(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105051729 B
(45)授权公告日 2020.09.04

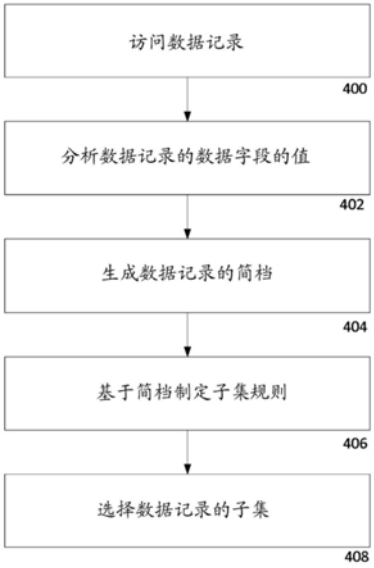
(21)申请号 201480004942.5
(22)申请日 2014.01.31
(65)同一申请的已公布的文献号
申请公布号 CN 105051729 A
(43)申请公布日 2015.11.11
(30)优先权数据
61/759,799 2013.02.01 US
13/827,558 2013.03.14 US
(85)PCT国际申请进入国家阶段日
2015.07.15
(86)PCT国际申请的申请数据
PCT/US2014/014186 2014.01.31
(87)PCT国际申请的公布数据
WO2014/121092 EN 2014.08.07
(73)专利权人 起元技术有限责任公司
地址 美国马萨诸塞州
(72)发明人 M.A.伊斯曼 R.A.爱泼斯坦 R.豪格 A.F.罗伯茨 J.罗尔斯顿 J.L.理查森 J.普尼奥沃
(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
代理人 刘新宇
(51)Int.Cl.
G06F 16/21(2019.01)
G06F 16/9535(2019.01)
G06F 11/36(2006.01)
(56)对比文件
US 2004260711 A1,2004.12.23,
US 2004049492 A1,2004.03.11,
US 2005055369 A1,2005.03.10,
US 6163774 A,2000.12.19,
US 2005114369 A1,2005.05.26,
US 2005071320 A1,2005.03.31,
CN 101271471 A,2008.09.24,
CN 101911069 A,2010.12.08,
CN 103348598 A,2013.10.09,
审查员 谢婉婉
权利要求书4页 说明书13页 附图5页

(54)发明名称

数据记录的选择

(57)摘要

一种计算机实现的方法,包括:访问多个数据记录,每一个数据记录具有多个数据字段。该方法还包括分析所述多个数据记录中的至少一些的一个或多个数据字段的值,基于所述分析,生成多个数据记录的简档。该方法还包括基于所述简档,制定至少一个子集规则;以及基于所述至少一个子集规则,从多个数据记录中选择数据记录的子集。



CN 105051729 B

1.一种计算机实现的方法,包括:

接收执行信息,所述执行信息指示在对数据记录的第一集合进行处理时数据处理应用的处理规则被执行的次数,其中在对特定数据记录的处理期间所述处理规则是否由所述数据处理应用执行直接或间接地取决于该特定数据记录中一个或多个数据字段中的各数据字段中的值;

分析数据记录的第二集合中的至少一些数据记录各自的一个或多个数据字段的值,所述分析包括生成数据记录的所述第二集合中的各数据记录的一个或多个数据字段中的各数据字段的简档,数据字段的简档表征该数据字段中的值;

获得基于(i)所生成的简档以及(ii)所述执行信息制定的子集规则,其中所述子集规则用于将所述数据字段中的特定的一个数据字段标识为目标数据字段;以及

根据所述子集规则而从数据记录的所述第二集合中选择数据记录的子集,其中该数据记录的子集的选择是基于数据记录的所述第二集合的目标数据字段的值进行的。

2.如权利要求1所述的方法,其中获得子集规则包括制定所述子集规则,制定所述子集规则包括基于所述数据字段中所标识的一个数据字段的基数将该一个数据字段标识为所述目标数据字段。

3.如权利要求2所述的方法,其中所述目标数据字段具有所述数据记录中的不同的值的集合,以及其中选择数据记录的子集包括:选择数据记录以使得在所选择的子集中有至少一个数据记录具有所述目标数据字段的各个不同的值。

