(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103874981 A
(43)申请公布日 2014.06.18

(21)申请号 201280050737.3
(22)申请日 2012.08.16
(30)优先权数据
13/211,031 2011.08.16 US
(85)PCT国际申请进入国家阶段日
2014.04.15
(86)PCT国际申请的申请数据
PCT/US2012/051044 2012.08.16
(87)PCT国际申请的公布数据
WO2013/025856 EN 2013.02.21
(71)申请人 全国学生资料库
地址 美国弗吉尼亚州
(72)发明人 道格拉斯·T·夏皮罗
黛安娜·吉勒姆
(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 杨黎峰 李欣
(51)Int.Cl.
G06F 7/02(2006.01)

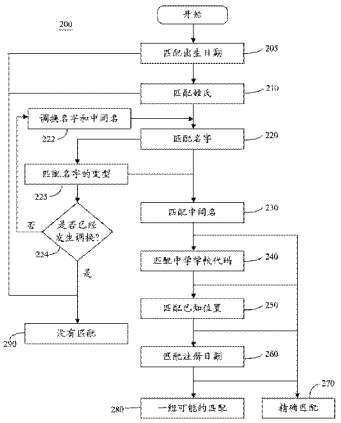
权利要求书2页 说明书6页 附图3页

(54)发明名称

高效率的学生记录匹配

(57)摘要

公开了一种用于有效地且智能地匹配学生注册记录的系统和方法。该方法例如可以用于追踪学生从中学机构到高等教育机构的进展情况且生成关于总的大学入学率的统计资料以通知政策决定。在示例性实施方式中，匹配算法分析学生姓名的常用变型以及中学机构与学生的当前已知地址之间的地理距离，以生成较高可信度的匹配。



CN 103874981 A

1. 一种用于匹配学生注册记录的计算机化的方法，所述方法包括：

使用处理器检索与学生相关联的注册记录；

确定所述学生的姓名的拼写的常用变型；以及

使用所述处理器来基于所述变型识别与所述学生相关联的其它注册记录。

2. 如权利要求 1 所述的方法，其中，该确定步骤包括：检查针对其他学生预先匹配的学生注册记录，并且识别学生的姓名的拼写的常用变型。

3. 如权利要求 1 所述的方法，其中，该识别步骤包括：

使用所述处理器调换所述学生的名字和中间名；以及

使用所述处理器来基于调换后的名字和中间名识别与所述学生相关联的其它注册记录。

4. 如权利要求 1 所述的方法，其中，该识别步骤包括：将与所述注册记录相关联的学校的位置与所述学生的当前已知地址进行匹配。

5. 如权利要求 1 所述的方法，其中，该识别步骤包括：将与所述注册记录相关联的注册日期和所述学生的来自另一注册记录的毕业日期进行比较。

6. 一种用于确定第一注册记录和第二注册记录是否与同一学生相关联的计算机化的方法，所述方法包括：

使用处理器将与所述第一注册记录相关联的学生的姓名的变型和与所述第二注册记录相关联的学生的姓名进行比较；

使用所述处理器确定与所述第一注册记录相关联的学校的位置；

使用所述处理器将与所述第一注册记录相关联的注册日期和与所述第二注册记录相关联的注册日期进行比较，以确定所述第一注册记录和所述第二注册记录是否与同一学生相关联；以及

