



(12)发明专利

(10)授权公告号 CN 104798052 B

(45)授权公告日 2018.11.27

(21)申请号 201380059197.X

(73)专利权人 达塔洛吉尔斯股份有限公司

(22)申请日 2013.10.04

地址 美国科罗拉多

(65)同一申请的已公布的文献号

(72)发明人 J·J·雷姆伯特
I·M·卡斯特里洛

申请公布号 CN 104798052 A

(43)申请公布日 2015.07.22

(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 11038
代理人 陈新

(30)优先权数据

13/644,736 2012.10.04 US

(85)PCT国际申请进入国家阶段日

2015.05.13

(51)Int.CI.

G06F 12/14(2006.01)

G06F 17/30(2006.01)

(86)PCT国际申请的申请数据

PCT/US2013/063470 2013.10.04

(56)对比文件

US 6061798 A, 2000.05.09, 全文.

(87)PCT国际申请的公布数据

W02014/055871 EN 2014.04.10

WO 2007051245 A1, 2007.05.10, 全文.

审查员 吴海旋

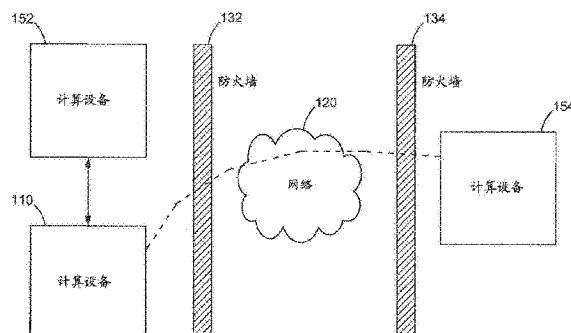
权利要求书4页 说明书15页 附图12页

(54)发明名称

消费者匹配的方法和装置

(57)摘要

在一些实施例中，方法包括从第一计算设备接收第一记录集和从第二计算设备接收第二记录集。来自第一记录集的每个记录都具有第一属性串；并且来自第二记录集的每个记录都具有第一属性串。所述方法包括定义第三记录集以包括来自第一记录集具有的第一属性串等于来自第二记录集的记录的第一属性串的每个记录。所述方法还包括对来自第一计算设备的第四记录集和来自第二计算设备的第五记录集重复以上的步骤，以进一步定义第三记录集。来自第四记录集的每个记录都具有第二属性串，并且来自第五记录集的每个记录都具有第二属性串。



1.一种进行数据记录匹配的方法,包括:

由第三计算设备从第一计算设备接收第一记录集,来自第一记录集的每个记录都包括:(1)加密后的第一标识串和(2)散列后的第一属性串;

由第三计算设备从第二计算设备接收第二记录集,来自第二记录集的每个记录都具有:(1)加密后的第二标识串和(2)散列后的第一属性串;

由第三计算设备从第一计算设备接收第四记录集,来自第四记录集的每个记录都包括:(1)加密后的第一标识串和(2)散列后的第二属性串;

由第三计算设备从第二计算设备接收第五记录集,来自第五记录集的每个记录都具有:(1)加密后的第二标识串和(2)散列后的第二属性串;以及

由第三计算设备通过以下操作中的一者或两者在不进行解密的情况下基于第一记录集、第二记录集、第三记录集和第四记录集生成匹配的记录集:

定义第三记录集以包括来自第一记录集的所具有的散列后的第一属性串等于来自第二记录集的记录的散列后的第一属性串的每个记录,来自第三记录集的每个记录都包括:(1)加密后的第一标识串和(2)加密后的第二标识串;以及

定义第六记录集以包括来自第四记录集的所具有的散列后的第二属性串等于来自第五记录集的记录的散列后的第二属性串的每个记录,来自第六记录集的每个记录都包括:(1)加密后的第一标识串和(2)加密后的第二标识串。

2.根据权利要求1所述的方法,进一步包括:

将第六记录集修改为修改后的状态,以排除来自第六记录集的所具有的加密后的第一标识串等于来自第三记录集的记录的加密后的第一标识串的每个记录;

将匹配的记录集定义为包括:(1)来自第三记录集的每个记录和(2)来自修改后的状态下的第六记录集的每个记录;以及

将指明匹配的记录集的信号发送到第二计算设备。

3.根据权利要求1所述的方法,进一步包括:

将第三记录集修改为修改后的状态,以排除来自第三记录集的所具有的加密后的第一标识串等于来自第六记录集的记录的加密后的第一标识串的每个记录;

将匹配的记录集定义为包括:(1)来自修改后的状态下的第三记录集的每个记录和(2)来自第六记录集的每个记录;以及

将指明匹配的记录集的信号发送到第二计算设备。

4.根据权利要求1所述的方法,其中,来自第一记录集的每个散列后的第一属性串基于电子邮件地址。

5.根据权利要求1所述的方法,进一步包括:

从第一计算设备接收第七记录集,来自第七记录集的每个记录都包括加密后的第一标识串和散列后的第三属性串;

从第二计算设备接收第八记录集,来自第八记录集的每个记录都具有加密后的第二标识串和散列后的第三属性串;以及

定义第九记录集以包括来自第七记录集的所具有的散列后的第三属性串等于来自第八记录集的记录的散列后的第三属性串的每个记录,来自第九记录集的每个记录都包括:(1)加密后的第一标识串和(2)加密后的第二标识串。

6. 根据权利要求5所述的方法,进一步包括:

将第六记录集修改为修改后的状态,以排除来自第六记录集的所具有的加密后的第一标识串等于来自第三记录集的记录的加密后的第一标识串的每个记录;

将第九记录集修改为修改后的状态,以排除:(1)来自第九记录集的所具有的加密后的第一标识串等于来自第三记录集的记录的加密后的第一标识串的每个记录以及(2)来自第九记录集的所具有的加密后的第一标识串等于来自第六记录集的记录的加密后的第一标识串的每个记录;

将匹配的记录集定义为包括:(1)来自第三记录集的每个记录、(2)来自修改后的状态下的第六记录集的每个记录和(3)来自修改后的状态下的第九记录集的每个记录;以及

将指明匹配的记录集的信号发送到第一计算设备。

7. 根据权利要求1所述的方法,其中,来自第四记录集的每个散列后的第二属性串基于邮政编码并且基于名字的一部分。

8. 一种进行数据记录匹配的装置,包括:

匹配模块,被配置为由第三计算设备从第一计算设备接收第一记录集,来自第一记录集的每个记录都包括:(1)加密后的第一标识串和(2)散列后的第一属性串;

匹配模块被配置为由第三计算设备从第二计算设备接收第二记录集,来自第二记录集的每个记录都具有:(1)加密后的第二标识串和(2)散列后的第一属性串;

匹配模块被配置为由第三计算设备从第一计算设备接收第四记录集,来自第四记录集的每个记录都包括:(1)加密后的第一标识串和(2)散列后的第二属性串;

匹配模块被配置为由第三计算设备从第二计算设备接收第五记录集,来自第五记录集的每个记录都具有:(1)加密后的第二标识串和(2)散列后的第二属性串;以及

匹配模块被配置为由第三计算设备通过以下操作中的一者或两者在不进行解密的情况下基于第一记录集、第二记录集、第三记录集和第四记录集生成匹配的记录集:

定义第三记录集以包括来自第一记录集的所具有的散列后的第一属性串等于来自第二记录集的记录的散列后的第一属性串的每个记录,来自第三记录集的每个记录都包括:(1)加密后的第一标识串和(2)加密后的第二标识串;以及

定义第六记录集以包括来自第四记录集的所具有的散列后的第二属性串等于来自第五记录集的记录的散列后的第二属性串的每个记录,来自第六记录集的每个记录都包括:(1)加密后的第一标识串和(2)加密后的第二标识串。

9. 根据权利要求8所述的装置,其中,匹配模块被配置为经由防火墙操作地耦接到第一计算设备。

10. 根据权利要求8所述的装置,其中,匹配模块被配置为:(1)被设置在第一防火墙后面,(2)操作地耦接到被设置在第一防火墙后面的第二计算设备,以及(3)经由第一防火墙操作地耦接到被设置在与第一防火墙不同的第二防火墙后面的第一计算设备。

11. 根据权利要求8所述的装置,其中,来自第一记录集的每个散列后的第一属性串基于电子邮件地址。

12. 根据权利要求8所述的装置,其中,修改后的状态下的匹配的记录集是在第一修改后的状态下的匹配的记录集,其中,

匹配模块被配置为将匹配的记录集修改为第二修改后的状态,以排除来自第一修改后

的状态下的匹配的记录集的每个这样的记录：(1) 所具有的加密后的第一标识串等于来自第一记录集的记录的加密后的第一标识串并且 (2) 与第三记录集和第四记录集相关联；

匹配模块被配置为将指明第二修改后的状态下的匹配的记录集的信号发送到第二计算设备。

13. 根据权利要求8所述的装置，其中，与来自第一记录集的每个散列后的第一属性串相关联的第一属性具有比与来自第三记录集的每个散列后的第二属性串相关联的第二属性更高的匹配率。

14. 一种存储着表示使处理器执行过程的指令的代码的非暂态处理器可读介质，表示使处理器执行过程的指令的所述代码包括执行以下操作的代码：

由第三计算设备从第一计算设备接收第一记录集，来自第一记录集的每个记录都包括：(1) 加密后的第一标识串和 (2) 散列后的第一属性串；

由第三计算设备从第二计算设备接收第二记录集，来自第二记录集的每个记录都具有：(1) 加密后的第二标识串和 (2) 散列后的第一属性串；

由第三计算设备从第一计算设备接收第四记录集，来自第四记录集的每个记录都包括：(1) 加密后的第一标识串和 (2) 散列后的第二属性串；

由第三计算设备从第二计算设备接收第五记录集，来自第五记录集的每个记录都具有：(1) 加密后的第二标识串和 (2) 散列后的第二属性串；以及

由第三计算设备通过以下操作中的一者或两者在不进行解密的情况下基于第一记录集、第二记录集、第三记录集和第四记录集生成匹配的记录集：

定义第三记录集以包括来自第一记录集的所具有的散列后的第一属性串等于来自第二记录集的记录的散列后的第一属性串的每个记录，来自第三记录集的每个记录都包括：(1) 加密后的第一标识串和 (2) 加密后的第二标识串；以及

