



(12) 发明专利申请

(10) 申请公布号 CN 103201748 A

(43) 申请公布日 2013. 07. 10

(21) 申请号 201180053956. 2

(74) 专利代理机构 北京市金杜律师事务所

(22) 申请日 2011. 09. 09

11256

代理人 王茂华

(30) 优先权数据

2010-250600 2010. 11. 09 JP

(51) Int. Cl.

G06F 21/62(2013. 01)

(85) PCT申请进入国家阶段日

2013. 05. 08

(86) PCT申请的申请数据

PCT/JP2011/070618 2011. 09. 09

(87) PCT申请的公布数据

W02012/063546 JA 2012. 05. 18

(71) 申请人 日本电气株式会社

地址 日本东京都

(72) 发明人 伊东直子 丰田由起

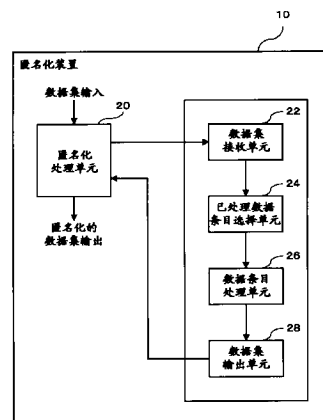
权利要求书2页 说明书10页 附图22页

(54) 发明名称

匿名化装置和匿名化方法

(57) 摘要

本发明即使在数据集可能被重复提供、其中后来添加的数据条目的属性信息显著偏离已知数据条目的属性信息的值范围的情况下仍然允许适当的泛化。对于具有多个条目的数据集的每个数据条目，基于预定泛化规则来泛化构成准标识符的至少一个数据属性，其中数据条目包括构成准标识符的至少一个属性，该准标识符是可以识别个人的信息，于是在数据集所包括的多个数据条目中，在基于泛化规则被泛化时使数据集不满足预定匿名性标准的数据条目以及由于在该数据条目与泛化对象之间共享属性数据值而使数据集满足预定匿名性标准的至少一个数据条目被选择，于是对于所选择的数据条目，泛化对象的属性数据值被修改为预定的共享值，而不考虑该预定泛化规则。



1. 一种匿名化装置,包括:

泛化单元,被配置为基于预定泛化规则而针对数据集的每个数据条目泛化形成准标识符的至少一个属性数据的值,所述数据集具有多个数据条目,每个所述数据条目包括形成所述准标识符的所述至少一个属性数据以及除所述准标识符之外的至少一个属性数据,所述准标识符是可以标识个人的信息;

条目选择单元,被配置为从所述数据集中包括的所述多个数据条目中选择在基于所述泛化规则被泛化时变成所述数据集不满足预定匿名性标准的因素的数据条目,以及至少一个数据条目,所述至少一个数据条目的泛化目标属性数据具有与该数据条目共同的值,从而使所述数据集能够满足所述预定匿名性标准;以及

条目处理单元,被配置为针对由所述条目选择单元选择的所述数据条目,将所述泛化目标属性数据的所述值改变为与所述预定泛化规则无关的预定共同值。

2. 根据权利要求1所述的匿名化装置,其中所述条目选择单元从所述数据集中所包括的所述多个数据条目中选择如下数据条目,当所述数据条目基于所述泛化规则而被泛化时使所述数据集无法满足所述预定匿名性标准,以及具有在形成所述准标识符的所述至少一个属性数据中的、未基于所述泛化规则而被泛化的属性数据的不同值的多个数据条目。

3. 根据权利要求1所述的匿名化装置,其中所述条目选择单元从所述数据集中所包括的所述多个数据条目中选择如下数据条目,当所述数据条目基于所述泛化规则而被泛化时使所述数据集无法满足所述预定匿名性标准,以及至少一个数据条目,所述至少一个数据条目使所述数据集满足所述预定匿名性标准,即使此至少一个数据条目被排除在所述数据集之外。

4. 根据权利要求3所述的匿名化装置,其中所述条目处理单元将所述泛化目标属性数据的所述值改变为所述预定共同值,以允许所述数据集满足所述预定匿名性标准,并且将形成所述准标识符的所述至少一个属性数据中的、除所述泛化目标属性数据之外的至少一个属性数据的值改变为预定共同值。

5. 根据权利要求1到4中任一项所述的匿名化装置,其中,当数据条目被新添加到所述数据集时,如果当添加的所述数据条目的值和所述数据条目中的、其属性数据的值已被改变的至少一个数据条目的改变之前的值基于所述泛化规则而被泛化时所述数据集满足所述预定匿名性标准,则所述条目处理单元针对此至少一个数据条目而将所述属性数据的值改变为通过基于所述泛化规则来泛化所述改变之前的所述值而获得的值。

6. 根据权利要求1到5中任一项所述的匿名化装置,其中,如果所述数据集由于从所述数据集删除至少一个数据条目而无法满足所述预定匿名性标准,所述条目处理单元向所述数据集添加伪数据条目以允许所述数据集满足所述预定匿名性标准。

7. 根据权利要求6所述的匿名化装置,其中,如果所述数据集由于向所述数据集新添加数据条目而即使将所述伪数据条目排除也满足所述预定匿名性标准,所述伪数据条目从所述数据集被删除。

8. 根据权利要求1到7中任一项所述的匿名化装置,还包括条目选择规则输入单元,被配置为输入由所述条目选择单元执行的对数据条目的选择的规则,其中

所述条目选择单元基于从所述条目选择规则输入单元输入的所述规则来选择所述数据条目。

9. 根据权利要求1到8中任一项所述的匿名化装置,还包括条目处理规则输入单元,被配置为输入由所述条目处理单元执行的所述数据条目的处理的规则,其中

所述条目处理单元基于从所述条目处理规则输入单元输入的所述规则来处理所述数据条目。

10. 根据权利要求1到8中任一项所述的匿名化装置,还包括泛化规则输入单元,被配置为输入由所述泛化单元执行的所述数据条目的泛化的规则,其中

所述泛化单元基于从所述泛化规则输入单元输入的所述规则来泛化所述数据条目。

匿名化装置和匿名化方法

技术领域

[0001] 本发明涉及一种匿名化装置和一种匿名化方法。

背景技术

[0002] 近年来,用于隐私保护数据公开以允许对公司所拥有的个人信息(微数据)进行二次使用同时保护用户隐私的技术引起了注意。非专利文档 1 提出了一种用于隐私保护数据公开的技术。在各种用户信息(微数据)中,通过与其他背景知识组合可以识别个人的属性信息的集合被称为准标识符。用户不希望公开的属性信息被称为敏感数据。在用于隐私保护数据公开的技术之一的匿名化中,不仅删除显式的用户标识符,而且使形成准标识符的属性信息是含糊的,以避免从这些种类的属性信息的组合识别个人,或者可以减弱准标识符与敏感数据之间的关联,由此改进用户信息的匿名性。

[0003] 用于匿名化的具体操作包括用于以较高级别的概念来替代数据的泛化(generalization)、用于抑制数据的抑制、用于分割表并且减弱标识信息与秘密信息关联的剖析(anatomization),用于在一组数据中交换在数据被泛化时期准标识符相同的标识信息与秘密信息的置换以及用于对数据添加噪声等的扰动。在作为此类操作中的最常见方法的泛化中,数据条目根据准标识符的属性被分组,针对每个分组对准标识符的属性值进行泛化,并且相同的泛化准标识符被给予属于相同准标识符组的数据条目。

[0004] 作为用于泛化的估计隐私保护的基本索引,有 k -匿名性。 k -匿名性表明:存在具有相同的泛化准标识符的 k 个或更多个数据条目。此外,称为 I -多样性的索引表明: I 类或更多类的敏感数据值存在于具有相同泛化准标识符的数据条目中。基本上, k 和 I 的值越大,认为隐私被越有力地保护。已经研究了在抑制信息丢失的同时实现泛化以增大 k 和 I 的值的方法。

[0005] k -匿名性和 I -多样性是关注泛化数据集的单个提供的隐私保护的索引。此外,非专利文档 2 提出了一种称为 m -不变性的索引,它将在数据被多次提供时由组合这些数据的泛化数据集而泄露隐私的风险纳入考虑。 m -不变性表明在连续发布的泛化数据集中所包括的所有准标识符分组中存在具有不同敏感数据值的 m 个或更多个数据条目,并且表明跨多个泛化数据集而存在的数据条目所属的泛化准标识符分组中包括的敏感数据值的集合是相同的。如果 m -不变性被确保,则 I -多样性被同时满足。为了确保 m -不变性,给出了一种在添加伪条目之后执行准标识符组的泛化的方法。

[0006] 非专利文档 1:Chen, B.; Kifer, D.; Lefevre, K.; Machanavajjhala, A., "Privacy-Preserving Data Publishing", Foundations and Trends in Databases, 2009 年,第二卷,第 1-167 页。

[0007] 非专利文档 2:X. Xiao and Y. Tao "m-invariance: Towards privacy preserving republication of dynamic datasets", Proceedings of the ACM SIGMOD International Conference on Management of Data, 2007。

[0008] 但是,例如当数据结合被重复提供时,后续添加的数据条目的属性信息可能大幅

度脱离初始假设的值的范围。

[0009] 当这些值是形成准标识符的属性时,利用传统的泛化方法难以保证 k -匿名性和应用有意义的泛化。因此,需要从目标数据移除添加的数据条目或执行具有相当高级别抽象的泛化。由此造成了信息丢失。

[0010] 还存在一个问题:每当数据集中发生改变时,适于该数据集特性的匿名化被执行,此时准标识符泛化的方法对于每个数据集而言是不同的,各个数据条目所属的分组是完全不同的,并且难以在时间序列中观察数据集的特性以及在时间序列中跟踪特定的数据条目。

