[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/30

[12] 发明专利说明书

专利号 ZL 00812084.6

[45] 授权公告日 2005 年 10 月 5 日

[11] 授权公告号 CN 1221911C

[22] 申请日 2000.7.14 [21] 申请号 00812084.6

[30] 优先权

[32] 1999. 7.20 [33] US [31] 09/357,301

[32] 1999.10. 6 [33] US [31] 09/412,970

[32] 2000. 7.14 [33] US [31] 09/617,047

[86] 国际申请 PCT/US2000/019195 2000.7.14

[87] 国际公布 WO2001/006414 英 2001.1.25

[85] 进入国家阶段日期 2002.2.26

[71] 专利权人 普瑞曼迪亚公司

地址 美国宾夕法尼亚

[72] 发明人 布约恩·J·格鲁恩沃尔德

审查员 刘 栩

[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所

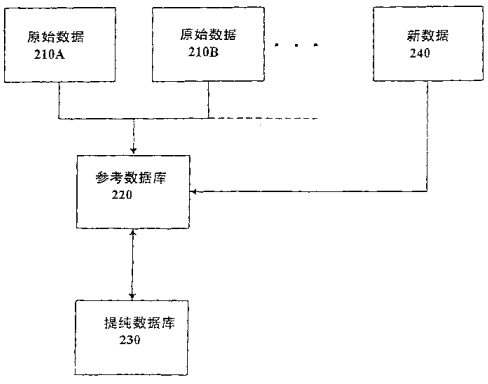
代理人 李 勇

权利要求书 3 页 说明书 31 页 附图 12 页

[54] 发明名称 用于组织数据的系统与方法

[57] 摘要

一种系统与方法，用于使用一种改进的机制来识别数据库中字段(如列)之间的重复数据，以组织来自一个或多个来源的原始数据。 字段可以是在一个单一数据库之内的类似字段，或者一对数据库之内相似或相同的字段，并且可以被组织为数组或字段向量。 本发明将每个字段向量排序，并且，如果需要的话，用通用值对其分区。 识别字段向量之间的重复数据所需要进行的比较的数目由于反馈所比较的值之间的差而得到减少。 这个差值被用于将索引调整到用于后续比较的字段向量。



ISSN1008-4274

1. 一种可在计算机上操作、用于将至少一个可由计算机访问的原始数据库的信息转换为可由计算机访问的提取数据库的方法，所述原始数据库包括多个原始记录，所述多个原始记录中的每个记录包括至少一个原始数据字段，每个原始数据字段包括多个数据元素，该方法包括以下步骤：

将一个原始记录中的至少一个原始数据字段转换为一个数字值，通过将该原始数据字段中的多个数据元素中的每个数据元素表示为一个数码系统中的一个数字，来生成所述数字值，根据该数据字段中的多个数据元素中的任一数据元素的可能的取值范围来选择该数码系统；

生成一个至少包括所述数字值的向量，所述向量代表该原始记录；并且

通过将所述向量与提取数据库中的至少一个其他向量进行比较，确定是否将所述代表该原始记录的向量包括到所述提取数据库中。

2. 权利要求 1 的方法，其中所述确定是否将所述代表该原始记录的向量包括到提取数据库中的步骤包括：将所述向量中的数字值与所述提取数据库中的至少一个其他向量的相应数字值在数值上进行比较。

3. 权利要求 1 的方法，其中所述确定是否将所述代表该原始记录的向量包括到提取数据库中的步骤包括：确定所述向量与所述提取数据库中的至少一个其他向量之间的内积。

4. 权利要求 1 的方法，其中所述确定是否将所述代表该原始记录的向量包括到提取数据库中的步骤包括：在所述向量与所述提取数据库中的至少一个其他向量之间进行模式识别分析。

5. 权利要求 1 的方法，其中所述确定是否将所述代表该原始记录的向量包括到提取数据库中的步骤包括：确定在至少所述向量与所述提取数据库中的至少一个其他向量之间的向量积。

6. 权利要求 1 的方法，其中所述确定是否将所述代表该原始记录的向量包括到提取数据库中的步骤包括：确定所述向量的向量大小。

7. 权利要求 1 的方法，其中所述确定是否将所述代表该原始记录的向量包括到提取数据库中的步骤包括：确定所述向量的向量方向。

8. 一种可在计算机上操作、用于将至少一个可由计算机访问的原始数据库的信息转换为可由计算机访问的提取数据库的方法，所述原始数据库包括多个原始记录，所述多个原始记录中的每个记录包括至少一个原始数据字段，每个原始数据字段包括多个数据元素，该方法包括以下步骤：

将一个原始记录中的至少一个原始数据字段中的每个字段转换为一个数字值，通过将该原始数据字段中的多个数据元素中的每个数据元素表示为一个数码系统中的一个数字，来生成每个数字值，根据该数据字段中的多个数据元素中的任一数据元素的可能的取值范围来选择该数码系统；

生成一个代表该原始记录的向量，所述向量包括为该原始记录中的至少一个原始数据字段中的每个字段所生成的数字值；并且

通过将所述向量与提取数据库中的至少一个其他向量进行比较，确定是否将所述代表该原始记录的向量包括到所述提取数据库中。

9. 一种可在计算机上操作、用于将至少一个可由计算机访问的原始数据库的信息转换为可由计算机访问的提取数据库的方法，所述原始数据库包括多个原始记录，所述多个原始记录包括至少一个原始数据字段，每个原始数据字段包括多个数据元素，所述提取数据库包括至少一个字段向量，所述至少一个字段向量包括多个数字值，该方法包括以下步骤：

通过将所述原始数据字段中的多个数据元素中的每个数据元素表示为一个数码系统中的一个数字，将至少一个原始数据字段中的每个字段转换为一个数字值，所述数字值在根据至少一个原始数据字段中的多个数据元素中的任一数据元素的可能的取值范围选定的数码系统内具有一个表示方式；并且

通过将所述数字值与所述至少一个字段向量中的多个数字值中的至少一个数字值在数值上进行比较，确定是否将对应于所述至少一个原始数据字段的所述数字值包括到所述至少一个字段向量中。

10. 一种可在计算机上操作、用于将至少一个可由计算机访问的原始数据库的信息转换为可由计算机访问的提取数据库的方法，所述原始数据库包括多个原始记录，所述多个原始记录中的每个记录包括一个原始数据字段，每个原始数据字段包括多个数据元素，所述提取数据库包括多个数字值，该方法包括以下步骤：

将一个数码系统中的数字分配给该原始数据字段中的多个数据元素中的每个数据元素，根据该数据字段中的一个数据元素的可能的取值范围来选择该数码系统；

由所述数字生成一个数字值，这个数字值在所述数码系统中代表所述原始数据字段；并且

通过将至少一个数字值与所述提取数据库中的多个数字值中的至少一个数字值进行比较，确定是否将至少一部分原始记录包括到所述提取数据库中。

11. 权利要求 10 的方法，其中所述数码系统具有一个至少等于所述数据字段中的一个数据元素的可能的取值数目的基数。

12. 权利要求 10 的方法，其中所述多个数据元素包括至少一个字母数字字符。

13. 权利要求 10 的方法，其中所述原始数据字段包括化学信息，并且其中所述数码系统具有一个至少等于包含在该化学信息中的化学结构数目的基数。

14. 权利要求 10 的方法，其中由所述数字生成一个数字值的步骤包括：由所述数字生成一个数字值，使得所述数字值在所述数码系统中可以作为原始数据字段被识别。

15. 权利要求 11 的方法，其中所述基数还使得可通过与计算机相结合的数据字来表示的数据元素的数目最大。

用于组织数据的系统与方法

本发明涉及数据库系统，并且，更具体地，涉及一种用于在数据库系统中组织和/或查找数据的系统与方法。

计算机化数据库系统已经被使用了很长时间，并且其基本概念已经广为人知。在 C. J. DATE, INTRODUCTION TO DATABASE SYSTEMS(Addison Wesley, 第 6 版, 1994)中可以找到对数据库系统很好的介绍。

通常，数据库系统被设计为以数据库中的数据有用的方式组织、储存并检索数据。例如，数据或数据的分区集合可以被搜索、排序、组织和/或与其他数据组合。在很大程度上，一个特定数据库系统的有用性取决于在该数据库系统中的数据的完整性（即准确性和/或正确性）。数据完整性受到所储存数据的“混乱”程度的影响。混乱可能会以错误或不完全的数据的形式出现，如重复的数据、破碎的数据、错误的数据等。在许多数据库系统中，有时现有的数据可能会被编辑并处理，结果就可能会引发额外的错误。在某些数据库系统中，新数据可能被引入。此外，由于数据库系统被升级到新的硬件和/或软件，可能需要进行数据转换，或者必不可少的附加字段。而在某些应用中，数据库中的数据可能只是过时了。

不管采用什么样的预防步骤，某种程度的混乱最终还是被引入了传统的数据库系统中。这种混乱程度随着时间以指数增长，直到最后在一个传统数据库中的数据变得完全无用了。作为结果，即使一个很小程度的混乱最终也会影响数据库系统的完整性。

不幸的是，在大型数据库系统中，在数据中识别并纠正混乱经常是困难的任务，即便不是不可能的任务。传统上，这样的任务是人工完成的，从而使得这些任务变得消耗时间、昂贵并且受到人为错误的影响。进而，由于该任务的这种性质，大量的混乱可能未被检测出来。所需要

的是一种用于在一个数据库系统中组织数据系统和方法，来克服这些及其他关联的问题。

本发明提供了一种用于在一个数据库系统中组织数据的系统与方法。本发明从由一个或多个原始数据源提取出的原始数据中导出一个准确数据的提取（distilled）数据库。原始数据被从其原始格式转换为数据格式。

依据本发明的一个实施例，原始数据表现为一个具有数字元素的向量。一旦原始数据被数字化地表现出来，就可以在这些向量上实施各种数学运算，如关联函数、模式识别方法或者其他类似的数字方法，以确定在一个特定向量中的内容如何与一个“提取”的或者参考数据库中的其他向量相对应。该提取数据库由一个或多个相关向量的集合构成，他们被认为针对其他集合是唯一的（如垂直的）。这些集合代表来自原始数据的最佳可用信息。在所有原始数据都被合并到提取数据库中去之后，新数据可以被屏蔽，以确保新的错误不被引入到提取数据库中去。新数据也可以被评估，以确定它是否是唯一的或者它是否包括比已经出现在提取数据库中的信息更好的信息。新数据从而被加入到提取数据库中。

依据本发明的一个实施例，根据一个带有适当基数的数码系统将原始数据转换为数字格式。一个适当的基数是根据在原始数据中所包含的信息类型来确定的。例如，对于通常由字母-数字字符组成的原始数据，一个适当的基数可以是大于或等于原始数据中出现的不同字母-数字字符的数目。使用这样一个数码系统可以使原始数据被以数字形式表现，允许通过各种广为人知的数学运算来进行操作。

依据本发明的一个实施例，该数码系统可以被选择，以便数字本身对于其所代表的原始数据保持语义上的意义。换句话说，在数码系统中的数字被选择，以使它们与原始数据相对应。例如，当原始数据由字母数字字符组成时，数字被选择，以与其所代表的字母数字字符相对应。当数码系统中的数字被随后显示时，它们表现为所代表的字母数字字符。

依据本发明的一个实施例，一旦原始数据在一个适当的数码系统中被以向量表示，则被表示的数据可以被有效地在数据库中使用各种已知的技术进行操作（如排序等）。进而，可以对向量进行各种已知的数学运算来分析数据内容。这些数学运算可以包括关联函数、特征向量分析、模式识别方法以及其他显而易见的方法。

依据本发明的一个实施例，原始数据被合并到一个提取数据库中。该提取数据库代表从原始数据中提取的最佳数据，没有任何数据混乱。

依据本发明的一个实施例，新数据可以与该提取数据库进行比较，以确定该新数据是否实际上包括任何尚未出现在提取数据库中的新信息或内容。任何尚未出现在提取数据库中的新信息被加入提取数据库中而不会增加任何混乱。以这种方式，提取数据库的完整性可以得到保持。

依据本发明，一种用于处理信息的方法包括以下步骤：根据包含在信息中的一个数据元素的可能值的一个范围来选择一个适当的数码系统，用在一个数码系统中的一个数字来表示所述数据元素；以及对该数码系统所代表的所述数据元素进行运算，以处理该信息。

依据本发明的一个实施例，选择一个适当数码系统的步骤包括选择一个数码系统的步骤，该数码系统带有一个基数，该基数至少等于并且近似相同于字母数字字符“0”-“9”和“A”-“Z”的序列。

依据本发明的一个实施例，选择一个适当的数码系统的步骤包括选择一个带有一个基数的数码系统的步骤，该基数大于一个字母数字字符“0”-“9”和“A”-“Z”的序列。

依据本发明的一个实施例，选择一个适当的数码系统的步骤包括选择一个带有一个基数的数码系统的步骤，该基数至少等于一个字母数字字符“0”-“9”、“A”-“Z”和“a”-“z”的序列。

依据本发明的一个实施例，选择一个适当的数码系统的步骤包括选择一个基数 40 数码系统的步骤。

依据本发明的一个实施例，该信息包括财务信息、科学信息、工业信息或者化学信息。

权利要求 16 的方法，其中分配数字的步骤还包括将数码系统中的

数位 A-Z 分别分配给字母数字字符“a”-“z”。

依据本发明的一个实施例，所述比较所述向量与一个提取矩阵的步骤包括进行一个特征向量分析，或者进行一个模式识别分析，或者确定在所述向量与所述提取矩阵中的一个向量之间的点积，或者确定在所述向量与所述提取矩阵中的一个向量之间的叉积，或者确定所述向量与所述提取矩阵中的一个向量之间的差，或者确定所述向量与所述提取矩阵中的一个向量的和，或者确定所述提取矩阵的一个决定因子（determinant），或者确定所述向量的一个量值（数值），或者确定所述向量的一个方向。