定义第六记录集以包括来自第四记录集的所具有的散列后的第二属性串等于来自第五记录集的记录的散列后的第二属性串的每个记录，来自第六记录集的每个记录都包括：(1) 加密后的第一标识串和 (2) 加密后的第二标识串。

15. 根据权利要求14所述的存储着表示使处理器执行过程的指令的代码的非暂态处理器可读介质，其中：

来自第一记录集的每个记录都包括：(1) 由第一计算设备使用加密函数加密的第一标识串和 (2) 由第一计算设备使用散列函数散列的第一属性串；以及

来自第二记录集的每个记录都包括：(1) 由第二计算设备使用所述加密函数加密的第一标识串和 (2) 由第二计算设备使用所述散列函数散列的第一属性串。

16. 根据权利要求14所述的存储着表示使处理器执行过程的指令的代码的非暂态处理器可读介质，表示使处理器执行过程的指令的所述代码进一步包括执行以下操作的代码：

将第六记录集修改为修改后的状态，以排除来自第六记录集的所具有的第一标识串等效于来自第三记录集的记录的第一标识串的每个记录；

将匹配的记录集定义为包括：(1) 来自第三记录集的每个记录和 (2) 来自修改后的状态下的第六记录集的每个记录；以及

将指明匹配的记录集的信号发送到第二计算设备。

17. 根据权利要求14所述的存储着表示使处理器执行过程的指令的代码的非暂态处理

器可读介质,表示使处理器执行过程的指令的所述代码进一步包括执行以下操作的代码:

从第一计算设备接收第七记录集,来自第七记录集的每个记录都包括第一标识串和第三属性串;

从第二计算设备接收第八记录集,来自第八记录集的每个记录都具有第二标识串和第三属性串;以及

定义第九记录集以包括来自第七记录集的所具有的第三属性串等效于来自第八记录集的记录的第三属性串的每个记录,来自第九记录集的每个记录都包括:(1)第一标识串和(2)第二标识串。

18. 根据权利要求17所述的存储着表示使处理器执行过程的指令的代码的非暂态处理器可读介质,表示使处理器执行过程的指令的所述代码进一步包括执行以下操作的代码:

将第六记录集修改为修改后的状态,以排除来自第六记录集的所具有的第一标识串等于来自第三记录集的记录的第一标识串的每个记录;

将第九记录集修改为修改后的状态,以排除:(1)来自第九记录集的所具有的第一标识串等于来自第三记录集的记录的第一标识串的每个记录以及(2)来自第九记录集的所具有的第一标识串等于来自第六记录集的记录的第一标识串的每个记录;

将匹配的记录集定义为包括:(1)来自第三记录集的每个记录、(2)来自修改后的状态下的第六记录集的每个记录和(3)来自修改后的状态下的第九记录集的每个记录;以及

将指明匹配的记录集的信号发送到第一计算设备。

19. 根据权利要求14所述的存储着表示使处理器执行过程的指令的代码的非暂态处理器可读介质,其中,来自第一记录集的每个第一属性串基于姓并且基于邮政编码。

20. 根据权利要求14所述的存储着表示使处理器执行过程的指令的代码的非暂态处理器可读介质,其中,第一记录集包括第一已知记录,第二记录集包括第一已知记录,第四记录集包括第二已知记录,第五记录集包括第二已知记录,表示使处理器执行过程的指令的所述代码进一步包括确认第三记录集包括第一已知记录并且第六记录集包括第二已知记录的代码。

消费者匹配的方法和装置

[0001] 相关申请

[0002] 本发明要求2012年10月4日提交的标题为“Method and Apparatus for Matching Consumers”的13/644,736号美国申请的优先权并且是其继续,其全部内容通过引用并入于此。

技术领域

[0003] 本文描述的一些实施例一般地涉及消费者匹配的方法和装置。

背景技术

[0004] 营销合作伙伴——例如营销实体、网站、线上和线下店铺以及数据分析实体——可以共享信息,以便计划、执行和测定营销和其他工作。用于信息共享、尤其是消费者信息共享和匹配的系统和方法已存在,但是往往可能触犯隐私策略和/或法律。不仅如此,这样的系统可能不会有效地将来自一个实体的信息与来自其他实体的信息匹配,因为它们可能使用有限的信息,这会导致错过存在的匹配。

[0005] 所以,存在着对改进的消费者匹配的方法和装置的需求。

发明内容

[0006] 在一些实施例中,方法包括从第一计算设备接收第一记录集。来自第一记录集的每个记录都包括加密后的第一标识串和散列后(hashed)的第一属性串。所述方法包括从第二计算设备接收第二记录集。来自第二记录集的每个记录都具有加密后的第二标识串和散列后的第一属性串。所述方法进一步包括定义第三记录集以包括来自第一记录集的所具有的散列后第一属性串等于来自第二记录集的记录的散列后第一属性串的每个记录。来自第三记录集的每个记录都包括加密后的第一标识串和加密后的第二标识串。所述方法还包括对来自第一计算设备的第四记录集和来自第二计算设备的第五记录集重复以上的步骤,以进一步定义第三记录集。来自第四记录集的每个记录都具有加密后的第一标识串和散列后的第二属性串,并且来自第五记录集的每个记录都具有加密后的第二标识串和散列后的第二属性串。

附图说明

[0007] 图1是根据实施例的被配置为匹配记录集的多台计算设备的示意性说明。

[0008] 图2是根据实施例的计算设备的框图。

[0009] 图3说明根据实施例的匹配记录集的方法的流程图。

[0010] 图4A至图4L说明根据实施例的匹配记录集的过程。

具体实施方式

[0011] 在一些实施例中,方法包括从第一计算设备接收第一记录集。来自第一记录集的

每个记录都包括加密后的第一标识串和散列后的第一属性串。所述方法包括从第二计算设备接收第二记录集。来自第二记录集的每个记录都包括加密后的第二标识串和散列后的第一属性串。在一些实施例中，散列后的第一属性串可以基于例如电子邮件地址、邮政编码、名字的部分等。基于第一记录集和第二记录集，第三记录集被定义为包括来自第一记录集的所具有的散列后第一属性串等于来自第二记录集的记录的散列后第一属性串的每个记录。来自第三记录集的每个记录都包括加密后的第一标识串和加密后的第二标识串。

[0012] 类似地，所述方法包括从第一计算设备接收第四记录集。来自第四记录集的每个记录都包括加密后的第一标识串和散列后的第二属性串。所述方法包括从第二计算设备接收第五记录集。来自第五记录集的每个记录都包括加密后的第二标识串和散列后的第二属性串。所述方法进一步包括定义第六记录集以包括来自第四记录集的所具有的散列后第二属性串等于来自第五记录集的记录的散列后第二属性串的每个记录。来自第六记录集的每个记录都包括加密后的第一标识串和加密后的第二标识串。

[0013] 在一些实施例中，所述方法可以进一步包括把第六记录集修改为修改后的状态，以排除来自第六记录集的所具有的加密后第一标识串等于来自第三记录集的记录的加密后第一标识串的每个记录。可以把匹配的记录集定义为包括来自第三记录集的每个记录和来自在修改后的状态下的第六记录集的每个记录。不仅如此，可以把指明匹配的记录集的信号发送到第一计算设备和/或第二计算设备。

[0014] 正如本文所用，模块可以是例如操作地耦接的电气组件的任何组装和/或集合，并且可以包括例如存储器、处理器、电气迹线、光学连接器、软件（在硬件中执行）等。正如本文所用，单数形式的“一”、“一个”和“该”包括复数引用，除非上下文清楚地表明并非如此。因此，例如，术语“一记录数据库”意在意味着单个数据库或具有类似功能的数据库的集合。不仅如此，正如本文所描述的，实体（例如与计算设备相关联的商业实体）可以是营销实体、网站和/或网站操作员、线上和/或线下店铺、数据分析实体等。

[0015] 图1是根据实施例的被配置为匹配记录集的多台计算设备110、152和154的示意性说明。正如图1所示，计算设备110被直接地或操作地耦接到计算设备152。计算设备110还经由至少第一防火墙132、网络120和第二防火墙134被操作地耦接到计算设备154。正如以下所描述的，计算设备110、152和154可以被配置为协同地执行使来自多个记录集的记录匹配的过程。

[0016] 防火墙132或134可以是任何基于软件的模块和/或基于硬件的设备，其被用于控制和过滤进入的和/或外出的网络流量。防火墙132或134可以用于使内部网络与外部网络分开，从而保持内部网络免于外部网络的危险。在图1所示的实例中，防火墙132把包括计算设备152和计算设备110的内部网络与外部网络120分开；防火墙134把包括计算设备154的内部网络与外部网络120分开。在一些实施例中，防火墙132或134可以是例如网络层防火墙（例如数据包过滤防火墙）、电路级防火墙、应用层防火墙、代理服务器等。

[0017] 网络120可以是对防火墙132后面的内部网络和防火墙134后面的内部网络为外部并且连接着这两个内部网络（通过防火墙132和134）的任何类型的网络。网络120可以是有线网络、无线网络或者有线/无线网络的组合。在一些实施例中，网络120可以是例如局域网（LAN）、广域网（WAN）、无线LAN（WLAN）、因特网等。

[0018] 计算设备（例如计算设备110、152或154）可以是被配置为产生、存储、操纵一个或

多个记录集和/或对这一个或多个记录集执行任何其他操作的任何设备。这样的计算设备可以是例如服务器、工作站、数据中心、数据处理计算机或任何其他类型的计算设备或者计算设备的组合。

[0019] 在一些实施例中,不同的计算设备可以被配置为执行不同的功能。在图1的实例中,计算设备152和154可以被配置为产生数据文件,这些数据文件包含着要与其他原始记录集匹配的原始记录集。具体来说,计算设备152和154可以被配置为例如访问和检索外部资源(如存储设备)的数据、基于检索到的数据定义一个或多个记录集、把所定义的记录集适当地联系在一起以产生一个或多个数据文件、在存储器中存储数据文件、把包括记录集的数据文件发送到其他设备(如计算设备110)等。另一方面,计算设备110可以被配置为使来自多个原始记录集的记录匹配以定义一个或多个匹配的记录集。具体来说,计算设备110可以被配置为例如(如从计算设备152和154)接收包含若干原始记录集的数据文件、比较和匹配来自多个原始记录集的记录以定义匹配的记录集、把匹配的记录集存储在存储器中、把匹配的记录集发送到其他设备(如计算设备152或154)等。