[0011] 例如,图 23 示出了原始数据集。在这个数据集中,形成准标识符的属性是性别和出生地。疾病名称是敏感数据。图 24 和图 25 中所示的用于出生地的泛化规则被应用于数据集,泛化由此被执行并且获得图 26 中所示的泛化之后的数据集。如图 26 所示,泛化之后的数据集满足 $k = 2$ 匿名性以及 $I = 2$ 多样性。

[0012] 图 27 示出后续添加到图 23 所示的数据集的数据条目。后续添加的数据条目的出生地的值是“伦敦”,它是无法根据图 24 和图 25 中所示的泛化规则被泛化的值。因此,需要用于泛化该值的新泛化规则。

[0013] 图 28 到图 30 示出了新泛化规则的一个示例。当值“伦敦”根据图 28-图 30 所示的规则被泛化时,泛化之后的数据条目是图 31 所示的数据条目。但是,图 31 所示的数据条目与图 26 所示的任何一个数据条目都不共享泛化的准标识符,并且不属于现有的泛化组。因此,为了获得泛化之后满足 $k = 2$ 匿名性和 $I = 2$ 多样性的数据集,除了省略所添加的数据条目之外别无他法。

[0014] 备选地,还需要对现有的数据条目应用将所添加的数据条目纳入考虑的新泛化规则。例如,如图 32 所示,需要引入覆盖所有出生地的概念“地球”,并且引入具有高抽象水平的泛化规则。存在一个问题:当基于规则的泛化被执行以保持 $k = 2$ 匿名性和 $I = 2$ 多样性时,如图 33 所示,所有数据条目的出生地的值是“地球”并且出生地的值是没有意义的。

[0015] 备选地,如图 34 和图 35 所示,也可在“地球”级别对数据条目的仅一部分应用泛化规则。在这种情况下,如图 36 所示,可以尽可能地保持出生地的值的意义。然而存在一个问题:当最佳泛化处理每次独立执行时,类似于第八数据条目,相同条目所属的分组对于每个快照而言是不同的,并且难以在时间序列中跟踪数据集的特性。

发明内容

[0016] 在这种情况下提出了本发明,并且本发明的目的是支持属性信息的适当泛化,即使是在数据集可能被重复提供并且后续添加的数据条目的属性信息大幅偏离已知数据条目的属性信息的值范围时。

[0017] 根据本发明的一个方面的匿名化装置包括:泛化单元,被配置为基于预定泛化规则而针对数据集的每个数据条目泛化形成准标识符的至少一个属性数据的值,所述数据集具有多个数据条目,每个所述数据条目包括形成所述准标识符的所述至少一个属性数据以及除所述准标识符之外的至少一个属性数据,所述准标识符是可以标识个人的信息;条目选择单元,被配置为从所述数据集中包括的所述多个数据条目中选择在基于所述泛化规则被泛化时变成所述数据集不满足预定匿名性标准的因素的数据条目,以及至少一个数据条

目,所述至少一个数据条目的泛化目标属性数据具有与该数据条目共同的值,从而使所述数据集能够满足所述预定匿名性标准;以及条目处理单元,被配置为针对由所述条目选择单元选择的所述数据条目,将所述泛化目标属性数据的所述值改变为与所述预定泛化规则无关的预定共同值。

[0018] 在本发明中,“单元”不是简单地意味着物理设备,而是包括由软件实现的“单元”的功能。一个“单元”或装置的功能可以由两个或更多物理设备或装置实现,或者两个或多个“单元”或装置的功能可以由一个物理设备或装置实现。

[0019] 根据本发明,即使在数据集可能被重复提供并且后续添加的数据条目的属性信息大幅偏离已知数据条目的属性信息的值范围时,也可以执行属性信息的适当泛化。

附图说明

[0020] 图 1 是示出根据本发明的一个实施方式的匿名化装置的配置示例的框图。

[0021] 图 2 是示出匿名化装置中的处理流程的示例的框图。

[0022] 图 3 是示出匿名化装置中的处理流程的示例的框图。

[0023] 图 4 是示出包括出生地的值被改变的数据条目的数据集的示例的框图。

[0024] 图 5 是示出出生地的值被改变的数据条目的示例的框图。

[0025] 图 6 是示出包括性别和出生地的值被改变的数据条目的数据集的示例的框图。

[0026] 图 7 是示出性别和出生地的值被改变的数据条目的示例的框图。

[0027] 图 8 是示出添加数据条目的数据集的示例的框图。

[0028] 图 9 是示出要添加的数据条目的示例的框图。

[0029] 图 10 是示出包括性别和出生地的值被改变为原始值的数据条目的数据集的示例的框图。

[0030] 图 11 是示出要添加的数据条目的示例的框图。

[0031] 图 12 是示出在时间 T 的原始数据集的示例的框图。

[0032] 图 13 是示出在时间 T+1 的原始数据集的示例的框图。

[0033] 图 14 是示出在时间 T+2 的原始数据集的示例的框图。

[0034] 图 15 是示出在时间 T 处理的数据集的示例的框图。

[0035] 图 16 是示出在时间 T+1 处理的数据集的示例的框图。

[0036] 图 17 是示出在时间 T+2 处理的数据集的示例的框图。

[0037] 图 18 是示出匿名化装置的另一配置示例的框图。

[0038] 图 19 是示出匿名化装置的又一配置示例的框图。

[0039] 图 20 是示出匿名化装置的又一个配置示例的框图。

[0040] 图 21 是示出匿名化装置的又一个配置示例的框图。

[0041] 图 22 是示出匿名化装置的又一个配置示例的框图。

[0042] 图 23 是示出原始数据集的示例的框图。

[0043] 图 24 是示出泛化规则的示例的框图。

[0044] 图 25 是示出泛化规则的结构示例的框图。

[0045] 图 26 是示出泛化数据集的示例的框图。

[0046] 图 27 是示出要添加的数据条目的示例的框图。

- [0047] 图 28 是示出泛化规则的示例的框图。
- [0048] 图 29 是示出泛化规则的结构示例的框图。
- [0049] 图 30 是示出泛化规则的结构示例的框图。
- [0050] 图 31 是示出泛化的数据条目的示例的框图。
- [0051] 图 32 是示出泛化规则的示例的框图。
- [0052] 图 33 是示出泛化的数据集的示例的框图。
- [0053] 图 34 是示出泛化规则的示例的框图。
- [0054] 图 35 是示出泛化规则的结构示例的框图。
- [0055] 图 36 是示出被泛化的数据集的示例的框图。

具体实施方式

[0056] 下面参考图来阐释本发明的实施例。

[0057] 图 1 是示出根据本发明的一个实施例的匿名化装置的配置示例的框图。匿名化装置 10 例如是将匿名化应用于图 23 中所示的数据集的装置，该数据集具有数据条目，数据条目包括可以识别个人的属性数据。匿名化装置 10 是信息处理装置，例如应用服务器。匿名化装置 10 包括处理器、存储器、输入设备以及存储设备。

[0058] 如图 1 所示，匿名化装置 10 包括匿名化处理单元 20、数据集接收单元 22、已处理数据条目选择单元 24、数据条目处理单元 26 以及数据集输出单元 28 作为功能单元。这些功能单元通过例如执行存储在存储器中的程序的处理器而被实现。

[0059] 匿名化处理单元 20（泛化单元）对输入的数据集执行匿名化处理，例如泛化、抑制和置换，并且输出经匿名化的数据集。例如，匿名化处理单元 20 根据预定泛化规则来执行对各个数据条目中所包括的属性数据的泛化。

[0060] 例如，在图 23 所示的数据集的情况下，通过与其他背景知识结合，性别和出生地是可以识别个人的信息集合。性别和出生地形成了准标识符。针对图 23 所示的数据集中的各个数据条目，匿名化处理单元 20 例如根据图 24 和图 25 所示的泛化规则，执行对形成准标识符的属性数据中的出生地的值的泛化。

[0061] 图 26 示出通过泛化图 23 所示的数据集获得的数据集的示例。例如，在第一数据条目中，出生地“名古屋”被泛化为“东海”。其他数据条目以相同方式被泛化，从而通过准标识符而形成泛化组。例如，在泛化之后在数据集中的第一数据条目和第二数据条目中，性别是“女”并且出生地是“东海”。第一数据条目和第二数据条目形成一个泛化组。匿名化处理单元 20 将用于识别泛化组的标识符给予通过泛化形成的泛化组。

[0062] 用于泛化的方法不限于词语含义的抽象。例如，以下处理可被用于泛化：用于增加数字值等的粒度的处理，例如将年龄转换为“30 岁”或“25 到 35 岁”，或用于将例如纬度和精度的位置信息转换为一个适当范围（区域）中的数据的处理。

[0063] 参考图 26 中所示的数据集，两个或更多数据条目存在于在每个泛化组中。满足 $k = 2$ 匿名性。作为敏感数目的“疾病名称”的两种或更多值被包括在每一泛化组中。满足 $I = 2$ 多样性。在匿名化装置 10 中，针对 k 匿名、 I 多样性等设置预定匿名性标准。在这个实施例中，假设在匿名化装置 10 中， $k = 2$ 匿名性以及 $I = 2$ 多样性被设置为预定匿名性标准。

