数据和包含cookie或设备标识符的其他网络流量。

[0018] 根据所公开实施例的第二方面,提供了一种使用系统来确定关联的方法,其中,该方法包括:

[0019] (i)使用计算机可读储存设备来存储将设备标识符与设备活动历史相关联的原始或预处理数据;

[0020] (ii)使用配对框架,其被配置为至少部分地基于设备活动历史的至少一部分来确定候选设备对;

[0021] (iii)使用特征向量生成框架以至少部分地基于原始或预处理数据内与确定的候选设备对的设备相关联的活动历史,来产生与确定的候选设备对相对应的多特征值特征向量,其特征在于:方法还包括:

[0022] (iv)布置特征向量生成框架至少部分地基于原始或预处理数据内与至少一个确定的候选设备对的第一设备标识符相关联的设备活动历史,及原始或预处理数据内与至少一个确定的候选设备对的第二设备标识符相关联的设备活动历史,采用至少一个规则来产生在对应于至少一个确定的候选设备对的特征向量内表示的至少一个特征的特征值;

[0023] (v)使用评分引擎以至少部分地基于所产生的与确定的候选设备对相关联的特征向量来确定与包括至少一个确定的候选设备对的确定的候选设备对相关联的分数;

[0024] (vi)使用计算机可读储存设备来存储图结构,其中,图结构内的节点表示设备标识符,包括至少一个确定的候选设备对的第一设备标识符和第二设备标识符,并且其中,图结构内的节点对之间的边表示确定的候选设备对;及

[0025] (vii)使用聚类引擎来识别图结构内表示相应设备组的相应两个或更多个节点的聚类。

[0026] 根据第三方面,提供了一种计算机程序产品,包括其上存储有计算机可读指令的非暂时性计算机可读储存介质,所述计算机可读指令可由包括执行第二方面的方法的处理硬件的计算机化设备执行。

[0027] 应当理解,在不脱离由所附权利要求限定的本发明的范围的情况下,本发明的特征易于以各种组合进行组合。

附图说明

[0028] 现在将参考以下附图仅通过示例的方式描述本公开内容的实施例,其中:

[0029] 图1A和1B是与本公开内容的不同物理实施例相关联的不同网络地址处的网络连接设备使用的图示;

[0030] 图2是根据本公开内容实施例的数据处理系统的示意图,其中数据处理系统包括数字身份配对引擎;

[0031] 图3是表示根据本公开内容实施例的图2的数据处理系统的数字身份配对引擎的配置的说明性过程流程图;

[0032] 图4A是根据本公开内容实施例的使用图2的配对引擎识别的多个示例设备对的图示;

[0033] 图4B是根据本公开内容实施例的与示例特征向量相关联的图4A的示例设备对的图示;

[0034] 图5是表示根据本公开内容实施例的图2的系统的特征向量生成引擎的配置的说明性过程流程图;

[0035] 图6A是表示根据本公开内容实施例的根据图5的过程应用的第一示例规则的示例细节的过程的说明性流程图,以产生作为与候选设备对相关联的特征信息的函数的示例特征值;

[0036] 图6B是表示根据本公开内容实施例的根据图5的过程应用的第二示例规则的示例细节的过程的说明性流程图,以产生作为与候选设备对相关联的特征信息的函数的示例特征值;

[0037] 图7是根据本公开内容实施例的提供用于产生特征值的输入储存系统内的低级特征信息和图2的系统的规则储存内的规则的示例的说明性图表;

[0038] 图8A是表示根据本公开内容实施例的图2的数据处理系统的评分引擎的配置的说明性流程图;

[0039] 图8B到8E是表示根据本公开内容实施例的图2的数据处理系统的评分引擎的替代配置的说明图;

[0040] 图9是表示根据本公开内容实施例的图8A的评分引擎(可选地还用于图8B至8E的评分引擎)的训练模块的细节的说明性流程图;

[0041] 图10A是表示根据本发明实施例的无监督预训练模块可操作以基于样本设备对特征向量将图7的设备对重映射到的设备对聚类的示例的说明图；

[0042] 图10B是表示根据本发明实施例的图9的训练模块的重映射模块的使用的说明图；

[0043] 图11是表示根据本发明实施例的包括在图2的数据处理系统的评分引擎内的计算机可读储存设备内发生的数据变换的数据流过程的说明图；

[0044] 图12是表示根据本发明实施例的图2的系统的聚类引擎的配置的说明性流程图；

[0045] 图13是表示根据本发明实施例的将图12的聚类引擎用于图4A和4B的设备对产生的示例图的说明图；

[0046] 图14是表示根据本发明实施例的在图13的图中使用图12的聚类引擎识别的示例提出的用户设备聚类用户设备聚类集合的说明图；

[0047] 图15是表示根据本发明实施例的基于图14的提出的用户设备聚类使用图12的聚类引擎产生的最终设备聚类的示例集合的说明图；

[0048] 图16是表示根据本发明实施例的包括在图2的数据处理系统的聚类引擎内的用户设备聚类期间发生的数据变换的数据流程图的说明图；

[0049] 图17是表示根据本公开内容实施例的被配置为在网络环境中操作的图2的数据处理系统的说明图。

[0050] 在附图中，下划线数字用于表示下划线数字所在的项目或下划线数字相邻的项目。未加下划线的数字涉及由将未加下划线的数字链接到项目的线标识的项目。当数字未加下划线并伴有相关箭头时，未加下划线的数字用于标识箭头指向的一般项目。

具体实施方式

[0051] 总的来说，下面提供对本公开内容的实施例的描述，其使得本领域的任何技术人员能够创建和使用数据处理系统来基于匿名互联网用户数据(即与互联网可连接设备相关联的“活动历史”)来关联属于同一给定用户的互联网可连接设备。对于本领域技术人员来说，对实施例的各种修改是显而易见的。相同的附图标记可能用于表示不同附图中相同项目的不同视图。下面引用的附图中的流程图用于表示过程。诸如包括一个或多个处理器和储存设备的计算机系统的机器被配置为执行这些过程。流程图表示使用计算机程序代码配置以执行参考流程图描述的操作的计算机系统的一个或多个处理器和/或储存设备的配置。

[0052] 本公开内容的实施例提供了一种数据处理系统，其可操作地使用原始或预处理数据来将设备与设备属性相关联，例如设备活动历史；可选地，原始或未处理的数据包括例如日志文件、比特流数据和包含cookie或设备标识符的其他网络流量。数据处理系统在操作中采用配对引擎，用于至少部分地基于前述设备属性的至少一部分来确定设备对。此外，数据处理系统采用特征向量生成引擎，用于至少部分地基于原始或预处理数据内与确定的设备对的设备相关联的特征值来产生对应于确定的设备对的特征向量。此外，数据处理系统包括评分引擎，用于至少部分地基于所产生的与确定的设备对相关联的特征向量来确定用于与确定的设备对相关联的分数。另外，数据处理系统在操作中采用图结构，该图结构包括表示确定的设备对的设备的节点，并且包括指示确定的设备对的节点对之间的边。在数据处理系统中，还采用聚类引擎来识别图结构内表示相应设备组的相应节点聚类。

[0053] 接下来,将参考图1A概述数字身份配对。在由150表示的给定的示例家庭A,有使用多个移动互联网连接数字设备与互联网101或类似的数据通信网络进行通信的活人(即“用户”)。具体地,笔记本计算机系统(151和161)、蜂窝电话(152和162)以及平板计算机系统157都是移动互联网连接数字设备,各自都具有唯一的设备标识符。这些移动设备可以并且经常与他们的一个或多个用户一起被带到提供无线(或有线)互联网接入的其他位置。例如,家庭A的给定用户X可以使用笔记本计算机系统151和蜂窝电话152,无论是在家庭A150还是在由110表示的工作场所W。

[0054] 当给定用户X将笔记本计算机系统151和蜂窝电话152的设备带到工作场所W 110时,那些设备不再能够使用Wi-Fi路由器165来接入互联网101。相反,给定用户X将笔记本计算机系统151连接到工作场所W 110处的局域网129,并配置蜂窝电话152以使用由无线接入点(例如工作场所W 110的)提供的本地Wi-Fi网络。利用这两个互联网连接,给定用户X然后将能够使用笔记本计算机系统151和蜂窝电话152通过工作场所W 110处的防火墙/代理服务器121接入互联网101。

[0055] 当给定用户X在家庭A150时,笔记本计算机系统151和蜂窝电话152都将使用Wi-Fi路由器165上的单个网际协议(IP)地址A163。类似地,当给出相同的给定用户X在工作场所W 110时,笔记本计算机系统151和蜂窝电话152都将使用防火墙/代理服务器121上的单个IP地址W 123。当然,给定用户X可以使用笔记本计算机系统151和蜂窝电话152中的一个或另一个或两者选择性地发送和接收信息。然而,通过笔记本计算机系统151与给定用户X通信的第三方可能不知道通过蜂窝电话152也可以到达给定用户X,通过蜂窝电话152与给定用户X通信的第三方可能不知道通过笔记本计算机系统151也可以到达给定用户X。

[0056] 然而,给定上面参考图1A描述的特定互联网使用模式数据,对于给定用户X一无所知并且不知道给定用户X拥有笔记本计算机系统151和蜂窝电话152两者的敏锐观察者可以合理地推断出笔记本计算机系统151和蜂窝电话152非常有可能由同一个人使用,因为笔记本计算机系统151和蜂窝电话152在家庭,即家庭A150和在工作场所W 110一起使用。在做出这样的推断之后,机器可以将与笔记本计算机系统151相关联的数字标识符和与蜂窝电话152相关联的数字标识符链接一起。这种不同数字身份与单个用户的配对被称为数字身份配对。

[0057] 接下来,将参考图2提供系统概述。图2中提供了表示根据本公开内容的一些实施例的系统200的说明性框图;系统200也称为数据处理系统。系统200包括用于储存设备标识符和与所识别的设备相关联的特征的标记的计算机可读特征储存设备208、配对引擎202、特征向量生成(FVG)引擎212、评分引擎204和聚类引擎206。配对引擎202基于在计算机可读特征储存设备208内指示为与候选设备对的设备相关联的低级别初步配对识别特征来识别候选设备对。应当理解,术语“候选”在本文中用于指示配对确定是初步的并且可以在评分引擎204和聚类引擎206中经历进一步评估。FVG引擎212响应于候选设备对作为与候选设备对的各个设备相关联的附加特征信息(即设备活动历史)的函数产生特征向量。具体而言,配对引擎202向FVG引擎212提供候选设备对标识符。响应于所接收的候选设备对标识符,FVG引擎212使用候选设备对的设备的各个身份来识别在计算机可读特征储存设备208内与该候选设备对的那些各个设备相关联的附加特征信息,即设备活动历史。FVG引擎212作为被识别为与所识别的候选设备对的各个设备相关联的附加特征信息(即设备活动历史)的

函数生成特征向量值以与所识别的候选设备对相关联。评分引擎204至少部分地基于使用FVG引擎212产生的特征向量来确定与所识别的候选设备对相关联的分数。聚类引擎206在计算机可读储存设备内产生表示所识别的候选设备对和确定的分数的至少一部分的图结构。聚类引擎206识别图结构内指示与同一用户相关联的设备聚类或设备组的数字身份聚类U.sub.1、U.sub.2、……、U.sub.N。

[0058] 根据一些实施例，存储在储存设备内的每个单独的数字身份聚类充当数字身份组通信结构U.sub.1、U.sub.2、……、U.sub.N，包括与同一用户相关联并且可用于与拥有分组设备的用户通信的一组设备。因此，例如，数字身份组通信结构U.sub.1包括设备D.sub.A1、D.sub.A2、……、D.sub.AK。具体而言，数字身份分组揭示了被分组的多个设备，这些设备可用于与分组设备的所有者通信。相反，在没有数字身份分组的情况下，希望与特定设备用户通信的第三方不能容易地识别与该特定设备用户通信的不同设备。因此，数字身份组通信结构识别与同一用户相关联的一组设备，以便扩展与该用户的可用通信途径，从而增加与用户通信的机会。应当理解，基于使用配对引擎202确定的数字设备配对，基于使用FVG引擎212产生的特征向量使用评分引擎204进行评分，以及使用聚类引擎206进行聚类，来确定对于希望与与组相关联的用户通信的某些第三方预先未知的设备身份组U.sub.1、U.sub.2、……、U.sub.N。

[0059] 具体而言，在本公开内容的一些实施例中，计算机可读特征储存设备208包括存储将设备标识符信息与低级别特征信息(即低级别设备活动历史)相关联的上述原始或预处理数据209和标签数据的一个或多个储存设备。低级别特征信息可以存储在计算机可读特征储存设备208本身内，或者可以存储在计算机可读特征储存设备208内指示的不同储存位置(未示出)中。例如可以从互联网连接设备请求的服务器日志获得低级别数据。系统200包括计算机可读规则储存设备213，其存储用于基于计算机可读特征储存设备208内指示的低级别特征信息生成高级别特征值的规则。具体而言，计算机可读规则储存设备213存储指令，用于配置FVG引擎212以基于与候选设备对的各个设备相关联的低级别特征，响应于配对引擎202，来识别附加高级别特征信息，即高级别设备活动历史，以与候选设备对相关联，并使用该附加特征信息与低级别特征信息一起产生在特征向量内使用的特征值。系统200包括缓冲器电路210、211，其被配置为分别作为输入接收某些低级别特征信息和标签数据。在一些实施例中，在缓冲器电路210内接收的低级别特征信息包括设备ID、公共源/目的地标识符(例如，IP地址)和时间戳。

[0060] 配对引擎202使用初步配对识别特征信息，例如设备ID、公共源/目的地标识符(例如，IP地址)和时间戳，以确定设备的初步配对。如本文所使用的，术语“设备对”指的是存储在储存设备中并且指示不同数字身份与单个用户的配对的信息结构。配对引擎202用作预过滤器，其产生表示设备的初步“候选”配对的设备对。配对引擎202还将初步对分数与每个设备对相关联，其提供该对中的设备实际上与同一用户相关联的可能性的指示。

[0061] FVG引擎212被配置为使用来自计算机可读规则储存设备213的规则集来作为FVG引擎212内与候选设备对的不同各个设备相关联的低级别特征信息，即低级别设备活动历史的函数，产生特征向量以与候选设备对相关联。FVG引擎212内指示的低级别特征信息包括与各个设备相关联的用户简档信息。计算机可读特征储存设备208内指示的低级别特征信息可以通过互联网收集，并且可以包括诸如诸如年龄、性别或婚姻状况的人口统计数据

和/或诸如用户购买某些物品的意图或诸如喜欢或不喜欢的个人兴趣的行为(英国英语：“行为”)数据的信息。

[0062] 评分引擎204作为输入接收候选设备对标识符(诸如对内的设备的设备ID)，配对引擎202产生的相关初步对分数，FVG引擎212产生的相关特征向量以及相关标签信息。在本公开内容的一些实施例中，与使用配对引擎202产生的候选设备对相关联的初步对分数被包括作为与候选设备对相关联的特征向量的特征值。评分引擎204至少部分地作为其相关联的特征向量的函数产生所接收的设备对的精确对分数。使用评分引擎204产生的精确对分数指示设备对实际上与同一用户相关联的可能性。由评分引擎204产生的对分数优选地取代使用配对引擎202产生的初步对分数。

[0063] 在操作中，聚类引擎206作为输入接收使用评分引擎204产生的设备对和相关联的精确对分数。聚类引擎206基于设备对在计算机可读储存设备内产生图结构(未示出)。图节点表示与设备标识符相关联的设备。图中的边表示候选设备对的设备的潜在配对。精度对分数与图的边相关联。聚类引擎206基于图内与边连接相关联的对分数来识别图内的用户设备聚类。聚类引擎206识别与公共用户相关联的用户设备组。聚类引擎206在计算机可读储存设备214中存储标识与相应用户相关联的相应设备组的信息。

[0064] 使用聚类引擎206识别的用作聚类中设备的相应所有者的数字身份组通信结构的示例用户设备聚类U.sub.1、U.sub.2、……、U.sub.N被示为存储在与相应用户标识符信息相关联的计算机可读储存设备214中。与不同的相应用户相关联的不同设备标识符组。尽管唯一用户的实际身份未知，但每个聚类都将设备ID与唯一用户相关联。例如，由设备ID(D.sub.A1、D.sub.A2、……、D.sub.AK)指示的第一用户设备聚类U.sub.1与唯一用户U.sub.1相关联。设备ID的聚类在本文可以称为“设备聚类”或“用户设备聚类”。

[0065] 在本公开内容的一些实施例中，聚类引擎206向评分引擎204提供反馈F。根据本公开内容的一些实施例，由评分引擎204考虑的一个示例特征值是图中指示为可能与候选设备对的设备配对的附加设备的数量的指示。该示例指示提供了不一定单独存在于三元组信息中的较大图的视图。该额外的潜在配对信息是“反馈”F的示例，通过该“反馈”F，聚类引擎206可以影响使用评分引擎204产生的精确分数，这又可以影响与图结构的边相关联的精度分数。

[0066] 此外，例如，聚类引擎206可以将先前的精确对分数作为反馈F提供给评分引擎204，以由评分引擎204确定稍后更新的精确对分数。使用评分引擎204产生的先前对分数可允许评分引擎204确定新的更新对分数以改进当关于候选设备对的更有限信息可用时(例如当较少的特征信息可用时)产生的先前对分数。应当理解，识别可能的设备对是时间上的迭代过程，随着更多信息变得可用于系统200，潜在地给定配对的预测正确的概率更大。

[0067] 接下来，将描述在操作中在系统200内采用的设备和设备标识符。如本文所使用的，术语“设备ID”和“设备标识符”和“数字身份”是指例如用户设备的标识符，存储在计算机可读储存设备中的用户设备程序，存储在计算机可读储存设备中的用户设备信息结构(例如，软件cookie)或设备用户身份。例如，如本文所使用的，“计算机可读”储存设备指的是非暂时性储存设备，可以使用可编程计算机或设备(例如智能电话)从该储存设备提取所存储的信息。设备ID的示例包括网络浏览器cookie、蜂窝电话设备标识符、MAC地址、用户id以及链接到特定客户端设备、客户端程序或设备用户的其他标识符。例如，如本文所使用

的,术语“设备”或“用户设备”通常用于指代数字实体,诸如笔记本计算机系统、台式计算机系统、蜂窝电话、平板计算机系统、智能手表、诸如互联网连接电器的智能设备和网络浏览器。本公开内容的教导可以与各种不同的设备ID一起使用。在将公开的数字身份配对的示例中,笔记本计算机系统、台式计算机上的网络浏览器cookie和蜂窝电话设备上的设备标识符被用作设备ID。然而,所公开的技术可以与可以用于识别特定客户端设备、网络浏览器、用户或其他数字身份的任何其他合适的设备一起使用。例如,如本文所使用的,字母“D”用于指代诸如移动电话的物理设备,字母“C”用于指代表示诸如笔记本计算机的物理设备的cookie。从前述内容可以理解,在计算机可读储存设备中编码的移动电话和浏览器cookie在本文中可以称为“设备”。

[0068] 接下来,将描述在系统200中操作中使用的特征和特征信息。如本文所使用的,术语“特征”指的是与设备相关联的属性。如本文所使用的,术语“特征信息”指的是指示一个或多个特征的信息,例如设备活动历史。根据本公开内容的一些实施例,通过观察互联网通信来收集特征信息。根据一些实施例,特征信息用于产生作为特征向量的组成分量的特征值。

[0069] 例如,低级别的特征信息包括例如以下信息:

[0070] (i)源设备属性,例如,iPad,已安装应用程序的ID(例如,安装名为“Words with Friends”的应用程序),

[0071] (ii)设备用户的末端,设备ID,地理位置,日期等,

[0072] (iii)目标设备属性(例如,安装的MacOSX,相关的三个其他设备等),

[0073] (iv)上下文属性(例如,网站主题等),以及

[0074] (v)IP属性(例如,在此IP地址上观察到多少个唯一设备,已观察到该IP地址的天数)。

[0075] 通常,与设备对设备相关的信息来自几大类:

[0076] (a)空间和时间观测,

[0077] (b)人口统计学,和

[0078] (c)卫星对。

[0079] 低级人口统计信息包括例如以下信息:(i)该对的C和D方的年龄、性别、家庭位置和消费者兴趣是什么?

[0080] (ii)C和D描述是否与一个唯一用户一致?

[0081] 低级卫星对信息包括例如以下信息:

[0082] (a)是否存在与C或D相关的其他对?(注意:可以理解,在图论中,如果它们共享一个顶点,则两个边被称为“关联”);

[0083] (b)这些对与当前对相比如何?