总的来说，本发明的特征就如在独立权利要求中所陈述的那样，而附属权利要求则包括了本发明的优选实施例。

参考下面的附图对本发明的优选实施例进行了描述。在图中，相同的参考号码表示同样的或者功能上相同的元素。另外，一个参考号码最左边的数位标志着第一次出现该参考号码的图号。

图 1 描绘了一个处理系统，在其中可以实施本发明。

图 2 描绘了由本发明的一个实施例处理数据的阶段。

图 3 是一个流程图，用于依据本发明的一个实施例将原始数据从其原始格式转换为一个数字格式。

图 4 描绘了一个适于与本发明一起使用的数据记录。

图 5 描绘了适于与本发明一起使用的原始数据表。

图 6 描绘了参考数据表，它表示依据本发明的一个实施例格式化的数据。

图 7 是一个流程图，用于依据本发明的一个实施例来分析参考数据。

图 8 描绘了提取数据表，它表示依据本发明的一个实施例被关联的相关数据。

图 9 描绘了一个集群在一个二维空间中的数据示例。

图 10 是一个流程图，用于在一对字段向量中标识重复数据。

图 11 是一个流程图，用于更详细地在一对字段向量中标识重复数

据。

图 12 描绘了在一对字段向量中标识重复数据的一个示例。

本发明针对一种用于在一个数据库系统中组织数据的系统与方法。下面将针对不同的示例性实施例来对本发明进行描述，特别是针对不同的数据库应用。然而，显而易见地，本发明的不同特性可以被扩展到其他领域。通常，本发明可以被应用于许多数据库应用中，其中有大量可能不相关的数据必须被编译、储存、操作、和/或分析，以便确定存在于该数据所代表的内容中的不同关系。更具体地，本发明提供了一种方法，用于实现并维护一个数据库系统中的数据完整性，即使在该数据开始就具有一种较高的混乱程度的情况下。正如这里所使用的，混乱是指重复的、错误的、不完全的、不正确的、虚假的或者非正确的或多余的数据。混乱可以以许多显而易见的方式出现在数据库系统中。

本发明的一个实施例被用于维护一个与应收账款相联系的数据库。在这个实施例中，一个公司可以从一个或多个来源收集关于不同个人、企业和/或账户的数据。这些来源可以包括，例如，信用卡公司、金融机构、银行、零售商和批发公司以及此类来源。当这些来源中的每一个都可以提供多种账户的数据时，每个来源可以根据其自身需要提供代表不同信息的数据。进而，可以以完全不同的方式来组织这种数据。例如，一个批发分销商可能具有对应于与公司账户相应的应收款的数据。这种数据可以按账号来组织，每个数据记录具有标识一个账号、一个与该账号关联的公司、一个该公司的地址和该账户所欠数额的数据字段。一个零售公司可以具有代表类似信息的数据记录，但它是基于与个人以及公司相应的账户的。

在本发明的其他实施例中，其他类型的来源可以提供不同的数据类型。例如，科研机构可以提供针对不同研究领域的科学数据。工业公司可以提供针对原始材料、制造、生产和/或供应的工业数据。法院或其他类型的法律机构可以提供针对法律状态、判决、破产和/或扣押物的法律数据。显而易见地，本发明可以使用来自多种来自的数据。

在本发明的另一个实施例中，一个数据库可以被维护，以实现一个

综合计账与订单控制系统。除了来自与上述来源相似的来源的信息之外，本实施例可以包括与库存相应的数据记录、与库存供应商相应的数据记录、以及与库存采购者相应的数据记录。库存数据可以按部件号码来组织，每个数据记录具有标识内部部件号码、外部部件号码（即供应商部件号码）、手头的数量、期望发出的数量、期望接收的数量、批发价及零售价的多个数据字段。供应商数据可以按一个供应商号码来组织；而客户数据可以按一个客户号码来组织。与这些记录中的每一个相应的数据记录可以包括标识部件号码、部件价格、订购数量、发货数据及其他此类信息的多个数据字段。

本发明的另一个实施例可以包括一个企业存储系统，该系统将公司信息从多个不同的来源统一在一起，并且使该信息在公司网络上对用户可用，不论数据类型、生成数据的计算机类型、或者请求数据的计算机类型。本发明的又一个实施例包括一个商业情报系统，它储存及推销信息，并且允许该信息被在线处理及分析。

本发明使从不同来源收集的原始数据可以被分析并提取为一个准确数据集合、以对一个特定应用有用的方式对其进行组织。使用上例的一个综合计账与订单控制系统，下面会对其进行更全面的解释，本发明可以产生一个提取数据库，其中相关数据如与一个特定供应商或者客户相关的数据可以被如此标识。在这个例子中，与相同供应商或客户相应的重复数据可以被识别出来和/或丢弃，而与供应商或者客户相关联的错误数据可以被识别、分析并可能被纠正。

总的来说，本发明可以实施在硬件或软件中，或者两者的组合中。理想地，本发明被实施为一种软件程序，它在一个可编程处理系统中执行，该处理系统包括一个处理器、一个数据存储系统以及输入输出设备。在图 1 中描绘了这样一个系统 100 的示例。系统 100 可以包括一个处理器 110、一个存储器 120、一个存储设备 130、以及一个 I/O 控制器 140，它们通过一个处理器总线 150 彼此相连。I/O 控制器 140 还通过一个 I/O 总线 160 与不同的输入与输出设备相连，例如键盘 170、鼠标 180 和显示器 190。显然，其他组件可以包括在系统 100 中。

图 2 描绘了本发明处理数据的不同形式。原始数据 210 可以从一个或多个来源收集，如原始数据 210A 与原始数据 210B。如这里所使用的，“原始数据”只是代表实际上从一个特定来源接收的数据。显然，原始数据 210 的附加来源可以被包括在其中。如下面解释的那样，来自不同来源的原始数据 210 最好被转换为一种数字格式并储存在一个参考数据库 220 中。使用这里称为“数据透析”的一种处理，本发明“净化”原始数据 210 来形成参考数据库 220 中的参考数据。参考数据库 220 包括在原始数据 210 中存在的所有信息，包括重复的、不完全的、不一致的及错误的数据。

储存在一个提取数据库 230 中的提取数据是从参考数据库 220 的参考数据导出的。提取数据代表原始数据 210 可用的“准确”数据。提取数据库 230 包括在原始数据 210 中存在的唯一信息。提取数据因而代表了原始数据 210 可用的最佳信息。

仍然如下面所解释的那样，本发明还为使用提取数据库 230 来分析并验证新数据 240 作了准备，这也可以被用于在适当的时候更新参考数据库 220 和提取数据库 230。

虽然本发明有大量的实施例，为了阐明其描述，在一个综合计账与订单控制系统的环境中，参考图 3-8 解释一个优选实施例。在该实施例中，原始数据 210 是收集自不同来源的一个数据集合，如订单处理、发货、接收、应付款与应收款等。这种原始数据 210 可以包括相关但具有不同数据字段的数据记录、重复数据记录、具有一个或多个错误数据字段的数据记录等。为了找出这种错误，本发明将原始数据 210 从其初始格式与数据结构（这可能根据来源而有所不同）转换为一个数字格式并将这个参考数据储存在参考数据库 220 中。

依据本发明，参考数据随后被比较并分析，以提取出可用的最佳信息。在本发明的一个实施例中，该最佳信息可以被储存为在提取数据库 230 中的提取数据。现在开始描述该过程。

收集原始数据

图 3 描绘了依据本发明的一个实施例，原始数据 210 被转换为参考

数据库 220 中的参考数据的过程。在步骤 310，原始数据 210 是从一个原始数据源收集的。如图 2 中所绘，原始数据 210 可以包括来自一个或多个来源的数据，如原始数据 210A 与原始 210B。如这里所使用的那样，“数据”是指信息物理上的数字表现，而数据“内容”是指其意义，其中所包含或者由该数据所代表的信息。在原始数据 210 中的不同记录可以包括相似类型的数据内容。例如，在计账环境中，原始数据 210 中的不同记录可以全都包括与一个特定账户相关的数据内容。

原始数据 210 将典型地被以数据记录 400 的形式接收，如图 4 所绘。每个数据记录 400 一般包括相关信息，如对于一个特定个人、公司或者账户的信息。每个数据记录 400 在一个或多个数据字段 410 中储存这种信息。可能的数据字段 410 的示例包括，例如，一个账号、姓、名、公司名称、账户余额等。然后每个数据字段 410 可以包括一个或多个数据元素 420，它们用于代表关于该特定记录与特定字段的信息。数据元素 420 可以以不同格式存在，例如字母数字、数字、ASCII（美国信息交换标准码）以及 EBCDIC（扩充二-十进制交换码），或者其他表现形式也是显而易见的。从不同来源收集的原始数据 210 可以被有差别地格式化。数据记录 400 可以包括不同数据字段 410，而包含在数据字段 410 中的信息可以使用数据元素 420 以不同格式来表示，这也是显而易见的。

在图 5 的原始数据表 510、520 和 530 中对原始数据 210 的示例进行的说明。数据记录，如数据记录 510-1 与数据记录 510-2 被描绘为原始数据表 510、520 和 530 的行，而数据字段 510-A、与数据字段 510-B 被描绘为原始数据表 510、520 和 530 的列。数据字段或者数据记录都能够被认为是普通数学向量或者张量，并因面可以被操作。图 5 中所描绘的表是可能存在于本发明不同实施例中的数据的示例。在另一个实施例中，数据可以来自许多来源并且可以被格式化为具有更大数目的数据记录和/或数据字段的数据库，这是显而易见的。

转换为数字格式

参考图 3，在步骤 320 中，本发明将原始数据 210 从其初始表现形式（可能是数字字母、数字、ASCII、EBCDIC 或者其他类似格式）转

换为一种数字表现形式。这保证了参考数据是以相同的方式被表现的。于是，参考数据，包括不同来源的数据在内，可以被同样地进行处理。

依据本发明，原始数据 210 被从其初始表现形式转换为一种适当的数字表现形式。一种适当的数字表现形式使用一个数码系统，其中数据元素 420 的每个可能值可以用该数码系统中的一个唯一数字或值来表示。换句话说，选择一个用于该数码系统的基数，以使该基数至少与对于一个特定数据元素的可能值的数目一样大。例如，在一个用于检测核酸中的腺嘌呤（A）、鸟嘌呤（G）、胞核嘧啶（C）与胸腺嘧啶（T）的核苷酸序列的生物技术应用中，每个数据元素可以是仅有的四种值中的一种：A、G、C 与 T。在这样一种应用中，关于该数码系统的基数为 4，就足以用一个唯一数字来代表每个数据元素。一个这样的数码系统可以包括数码 A、G、C 与 T。在本发明的某些实施例中，可以期望使用一个基数，它至少比数据元素 420 的不同可能值数目大 1，以便提供一个空字段的数码。在这种情况下，例如数码系统可以包括数码 A、G、C、T 与^，其中^是空字段值。

依据本发明的一个优选实施例，原始数据 210 中的数据元素 420 由字符组成，例如字母数字字符。在该优选实施例中，选择基数 40 来代表字母数字字符，如下表所示。（注意需要一个最小为 36 的基数。）这个基数被选定来容纳 10 个数字字符“0”-“9”和 26 个字母字符“A”-“Z”，以及允许若干附加字符。在本实施例中，不区分大小写字母。

如表 1 所示，基数 40 数码系统包括数码 0-9，接着是 A-Z，再接着是 4 个附加数码。这些数码中的一个被用于表示一个空字段。这个数码被用于表示为空或者没有值（与 0 值不同）的数据字段 410。其他数码可以被使用，例如，用来表示其他信息类型，如空格；或者被用作控制信息。

字母 数字	基数 10 数码	基数 40 数码	字母 数字	基数 10 数码	基数 40 数码
0	0	0	K 或 k	20	K
1	1	1	L 或 l	21	L
2	2	2	M 或 m	22	M
3	3	3	N 或 n	23	N
4	4	4	O 或 o	24	O
5	5	5	P 或 p	25	P
6	6	6	Q 或 q	26	Q
7	7	7	R 或 r	27	R
8	8	8	S 或 s	28	S
9	9	9	T 或 t	29	T
A 或 a	10	A	U 或 u	30	U
B 或 b	11	B	V 或 v	31	V
C 或 c	12	C	W 或 w	32	W
D 或 d	13	D	X 或 x	33	X
E 或 e	14	E	Y 或 y	34	Y
F 或 f	15	F	Z 或 z	35	Z
G 或 g	16	G	--	36	[
H 或 h	17	H	--	37	\
I 或 I	18	I	--	38	]
J 或 j	19	J	--	39	^

表 1

基数 40 格式中原始数据 210 的表现形式具有许多好处。一个好处是原始数据 210 可以用一种数字方式来表示，有助于直接进行数学操作。另一个好处是正确地选择基数与数码系统中的数字可以使所代表的内容保持语义意义，有助于以其数字格式的表现形式来识别原始数据 210 的内容。例如，4 个字母字符“J”“O”“H”“N”来表示的单词“JOHN”

可以在不同数码系统中被表示。一个这样的数码系统是一个基数 40 数码系统。使用表 1，以一个基数 40 数码系统来表示字母数字字符“JOHN”可以得到“四十进制”值'JOHN'，它等同于十进制值 1,255,103（$19*40^3+24*40^2+17*40^1+23*40^0$，其中基数 40 的'J'等于十进制的 19 等）。注意，基数 10 数码丢失了来自原始数据 210 的语义意义，而基数 40 数码保留了语义意义，如数码'JOHN'可被识别为内容“JOHN”。语义意义提供了一种数字表现形式的好处，而同时保持了传达语义内容的能力。