[0064] 再次参考图 1, 数据集接收单元 22 在泛化之前或泛化之后从匿名化处理单元 20 接收数据集, 并且将数据集输出到已处理数据条目选择单元 24。

[0065] 当数据条目在匿名化处理单元 20 中基于泛化规则而被泛化时, 已处理数据条目选择单元 24 从输入数据集中所包括的多个数据条目中选择使数据集无法满足预定匿名性标准的数据条目, 以及除了该数据条目之外的至少一个数据条目。当该数据条目基于泛化规则而被泛化时使数据集无法满足预定匿名性标准的数据条目例如是这样的数据条目, 该数据条目不属于数据集中的任何泛化组的, 并且当其准标识符基于泛化规则而被泛化时变成抑制的目标。所述至少一个数据条目例如是这样的多个数据条目, 其中在形成准标识符的多个属性数据中的未基于泛化规则而被泛化的属性数据的值是不同的; 或者是使数据集满足预定匿名性标准的至少一个数据条目, 即使该数据条目被从所述数据集被移除。下面参考具体的示例阐释细节。

[0066] 针对已处理数据条目选择单元 14 所选择的数据条目, 数据条目处理单元 26 将泛化目标属性数据的值改变为预定的共同值, 并且将该值经由数据集输出单元 28 输出到匿名化处理单元 20。例如, 数据条目处理单元 26 可以将所选数据条目的出生地的值改变为“*”。改变后的预定共同值例如可以是具有属性数据可采用的最高抽象水平的值。例如, 对于出生地, 改变后的预定共同值可以是“地球”。

[0067] 图 2 和图 3 是示出匿名化装置 10 中的处理的流程的示例。如图 2 所示, 对数据集的处理可以在匿名化处理单元 20 中在泛化之前被应用于数据集。如图 3 所示, 还可以在匿名化处理单元 20 中在泛化之后将所述处理应用于数据集。所述处理可以被在匿名化中途执行多次, 例如, 对数据集的处理可以被分开执行两次, 即, 在泛化之前和在泛化之后。

[0068] 在这个实施例所阐释的示例中, 处理在泛化之后被应用于数据集。首先, 假设在匿名化处理单元 20 中设置有图 24 所示的泛化规则。当图 23 所示的数据集被输入到匿名化处理单元 20 中时, 出生地的值基于泛化规则而被泛化并且获得图 26 所示的数据集。如上所述, 图 26 所示的数据集满足匿名化装置 10 中的匿名性标准。在这个状态的前提下, 下面描述关于数据处理的示例。

[0069] < 数据处理示例 1 >

[0070] 假设, 在获得图 26 所示数据集之后, 图 27 所示的数据条目被输入到匿名化处理单元 20, 以作为对数据集的附加条目。图 27 所示的数据条目的出生地是“伦敦”, 它无法根据图 24 所示的泛化规则而被泛化。因此, 当添加数据条目时, 不满足匿名化的标准。于是, 匿名化处理单元 20 向数据集接收单元 22 输出由图 26 所示的泛化之后的数据集和图 27 所示的数据条目形成的数据集。

[0071] 数据集接收单元 22 从匿名化处理单元 20 接收数据集并且将数据集输出到已处理数据条目选择单元 24。

[0072] 已处理数据条目选择单元 24 从数据集中所包括的多个数据条目中选择在数据集基于泛化规则而被泛化时使数据集无法满足匿名化的标准的数据条目, 以及具有形成准标识符的多个属性数据中的未基于泛化规则而被泛化的属性数据的不同值的多个数据条目。当数据条目基于泛化规则而被泛化时使数据集无法满足匿名化的标准的数据条目是图 27 中所示的数据条目。具有形成准标识符的多个属性数据中的、未基于泛化规则而被泛化的属性数据的不同值的多个数据条目的示例是图 4 中由虚线围绕的数据条目。换言之, 在形

成准标识符的属性数据中未被泛化的属性数据是性别。具有性别不同值的多个数据条目被选择。在图 4 所示的示例中,性别为“女”并且泛化组为“1”的数据条目以及性别为“男”并且泛化组为“4”的数据条目被选择。

[0073] 数据条目处理单元 26 将已处理数据条目选择单元 24 所选择的数据条目的出生地的值改变为例如图 4 和 5 所示的“*”。图 26 所示的对数据集的处理可以在图 27 所示的数据条目被添加之前预先执行,或者可以在图 27 所示的数据条目被添加的时刻被执行。

[0074] 数据集输出单元 28 将由数据条目处理单元 26 处理的数据集输出到匿名化处理单元 20。

[0075] 如图 4 和图 5 所示,根据数据条目处理单元 26 的数据处理,图 5 所示的数据条目的准标识符与图 4 所示数据集中的泛化组为“1”的准标识符相同。因此,在匿名化处理单元 20 中,例如,“1”被给予图 5 所示的数据条目的泛化组。由此,可以允许数据集满足匿名性标准,而无需省略所添加的数据条目也不用将出生地泛化为无意义的水平。

[0076] 换言之,当泛化被应用于数据集时,有意地向可被应用以较为具体的泛化规则的一部分数据条目应用较为抽象的泛化规则。由此,不管要添加的数据条目后续采用什么类型的值,可以在将该数据条目添加到泛化的数据集的同时保持匿名性标准。

[0077] 如图 4 所示,已处理数据条目选择单元 24 在泛化组单元中选择数据条目。由此,因为每个泛化组的数据条目的数目没有减少,可以防止不满足匿名性标准。

[0078] < 数据处理示例 2 >

[0079] 在这个示例中,如在上述示例中一样,假设在获得图 26 所示的数据集之后,图 27 中所示的数据条目被输入到匿名化处理单元 20 以作为数据集的附加条目。

[0080] 数据集接收单元 22 从匿名化处理单元 20 接收数据集,并且将数据集输出到已处理数据条目选择单元 24。

[0081] 已处理数据条目选择单元 24 从数据集中所包括的多个数据条目中选择这样的数据条目,当该数据条目基于泛化规则而被泛化时使数据集无法满足预定匿名性标准,以及使数据集满足匿名标准的至少一个数据条目,即使该数据条目被排除在数据集之外。使数据集满足预定匿名性标准(即使该数据条目被排除在数据集之外)的至少一个数据条目的示例是被图 6 中虚线环绕的数据条目。如图 6 所示,第三到第五数据条目的泛化组是“2”。在这些数据条目中,即使第五数据条目被排除,匿名性标准仍被第三和第四数据条目满足。类似地,在第六到第八数据条目中,其泛化组是“3”,即使第八数据条目被排除,也满足匿名性标准。

[0082] 数据条目处理单元 26 将已处理数据条目选择单元 24 所选择的数据条目的性别和出生地的值改变为例如图 6 和 7 所示的“*”。数据条目处理单元 26 可以分别将性别和出生地的值分别改变为其他预定的共同值。

[0083] 数据集输出单元 28 将数据条目处理单元 26 处理的数据集输出到匿名化处理单元 20。

[0084] 如图 6 和图 7 所示,根据数据条目处理单元 26 的数据处理,图 7 所示的数据条目的准标识符与从图 6 所示的数据集中选择的数据条目的准标识符相同。因此,在匿名化处理单元 20 中,“5”被给予这些数据条目。由此,可以允许数据集满足匿名性标准,而不用省略添加的数据条目也无需将出生地泛化为无意义的水平。

[0085] 数据条目处理单元 26 可以仅将作为泛化目标属性数据的出生地的值改变为“*”。但是,通过将作为准标识符中所包括的其他属性数据的性别的值也改变为“*”,可以增加由已处理数据条目和所添加数据条目形成新泛化组的可能性。

[0086] < 数据处理示例 3>

[0087] 此示例是一个这样的示例:数据条目还进一步添加到数据处理示例 2 所处理的数据集。图 8 示出了数据处理示例 2 所处理的数据集。在这个示例中,还应用图 28 所示的泛化规则“欧洲”。特别地,如图 8 所示,第十一条目的出生地的改变之前的值是通过根据图 28 中所示的泛化规则来泛化“伦敦”而得到的“欧洲”,“伦敦”是图 27 所示的数据条目的出生地的值。

[0088] 假设在获得图 8 所示的数据集之后,图 9 所示的数据条目被输入到匿名化处理单元 20 以作为对数据集的附加条目。图 9 所示的数据条目的出生地的值是“巴黎”。当该数据条目被匿名化处理单元 20 泛化时,出生地的值被泛化为“欧洲”并且匿名性标准未被满足。因此,匿名化处理单元 20 向数据集接收单元 22 输出由图 8 中所示的泛化之后的数据集和图 9 中所示的数据条目所形成的数据集。

[0089] 数据集接收单元 22 从匿名化处理单元 20 接收数据集,并且将数据集输出到已处理数据条目选择单元 24。

[0090] 已处理数据条目选择单元 24 从数据集中所包括的多个数据条目中选择所添加的条目,即,当该数据条目基于泛化规则而被泛化时使得数据集无法满足预定匿名性标准的数据条目,以及如下数据条目,如果该数据条目的属性数据的值被返回成处理之前的值,则该数据条目与所添加的条目一起形成泛化组,由此满足匿名性标准。

[0091] 参考图 8 所示的数据集,在第 11 数据条目中,性别和出生地的处理之前的值分别是“女”和“欧洲”,并且敏感数据的值是“消化不良”。属性数据处理之前的值例如可以通过根据泛化规则来泛化之前的数据条目的属性数据而获得。例如,可以提供存储单元,其独立于泛化之前的数据而存储与其属性数据被处理的条目相关联的处理之前的属性数据的值。

[0092] 在图 9 所示的数据条目中,性别和出生地的值分别是“女”和“欧洲”,并且敏感数据的值是“支气管炎”。特别地,如果图 8 所示的数据集中的第十一数据条目的性别和出生地的值分别被返回成处理之前的“女”和“欧洲”,并且图 9 所示的数据条目的出生地被泛化,由这两个数据条目形成了满足匿名性标准的新泛化组。

[0093] 因此,数据条目处理单元 26 将由已处理数据条目选择单元 24 选择的图 8 所示的数据集中的第十一数据条目的性别和出生地的值分别改变为处理之前的“女”和“欧洲”。图 10 示出了已处理的数据集。