如果确定所述第一注册记录和所述第二注册记录与同一学生相关联，则将匹配的指示存储在数据库中。

7. 如权利要求 6 所述的方法，其中，与所述第一注册记录相关联的所述学生的姓名的变型是基于与预先匹配的注册记录相关联的学生的姓名的变型而确定的。

8. 如权利要求 6 所述的方法，还包括：将与所述第一注册记录相关联的学校的位置和所述学生的当前已知地址进行匹配。

9. 一种用于追踪学生注册记录的计算机系统，所述系统包括：

存储设备，所述存储设备包含中学注册记录和高等教育注册记录；和

处理器，所述处理器配置成：

检索与学生相关联的注册记录；

确定所述学生的姓名的拼写的常用变型；以及

基于所述变型识别与所述学生相关联的其它注册记录。

10. 如权利要求 9 所述的系统，其中，所述处理器还配置成检查预先匹配的学生注册记录，并且识别学生的姓名的拼写的常用变型。

11. 如权利要求 9 所述的系统，其中，所述处理器还配置成：

调换所述学生的名字和中间名；以及

基于调换后的名字和中间名识别与所述学生相关联的其它注册记录。

12. 如权利要求 9 所述的系统，其中，所述处理器还配置成计算所述学生就读的学校与所述学生的当前已知地址之间的距离。

13. 如权利要求 9 所述的系统，其中，所述处理器还配置成将与所述注册记录相关联的注册日期和所述学生的来自另一注册记录的毕业日期进行比较。

高效率的学生记录匹配

背景技术

[0001] 确保从高中毕业的所有学生都为大学作好准备是国家的必要的事。存在大的有待拉近的成绩差距，并且我们的国家需要提高对于所有学生的教育标准以保持竞争力。虽然要达到这些目标存在一系列挑战，但是从联邦政府，许多州、地区和教育改革者以及慈善家的努力中显现出对于改变的强劲势头。

[0002] 一些群体的目标是确保从高中毕业的高中学生中的 80% 为大学作好准备，在达到该目标的过程中，着重于支持低收入和少数民族的年轻人。该策略的关键因素是确保丰富和可靠的数据的可得，利用这些数据在从校舍到州议会大厦的所有层面上制定出合理的政策和实践决策，这对学生的成绩和成果有很大影响。该数据既提供了 K-12 教育系统的性能的成果数据，而且还为项目的评估提供了强大的数据集。进一步地，用于测量大学准备策略的成功的最有可能的方式将是评估学生的高等教育的成绩和成果。

[0003] 一些需要答复的独特问题如下：

[0004] 我们高中毕业生在毕业后直接上大学或者在毕业后的两年内上大学的百分比是多少？

[0005] 我们学生团体年年保持登记入读的百分比是多少以及获得学位的百分比是多少？

[0006] 如何使对这些问题的答复针对各个地区、高中学校和关键的学生群体而不同？如何使对这些问题的答复针对各种类型的高等教育制度（例如，公共 / 私人，2 年制 /4 年制，选择性的 / 非选择性的）而不同？

[0007] K-12 项目（例如，所完成的高中课程、国家成绩测试分数）和高等教育成果之间的关系是什么？

[0008] 什么高等教育成果与我们的特定 K-12 项目（尤其是为了提高大学准备率和就读率而设计的项目）相关联？

[0009] 寻找这些问题的答案呈现出许多独特的且具有挑战性的难题。必须记录和维护的数据量本身就是挑战，即使是在这个数字化时代。然而，如果适当地组织数据且向数据提供适当的用于索引的标识符，则当前的数据库技术允许大量数据的管理。适当的索引允许有效地且高可靠性地进行检索。利用所积累的数据和适当的索引，可以答复部分上述问题，并且为我们的一些教育难题找到了解决方案。

[0010] 目前，国内大多数大学生的记录由全国学生资料库存储和保管。该机构通过维护来自学院和大学的反映其当前招生情况的更新信息的数据库，来提供许多政府职能所需的服务（例如，大学贷款服务）。该数据库目前持有许多记录，这些记录反映了从 1993 年以来的大学招生数据。全国的高中学校地区具有关于其学生的信息，包括学生在标准化考试中的成绩如何，学生的种族和影响教育的其它因素。

[0011] 因为在中学记录和高等教育记录之间出现的差异，导致将来自高中学校的记录匹配到大学注册记录的能力是难以满足的。这些问题可以包括简单的笔误、学生改变其姓名、学生的姓名通过什么形式记录。例如，在中学记录上，学生的姓名可以是 Jimmy Johnson，但

是在高等教育记录中，他的姓名被记录为James Johnson或者Jim Johnson。在这种情况下，即使一名学生在两个数据库中都有记录，匹配中学记录和高等教育记录也可能是困难的或者无法实现的。这会导致Jimmy没有去上大学的错误结论。

[0012] 为了有助于防止出现不匹配，用户将经常想要使用学生的社会安全号码(Social Security Number，SSN)作为标识符来获取匹配。然而，在当前的隐私法(例如家庭教育权利和隐私法(FERPA))下，不允许研究员出于研究目的而使用SSN来匹配记录。因此，代理处和研究员可被迫使用姓名和出生日期来进行搜索，导致准确度低的结果。