[0084] 可选地,卫星对信息是从设备图(下面更全面地描述)的角度将信息注回到配对预测过程中。

[0085] FVG引擎212产生特征向量,其表示作为低级别特征信息的函数的关于候选设备对的高级别特征信息。具体而言,在一些实施例中,FVG引擎212被配置为使用规则来作为设备使用信息(例如设备活动历史)的函数产生特征向量。根据本公开内容的一些实施例,基于空时信息开发了许多高级别特征。在一些实施例中,特征FVG引擎212使用空时启发法评估

低级别特征(例如,在IP地址C上收集设备A的所有观测,设备A在夜间或周末期间在IP地址C上出现的频繁程度)。

[0086] 接下来,将更详细地描述配对引擎202。可选地,配对引擎202根据如在2013年5月10日提交的2014年4月3日公开的美国专利申请US 1014/0095320中描述的一些实施例来实施;然而,应当理解,配对引擎202的其他类型的实施方式也在本公开内容的范围内,如下面将更详细地阐述的。在图3中,示出了表示根据本公开内容的一些实施例的图2的系统的数字身份配对引擎的示例配置的说明性过程流程图。最初,模块510从计算机可读特征储存设备208收集用于许多不同数字身份的互联网使用数据。在本公开内容的一些实施例中,为了分析而收集的互联网使用数据包括数据三元组,其包括客户端设备ID、网络源/目的地地址标识符(例如,IP地址)和时间戳。

[0087] 在本公开内容的实施例中,公共源/目的地标识符是如果两个客户端设备相关,则客户端设备(如由它们的设备ID所标识的)可能共同具有的一些源或目的地的身份。在图1A中所示的情况下,IP地址是可用于链接相关客户端设备的公共源/目的地标识符。具体地,数字身份配对系统可以使用笔记本计算机系统151(例如给定的笔记本计算机系统)和蜂窝电话152(例如蜂窝电话)在工作时共享单个IP地址W 123,在家时共享单个IP地址A163的事实来推断出笔记本计算机系统151和蜂窝电话152是相关的客户端设备。

[0088] 时间戳在每个数据三元组中可用于确保所使用的数据是相关的。给定的互联网连接设备的所有权可能随时间而改变,使得应当可选地不使用非常旧的互联网使用数据,例如设备活动历史。在本公开内容的一些实施例中,使用IP(网际协议)地址来实现网络源/目的地地址。许多IP地址是“动态地址”,可以由不同实体在不同时间使用。因此,互联网使用数据观察应该具有相对紧密的时间关系(当在本公开内容的实施例中使用时),以便提供准确的数字身份配对结果;稍后将参考图8B到8E更详细地描述这种方法。除了确保互联网使用观察在时间上接近之外,所公开的系统200的某些实施例以更复杂的方式使用互联网使用数据三元组中的时间戳,如将在本公开内容的稍后部分中公开的。

[0089] 上述互联网使用数据的三元组(设备ID、公共源/目的地标识符和时间戳)可以由跟踪每个接收的互联网服务器请求的互联网网络服务器收集。可替换地,例如,在本公开内容的一些实施例中,在客户端设备上运行并向互联网上的服务器报告使用信息的各个应用程序(诸如游戏、媒体聚合器、实用程序等)也可以是使用数据的源。例如,采用cookie型软件提供为系统200收集的数据,即提供设备活动历史数据是可行的。

[0090] 再次参考图3,在收集互联网使用数据之后,根据本公开内容的一些实施例,数字身份配对的下一步骤是确定候选数字身份对集合。通常,在互联网环境中,每天有数百万个不同的数字身份涉及互联网活动。试图作为潜在的数字身份分析数字身份的每种可能的排列将是一项极其困难且可能是徒劳的计算任务。因此,为了减小数字身份配对问题的大小,模块520分析所收集的互联网使用数据以识别具有是相关的适合(即合理)概率的数量少得多的候选数字身份对。即,通过将预过滤器应用于搜索空间来限制用于确定互联网请求和设备ID之间的关联的搜索空间的大小;这种方法在很大程度上减小了系统200中的计算工作量。

[0091] 在使用IP地址作为公共源/目的地标识符的本公开内容的实施例中,在系统200内可选地使用两种不同的技术来选择潜在的数字身份对以供进一步分析。第一种策略是检查

已知使用相同IP地址的不同数字身份的数量。具体地，如果已知小于阈值数量的数字身份使用特定IP地址，则来自该单个IP地址的所有不同数字身份逻辑配对可被视为潜在数字身份对。一个相关的推理是，如果只有少量不同数字身份与单个公共IP相关，那么很可能这些不同数字身份中的一些与同一人相关联，并且该人能够在统计上链接属于该同一人的数字身份。例如，共享单个互联网帐户的家庭成员可能具有使用可以在本公开内容的实施例中进行统计链接的多于一个数字身份的家庭成员。

[0092] 在本公开内容的一些实施例中，例如，阈值被设置为6，使得如果在特定IP地址处见到六个或更少的数字身份，那么这六个或更少数字身份的各种逻辑组合可被认为是潜在的数字身份对。例如，在图1A中，家庭A 150仅具有六个不同的数字设备(151、152、161、162、157和159)，其通过Wi-Fi路由器165上的单个IP地址A 163耦合到互联网101，使得家庭A 150中的各种数字设备可以被认为是候选数字身份对。相比之下，非常大量的数字设备通过工作场所W 110处的单个IP地址W 123耦合到互联网101，使得数字身份配对系统不立即将工作场所W 110处的数字设备的所有组合认为是潜在的数字身份对。实际上，系统200识别家庭住户(其通常具有少于六个互联网连接设备)，然后尝试配对来自同一用户使用的家庭住户的数字设备。因此，这种方法采用简单的规则，该规则允许在系统200中识别给定客户端的设备。

[0093] 在本公开内容的另一个实施例中，数字身份配对系统考虑特定IP地址源并确定该IP地址是否是可能找到配对数字身份的地址(例如，如上所述的家庭住宅)。互联网上的静态IP地址通常由互联网名称与数字地址分配机构(ICANN)分配。通过检查谁拥有特定IP地址，可以确定是否容易识别可能位于该IP地址的相关数字身份。因此，例如，互联网服务提供商(ISP)用于提供住宅互联网服务的IP地址可以是在识别潜在数字身份对时使用的良好IP地址。还可以使用识别住宅家庭IP地址的各种其他系统。此外，除了用于识别住宅家庭的系统之外或代替用于识别住宅家庭的系统，还可以在模块520中使用识别可能的数字身份对的其他技术。

[0094] 在使用模块520选择潜在数字身份对的集合之后，模块540然后处理所收集的互联网使用数据，即所收集的特征信息，具体是三元组信息，以确定候选数字身份配对的初步对分数。由于观察到的互联网使用数据将随时间变化并且某些机会活动可能导致检测到错误的数字身份关联，因此可以对使用配对引擎202产生的初步对分数进行后处理以去除噪声。例如，可以使用各种技术随时间平滑关联分数，例如通过采用时间移动平均。因此，在阶段560，随时间生成的关联分数数据可以是使用模块540产生的后处理对分数，使得很大程度上滤除异常数据点。阶段560，例如实施为模块，产生高概率数字身份配对的集合。

[0095] 应当理解，在本公开内容的一些实施例中，使用配对引擎202产生的初步对分数提供的初步对分数比使用评分引擎204产生的对分数更粗略，即更不精确。在本公开内容的一些实施例中，在确定精确对分数(即，“置信度分数”)时，评分引擎204考虑使用配对引擎202产生的初步对分数。配对引擎202将这些候选设备对及其相关的初步对分数识别到评分引擎204。此外，如下面更充分说明的，所识别的候选设备对的设备ID与计算机可读特征储存设备208内的原始或预处理数据209一起使用，以检索附加特征信息，即附加设备活动历史，供FVG引擎212使用以产生对应于候选设备对的特征值。参考图1B，示出了图1A的家庭A 150和工作场所W 110以及额外的家庭B 180和网吧C 190的图示。在家庭A150中，有两个数字设

备用户:用户X和用户Y。用户X经常使用笔记本计算机系统151和蜂窝电话152,从而笔记本计算机系统151和蜂窝电话152的数字身份可被识别为配对的数字身份。使用网络浏览器上的cookie(C)数字识别用户X的笔记本计算机系统151,使得它被表示为C.sub.X,并且以设备ID(D)数字识别用户X的蜂窝电话b,使得它被表示为D.sub.X。该数字设备参考指示命名法也将与其他笔记本电脑和电话一起使用。家庭A150的用户Y经常使用笔记本计算机系统C.sub.Y 161和蜂窝电话D.sub.Y 162,使得C.sub.Y161和D.Yub 162也可识别为相关的数字身份对。可以理解的是,家庭A 150中的所有数字设备(C.sub.X 151、D.sub.X 152、C.sub.Y 161和D.sub.Y 162)可操作以使用当在家庭A 150处使用那些数字设备时,分配给Wi-Fi路由器165的相同单个IP地址A 163。此外,应当理解,尽管该示例实施例基于使用网络浏览器cookie和移动设备标识符,可任选地使用可与数字设备相关联的任何其他类似标识符。

[0096] 居住在家庭B 180的用户Z经常使用笔记本计算机系统C.sub.Z 181181和蜂窝电话D.sub.Z 182182182。当在家庭B 180时,C.sub.Z 181181和D.sub.Z 182182182将使用分配给在家庭B 180处使用的Wi-Fi路由器185的IP地址B 183。用户X和用户Z在工作场所W 110一起工作,使得C.sub.X151、D.sub.X 152、C.sub.Z 181181和D.sub.Z 182182182经常在工作场所W110使用。当在工作场所W 110时,那些数字设备将全部使用在工作场所W 110处分配给防火墙/代理服务器121的单个IP地址W 123。许多其他数字设备(111、112、113、114、115和116)也将使用工作场所W 110处的单个IP地址W 123。

[0097] 最后,参考图1B,示出了网吧C 190的图示,其向网吧C 190的顾客提供免费的Wi-Fi服务。用户Z和用户Y经常访问网吧C 190,使得在网吧C 190处示出了C.sub.Z 181181、D.sub.Z 182182182和D.Yub 162,其中在Wi-Fi路由器195上使用IP地址C 193。可以理解,许多其他访客(未示出)也经常访问网吧C 190。然而,仅在网吧C 190一起看到的各种数字设备将不被认为是潜在的数字身份对,因为在网吧C 190处一起看到太多的数字身份配对。

[0098] 在收集互联网使用数据之后(如参考模块510所解释的),识别数字身份对的下一步骤是选择潜在数字身份对集合,如参考图3的520所解释的。如前一部分所述,可以选择与具有六个或更少数字身份的IP地址相关联的数字身份的各种组合作为潜在数字身份配对。在图1B的示例中,家庭A 150和B 180处的数字设备的各种组合因此被认为是候选数字身份对。(工作场所W 110和网吧C 190具有与它们相关联的太多不同的数字身份,因此不能为数字身份对提供良好的候选。)为简单起见,在该示例中将仅分析来自家庭A 110的可能的数字身份配对(C.sub.X,D.sub.X),(C.sub.X,D.sub.Y),(Csub.Y,D.sub.X)和(C.sub.Y,D.sub.Y)。

[0099] 在本公开内容的一个特定实施例中,例如,配对引擎202使用贝叶斯概率分析的变化来计算每个潜在cookie和设备ID数字身份对的初步对分数,也称为“关联”分数。在一些实施例中,还确定“支持”分数和“置信度”分数。支持、置信度和关联分数可以定义如下:

[0100] 支持=P(cookie,设备ID)

[0101] 置信度=P(cookie\设备ID)

[0102] 关联(cookie.fwdarw.设备ID)=P(cookie|sheb ID)/P(cookie)

[0103] 这三个分数可用于识别数字身份配对并评估已经进行的数字身份配对的置信度。支持分数表示分析此特定cookie和设备ID对的数据支持量。置信度分数表示初步对分数

(也称为关联分数)中有多少置信度。使用配对引擎202产生的初步对分数提供了cookie和设备ID的关联密切程度的评级。在本公开内容的一些实施例中,使用配对引擎202产生的初步对分数被用作评分引擎204中的特征。

[0104] 在本公开内容的一些实施例中,可选地使用上述互联网使用、三元组特征信息来计算支持分数、置信度分数和初步对(关联)分数。以下关系描述了根据一些实施例如何使用互联网使用信息来计算支持、置信度和初步对(关联)分数:

[0105] 共现(cookie,设备ID)=给定cookie和给定设备ID与相同的位置有关(即,相同的源标识符IP地址)的次数。

[0106] P(cookie,设备ID)=共现(cookie,设备ID)/总样本量

[0107] P(cookie\设备ID)=共现(cookie,设备ID)/出现(设备ID)

[0108] P(cookie)=出现次数(cookie)/总样本量

[0109] 接下来参考图4A,示出了表示使用图2的配对引擎202识别的多个示例候选设备对和初步对分数的说明图。例如,设备对数据结构将初步对分数S3与具有设备标识符D1并具有cookie标识符C2的候选设备对相关联。应当理解,根据本公开内容的一些实施例,初步对分数S3是0和1之间的值;然而,在某些情况下,初步对分数S3可以大于1。候选设备对及其对应的初步分数作为输入提供给评分引擎204。在本公开内容的一些实施例中,作为特征向量中的特征值提供初步分数。

[0110] 接下来,将更详细地描述在系统200的操作中采用的特征向量。如本文所使用的,特征向量指的是包括与候选设备对相关联的多个特征值的集合。此外,特征向量是存储在计算机可读储存器中的信息结构。每个特征值表示与候选设备对相关联的一个或多个特征。

[0111] 根据本公开内容的一些实施例,特征向量用于在评分引擎204的训练期间评估所识别的设备对之间的相关性。FVG引擎212作为特征信息(即活动历史)的函数产生包括在特征向量内的特征值,其可以与候选设备对的各个设备相关联,或者与设备对相关联,并且与一个或多个规则相关联。因此,在本公开内容的一些实施例中,例如,每个特征向量表示N个特征(例如,F.sub.1、F.sub.2、F.sub.3、F.sub.4、……、F.sub.N)。在本公开内容的一些实施例中,每个候选设备对与特征向量(V)相关联,其中每个特征具有整数值(F.sub.value1、F.sub.value2、……、F.sub.valueN),将该特征的状态表示为与相关的设备对有关;可选地,每个特征可以具有分数值,例如概率指示值,并且实施规则以考虑这样的分数值。因此,例如,FVG引擎212使用规则来基于与关联于设备对的F.sub.1特征相关联的特征信息(即活动历史)产生给定特征向量值F.sub.value1;然后,使用不同的规则来基于与关联于设备对的F.sub.2特征相关联的不同特征信息产生特征向量值F.sub.value2;以此类推。

[0112] 接下来,将更详细地描述特征向量生成引擎。在图5中,示出了表示根据本公开内容的一些实施例的FVG引擎212的配置的说明性过程流程图300。该图的模块对应于计算机程序代码(未示出),计算机程序代码可操作以配置包括一个或多个处理器和/或控制组件的计算机系统,以导致由不同模块表示的特定操作的执行。在操作中,模块302选择候选设备对。此外,模块303使用存储在计算机可读特征存储设备208内的原始或预处理数据209来检索与所选对的各个设备ID相关联的存储的特征信息,即设备活动历史。此外,模块304从计算机可读规则储存设备213中选择规则。另外,模块306基于检索到的特征信息(即活动历

史)和当前选择的规则产生特征值以供特征向量中使用。

[0113] 决策模块308确定是否存在要应用于与所选候选设备对的设备相关联的当前特征信息(即活动历史)的附加规则。响应于模块308的将附加规则应用于当前选择的候选设备对的决定(基于应用规则),控制流回模块304并且选择另一规则。例如,哪些规则应用于候选设备对可取决于该对可用哪些特征信息(即活动历史)。相反,响应于通过模块308决定不将附加规则应用于当前所选候选设备对的确定,模块310向评分引擎204提供包括基于与当前所选候选设备对的各个设备相关联的单个特征信息产生的特征值的特征向量,并且决策模块312确定计算机可读特征存储设备208是否包括要评估的附加设备对。响应于决策模块312确定要评估附加设备对,控制流到模块302并且选择另一个候选设备对。相反,响应于决策模块312确定不评估附加设备对,该过程结束并等待另外的候选设备对的识别。

[0114] 应当理解,根据本公开内容的一些实施例,配对引擎202提供候选设备对的初步确定。初始候选对确定用作索引,以用于在计算机可读特征储存设备208中搜索与当前所选候选设备对的单独的不同设备相关联的单独且可能不同的特征信息,活动历史。FVG引擎212将来自计算机可读规则储存设备213的一个或多个规则应用于使用配对引擎提供的索引信息识别的单独且可能不同的信息,以便产生用于产生与所选候选设备对相关联的特征向量的特征值。

[0115] 例如,候选设备对可以包括笔记本计算机和智能电话。可以针对笔记本计算机和智能电话单独收集单独的互联网使用信息,本文也称为低级别特征信息。然后,这种原始或预处理数据209用于识别为笔记本计算机收集的使用信息,并识别为智能电话收集的单独收集的使用信息。在模块306中,FVG引擎212基于针对该对中的两个设备单独收集的使用信息,应用一个或多个规则以产生设备对的特征向量值。例如,使用信息可选地包括设备活动历史。

[0116] 还将理解,如将在下面更充分解释的,所产生的许多候选设备对的特征向量用于产生许多对应候选设备对的精确对分数。图4B是表示根据本公开内容的一些实施例的与示例特征向量VI至V33相关联的图4A的示例设备对的说明图。在本公开内容的一些实施例中,图4A中所示的加权值(即“权重”)可以被包含为特征向量内的单个特征值。具体而言,通过监督和无监督训练,评分引擎204产生精确对分数,以与配对引擎202识别的候选设备对相关联。使用评分引擎204产生的精确对分数又由聚类引擎206用于促进用户设备的聚类。

[0117] 图6A是根据本公开内容的一些实施例的过程650的说明性流程图,表示由模块306应用的示例第一规则根据原始或预处理数据209中与使用配对引擎202识别的示例候选设备对的两个设备相关联的低级别特征信息产生示例高级别特征值的示例细节。应当理解,原始或预处理数据209本质上可以是多样的,并且不以任何方式类似于任何形式的索引结构;例如,数据209包括作为互联网事件的时间序列的设备活动历史。此外,应当理解,模块306根据当前使用模块304选择的规则的细节来配置FVG引擎212;由此基于低级别特征信息产生特征值。模块306不同地配置FVG引擎212以应用不同的规则。具体而言,根据一些实施例,模块306使用存储在计算机可读规则储存设备213中的指令来根据使用模块304选择的规则配置FVG引擎212,以基于使用模块303选择的与使用模块302选择的候选设备对的设备相关联的低级别特征信息来确定高级别特征的特征值。如本文所使用的,“高”级别特征是其值取决于一个或多个其他特征的值的特征。产生的特征值表示设备C和设备D之间的关系

的属性。

[0118] 图6A的示例第一规则过程650使用在原始或预处理数据209中指示的多条低级别特征信息来为指示包括设备C和设备D的候选设备对之间的关系强度的特征产生特征值。如本文所解释的，向量包括多个特征值，使用不同的规则产生每个特征值。示例规则为示例特征产生五个可能的特征值之一：0、1、2、3和4。使用示例规则产生的高级别特征值包含在特征向量中，如下面更充分说明的。例如，规则是分层的。它包括分支确定的层次结构，其中一些确定是其他依赖性确定的谓词。参考图6A，第一决策模块652确定示例候选设备对C和D的两个设备是否访问来自相同IAB类别的信息，例如交互式广告局(IAB)类别。响应于确定C和D未访问具有相同类别的内容，第一决策模块652在操作中产生特征值0。如果C和D未访问具有相同类别的内容，则第一规则过程650以产生特征值0而结束。

[0119] 响应于第一决策模块652在操作中确定示例候选设备对的两个设备访问来自相同类别的信息，第二决策模块654确定该类别是否是体育。第二决策模块654响应于确定尽管C和D都访问具有相同类别的内容，但是设备C和D的估计地理位置相隔超过一百英里而产生特征值1。如果C和D访问具有相同类别的内容，但是设备C和D相距超过一百英里，则第一规则过程650以产生特征值1而结束。

[0120] 响应于第二决策模块654确定两个设备都访问体育类别，第三决策模块656确定两个设备中的任何一个是否在周末时间或晚上时间期间访问体育内容。响应于确定设备C和设备D在周末或晚上期间都不访问体育内容，第三决策模块656产生特征值2。如果设备C和设备D都访问体育内容但是它们都不在周末或晚上期间访问体育内容，则第一过程650以产生特征值2而结束。

[0121] 响应于第三决策模块656确定至少一个设备在周末或晚上访问体育类别，第四决策模块658确定两个设备中是仅一个还是两个在周末或者晚上访问体育类别。第四决策模块658响应于仅确定两个设备中仅一个或另一个在周末或晚上访问体育类别而产生特征值3。响应于确定两个设备都在周末或晚上访问体育类别，第四决策模块658产生特征值4。

[0122] 图6B是表示根据本公开内容的一些实施例的由模块306应用的示例第二规则的示例细节的示例第二规则过程660的说明性流程图。图6B的示例第二规则过程660应用于与图6A中相同的候选对C和D，为示例特征产生六个可能的特征值之一：0、1、2、3、4和5。第一决策模块662确定示例候选设备对C和D的两个设备是否访问信息来自同一类别。响应于确定C和D未访问具有相同类别的内容，第一决策模块662产生特征值0。如果C和D未访问具有相同类别的内容，则示例第二规则过程660以产生特征值0而结束。

[0123] 响应于第一决策模块662确定示例候选设备对中的两个设备都访问来自相同类别的信息，第二决策模块664确定设备C和D是否在地理上位于彼此的一百英里以内。第二决策模块664响应于确定尽管C和D都访问具有相同类别的内容，但是它们没有位于彼此的一百四十公里内，而产生特征值1。如果C和D访问具有相同类别的内容，但没有位于彼此的一百英里内，则示例第二规则过程660以产生特征值1而结束。

[0124] 响应于第二决策模块664确定设备位于彼此的一百四十公里内，第三决策模块666确定该类别是否与体育有关。响应于确定该类别不与体育相关，第三决策模块666产生特征值2。如果访问相同类别的设备C和设备D位于彼此的一百四十公里(100英里)内，但该类别与体育无关，则示例第二规则过程660以产生特征值2而结束。