在本发明的某些实施例中，一个基数及其相应数码系统的选择可以取决于处理器 110 所使用的位数。处理器 110 使用的位数以及为数码系统选定的基数规定了能够被处理器 110 中的一个数据字所表示的数码字符。这种关系由下列等式决定:

$$N=B*\ln(2)/\ln(R)$$

其中 N 是处理器 110 的一个数据字所表示的全部字符的数目，B 是每个数据字的位数，而 R 是所选的基数。这种关系限制了可以装入一个数据字中的原始数据 210 的数据元素 420 的数目。例如，在一台 32 位机器中，可以装入一个使用基数 40 数码系统的数据字中的最大字符数目为 6（$32*\ln(2)/\ln(40)=6.013$）。可以装入一个使用基数 41 数码系统的数据字中的最大字符数目只有 5（$32*\ln(2)/\ln(41)=5.973$）。于是，在本发明的某些实施例中，除了具有一个大到足够保持语义的基数之外，该基数还可以被选择，以使单个数据字所表示的字符数目最大化，和/或有助于根据各种不同的处理器的优点或特别设计进行快速数学运算。在原始数据由字母数字字符组成的实施例中，一个适当的基数可以从 36 到 40。这个范围保持了语义意义并使 32 位数据字所表示的字符数目最大化。在本发明的其他实施例中，其他类型的原始数据和其他大小的数据字可以规定其他适当的基数范围。

本发明上述的实施例并不区分大小写字符。然而，本发明的其他实施例可以区分这些类型的字符。因此，为了区分这些字符，显然一种基数 64 的表示法（“0”-“9”，“A”-“Z”，“a”-“z”，以及两个

其他值）是合适的。

在每个数据字段 410 中的数据元素 420 还规定在处理器 110 中表示的数码所需的精度。如上所述，对于在一个 32 位机器中的单精度运算来说，每个数据字段 410 宽度可以只有 6 个字符或者数据元素 420。在本发明的某些实施例中，这可能是不够的。在这些实施例中，可能需要两倍、三倍或者甚至是四倍精度来将整个数据字段 410 表示为一个单一值。双倍精度数码对于最高 12 个字符的数据字段 410 是足够的；三倍精度数码对于最高 18 个字符是足够的；而四倍精度数码对于最高 24 个字符是足够的。

本发明的替换实施例可以通过将一个大型数据字段断开为一个或多个更小的数据字段来容纳大型数据字段。大型数据字段可以在自然边界被切断，如由空格定义的边界。例如，一个表示“123 West Main Street”这样一个地址的数据字段可以被断开为 4 个更小的数据字段：'123', 'West', 'Main', 和'Street'。大型数据字段也可以在数据字边界被切断。在上面的地址示例中，更小的数据字段可以是'123We', 'st\Mai', 'n\Stre', 和'et'，其中数码'\'被用于表示一个空格。显然，本发明的其他实施例可以以其他方式来容纳大型数据字段。

数据结构转换

如图 3 中所示，在步骤 330，被表示为一个数码的原始数据 210 被储存在一个预定的数据结构中。在本发明的一个实施例中，这个数据结构是一个单字段表，如图 6 的表 610-670 所示。这种数据结构可以是多样的。例如，在本发明的其他实施例中，取代单字段表，该数据结构可以是一个多字段表。在这些实施例中，数据结构可以实施为一些标准特性，例如表头和索引，如下面要详细解释的那样，还可以包括对每个记录的概率值。这些概率值代表该记录中的数据是完全的可能性。更高的概率值可以表示完全性概率，而更低的概率值同样地可以表示更低的完全性概率。这会在下面进一步详细描述。一开始，概率值被设为 0。其他实施例也可以包括键数码或者标识数码来帮助进行排序并维护数据记录之间的关系。

在本发明的一个优选实施例中，图 5 中描绘的原始数据 210 包括三个表 510、520 和 530。表 510 可以代表来自例如一个公司的应收款系统的原始数据 210。表 510 的各列代表对于账号、姓、名字的开头字母的数据字段和用于列出为一个特定个人处理的不同订单的附加字段。表 510 的各行（如 510-1 和 510-2）代表对于不同个人的数据记录。表 520 和 530 可以代表由信用卡公司维护的原始数据 210。表 520 和 530 的各列代表对于账号、姓、名和地址的数据字段。表 520 和 530 的各行代表对于特定账户的数据记录。

在优选实施例中，步骤 330 将原始数据 210 从图 5 中所示的格式转换为图 6 中所示的格式。图 6 描绘了从图 5 的不同原始数据表 510、520、530 组合出的原始数据 210，它被表示为一个基数 40 数码系统中的数码，并且被格式化为新表（表 610-670），这些共同组成参考数据库 220。

每个参考数据库表 610-670 与一个来自图 5 的原始数据表 510、520 和 530 的单独的字段相对应。更明确地，参考数据表 610-670 的数据记录对应于原始数据表 510 的数据记录，跟着是原始数据表 520 的数据记录，跟着是原始数据表 530 的数据记录。在本发明的一个实施例中，其中一个原始数据表记录没有关于在一个参考表 610-670 中表示的一个特定数据字段 410 的信息，则一个空字段值被输入到参考表中的该字段中。例如，表 510 的第一数据记录 510-1 没有关于地址的信息，于是一个空字段值被置入表 670 的第一位置。

数据最好是以这样的方式储存在参考数据库 220 中，即使所有对应于原始数据表中的一个单一数据记录的数据很容易被识别出来。在图 5 和 6 中表示的实施例中，例如，对应于原始数据表（表 510、520、530）的任何特定数据记录最好在参考表 610-670 中被表示为储存在跨参考表 610-670 的索引 i 中的数字数据的一个“向量”。例如，对应于原始数据表 520 的第六条记录 520-6（描绘为属于“Jennifer Brown”的账号“A60”，住在“51 Fourth Street”）在参考数据库表 610-670 中被表示为一个向量，它具有由表 610-670 的第十条记录 610-10、620-10、630-10、640-10、650-10、660-10 和 670-10 构成的系数。

如图 6 中所示，参考数据库 220 包括一个新表 610，它并不与图 5 中所示的原始数据 210 中的任何数据字段 410 相对应。这个表是一个“键表”，它标识在这些数据向量中的相关数据。如下面所述，由图 6 中所示的表组成的参考数据库 220 可以包括对于数据字段的附加键表。这些可以包括个人识别号码（“PIDN”）、账户识别号码（“AIDN”）、或者其他类型的识别号码。这些键表或者识别号码可以被用于识别在参考数据库 220 中相关数据向量的集合。

在本例中，键表 610 具有一个单一字段“PIDN”，其代表个人识别号码。键表 610 提供一个唯一标识符，这样一个特定的 PIDN 号码永远不会指向原始数据 210 中表示的一个以上的个人。换句话说，PIDN 号码反映了这样的事实，即原始数据 210 中的多条记录可以指向相同的个人。

理想地，键表 610 中的每条数据记录一开始就对应于原始数据表 510、520 和 530 表示的一条不同的数据记录。例如，在图 6 中，键表 610 中的数据记录 610-10 被这样实施，使它包括对于参考表 620-670 中相应数据的标识符（如指针或索引），这些共同对应于原始数据表 520 中的一条单一记录 520-6。

一开始，虽然一个单一的 PIDN 并未指向多个个人，而一个单一个人可以对应于多个 PIDN。例如，在图 6 中，向量 4（由 PIDN4 定义）与向量 9（由 PIDN9 定义）看来指向相同的个人，而如所描绘的那样，这个个人一开始就被分配给两个 PIDN 号码—PIDN4 和 PIDN9。如下所述，本发明支持对 PIDN4 和 PIDN9 是否事实上指向相同个人进行一个判断，并且，如果是的话，就向该个人分配一个单一 PIDN。替换地，某些实施例可以分配一个新 PIDN 号码给被如此判断的个人，并且可以保留对旧 PIDN 号码的一个参照。

如上所述，在本实施例中，记录在参考数据库表 610-670 中被表现为向量，它们具有跨 8 个单字段表的基数 40 数码系数。该数字表示法允许使用直接数学运算对数据进行分析，该数学运算被用于，例如，产生关联、计算特征向量、执行不同的坐标变换、并应用不同的模式识别

分析。然后，这些运算可以被用于提供或者导出关于记录及其相互关系的信息。通过使用小型的单字段表，这些运算可以被快速执行。另外，如将要描绘的那样，对包括字母数字字符的原始数据 210 的基数 40 数码表示法使得原始数据 210 的内容保留了其语义意义。

数据透析

回过来参考图 2，一旦参考数据库 220 如图 6 中所示被生成，一个数据透析处理 700 被应用，以提取出最准确的数据，从而将其包括在提取数据库 230 中。现在参考图 7 来描述数据透析 700。

参考数据分区

在步骤 710 中，最好根据某些准则将参考数据库 220 分区或排序为多个集合。这些排序准则可以是多样的。例如，如图 8 的表 810 中所示，在本实施例中，可以根据姓将数据记录排序为多个集合，其值按数码升序排列（记起原始数据的内容现在在参考数据库 220 中被表现为基数 40 数码）。表 810 是从图 6 中所绘的参考数据库表 620 导出的，表 810 的每个条目是由一个唯一的姓氏来定义的并且具有与该姓氏相匹配的表 620 记录的相应集合。在所述的表示法中，表 810 包括一个字段，用于定义该集合（在这种情况下是一个姓氏），以及该集合成员的标识符（如索引、指针或者其他适当的参考—这里是 PIDN）。

在本发明的某些实施例中，不是参考数据库 220 中所有的向量都将具有该字段的数据，其中集合是基于该字段的。这种向量可以以各种方式进行处理。例如，在参考数据库 220 中，所有没有该数据字段数据的向量可以被看作一个单一的附加集合的成员。替换地，参考数据库 220 中，每个没有该数据字段数据的向量可以被看作其自身集合的单一成员。

识别重复数据

回到图 7，在步骤 720 中，在被识别为重复的分区集合之内的数据记录被标记出来。在本发明的某些实施例中，重复数据可能是不必要的并且可以被丢弃。在其他实施例中，所有信息保留中参考数据库 220 中，因为所有信息，即使是错误的、不完全的或者重复的信息也可能要好于

没有信息，并且可能对于某些目的来说是有用的，如识别欺诈或偷窃。

在本发明的某些实施例中，比较一对向量可以识别出重复现象。显然，可以使用各种不同的运算。在一个简单的示例中，可以执行一个直接向量减法来测量两条记录之间的相似程度。其他技术也可以使用，以识别重复向量，例如使用“查找”表来识别普通名称、昵称、缩写等。

图 8 的表 810 描绘的是，姓氏“Smith”对应于 PIDN2、4、8、9 和 11，表示由图 6 中所绘的参考数据库表 610-670 的条目 2、4、8、9 和 11 构成的向量:

对于 PIDN2: [SMITH, J, 98-002, A40, A60, ^]

对于 PIDN4: [SMITH, J, 98-004, A50, B10, ^]

对于 PIDN8: [SMITH, Jennifer, ^, A40, ^, 300 Pine St.]

对于 PIDN9: [SMITH, John, ^, A50, ^, 37 Hunt Dr.]

对于 PIDN11: [SMITH, Jhon, ^, B10, ^, 85 Belmont Ave.]

比较向量的向量（或矩阵）运算以及用于确定何时两个条目足够相似以至可以被认定为重复的阈值可以针对不同实施例进行适当的定义。在一个简单的示例中，一对向量的相应系数之间的绝对差可以显示相应的记录对之间的相似性。如果一个第一向量与一个第二向量任意字段不是不一致的，则这对向量可以被认为是重复的，并且不提供任何附加数据。在本实施例中，附加规则也会被定义，例如，为比较不同长度的条目（如对应于数字的右对齐字符串、以及对应字母的左对齐字符串），为了一般地识别单词的拼写错误或者拼写变体，以及为了识别单词中顺序颠倒的字母。显然，这种处理可以通过各种机制来进行。在图 8 的表 810 的示例中，没有数据记录是严格的重复的，所以在步骤 720 中没有进行标记。

数据相关（Correlating）

回到图 7，在步骤 730 中，本发明的优选实施例对保留在每个集合内的数据记录进行相关，并且在步骤 740 中，进一步将数据记录分区为独立的数据记录子集。一般来说，两个向量之间的相关是对一个与另一个的关联有多么紧密的一种度量，具体的相关方法根据预期的应用会有

所不同。对于相关函数的一般描述与示例可以在参考材料中找到，如 NUMERICAL RECIPES IN C; THE ART OF SCIENTIFIC COMPUTING（剑桥大学出版社，第二版，1992），William H. Press 等著。其他技术与示例可以在 Donald E. Knuth 所著的 THE ART OF COMPUTER PROGRAMMING（Addison-Wesley Pub., 1998）中找到。

作为一个示例，向量之间相关的一种简单度量是其点积，它可以被赋与适当的权值。根据该应用，可以只对向量系数的一个子集计算点积，或者可以定义点积以便不仅比较相应系统，而且比较其他被确定为在相关字段中的系数对（即比较一个第一向量的“名”系数与一个第二向量的“中间名”系数）。对于识别重复数据的运算，相关函数可以根据其预期应用而进行适当的调整。例如，一个相关函数可以被定义为适当比较不同长度的条目并适当区分显著的与不显著的差别，这是显而易见的。