[0013] 因此，需要这样的系统，该系统在符合保护学生信息的法律的同时，允许在匹配高等教育数据库以提供与中学学生信息的匹配上增大可靠性和效率。

发明内容

[0014] 公开了一种用于有效地且智能地匹配学生注册记录的计算机系统和基于计算机的方法。该方法例如可以用于利用计算机追踪学生从中学机构到高等教育机构的进展情况且生成关于总的大学注册率的统计资料以报告政策决定。在示例性实施方式中，可将来自中学机构(例如高中学校)的记录与来自高等教育机构(例如大学和学院)的记录进行匹配。

[0015] 可以使用基于计算机的在计算机处理器中实现的匹配算法来匹配来自各个机构的记录，该匹配算法基于姓名、姓名的变型、出生日期、地理位置、注册日期和中学机构代码来匹配记录。在示例性实施方式中，该匹配算法可以基于学生姓名的拼写的常用变型来匹配姓名和出生日期。姓名的常用变型可以通过检查预先匹配的学生注册记录并识别学生姓名的拼写的常用变型来确定。该匹配算法也可以调换记录的名字和中间名，以将该记录与其它记录进行匹配。附加地，该匹配算法可以计算中学机构和学生的当前已知地址之间的地理距离，以生成较高可信度的匹配。此外，该匹配算法可以将与高等教育机构记录相关联的注册日期与学生从其中学机构毕业的日期进行比较。

附图说明

[0016] 图1示出中学记录数据库和高等教育记录数据库以及其中的个人记录；

[0017] 图2示出用于在中学数据库和高等教育数据库之间匹配记录的匹配算法；以及

[0018] 图3示出用于匹配的硬件系统。

具体实施方式

[0019] 在下面的详细描述中，参照附图，这些附图形成本发明的一部分并且以说明性的方式示出本发明的具体实施方式。足够详细地描述这些实施方式以使本领域的技术人员能够实现这些实施方式，并且应当理解，可以利用其它实施方式且可进行符合逻辑的处理变化。

[0020] 图1示出中学(高中学校)记录数据库100，高等教育记录数据库120。中学记录数据库100包括记录110、记录112、记录114、记录116，其中，记录110、记录112、记录114、记录116每个都包含单个学生的信息。记录110、记录112、记录114、记录116可以包括名字和姓氏、中间名或者首字母、SSN、出生日期(DOB)、中学学校的毕业日期、以及中学学校代码。或者，记录110、记录112、记录114、记录116可以包括学生的SSN、名字和姓氏、种族和

高中学校代码。一些记录 110、记录 112、记录 114、记录 116 可以包括 SSN，其它记录可以不包括 SSN。进一步地，个人记录 110、个人记录 112、个人记录 114、个人记录 116 每个都可以包括除名字和姓氏以及出生日期以外的不同信息。本质上，记录 110、记录 112、记录 114、记录 116 可以包括学生的许多标识符和属性，且不应当被限制于所给出的示例。

[0021] 高等教育数据库 120 包括学生记录 130、学生记录 132、学生记录 134、学生记录 136、学生记录 138。这些记录 130、记录 132、记录 134、记录 136、记录 138 中的每个都可以包括与中学记录数据库 100 中的记录 110、记录 112、记录 114、记录 116 相同的信息，例如，学生的名字和姓氏、中间名或者首字母、SSN、DOB 和其它标识信息。应该理解，记录 130、记录 132、记录 134、记录 136、记录 138 可以包括其它信息且这些信息可以不限于所给出的示例。高等教育记录 130、记录 132、记录 134、记录 136、记录 138 由高等教育机构提供且反映这些机构的各自的招生记录。进一步地，当在高等教育数据库 120 中创建高等教育记录 130、记录 132、记录 134、记录 136、记录 138 时，这些记录中的每个记录都被提供独特的或特定的高等教育标识符。

[0022] 中学记录数据库 100 和高等教育记录数据库 120 可以配置在任何允许有效地存储和检索数据库的数据库结构中。中学数据库 100 可以包括为具有特定属性而已经被预先选择的记录。例如，中学数据库 100 中的所有记录可以是进入某高中学校或者进入某地区的高中学校的学生。进一步地，中学数据库 100 可以包括特定种族的学生的记录或者在特定的标准化考试中得分高于或低于某阈值分数的学生的记录。这些因素（例如，与追踪教育成绩相关的因素）的任何组合可以用于确定什么样的记录包括在中学数据库 100 中。