[0125] 响应于第三决策模块666确定所访问的内容与体育有关，第四决策模块668确定两个设备中是没有一个、仅一个还是两个在周末或晚上访问足球运动类别。第四决策模块668响应于确定没有一个设备在周末或晚上访问足球运动相关内容而产生特征值3。第四决策模块668响应于确定两个设备中仅一个或另一个在周末或晚上访问体育类别而产生特征值4。第四决策模块668响应于确定两个设备在周末或晚上都访问体育类别而产生特征值5。

[0126] 应当理解，图6A至6B的规则过程。提供了由FVG引擎212应用的两个示例规则，为基于低级别特征信息产生的高级别特征产生示例特征值。将理解，FVG引擎212被配置为应用许多规则以基于所观察到的低级别特征信息来确定要包括在特征向量中的许多对应特征值。

[0127] 因此，FVG引擎212使用低级别特征信息和预定义规则集产生高级特征值(或示例，0、1、2、3、4、5)。在一些实施例中，生成高级别特征值涉及使用异构低级别特征，诸如指示如下的特征；

[0128] (i)一个或多个类别(例如，IAB类别)；

[0129] (ii)一个或多个地理位置(例如，相互在一百四十公里(100英里)内)；和

[0130] (iii)一个或多个时间(例如，在周末或晚上)。

[0131] 此外，在本公开内容的一些实施例中，生成高级特征值涉及以连续顺序应用多个规则(例如，第一决策模块662、第二决策模块664、第三决策模块666和第四决策模块668各自以预定义顺序实施预定义规则)。

[0132] 示例序列中的每个规则都会生成高级别的特征。当FVG引擎212应用示例规则序列时，低级别特征与较高级别特征混合。具体地，例如，第一决策模块662可以仅基于一个低级别特征(即类别)产生特征值0。第二决策模块664可以产生值1以表示代表低级别类别特征和地理邻近特征的组合的高级别特征。第三决策模块666可以基于代表低级别类别特征、地理邻近特征和体育类别特征的组合的高级别特征来产生特征值2。第四决策模块668可以基于低级别类别特征、地理邻近特征、体育类别特征和时间范围特征的组合来产生特征值3、4或5。

[0133] 在图7中，示出了根据本公开内容的一些实施例的提供用于产生特征值的图2的系统的计算机可读特征储存设备208内的低级别特征信息和计算机可读规则储存设备213内的规则的示例的说明性图。表中的许多示例规则是设备C、设备D和IP地址的观察历史的相对粗略的描述符。应当理解，并非所有规则都适用于每个潜在的设备对，并且只有规则的某些部分可适用于任何给定的设备对。

[0134] 接下来，将更详细地阐明在系统200中操作中使用的标签。存储在计算机可读特征储存设备208中的标签信息(即ID对)用于训练评分引擎204以作为设备对特征向量的函数生成设备对分数。具体而言，给定特征向量包括与对应设备标识符相关联的标签值。有标签数据用于训练评分引擎204以确定无标签设备对的两个设备与同一给定用户相关联的可能性。与有标签设备对相关联的特征向量通常被称为“基础事实”。具体地，提供标签数据以指示可验证地已知与同一给定用户相关联的特征向量，并且还提供标签数据，其指示可验证地已知不与同一给定用户相关联的特征向量。这些已知设备对关系在本文中称为“有标签的”。根据本公开内容的一些实施例，每个有标签设备对与标签-1或标签+1相关联。所有其他设备对，本文称为“无标签的”，与标签0相关联。如前面参考计算机可读特征储存设备208

内的特征信息(即活动历史)和计算机可读规则储存设备213内的规则所阐明的,相应候选设备对与相应特征向量相关联。标签＝-1的有标签设备对与已知指示不与同一用户相关联的一对设备的特征向量相关联。标签＝+1的有标签设备对与已知指示与同一用户相关联的一对设备的特征向量相关联。有标签设备对在系统200中用于在训练阶段期间学习以评估相应无标签候选设备对的特征向量是否指示相应候选设备对与同一用户相关联。

[0135] 接下来,将更详细地描述评分引擎204。在图8A中,示出了表示根据本公开内容的一些实施例的评分引擎204的配置的说明性流程图700。图8A中图的模块对应于计算机程序代码(未示出),该代码配置包括一个或多个处理器和/或控制组件的计算机系统,以导致执行由不同模块表示的特定操作。模块702将使用配对引擎202识别的候选设备对与使用特征向量生成(FVG)引擎212产生的特征向量相关联。在本公开内容的一些实施例中,使用配对引擎202产生的初步对分数被合并到用于确定相关候选对的精确对分数的特征向量中。模块702还接收与候选对和特征向量相关联的标签信息。每个候选设备对也与初步对分数相关联。例如,具有设备标识符D1和cookie标识符C2的候选设备对与特征向量V1相关联并且与初始对分数S3相关联。

[0136] 再次参考图8A,模块704在操作中在从配对引擎202接收的候选设备对中识别无标签和有标签的候选设备对。模块704将有标签设备对提供给标签数据选择模块705。模块704提供无标签候选设备对以训练训练模块706。标签数据选择模块705指定第一有标签候选设备对(称为“训练集”)以提供给训练模块706,用于训练使用评分模块708实现的评分模型,指定第二有标签候选设备对(称为“测试集”),以提供给分数测试模块709,用于测试使用评分模型产生的分数的质量。在本公开内容的一些实施例中,有标签数据选择模块705指定接收的有标签候选设备对的百分之六十到百分之九十,例如约百分之八十用于训练,并指定接收的有标签候选设备对的约百分之二十用于测试。

[0137] 训练模块706评估无标签候选设备对的特征向量与已指定用于训练的第一有标签候选设备对的特征向量之间的相关性,并基于相关性评估确定要与第一候选设备对关联的对分数。如上所述,与无标签候选对相关联的特征向量具有标签＝0,并且与有标签候选设备对相关联的特征向量具有标签＝-1或标签＝+1。训练模块706产生由评分模块708使用的评分模型。训练模块706产生将模型参数与向量特征相关联的评分模型。训练模块706使用训练数据以确定模型参数,该训练数据包括与候选设备对相关联的特征向量的阵列和对应标签。根据本公开内容的一些实施例,所确定的模型参数指示向量内的特征对于确定候选设备对的设备是否实际上与同一用户相关联的重要性。训练的目标是产生训练模型,该训练模型使模型预测和训练标签数据中的观察值之间的残差最小化。

[0138] 使用训练模块706产生的由评分模型使用的模型参数用于配置基于特征的评分模块708。评分模块708使用评分模型中的模型参数,结合与候选设备对相关联的特征向量,为无标签候选设备对产生0到1之间的精确对分数。候选设备对的精确对分数表示比对应初步对分数更准确的候选设备对的设备实际上与同一用户相关联的可能性的估计。根据本公开内容的一些实施例,将精确对分数确定为与每个设备对相关联的特征向量(其可以包括指示初步对分数的特征值)和训练评分模型内的模型参数的函数。因此,系统200可选地采用基于以下的迭代过程:

[0139] (i)提出初始候选设备对;

[0140] (ii)基于设备活动历史确定关联概率;和

[0141] (iii)基于一对设备的给定关联的概率高于一个或多个阈值标准进行选择,并且重复到(ii)。

[0142] 分数测试模块709评估使用评分模型确定的精确对分数的质量。具体地,分数测试模块709作为候选设备对的特征向量与由标签数据选择模块705指定用于测试的第二有标签候选设备对之间的相关性的函数,来确定候选设备对的精确对分数质量。精确对分数通过分数质量阈值(即上述“一个或多个阈值标准”)的候选设备对被传递到模块710,模块710将候选设备对及其精确对分数传送到聚类引擎206。从基于特征的评分模块708中去除精确对分数未通过分数质量阈值的候选设备对。

[0143] 在本公开内容的另一示例实施例中,如图8B至8D所示,更详细地示出了确定设备配对的过程。在系统200的操作中,设备活动历史有利地在时间上是本地的,因为旧设备活动历史可能具有误导性(例如,给定用户将他/她的计算机出售给另一个人,或者购买新的智能电话)。因此,系统200在操作中采用逻辑时间T,其具有增量时间T+ΔT。在逻辑时间T,以下条件适合:

[0144] (i)对于数据集A,针对逻辑时间T识别无标签对;例如,在逻辑时间T提取在Pippio数据不能确定的逻辑时间T产生的候选对;“Pippio数据”涉及从美国Arbor Technologies Inc.提供的数据分析软件产品输出的分析数据;

[0145] (ii)对于数据集A,针对逻辑时间T识别有标签对(即,在逻辑时间T生成的候选对,其可以由在逻辑时间T提取的Pippio数据确定);由此确定数据B中的B+正对和数据B中的B-负对;

[0146] (iii)系统200随后在逻辑时间T生成ML模型,在逻辑时间T使用B+正对和B-负对,并用于生成数据集A的分数以获得分析结果C;

[0147] (iv)在逻辑时间T+ΔT,对于数据集D,识别无标签对,即不能从在逻辑时间T+ΔT提取的Pippio数据确定的候选对;

[0148] (v)在逻辑时间T+ΔT,对于数据集E,识别标签对,即不能从在逻辑时间T+ΔT提取的Pippio数据确定的候选对,其中有数据集E中的E+正对和数据集E中的E-负对;

[0149] (vi)从上面的(i)到(v),对于数据集F,接下来计算在逻辑时间T处在分析结果C中有标签对,即可以由在逻辑时间T+ΔT处提取的Pippio数据确定的逻辑时间T处的分数对C,其中存在F+低置信度分数但具有正标签,以及F-高置信度分数但具有负标签;和

[0150] (vii)使用所识别的对E+、F+、E-、F-在逻辑时间T+ΔT开发新ML模型,然后该模型用于生成数据集D的分数;因此,如上所述,通过使用F+和F-对以迭代方式重新训练ML模型,提供了反馈回路。如用于实现与上述图8B至8E中所示的(i)至(vii)相关的过程的模块。

[0151] 参考图8B,示出了设备连接覆盖引擎的概述。服务器日志、SDK日志、出价流等由数据2000表示。将数据2000提供给数据收集器2010。来自数据收集器2010的经处理输出数据被提供给连接生成器2020,其在操作中生成连接覆盖2030;连接覆盖2030依次包括设备内图2040、设备间图2050、身份图2060和家庭图2070。通过分析这些图2040、2050、2060、2070,可以理解可以高度确定地识别出设备配对。来自连接覆盖2030的输出在操作中以迭代方式反馈到连接生成器2020,以提供确定性更大(即代表可靠性)的配对结果。

[0152] 接下来参考图8C,示出了在操作中在数据收集器2010中发生的数据收集的概述。

从给定数据源，例如由数据源X 2100表示，将活动历史数据提供给ETL模块2110，ETL模块2110将相应经处理的数据提供给Botnoise过滤器2120，其在操作中将经过滤的数据提供给归一化和采样模块2130。然后将来自归一化和采样模块2130的输出提供给事件存储器X 2140，其条目用作对应用于系统200中以识别用户设备对之间的关联的规则的输入数据。

[0153] 接下来参考图8D，示出了数据选择的过程，执行该过程用于生成用于确定用户设备的配对的事件数据集。数据存储器X 2200向事件访问控制系统2210提供数据，该事件访问控制系统2210应用各种策略规则，例如策略I 2220，以生成事件集I 2230，该事件集I 2230在确定指示设备之间的关联的配对时使用；“I”是整数。策略规则用于确定使用哪些类别的数据来查找设备之间的关联，例如：

[0154] (i)由所有者x、y、z；

[0155] (ii)按国家A，和

[0156] (iii)按2017/10/20至2017/11/20之间的时间。

[0157] 应当理解，当确定设备的配对时，在系统200的操作中作为替代或添加可选地使用其他类型的类别。

[0158] 接下来参考图8E，示出了用于执行设备内图构建的模块和过程。例如从事件集I 2230导出的事件集I 2300将事件数据提供给特征向量生成器(FVG)2310，其输出数据使候选对生成器2320能够识别设备的潜在配对。将来自候选对生成器2320的输出对数据提供给决策模块2330，决策模块2330可操作以使用特征向量的子集执行确定性配对，以提供低置信度(由“Z1”表示)和高置信度(由“Z2”表示)数据结果。

[0159] 在使用特征向量的子集的ML评分模块2340的操作中提供低置信度结果(Z1)。在ML评分模块2340生成低置信度数据结果的情况下，这些结果用于指示可能的设备间候选对2380。作为ML评分模块2340的结果和高置信度数据结果(Z2)的函数，生成可能的设备内对2350，将其提供给聚类模块2360，用于生成例如设备内图类型1(新片段)2370；然而，由此可以生成其他类型的图，例如如图8B所示。

[0160] 图9是表示根据本公开内容的一些实施例的用于实施图8A的训练模块706的评分引擎204的配置的细节的说明性流程图800。关于图8B至8E采用了大致类似的方法。训练模块706实施半监督学习和监督学习过程。无监督预训练模块802基于有标签和无标签设备对的特征向量执行预训练；如果涉及标签，则有利地采用监督学习，其中，如果有标签和无标签的数据混合在一起，则采用半监督学习。

[0161] 根据一些实施例，无监督预训练模块802产生映射，重映射模块804将使用该映射将有标签设备对的特征向量中的特征重映射到由无监督预训练模块802确定的设备重映射对聚类。重映射模块804映射候选设备对特征向量以重映射存储在计算机可读储存设备中并且用于促进监督学习的聚类。重映射模块804将候选设备对特征向量映射到重映射聚类，以便将使用无监督预训练过程确定的具有更大特征向量相似性的设备对聚类在一起。监督训练模块808基于有标签设备对的特征确定与设备对特征向量相关联的分数。所确定的分数指示对的两个设备都与同一用户相关联的可能性。使用重映射模块804执行的设备对重映射聚类通过将有标签特征向量改进地以它们具有最大亲和度的特征向量重映射聚类为目标来促进监督学习。具体而言，无监督预训练模块802和重映射模块804用于将对应于候选设备对的特征向量映射到重映射聚类，以便基于候选设备对在特征空间中的表示来聚类

候选设备对。如上所述，有标签设备对具有-1或+1的标签值，并且无标签设备对具有标签值0。应当理解，根据本公开内容的一些实施例，无监督预训练以非常稀疏的数据重构特征向量数据以减少维度。

[0162] 监督训练模块808基于有标签设备对的特征确定与候选设备对特征向量相关联的精确对分数。所确定的分数指示候选设备对的两个设备与同一用户相关联的可能性。使用重映射模块804执行的设备对重映射聚类通过将有标签特征向量改进地以它们具有最大亲和度的特征向量重映射聚类为目标来促进监督学习。

[0163] 图10A是表示根据本公开内容的一些实施例的重映射模块804基于候选设备对特征向量将图7的设备对重映射到设备对重映射聚类的示例的说明图。每个候选设备对重映射聚类与特征向量集相关联，该特征向量集不同于使用无监督训练过程确定的其他设备对重映射聚类的特征向量集。到设备对重映射聚类的这个中间映射通常被称为“特征空间嵌入”，并且使用无标签示例的过程通常被称为半监督学习过程的“无监督预训练”步骤。

[0164] 在本公开内容的一些实施例中，无监督预训练模块802在操作中使用贝叶斯网络来训练重映射模块804以产生设备对重映射聚类。在本公开内容的备选实施例中，无监督预训练过程可以使用例如PCA、树嵌入、自动编码器或RBM。具体而言，根据本公开内容的一些实施例，监督学习过程产生重映射的训练数据集806。在图10B中，示出了表示重映射模块804用于根据一般非线性嵌入函数E将示例特征向量Vin从跨度为{x.sub.1、x.sub.2、……、x.sub.n}的原始特征空间重映射到跨度为{l.sub.1，l.sub.2、……、l.sub.k}的潜在空间的说明图，其中k不一定>n或<n。潜在空间表示E(V)允许更有效的学习。

[0165] 监督训练模块808从有标签训练数据推断相关函数，这可以用于估计无标签设备对属于同一用户的可能性。在本公开内容的一些实施例中，监督学习过程使用贝叶斯网络或树集合来产生设备对分数并产生评分模型。在本公开内容的备选实施例中，监督学习过程可以使用回归模型或神经网络。该过程的目标是将可能的设备对的模型估计与有标签数据之间的残差减到最小程度。监督训练模块808产生数学评分模型，其由在重映射模块804中定义的特征空间中的模型参数组成。评分模型用于配置上述评分模块708以对所有其他候选设备对进行评分。

[0166] 接下来，将更详细地描述在评分期间执行的数据变换。参考图11，示出了表示根据本公开内容的一些实施例的包括在评分引擎204内的计算机可读储存设备内发生的数据变换的数据流过程1000的说明图。在操作中，评分引擎204接收候选设备对数据结构1002作为来自配对引擎202的输入。应当理解，在训练阶段期间，候选设备对用于训练评分引擎204。候选设备对数据结构包括标识设备对的每个设备的设备ID，并包括配对设备的相应初步对分数。在第一数据变换中，特征向量与候选设备对相关联，从而产生将候选设备对、其初步对分数与其相关联的特征向量1004相关联的数据结构。在本公开内容的一些实施例中，将初步对分数合并在特征向量内。在第二数据变换中，训练标签数据1005与候选设备对相关联，从而产生候选设备对、其初步对分数、其相关联的特征向量及其相关联的训练标签1006。此处应当理解，有标签特征对的标签值为-1或+1，无标签设备对的标签值为0。在第三数据变换中，涉及无监督预训练模块802和重映射804的无监督预训练使用设备对特征向量及其相关联的训练标签来产生创建聚类的候选设备对聚类(A，B，C，D)1008的亲和力分数，然后将其附加到特征向量。在第四数据变换中，涉及监督训练模块808的监督训练产生与候

选设备对1010相关联的精确对分数并确定用于评分模型的参数。

[0167] 在第四数据变换期间分数测试模块709使用测试标签数据1012来执行分配给设备对1008的分数的精确评估。在精确评估期间使用的测试标签数据1012与训练标签数据1005不相交。即，用作测试标签数据1012的有标签设备对与用于候选设备对数据结构1002的有标签设备对不相交。将设备对放入不同的量化对分数桶中。在第五数据变换中，通过测量每个分数桶中的精度和召回率，将设备对模型分数变换为精确分数。将低于质量阈值的桶的精确对分数中的设备对滤除。从用于配置图8A的基于特征的评分模块708的训练结果的语料库中移除具有无效对分数的最终设备对1014。举例来说，示出了设备对D.subil-D.sub.jl被精确过滤器移除。

[0168] 接下来，将更详细地描述聚类引擎。在图12中，示出了表示根据本公开内容的一些实施例的聚类引擎206的配置的说明性过程流程图1100。该图的模块对应于计算机程序代码(未示出)，该代码配置包括一个或多个处理器和/或控制组件的计算机系统(例如数据处理系统)，以导致执行由不同模块表示的特定操作。模块1102从评分引擎204接收候选设备对和相关联的精确对分数。应当理解，使用配对引擎202识别的一些第一候选对可以由评分引擎204滤除，使得可以由聚类引擎206接收不同的第二候选设备对集合。模块1104组装包含所接收的设备对的图。在本公开内容的一些实施例中，模块1104配置计算机系统充当图生成器。图13是表示根据本公开内容的一些实施例的将图12的聚类引擎用于图4A至4B的候选设备对产生的示例图的说明图。应当理解，该图使用聚类引擎206的评分模块708产生的精确对分数而不是使用配对引擎202产生的初步对分数。图节点对应于设备标识符。连接图节点的图边指示设备对。与图边相关联的精确对分数指示设备对的关联设备实际上与同一用户相关联的可能性。

[0169] 在生成图的过程中，模块1104从设备图中修剪一些图边。具体而言，根据一些实施例，模块1104实施局部图稀疏化过程以在执行图聚类之前清理图。对局部图稀疏化过程的输入包括具有使用评分模块708提供的相关精确对分数的设备对(两个设备ID的集合)。来自图稀疏化过程的输出包括明智选择的较少对。在进行聚类之前选择要移除的对的一种简单方法是在分数上施加平坦切割。例如，将丢弃所有精度分数<0.1的对。尽管平坦切割方法简单，但它有时并不是最佳的，因为相比于它在图的过密集部分上，它往往在图的不太密集的部分上比过于苛刻，破坏了我们保持许多良好聚类的能力。可替换地，局部图稀疏化在平坦切割方法上进行了改进。明智修剪的图可以更好地指示设备对之间的关系。具体地，例如，与共享共同的设备ID的不同设备对相关联的设备也共享图中共同的节点。此外，边分数表示不同设备对的相对强度。

[0170] 模块1106选择图中的潜在用户设备聚类，以评估它们是否实际包括一个或多个用户设备聚类。图14是表示根据一些实施例的示例性提出的用户设备集聚类使用图13的图识别的用户设备聚类U1.sub.A-U9.sub.A的示意图。应当理解，已经移除了多个图边以产生所提出的用户设备聚类。例如，在图14中不存在图13中所示的设备D4和Cookie C2之间的图边。

[0171] 模块1108修改所提议的用户设备聚类以尝试基于一个或多个聚类适合度要求来识别有效的用户设备聚类。用户设备聚类修改可以涉及在潜在设备聚类内添加、删除或组合边，以尝试满足聚类适合度要求。决策模块1110确定通过模块1108的修改是否已经达到

适合度要求。响应于决策模块1110判定尚未达到适合度要求的确定，决策模块1112确定是否继续修改潜在的聚类以满足适合度要求。响应于决策模块1112确定继续修改，控制流回到模块1108。响应于决策模块1112确定不继续修改，模块1114放弃所提出的聚类。