[0020] 在计算设备152、154或110处所定义和/或处理的记录集可以是在数据文件中存储数据的任何类型的数据结构。原始记录集可以在计算设备152或计算设备154处被定义,并且在计算设备110处被进一步处理(如与其他原始记录集比较以定义匹配的记录集)。记录集(如原始记录集、匹配的记录集)可以是例如数组、列表、表格、队列、树、图、图形或任何其他适合类型的数据结构。在记录集中存储的数据可以例如与服务的用户、公司的客户、在线论坛的注册浏览者、产品的购买者等相关联。这样的数据可以包括例如购买者的电子邮件地址、在线浏览者的登录身份、用户名(如姓、名)、公司的地址(如邮政编码、街道地址)等。

[0021] 在一些实施例中,计算设备152和110可以与第一实体相关联,第一实体诸如例如公司、组织、个体等。在这样的实施例中,计算设备152和110可以被包括在与第一实体相关联并通过防火墙132与外部网络120(如因特网)分开的第一内部网络(图1未显示)中。计算设备152和计算设备110可以在第一内部网络内彼此直接地耦接或共同位于单个设备(如数据中心)上。另选地,计算设备152和计算设备110可以是分开的设备,它们经由一台或多台其他设备(诸如例如服务设备、路由设备、另一台计算设备等)彼此操作地耦接。在一些实施例中,计算设备152和计算设备110可以是一台计算设备,在这样的实施例中,这样的计算设备可以定义原始记录集,并且还基于在该计算设备处定义的和/或从其他计算设备(如计算设备154)接收到的多个原始记录集,来定义匹配的记录集。另外,在计算设备152处定义的记录集可以基于与第一实体相关联的数据。例如,第一实体可以是线上公司,并且在计算设备152处定义的记录集可以包括与该公司网站的注册用户相关联的数据。

[0022] 类似地,计算设备154可以与和第一实体不同的第二实体(如公司、组织、个体等)相关联。因此,计算设备154可以被包括在第二内部网络(图1未显示)中,它与第二实体相关联并通过防火墙134与外部网络120(如因特网)分开。在这样的实施例中,正如图1所示,计算设备154可以经由至少防火墙134、网络120和防火墙134操作地耦接到计算设备110。另外,在计算设备154处定义的记录集可以基于与第二实体相关联的数据。例如,第二实体可以是数据分析实体,在计算设备154处定义的记录集可以包括与一个或多个零售连锁店的顾客相关联的数据。

[0023] 在一些实施例中,尽管图1未显示,但是第一内部网络和/或第二内部网络可以通过多于一个防火墙或不通过防火墙而与外部网络120分离。在这样的实施例中,计算设备110和计算设备154可以经由任何数量的防火墙彼此耦接(直接地或操作地),或者不通过任何防火墙被分离。不仅如此,在一些实施例中,计算设备110可以对于第一内部网络和第二内部网络都在外部。例如,计算设备110可以与第三实体相关联,它与第一实体和第二实体不同。结果,计算设备110可以被包括在防火墙132与防火墙134之间的网络120中。在这样的实施例中,计算设备110可以经由防火墙132操作地耦接到计算设备154,并且经由防火墙134操作地耦接到计算设备152。

[0024] 在一些实施例中,尽管图1未显示,但是计算设备110可以经由防火墙132、网络120和/或第三防火墙操作地耦接到第四计算设备。第四计算设备可以与第三实体相关联,它与第一实体和第二实体不同。因此,计算设备110可以从计算设备152、计算设备154和第四计算设备接收记录集,并且执行过程来以与本文关于图3所描述的类似方式匹配来自接收到的记录集的记录。

[0025] 图2是根据实施例的计算设备200的框图。计算设备200可以是被配置为匹配从其他计算设备接收到的记录集的计算设备,类似于图1的计算设备110。正如图2所示,计算设备200包括通信接口230;存储器210,它包含记录数据库212(如用于一个或多个数据文件、关系数据库等的存储设备);以及处理器250,它包含匹配模块254。通信接口230的操作(如传送/接收数据文件)和匹配模块254的操作(如比较记录集、产生匹配的记录集)以及对记录数据库212的操纵(如存储数据文件、删除数据文件)或者对存储器210的任何其他部分的操纵,都可以由处理器250控制。

[0026] 在一些实施例中,计算设备200的通信接口230可以与计算设备200的一个或多个端口(用于有线连接,图2未显示)和/或天线(用于无线连接,图2未显示)相关联。通信接口230以及相关的端口和/或天线可以用于实现计算设备200与其他计算设备(如图1的计算设备152、154)或其他设备(如显示器设备、存储设备)之间的一个或多个有线和/或无线连接。在这些连接当中,有线连接可以是例如经由电缆的双绞线电信号、经由光缆的光纤信号等;而无线连接可以基于任何适合的无线通信协议(如蓝牙协议、Wi-Fi协议等)。因此,计算设备200可以被配置为通过与通信接口230相关联的一个或多个端口和/或天线,从其他计算设备(如图1的计算设备152、154)和/或其他设备接收数据(如包含记录集的数据文件、软件更新和/或诊断工具等)以及/或者向这些其他计算设备和/或其他设备发送数据。在一些实施例中,通信接口230可以允许由例如计算设备152、154对处理器250的远程访问,以便允许软件更新和/或诊断活动。具体来说,在一些实施例中,一个或多个防火墙(如图1的防火墙132或134)可以在通信接口230处实现,使得通过通信接口230传输的数据可以被适当地过滤。

[0027] 处理器250可以是任何适合的处理器,被配置为运行和/或执行处理器250中所包括的模块。处理器250中的每个模块都可以是以下各项的任何组合:基于硬件的模块(如现场可编程门阵列(FPGA)、专用集成电路(ASIC)、数字信号处理器(DSP))以及/或者基于软件的模块(如在存储器中所存储的和/或在处理器250处所执行的计算机代码的模块),它们能够执行与该模块相关联的一个或多个特定功能。具体来说,匹配模块254可以被配置为执行记录集的匹配过程,正如关于图3和图4A至图4L所详细描述的。此外,在一些实施例中,处理

器250可以包括其他模块(图2未显示),其被配置为执行计算设备200的其他功能。例如,处理器250可以包括被配置为从其他计算设备(如图1的计算设备152、154)检索原始记录集并且把匹配的记录集传送到其他计算设备的模块。对于另一个实例,处理器250可以包括被配置为在显示器设备上显示记录集的模块,其中显示器设备嵌入在计算设备200内或耦接到计算设备200。

[0028] 在一些实施例中,存储器210可以是例如随机存取存储器(RAM)(如动态RAM、静态RAM)、闪存、可移动存储器等等。与执行记录集的匹配过程相关联的数据和信息可以在存储器210中存储、保持和更新。具体来说,记录集(包括要被匹配的原始记录集和作为执行匹配过程的结果的匹配后的记录集)可以在存储器210内的记录数据库212中存储和更新。另外,尽管图2中未显示,与执行记录集的匹配过程相关联的其他数据和信息也可以存储在存储器210的其他部分中。例如,与执行匹配过程相关联的指令可以存储(如作为指令集)在存储器210内的非暂态过程可读(process-readable)介质中。

[0029] 图3说明根据实施例的匹配记录集的方法300的流程图。方法300可以在结构上和功能上类似于关于图1和图2所示出和描述的计算设备110和计算设备200的计算设备处执行。具体来说,与执行方法300相关联的指令可以存储在计算设备的存储器(如图2中计算设备200的存储器210)中,并且在计算设备的处理器中的匹配模块(如图2中计算设备200的处理器250中的匹配模块254)处执行。执行方法300的实例关于图4A至图4L进行详细说明。

[0030] 在302,匹配模块可以被配置为从第一计算设备接收第一记录集。例如,第一记录集可以被包括在从第一计算设备向宿有(host)匹配模块的计算设备所发送的数据文件中。第一计算设备可以在结构上和功能上类似于关于图1所示出和描述的计算设备152和154。在一些实施例中,第一记录集可以在第一计算设备处定义并存储。在一些其他实施例中,第一记录集可以在另一设备处定义并接着存储在第一计算设备中。

[0031] 第一记录集可以基于与例如拥有或控制第一计算设备的第一实体(如公司、组织、个体)相关联的数据被定义(如在第一计算设备处)。不仅如此,来自第一记录集的每个记录都可以包括与第一实体相关联的至少第一标识串和第一属性串。

[0032] 每个第一标识串都可以是例如文本串(如“A用户”)、数字(如“999”)、代码(如“101a”)、符号(如“#”)、上述各项的组合以及/或者可以包括在记录中的任何其他适合的格式。在一些实施例中,每个第一标识串都可以用于唯一地标识例如第一实体的用户。在其他实施例中,多于一个的第一标识串可以与例如第一实体的公共用户相关联。

[0033] 第一记录集中所包括的每个第一属性串都可以是表示与例如第一实体的用户相关联的第一属性的数据。该数据可以是例如文本串(如“Lionel Messi”)、数字(如“21000”)或者适于表示第一属性的任何其他格式。第一属性可以是例如用户的电子邮件地址、用户名或用户名的一部分(如姓、名)、递送点(如与递送点条形码和/或智能邮寄条形码相关联的递送点)、完整地址或地址的部分(如邮政编码、城市名)、各种项(如邮政编码和姓)的组合等。因此,来自第一记录集的记录中的每对第一标识串和第一属性串与例如第一实体的用户相关联。

[0034] 在304,匹配模块可以被配置为从与第一计算设备不同的第二计算设备接收第二记录集。例如,第二记录集可以被包括在从第二计算设备向宿有匹配模块的计算设备所发送的数据文件中。类似于第一计算设备,第二计算设备可以在结构上和功能上类似于关于

图1所示出和描述的计算设备152和154。在一些实施例中，第二记录集可以在第二计算设备处定义并存储。在一些其他实施例中，第二记录集可以在另一设备处定义并接着存储在第二计算设备中。

[0035] 类似于第一记录集，第二记录集可以基于与例如拥有或控制第二计算设备的第二实体(如公司、组织、个体)相关联的数据被定义(如在第二计算设备处)。第二实体可以与第一实体不同。不仅如此，来自第二记录集的每个记录都可以包括与第二实体相关联的至少第二标识串和第一属性串。