[0023] 为了追踪教育进展，需要确定中学数据库 100 中这些预先选择的学生中哪些学生继续进入高等教育机构。这样做，使得中学数据库 100 中的记录与高等教育数据库 120 中的记录相匹配，如图 2 所示的匹配过程 200 所示。

[0024] 图 2 示出根据示例性实施方式的可用来匹配记录的计算机化的匹配过程 200。具体地，过程 200 可以用于匹配学生记录。计算机化的过程 200 可以在使用计算机可读代码的处理器上实现，该处理器例如为服务器（例如，图 3 中的服务器 310）。计算机化的过程 200 可以实现成匹配存储在同一数据库中或不同数据库中的学生记录。例如，过程 200 可以用于将来自中学数据库 100 的记录 110 与其在高等教育数据库 120 中的对应记录进行匹配。

[0025] 计算机化的过程 200 开始于获取将要匹配到数据库中的记录的未匹配的记录。例如，未匹配的记录可以是学生记录 110，该学生记录 110 包括关于学生的信息组，例如学生的名字和姓氏、学生的中间名或首字母、DOB、中学学校代码、中学学校邮政编码以及中学学校毕业日期，并且该数据库可以是高等教育数据库 120。

[0026] 接着，在计算机处理步骤 205 处，计算机化的过程 200 将未匹配的记录的 DOB 与该数据库中的记录的 DOB 进行匹配。计算机化的过程 200 首先执行精确的字符匹配功能，该功能要求匹配是精确的。如果发现一个或者多个包含精确姓氏的记录，则计算机化的过程 200 进入计算机处理步骤 210。如果没有找到匹配，则计算机化的过程 200 进行局部的匹配，从而允许一个或者多个字符是不准确的。例如，计算机化的过程 200 可以发现未匹配的记录的为 1988 年 1 月 31 的 DOB 与数据库中的记录的为 1988 年 1 月 21 的 DOB 相匹配。应该理解，匹配记录中的 DOB 所需的匹配字符的数量可以改变。

[0027] 如果没有识别到匹配，则计算机化的过程 200 进入计算机处理步骤 290，并且指示

出没有找到匹配。如果一个或者多个记录的 DOB 与未匹配的记录的 DOB 相匹配，则计算机化的过程 200 进入计算机处理步骤 210。

[0028] 在计算机处理步骤 210 处，计算机化的过程 200 将未匹配的记录的姓氏和数据库中的记录的姓氏进行匹配。计算机化的过程 200 首先执行严格的字符匹配功能，该功能要求匹配是精确的。如果发现一个或者多个包含精确姓氏的记录，则计算机化的过程 200 进入计算机处理步骤 220。如果没有找到匹配，则计算机化的过程 200 进行局部的匹配，从而允许一个或者多个字符是不准确的。例如，计算机化的过程 200 可以发现未匹配的记录的为 Weinstein 的姓氏与数据库中的记录的为 Wienstein 的姓氏相匹配。应该理解，匹配记录中的姓氏所需的匹配字符的数量可以改变。

[0029] 如果没有识别到匹配，则计算机化的过程 200 进入计算机处理步骤 290，并且指示没有找到匹配。如果一个或者多个记录的姓氏与未匹配的记录的姓氏相匹配，则计算机化的过程 200 进入计算机处理步骤 220。在计算机处理步骤 220 处，计算机化的过程 200 使用精确的匹配来将未匹配的记录的名字与在计算机处理步骤 210 处所匹配到的记录的名字进行匹配。如果没有找到精确的匹配，则计算机化的过程 200 将未匹配的记录的名字与在计算机处理步骤 210 处所匹配到的记录进行局部匹配。计算机化的过程 200 可以遵循与在计算机处理步骤 210 处所使用的局部匹配相同的标准或者不同的标准。如果找到一个或者多个匹配，则计算机化的过程 200 进入计算机处理步骤 230。