[0172] 响应于决策模块1110确定已经达到聚类适合度要求，聚类准确度过滤器模块1115确定所提出的用户设备聚类的哪些设备标识符将与最终用户设备聚类相关联以及要移除所提出的用户设备聚类的哪些设备。模块1116输出聚类作为最终聚类。在模块1114之后或在模块1116之后，取决于给定潜在用户设备聚类的控制流程，控制流到决策模块1118，决策模块1118确定图中是否存在要评估的更多潜在设备聚类。响应于确定图中存在要评估的更多潜在设备聚类，控制流回到模块1106并且识别另一个潜在用户设备聚类。响应于确定不存在要评估的额外设备聚类，控制流到模块1120，这导致决策模块等待新的潜在设备聚类。

[0173] 图15是表示根据一些实施例的基于图14的所提出的用户设备聚类使用图12的聚类引擎产生的示例性最终设备聚类U1.sub.B-U9.sub.B集合的说明图。相对于所提出的用户设备聚类修改图14的最终设备聚类中的一个或多个，以移除由第二模型训练过程确定的设备指示符。具体地，例如，cookie C13的设备标识符存在于所提出的用户设备聚类U5.sub.A中，但是在用户设备聚类U5.sub.B中缺失。最终设备聚类U1.sub.B-U9.sub.B可以充当数字身份组通信系统，用于与拥有聚类内标识的设备或与聚类内标识的设备相关联的用户进行通信。即，聚类充当数字身份组通信系统，用于识别可用于与相关用户通信的多个设备。

[0174] 接下来，将更详细地描述利用标签传播的聚类适合度确定。根据本公开内容的一些实施例，模块1108至1116经由标签传播执行用户设备聚类，以达到包括分配给唯一用户标识符的节点集的最终用户设备聚类。根据本公开内容的一些实施例的标签传播涉及最初为配对图中的每个顶点分配唯一标签，然后通过相连的边将标签传播到其他顶点，并更新与图中的每个顶点相关联的标签。迭代地执行标签传播和标签更新，直到它不再更新图中每一个节点的标签(如果没有传入标签的聚合分数比当前标签的分数更好，则节点不会更新其标签)。一旦标签传播停止，具有相同标签的任何节点/顶点将被视为属于同一用户。

[0175] 为了传播标签，根据一些实施例，图的每个节点将其标签发送到其邻近(英国英语:“邻近”)节点，并且同时它还将接收从其邻居(英国英语:“邻居”)节点发送的标签。除了发送和接收标签之外，每个顶点还基于其当前标签以及它从其邻居接收的所有标签更新其标签。整个过程在每个单独的节点上并行执行。在每个节点中，有三个参数控制将标签发送到其他节点的行为(英国英语:“行为”)。首先是百分比数，它控制相邻节点的多大部分将从当前节点接收标签，第二个是整数，它限制从当前节点接收标签的最大节点数，第三个是整数，它控制标签在图中的移动距离。

[0176] 有关控制接收标签的邻居节点的部分的百分比参数的详细信息如下:每个节点首先计算与将节点连接到其邻居的所有边相关联的对分数的最大值，然后仅将标签发送到那些分数高于最大值百分比的邻居。

[0177] 关于第二个参数，即使有许多邻居有资格从当前节点接收标签，该参数也用于将其限制在前几个节点。我们将按降序对所有有资格的边进行排序，并仅通过由该参数确定的前几个边传播标签。

[0178] 第三个参数控制标签在整个图中的穿行距离。在操作中，跟踪到目前为止标签已

经行进了多少边，并且一旦其行进的边的数量超过该整数限制，就不允许标签进一步行进。

[0179] 当给定标签穿过图时，它会沿着它行进的路径收集对分数。每个标签以初始分数1开始。对于源自一个节点(例如节点A)到另一个节点(例如节点B)的标签，将该标签的分数定义为节点A处的标签的分数乘以连接节点A和B的边的分数，再除以标签到达该节点所行进的边数。每个节点将聚集其接收到的所有标签的分数，将挑选具有最高分数的标签，如果新分数大于当前标签的分数，则将该标签分配给自身。新标签的分数也将保留用于下一次迭代。如果标签改变，则当前节点会将其传播到其邻居，以便它们在下一次迭代中使用。

[0180] 应当理解，标签传播模块的最终输出取决于所有上述参数的联合行为。虽然很难预测这些参数的哪种组合将提供最佳性能，但实际上已经定义了用户聚类性能指标(例如，用户聚类的精度)，分配了单独的配对输入集合作为训练数据，并实施了网格搜索，以找到在输出用户聚类上产生最佳性能的所有参数的组合。然后将这些参数的最佳值应用于未来的配对输入并生成最佳用户聚类。

[0181] 接下来，将更详细地描述利用模拟退火的聚类适合度确定。可替换地，根据一些实施例，模块1108到1116经由标签模拟退火来执行用户设备聚类。模拟退火是迭代的、概率性的、聚类适合度驱动的算法。退火在图中的每个节点(设备)执行一次，因为每个节点充当最终用户设备聚类的“种子”。当以下两者同时出现时，聚类适合度函数最大：

[0182] (i)聚类的成员是强相互关联的，这意味着观察到成员节点之间的大多数可能的对存在并且它们具有非常高的分数；和

[0183] (ii)聚类与聚类成员之外的设备的连接非常弱。

[0184] 模拟退火过程建议随机地在设备之间添加或去除边(对)，并且以与所得到的聚类的适合度相关的概率接受这些提议。模拟退火过程迭代并且接受概率根据模拟温度调度演变(从更随机化的“高温”阶段开始演变为更为优化集中的“低温”阶段)。当温度低于某个值并且聚类适合度函数的值已经稳定时，该过程停止。

[0185] 模拟退火是“适合度驱动”聚类方法类的一个示例。“适合度驱动”方法是聚类算法的三元组分类之一，即“分裂”、“聚合”或“适合度驱动”，当然，混合可以通过组合任何这些的示例来形成。退火的过程和最终结果需要聚类适合度函数的规范，其中，与流行示例的直觉相反，定制算法有很大的自由度。如前所述，通常使用的聚类适合度函数是在以下情况下最大化的两个因子的乘积：

[0186] (i)所发现的聚类内的设备最大程度地相互关联；和

[0187] (ii)给定聚类内的设备与给定聚类外的设备最大程度地隔离。

[0188] 鉴于聚类适合度的这个定义，退火算法中几乎没有自由参数(即剩下的是温度调度)，因此看起来没有什么可以调整的，并且你从应用退火得到的聚类是“退火聚类”。

[0189] 然而，在特定问题域中，通过手工将附加项添加到聚类适合度函数以惩罚或鼓励在末端聚类中看到的各种效果/度量通常是有益的。由于这些项是手工添加的，因此不知道其各个强度中的哪些应该相互比较，或者与上述原始基线适合度相比较，以便获得最佳聚类。因此，通常引入一个这样的自由参数，其需要针对适合度函数中的每个附加项进行调整。

[0190] 例如，在上面描述的示例图中，可以有几种不同的“类型”的候选设备对(例如，桌面cookie到桌面cookie，移动设备到移动设备，或移动网络cookie到移动设备等)，每个都

有自己的对评分模型。在对类型之间，作为分数的函数的性能是完全不同的，这是由于收集关于各种“设备”的数据的系统性效果的许多差异(例如，笔记本电脑相比于智能电话的不同活动率)。因此，为聚类适合度函数添加附加因子以解决这些差异是合适的(例如，如果你的cookie设备模型是两者中性能更高的，则相比于主要基于cookie-cookie对构建的聚类，更多地信任基于cookie设备对的良好混合构建的聚类)。在这种情况下，完整的适合度函数可以写为：

[0191] $$\text{Fitness}(C)=\sum_{a}\lambda_a\left[\left(\frac{C^a_{in}}{2}\right)^{-1}\sum_{i}^{C^a_{in}}s_i\right]\cdot\left[\frac{\sum_{i}^{C^a_{in}}s_i}{\sum_{i}^{C^a_{in}}s_i+\sum_{j}^{C^a_{ex}}s_j}\right]$$

[0192] 其中“a”的总和是对类型(cookie设备、cookie-cookie等)的总和，各种C表示类型“a”的内部或外部对的数量，λ是我们需要调整以找到最佳性能的自由参数。

[0193] 调整这些参数是作为单独的步骤完成的，即通过优化聚类精度的组合(使用有标签数据)和通过使用整个设备图的子图使用启发式度量。在实践中，聚类精度是选择所采用的最佳参数的主要因素，但是可选地用诸如平均聚类适合度、设备/聚类的平均数量等的简单度量来补充它是可行的。在给定要优化的期望量的情况下找到最佳参数实际上是常规优化问题。解决此问题的一些最简单的技术包括网格搜索和蒙特卡罗方法(其中，搜索区域由限制搜索区域的大小的边界条件定义)。对于自由参数远多于四个的适合度函数，计算通常非常昂贵并且通常必需做出更“智能”的事以找到良好的最佳值(例如马尔可夫链、蒙特卡罗或其贝叶斯变型)。

[0194] 在该优化阶段中的数据划分类似于基本监督学习的划分：选择整个图的随机子图作为给定“测试”数据集，选择另一个子图作为给定“训练集”，由此存在使用训练集中的聚类优化的度量，然后在继续将最佳聚类应用于整个设备图之前，在测试集上验证此性能。

[0195] 接下来，将更详细地描述在设备聚类期间采用的数据变换。在图16中，提供了表示数据流过程1500的说明，图，该数据流过程1500包括在本公开内容的一些实施例中在聚类引擎206内的用户设备聚类期间发生的数据变换。聚类引擎206可操作以作为输入从评分引擎204接收与特征向量和精确对分数1502相关联的候选设备对。在第一数据变换中，创建图1504，其中用户设备由图节点表示，设备对由图边表示，设备对分数与对应的图边相关联。在第二数据变换中，执行图修改以产生所提出的用户设备聚类1506。示出了图14的示例性所提出的用户设备聚类。在第三数据变换中，测试标签数据1510用于确定所提出的用户设备聚类内的设备对是否满足第二精度阈值精度水平。产生最终用户设备聚类1512，其中，将如果用户设备对不满足第二精度阈值水平，则已将其移除。示出了图15的最终设备聚类1512，其中，例如从最终聚类集合中移除聚类U.sub.5B。

[0196] 接下来，将描述在配对和聚类之间找到平衡。使用配对引擎202和在评分引擎204中配对候选设备的方法被设计为产生尽可能接近“Oracle精度”的对分数，意味着其设备真正由同一人拥有的所有对具有分数＝1和所有不是如此的对具有分数＝0。实际上，当然，评

分模型并不完美，绝大多数对的分数都在这些极值之间。这种缺陷的一部分是无法减轻的(无法预测的纯随机噪声，通常称为“方差”)，但大部分缺陷是可减轻的(通常称为“偏差”)。

[0197] 还存在配对引擎中未包含的许多其他有用的信息。原则上，对任何这些信息进行特征化并将其直接输入到配对引擎中以改进系统200在操作中提供的预测是可行的。这些信息中的大部分与关于未包括在该对(即AB)自身中的设备的信息有关，问题如下：

[0198] (i)A的朋友中有多少也是B的朋友?

[0199] (ii)在与B相同的IP上看到多少A的朋友?

[0200] (iii)A和B分别有多少朋友?

[0201] (iv)A的好朋友中有多少也是B的朋友?

[0202] 这些信息中的一些实际上已经被馈送到配对引擎中(迄今被称为“设备图反馈到配对”或“卫星特征”}，等等。

[0203] 应当理解，更加有效的理解该信息的方式是显而易见的：寻求给配对引擎提供可以围绕该对建立的设备图的本地属性的视图。从设备图的角度来看，可以一次性回答所有这些问题，并且相对容易，而在设备图中提供信息的综合编码作为配对引擎的特征所需的工作量是令人望而却步的。因此，作为效率和实用性的问题，比方说，关于向配对引擎202和/或评分引擎204添加多少信息以产生良好的对分数和有多少信息将仅在构造设备图之后在聚类引擎206中考虑“划了界线”。

[0204] 采用这种模块化结构“配对引擎＝>聚类引擎”，而不是试图构造一个超出便利性的大模型，实际上改善了最终结果。事实是，虽然使用配对引擎202和/或评分引擎204的配对候选设备具有完全消除我们的预测中的可减少误差(偏差)的理论能力，但改善配对引擎202和/或评分引擎204的性能所需的工作量通常随着期望的性能改进或增强而指数性地增长。停止短的并将近似配对模型的对分数发送到单独的聚类引擎206允许以少得多的计算量来补偿大部分可能的偏差减少。

[0205] 接下来，将关于其网络环境描述系统200。在图17中，提供了表示根据本公开内容的一些实施例的被配置为在网络环境中操作的图2的系统200的说明图。网络环境1600中的系统200包括第一服务器系统1602，用于从用户接收移动应用程序使用信息和页面查看信息。第一服务器系统1602包括目标引擎1606，用于使信息传输跨多个设备以用户为目标，并提供基于用户的频率上限，用于跨多个设备传递给用户的信息，洞察引擎1608，以提供跨用户的多个设备的跨设备属性，并向用户提供离线-在线属性，优化引擎1610，用于产生基于用户的优化和基于平台的优化。

[0206] 系统200包括第二服务器系统1604，用于产生指示与用户相关联的设备组的用户设备聚类。第二服务器系统1604包括配对引擎202、评分引擎204、聚类引擎206、FVG引擎212、原始或预处理数据209和计算机可读规则储存设备213。根据本公开内容的一些实施例，第二服务器被配置为根据Hadoop框架充当分布式大规模并行处理和分布式储存系统。因此，实际上，配对引擎202、评分引擎204和聚类引擎206引擎的多个实例(如图所示)并行运行。

[0207] 网络环境1600包括第一服务器系统1602和第二服务器系统1604之间的第一接口(在虚线内指示)1612。第一接口1612包括第三服务器1614，其从第一服务器系统1602收集包含与用户与设备的交互有关的信息的日志，例如移动应用程序使用和页面查看。该日志

可选地包括按时间排序的仅附加的有序记录序列。通常，为每个条目分配唯一的顺序日志条目号。由于日志条目是按时间排序的，因此日志条目号可以充当条目的“时间戳”。第三服务器1614被配置为处理高吞吐量、低延迟的实时数据馈送。根据一些实施例，第三服务器1614被配置为充当消息代理，其调解不同应用程序之间的通信，例如向第一服务器系统1602推送出价模型并将日志文件从优化引擎1610推送到第二服务器系统1604。根据本公开内容的一些实施例，第三服务器1614根据日志收集系统文档来实施，但是可以替代地采用其他类型的系统文档。第一接口1612还包括第二服务器1604内的日志提取器模块1616，其从第三服务器1614提取日志并将日志提供给配对引擎202。日志提取器模块1616充当数据归一化器，其将在第三服务器1614中接收的非结构化数据转换为例如适合于输入到配对引擎202的诸如键值对的结构化数据。可以理解，可以从不同地构造数据的不同用户设备平台接收日志内的信息。

[0208] 网络环境1600包括第一服务器系统1602和第二服务器系统1604之间的第二接口(在虚线内表示)1618。第二接口1618包括第二服务器系统1604内的用户简档模块1620，其接收并存储由聚类引擎206产生的用户设备聚类信息，其指示与用户相关联的设备组。如上所述，聚类引擎206产生用户设备聚类结构，其将用户设备ID的聚类与唯一用户相关联。用户简档模块1620存储将用户设备ID聚类与唯一用户相关联的信息。例如在本公开内容的一些实施例中，用户简档模块1620还从日志提取模块1616接收信息，诸如包括性别、年龄、收入、位置和行为模式(例如，搜索模式)中的一个或多个的用户人口统计信息。第二接口1618还包括第四服务器1622，第四服务器1622从用户简档模块1620获得用户设备聚类结构，并将其提供给第一服务器。第四服务器1622被配置为服务于许多并发用户。根据本公开内容的一些实施例，第四服务器1622被配置为存储、检索和管理面向文档的信息，有时称为半结构化数据。面向文档的数据库是一类NoSQL数据库，它是围绕“文档”的抽象概念设计的。根据一些实施例，第四服务器1622使用储存数据服务器来实现，储存数据服务器可以从单个机器聚类到跨越许多机器的非常大规模的部署，并且被配置为以低延迟和高持续吞吐量提供可伸缩的键值或文档访问。

[0209] 网络环境1600包括第一服务器和第二服务器之间的第三接口(在虚线内指示)1624。第三接口1624包括第二服务器系统1604内的活动传递和执行模块1626，其接收并存储指示与用户相关联的设备组的用户设备聚类结构的计数。例如在本公开内容的一些实施例中，活动传递和执行模块1626还从日志提取模块1616接收信息，例如，用户的桌面设备上的广告印象(即，“广告印象”)数，在用户的移动设备上接收的印象数，以及在用户的多个设备上接收的印象数。根据本公开内容的一些实施例，根据基于Hadoop的Hive数据仓库基础设施来配置活动传递和执行模块。例如，Hive基础设施最适合于大型数据集(例如广告活动数据)上的批量作业。第三接口1624还包括第五服务器1628，其从活动传递和执行模块1626中提取活动执行信息并将其提供给第一服务器系统1602。根据本公开内容的一些实施例，第五服务器1628包括SQL服务器，其提供第二服务器对与基于用户的活动和执行范围有关的信息的访问。

[0210] 在不脱离由所附权利要求限定的本发明的范围的情况下，可以对前面描述的本发明的实施例进行修改。用于描述和要求保护本发明的诸如“包括”、“包含”、“结合”、“由……组成”、“具有”、“是”的表达旨在以非排他的方式解释，即允许未明确描述的项目、组件或元

素存在。对单数的引用也应被解释为涉及复数。所附权利要求中的括号内包括的数字旨在帮助理解权利要求,并且不应以任何方式解释为限制这些权利要求所要求保护的主题。

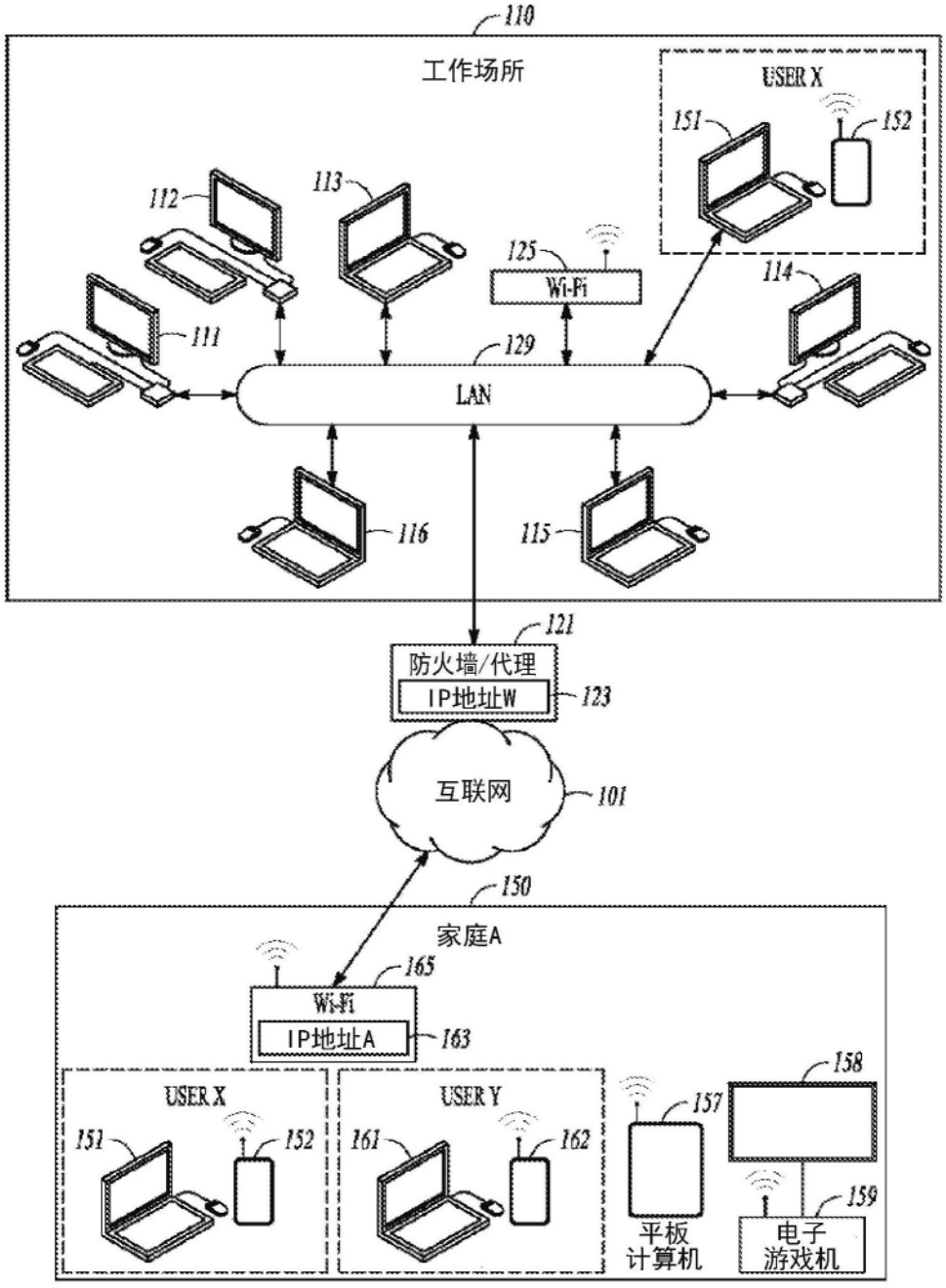

110
工作场所
USER X
151
152
112
113
125
Wi-Fi
114
111
129
LAN
116
115
121
防火墙/代理
IP地址W
123
互联网
101
150
家庭A
165
Wi-Fi
IP地址A
163
USER X
151
152
USER Y
161
162
157
158
159
平板
计算机
电子
游戏机