[0036] 在第二记录集中所包括的第二标识串可以在结构上类似于在第一记录集中所包括的第一标识串。在一些实施例中，每个第二标识串都可以用于唯一地标识例如第二实体的用户。在其他实施例中，多于一个的第二标识串可以与例如第二实体的公共用户相关联。

[0037] 类似于在第一记录集中所包括的第一属性串，在第二记录集中所包括的每个第一属性串都可以是表示与例如第二实体的用户相关联的第一属性的数据。与第二实体的用户相关联的第一属性同与第一实体的用户相关联的第一属性相同。这样的第一属性可以例如是第一实体的用户或第二实体的用户的电子邮件地址、第一实体的用户或第二实体的用户的邮政编码与姓的组合等等。来自第二记录集的记录中的每对第二标识串和第一属性串与例如第二实体的用户相关联。不仅如此，来自第一记录集的记录中的第一属性串可以与来自第二记录集的记录中的第一属性串一致。

[0038] 在一些实施例中，匹配模块可以被配置为经由一个或多个防火墙(如图1的防火墙152、154)从第一计算设备接收第一记录集，以及从第二计算设备接收第二记录集。在这样的实施例中，在匹配模块处接收到的第一记录集和第二记录集可以被加密。例如，第一记录集中的每个第一标识串可以是加密的值，其是作为使用第一加密密钥加密原始第一标识串的结果；并且第二记录集中的每个第二标识串可以是加密的值，其是作为使用第二加密密钥加密原始第二标识串的结果。第一加密密钥可以与第二加密密钥不同。对于另一个实例，第一记录集中的每个第一属性串可以是散列值，其是作为对与第一实体用户相关联的第一属性串的原始数据执行散列函数(hash function)的结果；并且第二记录集中的每个第一属性串可以是散列值，其是作为对与第二实体用户相关联的第一属性串的原始数据执行同一散列函数的结果。具体来说，作为对第一记录集和第二记录集的第一属性串应用同一散列函数的结果，当且仅当它们第一属性串的对应原始数据一致时，第一记录集中的散列后的第一属性串与第二记录集中散列后的第一属性串一致。这样的第一记录集和第二记录集的实例关于图4A至图4L来说明。

[0039] 在一些实施例中，在匹配模块接收到第一记录集和第二记录集之前可以对它们实施多于一个等级的加密。在一些实施例中，加密密钥或散列函数在宿有匹配模块的计算设备处不可用。结果，第一记录集中的对应数据和第二记录集中的对应数据无法在计算设备处解密或恢复。因此，在这样的实施例中，与第一实体的用户或第二实体的用户相关联的原始数据(如原始第一标识串、原始第二标识串、第一属性串的原始数据)在宿有匹配模块的计算设备处不可用。

[0040] 在一些实施例中，与记录集(如第一记录集和第二记录集)相关联的附加信息可以连同该记录集从计算设备(如第一计算设备、第二计算设备)发送到匹配模块。这样的附加信息可以包括例如与记录集相关联的属性(如第一属性)的优先级。例如，“电子邮件地址”

作为属性具有1的优先级(即最高优先级);“邮政编码与姓的组合”作为属性具有2的优先级(即第二高优先级);而“邮政编码”作为属性具有3的优先级(即第三高优先级(或最低优先级))。具体来说,在方法300的实例中,第一记录集的优先级与第二记录集的优先级相同,因为与第一记录集相关联的属性(即第一属性)和与第二记录集相关联的属性相同。

[0041] 在一些实施例中,附加信息(诸如优先级)可以与记录集分离地被发送到匹配模块。在其他实施例中,这样的附加信息可以包括在记录集中(例如,连同其他两项:标识串和属性串,作为每个记录中的第三项),从而在记录集被发送给匹配模块时被发送到匹配模块。

[0042] 在306,匹配模块可以被配置为基于接收到的第一记录集和第二记录集定义第三记录集,使得第三记录集包括来自第一记录集的具有的第一属性串等于来自第二记录集的记录的第一属性串的每个记录。对于来自第一记录集的每个这样的记录,第三记录集包括以下记录,其包含来自第一记录集的该记录的第一标识串以及来自第二记录集的对应记录(即来自第二记录集的记录具有的第一属性串等于来自第一记录集的该记录的第一属性串)的第二标识串。

[0043] 匹配模块可以被配置为比较第一记录集和第二记录集以便用各种方法定义第三记录集。在一些实施例中,例如,匹配模块可以被配置为把来自第一记录集的每个记录中的第一属性串与来自第二记录集的每个记录中的第一属性串进行比较。如果这两个第一属性串相等,则匹配模块可以被配置为检索来自第一记录集的记录的第一标识串和检索来自第二记录集的记录的第二标识串,并接着定义第三记录集中的新记录以包括检索到的第一标识串和检索到的第二标识串。因此,在对每对来自第一记录集的记录和来自第二记录集的记录执行了这样的方法之后定义第三记录集。在其他实施例中,可以以任何其他适合的方法定义第三记录集。

[0044] 在一些实施例中,正如以上所描述的,来自第一记录集和第二记录集的记录中所包括的全部数据(如标识串、属性串)都是加密后的数据(如由加密密钥加密的、由散列函数散列的)。在这样的实施例中,在匹配模块处所执行的操作(如比较、匹配)是对加密数据执行的。结果,来自第三记录集的记录中所包括的数据(如标识串)也是加密后的数据。

[0045] 在308,类似于步骤302,匹配模块可以被配置为从第一计算设备接收第四记录集。第四记录集可以基于与第一实体相关联的数据定义。来自第四记录集的每个记录可以包括与第一实体相关联的至少第一标识串和第二属性串。第四记录集中所包括的记录数量可以与第一记录集中所包括的记录数量不同。第四记录集中所包括的第一标识串的部分可以与第一记录集中所包括的第一标识串的部分一致;而第四记录集中所包括的第二属性串与第一记录集中所包括的第一属性串不同,因为第二属性与第一属性不同。另外,第二属性具有的优先级与第一属性的优先级不同。例如,第一属性可以是“电子邮件地址”,它具有1的优先级;而第二属性可以是“邮政编码和姓的组合”,它具有2的优先级。

[0046] 在310,类似于步骤304,匹配模块可以被配置为从第二计算设备接收第五记录集。第五记录集可以基于与第二实体相关联的数据定义。来自第五记录集的每个记录可以包括与第二实体相关联的至少第二标识串和第二属性串。第五记录集中所包括的记录数量可以与第二记录集中所包括的记录数量不同。第五记录集中所包括的第二标识串的部分可以与第二记录集中所包括的第二标识串的部分一致;而第五记录集中所包括的第二属性串与第

二记录集中所包括的第一属性串不同。另外,在一些实施例中,第四记录集和第五记录集中所包括的数据可以被加密,类似于第一记录集和第二记录集中所包括的数据。

[0047] 在312,类似于步骤306,匹配模块可以被配置为基于接收到的第四记录集和第五记录集定义第六记录集,使得第六记录集包括来自第四记录集的具有的第二属性串等于来自第五记录集的记录的第二属性串的每个记录。对于第四记录集的每个这样的记录,第六记录集包括以下记录,其包含来自第四记录集的该记录的第一标识串以及来自第五记录集的对应记录(即来自第五记录集的记录具有的第二属性串等于来自第四记录集的该记录的第二属性串)的第二标识串。不仅如此,类似于第三记录集,来自第六记录集的记录中所包括的数据(如标识串)可以是加密后的数据。

[0048] 在一些实施例中,基于两个原始记录集(如第一和第二记录集、第四和第五记录集)定义记录集(如第三记录集、第六记录集)的方法可以对原始记录集的多个对重复多次。在这样的实施例中,原始记录集可以基于实体的用户的不同属性配对,这些属性可以与不同的优先级相关联。例如,第一组合的记录集可以对于优先级为1的第一属性“电子邮件地址”基于第一对原始记录集定义;第二组合的记录集可以对于优先级为2的第二属性“邮政编码与姓的组合”基于第二对原始记录集定义;第三组合的记录集可以对于优先级为3的第三属性“邮政编码”基于第三对原始记录集定义;诸如此类。

[0049] 在一些实施例中,匹配模块可以被配置为基于诸如第三记录集和第六记录集的两个记录集(即包括第一标识串和第二标识串两者、并且基于两个原始记录集所定义的记录集)来执行匹配过程。具体来说,例如,匹配模块可以被配置为基于第三记录集和第六记录集定义匹配的记录集,使得1)来自匹配记录集的每个记录是来自第三记录集的记录或者来自第六记录集的记录,以及2)匹配记录集包括来自第三记录集和第六记录集的记录的全部或部分。这样的匹配记录集可以基于与第三记录集和第六记录集相关联的属性的优先级定义,使得包括第一标识串(和第二标识串)并与更低优先级相关联的记录被排除在匹配记录集之外——如果包括相同第一标识串(和第二标识串)并与更高优先级相关联的另一记录被包括在匹配的记录集中的话。不仅如此,来自第三记录集和第六记录集的每个记录都包括在匹配的记录集中——如果该记录未被以上准则排除的话。

[0050] 例如,如果第一属性(如“电子邮件地址”)与更高优先级相关联,第二属性(如“邮政编码”)与更低优先级相关联,那么来自第三记录集的每个记录(包括第一属性串)都被包括在匹配的记录集中。对于来自第六记录集的每个记录(包括第二属性串),如果该记录具有的第一标识串被包括在来自第三记录集的记录中,那么来自第六记录集的该记录被排除在匹配的记录集之外;否则来自第六记录集的该记录被包括在匹配的记录集中。

[0051] 另外注意,第一标识和第二标识在这样的匹配过程中是可交换的。也就是,接收第一对记录集(即第一记录集和第二记录集)和接收第二对记录集(即第四记录集和第五记录集)的顺序——等效于第一标识和第二标识的顺序——可以是可交换的。换言之,匹配过程也可以以如下方式执行,使得包括第二标识串(和第一标识串)并与更低优先级相关联的记录被排除在匹配的记录集之外——当且仅当包括相同第二标识串(和第一标识串)并与更高优先级相关联的另一记录被包括在匹配的记录集中时。

[0052] 在一些实施例中,以上描述的这种匹配过程可以在匹配模块处实现,以基于与属性的各种优先级相关联的多于两个的组合记录集来定义匹配的记录集。结果,包括第一标