[0030] 如果没有找到匹配，则计算机化的过程 200 进入计算机处理步骤 225，并且使用来自姓名变型数据库的姓名来执行名字的匹配。姓名变型数据库提供且排列名字的已知变型。该变型数据库可以包括使用数据库内的记录所编辑的姓名历史变型的经验性分析。也可以基于使用计算机化的过程 200 所匹配的或者由人类分析师所匹配的记录来填充该变型数据库。可以基于数据库中的新记录和使用计算机化的过程 200 所进行的或者由人类分析师所进行的新匹配，来连续地或者周期性地更新该变型数据库。例如，对于名字 Lyndsey，该变型数据库可以包括一组变型，例如 Lindsey、Lyndsay、Lindsay、Lindsi 等。作为另一个示例，对于名字 Cami，该变型数据库可以包括一组变型，例如 Camille、Camile、Camilla、Camill 等。作为另一个示例，对于名字 Christopher，该变型数据库可以包括这些变型，例如 Chris、Cris、Christofer 等。

[0031] 为了使用姓名变型数据库中的姓名来执行名字的匹配，计算机化的过程 200 首先识别姓名变型数据库中包括未匹配的记录的名字的一组变型。如果姓名变型数据库中不存在包括未匹配的记录的名字的一组变型，则计算机化的过程 200 进入计算机处理步骤 224。否则，计算机化的过程 200 将该组变型中的所有姓名与数据库中的在计算机处理步骤 210 处所匹配到的记录的名字进行比较。如果数据库中的在计算机处理步骤 210 处所匹配到的姓名都不包括该组变型中的任一姓名，则计算机化的过程 200 进入计算机处理步骤 224。否则，计算机化的过程 200 可以根据匹配到的姓名的排序来排列与该组变型中的姓名的匹配。在完成匹配后，计算机化的过程 200 进入计算机处理步骤 230。

[0032] 在计算机处理步骤 224 处，计算机化的过程 200 确定名字和中间名是否已被调换。如果名字和中间名已被预先调换，则计算机化的过程 200 进入计算机处理步骤 290，并且指示没有找到匹配。如果名字和中间名没有被预先调换，则计算机化的过程 200 进入计算机处理步骤 222。在计算机处理步骤 222 处，将未匹配的记录的名字和中间名调换，从而出于

匹配目的使中间名变成名字。在调换名字和中间名之后，计算机化的过程 200 返回到计算机处理步骤220，并尝试将未匹配的记录的中间名与数据库中的记录的名字进行匹配。如果在计算机处理步骤 220 处进行匹配，则计算机化的过程 200 进入计算机处理步骤 270。

[0033] 在计算机处理步骤 230 处，计算机化的过程 200 将未匹配的记录的中间名或者中间名首字母与在计算机处理步骤 220 处使用精确匹配所匹配到的记录的中间名或者中间名首字母进行匹配。如果没有找到精确的匹配，则计算机化的过程 200 将未匹配的记录的中间名或者中间名首字母与在计算机处理步骤 220 处所匹配到的记录进行局部匹配。过程 200 可以遵循与在计算机处理步骤 210 和计算机处理步骤 220 处所使用的局部匹配相同的标准或者不同的标准。如果仅找到一个匹配，则计算机化的过程 200 进入计算机处理步骤 230，并且指示发现了精确匹配。如果计算机处理步骤 230 产生不止一个匹配，则计算机化的过程 200 进入计算机处理步骤 240。

[0034] 在计算机处理步骤 240 处，计算机化的过程 200 将未匹配的记录的中学学校代码与在计算机处理步骤 230 处所匹配到的记录的中学学校代码进行匹配。如果仅有一个精确匹配，则计算机化的过程 200 进入计算机处理步骤 270。如果没有匹配或者存在不止一个匹配，则计算机化的过程 200 进入计算机处理步骤 250。

[0035] 在计算机处理步骤 250 处，计算机化的过程 200 利用区域映射数据库来将与未匹配的记录中的中学学校相关联的邮政编码和数据库的记录中的学生的邮政编码相匹配。区域映射数据库包括将未匹配的记录中的中学学校周围的邮政编码关联到中学学校代码的数据。于是可将中学学校的关联的邮政编码与在计算机处理步骤 240 处或在计算机处理步骤 230 处(如果在计算机处理步骤 240 处没有出现匹配)所匹配到的记录中的学生邮政编码相匹配。例如，未匹配的记录可以包括中学学校代码，基于区域映射数据库，该中学学校代码与如下邮政编码相关联 :22040、22041、22042、22043、22044 和 22046。区域映射数据库可以包括与使用数据库内的记录所编辑的学校相关的邮政编码的经验性分析。也可以基于使用计算机化的过程 200 所匹配的或者由人类分析师所匹配的记录来填充该区域映射数据库。可以基于数据库中的新记录和使用计算机化的过程 200 所进行的或者由人类分析师所进行的新匹配，来连续地或者周期性地更新该区域映射数据库。