图1A

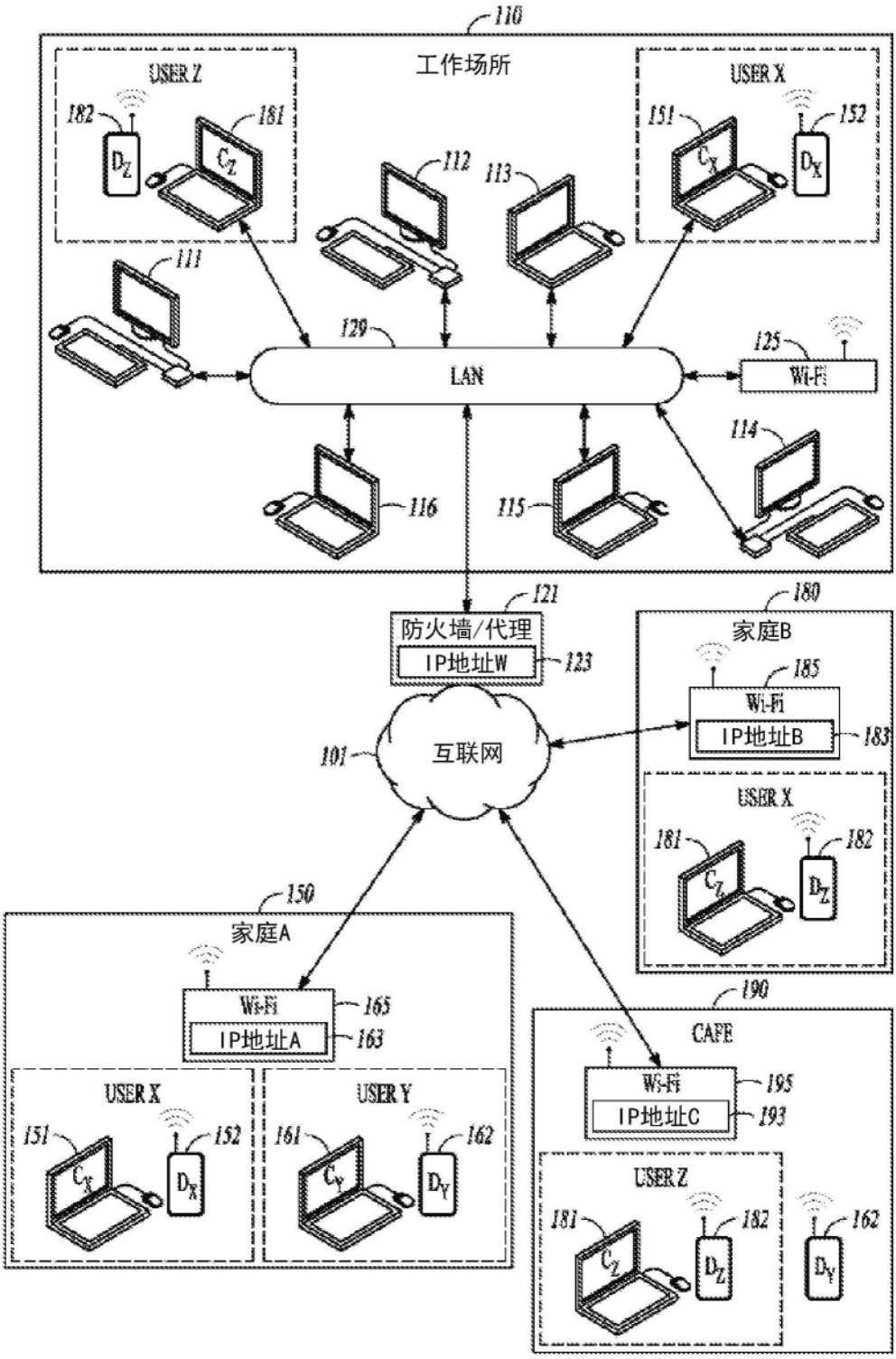

110
工作场所
USER Z
182
181
USER X
151
152
112
113
111
129
LAN
125
Wi-Fi
114
116
115
121
防火墙/代理
IP地址W
123
180
家庭B
185
Wi-Fi
IP地址B
183
101
互联网
USER X
181
182
150
家庭A
Wi-Fi
165
IP地址A
163
USER X
USER Y
151
152
161
162
190
CAFE
Wi-Fi
195
IP地址C
193
USER Z
181
182
162

图1B

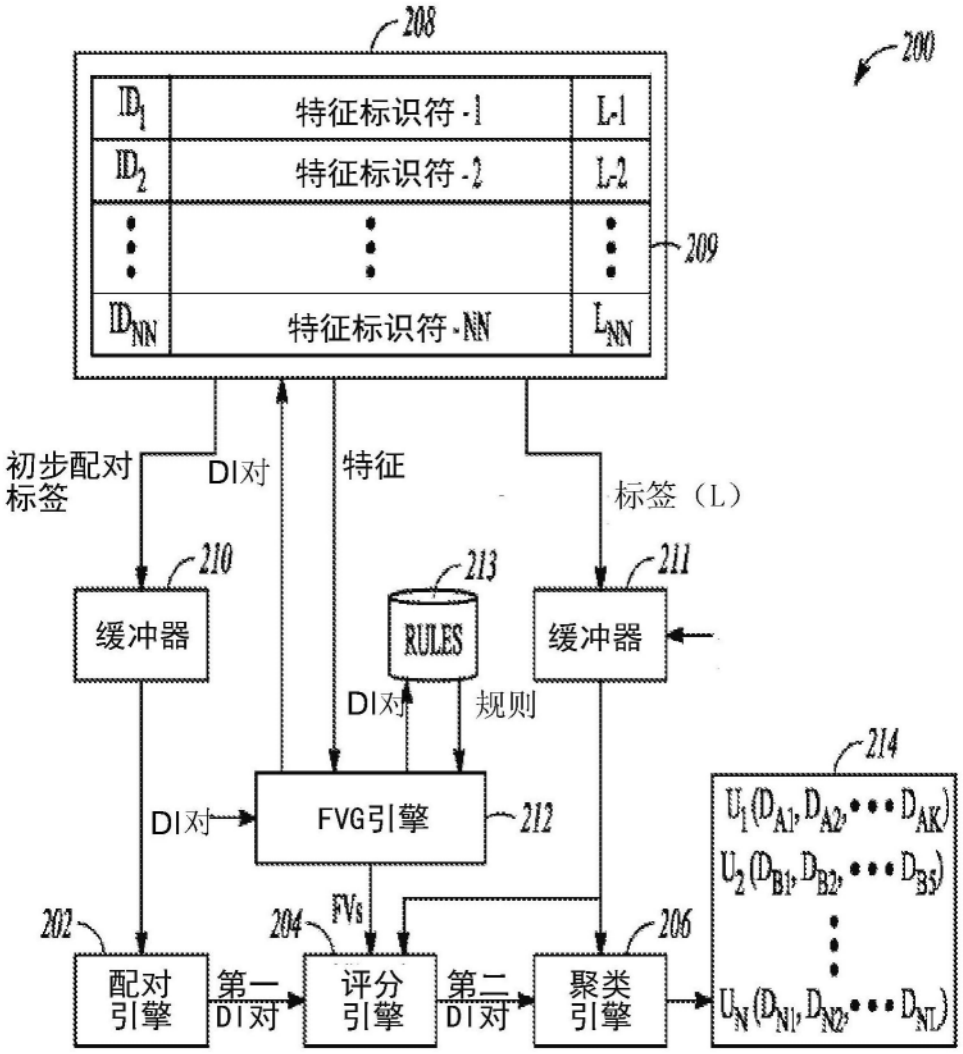

208
200
ID1
特征标识符-1
L-1
ID2
特征标识符-2
L-2
IDNN
特征标识符-NN
LNN
209
初步配对标签
DI对
特征
标签（L）
210
缓冲器
213
RULES
211
缓冲器
DI对
规则
214
DI对
FVG引擎
212
202
204
FVs
206
配对引擎
第一DI对
评分引擎
第二DI对
聚类引擎

图2

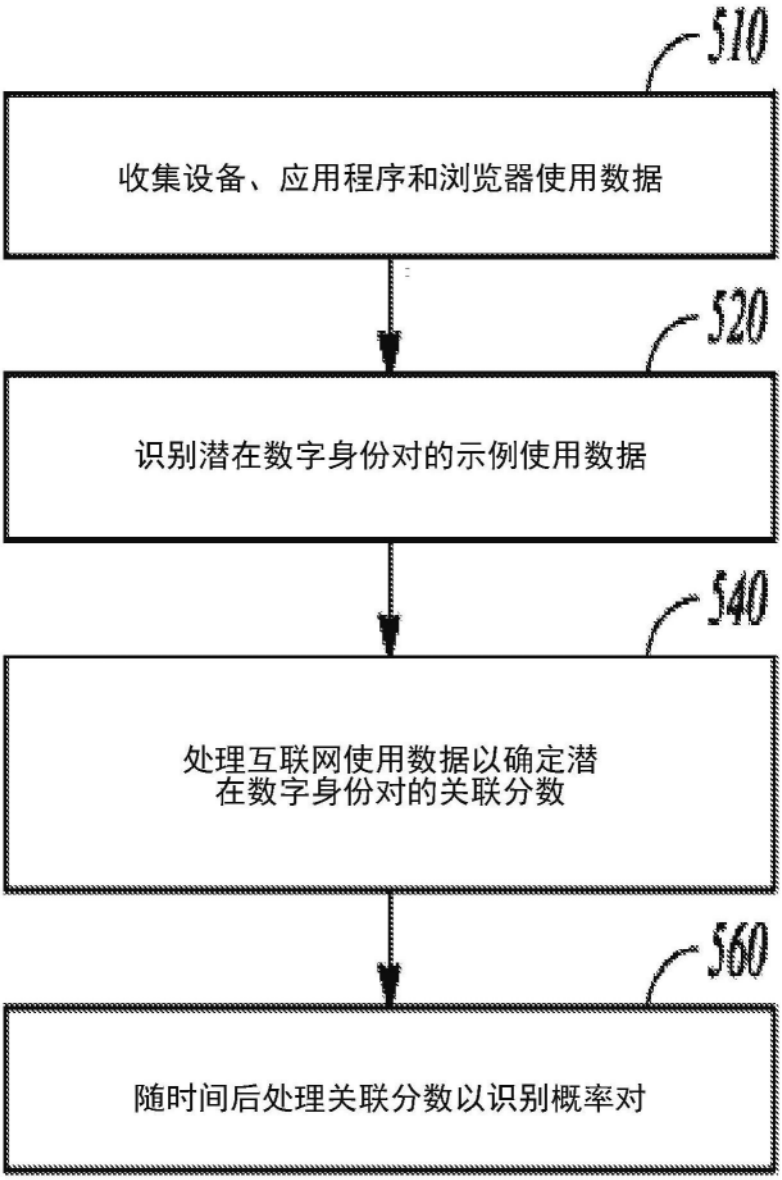

510
收集设备、应用程序和浏览器使用数据
520
识别潜在数字身份对的示例使用数据
540
处理互联网使用数据以确定潜在数字身份对的关联分数
560
随时间后处理关联分数以识别概率对

图3

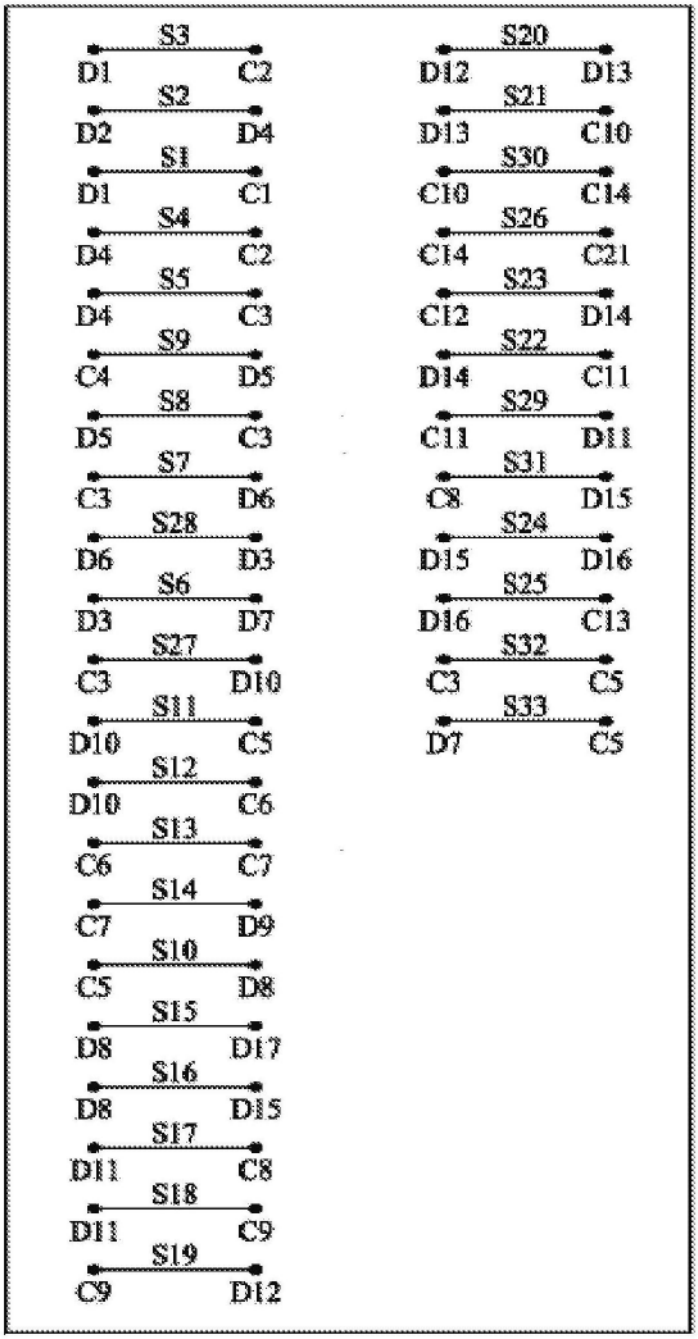

S3
D1 C2
S2
D2 D4
S1
D1 C1
S4
D4 C2
S5
D4 C3
S9
C4 D5
S8
D5 C3
S7
C3 D6
S28
D6 D3
S6
D3 D7
S27
C3 D10
S11
D10 C5
S12
D10 C6
S13
C6 C7
S14
C7 D9
S10
C5 D8
S15
D8 D17
S16
D8 D15
S17
D11 C8
S18
D11 C9
S19
C9 D12
S20
D12 D13
S21
D13 C10
S30
C10 C14
S26
C14 C21
S23
C12 D14
S22
D14 C11
S29
C11 D11
S31
C8 D15
S24
D15 D16
S25
D16 C13
S32
C3 C5
S33
D7 C5

图4A

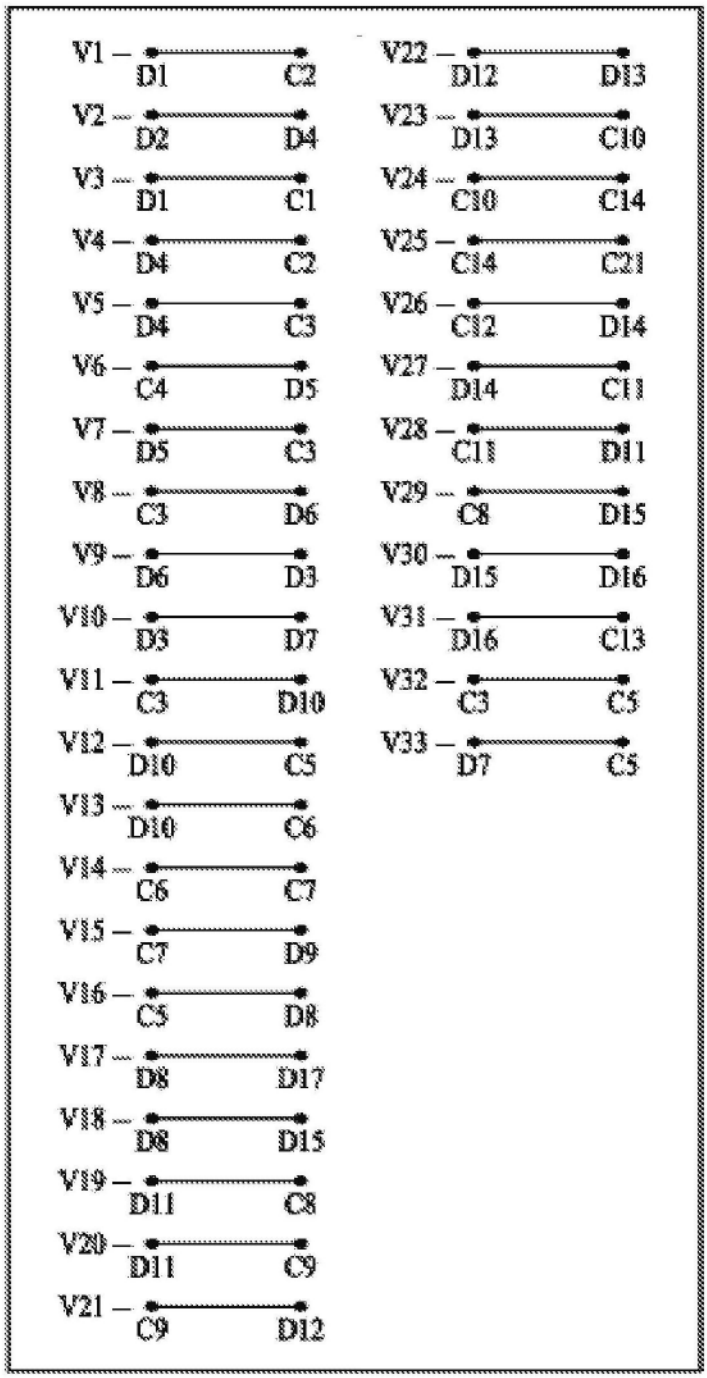

V1 D1 C2
V2 D2 D4
V3 D1 C1
V4 D4 C2
V5 D4 C3
V6 C4 D5
V7 D5 C3
V8 C3 D6
V9 D6 D3
V10 D3 D7
V11 C3 D10
V12 D10 C5
V13 D10 C6
V14 C6 C7
V15 C7 D9
V16 C5 D8
V17 D8 D17
V18 D8 D15
V19 D11 C8
V20 D11 C9
V21 C9 D12
V22 D12 D13
V23 D13 C10
V24 C10 C14
V25 C14 C21
V26 C12 D14
V27 D14 C11
V28 C11 D11
V29 C8 D15
V30 D15 D16
V31 D16 C13
V32 C3 C5
V33 D7 C5

图4B

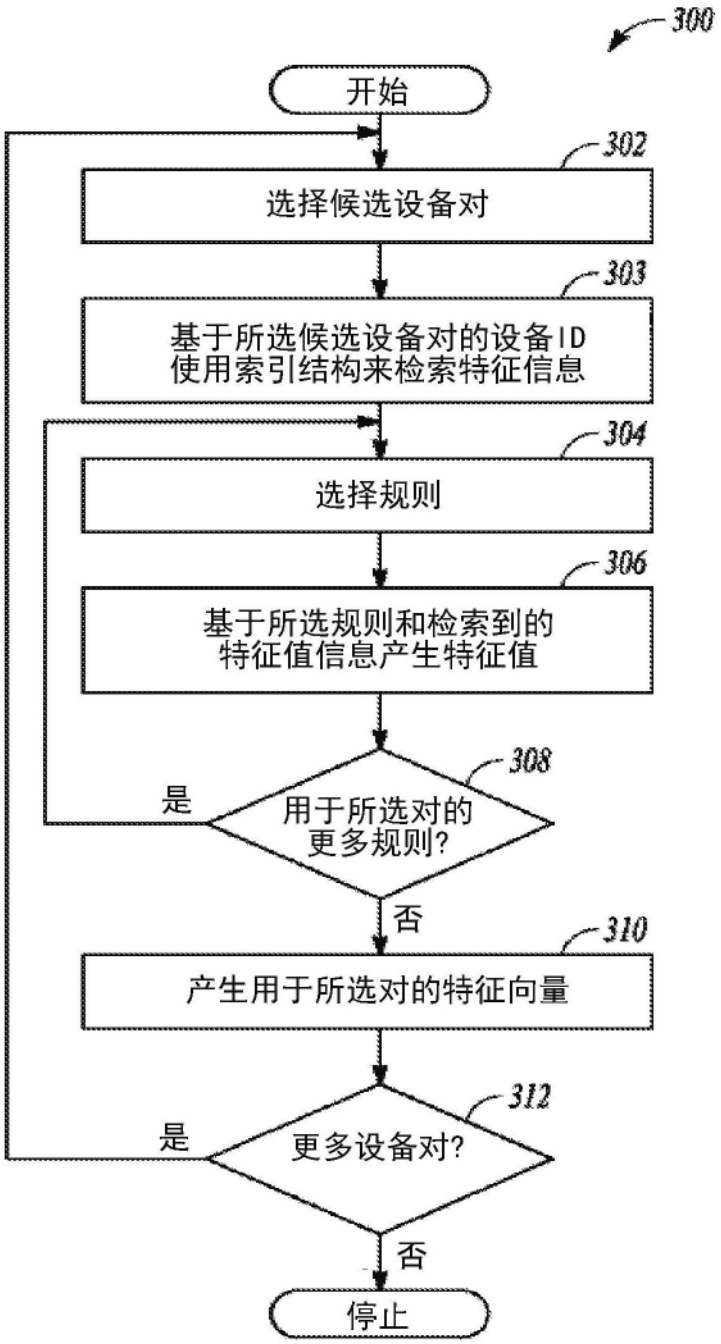

300
开始
302
选择候选设备对
303
基于所选候选设备对的设备ID
使用索引结构来检索特征信息
304
选择规则
306
基于所选规则和检索到的
特征值信息产生特征值
308
是
用于所选对的
更多规则?
否
310
产生用于所选对的特征向量
312
是
更多设备对?
否
停止