参考图 5、6 和 8 的表所解释的实施例中，一个相关函数的示例比较对应于共享相同姓氏的一个集合的成员的向量，以标识出独立的向量子集。再一次地，这种判断可以根据因应用而异的准则来进行。在本例中，独立子集可以被定义为那些代表不同个人的向量。

作为应用相关函数的结果，一个反映一对向量独立程度的相关参数被赋值。例如，可以赋以高值来表示高度的相似程度，而赋以低值来表示有限的相似程度。相关值随后被与一个预定的阈值进行比较以判断对应于那些向量的两条记录是否被认为是独立的—再次注意，该阈值在不同的应用中可以变化。

根据相关值，在步骤 740 中，优选实施例在每个集合之内将数据记录分区为独立的数据记录子集。在图 5、6 和图 8 的表 810 的示例中，一个独立子集的成员可以被标识为那些成员，它们具有：相同的姓氏（考虑到拼写错误和拼写变体）；相对近似的名（考虑到拼写错误、拼写变体、昵称和名、中间名与首字母的组合）；一个或多个匹配的账号；以及不超过 3 个地址（允许工作与家庭地址，以及一个地址变化）。

应用这样一种函数的结果被描绘在图 8 的表 820 中。识别出来的个

人是:

Jennifer Brown, PIDN 10;

Howard Lee, PIDN 3 和 6;

Carole Lee, PIDN 7;

Jennifer Smith, PIDN 2 和 8;

John Smith, PIDN 4 和 11;

John Smith, PIDN 9;

Ann Zane, PIDN 1、5 和 12; 以及

Molly Zane, PIDN 12。

其他用于对向量进行相关的运算也是可用的。这些可以包括计算点积、叉积、长度、方向向量以及很多依据已知技术用于评估的其他函数与算法。

图 9 描绘了一个称为集群的概念的二维示例，它被用于在概念上描述本发明的某些一般方面。在图 9 中，四个集群作为一个二维点的集合而存在。这些集群被标识为：(a, b), (c, d), (e, f), 及(g, h)。如所示的那样，每个集群由一个或多个二维空间中的点构成。每个点对应于一条代表（具有或多或少的准确性）在该空间中集群的“真”值的数据记录。如所示的那样，集群(a, b)和(c, d)相当容易相互区分，并与集群(e, f)和(g, h)相区分。然而，在本简例中，集群(e, f)和(g, h)并不容易相互区分。扩展空间（即对向量增加附加数据字段）可以增大集群如(e, f)与(g, h)之间的区别，以使它们相互之间变得更容易区分。替换地，扩展该空间可以表明(g, h)是一个属于集群(e, f)或者甚至是集群(c, d)的点。在理论上，在一个具有各种已知特征的希尔伯特空间中，空间可以无限扩展。显然，对于大量的，即使不是无限的向量，这些特征可以为本发明所利用。

进而，对向量增加附加数据字段（即扩展空间）可以将集群相互分离，从而有助于其相关，而从向量中删除数据字段（即缩减空间）也可以识别某些相关。在本发明的某些实施例中，缩减空间可以识别某些实际上代表相同人个或其他唯一实体的集群。例如，在一个数据库中的一条记录可以具有 10 个数据字段，它们与该数据库中的一个第二记录中

同样的 10 个数据字段完全一致。这些数据字段可以对应于名、出生日期、地址、母亲未婚时的姓氏等。但是，这两条记录可以有两个字段不同。这两个字段可以对应于姓氏和社会保险号码。在某些情况下，这些记录可以对应于相同的个人。本发明简化了识别这些类型的记录的处理，使用传统方法来检测即使不是不可能的也会是困难的。

于是，从一个向量中移除一个或多个特定数据字段并缩减相应空间可以展现那些不这样做就不是很明显的集群。对传统上被用于识别目的的数据字段（如姓氏、社会保险号码等）进行这种处理可以展现数据库中的重复记录。这对于识别欺诈是尤其有用的。移去其中包括一个值为空的数据字段的向量的数据字段也可以展现那些不这样做就不是很明显的集群。

进而，一旦集群被识别为代表相同个人或者实体，对于该个人或实体的最佳信息可以被从每条记录或“黑点”所提供的信息中提取出来。

本发明的原理可以被扩展超出简单向量与数据字段。例如，可以通过使用代表一个多维空间中的对象的张量来扩展本发明。在这种方式中，本发明可以被用于表示不同物理现象的参数，以深入洞察其操作与效果。这种应用对于解释人类基因并支援如人类基因组项目这样的计划是尤其有用的。

处理孤立（stranded）数据

再参考图 7，在步骤 750，本发明的优选实施例评估“孤立”数据。孤立数据是指那些来自参考数据库 220 的没有在步骤 710 中被分区到任何集合中去的记录。在某些实施例中，参考数据库 220 可以包括大量与数据字段相应的表和大量具有不同字段组合的数据的向量。例如，在一个参考数据库 220 包括对于不同数据字段的 20 个表及由每个表的相关数据记录定义的 1000 个向量的实施例中，假设在 1000 个向量中只有 800 个具有对于字段“姓氏”的数据，在步骤 710 中通过该字段生成集合。步骤 710 可能没有将没有“姓氏”数据的 200 个向量分区到任何集合，或者将这 200 个向量的每一个分入其自身的集合。在任何一种情况下，结果都是，这 200 个向量并未在步骤 720、730 和 740 中被与其他向量

相关。步骤 750 可以评估这些向量。

评估的方法可以是多样的。例如，一个实施例可以将每个孤立条目与步骤 740 中识别出的每个子集的一个成员相关。根据得出的相关值，该向量可以被加入相关度高的那个子集，或者可以定义一个新子集。替换地，在某些实施例中，可以判断这样的评估是否太耗费时间和/或太昂贵，而步骤 750 可以被完全跳过。

重复相关处理

对于特定的实施例，步骤 710-750 可以根据需要被重复进行。如上面注意到的，某些实施例将具有包含大量字段和大量条目的参考数据 220，其中有许多条目的数据只有一个字段子集。在这样一种情况下，在一个单一字段上执行步骤 710-750 未必能够导出所有相关信息。即使在参考图 5、6 和 8 解释的简单示例中，在单一字段“姓氏”上进行的相关只可以提供关于那些条目之间相互关系的部分信息。例如，对应于图 6 中 PIDN2 和 8 的 Jennifer Smith 可能与对应于 PIDN10 的 Jennifer Brown 是同一个人，因为 PIDN2 和 10 共用一个通用账号。在姓氏字段上执行相关可能不能将这些 PIDN 标识为与相同的个人相对应，因为它们只是针对共享相同姓氏的其他 PIDN 来评估的。在账号字段上执行相关可以提供关于这些 PIDN 是否相关的附加信息。

于是，跨不同字段的相关对于完全评估参考数据库 220 中数据的相关性程度是必不可少的。

使用相关结果来更新参考数据

一旦完成了步骤 710-760，参考数据库 220 就已经被提取为一个提取数据库 230，如图 2 所示。在本发明的某些实施例中，这两个数据库被分别处理并相互共存。在本发明的其他实施例中，一个单一数据库存在被标记为或者被标识为属于参考数据库 220 或提取数据库 230 的记录。这可以通过使用记录不同的 PIDN 范围来在两个数据库中进行分配而实现。进而，在两个数据库中记录之间的关系可以通过对参考数据库 220 中记录的 PIDN 增加一个常量以生成在提取数据库 230 中记录的 PIDN 而维护。例如，参考数据库 220 中一条 PIDN 为 12345 的记录可

以在提取数据库 230 中具有一个 PIDN9012345。以这种方式，两个数据库可以被看作一个单一数据库的不同部分。

使用提取数据库

一旦完成了数据透析过程 700，提取数据库 230 将来自参考数据库 220 的数据记录子集标识为相关记录，如上所述，可以确定参考数据库 220 中字段的概率来提供一个对其完整性的定性度量。这可以通过为各个数据字段的每一个赋与一个完整性概率然后使用它们来计算该数据记录的整体完整性概率来完成。例如，对于代表名的数据字段，值'J'可以被赋与低概率（如 0 或 0.1），值'JOHN'可以被赋与更高的概率（如 0.7 或 0.8），而值'JONATHAN'可以被赋与最高概率（如 0.9 或 1.0）。这些值可以被略为随意地赋值，或者依据某种结构前提来赋值。然而，这些值有助于识别集合中的哪些数据字段最可能包括最完全的信息或者换句话说，最可能的数据。

使用本发明可以确定大量关于记录及其相互关系的信息，并且可以专门为特定的应用进行定制。进而，使用标准数据库操作，提取数据库 230（参考参考数据库 220 的数据）可以被操作，以根据需要提供格式化的报告。例如，一个实施例可以被定制为生成一个列出相关记录子集的报告，一个子集的记录提供关于一个特定个人或实体的信息。在这样一个子集中的记录可以提供例如关于信息的不同字段的信息；名字的别名和/或变体、地址、社会保险号码等；以及字段—如职业、地址和账号，对于它们，该个人可能具有不止一个条目。

由于所有数据是以数字基数 40 格式来表示的，子集在报告中可以按数字来排序。基数 40 格式提供额外的好处，即将字母字符表示为其各自的字母（如上面的转换表所示）。这样，报告将会以数字表示法来显示，该表示法保持了其所代表的数据的语义意义，允许数据被人工地读取和分析。例如，如果该报告显示了一个具有包括 J SMITH, JOHN SMITH, JOHN G SMITH, G SMITHT 和 GERALD SMITH 条目的个人的记录，阅读该报告的一个人会理解到，这个人使用了不同的名，包括他的名或首字母、中间名或首字母、或它们的某种组合。

增加新数据

对于传统的数据库应用，新数据可以时常被加入。如图 2 所示，本发明解决了增加新数据 240，这将影响参考数据库 220 和提取数据库 230。

通常，新数据记录 240 可以参照图 3 所述来格式化，并被输入到现有的参考数据库 220 中。另外，新数据记录 240 可以针对提取数据库 230 被度量，以判断新信息或内容在新数据记录 240 中是否可用。例如，一条新数据记录 240 可以与来自提取数据库 230 的数据记录进行相关，以判断该新数据记录 240 是否与任何已经存在于提取数据库 230 中的数据记录相关。如果是，并且新数据记录 240 包含没有出现在提取数据库 230 中的信息或内容，则新数据记录 240 可以被用于更新提取数据库 230。例如，如果新数据记录 240 包括一个名叫 John Smith 的个人信息，他对应于已经出现在提取数据库 230 中的数据记录，但提供了附加信息，即 Smith 先生的中间名是 Greg，则该附加信息可以被适当地加入提取数据库 230。

对参考数据库 220 与提取数据库 230 中数据记录的改变可以使用标准数据库保护操作来处理，在参考资料如 C. J. DATE, INTRODUCTION TO DATABASE SYSTEMS（Addison Wesley，第 6 版，1994）（特别参见 Part IV）中所述，参照上述内容。例如，在由一个授权数据库管理员对参考数据库 220 进行改变时，参考数据库 220 中的相关数据记录被根据标准关系定义所做的判断来更新，并且其中适当地根据提取数据库 230 中定义的关系。

在字段向量之间识别重复数据

传统数据库的一个问题是在合并来自一个第一数据库，如原始数据 210A 的数据记录与来自一个第二数据库，如原始数据 210B 的数据记录时的困难。在具有共享或重复数据的这些数据库中的记录需要被识别，以使其中包含的内容可以被合并为在一个数据库中的一个单一记录，如参考数据库 220 或提取数据库 230。例如，两个数据库 210 可以包括对于 JOHN SMITH 的一个或多个条目。如果数据库 210 中的相应记录代

表相同的个人 John Smith，则每条记录的内容应当被合并为一条单一记录，例如在提取数据库 230 中。

用于在这些数据库中识别这种重复数据的传统穷举方法包括比较一个来自第一数据库的数据记录与第二数据库中的每条数据记录，并对第一数据库中的每条记录重复这个过程。这个过程耗费时间、计算密集，并且因而是昂贵的。事实上，计算量与两个数据库中每一个中的记录数量是几何相关的。

用于减少在数据库 210 中识别重复数据所需的计算量与时间的一种处理在下面参照图 10-12 来描述。在下面所述的过程中，选择一个在数据库中普通或类似的特定字段，例如姓名字段或地址字段。这个字段被对于每个数据库排列为一个表或者一个数组，它包括每条记录所选字段的值。例如，如上所述，610-670 的每个表代表一个数据库中每条数据记录的一个特定字段。出于讨论的目的，这些表被称为字段向量。

依据本发明，每个字段向量被以数字顺序排序，并且如果必要的话，将其分区为相同数据的集合，如前面针对图 7 和 8 所作的叙述。例如，多条与 JOHN SMITH 相关的记录在字段向量之内会被分区到一起。最好能够储存关于集合之间分区位置的信息。

一旦字段向量被排序并分区，一个第一字段向量的第一元素的值被与一个第二字段向量的第一元素值进行比较。本质上，如果第一字段向量中的值大于第二字段向量中的值，则向第二字段向量中增加一个索引或者将索引调整到下一个分区集合内的一个位置，以获得在第二字段向量中的下一个值。第二字段向量中的这下一个值随后被与第一字段向量中的值进行比较。只要第一字段向量中的值大于第二字段向量中的值，这个处理就继续下去。

另一方面，如果第一字段向量的值小于第二字段向量的值，则第一字段向量中索引被增加或者将索引调整到下一个分区集合的一个位置，以获得第一字段向量的下一个值。第一字段向量的这下一个值随后被与第二字段向量中的值进行比较。只要第一字段向量中的值小于第二字段向量中的值，这个处理就继续进行下去。