[0036] 为了将与未匹配的记录相关联的邮政编码和数据库中的学生记录的邮政编码进行匹配，计算机化的过程 200 首先识别区域映射数据库中的一组与未匹配的记录的中学学校代码相关联的邮政编码。如果区域映射数据库中没有与未匹配的记录的中学学校代码相关联的邮政编码，则计算机化的过程 200 进入计算机处理步骤 260。否则，计算机化的过程 200 将该组与未匹配的记录相关联的邮政编码与数据库中的在计算机处理步骤 240 或者步骤 230 处所匹配到的记录的学生邮政编码相比较。如果仅有一个精确匹配，则计算机化的过程 200 进入计算机处理步骤 270。如果不存在匹配或者存在不止一个匹配，则计算机化的过程 200 进入计算机处理步骤 260。

[0037] 在计算机处理步骤 260 处，计算机化的过程 200 将未匹配的记录的中学学校毕业日期与在计算机处理步骤 250 处、或者在计算机处理步骤 240 处(如果在计算机处理步骤 250 处没有出现匹配)、或者在计算机处理步骤 230 处(如果在计算机处理步骤 240 和计算机处理步骤 250 处没有出现匹配)所匹配到的记录的注册日期在可接受的范围内进行匹配。例如，如果中学学校毕业日期是 2008 年 6 月 15 日，则其可以与 2008 年 8 月或者 9 月中的注

册日期相匹配。如果仅有一个精确匹配，则计算机化的过程 200 进入计算机处理步骤 230。如果不存在匹配或者存在不止一个匹配，则计算机化的过程 200 进入计算机处理步骤 280。在计算机处理步骤 280 处，计算机化的过程 200 报告所有可能的匹配。

[0038] 应该理解，计算机化的过程 200 可以在范围上进行改变，并不应该被限制于所描述的具体过程。可将一个或者多个步骤从过程 200 中省略，也可添加额外的步骤。匹配的计算机化的过程 200 提供了多种优势。例如，提供了使用姓名的多种拼写和变换来匹配记录的过程。该过程也提供了使用除仅姓名的匹配之外的信息和使用由中学记录所提供的有限信息而减小可能的匹配范围的能力。

[0039] 图 3 示出包括第一数据存储器 320 和第二数据存储器 322 的系统 300。系统 300 还包括连接到数据存储器 320 和数据存储器 322 的服务器 310。在一个实施方式中，服务器 310 是来自 IBM 公司的型号为 3650 的数据库服务器，该数据库服务器运行甲骨文（Oracle）公司的软件。数据存储器 320 和数据存储器 322 可以是 IBM 公司的 DS4800 存储系统的一部分。在一个实施方式中，中学数据库 100 位于数据存储器 320 中，高等教育数据库 120 位于数据存储器 322 中。服务器 310 与数据存储器 320、数据存储器 322 进行通信，并且在数据库之间传输信息。进一步地，服务器 310 运行用于确定一个数据库的记录是否与另一个数据库的记录相匹配的算法。在另一个实施方式中，数据库 100 和数据库 120 都位于同一个数据存储器中，但是位于该数据存储器的不同部分。

[0040] 上述描述和附图示出了实现本发明的目标、特征和优势的优选实施方式。尽管上文已经描述了某些优势和优选实施方式，但是本领域的技术人员将意识到，可以进行代替、增加、删除、修改和 / 或其它改变，而不脱离本发明的精神或范围。因此，本发明不由以上描述所限定，而仅由任何后续的要求其优先权的非临时申请的权利要求的范围所限定。