图5

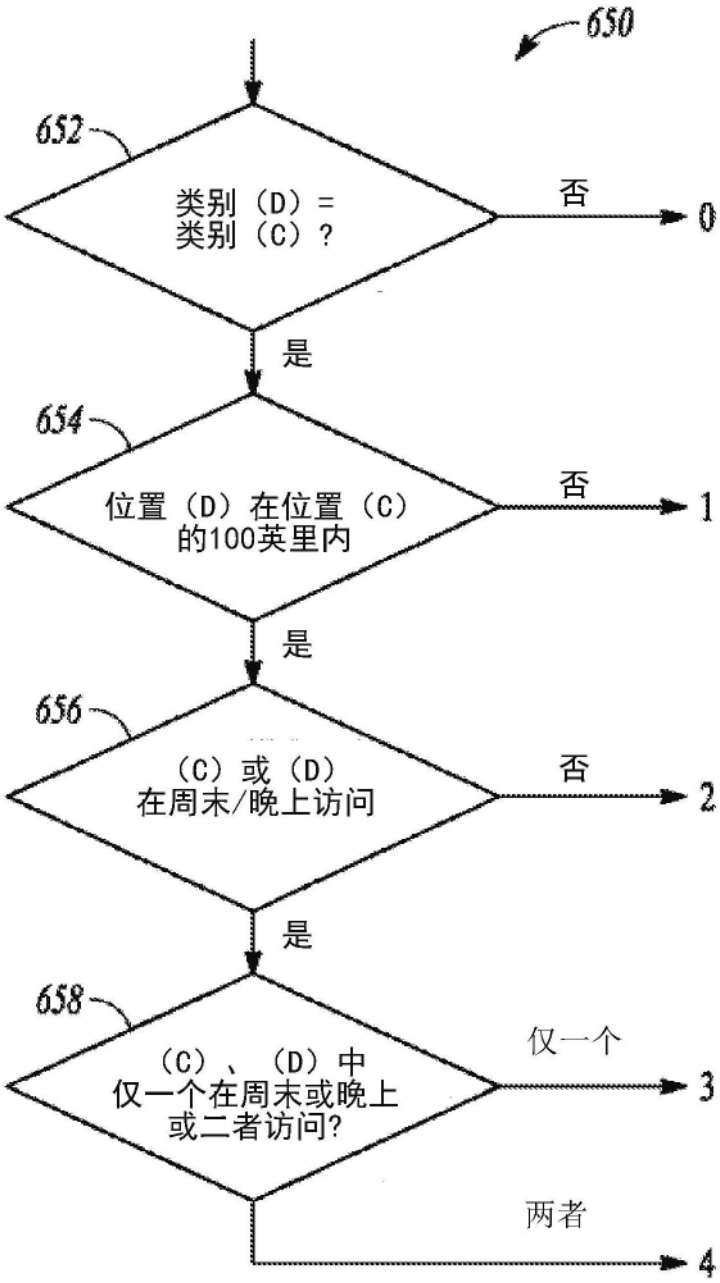

650
652
类别（D）=
类别（C）？
否
0
是
654
位置（D）在位置（C）
的100英里内
否
1
是
656
（C）或（D）
在周末/晚上访问
否
2
是
658
（C）、（D）中
仅一个在周末或晚上
或二者访问?
仅一个
3
两者
4

图6A

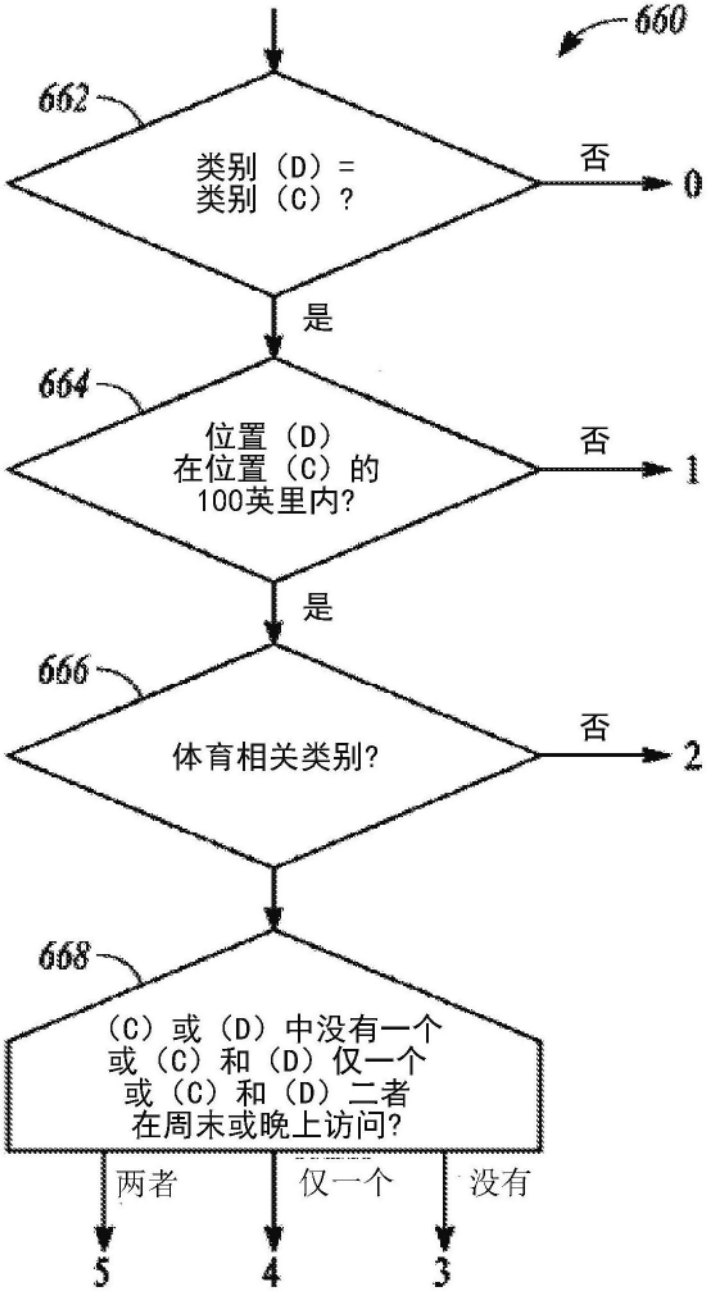

660
662
类别（D）=
类别（C）？
否
0
是
664
位置（D）
在位置（C）的
100英里内?
否
1
是
666
体育相关类别?
否
2
668
（C）或（D）中没有一个
或（C）和（D）仅一个
或（C）和（D）二者
在周末或晚上访问?
两者
仅一个
没有
5
4
3

图6B

特征名称	描述
L	（“标签”）我们能C和D属于唯一用户吗？ 或者它们不属于
NT	什么类型的设备ID是D？(IDFA, RAW, ANDROID, APPLE, ...)
01	我们对于D总共有多少观察？
07	我们对于C总共有多少观察？
CO	有多少仅C或D出现的IP？ （IP集对称差分）
GEN	CD是相同性别？ 相反性别？无性别信息？
AGE	CD是相同年龄？他们年龄差多少？
GEO	CD家庭地点相同？他们的推断家庭地点 匹配接近程度或匹配程度差

图7

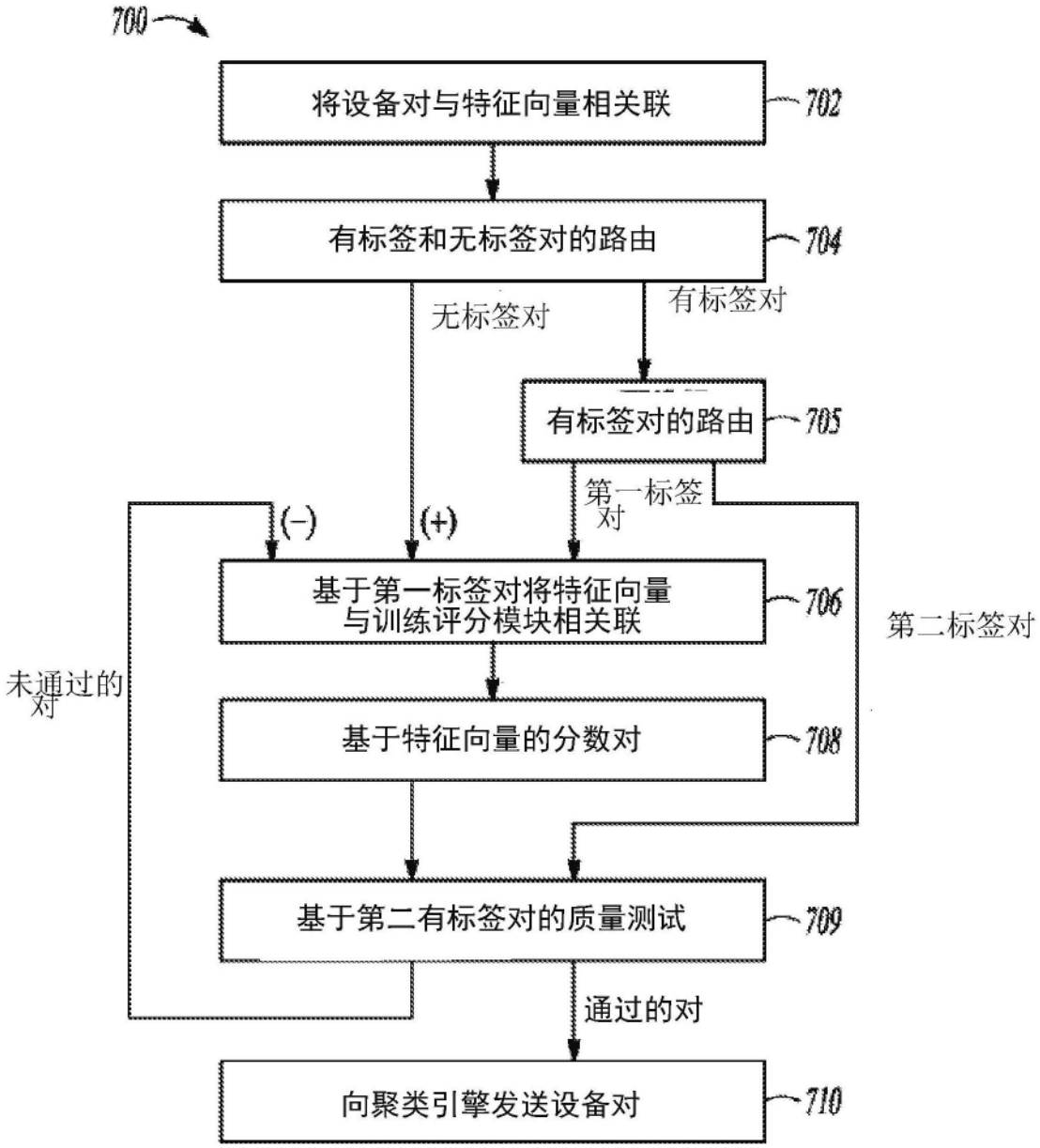

700
将设备对与特征向量相关联
702
有标签和无标签对的路由
704
无标签对
有标签对
有标签对的路由
705
第一标签对
(-)
(+)
基于第一标签对将特征向量与训练评分模块相关联
706
第二标签对
未通过的对
基于特征向量的分数对
708
基于第二有标签对的质量测试
709
通过的对
向聚类引擎发送设备对
710

图8A

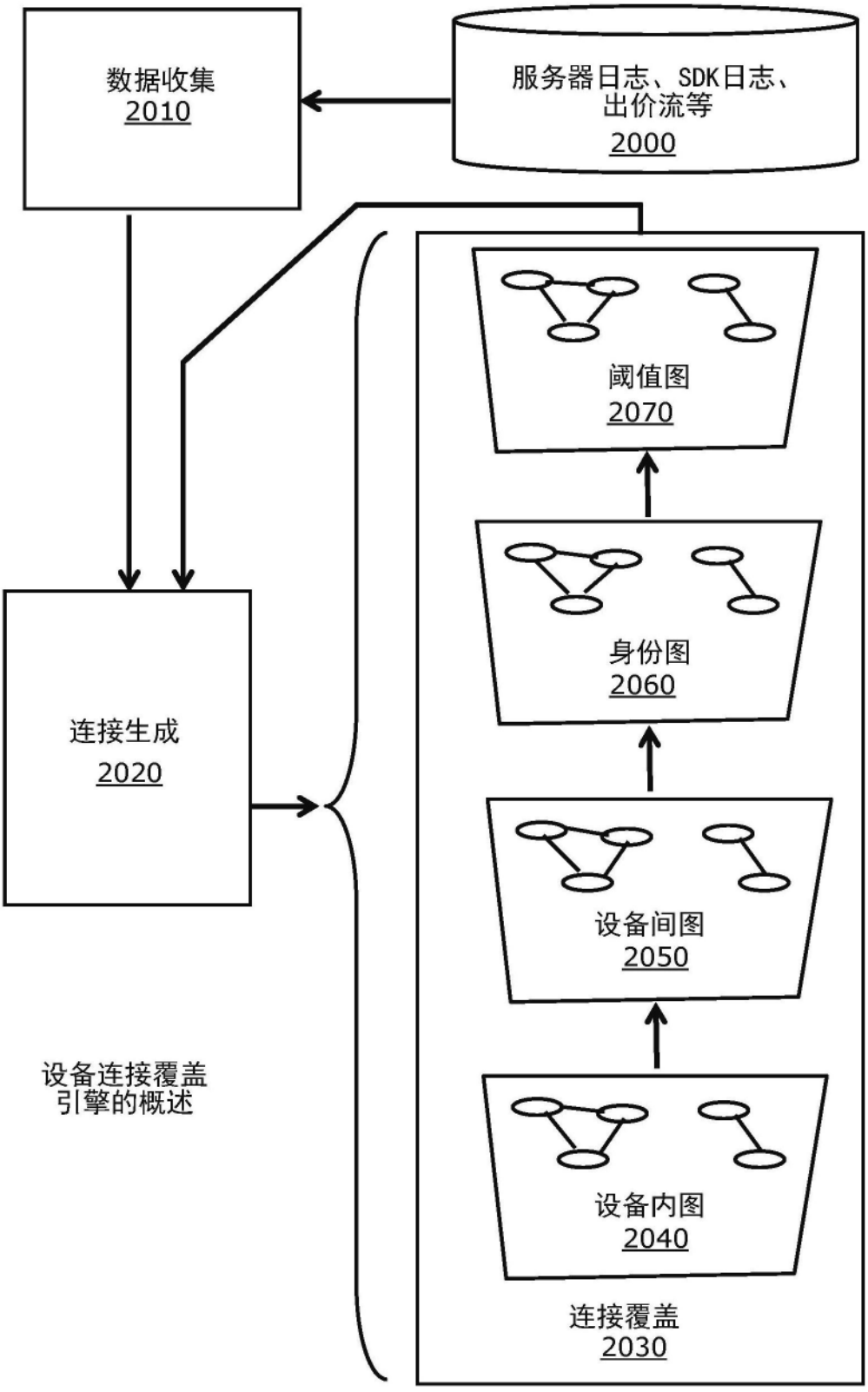

数据收集
2010
服务器日志、SDK日志、
出价流等
2000
阈值图
2070
身份图
2060
连接生成
2020
设备间图
2050
设备连接覆盖
引擎的概述
设备内图
2040
连接覆盖
2030