识串并与(来自多个优先级的)相对更低的优先级相关联的每个记录都被排除在匹配的记录集之外——当且仅当包括相同第一标识串并与(来自多个优先级的)相对更高的优先级相关联的另一记录被包括在匹配的记录集中时。

[0053] 匹配模块可以被配置为以各种方式实现以上描述的匹配过程。在一些实施例中，匹配模块可以被配置为实现“去重复接着组合”方法。例如，为了匹配和组合具有更高优先级属性(即第一属性)的第三记录集与具有更低优先级属性(即第二属性)的第六记录集，匹配模块可以被配置为把第六记录集修改为修改后的状态，以便从第六记录集排除具有的第一标识串等于来自第三记录集的记录的第一标识串的每个记录。匹配模块接着可以被配置为组合第三记录集与修改后状态下的第六记录集以定义匹配的记录集。

[0054] 对于另一个实例，为了匹配和组合具有优先级1(即最高优先级)的第三记录集、具有优先级2(即第二高优先级)的第六记录集和具有优先级3(即第三高优先级)的第七记录集(即组合后的记录集)，匹配模块可以被配置为：1)把第六记录集修改为修改后的状态，以便从第六记录集排除具有的第一标识串等于来自第三记录集的记录的第一标识串的每个记录；2)把第七记录集修改为修改后的状态，以便从第七记录集排除具有的第一标识串等于来自第三记录集的记录的第一标识串的每个记录，以及从第七记录集排除具有的第一标识串等于来自第六记录集的记录的第一标识串的每个记录；以及3)组合第三记录集、修改后状态下的第六记录集和修改后状态下的第七记录集以定义匹配的记录集。

[0055] 在一些其他实施例中，匹配模块可以被配置为实现“组合接着去重复”方法。例如，为了匹配和组合具有更高优先级属性(即第一属性)的第三记录集与具有更低优先级属性(即第二属性)的第六记录集，匹配模块可以被配置为组合第三记录集和第六记录集以定义初始状态下的匹配的记录集。接着匹配模块可以被配置为把初始状态下的匹配的记录集修改为最终状态，从初始状态下的匹配的记录集排除这样的每个记录：1)具有的第一标识串等于来自第三记录集的记录的第一标识串，并且2)与第六记录集相关联(或者等效地，与第二属性相关联)。

[0056] 对于另一个实例，为了匹配和组合具有优先级1(即最高优先级)的第三记录集、具有优先级2(即第二高优先级)的第六记录集和具有优先级3(即第三高优先级)的第七记录集(即组合后的记录集)，匹配模块可以被配置为：1)组合第三记录集、第六记录集和第七记录集以定义第一状态下的匹配的记录集；2)把第一状态下的匹配的记录集修改为第二状态，以便从第一状态下的匹配的记录集排除这样的每个记录：(i)具有的第一标识串等于来自第三记录集的记录的第一标识串，并且(ii)与第六记录集相关联(或者等效地与第二高优先级相关联)或与第七记录集相关联(或者等效地与第三高优先级相关联)；以及3)把第二状态下的匹配的记录集修改为第三状态(即最终状态)，以便从第二状态下的匹配的记录集排除这样的每个记录：(i)具有的第一标识串等于来自第六记录集的记录的第一标识串，并且(ii)与第七记录集相关联(或者等效地与第三高优先级相关联)。

[0057] 以上描述的两种方法是如何对两个或更多组合后的记录集实现匹配过程的实例。在一些实施例中，这两种方法在实现中可以组合。在其他实施例中，匹配过程可以以任何其他适合的方法实现。在一些实施例中，匹配模块可以被配置为把指明匹配的记录集的信号发送到向匹配模块提供原始记录集的一台或多台计算设备(如提供第一记录集和第四记录集的第一计算设备、提供第二记录集和第五记录集的第二计算设备)。这样的信号可以经由

例如宿有匹配模块的计算设备的通信接口(如图2中计算设备200的通信接口230)发送。另外,在一些实施例中,匹配模块可以被配置为以类似方法向第一计算设备和/或第二计算设备发送组合后的记录集,诸如第三记录集或第六记录集。

[0058] 在一些实施例中,与属性相关联的优先级可以被改变使得具有各种优先级的多个组合后的记录集可以被匹配,以便使用同一方法定义不同的匹配的记录集。因此,多个匹配的记录集可以在匹配模块处基于不同的顺序或与属性相关联的优先级定义。多个匹配的记录集可以在匹配模块处进一步进行比较并且可以确定最佳匹配的记录集。在一些实施例中,匹配模块可以被配置为把最佳匹配的记录发送到向匹配模块提供原始记录集的一台或多台计算设备(如图1的计算设备152、154)。

[0059] 在一些实施例中,匹配模块可以定义包括匹配过程的特性的报告。在这样的实施例中,该报告可以包括例如每个优先级的匹配率(如作为总记录的百分比的匹配数量)、每个优先级的匹配数量、特定优先级的新匹配数量(如未作为重复被排除的匹配)、添加每个优先级时的累积匹配数量,以及/或者添加每个优先级时的累积匹配率。在这样的实施例中,匹配模块可以被配置为向一台或多台计算设备发送指明报告的信号。在这样的实施例中,接收该报告的计算设备可以被配置为例如在每个优先级的匹配率低于该优先级(或相关联的属性)或者优先级(或相关联的属性)的特定组合的预定阈值时,启动对匹配的总数在该优先级(或相关联的属性)或者优先级(或相关联的属性)的特定组合的预定阈值之下等的报警和/或其他通知。在一些实施例中,这样的报告可以用于手工地或自动地选择在最终匹配的记录集中使用哪些优先级以及/或者选择哪个顺序来对匹配的记录集去重复,如本文中所描述的。

[0060] 图4A至图4L说明根据实施例的匹配记录集的过程。图4A至图4L说明的过程是参考关于图3所示出和描述的方法300的实例。该过程在类似于图2的匹配模块254和关于图3所描述的匹配模块的匹配模块处、与关于图1至图3所示出和描述的计算设备(如图1的计算设备152、154、110,图2的计算设备200)协作地执行。

[0061] 图4A至图4D说明在计算设备处定义的以及/或者从计算设备向匹配模块发送的记录集的第一组。具体来说,图4A示出了在第一计算设备(如图1的计算设备152)处定义的包括原始数据(即未加密或散列)的第一记录集。第一记录集包括与第一实体的用户(如网站的注册用户)的第一属性——电子邮件地址相关联的数据。正如图4A所示,第一记录集至少包括索引列和两个内容列:索引列在最左位置,它包括在第一记录集中存储的每个记录的索引(如7、8);第一属性串(如sally.doe@test.com, jane.doe@test.com)的列A,第一属性串是第一实体用户的电子邮件地址;以及第一标识串(如444444444, 222222222)的列B,第一标识串是与第一实体对应用户相关联的标识。因此,第一记录集中的每个记录都包括与第一实体用户相关联的第一属性串(如mary.doe@test.com)和第一标识串(如333333333)。

[0062] 图4C示出了第一加密记录集,它是从图4A中第一记录集加密的。具体来说,通过在第一计算设备处使用第一加密密钥,对第一记录集的记录中的每个第一标识串(如5 5 5 5 5 5 5 5) 加密 以 产 生 加 密 的 第 一 标 识 串 (例 如 , “RQbe7d1bPVe4aQFDI4vL25QJhIMIJjem0IWjY4eGAVs=”), 它被存储在第一加密记录集的对应记录的列B中。通过在第一计算设备处使用散列函数,第一记录集的记录中的每个第一属性串(如john.doe@test.com)都被散列以产生散列的第一属性串(例如,

“c7b57cle90c710de01c353b161df24c2c7b593a8”），它被存储在第一加密记录集的对应记录的列A中。

[0063] 类似于图4A，图4B示出了在第二计算设备（如图1中的计算设备154）处定义的包括原始数据（即未加密）的第二记录集。第二记录集包括与第二实体的用户（如在零售连锁店的购物者）的第一属性相关联的数据。正如图4B所示，第二记录集至少包括索引列和两个内容列：索引列在最左位置，它包括在第二记录集中存储的每个记录的索引（如7、8）；第一属性串（如sally.doe@test.com, jane.doe@test.com）的列A，第一属性串是第二实体用户的电子邮件地址；以及第二标识串（如666666666, 777777777）的列B，第二标识串是与第二实体的对应用户相关联的标识。因此，第二记录集中的每个记录都包括与第二实体的用户相关联的第一属性串（如mary.doe@test.com）和第二标识串（如999999999）。

[0064] 类似于图4C，图4D示出了第二加密记录集，它是从图4B中的第二记录集加密的。具体来说，通过在第二计算设备处使用第二加密密钥（可以与在第一计算设备所使用的第一加密密钥不同），对第二记录集的记录中的每个第二标识串（如777777777）加密以产生加密的第二标识串（例如，“FEJJ+1K5zwwbG2RQYjsDnGd6fz/Dg17QP2WDscfsWYg=”），它被存储在第二加密记录集的对应记录的列B中。通过在第二计算设备处使用散列函数（与在第一计算设备处所使用的相同），第二记录集的记录中的每个第一属性串（如john.deo@test.com）被散列，以便产生散列的第一属性串（例如，“c7b57cle90c710de01c353b161df24c2c7b593a9”），它被存储在第二加密记录集的对应记录的列A中。

[0065] 类似于图4A至图4D，图4E至图4H说明在计算设备处定义的以及/或者从计算设备向匹配模块发送的记录集的第二组。具体来说，类似于图4A，图4E示出了包括在第一计算设备处定义的原始数据的第三记录集。第三记录集包括与第一实体的用户的第二属性——名字和地址（具体来说，姓、名、城市和州的组合）相关联的数据。正如图4E所示，第三记录集至少包括索引列和两个内容列：索引列在最左位置，它包括在第三记录集中存储的每个记录的索引；第二属性串的列A，第二属性串是第一实体的用户名字和地址；以及第一标识串的列B，第一标识串是与第一实体的对应用户相关联的标识。因此，第三记录集中的每个记录都包括与第一实体的用户相关联的第二属性串和第一标识串。