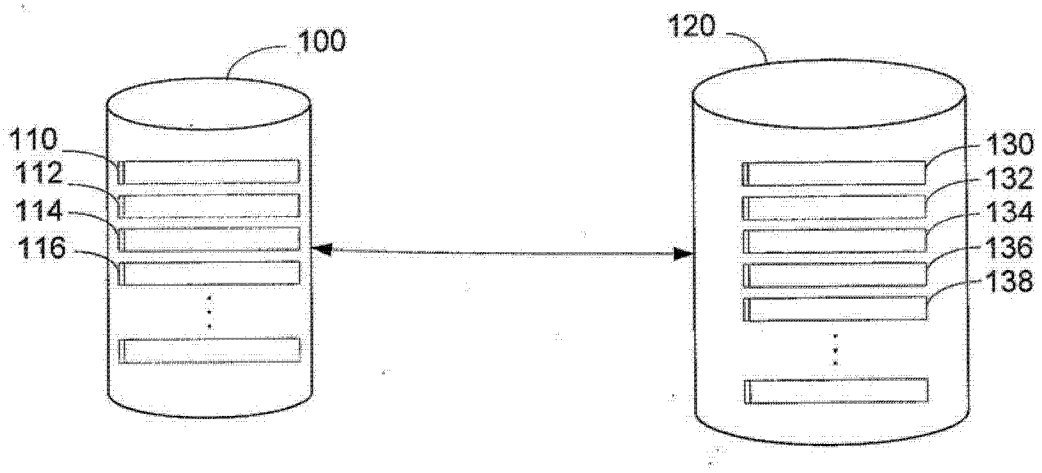


图 1

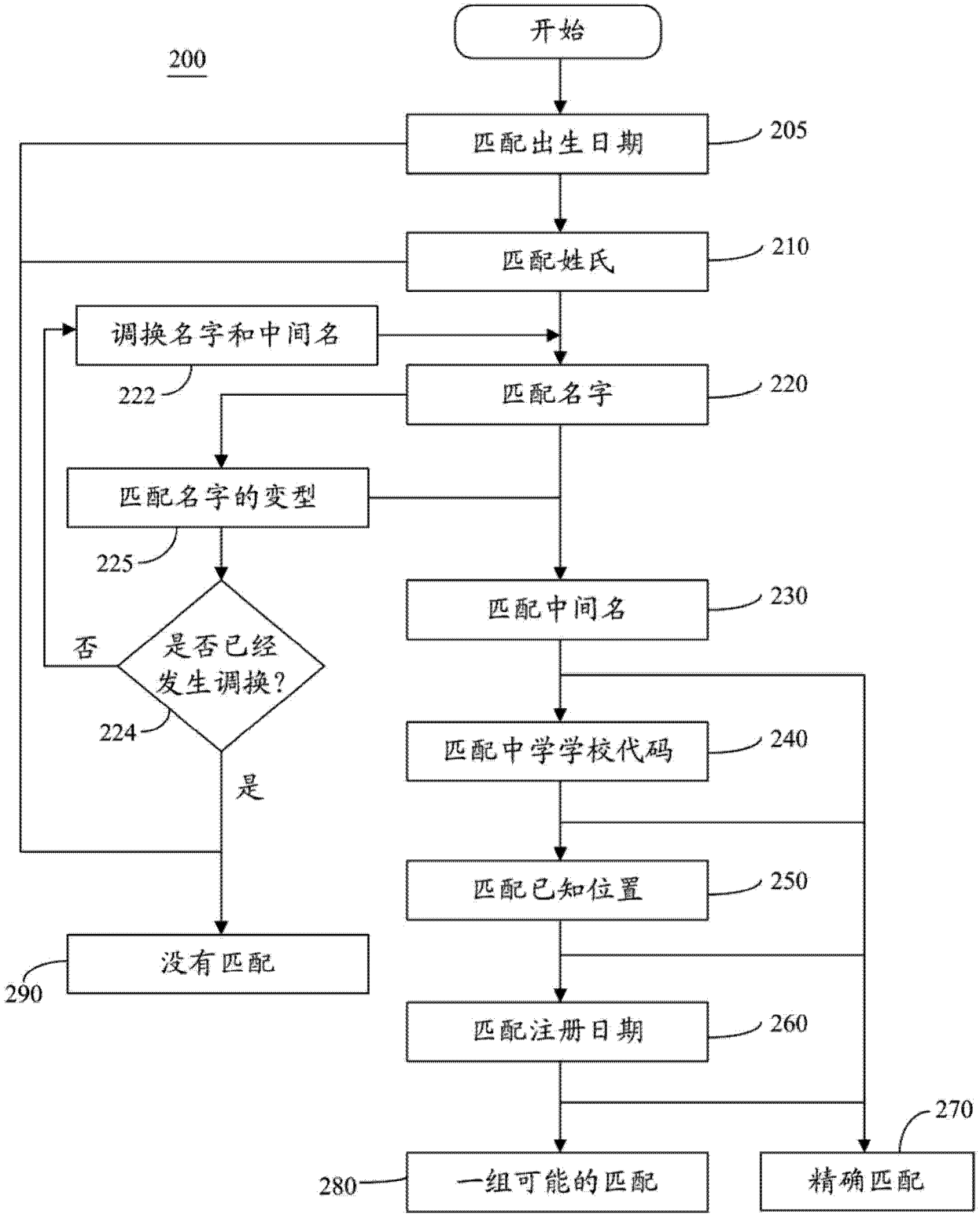

200
开始
匹配出生日期
205
匹配姓氏
210
调换名字和中间名
222
匹配名字
220
匹配名字的变型