4.如权利要求1所述的方法,其中生成简档包括对数据记录的所述第二集合中的数据记录的第一数据字段的值进行分类;以及

其中获得子集规则包括制定所述子集规则,制定所述子集规则包括基于所述分类,将所述第一数据字段标识为所述目标数据字段。

5.如权利要求4所述的方法,其中所述目标数据字段具有数据记录的所述第二集合中的不同的值的集合,其中选择数据记录的子集包括:选择数据记录以使得在所选择的子集中有至少一个数据记录具有所述目标数据字段的各个不同的值。

6.如权利要求1所述的方法,其中所述子集规则将第一数据字段标识为第一目标数据字段以及将第二数据字段标识为第二目标数据字段。

7.如权利要求6所述的方法,其中选择数据记录的子集包括:基于所述第一目标数据字段的不同的值的第一集合和所述第二目标数据字段的不同的值的第二集合的组合,选择数据记录的子集。

8.如权利要求1所述的方法,其中生成简档包括标识数据记录的所述第二集合中的、经由第一数据字段的值相关的数据记录之间的关系;以及

其中至少一个子集规则包括关系的标识。

9.如权利要求8所述的方法,其中选择数据记录的子集包括:

选择第一数据记录;以及

选择经由在子集规则中标识的关系与第一数据记录相关的一个或多个第二数据记录。

10.如权利要求8所述的方法,其中,数据记录之间的关系包括在数据记录的所述第二集合中的数据记录和数据记录的第三集合中的数据记录之间的关系。

11.如权利要求1所述的方法,其中生成简档包括:

对于所述数据记录中的至少一些数据记录生成伪字段;以及

使用累计值填充每个相应数据记录的伪字段,其中,基于第一数据记录和与第一数据记录相关的至少一个其它数据记录确定第一数据记录的累计值,

其中,第一数据记录与所述至少一个其它数据记录经由第一数据字段的值相关。

12.如权利要求11所述的方法,包括基于第一数据记录中的第二数据字段的值和每一个相关的其它数据记录的第二数据字段的值的总和确定累计值。

13.如权利要求1所述的方法,其中获得子集规则包括接收所述子集规则。

14.如权利要求1所述的方法,包括将数据记录的所选子集提供给数据处理应用。

15.如权利要求14所述的方法,包括:

基于所述数据处理应用的结果制定第二子集规则;以及

基于所述第二子集规则选择数据记录的第二子集。

16.一种计算系统,包括:

至少一个处理器,配置为:

接收执行信息,所述执行信息指示在对数据记录的第一集合进行处理时数据处理应用的处理规则被执行的次数,其中在对特定数据记录的处理期间所述处理规则是否由所述数据处理应用执行直接或间接地取决于该特定数据记录中一个或多个数据字段中的各数据字段中的值;

分析数据记录的第二集合中的至少一些数据记录各自的一个或多个数据字段的值,所述分析包括生成数据记录的所述第二集合中的各数据记录的一个或多个数据字段中的各数据字段的简档,数据字段的简档表征该数据字段中的值;

获得基于(i)所生成的简档以及(ii)所述执行信息制定的子集规则,其中所述子集规则用于将所述数据字段中的特定的一个数据字段标识为目标数据字段;以及

根据所述子集规则而从数据记录的所述第二集合中选择数据记录的子集,其中该数据记录的子集的选择是基于数据记录的所述第二集合的目标数据字段的值进行的。

17.如权利要求16所述的计算系统,其中获得子集规则包括制定所述子集规则,制定所述子集规则包括基于所述数据字段中所标识的一个数据字段的基数将该一个数据字段标识为所述目标数据字段。

18.如权利要求17所述的计算系统,其中所述目标数据字段具有所述数据记录中的不同的值的集合,以及其中选择数据记录的子集包括:选择数据记录以使得在所选择的子集中有至少一个数据记录具有所述目标数据字段的各个不同的值。

19.如权利要求16所述的计算系统,其中生成简档包括对数据记录的所述第二集合中的数据记录的第一数据字段的值进行分类;以及

其中获得子集规则包括制定所述子集规则,制定所述子集规则包括基于所述分类,将所述第一数据字段标识为所述目标数据字段。

20.如权利要求19所述的计算系统,其中所述目标数据字段具有数据记录的所述第二集合中的不同的值的集合,其中选择数据记录的子集包括:选择数据记录以使得在所选择的子集中有至少一个数据记录具有所述目标数据字段的各个不同的值。