图8B

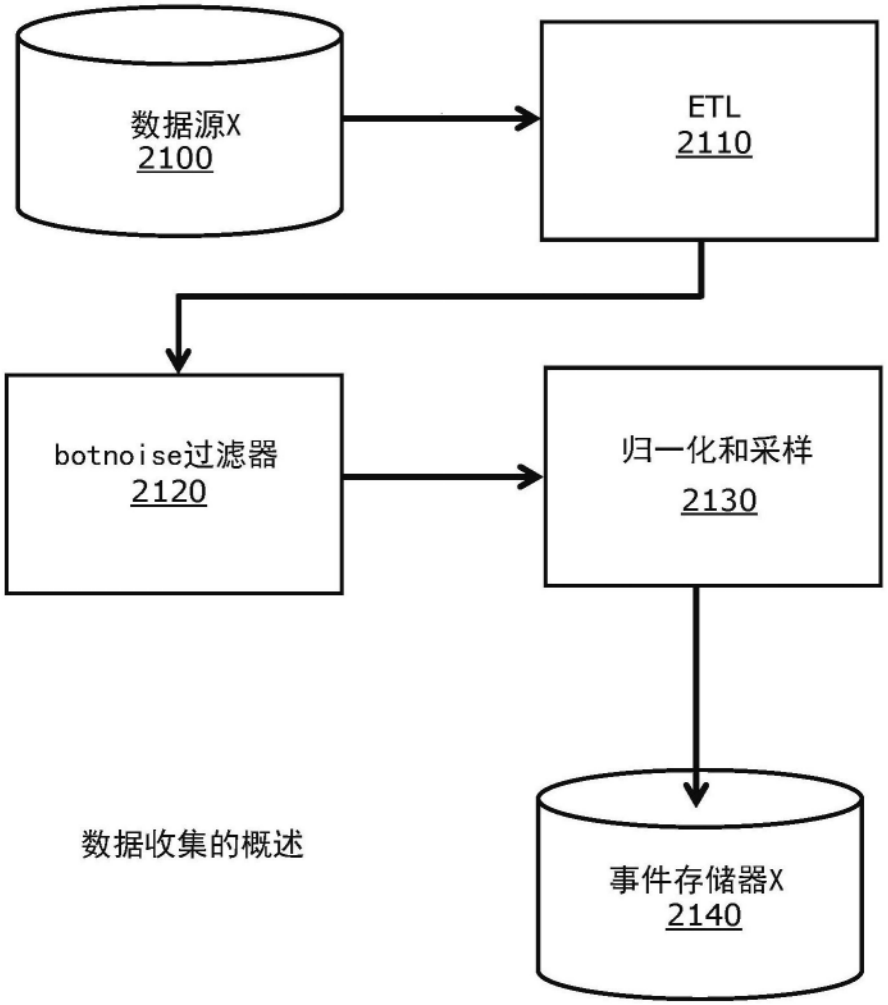

数据源X
2100
ETL
2110
botnoise过滤器
2120
归一化和采样
2130
事件存储器X
2140
数据收集的概述

图8C

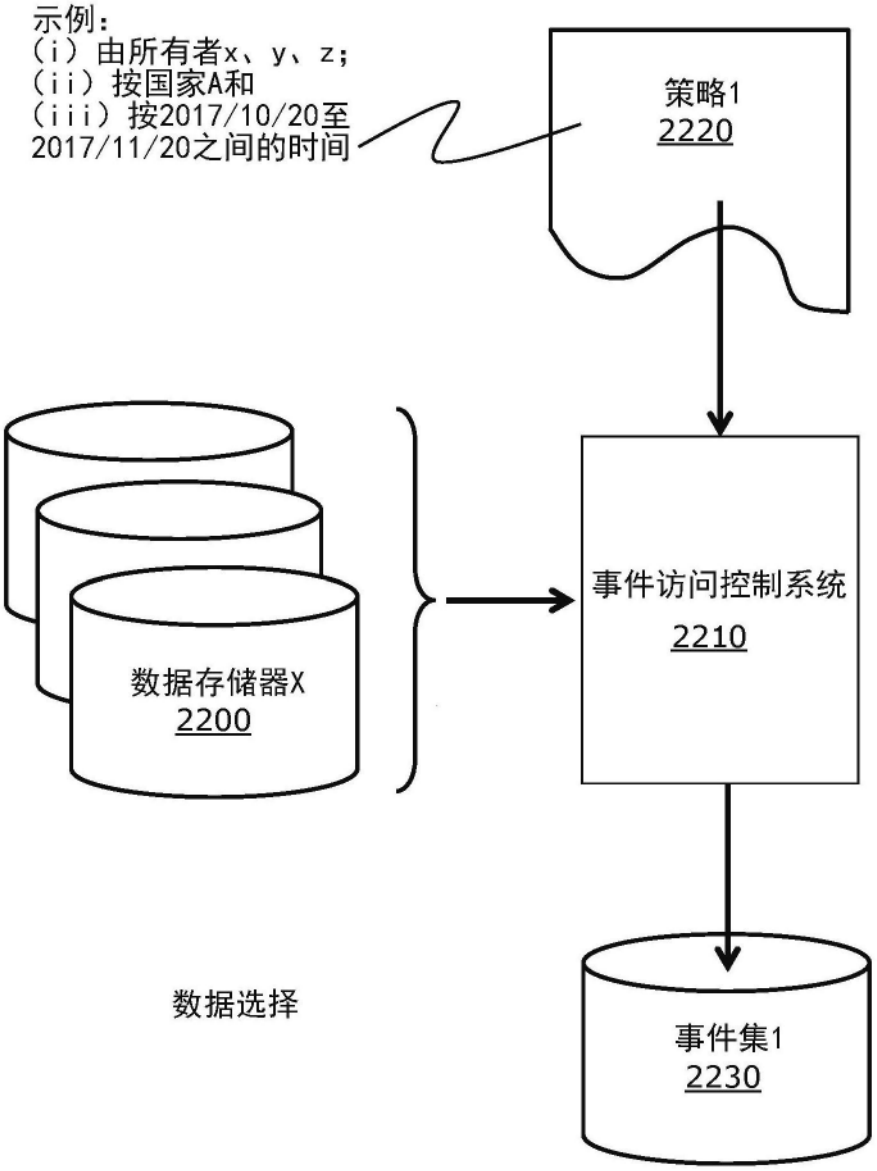
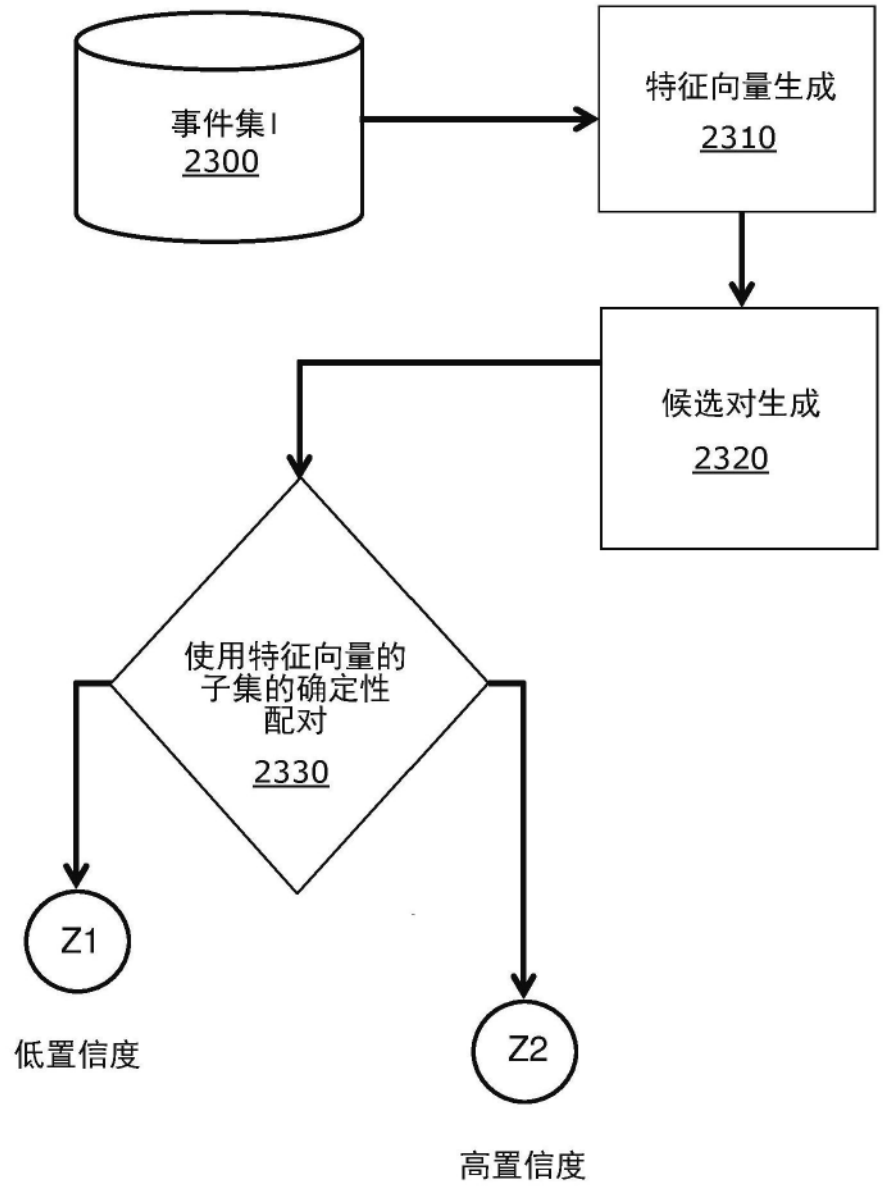

示例：
（i）由所有者x、y、z；
（ii）按国家A和
（iii）按2017/10/20至
2017/11/20之间的时间
策略1
2220
数据存储器X
2200
事件访问控制系统
2210
数据选择
事件集1
2230

图8D


事件集I
2300
特征向量生成
2310
候选对生成
2320
使用特征向量的
子集的确定性
配对
2330
Z1
低置信度
Z2
高置信度

设备内图构建

图8E

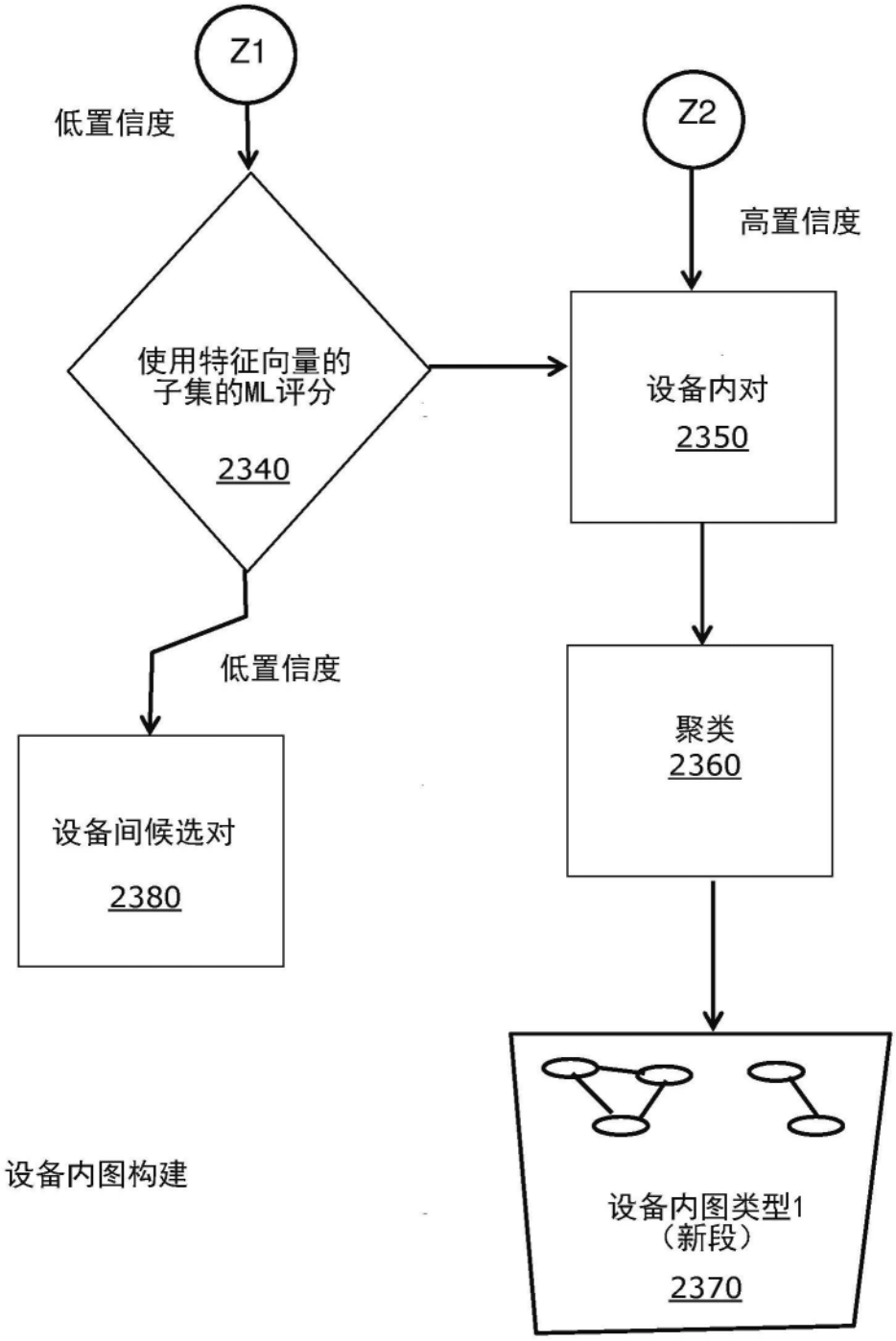

Z1
低置信度
Z2
高置信度
使用特征向量的
子集的ML评分
2340
设备内对
2350
低置信度
聚类
2360
设备间候选对
2380
设备内图构建
设备内图类型1
（新段）
2370

图8E(续)

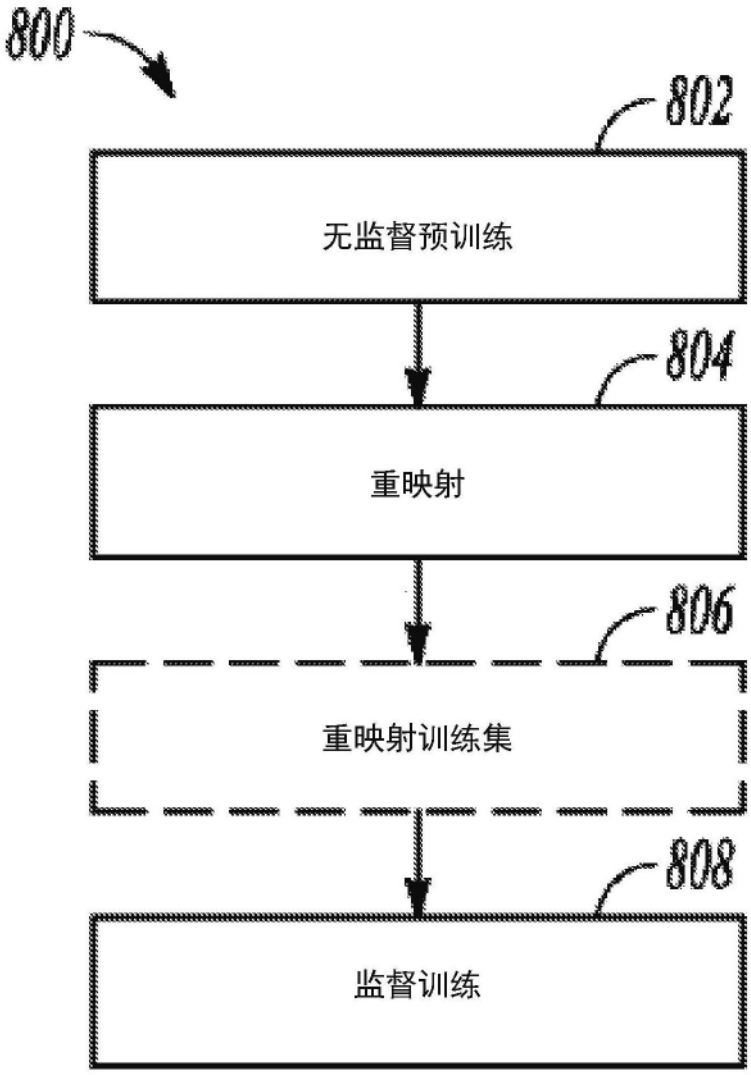


图9

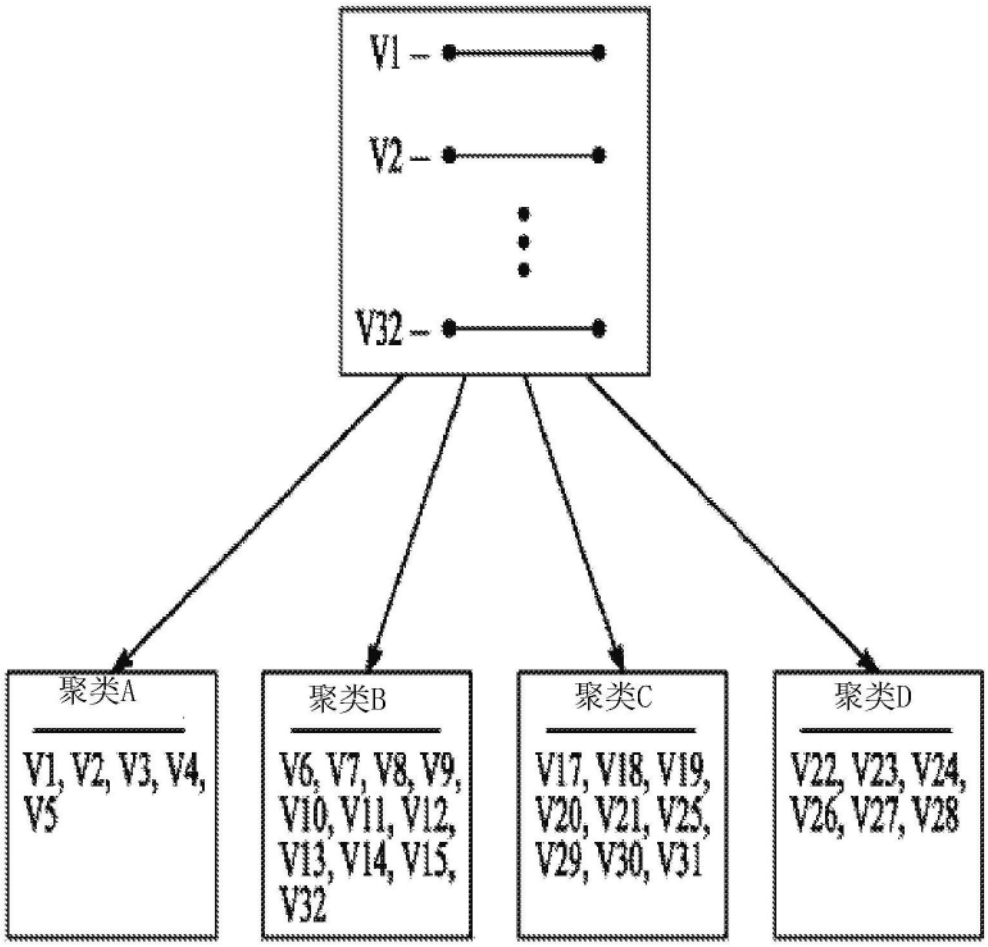

V1
V2
V32
聚类A
V1, V2, V3, V4, V5
聚类B
V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V32
聚类C
V17, V18, V19, V20, V21, V25, V29, V30, V31
聚类D
V22, V23, V24, V26, V27, V28

图10A

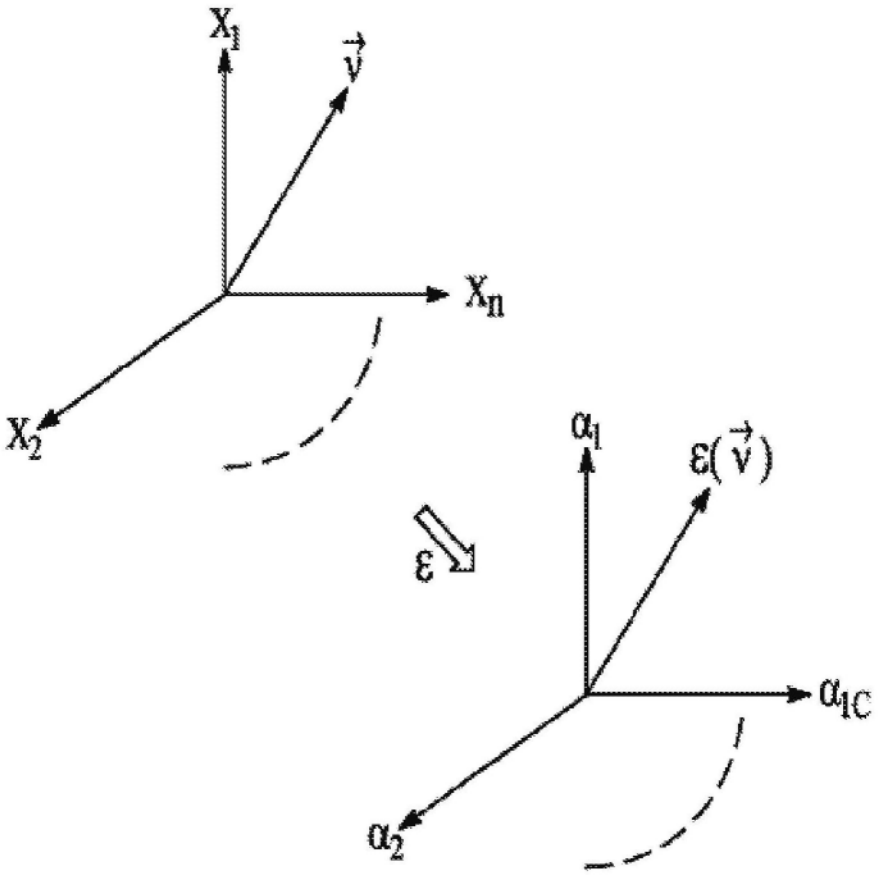


图10B

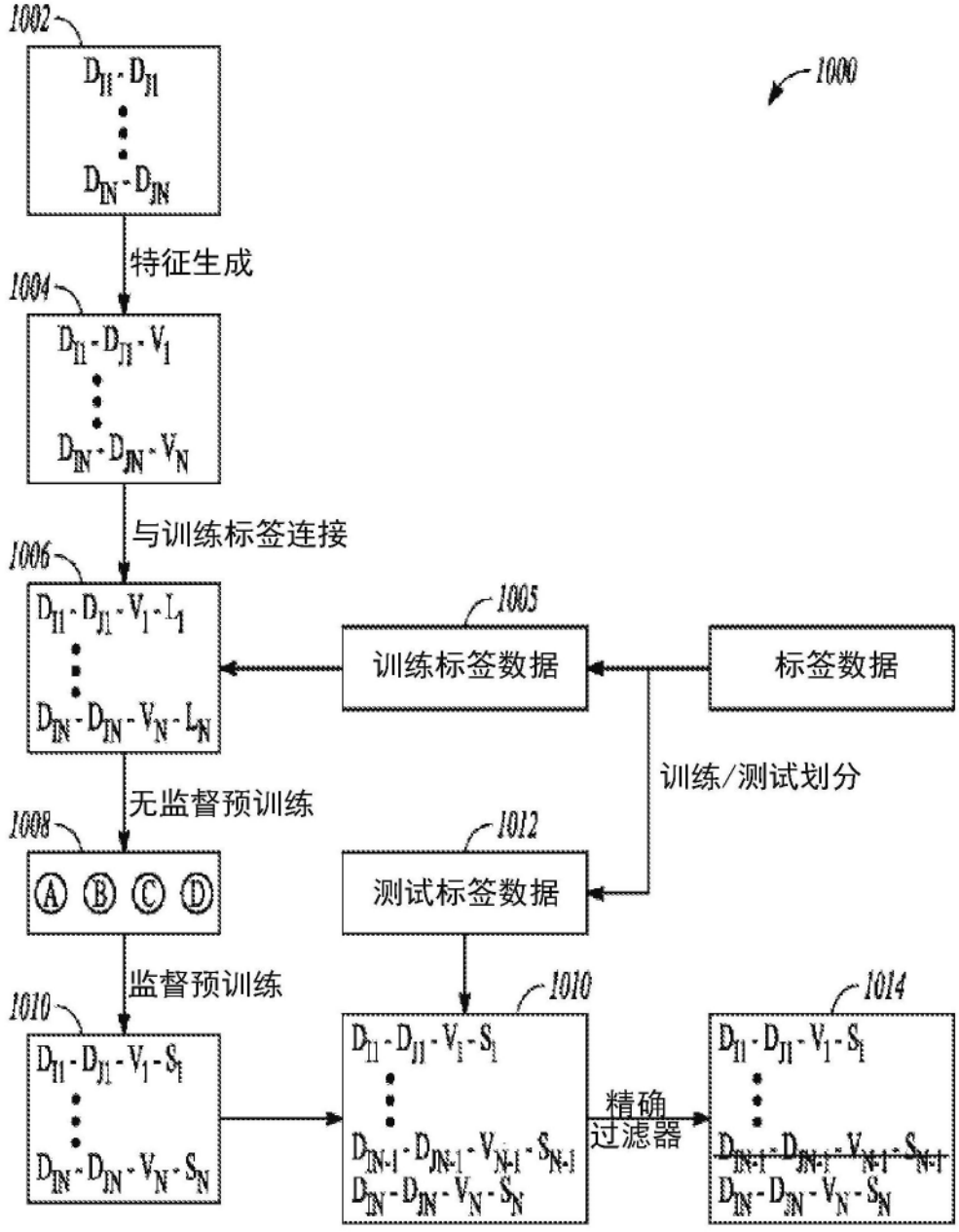

1002
1000
特征生成
1004
与训练标签连接
1006
1005
训练标签数据
标签数据
训练/测试划分
无监督预训练
1008
1012
测试标签数据
监督预训练
1010
1010
精确
过滤器
1014