当第一字段向量的值等于第二字段向量中的值时，过程就识别出了重复数据，随后最好将其储存在一个通用字段向量中。储存了识别出的重复数据之后，对第一字段向量和第二字段向量的索引都被增加或者调整到其各自字段向量的下一个分区集合之内的一个位置。

如此描述的处理过程可以被看作反馈控制机制，它根据字段向量中值之间的差来调整对数组之中的索引。在上述实施例中，一个正差产生了一个对第二字段向量索引的调整，而一个负差则产生一个对第一字段向量索引的调整。这个过程导致了一个在字段向量中值的数量与所需的计算量（即比较）之间的线性关系，这与传统方法的几何关系相反。

本发明也可以被扩展为排序机制。在一个特定值必须根据向量中的值的顺序（如字母、数字顺序等）被插入到一个字段向量中去时（即一条记录必须被插入到一个数据库中去），计算该特定值与向量中的元素之一的值之间的差。这个差值被“反馈”，以调整向量之中的索引来生成来自该向量的下一个值。使用精心建立的控制理论方法，索引调整可以被积分，以判断要被插入的值的正确位置。除了积分器之外，显然可以对差值加上一个比例增益，以建立一个期望系统性能。

现在参照图 10-12 对本发明进行描述。图 10 是一个识别一对字段向量内重复数据的流程图。字段向量可以来自一个单一来源，如原始数据 210A（如当在一个单一数据库中比较一个居住地址与一个邮件地址时），或者来自多个来源，如原始数据 210A 和原始数据 210B（如当比较两个数据库之间的一个姓名字段时）。

为了描述的目的，字段向量对分别被称为第一字段向量（“FV1”）和第二字段向量（“FV2”）。这些字段向量中的数据最好是上述表示字母数字数据的基数 40 数码。然而，在本发明的某些实施例中，该数据也可以存在于其他形式之中。

在步骤 1010，第一字段向量被按照数字顺序进行排序。在步骤 1020 中，第二字段向量也被按照数字顺序进行排序。在本发明的一个实施例中，向量是按照数字升序排序的，虽然本发明的其他实施例显然可以按降序对向量进行排序。

在步骤 1030 中，识别出在第一字段向量之内具有通用值的分区集合。同样地，在步骤 1040，也识别出在第二字段向量之内具有通用值的分区集合。步骤 1010-1040 前面参考图 7 和 8 描述的参考数据库 220 分区步骤执行了一个相似的功能。在本发明的某些实施例中，字段向量可以不包括任何分区集合，因为每个字段向量之内的通用值可能已经被去除了。但是，在本发明的一个优选实施例中，一个特定字段向量之内的通用值被保留下来。

在步骤 1050，一个标识第一与第二字段向量之间的通用值的通用值向量被确定，最好使用分区集合。参照图 11 对步骤 1050 进行了更详细的描述。

图 11 是识别一对字段向量之间通用值的一个流程图。在一个步骤 1110 中，三个向量索引被初始化。一个第一向量索引 I 是对于第一字段向量 FV1 的索引；一个第二向量索引 J 是对第二字段向量 FV2 的索引；第三向量索引 K 是对通用值向量（“CV”）的索引。如前面所提到的，通用值向量包括第一与第二字段向量共享的值。索引 I 和 J 被初始化，以在第一与第二字段向量的每一个中分别定位一个第一位置。索引 K 被初始化，为下一个会被包括在通用值向量中的通用值定位一个位置。

在判定步骤 1120，本发明判断在第一字段向量的第 I 个位置中的值是否大于或等于第二字段向量的第 J 个位置的值。如果是，则过程继续进行到判定步骤 1130；否则，过程继续进行到步骤 1170。当第一字段向量的第 I 个位置中的值小于第二字段向量的第 J 个位置的值时，步骤 1170 被有效地执行。在步骤 1170 中，第一索引 I 被调整，以定位第一字段向量中下一个分区集合的开始位置。步骤 1170 之后，过程继续进行到判定步骤 1160。

在判定步骤 1130 中，本发明判断第一字段向量的第 I 个位置中的值是否与第二字段向量的第 J 个位置的值相等。如果是，过程继续进行到判定步骤 1140；否则过程继续进行到步骤 1180。当第一字段向量的第 I 个位置中的值大于第二字段向量的第 J 个位置的值时，步骤 1180 被有效地执行。在步骤 1180，第二索引 J 被调整，以定位第二字段向量

中下一个分区集合的开始位置。步骤 1180 之后，过程继续进行到判定步骤 1160。

当第一字段向量的第 I 个位置中的值与第二字段向量的第 J 个位置的值相等时，步骤 1140 被有效地执行。在步骤 1140 中，包括在第一与第二字段向量中的值被置入通用值向量中。

在步骤 1150 中，第三索引 K 被增量，以定位下一个要被识别的通用值的通用值向量中的位置。第一索引 I 被调整，以定位第一字段向量中下一个分区集合的开始位置。第二索引 J 被调整，以定位第二字段向量中下一个分区集合的开始位置。

在判定步骤 1160 中，本发明判断是否有附加分区集合存在于第一字段向量和第二字段向量之中。如果是，过程继续进行到步骤 1120。如果没有分区集合保留在第一字段向量或者第二字段向量中，过程结束。当过程结束时，通用值向量包括所有在第一与第二字段向量之间识别出来的重复数据。

图 12 描绘了依据本发明在字段向量之间识别重复数据的一个示例。步骤 1010 与 1030 排序并分区字段向量 1（“FV1”），而步骤 1020 和 1040 排序并分区字段向量 2（“FV2”）。现在参考步骤 1110-1180 来描述步骤 1050 的操作，其中穿过步骤 1120 到 1160 并返回到步骤 1120 被称为一个“循环”。

在一个第一循环中，FV1 的第一元素（即第 0 个位置）与 FV2 的第一元素进行比较。（这在图 12 中被描绘为 FV1 与 FV2 之间的一条两端带有箭头的线，并被标注为 1）。在本例中，FV1 的值'8'与 FV2 的值'8'相比较。判定步骤 1120 与 1130 判断这些值相等，并且在步骤 1140 中，值'8'被置入通用值向量。（这在图 12 中被描绘为 FV2 与通用值向量之间的一条两端带有箭头的线，并被标注为 1'）步骤 1150 调整两个字段向量的索引，以指向下一个分区集合。判定步骤 1160 判断更多的分区集合存在于字段向量中，并且一个第二循环被启动。

在第二循环中，FV1 的下一元素被与 FV2 的下一个元素进行比较。在本例中，FV1 的一个值'9'被与 FV2 的一个值'9'进行比较。这些值再

次被判断为相等，并且值'9'被置入通用值向量。象前面一样，步骤1150调整两个索引，以指向其各自字段向量中的下一个分区集合。判定步骤1160判断更多的分区集合存在于两个字段向量之中，并且一个第三循环被启动。

在第三循环中，FV1 的下一个元素被与 FV2 的下一个元素进行比较。在本例中，FV1 的一个值'10'与 FV2 的一个值'12'进行比较。判定步骤1120判断 FV1 中的值不大于或等于 FV2 中的值，并且在步骤1170中，FV1 的索引被调整，以指向其中的下一个分区集合。判定步骤1160判断有更多的分区集合存在于两个字段向量中，并且一个第四循环被启动。

在第四循环中，FV1 的下一个元素与 FV2 的前一个元素进行比较。在本例中，FV1 的一个值'12'与 FV2 前一个比较过的值'12'进行比较。判定步骤1120与1130判断该值相等，并且在步骤1140中，值'12'被置入通用值向量中。步骤 1150 调整两个索引来指向其各自字段向量中的下一个分区集合。判定步骤 1160 判断有更多的分区集合存在于两个字段向量之中，并且一个第五循环被启动。

在第五循环中，FV1 的下一个元素与 FV2 的下一个值进行比较。在本例中，FV1 的一个值'15'与 FV2 的一个值'18'比较。判定步骤1120判断 FV1 中的值不大于或等于 FV2 中的值，并且在步骤1170中，对 FV1 的索引被调整，以指向其中的下一个分区集合。因为没有更多的分区集合存在于 FV1 中，过程结束。

在本例中，需要每个循环最多进行两次比较的 5 个循环来识别两个字段向量之间的三个通用值。在穷举方法中，需要 132 次比较（12*11）。

预编码信息

在本发明的不同实施例中，在将数据从其初始格式转换为数字格式之前，或者在某些实施例中与其同时，将数据预编码为一种中间编码格式。这种预编码还减少或压缩了初始格式到编码格式的信息。一旦进入编码格式，数据能够随之被以一种适当的数字格式来表示，如上所述。本发明的这些实施例用示例的方式很好地进行了描述。

在本发明的一个实施例中，音素被用于将初始格式的数据表示为编码格式。在本实施例中，音素可以被用于对单词、单词的某些部分（如音节）或者单词的词组进行编码。于是，发音一致或类似的单词或者音节被使用相同的音素来表示。例如名字“John”或者“Jon”会被使用相同的音素来表示。在某些实施例中，名字“Joan”也可以使用与名字“John”和“Jon”相同的音素来表示。依据本发明每个音素随后被部分地根据所用音素，以一种适当的数码系统表示为一个数字。

例如，一种特定语言可以被分割为其有限数目的“发音”或者音素，并表示为一个适当数码系统之内的数字。以这种方式，文本可以根据语音而不是特定拼写进行编码，从而使拼写错误的影响最小化，例如使用搜索引擎时的拼写错误。

本发明的这些实施例可以扩展到语音、语音识别和人工语音表现机制。尤其是，听觉语音音素（与相应的文本音素相对）也可以在一个适当的数码系统中如上述所表示，并被用于简化上述的语音识别与语音表现。

在本发明的其他实施例中，单词、词组、习惯用语、句子、和/或想法可以被预编码，然后被表示了一个适当数码系统中的数码，如上所述。这样的实施例可以被用于，例如，改进自动语言翻译系统。这些实施例还可以被用于改进搜索引擎。被称为一个或多个想法或概念的大型文本可以根据所传达的想法或概念的每一条来进行预编码。这些实施例提供概念性搜索，相对于识别和/或定位可能不出现在段落中的特定单词或词组。

在本发明的另一个实施例中，原始地址信息被预编码为坐标表示，例如，根据经度和纬度，并随后在一个适当的数码系统中被表示出来，例如，在一个基数 60 的数码系统中。这样一个系统可能对绘图操作、导航系统或者跟踪系统特别有用。

在本发明的另一个实施例中，原始指纹数据被预编码为不同参数、记录点（registration points）或者其他适于分类指纹的识别标记，它们每一个随后被表示为一个适当数码系统中相应的数字。于是，每个指纹

可以以一个字段中的值来表示，或者替换地，每个指纹可以被表示为一个字段向量。可以出于多种目的（即犯罪的和非犯罪的目的）在一个这种信息的数据库中根据从个人收集的指纹对结果数据进行组织和维护，这些可以包括由法庭专家、保安人员、背景调查员等收集的指纹。本发明理想地适用于净化现有指纹数据库、将那些数据库合并到一个参考数据库中、当变得可用时，增加新指纹信息、并将指纹信息与参考数据库中的信息进行匹配。

可以理解，在使用预编码的实施例中，在许多情况下，基本的初始数据必须被预处理为中间格式。这样，为了本发明可以被使用在一个搜索环境中，被搜索的信息必须被预编码或者“预处理”。在某些情况下，这种处理可能会导致语义意义的损失，如上面针对本发明的其他实施例所作的叙述那样。

示例性实施例

本发明不同实施例可以被用于许多不同应用，它们中的一些已经在上面描述和/或侧面提到了。例如，在上述应用中，本发明可以被用于组合从多个来源收集的计账信息来导出一个提取数据库，在该提取数据库中有关数据记录被识别而且重复与错误数据记录被去除。如所建议的那样，这可能会在例如涉及欺诈的情况下特别有用。典型地，使用信用卡或者其他形式的零售欺诈的人对其某些个人信息进行微小的改变，而让大多数信息保持相同。例如，一个社会保险号码中的数字时常可能被颠倒或者使用一个别名。然而，其他信息如个人地址、出生日期、母亲未婚时的姓氏等也经常被相同地使用。这些类型的欺诈很容易为本发明所识别，虽然它们难以为人类分析所识别。

其他可能的应用包括在电话推销中，来编辑一个目标个人或地址的列表；在邮购目录中，来减少大量发送给相同个人或家庭的目录；或者合并来自销售类似数据库的不同销售商的记录。还有另一个潜在的应用是在医学研究或者诊断领域中，其中核酸中的腺嘌呤（A）、鸟嘌呤（G）、胞核嘧啶（C）与胸腺嘧啶（T）的核苷酸序列可以被识别。另一个由税务组织如国家税务局、州与地方政府等使用的应用组织并维护准确的名

单与税务基本信息。

在其他实施例中，本发明可以在开始被用作一个特定数据库的门卫，以便从一开始就维护数据库的完整性，而不是在晚些时候才在数据库中实现完整性。在这些实施例中，没有原始数据 210 出现，并且只有新数据 240 存在。在新数据 240 被增加到数据库之前，针对提取数据库 230 进行度量，以判断新数据 240 是否包括附加信息或者内容。如果是，则只有新信息或内容通过更新提取数据库 230 中的一条现有记录被加入提取数据库 230，以反映新信息或数据，这是显而易见的。

在本发明的另一个实施例中，一个邮件服务，如美国邮政局，或者一个快件递送服务，如 Airborne Express, Federal Express, United Parcel Service 等，它们使用本发明来维护一个有效递送地址列表。一个与一件要被递送的物品相关联的地址针对一个地址参考数据库进行检查，以识别在该地址中的任何不准确性。不准确的地址可以或者被纠正（如对颠倒的号码等），或者对其进行联系以验证该地址。新地址可以在变得可用时被加入到参考数据库，例如，在物品被成功递送时。另外，某些发件人可以被识别为倾向于寄错物品或者提供了不正确的地址。如果合适的话，这些发件人可以被通知。