21.如权利要求16所述的计算系统,其中所述子集规则将第一数据字段标识为第一目标数据字段以及将第二数据字段标识为第二目标数据字段。

22.如权利要求21所述的计算系统，其中选择数据记录的子集包括：基于所述第一目标数据字段的不同的值的第一集合和所述第二目标数据字段的不同的值的第二集合的组合，选择数据记录的子集。

23.如权利要求16所述的计算系统，其中生成简档包括标识数据记录的所述第二集合中的、经由第一数据字段的值相关的数据记录之间的关系；以及

其中至少一个子集规则包括关系的标识。

24.如权利要求23所述的计算系统，其中选择数据记录的子集包括：

选择第一数据记录；以及

选择经由在子集规则中标识的关系与第一数据记录相关的一个或多个第二数据记录。

25.如权利要求16所述的计算系统，其中生成简档包括：

对于所述数据记录中的至少一些数据记录生成伪字段；以及

使用累计值填充每个相应数据记录的伪字段，其中，基于第一数据记录和与第一数据记录相关的至少一个其它数据记录确定第一数据记录的累计值，

其中，第一数据记录与所述至少一个其它数据记录经由第一数据字段的值相关。

26.如权利要求25所述的计算系统，其中所述至少一个处理器被配置为基于第一数据记录中的第二数据字段的值和每一个相关的其它数据记录的第二数据字段的值的总和确定累计值。

27.如权利要求16所述的计算系统，其中所述至少一个处理器被配置为将数据记录的所选子集提供给数据处理应用。

28.如权利要求27所述的计算系统，其中所述至少一个处理器被配置为：

基于所述数据处理应用的结果制定第二子集规则；以及

基于所述第二子集规则选择数据记录的第二子集。

29.一种计算系统，包括：

用于接收执行信息的装置，所述执行信息指示在对数据记录的第一集合进行处理时数据处理应用的处理规则被执行的次数，其中在对特定数据记录的处理期间所述处理规则是否由所述数据处理应用执行直接或间接地取决于该特定数据记录中一个或多个数据字段中的各数据字段中的值；

用于分析数据记录的第二集合中的至少一些数据记录各自的一个或多个数据字段的值的装置，用于分析的所述装置包括生成数据记录的所述第二集合中的各数据记录的一个或多个数据字段中的各数据字段的简档的装置，其中数据字段的简档表征该数据字段中的值；

用于获得基于(i)所生成的简档以及(ii)所述执行信息制定的子集规则的装置，其中所述子集规则用于将所述数据字段中的特定的一个数据字段标识为目标数据字段；以及

用于根据所述子集规则而从数据记录的所述第二集合中选择数据记录的子集的装置，其中该数据记录的子集的选择是基于数据记录的所述第二集合的目标数据字段的值进行的。

30.一种计算机可读介质，用于存储指令以使计算机系统：

接收执行信息，所述执行信息指示在对数据记录的第一集合进行处理时数据处理应用的处理规则被执行的次数，其中在对特定数据记录的处理期间所述处理规则是否由所述数

据处理应用执行直接或间接地取决于该特定数据记录中一个或多个数据字段中的各数据字段中的值；

分析数据记录的第二集合中的至少一些数据记录各自的一个或多个数据字段的值，所述分析包括生成数据记录的所述第二集合中的各数据记录的一个或多个数据字段中的各数据字段的简档，数据字段的简档表征该数据字段中的值；

获得基于(i)所生成的简档以及(ii)所述执行信息制定的子集规则，其中所述子集规则用于将所述数据字段中的特定的一个数据字段标识为目标数据字段；以及

根据所述子集规则而从数据记录的所述第二集合中选择数据记录的子集，其中该数据记录的子集的选择是基于数据记录的所述第二集合的目标数据字段的值进行的。

数据记录的选择

[0001] 优先权声明

[0002] 本申请要求提交于2013年2月1日的美国专利申请序列号61/759799以及提交于2013年3月14日的美国专利申请序列号13/827558的优先权，这两者的全部内容通过引用并入本文。