[0066] 类似于图4C，图4G示出了第三加密记录集，它是从图4E中的第三记录集加密的。具体来说，通过在第一计算设备处使用第一加密密钥，对第三记录集的记录中的每个第一标识串加密以产生加密的第一标识串，它被存储在第三加密记录集的对应记录的列B中。通过在第一计算设备处使用散列函数，第三记录集的记录中的每个第二属性串被散列以产生散列的第二属性串，它被存储在第三加密记录集的对应记录的列A中。

[0067] 类似于图4B和图4E，图4F示出了在第二计算设备处定义的包括原始数据的第四记录集。第四记录集包括与第二实体的用户的第二属性相关联的数据。正如图4F所示，第四记录集至少包括索引列和两个内容列：索引列在最左位置，它包括在第四记录集中存储的每个记录的索引；第二属性串的列A，第二属性串是第二实体的用户名字和地址；以及第二标识串的列B，第二标识串是与第二实体的对应用户相关联的标识。因此，第四记录集中的每个记录都包括与第二实体的用户相关联的第二属性串和第二标识串。

[0068] 类似于图4G和图4D，图4H示出了第四加密记录集，它是从图4F中的第四记录集加

密的。具体来说,通过在第二计算设备处使用第二加密密钥,对第四记录集的记录中的每个第二标识串加密以产生加密的第二标识串,它被存储在第四加密记录集的对应记录的列B中。通过在第二计算设备处使用散列函数,第四记录集的记录中的每个第二属性串被散列以产生散列的第二属性串,它被存储在第四加密记录集的对应记录的列A中。

[0069] 图4I说明了存储第一属性(即电子邮件)和第二属性(即姓、名、城市和州的组合)的信息的配置文件(图4I中示出为waterfall.list)的屏幕截图。在一些实施例中,与第一属性和第二属性的优先级相关联的信息也可以被存储在这样的配置文件中。在图4I的实例中,第一属性与索引1相关联,第二属性与索引2相关联,这在一些实施例中指明第一属性的优先级高于第二属性的优先级。在一些实施例中,这样的配置文件可以存储在例如与匹配模块相关联的存储器(如与图2中的匹配模块254相关联的存储器210)中,并且由匹配模块检索以便于执行匹配方法。

[0070] 图4J至图4L说明对图4A至图4H所示的记录集执行图3所描述的匹配过程的结果(图4J和图4K中的中间组合记录集以及图4L中的最终匹配记录集)。注意,图4J至图4L所示的记录集是为了说明和解释目的的解密版本。典型地,这样的解密记录集在匹配模块或任何其他计算设备(如第一计算设备、第二计算设备)处不可用,因为第一加密密钥和第二加密密钥典型地在任何单一计算设备处不同时可用。

[0071] 第一计算设备可以向匹配模块发送第一加密记录集(图4C所示)和第三加密记录集(图4G所示);第二计算设备可以向匹配模块发送第二加密记录集(图4D所示)和第四加密记录集(图4H所示)。根据图3中步骤306和312所描述的方法,匹配模块可以被配置为比较和组合第一加密记录集和第二加密记录集以定义第一组合后的加密记录集(图中未显示)。图4J示出了第一组合后的加密记录集的解密版本。正如图4A和图4B所示,由于第一记录集中具有索引9的记录中的第一属性串(图4A以圆圈突出的电子邮件地址john.doe@test.com)与第二记录集中具有索引9的记录中的第一属性串(图4B以圆圈突出的电子邮件地址john.deo@test.com)不同,所以这两个记录不匹配。第一记录集中的每个其他记录与第二记录集中的记录匹配(按照第一属性串),反之亦然。结果,来自第一记录集的匹配记录中的加密后的第一标识串和来自第二记录集的匹配记录中的加密后的第二标识串被包括在第一组合后的加密记录集中,其解密版本示出在图4J中。

[0072] 类似地,根据图3中步骤306和312所描述的方法,匹配模块可以被配置为比较和组合第三加密记录集和第四加密记录集以定义第二组合后的加密记录集(图中未显示)。图4K示出了第二组合后的加密记录集的解密版本。正如图4E和图4F所示,由于第三记录集中具有索引11的记录中的第二属性串(图4E以圆圈突出的“doe,sally,boulder,co”)与第四记录集中具有索引11记录中的第二属性串(图4F以圆圈突出的“doe,saly,boulder,co”)不同,所以这两个记录不匹配。第三记录集中的每个其他记录与第四记录集中的记录匹配(按照第二属性串),反之亦然。结果,来自第三记录集的匹配记录中的加密后的第一标识串和来自第四记录集的匹配记录中的加密后的第二标识串被包括在第二组合后的加密记录集中,其解密版本示出在图4K中。

[0073] 不仅如此,根据关于图3所描述的匹配方法,匹配模块可以被配置为匹配和整合第一组合后的加密记录集(其解密版本示出在图4J中)和第二组合后的加密记录集(其解密版本示出在图4K中),以定义最终匹配的记录集(其解密版本示出在图4L中)。具体来说,因为

根据图4I中的配置文件,第一属性(即电子邮件地址)具有比第二属性(即名字和地址)更高的优先级,所以来自第一组合后的加密记录集(它与第一属性相关联)的记录比来自第二组合后的加密记录集(它与第二属性相关联)的记录具有更高的优先级。结果,来自第一组合后的加密记录集的每个记录都被包括在最终匹配的记录集中;而当且仅当来自第二组合后的加密记录集的每个记录具有的加密后第一标识串不等于来自第一组合后的加密记录集的记录中的加密后第一标识串时,该记录才被包括在最终匹配的记录集中。正如图4J至图4L中的解密版本所示,来自图4J的解密记录集的每个记录都被包括在图4L的解密记录集中;而来自图4K的解密记录集的仅具有索引14的记录被包括在图4L的解密记录集中,因为来自图4K的解密记录集的其他三个记录(具有索引13、15和16)具有的第一标识串(即999999999,666666666,777777777)等于来自图4J的解密记录集的记录中的第一标识串。

[0074] 虽然以上关于图3至图4L示出和描述了基于两个原始记录集(如在方法300中所描述的第一和第二记录集、在方法300中所描述的第四和第五记录集)来定义组合的记录集(如图3的方法300中所描述的第三记录集或第六记录集),但是在其他实施例中,可以以类似方法基于多于两个的原始记录集来定义这样的组合记录集。在这样的实施例中,组合记录集可以包括多于两个的标识串。例如,组合的记录集可以基于三个原始记录集,通过定义组合记录集的每个记录都包括来自第一原始记录集的记录中的第一标识串、来自第二原始记录集的记录中的第二标识串和来自第三原始记录集的记录中的第三标识串来定义,其中这三个记录的每个记录都包括公共属性串。

[0075] 虽然关于图4A至图4L示出和描述的记录集包括索引列,但是在一些实施例中,记录集可以排除索引列。在这样的实施例中,匹配的记录集中的记录的优先级可以基于与任何列和/或行索引无关的记录集中的记录的顺序来指明。在一些实施例中,记录和/或记录项可以包括指明优先级和/或相关联属性的元数据。在这样的实施例中,是否在组合记录集中包括记录(如是否删除重复项,以及删除哪个项)可以基于该记录和/或记录项的顺序以及/或者相关联的元数据。

[0076] 虽然本文所描述的记录集包括与消费者和/或用户相关联的记录,但是在一些实施例中,记录集可以包括制作的和/或其他方式已知的记录,使得匹配模块可以检查匹配的准确度。例如,第一实体可以在向匹配模块发送的记录集中包括已知记录。类似地,第二实体可以在向匹配模块发送的记录集中包括相同的已知记录。以这种方式,在匹配模块比较了来自第一实体的记录集与来自第二实体的记录集之后,匹配模块可以确认已知记录被包括在匹配的记录集中。在这样的实施例中,已知记录在匹配的记录集中的存在可以确认匹配方法正在起作用,对记录集中的记录使用的散列函数被正确地实现,以及/或者对包括记录集的数据文件使用的加密被正确地实现。在一些实施例中,每对记录集(即与优先级相关联的每对记录集)能够包括不同的已知记录。

[0077] 本文在多个实施例中描述了记录集、匹配的记录集、数据文件等包括原始数据、加密数据和/或散列数据。在一些实施例中,数据可以无任何加密和/或散列地在实体之间传递,具有原始数据(未加密和/或散列)的加密和/或散列数据文件以及/或者具有加密和/或散列数据的未加密数据文件。例如,数据文件可以从一个实体(加密或未加密地)传送到另一个实体;并且该数据文件可以包括具有散列的、加密的和/或原始的标识串的列表的记录集,每个标识串都与散列的、加密的和/或原始属性串相关联。在一些实施例中,第一实体

(例如数据分析实体)可以把数据与多于一个的其他实体匹配。在这样的实施例中,数据分析实体可以具有对每个其他实体的唯一散列要素(hash salt)的访问权,每个其他实体仅可以具有其唯一的散列要素。

[0078] 在本文描述的一些实施例中,记录集可以在第一计算设备准备并发送给第二计算设备。例如,第一计算设备可以准备包括与第一属性串相关联的第一标识串的第一列表的第一记录集,并且准备包括与第二属性串相关联的第一标识串的列表的至少一部分的第二记录集。第一计算设备可以散列和/或加密(或保持原始)第一标识串、第一属性串和/或第二属性串,并且可以加密(或保持未加密)第一记录集和/或第二记录集,并且可以把记录集发送给另一台计算设备,包括具有匹配模块的计算设备。在其他实施例中,第一计算设备可以发送一个或多个未准备的数据文件,未准备的数据文件包括包含第一标识串、相关联的第一属性串和相关联的第二属性串一个或多个记录集。在这样的实施例中,接收这一个或多个未准备数据文件的计算设备可以组合数据文件(如果必要的话),并且可以准备第一记录集,以包括第一标识串和相关联的第一属性串,并且可以准备第二记录集,以包括第一标识串和相关联的第一属性串,并且可以把第一记录集和/或第二记录集发送到匹配模块。在一些实施例中,具有匹配模块的计算设备可以接收准备的记录集(如准备以被匹配的记录集)以及未准备的记录集(如未准备以被匹配的记录集)。在一些实施例中,准备记录集可以包括向记录集中的每个记录添加散列要素,如向记录集中的每个属性串添加散列要素。在这样的实施例中,一对要被匹配的记录集中的记录(具体来说是属性串)可以包括相同的散列要素,使得具有相同第一属性串的记录将匹配。