除了使用本发明来匹配如上所述的 DNA 序列片断之外，遗传学研究者（如药品公司、种子公司、动物饲养员等）还可以使用本发明来表示在一个集合中个人明显的、切实的和/或客观的特征，并使用这种信息来识别造成这些特征的个人基因或基因序列。

在另一个实施例中，本发明被用于在一个网络如因特网上的信号（数据包）交换与路由数据。为一个目的地址和序列信息检查一个进入包并将数据包以适当的顺序排序为一个适当的输出队列。在本实施例中，本发明对数码排序的能力提供了对传统系统独特的优点。使用一个替换数码系统（与现在所使用的一个传统数码系统相对）会产生一个扩充的地址空间，这提供了对网络寻址与通信协议方法的改进。

在另一个实施例中，本发明被用于在一个三维环境中表现并显示一个对象。这些行为需要庞大数量的排序来判断哪些对象在前景显示而哪

些对象相应淡化地背景上，以及判断每个对象的亮度特性（即阴影等）。

虽然在一个优选实施例中描述了本发明，其他实施例与变化都在后面的权利要求范围之内。例如，格式化过程 300 可以使用不同的基数或者其他字符集来格式化数据，并且可以使用不同的数据结构。数据结构可以表示多字段；根据应用，数据结构将表示多种字段。例如，在一个信用应用中，除了关于账户持有人的个人信息之外，字段还可以包括账户状态、账号和法律状态。在一个医疗诊断应用中，字段可以包括等位基因或者其他在组织样本中检测出的基因特征。

图 1

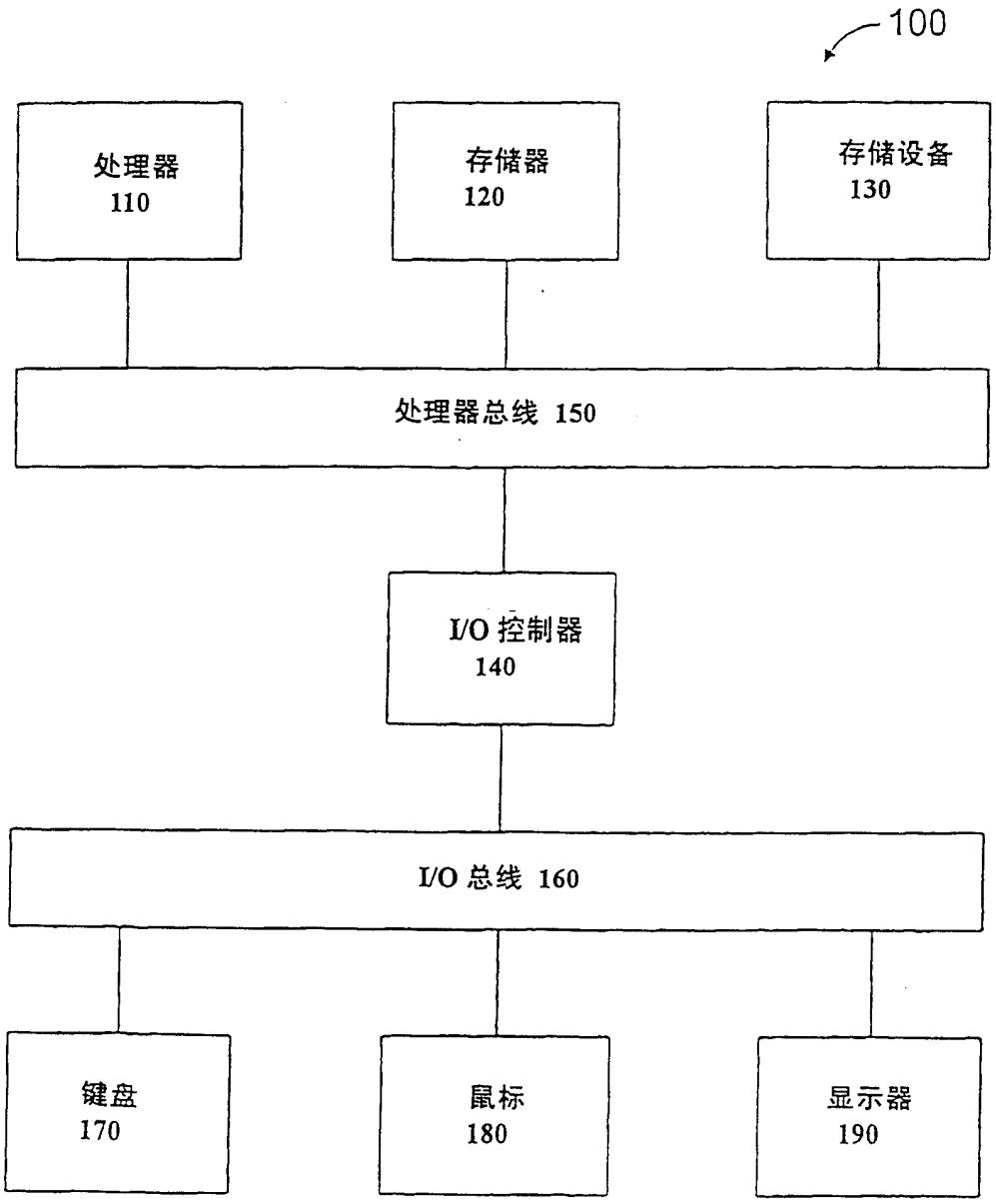

100
处理器
110
存储器
120
存储设备
130
处理器总线 150
I/O 控制器
140
I/O 总线 160
键盘
170
鼠标
180
显示器
190