技术领域

[0003] 本申请涉及数据记录的选择。

背景技术

[0004] 存储的数据集通常包括事先不知道其各种特性的数据。例如，数据集的典型值的数值范围、数据集中不同字段的关系、或是不同字段中的值之间的函数依赖，可能是未知的。数据简档(data profiling)可以涉及检查数据集的源，以便确定这些特性。

发明内容

[0005] 在数据处理应用的开发期间，开发人员可能在生产环境之外工作，并且可能无法访问生产数据。为了确保数据处理应用(在本文中称为“应用”)将在生产中适当地执行实际数据，可以在该应用的执行与测试期间使用真实的数据。应用通常包括执行依赖于一个或多个变量的值的规则。这些变量可以是对应于输入数据的输入变量，可以是依赖于一个或多个输入变量的派生变量，等等。可以从将被用于应用的开发和测试的实际生产数据中选择数据记录的子集。通常选择这些数据记录使得该输入数据对将被执行的应用中的每一条规则是足够的(例如，以使得达到在应用中的完整代码覆盖率)。

[0006] 在一般方面中，一种计算机实现的方法，包括：访问多个数据记录，每一个数据记录具有多个数据字段。该方法还包括为所述多个数据记录中的至少一些分析一个或多个数据字段的值，并且基于该分析生成多个数据记录的简档。该方法还包括基于所述简档制定至少一个子集规则；以及基于所述至少一个子集规则，从所述多个数据记录中选择数据记录的子集。

[0007] 实施例可包括下面的一项或多项。

[0008] 制定至少一个子集规则包括基于第一数据字段的基数(cardinality)，将第一数据字段标识为目标数据字段。在一些情况下，目标数据字段具有多个数据记录中的不同的值的集合，并且其中选择数据记录的子集包括：选择数据记录以使得在所选择的子集中有至少一个数据记录具有目标数据字段的各个不同的值。

[0009] 生成简档包括对多个数据记录中的第一数据字段的值进行分类。制定至少一个子集规则包括基于所述分类，将第一数据字段标识为目标数据字段。在一些情况下，目标字段具有多个数据记录中的不同的值的集合，并且其中选择数据记录的子集包括选择数据记录以使得在所选择的子集中有至少一个数据记录具有目标数据字段的各个不同的值。

[0010] 制定至少一个子集规则包括将第一数据字段标识为第一目标数据字段以及将第

二数据字段标识为第二目标数据字段。在一些情况下，选择数据记录的子集包括基于所述第一目标数据字段的不同的值的第一集合和所述第二目标数据字段的不同的值的第二集合的组合选择数据记录的子集。

[0011] 生成简档包括标识经由第一数据字段的值相关的数据记录之间的关系。至少一个子集规则包括关系的标识。在一些情况下，选择的数据记录的子集包括选择第一数据记录；以及选择经由在子集规则中标识的关系与第一数据记录相关的一个或多个第二数据记录。在一些情况下，数据记录之间的关系包括在数据记录的第一集合中的数据记录和数据记录的第二集合的数据记录之间的关系。

[0012] 生成简档包括对多个数据记录中的至少一些生成伪字段；以及使用累计值填充每个相应数据记录的伪字段。基于第一数据记录以及与第一数据记录相关的至少一个其它数据记录来确定用于第一数据记录中的累计值。第一数据记录经由第一数据字段的值与所述至少一个其它数据记录相关。在一些情况下，该方法包括基于第一数据记录中的第二数据字段的值以及每个其它想的数据记录的第二数据字段的值的总和确定累计值。

[0013] 该方法包括接收子集规则。

[0014] 该方法包括将所选的数据记录的子集提供给数据处理应用。在一些情况下，该方法包括基于数据处理应用的结果制定第二子集规则；以及基于所述第二子集规则选择数据记录的第二子集。

[0015] 在一般方面中，存储在计算机可读介质的软件包括导致计算系统访问多个数据记录的指令，每一个数据字段具有多个数据记录。该软件包括导致计算系统分析所述多个数据记录中的至少一些的一个或多个数据字段的指令；以及基于所述分析生成多个数据记录的简档。该软件还包括用于导致计算系统基于所述简档，制定至少一个子集规则的指令；以及基于所述至少一个子集规则从多个数据记录中选择数据记录的子集。