225
匹配中间名
230
否
是否已经发生调换?
224
匹配中学学校代码
240
是
匹配已知位置
250
没有匹配
290
匹配注册日期
260
270
280
一组可能的匹配
精确匹配

图 2

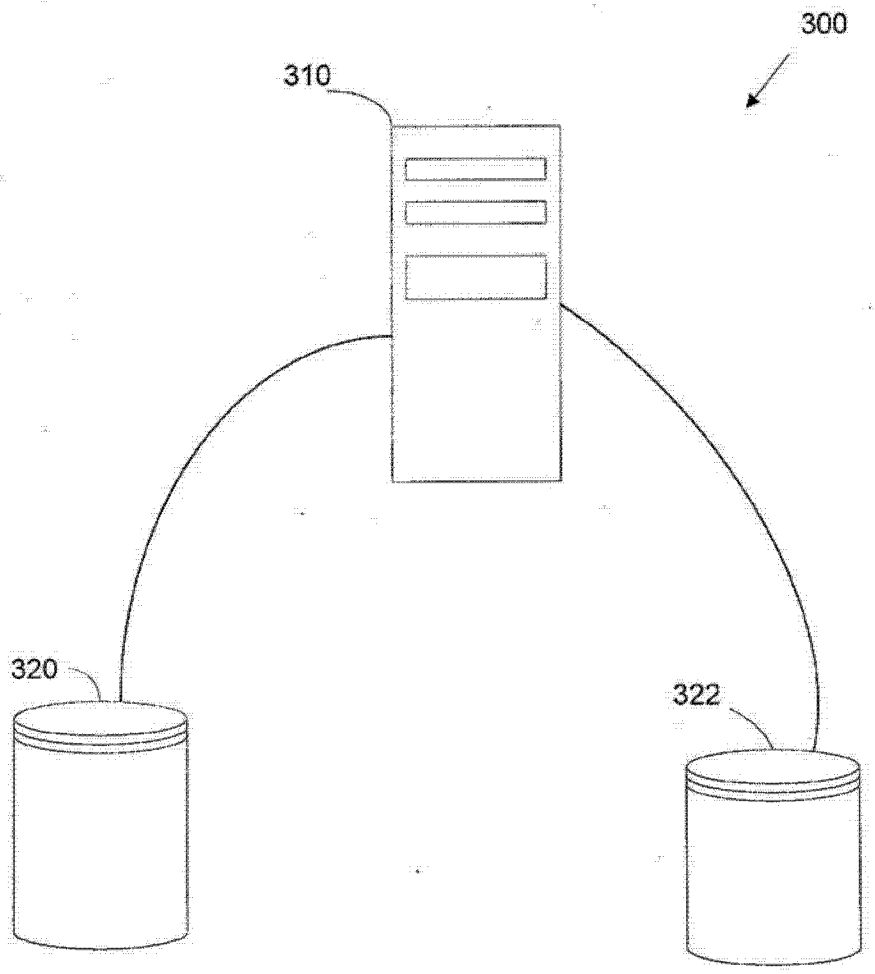


图 3