[0094] 数据集输出单元 28 将数据条目处理单元 26 处理的数据集输出到匿名化处理单元 20。

[0095] 当图 9 所示的数据条目在匿名化处理单元 20 中被泛化时,获得图 11 所示的数据条目。新泛化组由该数据条目和图 10 所示的数据集的第十一数据条目形成。换言之,例如,“6”被给予这些数据条目的泛化组。

[0096] < 数据处理示例 4>

[0097] 在此示例中描述执行添加和删除数据条目的示例。图 12- 图 14 示出匿名化之前在时间 T 到 T+2 的数据集的示例。

[0098] 首先,通过向图 12 中所示的数据集应用对出生地值的泛化以及按照与数据处理示例 1 相同的方式执行数据处理而获得的数据集被显示在图 15 中。

[0099] 假设原始数据在时间 T+1 改变,如图 13 所示。特别地,在时间 T 的原始数据集的“千代”、“洋子”、“正”和“三郎”的数据条目被删除,并且“爱丽丝”的数据条目被添加。在这种情况下,如图 16 所示,对于“爱丽丝”的数据条目,如在数据处理示例 1 的情况下,出生地的值被改变为“*”。“爱丽丝”的数据条目在同一泛化组中设置为“花子”。但是,因为“千代”的数据条目被删除并且“花子”和“爱丽丝”二者的数据条目的“疾病名称”都是“消化不良”,因此不满足匿名性标准。因此,数据条目处理单元 26 添加伪条目数据,其“疾病名称”是“支气管炎”,如图 16 所示,使得匿名性标准得以满足。

[0100] 此外,假设原始数据集在时间 T+2 如图 14 所示那样改变。特别地,“索菲”的数据条目被添加到时间 T+1 的原始数据。在这种情况下,如图 17 所示,类似“爱丽丝”的数据条目,对于“索菲”的数据条目,出生地的值被改变为“*”。“索菲”的数据条目与“花子”和“爱丽丝”设置在同一泛化组中。因为“索菲”的数据条目的“疾病名称”是“支气管炎”,因此即使图 16 所示的伪数据条目被移除,仍然满足匿名性标准。因此,如图 17 所示,伪数据条目被数据条目处理单元 26 删除。

[0101] 如上所述,利用此实施例中的匿名化装置 10,甚至当数据集可能被重复提供并且后来添加的数据条目的属性信息答复偏离已知数据条目的属性信息的值范围时,仍然可以执行对属性信息的适当泛化

[0102] 这个实施例是用于帮助理解本发明的实施例并且不是要限制地解释本发明。本发明可以被改变/改进而不背离本发明的精神。本发明的等效项包括在本发明中。

[0103] 例如,如图 18 所示,匿名化装置 10 可以包括已处理数据条目选择规则输入单元 30。换言之,用于选择数据条目的规则不必固定在已处理数据条目选择单元 24 中,而是可以根据来自已处理数据条目选择规则输入单元 30 的输入而改变。

[0104] 例如,如图 19 所示,匿名化装置 10 可以包括数据条目处理规则输入单元 32。换言之,用于处理数据条目的规则不必固定在数据条目处理单元 26 中,而是可以根据来自数据条目处理规则输入单元 32 的输入而改变。

[0105] 此外,例如如图 20 所示,匿名化装置 10 可以包括已处理数据条目选择规则输入单元 30 和数据条目处理规则输入单元 32。

[0106] 如图 21 所示,匿名化装置 10 可以包括匿名评估单元 34,用于由匿名化处理单元 20 评估通过匿名化产生的匿名化数据集的匿名性。在这种情况下,匿名评估单元 34 可以基于匿名性的评估结果来控制已处理数据条目确定规则输入单元 30 和数据条目处理规则输入单元 32,使得匿名性满足预定标准。

[0107] 如图 22 所示,匿名化装置 10 可以包括匿名化规则输入单元 36。特别地,将匿名化处理应用于数据条目的规则可以根据来自匿名化规则输入单元 36 的输入而改变,而不是被固定在匿名化处理单元 20 中。例如,当匿名化处理单元 20 没有图 28 和图 29 所示的“欧洲”的匿名化规则时,可以通过使用匿名化规则输入单元 36 来添加匿名化规则。

[0108] 本申请要求基于 2010 年 11 月 9 日提交的日本专利申请号 2010-250600 的优先权,在此并入其全部内容。

[0109] 以上参考实施例描述了本发明。但是本发明不限于实施例。可以在本发明的范围

中对本发明的配置和细节做出本领域技术人员可理解的各种改变。

[0110] 一部分实施例或全部实施例也可以描述为下列注解所指示的。但是,本发明不限于以下。

[0111] (注解 1) 一种匿名化装置,包括:泛化单元,被配置为基于预定泛化规则而针对数据集的每个数据条目泛化形成准标识符的至少一个属性数据的值,所述数据集具有多个数据条目,每个所述数据条目包括形成所述准标识符的所述至少一个属性数据以及除所述准标识符之外的至少一个属性数据,所述准标识符是可以标识个人的信息;条目选择单元,被配置为从所述数据集中包括的所述多个数据条目中选择在基于所述泛化规则被泛化时变成所述数据集不满足预定匿名性标准的因素的数据条目,以及至少一个数据条目,所述至少一个数据条目的泛化目标属性数据具有与该数据条目共同的值,从而使所述数据集能够满足所述预定匿名性标准;以及条目处理单元,被配置为针对由所述条目选择单元选择的所述数据条目,将所述泛化目标属性数据的所述值改变为与所述预定泛化规则无关的预定共同值。

[0112] (注解 2) 根据注解 1 所述的匿名化装置,其中所述条目选择单元从所述数据集中所包括的所述多个数据条目中选择如下数据条目,当所述数据条目基于所述泛化规则而被泛化时使所述数据集无法满足所述预定匿名性标准,以及具有在形成所述准标识符的所述至少一个属性数据中的、未基于所述泛化规则而被泛化的属性数据的不同值的多个数据条目。

[0113] (注解 3) 根据注解 1 所述的匿名化装置,其中所述条目选择单元从所述数据集中所包括的所述多个数据条目中选择如下数据条目,当所述数据条目基于所述泛化规则而被泛化时使所述数据集无法满足所述预定匿名性标准,以及至少一个数据条目,所述至少一个数据条目使所述数据集满足所述预定匿名性标准,即使此至少一个数据条目被排除在所述数据集之外。

[0114] (注解 4) 根据注解 3 所述的匿名化装置,其中所述条目处理单元将所述泛化目标属性数据的所述值改变为所述预定共同值,以允许所述数据集满足所述预定匿名性标准,并且将形成所述准标识符的所述至少一个属性数据中的、除所述泛化目标属性数据之外的至少一个属性数据的值改变为预定共同值。

[0115] (注解 5) 根据注解 1 到 4 中任一项所述的匿名化装置,其中,当数据条目被新添加到所述数据集时,如果当添加的所述数据条目的值和所述数据条目中的、其属性数据的值已被改变的至少一个数据条目的改变之前的值基于所述泛化规则而被泛化时所述数据集满足所述预定匿名性标准,则所述条目处理单元针对此至少一个数据条目而将所述属性数据的值改变为通过基于所述泛化规则来泛化所述改变之前的所述值而获得的值。

[0116] (注解 6) 根据注解 1 到 5 中任一项所述的匿名化装置,其中,如果所述数据集由于从所述数据集删除至少一个数据条目而无法满足所述预定匿名性标准,所述条目处理单元向所述数据集添加伪数据条目以允许所述数据集满足所述预定匿名性标准。

[0117] (注解 7) 根据注解 6 所述的匿名化装置,其中,如果所述数据集由于向所述数据集新添加数据条目而即使将所述伪数据条目排除也满足所述预定匿名性标准,所述伪数据条目从所述数据集被删除。

[0118] (注解 8) 根据注解 1 到 7 中任一项所述的匿名化装置,还包括条目选择规则输入

单元,被配置为输入由所述条目选择单元执行的所述数据条目的选择的规则,其中所述条目选择单元基于从所述条目选择规则输入单元输入的所述规则来选择所述数据条目。

[0119] (注解 9) 根据注解 1 到 8 中任一项所述的匿名化装置,还包括条目处理规则输入单元,被配置为输入由所述条目处理单元执行的所述数据条目的处理的规则,其中所述条目处理单元基于从所述条目处理规则输入单元输入的所述规则来处理所述数据条目。

[0120] (注解 10) 根据注解 1 到 8 中任一项所述的匿名化装置,还包括泛化规则输入单元,被配置为输入由所述泛化单元执行的所述数据条目的泛化的规则,其中所述泛化单元基于从所述泛化规则输入单元输入的所述规则来泛化所述数据条目。