[0016] 在一般方面中，一种计算系统包括被配置为访问多个数据记录的至少一个处理器，每一个数据记录具有多个数据字段。处理器被配置为分析多个数据记录中的至少一些的一个或多个数据字段的指令，以及基于所述分析生成多个数据记录的简档。处理器也被配置为基于所述简档制定至少一个子集规则；以及基于所述至少一个子集规则从多个数据记录中选择数据记录的子集。

[0017] 在一般方面中，一种计算系统包括用于访问多个数据记录的装置，每一个数据记录具有多个数据字段。计算系统包括用于分析所述多个数据记录中的至少一些的一个或多个数据字段的装置；以及用于基于所述分析生成多个数据记录的简档的装置。该计算系统还包括用于基于所述简档，制定至少一个子集规则的装置；以及用于基于所述至少一个子集规则从多个数据记录中选择数据记录的子集的装置。

[0018] 在一般方面中，一种计算机实现的方法包括：访问多个数据记录，每一个数据记录具有多个数据字段，以及从所述多个数据记录中选择数据记录中的第一子集。该方法包括将数据记录的第一子集提供给实现多个规则的数据处理应用，以及接收指示至少一个规则被数据处理应用执行的次数的报告。该方法包括，基于该报告，从多个数据记录中选择数据记录的第二子集。

[0019] 实施例可包括以下的一项或多项。

[0020] 该方法包括将数据记录的第二子集提供给数据处理应用。

[0021] 该方法包括，基于该报告，标识未由数据处理应用执行的一个或多个未执行的规则。选择数据记录的第二子集包括基于所述标识选择数据记录。

[0022] 该方法包括，基于该报告，标识各自执行少于相应最大次数的一个或多个规则。选择数据记录的第二子集包括基于所述标识选择数据记录。

[0023] 该方法包括，基于该报告，标识各自执行大于相应最小阈值次数的一个或多个规则。选择数据记录的第二子集包括基于所述标识选择数据记录。

[0024] 选择数据记录的第一子集包括：基于第一子集规则选择数据记录的第一子集。在一些情况下，基于第一子集规则选择数据记录的第一子集包括：选择数据记录的第一子集以使得该子集中的至少一个数据记录具有目标数据字段的不同的值的集合的每一个。在一些情况下，基于第一子集规则选择数据记录的第一子集包括：选择第一数据记录，以及选择经由所述第一子集规则中标识的关系与第一数据记录相关的一个或多个第二数据记录。在一些情况下，选择数据记录的第二子集包括：基于与第一子集规则不同的第二子集规则选择数据记录的第二子集。

[0025] 该报告包括指示触发数据处理应用的一个或多个规则执行的变量的值的数据。该方法包括：基于该变量将一个或多个数据字段标识为目标数据字段，其中所述变量取决于所标识的一个或多个数据字段。

[0026] 数据记录的第二子集包括数据记录的第一子集。

[0027] 该方法包括：迭代地选择数据记录的子集，并将数据记录的子集提供给数据处理的应用，直到报告指示数据处理应用已经执行了至少阈值数目的规则。

[0028] 在一般方面中，存储在计算机可读介质的软件包括导致计算系统来访问多个数据记录的指令，每一个数据记录具有多个数据字段，以及从多个数据记录中选择数据记录的第一子集。该软件包括导致该计算系统将数据记录的第一子集提供给实现多个规则的数据处理应用并且接收指示至少一个规则已经被数据处理应用执行的次数的报告的指令。该软件包括导致所述计算系统基于报告从所述多条数据记录中选择数据记录的第二子集的指令。

[0029] 在一般方面中，一种计算系统包括被配置来访问多个数据记录的至少一个处理器，每一个数据记录具有多个数据字段，以及从所述多个数据记录中选择数据记录的第一子集。所述处理器被配置为将数据记录的第一子集提供给实现多个规则的数据处理应用，以及接收指示至少一个规则已经被数据处理应用执行的次数的报告。该处理器被配置为基于所述报告从多条数据记录中选择数据记录的第二子集。