[0079] 虽然以上描述的匹配模块基于来自一个记录集的具有与来自另一记录集的记录相等的属性串的记录来匹配记录,但是在其他实施例中,匹配模块可以基于来自一个记录集的具有与来自另一记录集的记录概率上等效的属性串的记录来匹配记录。例如,与johndoe434@firstprovider.com相关联的属性串可以和与johndoe434@secondprovider.com相关联的属性串概率上等效。换一种说法,与johndoe434@firstprovider.com相关联的用户可能(如高于预定的确定性)与作为johndoe434@secondprovider.com的用户相同,即使属性串不一致。其他实例可以包括使昵称与全名匹配等等。

[0080] 虽然以上描述的方法涉及具有两种或三种属性和相关联的优先级,但是在其他实施例中,可以使用更多或更少的属性和相关联的优先级。例如,方法可以包括每个都与优先级相关联的五个属性。在这样的实例中,匹配模块可以接收和比较五对原始记录集以定义五个组合后的记录集。继续这个实例,匹配模块可以使用五个组合后的记录集的任何组合来定义多个匹配记录集,并且可以比较来自该多个匹配的记录集的每个匹配的记录集并且根据预定准则选择一个匹配的记录集。在这样的实例中,预定准则可以包括例如匹配的记录集中记录的最大或最小数量(如至少1000个匹配记录)、匹配的记录集中所包括的优先级的最大或最小量(如不多于5个优先级中的3个)、要求的优先级等,以及这些的组合。例如,预定准则可以包括:匹配集必须包括与优先级1和2相关联的组合后的记录集、5个组合后的记录集中的不多于4个组合后的记录集、以及满足头两个准则的记录数量最大的匹配记录集。

[0081] 本文描述的系统和方法意在可以通过软件(在存储器中存储的和/或在硬件上执

行的)、硬件或其组合执行。硬件模块可以包括例如通用处理器、现场可编程门阵列 (FPGA) 以及/或者专用集成电路 (ASIC)。软件模块(在硬件上执行的)可以以多种多样的软件语言(如计算机代码)表示,包括Unix实用程序、C、C++、JavaTM、Ruby、Visual BasicTM以及其他面向对象的、过程的或其他编程语言和开发工具。计算机代码的实例包括但是不限于微代码或微指令、机器指令(诸如由编译器产生的)、用于产生网络服务的代码以及包含由计算机使用解释器执行的更高级指令的文件。计算机代码的附加实例包括但是不限于控制信号、加密代码以及压缩代码。

[0082] 本文描述的一些实施例涉及具有非暂态计算机可读介质(也可以被称为非暂态处理器可读介质或存储器)的设备(如无线接入点、移动通信设备),非暂态计算机可读介质上具有用于执行多种计算机实施的操作的指令或计算机代码。计算机可读介质(或处理器可读介质)在它不包括暂态传播信号本身(如在诸如空间或电缆的传输媒介上携带信息的传播电磁波)的意义上是非暂态的。介质和计算机代码(也可以称为代码)可以为专用目的而设计和构建。非暂态计算机可读介质的实例包括但是不限于:磁存储介质,诸如硬盘、软盘和磁带;光学存储介质,诸如光盘/数字视频盘(CD/DVD)、光盘只读存储器(CD-ROM)以及全息(holographic)设备;磁光存储介质,诸如光盘;载波信号处理模块;以及专门配置为存储和执行程序代码的硬件设备,诸如专用集成电路(ASIC)、可编程逻辑器件(PLD)、只读存储器(ROM)以及随机存取存储器(RAM)设备。本文描述的其他实施例涉及计算机程序产品,其可以包括例如本文讨论的指令和/或计算机代码。

[0083] 虽然以上已经描述了各种实施例,但是应当理解,呈现它们仅仅作为实例而不是限制。在以上描述的方法和步骤指明以特定顺序发生某些事件时,某些步骤的顺序可以修改。另外,某些步骤在可能时可以在并行过程中同时执行,以及如以上所描述地顺序执行。尽管各种实施例已经被描述为具有具体的特征和/或组件的组合,但是具有根据本文所描述的任何实施例的任何特征和/或组件的任何组合或子组合的其他实施例也是可能的。不仅如此,尽管各种实施例被描述为具有与特定计算设备相关联的特定实体,但是在其他实施例中,不同实体可以与其他和/或不同的计算设备相关联。例如,虽然计算设备152和计算设备110被描述为与线上实体相关联,计算设备154被描述为与数据分析实体相关联,但是在其他实施例中,计算设备152和计算设备110可以与数据分析实体相关联,计算设备110可以与线上实体相关联。