[0121] 10 匿名化装置

[0122] 20 匿名化处理单元

[0123] 22 数据集接收单元

[0124] 24 已处理数据条目选择单元

[0125] 26 数据条目处理单元

[0126] 28 数据集输出单元

[0127] 30 处理的数据条目选择规则输入单元

[0128] 32 数据条目选择规则输入单元

[0129] 34 匿名评估单元

[0130] 36 泛化规则输入单元

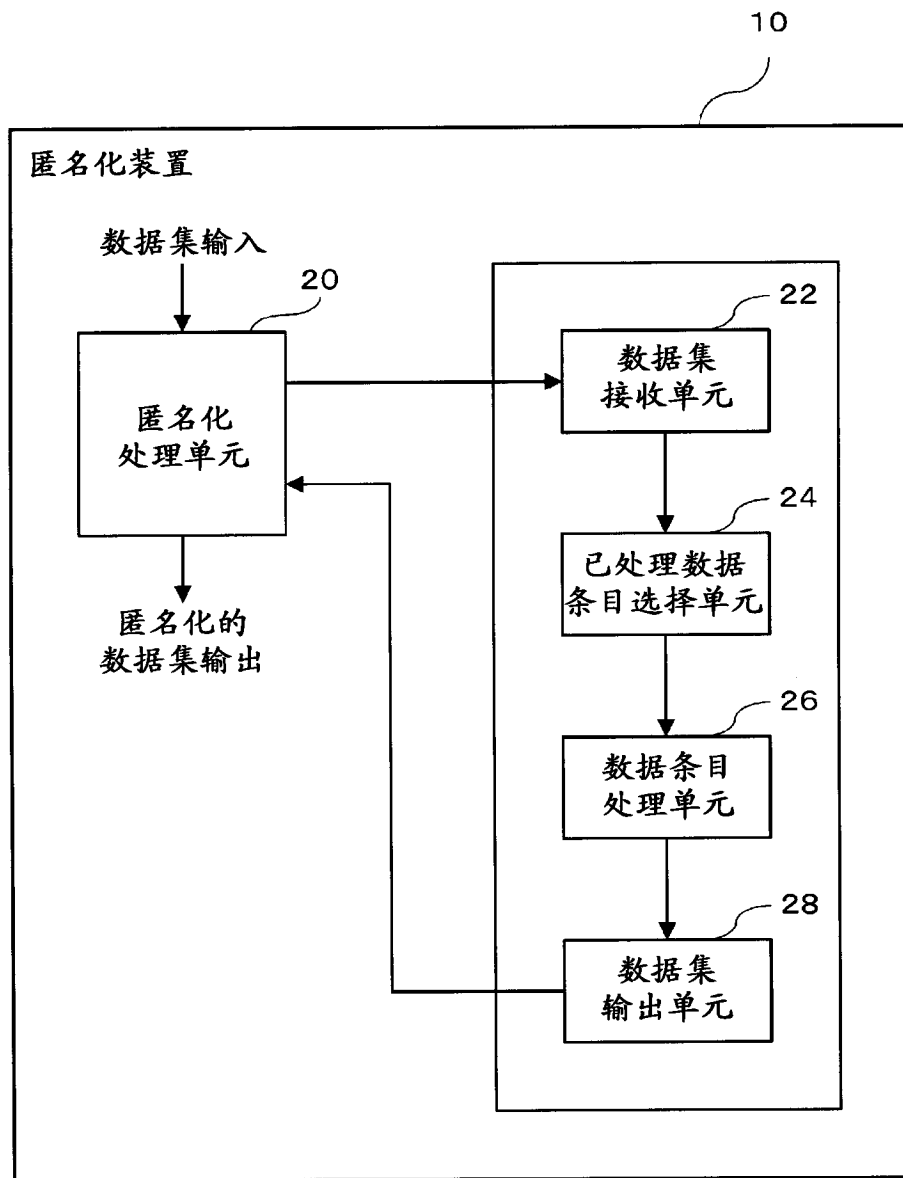


图 1

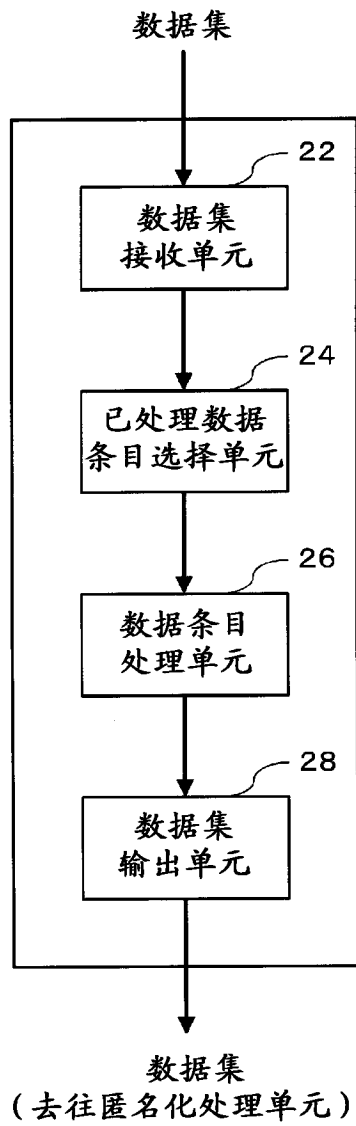


图 2

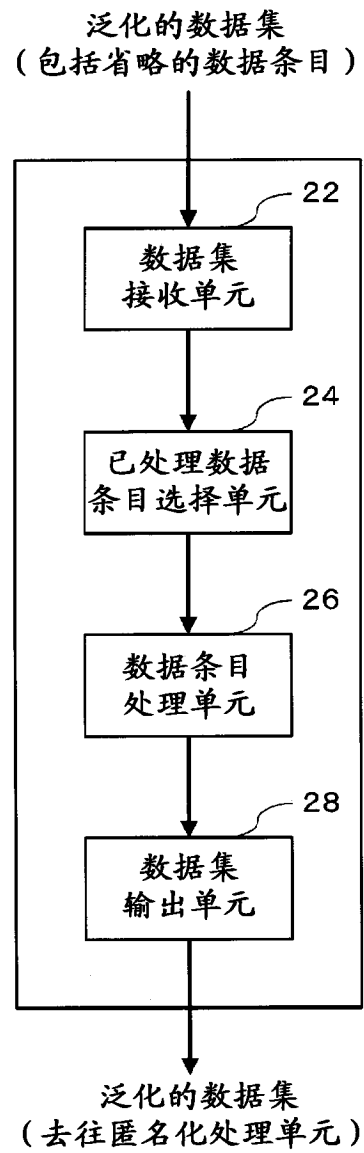


图 3

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
1	女性	东海 *	消化不良
1	女性	东海 *	支气管炎
2	男性	东海	湿疹
2	男性	东海	消化不良
2	男性	东海	湿疹
3	女性	关东	湿疹
3	女性	关东	消化不良
3	女性	关东	消化不良
4	男性	关东 *	消化不良
4	男性	关东 *	湿疹

图 4

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
1	女性	*	消化不良

图 5

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
1	女性	东海	消化不良
1	女性	东海	支气管炎
2	男性	东海	湿疹
2	男性	东海	消化不良
2 5	男性*	东海*	湿疹
3	女性	关东	湿疹
3	女性	关东	消化不良
3 5	女性*	关东*	消化不良
4	男性	关东	消化不良
4	男性	关东	湿疹

图 6

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
5	女性*	欧洲*	消化不良

图 7

准标识符			敏感数据
泛化组	性别	出生地	疾病名称
1	女性	东海	消化不良
1	女性	东海	支气管炎
2	男性	东海	湿疹
2	男性	东海	消化不良
2 5	男性*	东海 *	湿疹
3	女性	关东	湿疹
3	女性	关东	消化不良
3 5	女性*	关东 *	消化不良
4	男性	关东	消化不良
4	男性	关东	湿疹
5	女性*	欧洲 *	消化不良

图 8

准标识符			敏感数据
泛化组	性别	出生地	疾病名称
索菲	女性	巴黎	支气管炎

图 9

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
1	女性	东海	消化不良
1	女性	东海	支气管炎
2	男性	东海	湿疹
2	男性	东海	消化不良
2 5	男性*	东海 *	湿疹
3	女性	关东	湿疹
3	女性	关东	消化不良
3 5	女性*	关东 *	消化不良
4	男性	关东	消化不良
4	男性	关东	湿疹
5 6	女性	欧洲	消化不良

图 10

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
6	女性	欧洲	支气管炎

图 11

时间 T	准标识符		敏感数据
	姓名	性别	出生地
	花子	女性	名古屋
	千代	女性	岐阜
	诚	男性	丰田
	一郎	男性	静岡
	次郎	男性	岐阜
	幸子	女性	东京
	由纪子	女性	横滨
	阳子	女性	千叶
	正	男性	川崎
	三郎	男性	埼玉
			疾病名称
			消化不良
			支气管炎
			湿疹
			消化不良
			湿疹
			湿疹
			消化不良
			消化不良
			消化不良
			湿疹

图 12

时间 T+1	准标识符		敏感数据
	姓名	性别	出生地
	花子	女性	名古屋
	诚	男性	丰田
	一郎	男性	静冈
	次郎	男性	岐阜
	幸子	女性	东京
	由纪子	女性	横滨
	爱丽丝	女性	伦敦

图 13

时间 T+2	准标识符		敏感数据
	姓名	性别	出生地
	花子	女性	名古屋
	诚	男性	丰田
	一郎	男性	静冈
	次郎	男性	岐阜
	幸子	女性	东京
	由纪子	女性	横滨
	爱丽丝	女性	伦敦
	索菲	女性	巴黎

图 14

时间 T		准标识符		敏感数据
姓名	泛化组	性别	出生地	疾病名称
花子	1	女性	*	消化不良
千代	1	女性	*	支气管炎
诚	2	男性	东海	湿疹
一郎	2	男性	东海	消化不良
次郎	2	男性	东海	湿疹
幸子	3	女性	关东	湿疹
由纪子	3	女性	关东	消化不良
阳子	3	女性	关东	消化不良
正	4	男性	*	消化不良
三郎	4	男性	*	湿疹

图 15

时间 T+1		准标识符		敏感数据
姓名	泛化组	性别	出生地	疾病名称
花子	1	女性	*	消化不良
<u>伪条目</u>	1	女性	*	支气管炎
诚	2	男性	东海	湿疹
一郎	2	男性	东海	消化不良
次郎	2	男性	东海	湿疹
幸子	3	女性	关东	湿疹
由纪子	3	女性	关东	消化不良
爱丽丝	1	女性	*	消化不良