[0030] 在一般方面中，一种计算系统包括：用于访问多个数据记录的装置，每个数据记录具有多个数据字段；用于从所述多个数据记录中选择数据记录的第一子集的装置。该计算系统包括：用于将数据记录的第一子集提供给实现多个规则的数据处理应用的装置，以及用于接收指示至少一个规则由数据处理应用执行的次数的报告的装置。该计算系统包括：用于基于所述报告从多条数据记录中选择数据记录的第二子集的装置。

[0031] 本文所述的技术可以具有以下优点中的一个或多个。例如，完整的生产数据记录集可以是大规模的，使用这样大的记录测试数据处理应用可能是缓慢和不切实际的。通过仅使用所选的可表示与数据处理应用的操作相关的完整的数据记录集的特征的数据记录的子集，可以实现彻底和有效的测试。可以经由对完整数据记录集的自动简档的分析以及

执行数据处理应用的反馈，来实现准确地选择用于有效测试应用的最小数目的数据记录。

[0032] 其它特征和优点从下面的描述和权利要求书中是显而易见的。

附图说明

[0033] 图1是数据处理系统的框图。

[0034] 图2A是客户交易记录示例集的一小部分。

[0035] 图2B是人口统计数据记录的示例集的一小部分。

[0036] 图3是用于基于目标数据字段选择数据记录的子集的示例过程的流程图。

[0037] 图4是用于选择数据记录的示例过程的流程图。

[0038] 图5是用于选择数据记录的另一个示例过程的流程图。

具体实施方式

[0039] 在数据处理应用的开发期间，开发人员可能在生产环境之外工作，并且可能无法访问生产数据。为了确保数据处理应用将在生产中适当地执行实际数据，可以在该应用的开发与测试期间使用真实的数据。应用通常实现其执行依赖于(例如，由其触发)一个或多个变量的值的规则。这些变量可以是对应于输入数据的输入变量，可以是取决于一个或多个输入变量的派生变量，等等。为了有效地测试应用，可以提供足以导致应用中的每个逻辑规则得以执行(例如以使得实现应用中的完整代码覆盖率)的输入数据，以使得每一个逻辑规则被执行至少一个相应最小次数，和/或使得每个逻辑规则被执行不超过相应最大次数。

[0040] 将要提供给应用的数据记录的子集，通常从一个或多个较大的数据记录集合中选择(例如，从实际生产数据集)。所述子集可以基于由用户来指定的子集规则选择，其基于对数据记录简档的分析制定，基于应用的执行的反馈制定，等等。例如，可以选择包括可能会导致正在测试的应用的一些或所有的规则被执行的数据的数据记录用于子集。

[0041] 提供所选的数据记录给使用所选的数据记录作为输入数据执行的应用。所述应用实现一个或多个规则，也即，当规则对应的条件表达式得以满足时，则应用实现的每一个规则可以由应用执行，并且如果相应条件表达式不满足时，则应用不执行该规则。规则由包括至少一个条件表达式和执行表达式的规范指定。当满足条件表达式时(例如，对条件表达式评估的结果为真)，则对执行表达式求值。条件表达式可以取决于(例如，由其触发)一个或多个变量的值，其可以是对应于输入数据的输入变量，可以是取决于一个或多个输入变量的派生变量，等等。在一些示例中，应用执行所有已触发的规则。在一些示例中，应用执行比所有已触发的规则较少的规则，如某些规则或仅一个规则(例如，被触发的第一规则)。至少在提交于2007年4月10日的美国专利号80691295的第61行第五列-第11行第6列中更详细地描述了规则，其内容通过引用全部并入本文。

[0042] 在执行之后，可以提供包含指示应用的执行的数据(例如，应用中执行或没有执行的规则，应用中的每个逻辑规则被执行的次数，或其它执行数据)。基于该报告，可以标识附加输入数据，例如，会导致未执行的规则被执行的输入数据、会导致特定的逻辑规则执行指定次数的输入数据、或是会导致其它期望的执行结果的输入数据。可以执行纠正动作，例如，可以制定附加子集规则，并且可以根据这些附加子集规则选择更新的数据记录的子集。该更新的数据记录的子集可以包括足以导致先前未执行的规则的一些或全部执行的数据