图11

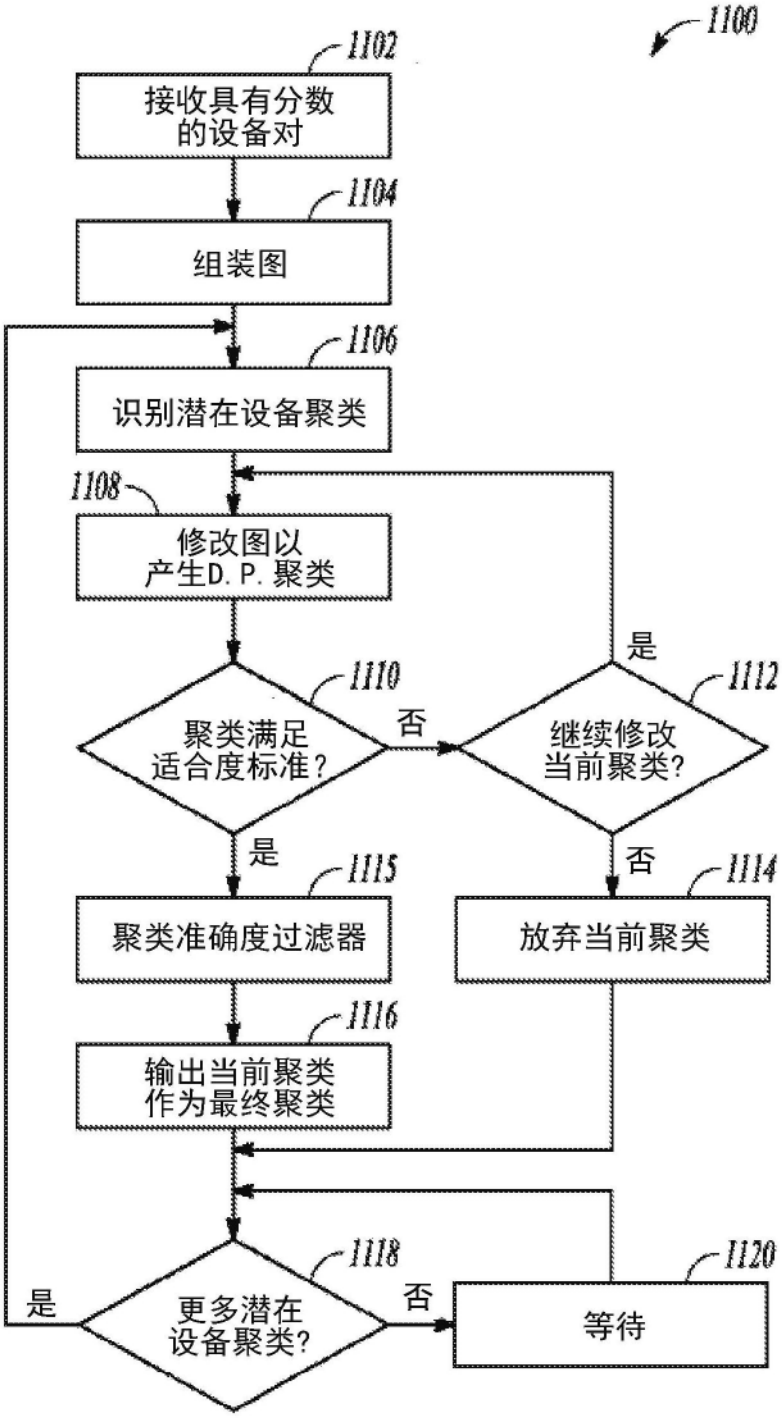

1100
1102
接收具有分数的设备对
1104
组装图
1106
识别潜在设备聚类
1108
修改图以产生D. P. 聚类
1110
聚类满足适合度标准?
否
1112
继续修改当前聚类?
是
是
否
1115
聚类准确度过滤器
1114
放弃当前聚类
1116
输出当前聚类作为最终聚类
1118
更多潜在设备聚类?
是
否
1120
等待

图12

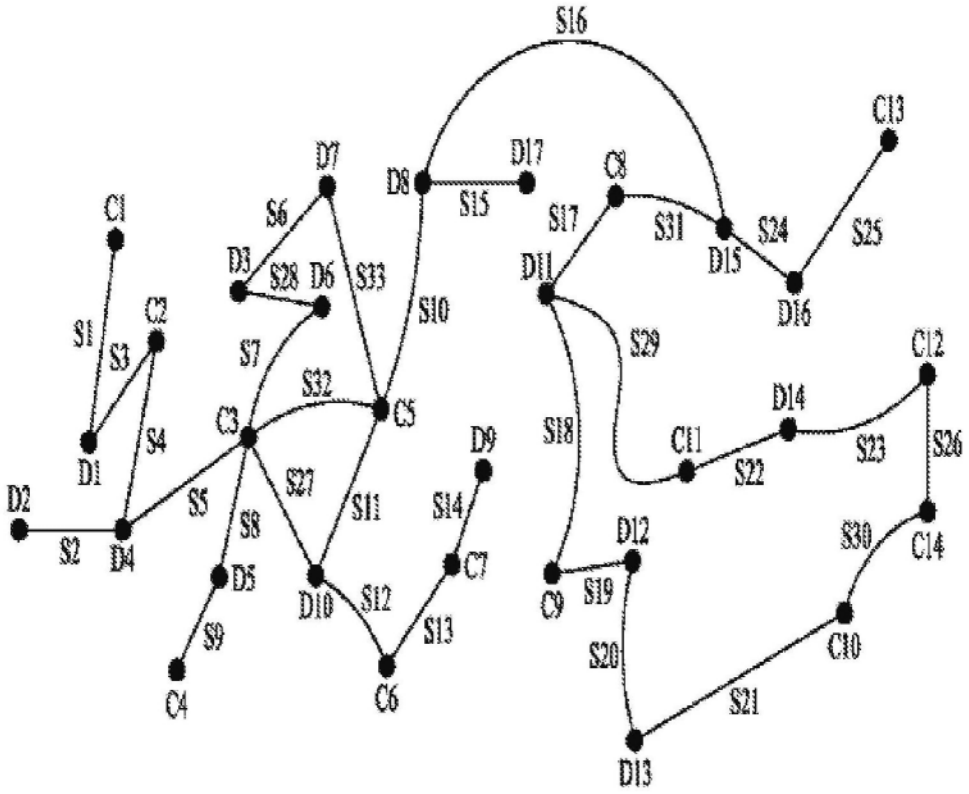

S16
C13
D7
D17
D8
C8
C1
S6
S15
S17
S31
S24
S25
D3
S28
D6
S33
D11
D15
D16
S10
C2
S1
S3
S7
S29
C12
S32
D14
C3
C5
S18
D9
C11
S23
S26
D1
S4
S22
D2
S5
S27
S11
S14
D12
S30
S8
S2
D4
C7
C14
D5
D10
S12
C9
S19
S9
S13
C10
S20
C4
C6
S21
D13

图13

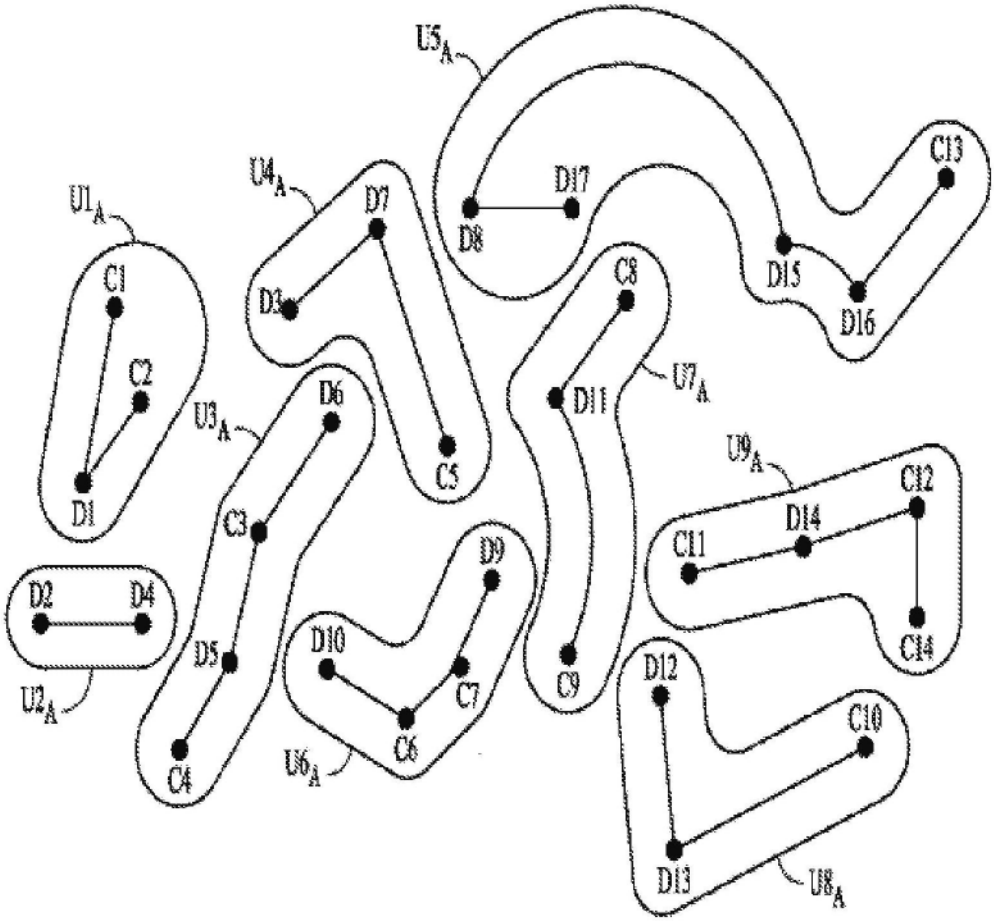

U5A
U4A
U1A
D17
D8
C13
D7
D15
C8
D3
D16
C1
C2
U7A
D6
U3A
D11
C5
U9A
C12
D14
C11
C3
D9
D2
D4
D10
C14
D5
C7
C9
D12
C10
C6
C4
U6A
D13
U8A
U2A

图14

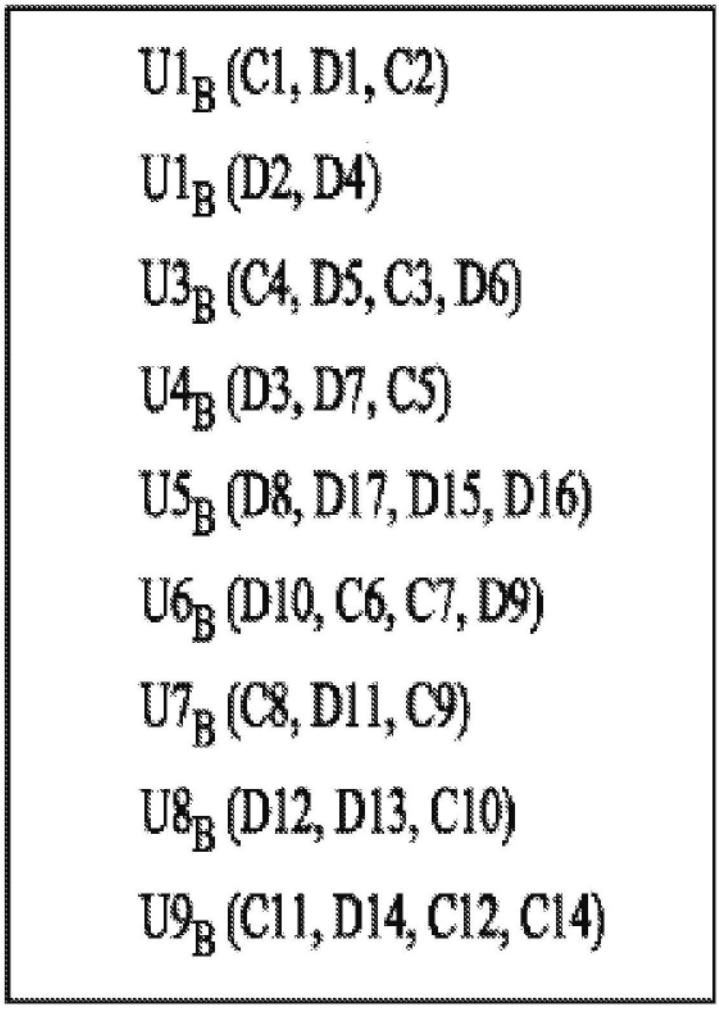

$U1_B$ (C1, D1, C2)
$U1_B$ (D2, D4)
$U3_B$ (C4, D5, C3, D6)
$U4_B$ (D3, D7, C5)
$U5_B$ (D8, D17, D15, D16)
$U6_B$ (D10, C6, C7, D9)
$U7_B$ (C8, D11, C9)
$U8_B$ (D12, D13, C10)
$U9_B$ (C11, D14, C12, C14)

图15

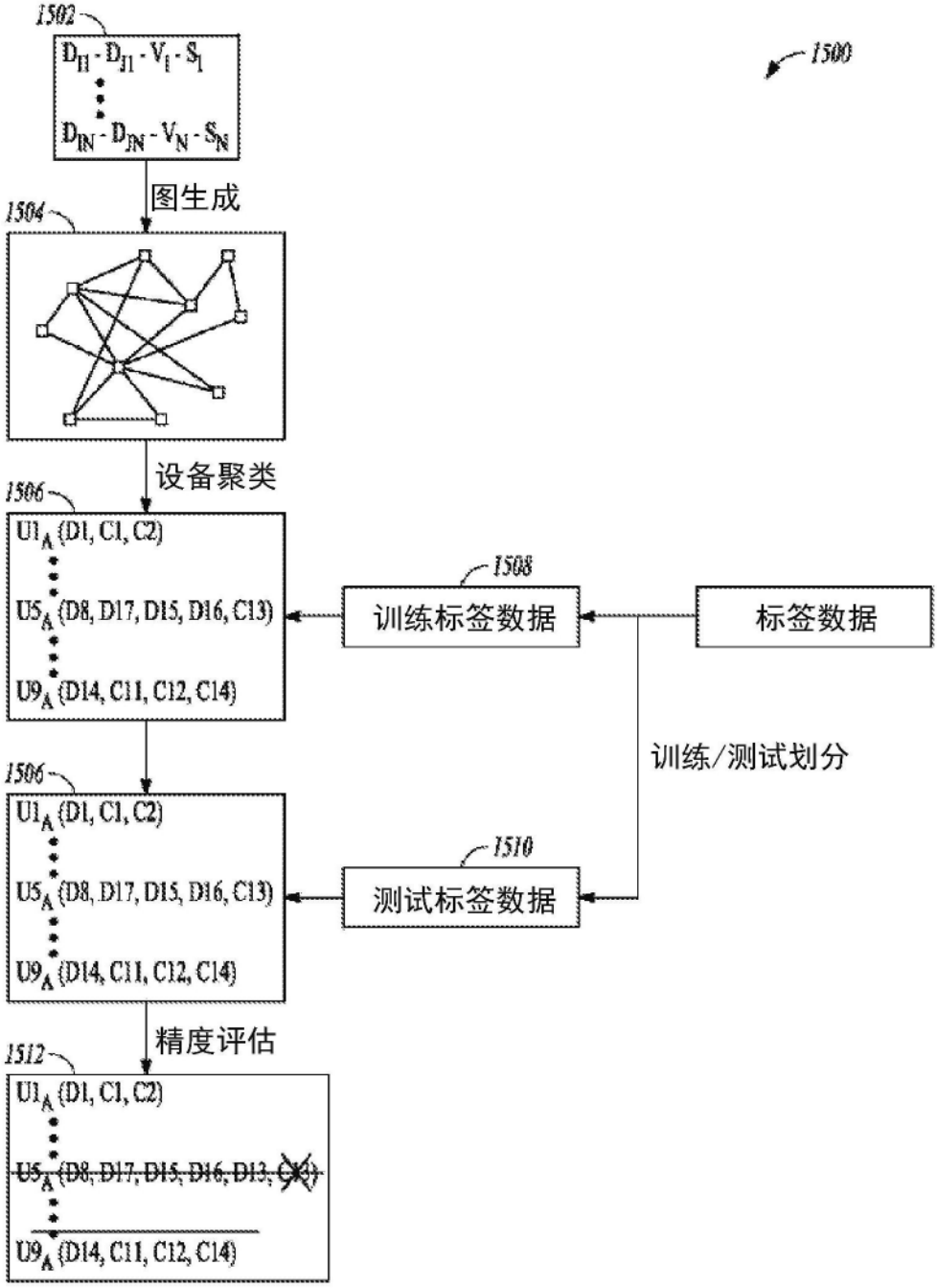

1502
D_{I1} - D_{J1} - V_1 - S_1
D_{IN} - D_{JN} - V_N - S_N
1500
图生成
1504
设备聚类
1506
$U1_A$ {D1, C1, C2)
$U5_A$ {D8, D17, D15, D16, C13)
$U9_A$ {D14, C11, C12, C14)
1508
训练标签数据
标签数据
训练/测试划分
1506
$U1_A$ {D1, C1, C2)
$U5_A$ {D8, D17, D15, D16, C13)
$U9_A$ {D14, C11, C12, C14)
1510
测试标签数据
精度评估
1512
$U1_A$ {D1, C1, C2)
$U5_A$ {D8, D17, D15, D16, D13, C13)
$U9_A$ {D14, C11, C12, C14)

图16

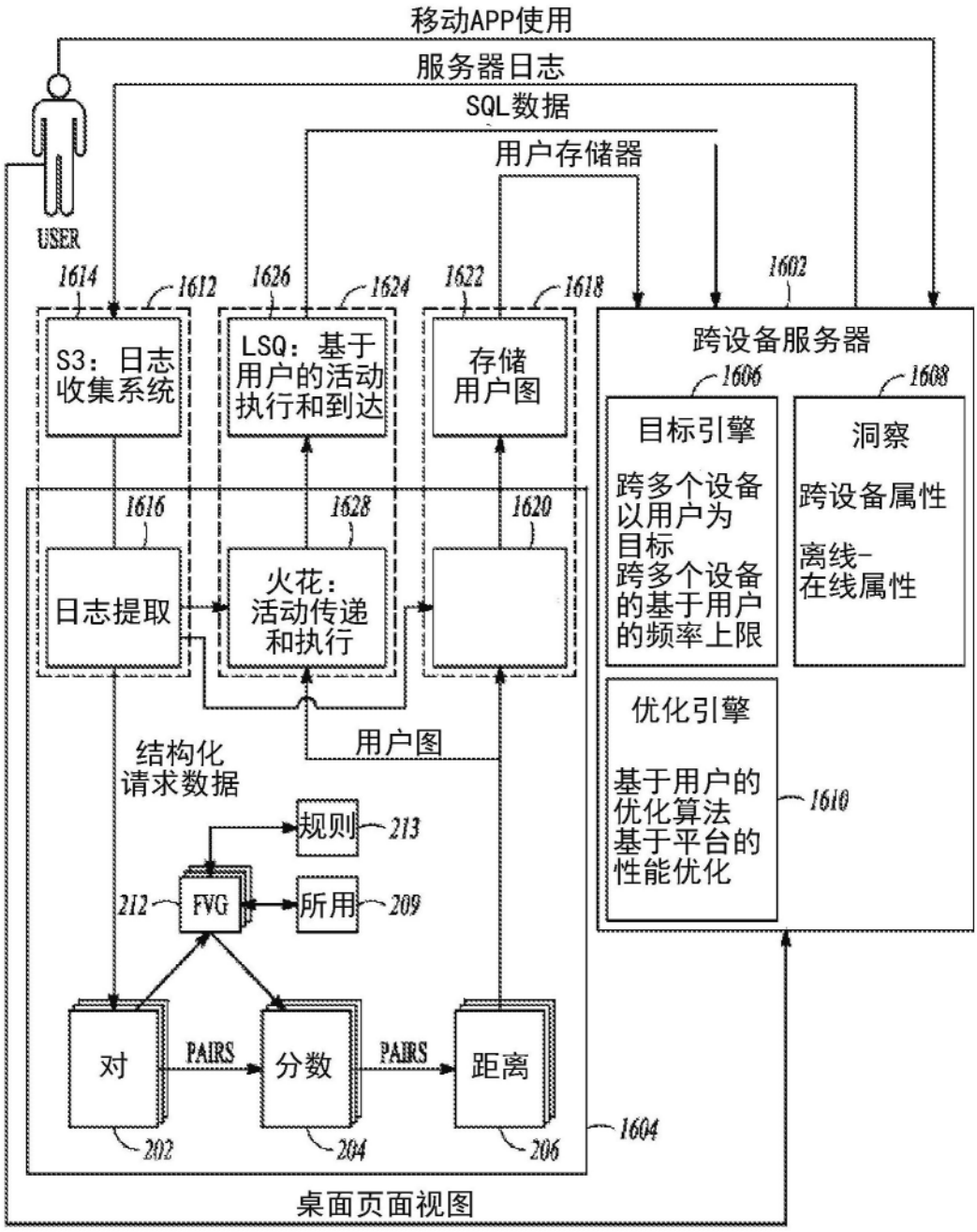

移动APP使用
服务器日志
SQL数据
用户存储器
USER
1614
1612
1626
1624
1622
1618
1602
S3：日志收集系统
LSQ：基于用户的活动执行和到达
存储用户图
跨设备服务器
1606
1608
目标引擎
跨多个设备以用户为目标
跨多个设备的基于用户的频率上限
洞察
跨设备属性
离线-在线属性
1616
1628
1620
日志提取
火花：活动传递和执行
优化引擎
基于用户的优化算法
基于平台的性能优化
1610
用户图
结构化请求数据
规则
213
212
FVG
所用
209
对
PAIRS
分数
PAIRS
距离
1604
202
204
206
桌面页面视图

图17