图 2

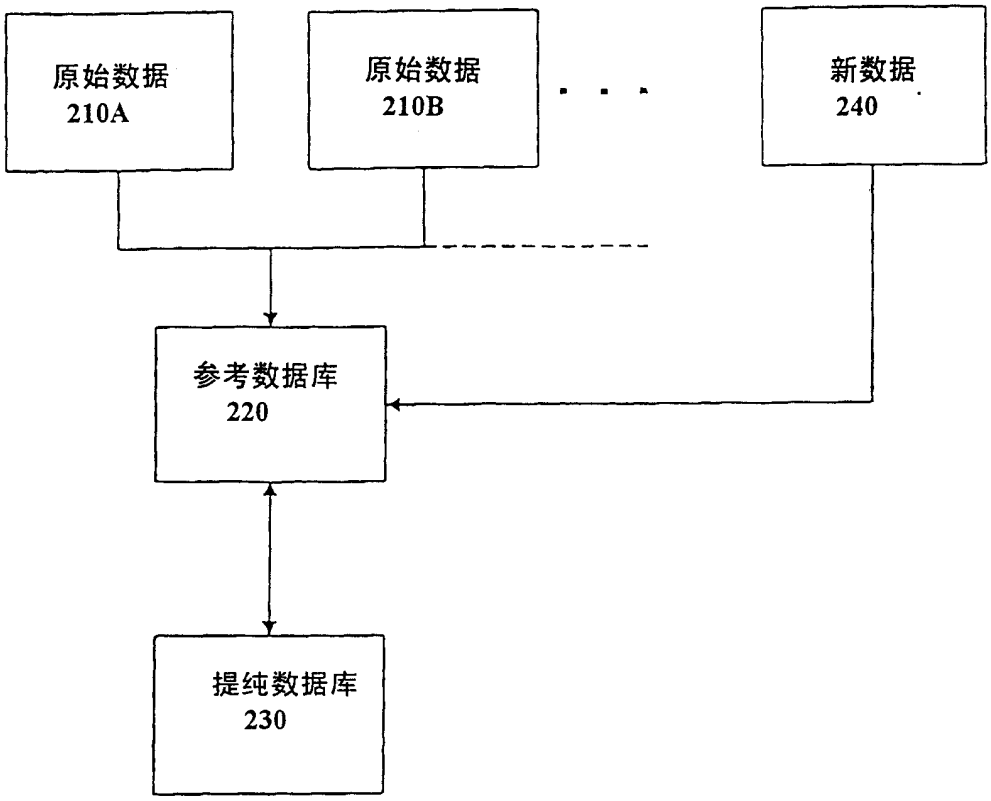


图 3

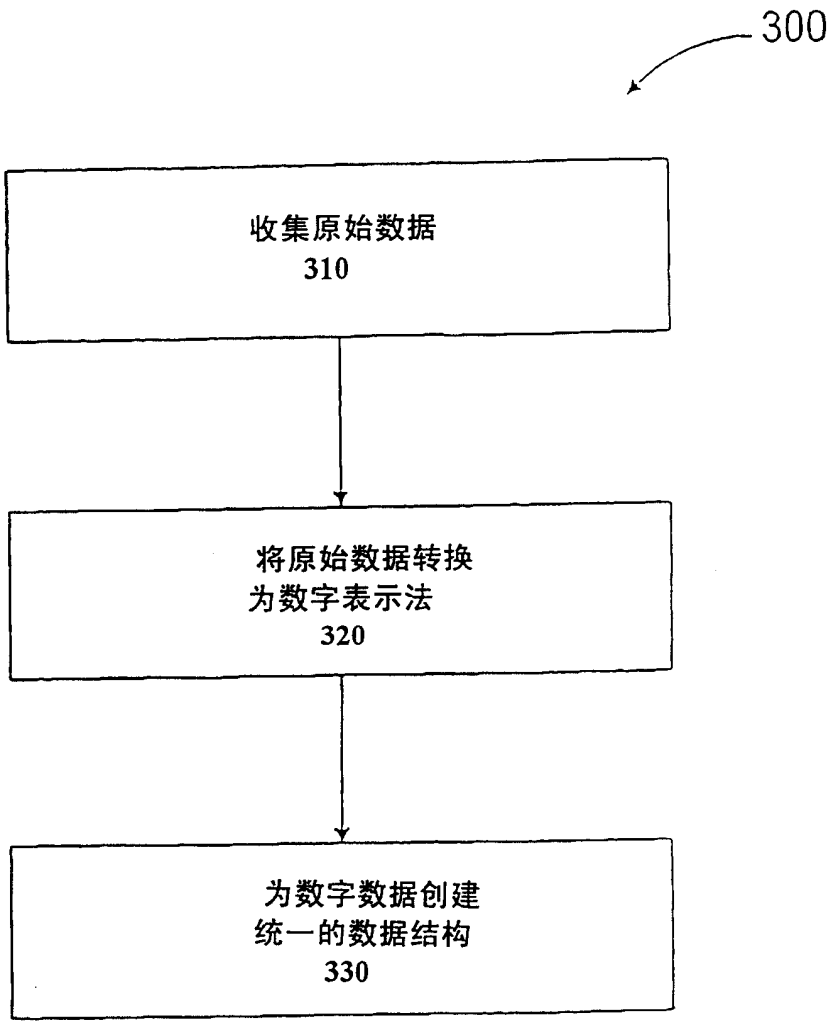


图 4

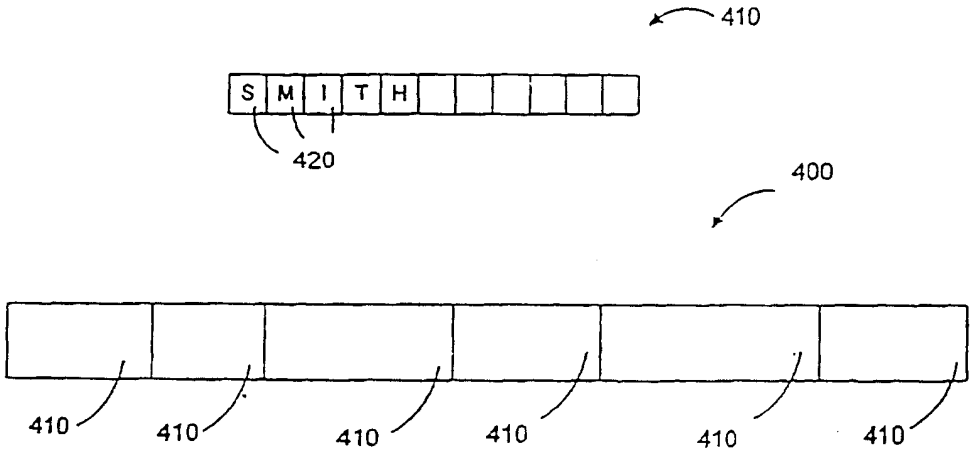
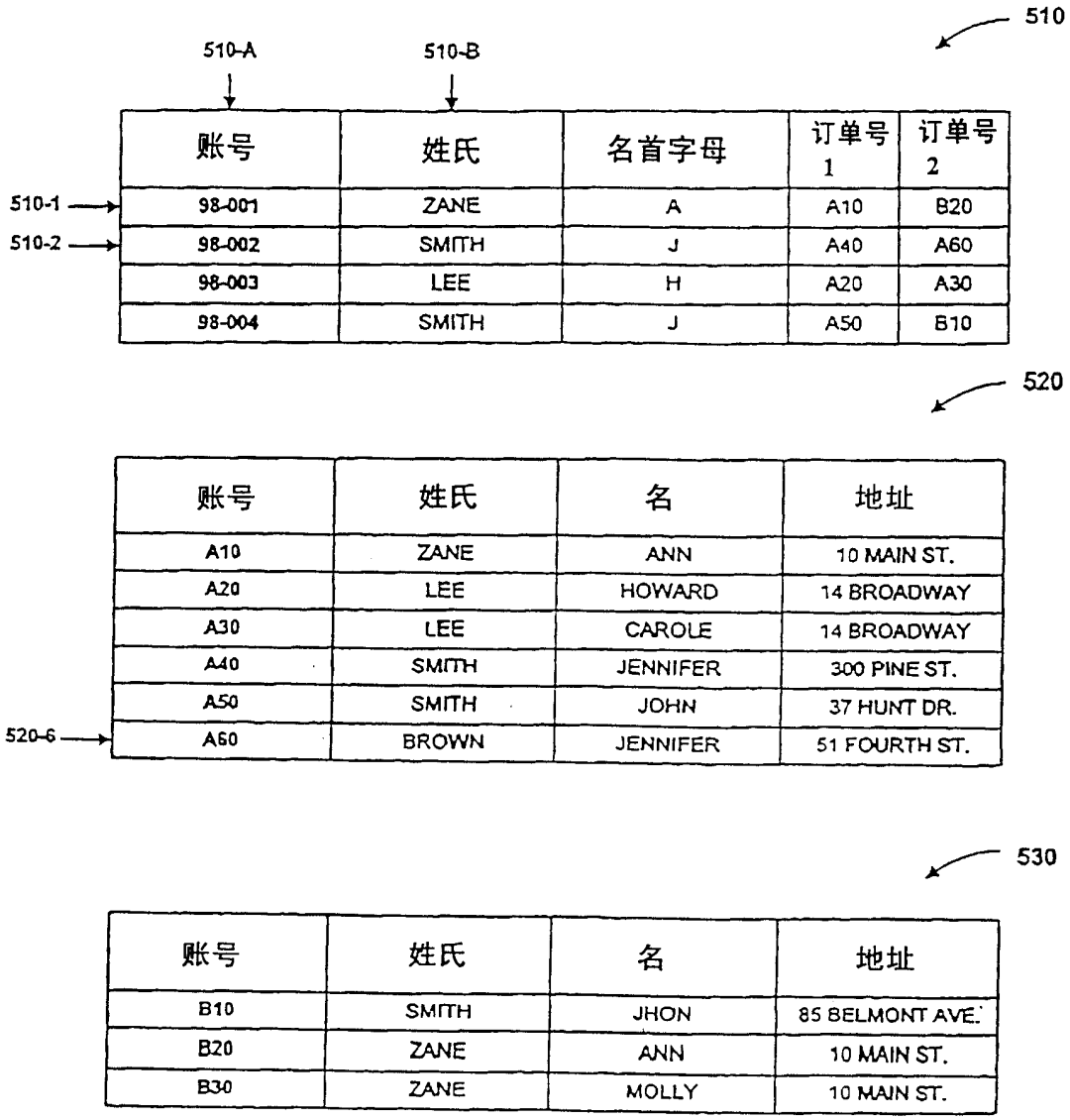


图 5



账号	姓氏	名首字母	订单号 1	订单号 2
98-001	ZANE	A	A10	B20
98-002	SMITH	J	A40	A60
98-003	LEE	H	A20	A30
98-004	SMITH	J	A50	B10

账号	姓氏	名	地址
A10	ZANE	ANN	10 MAIN ST.
A20	LEE	HOWARD	14 BROADWAY
A30	LEE	CAROLE	14 BROADWAY
A40	SMITH	JENNIFER	300 PINE ST.
A50	SMITH	JOHN	37 HUNT DR.
A60	BROWN	JENNIFER	51 FOURTH ST.

账号	姓氏	名	地址
B10	SMITH	JHON	85 BELMONT AVE.
B20	ZANE	ANN	10 MAIN ST.
B30	ZANE	MOLLY	10 MAIN ST.

图 6

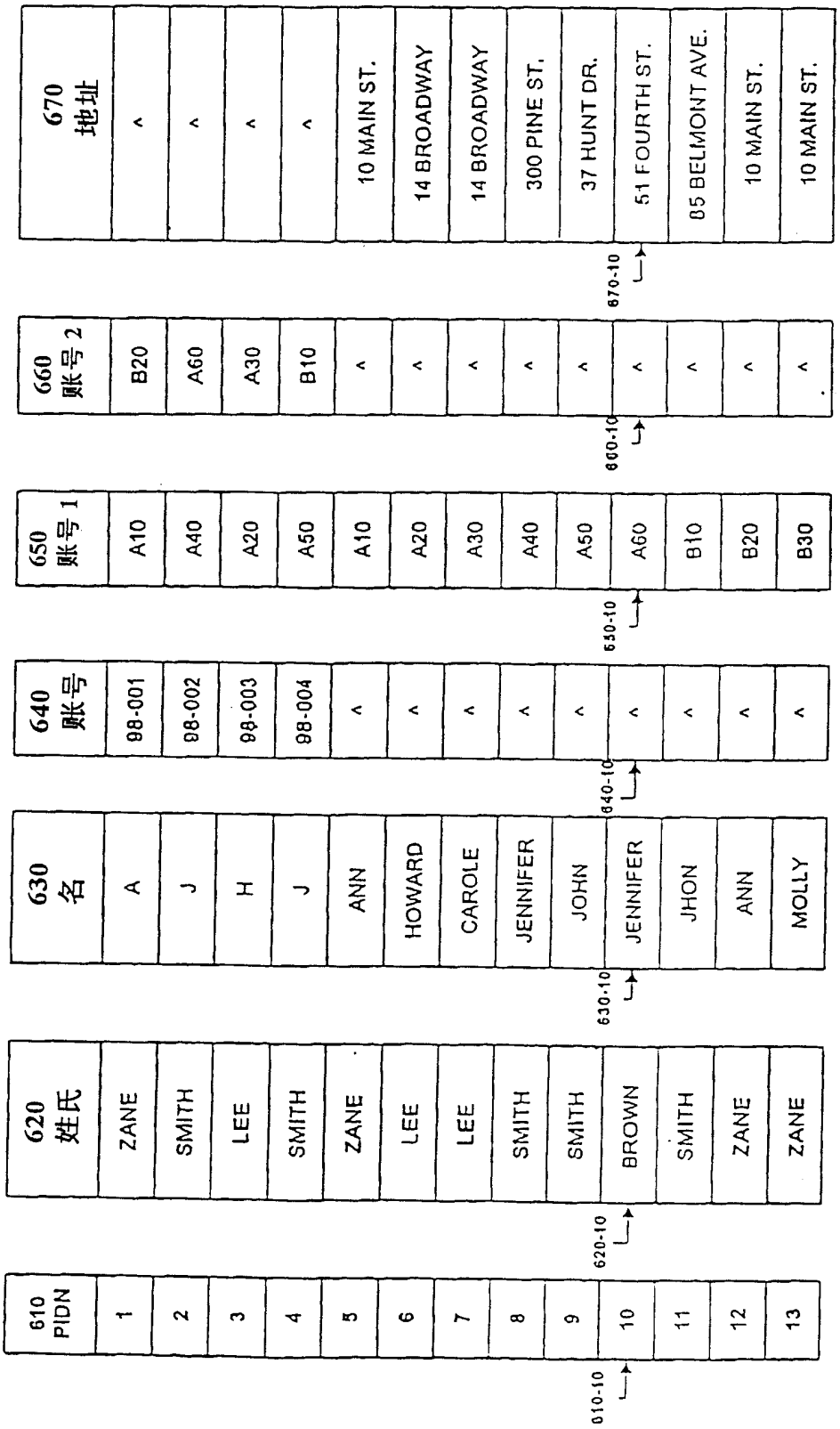

610 PIDN
620 姓氏
630 名
640 账号
650 账号 1
660 账号 2
670 地址
610-10
620-10
630-10
640-10
650-10
660-10
670-10
1 ZANE A 98-001 A10 B20 ^ 10 MAIN ST.
2 SMITH J 98-002 A40 A60 ^ 14 BROADWAY
3 LEE H 98-003 A20 A30 ^ 14 BROADWAY
4 SMITH J 98-004 A50 B10 ^ 300 PINE ST.
5 ZANE ANN ^ A10 ^ ^ 37 HUNT DR.
6 LEE HOWARD ^ A20 ^ ^ 51 FOURTH ST.
7 LEE CAROLE ^ A30 ^ ^ 85 BELMONT AVE.
8 SMITH JENNIFER ^ A40 ^ ^ 10 MAIN ST.
9 SMITH JOHN ^ A50 ^ ^ 10 MAIN ST.
10 BROWN JENNIFER ^ A60 ^ ^
11 SMITH JHON ^ B10 ^ ^
12 ZANE ANN ^ B20 ^ ^
13 ZANE MOLLY ^ B30 ^ ^

图 7

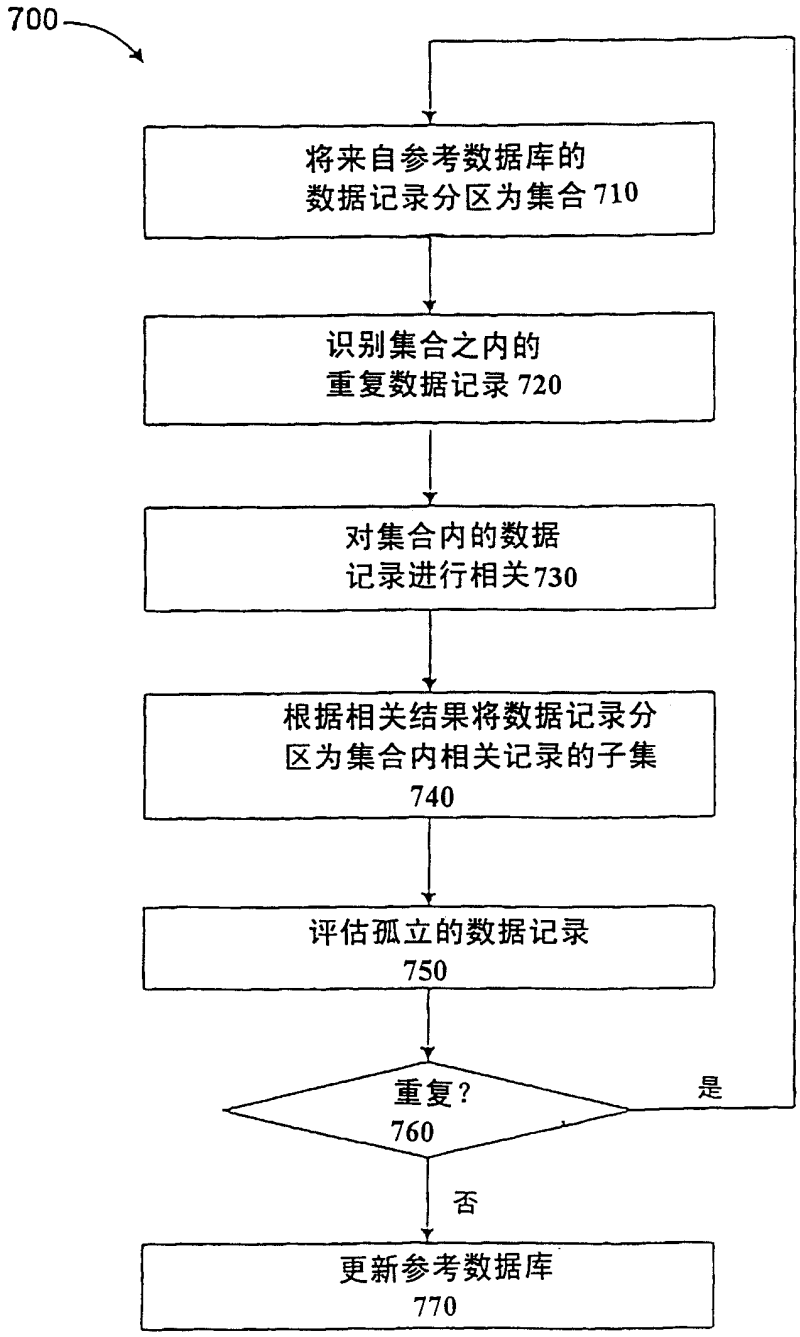
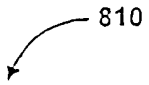
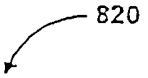

700
将来自参考数据库的
数据记录分区为集合 710
识别集合之内的
重复数据记录 720
对集合内的数据
记录进行相关730
根据相关结果将数据记录分
区为集合内相关记录的子集
740
评估孤立的数据记录
750
重复?
760
是
否
更新参考数据库
770