记录、足以导致部分或全部规则执行指定次数的数据记录、或足以导致其它期望的执行结果的数据记录。

[0043] 参照图1，数据处理系统包括托管在服务器102a中的记录选择子系统102。记录选择子系统102从一个或多个数据记录集(例如，生产数据记录)中选择数据记录。所选择的数据记录被提供给数据处理应用106，例如，正在进行测试或开发的应用。在一些示例中，应用106可以对于记录选择子系统102是本地的，例如，托管在相同的服务器102a上。在一些示例中，应用106对于记录选择子系统102是远程的，例如，托管在通过诸如局域或广域数据网络118(例如，因特网)的一个或多个网络访问的远程服务器106A上。

[0044] 数据记录存储在托管在一个或多个服务器104a、104b、104c、104d和相应存储设备108a、108b、108c、108d上的数据源104中。数据源104可以包括任何的各种数据源，诸如数据库109、电子表格文件110、文本文件112、大型机使用的本机格式文件114、或其它类型的数据源。数据源的一个或多个对于记录选择子系统102可以是本地的，例如，托管在同一计算机系统(例如，服务器102a)上。数据源的一个或多个对于记录选择子系统102可以是远程的，例如，存储在通过网络118、多个网络等访问的远程计算机(例如，服务器104a、104b、104c、104d)上。

[0045] 存储在数据源104中的数据记录包括一个或多个数据记录集。例如，数据记录可以包括客户交易记录、客户人口统计数据记录、财务交易记录、电信数据或其它类型的数据记录。每个数据记录具有一个或多个数据字段，并且每个数据记录的每个数据字段具有特定值(或缺乏)，例如数值、字母数字值、空值，等等。例如，在客户交易记录集中，每个记录可具有存储客户标识符、购买价格、事务类型以及其它数据的数据字段。

[0046] 记录选择子系统102中的子集模块120可以提供诸如根据一个或多个子集规则从存储在一个或多个数据源104中的一个或多个数据集中选择数据记录的子集的各种操作。子集规则是可由计算机执行的规则，根据其将从一个或多个数据集中选择数据记录的子集。子集规则还可以由子集模块120基于对由简档模块126生成的一个或多个数据集中的简档的分析制定。子集规则可以由子集模块120基于对由覆盖分析模块128提供的应用执行的结果(例如，基于报告)的分析制定。子集规则可以例如基于用户对数据记录和/或正在测试的应用106的理解，由用户经由用户界面124指定。子集规则还可以从诸如硬盘的存储介质中读出，或经由诸如因特网的网络接收。

[0047] 各种各样的子集规则是可能的，并且可以单独或组合应用。子集规则可以是确定性的(例如，该规则可以指定将选择所有匹配特定标准的记录)或非确定性的(例如，规则可以指定在所有匹配特定标准的记录中，将随机地选择这些记录中的两个)。

[0048] 在一些示例中，子集规则指派一个或多个目标数据字段，并指定目标数据字段的每一个不同的值或值的分类将被包括在所选的数据记录的子集的数据记录中的至少一个之中。子集模块120标识在一个或多个数据记录集中的目标数据字段的每一个不同的值，并且选择数据记录以满足子集规则。例如，对于五十个州的每一个具有不同的值的州数据字段，以及具有两个不同的值的性别数据字段，可以被标识为目标数据字段。选择用于子集的数据记录以使得州的五十个值中的每一个以及性别的两个值的每一个都包括在子集中的至少一个数据记录中。

[0049] 在一些示例中，子集规则指定在相同数据记录集或不同数据记录集之间的数据记

录之间的关系类型。子集模块120基于数据记录与为子集所选的其它数据记录的关系选择数据记录。例如，可以为子集选择共享客户标识符(cust_id)的字段公共值的数据记录。子集规则的其它示例也是可行的，如过滤。在一些示例中，可使用子集规则的组合选择数据记录的子集。