图 16

时间 T+2		准标识符		敏感数据
姓名	泛化组	性别	出生地	疾病名称
花子	1	女性	*	消化不良
诚	2	男性	东海	湿疹
一郎	2	男性	东海	消化不良
次郎	2	男性	东海	湿疹
幸子	3	女性	关东	湿疹
由纪子	3	女性	关东	消化不良
爱丽丝	1	女性	*	消化不良
索菲	1	女性	*	支气管炎

图 17

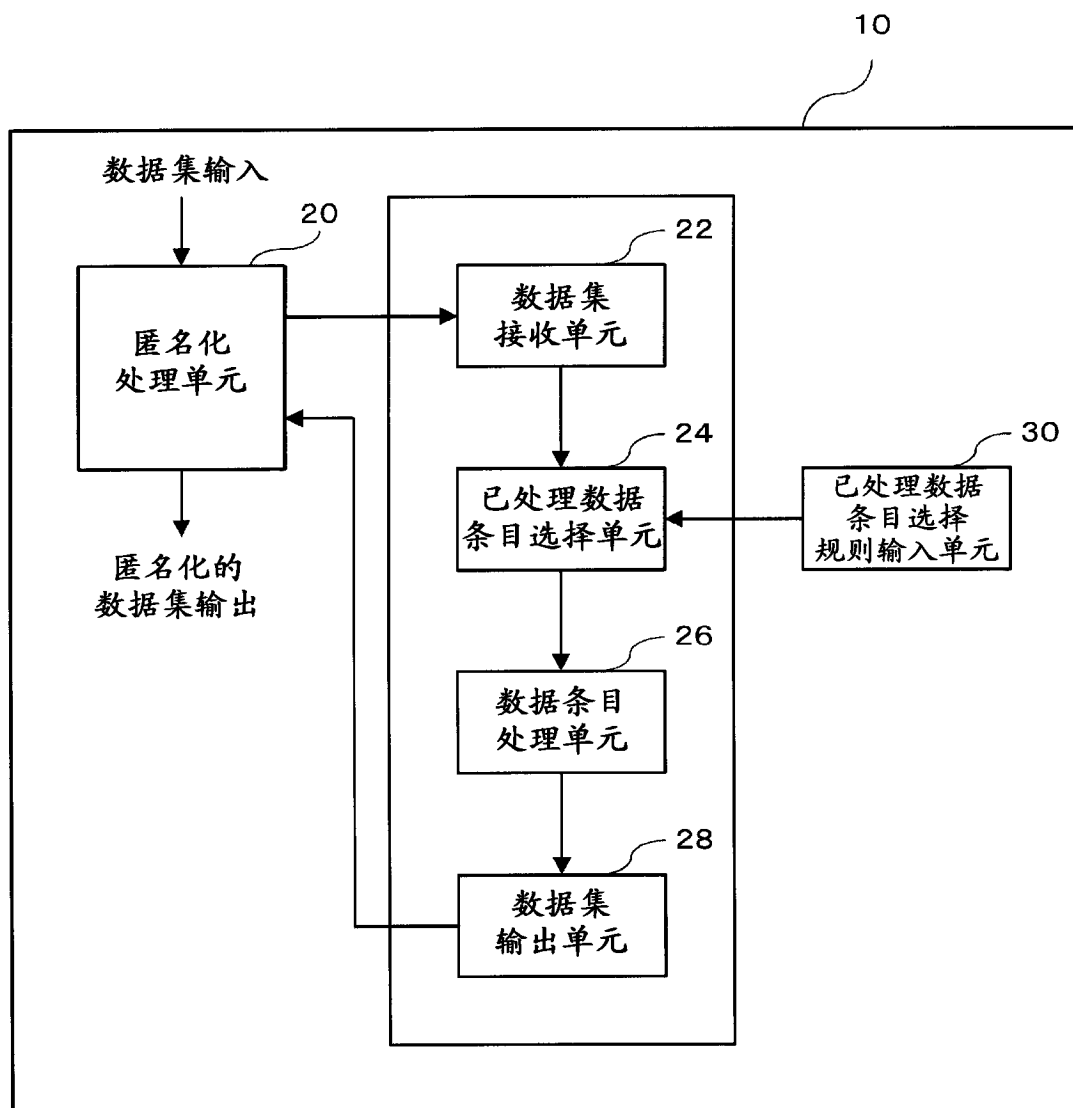


图 18

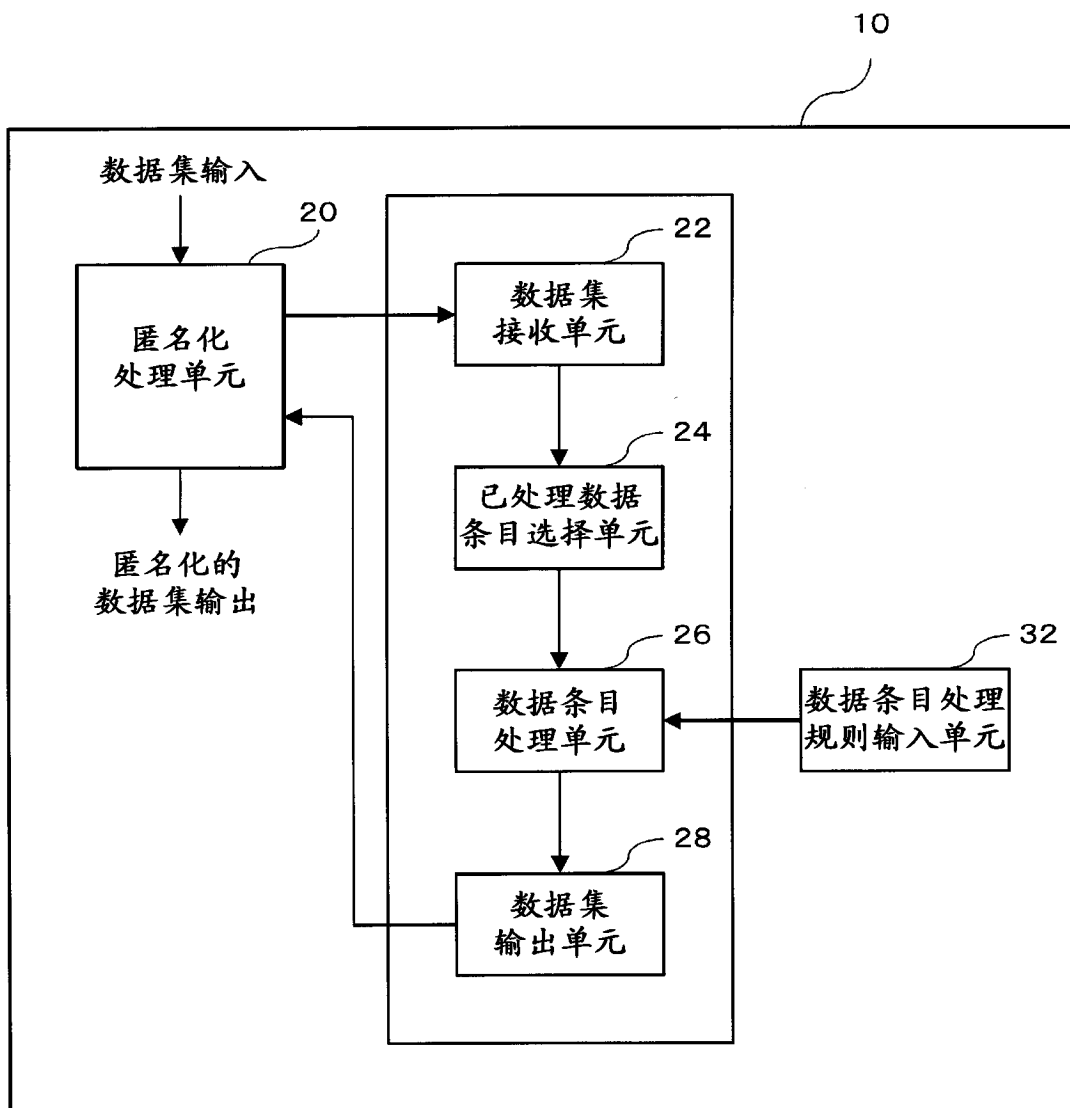


图 19