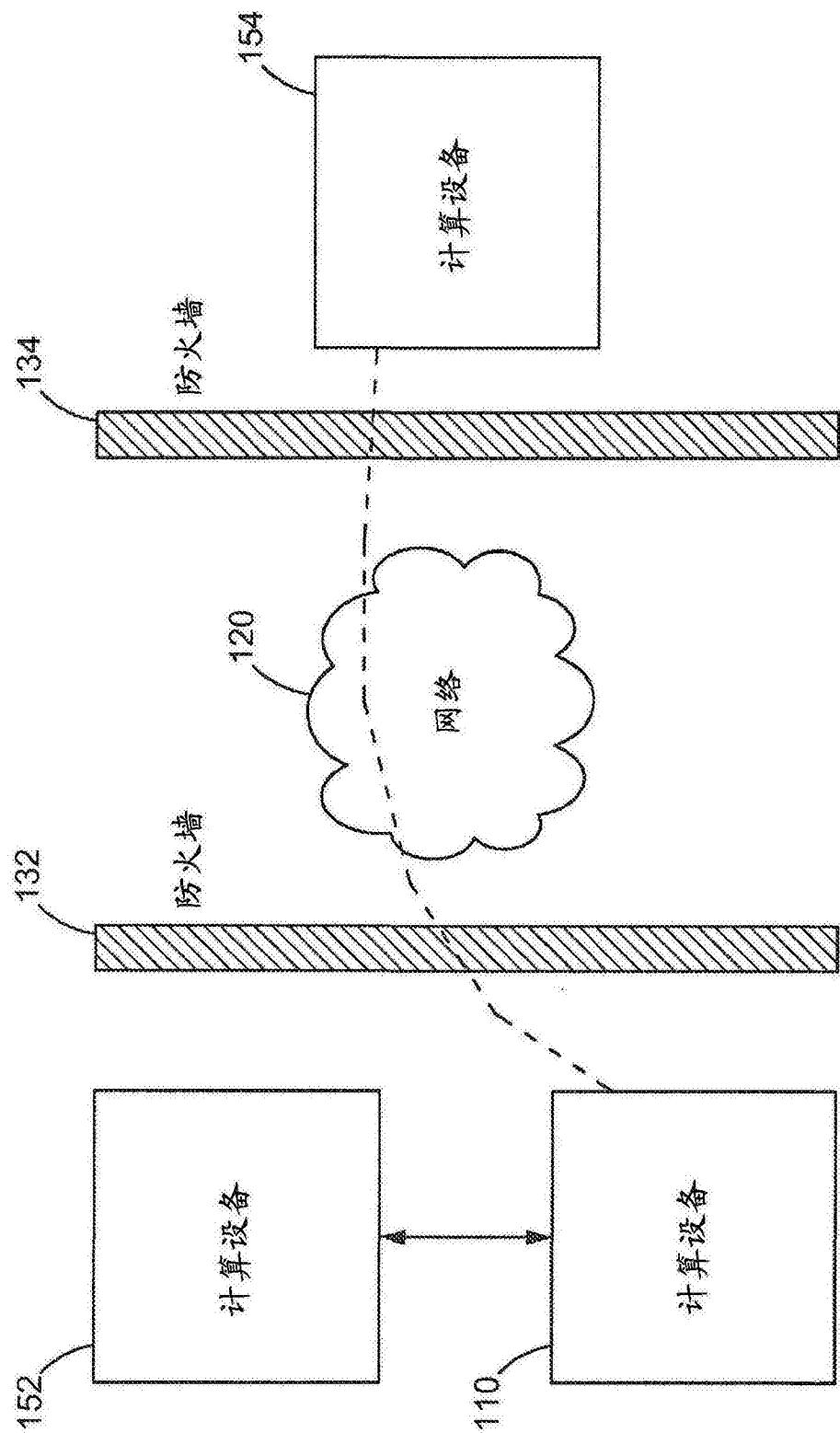


图1

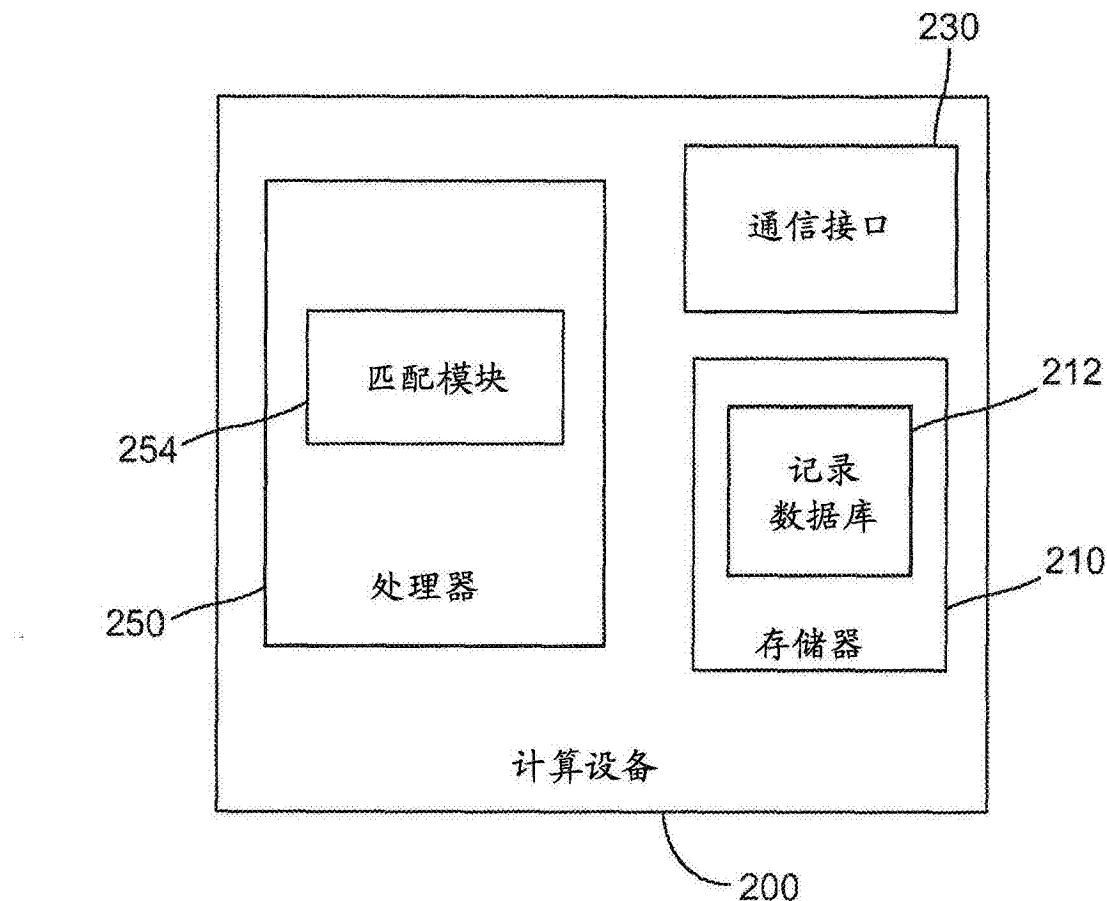


图2

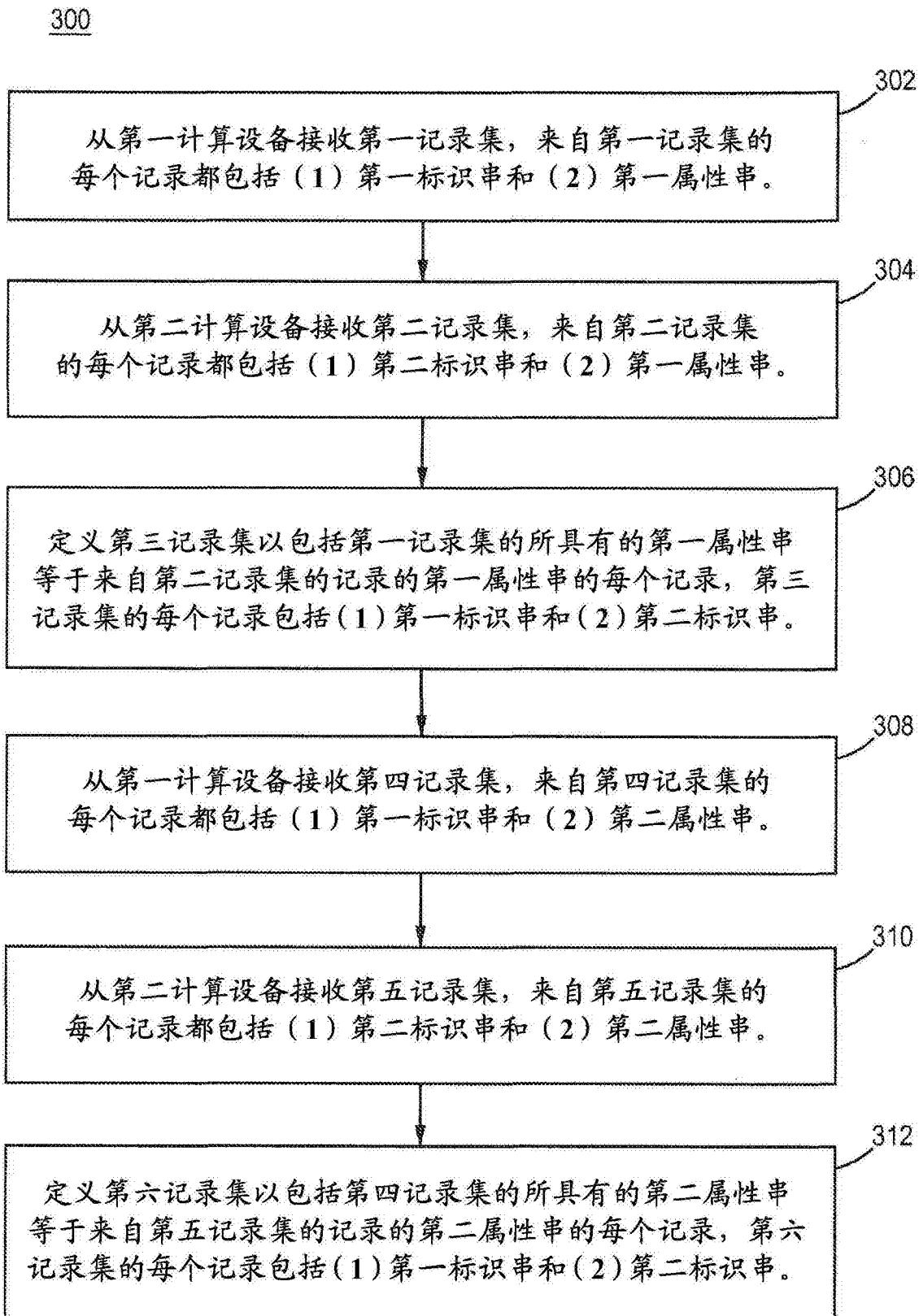


图3

△	A	B
7	sally.doe@test.com	6666666666
8	jane.doe@test.com	7777777777
9	john.doe@test.com	8888888888
10	mary.doe@test.com	9999999999
11	harry.doe@test.com	101010101

图4A

△	A	B
7	sally.doe@test.com	6666666666
8	jane.doe@test.com	7777777777
9	john.doe@test.com	8888888888
10	mary.doe@test.com	9999999999
11	harry.doe@test.com	101010101

图4B

	A	B
1	e170c5e297a6e342fa7c6b24fb4c7e9c1015e983	8X0bl5+TTayN4FxUDOeiPqw6XUL1R5B4X8YlakVKu2 =
2	780b026245305e4a93fc2d189e4bcd03c405b924	oFkW1A1TbCGQsTa8vWYBSIn5TPBgyyWuNO0upR9GI+4=
3	c7b57c1e90c710de01c353b161df24c2c7b593a8	RQbe7d1bPVe4aQFDI4vL25QJhIMUjemOlWjY4eGAVse=
4	9842664fab1b267737e7d23559eb7da46df806	bIB4pcFVWOOejEB/7XVwpaADN2vIN2toS8uefG0clCg=
5	0e856b5710b6da18eb864d1c9a4a7042832ac725	3WJg+/nyBnhrrYRMz1ydhcz+3UJ8avWF89yaap0CGq38=

图4C

	A	B
1	e170c5e297a6e342fa7c6b24fb4c7e9c1015e983	O0Jx0YMR6BoKlcIgs6kRPQ9ahqDYmpRSbPEq8da7aTw=
2	780b026245305e4a93fc2d189e4bcd03c405b924	ZRyskh6pL66OEFcXcZShwF8oLkjwn9RwpOpOWIEZe25l=
3	c7b57c1e90c710de01c353b161df24c2c7b593a9	FEJJ+IK5zwwbG2RQYjsDnGd6fz/DgI7QP2WDscfsWYg=
4	9842664f6ab1b267737e7d23559ebe7da46df806	2IZ+VcJc8K0sUV8BnHmcrxJJ7eM6lue6GqGduxAI96o=
5	0e856bb5710b6da18eb864d1c9a4a7042832act25	PAcqiAIDom8ugBc35Hyq96gReiSSL/GwWDDyLi/0OM=

图4D

A	B
7 doe,john,denver,co	6666666666
8 doe,mary,westminster,co	7777777777
9 doe,harry,arvada,co	8888888888
10 doe,jane,golden,co	9999999999
11 doe,sally,boulder,co	1010101010

图4E

A	B
7 doe,john,denver,co	6666666666
8 doe,mary,westminster,co	7777777777
9 doe,harry,arvada,co	8888888888
10 doe,jane,golden,co	9999999999
11 doe,sally,boulder,co	1010101010

图4F

	A	B
1	c5c61326bf2e9f37c9dba42ea059dd1e891af837	RQbe7d1bPVe4aQFDI4vL25QJhlMijemOlWjY4eGAVs=
2	b070c9191116c5fcb1031b66930ff0579a757190	blB4pcFVWOOeJEB/7XVwpaaADN2viW2tS8uefG0clCg=
3	7a16bb75a05715a3970692e56b926c22ffd4c434	3WJg+/nyBnhrRyMz1ydhZ+3El8avWF89yaap0CGq38=
4	7f1c3f7b9ee5fb40ded2e46640f11a90f3afb196	oEkwM1A1hCGOsTa8vWYBSIn5TPBgyyWuNO0upR9Gi+4=
5	hb40702eebac0fdc23b771cc3b0e436a99b3dd8	8X0bl5+TTayN4FxUDOeiPqW6XUL1R5B4X8YiakVKu2I=

图4G

	A	B
1	c5c61326bf2e9f37c9dba42ea059dd1e891af837	0OJx0YMR6BoKlcIgs6kRfPQ9ahqDYmpRSbPEq8da7aTw=
2	b070c9191116c5fcf1031b66930ff0579a757190	ZRyskh6pl66OEfcXcZShwf8oLkjwn9RwbpoOWIEZe25j=
3	7a16bb75a05715a3970692e56b926c22ffd4c434	FEJJ+IK5zwmwbG2RQYjsDnGd6fzIDgj7QP2WDscfsWYg=
4	7f1c377b9ee5f940ded2e46640f11a90f3afb196	2l7+VcJc8K0sUV8BnHmcrxLJ7eM6lue6GqGduxA1g6o=
5	bb40702eebac0fac23b771cc3B0e436a99b3dd9	PAccqjAIarovt8ugBc35Hyqq96gReiSSL/GwWDyLI/0OM=

图4H

	waterfall.list
1	email.csv
2	lastname,firstname,city,state.csv

图4I

	A	B
13	999999999	222222222
14	888888888	111111111
15	666666666	555555555
16	777777777	333333333

图4J

	A	B
13	999999999	222222222
14	888888888	111111111
15	666666666	555555555
16	777777777	333333333

图4K

∠	A	B
13	999999999	333333333
14	888888888	111111111
15	666666666	444444444
16	101010101	111111111
17	777777777	222222222

图4L