图 8



姓氏	PIDNs
BROWN	10
LEE	3, 6, 7
SMITH	2, 4, 8, 9, 11
ZANE	1, 5, 12, 13



姓氏	PIDNs
BROWN	10
LEE	3, 6
LEE	7
SMITH	2, 8
SMITH	4, 11
SMITH	9
ZANE	1, 5, 12
ZANE	13

图 9

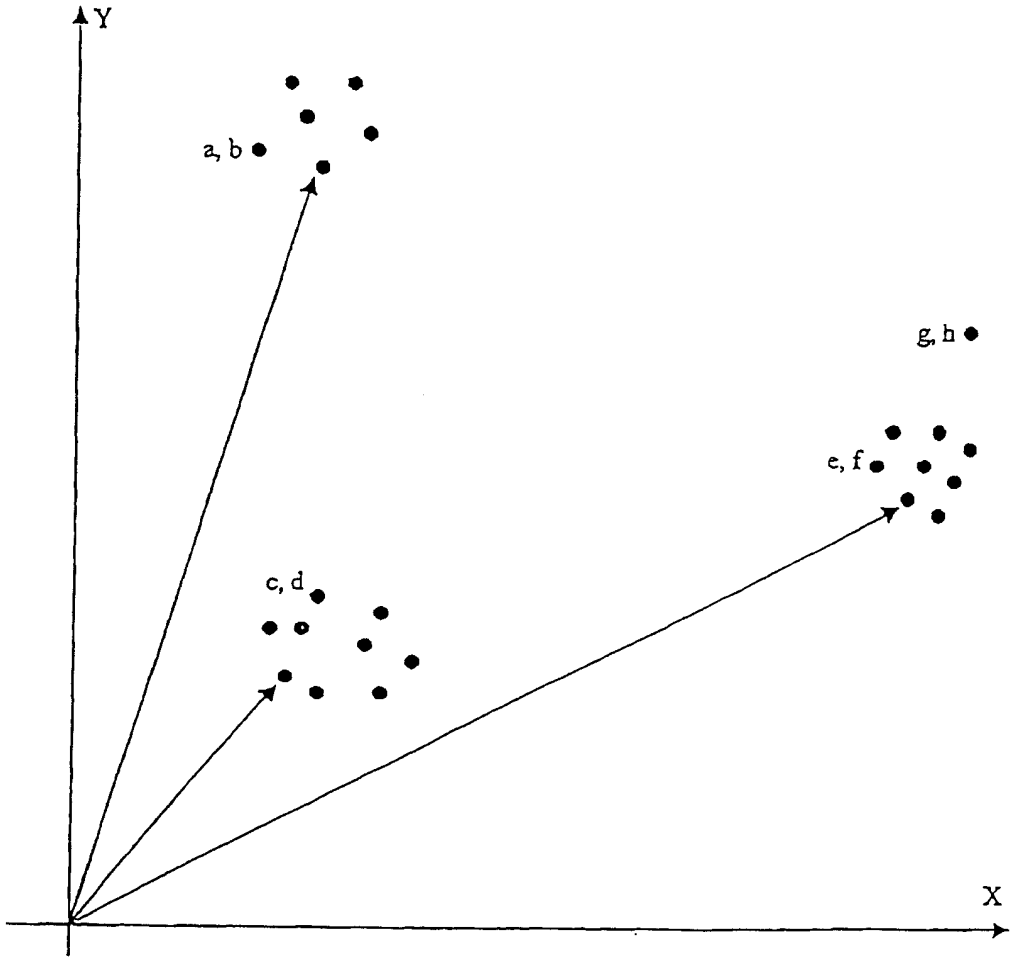


图 10

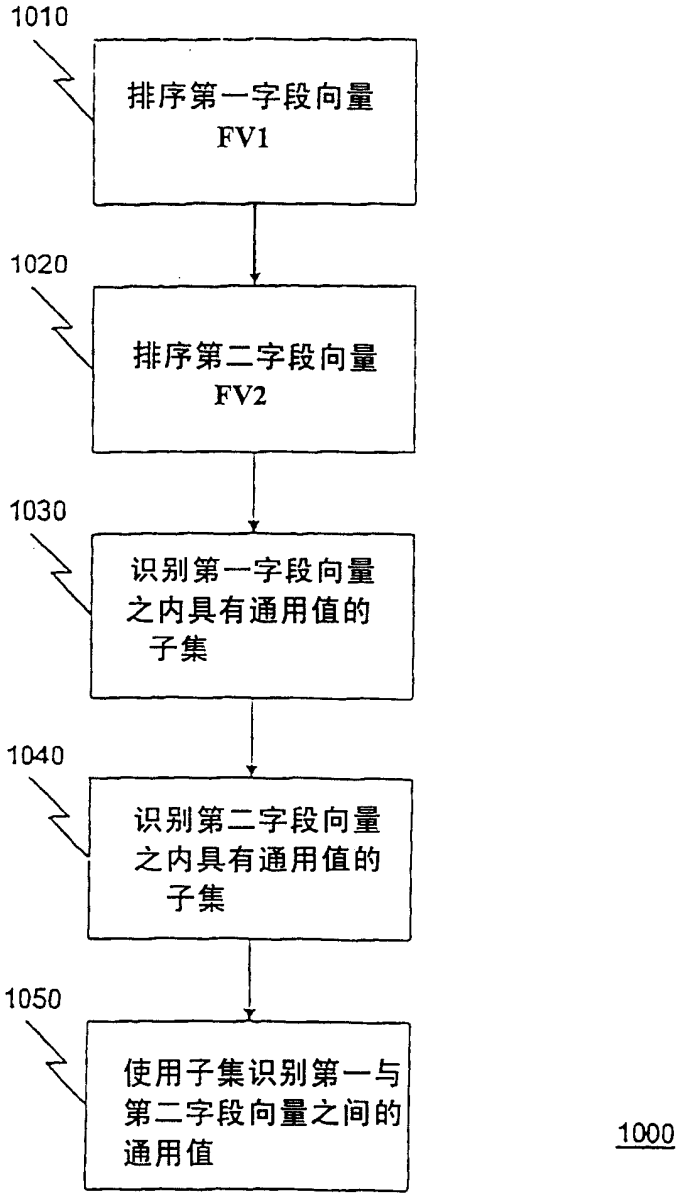

1010
排序第一字段向量 FV1
1020
排序第二字段向量 FV2
1030
识别第一字段向量之内具有通用值的子集
1040
识别第二字段向量之内具有通用值的子集
1050
使用子集识别第一与第二字段向量之间的通用值
1000

图 11

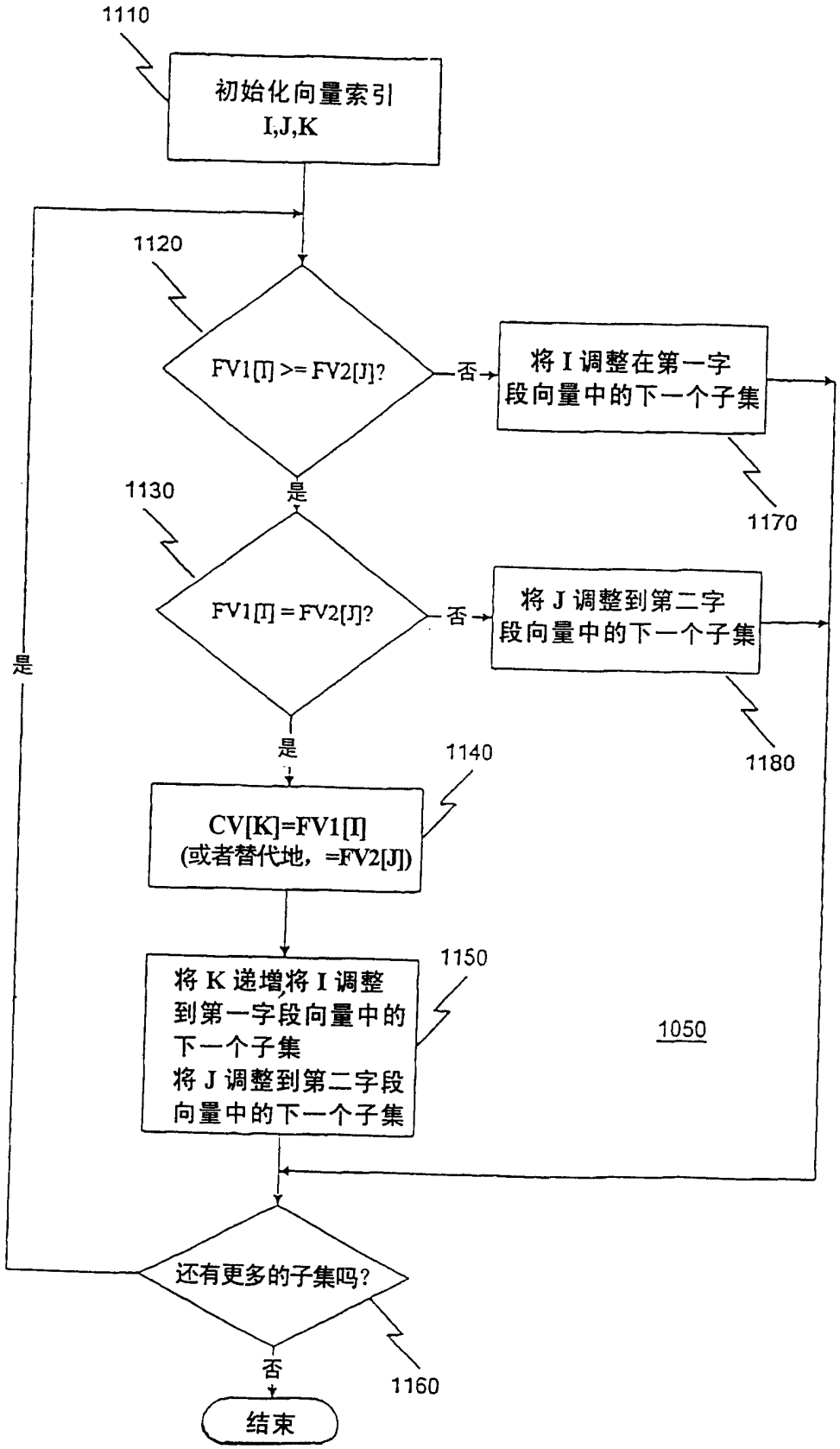
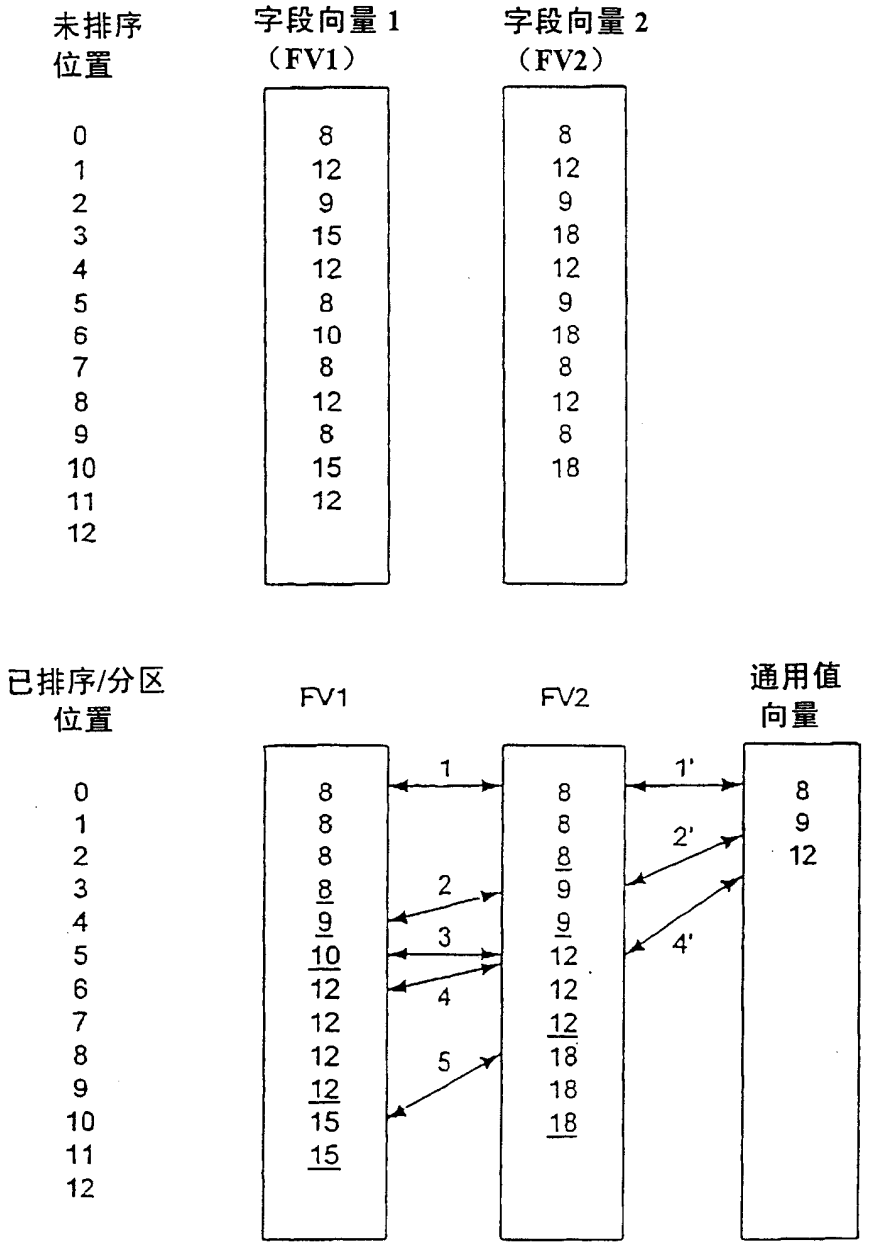

1110
初始化向量索引
I,J,K
1120
FV1[I] >= FV2[J]?
否
将 I 调整在第一字段向量中的下一个子集
1170
是
1130
FV1[I] = FV2[J]?
否
将 J 调整到第二字段向量中的下一个子集
1180
是
1140
CV[K]=FV1[I]
(或者替代地，=FV2[J])
1150
将 K 递增,将 I 调整到第一字段向量中的下一个子集
将 J 调整到第二字段向量中的下一个子集
1050
是
还有更多的子集吗?
否
1160
结束

图 12


未排序
位置
0
1
2
3
4
5
6
7
8
9
10
11
12
字段向量 1
（FV1）
8
12
9
15
12
8
10
8
12
8
15
12
字段向量 2
（FV2）
8
12
9
18
12
9
18
8
12
8
18
已排序/分区
位置
0
1
2
3
4
5
6
7
8
9
10
11
12
FV1
8
8
8
8
9
10
12
12
12
12
15
15
FV2
8
8
8
9
9
12
12
12
18
18
18
通用值
向量
8
9
12
1
2
3
4
5
1'
2'
4'