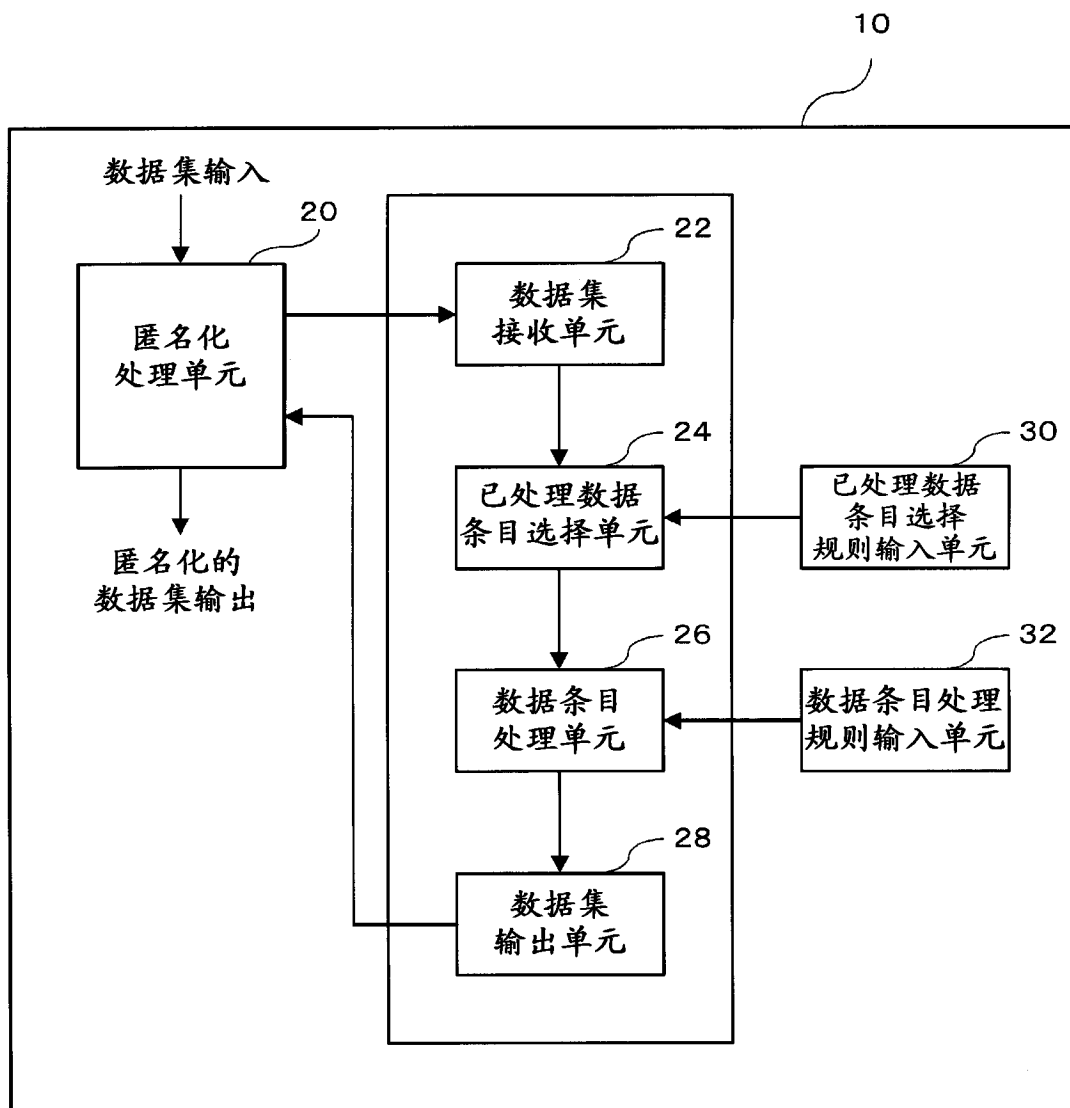


图 20

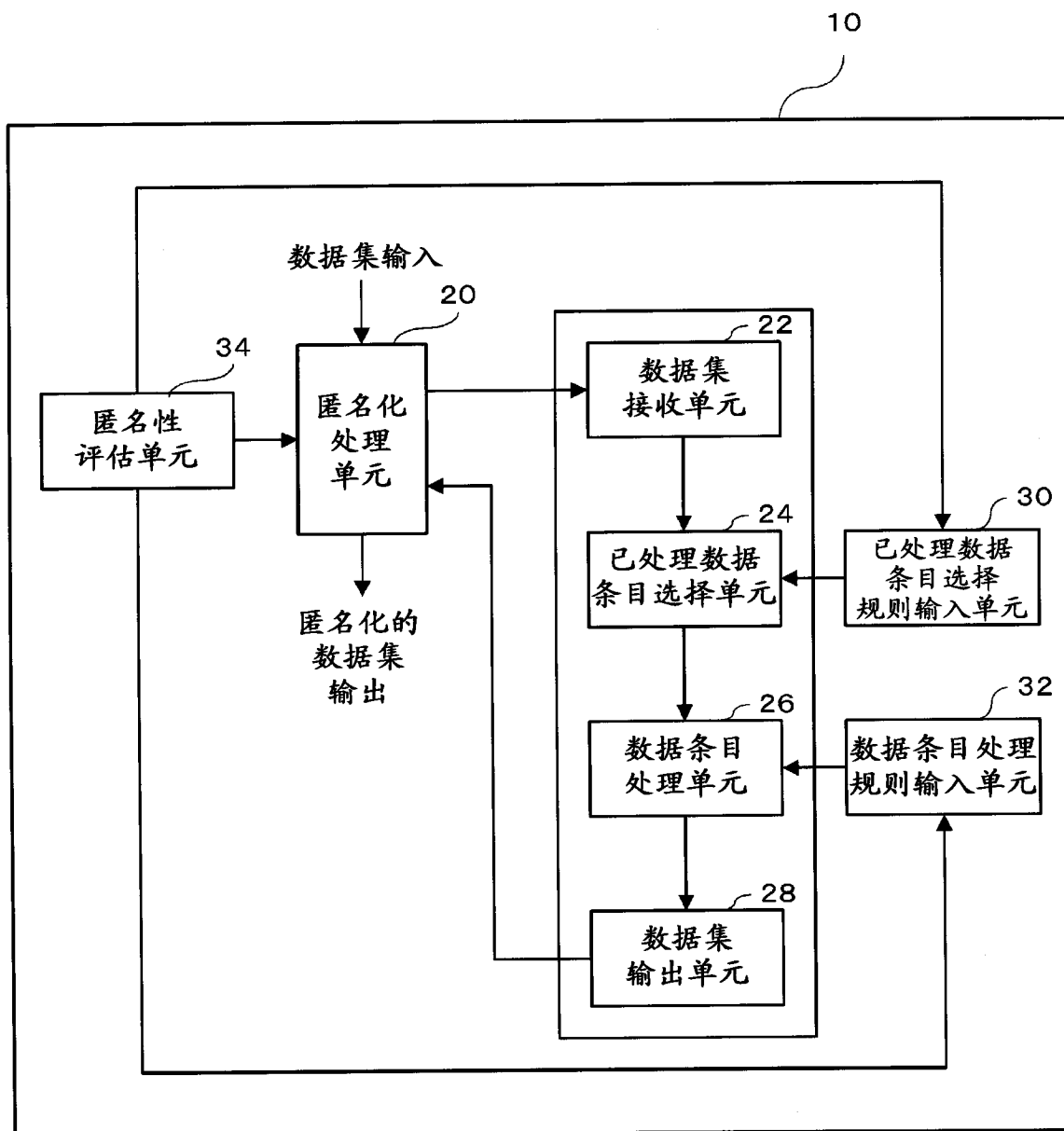


图 21

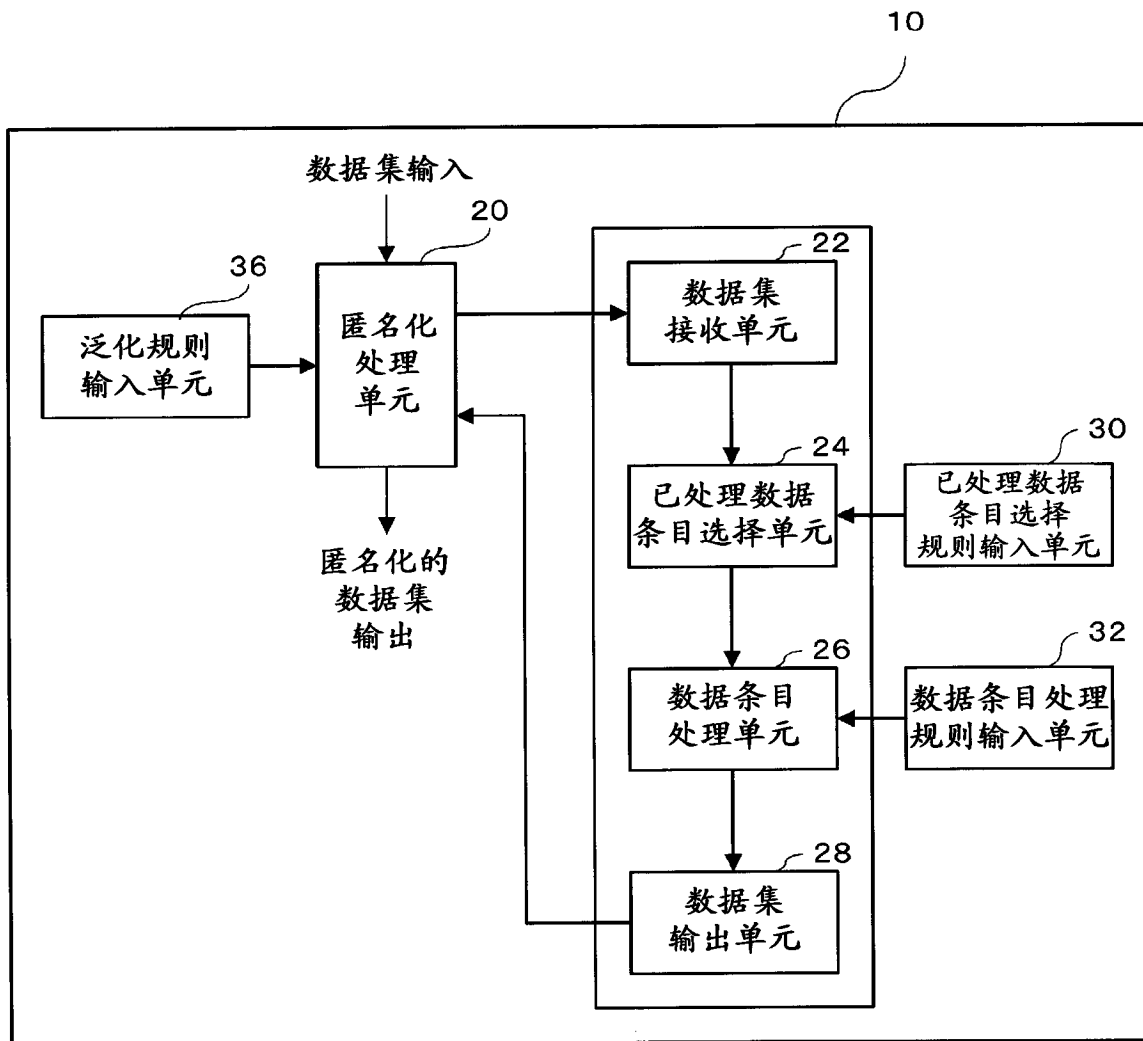


图 22

准标识符			敏感数据
姓名	性别	出生地	疾病名称
花子	女性	名古屋	消化不良
千代	女性	岐阜	支气管炎
诚	男性	丰田	湿疹
一郎	男性	静冈	消化不良
次郎	男性	岐阜	湿疹
幸子	女性	东京	湿疹
由纪子	女性	横滨	消化不良
阳子	女性	千葉	消化不良
正	男性	川崎	消化不良
三郎	男性	埼玉	湿疹

图 23

出生地	泛化之后的出生地
名古屋	东海
岐阜	东海
丰田	东海
静冈	东海
东京	关东
横滨	关东
千葉	关东
川崎	关东
埼玉	关东

图 24

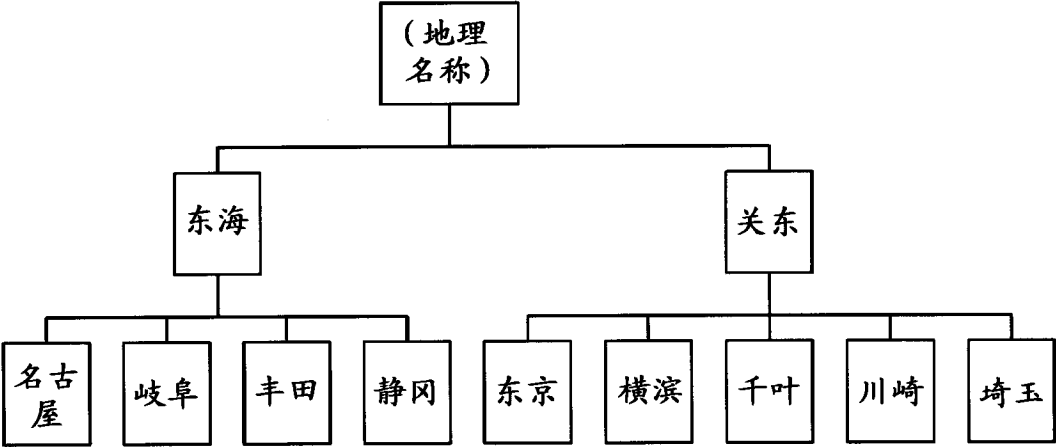


图 25

准标识符		敏感数据	
泛化组	性别	出生地	疾病名称
1	女性	东海	消化不良
1	女性	东海	支气管炎
2	男性	东海	湿疹
2	男性	东海	消化不良
2	男性	东海	湿疹
3	女性	关东	湿疹
3	女性	关东	消化不良
3	女性	关东	消化不良
4	男性	关东	消化不良
4	男性	关东	湿疹

图 26

准标识符

敏感数据

姓名	性别	出生地	疾病名称
爱丽丝	女性	伦敦	消化不良

图 27

出生地	泛化之后的出生地
伦敦	欧洲
巴黎	欧洲
柏林	欧洲

图 28

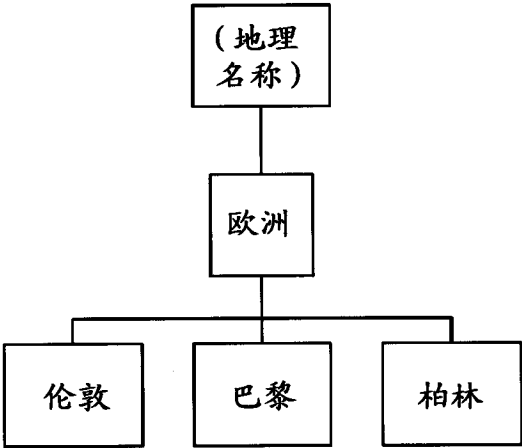


图 29

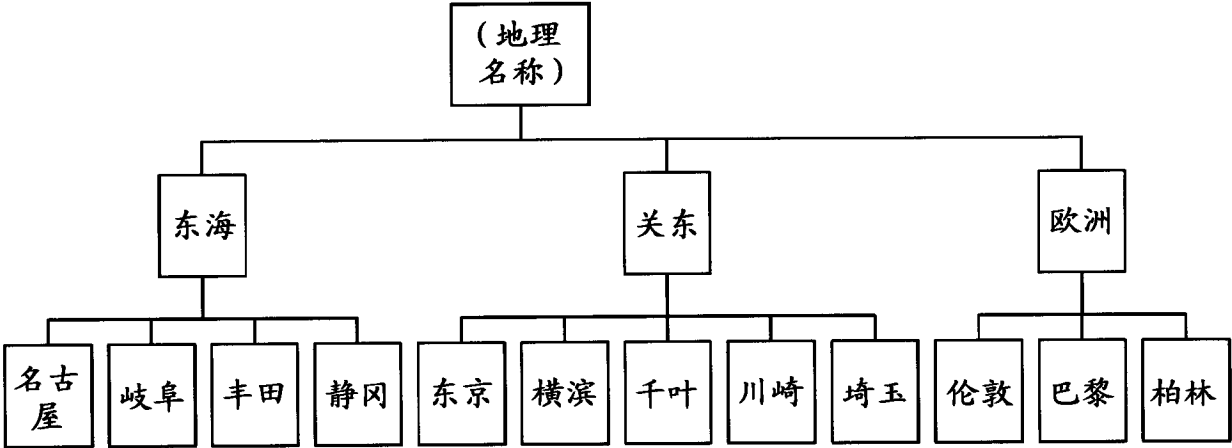


图 30

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
?	女性	欧洲	消化不良

图 31

出生地	泛化之后的出生地
名古屋	地球
岐阜	地球
丰田	地球
静岡	地球
东京	地球
横滨	地球
千叶	地球
川崎	地球
埼玉	地球
伦敦	地球

图 32

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
1	女性	地球	消化不良
1	女性	地球	支气管炎
2	男性	地球	湿疹
2	男性	地球	消化不良
2	男性	地球	湿疹
1	女性	地球	湿疹
1	女性	地球	消化不良
1	女性	地球	消化不良
2	男性	地球	消化不良
2	男性	地球	湿疹
1	女性	地球	消化不良

图 33

出生地	泛化之后的出生地
名古屋	东海
岐阜	东海
丰田	东海
静岡	东海
东京	关东
横滨	关东
千葉	地球
川崎	关东
埼玉	关东
伦敦	地球
巴黎	地球
柏林	地球

图 34

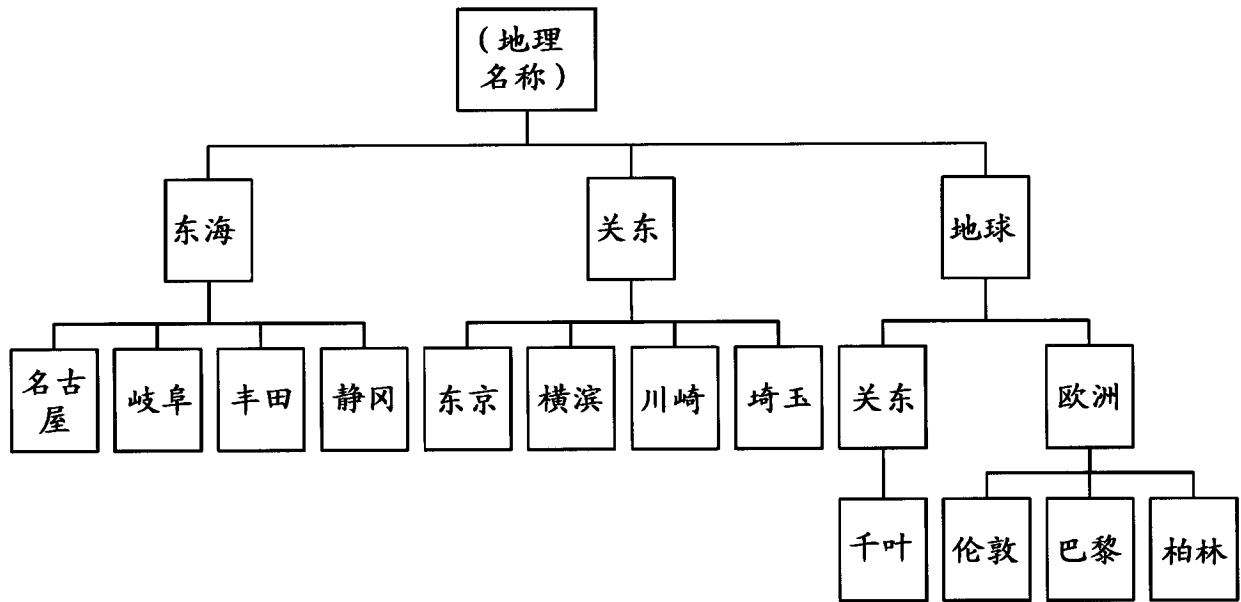


图 35

准标识符

敏感数据

泛化组	性别	出生地	疾病名称
1	女性	东海	消化不良
1	女性	东海	支气管炎
2	男性	东海	湿疹
2	男性	东海	消化不良
2	男性	东海	湿疹
3	女性	关东	湿疹
3	女性	关东	消化不良
5	女性	地球	消化不良
3	男性	关东	消化不良
4	男性	关东	湿疹
5	女性	地球	消化不良

图 36