[0050] 在一些示例中，由诸如数据分析师或应用开发者的用户提供子集规则。例如，用户可以标识目标字段、指定数据记录之间的关系、或以其它方式指示子集规则。

[0051] 在一些示例中，子集规则由子集模块120基于对由简档模块126自动生成的数据记录的简档的分析制定。简档模块126可以访问一个或多个数据记录集，并通过分析单个数据集的个别数据记录和/或分析在一个数据记录集内或跨越不同数据记录集的数据字段之间的关系生成数据记录的简档。

[0052] 数据记录集的简档是在对该数据记录集中的数据的诸如逐个字段的总结。简档可以包括表征该数据记录集中的数据的信息，例如在数据记录中的一个或多个数据字段的基数(cardinality)、一个或多个数据字段中的值的分类、在个别数据记录中的数据字段之间的关系、数据记录之间的关系、或其它表征数据记录集中的数据的信息。数据记录集的简档还可以包括表征伪字段的信息，所述伪字段是由简档模块126生成并使用通过操纵在相关数据记录中的一个或多个数据字段的值确定的值填充的数据字段。

[0053] 基于生成的数据记录的简档，子集模块120可以标识数据记录的可能与实现应用106的良好代码覆盖的数据记录的子集的选择相关的特征。例如，基于数据记录的简档，子集模块120可以标识有可能与应用的输入变量和派生变量相关的一个或多个数据字段或数据字段的组合。在一些情况下，子集规则也可以基于从用户或从计算机存储介质接收的输入和/或基于应用106的执行结果(例如，基于从覆盖分析模块128接收到的输入)制定。

[0054] 子集模块120可以执行一种或多种分析的类型的操作以指定子集规则。子集模块120可以基于对个别数据记录内的字段的分析，例如，通过确定哪些数据字段可能与应用106中的变量相关，指定一个或多个子集规则。在一些示例中，子集模块120标识基于简档中指示的目标数据字段的基数(即，跨所有的数据记录集的数据字段的不同的值或值的分类的数量)，标识目标数据字段。例如，性别数据字段(基数为二)可以被标识为目标数据字段而电话号码数据字段(基数取决于数据记录的总数量的数量级)不大可能被标识为目标数据字段。在一些示例中，子集模块120将使用操纵一个或多个数据字段中的数据得到的数据填充的伪字段标识为目标数据字段。例如，收入数据字段中的数据可以被分类成类别(例如，高、中、或低)，使用收入数据字段的分类(inc_range)填充的伪字段可以被标识为目标数据字段。在一些示例中，子集模块120基于如简档中指示的目标数据字段和相同记录中的一个或多个其它数据字段之间的关系标识目标数据字段。例如，简档可以指示数据字段州和邮政编码不是独立的；基于此依赖，子集模块120可以仅考虑这些数据字段中的一个作为可能的目标数据字段。子集模块120还可以基于如简档中指示的对一组数据记录内的和/或跨不同组的数据记录的不同数据记录之间的关系的分析，指定一个或多个子集规则。例如，简档可以指示数据记录可以经由数据字段的公共值(例如，cust_id数据字段的值)关联。数据记录的其它分析也是可能的。

[0055] 一旦子集模块120选择了数据记录的子集，则将指示数据记录的所选子集的数据提供给正在测试的应用106。例如，可以将数据记录的所选子集的标识符以及数据记录的地

址提供给应用106。也可以将包含数据记录的所选子集的文件提供给应用106。

[0056] 数据处理应用106使用数据记录的子集作为输入数据进行执行。在执行后，提供报告给记录选择子系统102中的覆盖分析模块128。也可以将报告提供给用户。该报告包含指示该应用执行(例如，应用中的规则执行或没有执行、应用中的每一个逻辑规则被执行的次数、或其它执行数据)的数据。在一些示例中，报告直接标识那些执行或没有执行的规则。报告还可以包含关于应用106的执行的附加信息，诸如每一个逻辑规则被执行的次数、在执行期间应用的每一个变量的值、或其它信息。

[0057] 对于应用中的每一个没有执行的逻辑规则，覆盖分析模块128标识应用106的与该逻辑规则相关的一个或多个变量。覆盖分析模块128可以基于包括在报告中的数据(例如，指示通过应用106数据流的数据)、基于关于应用的预载的信息等等标识变量。在一些情况下，覆盖分析模块128还标识将导致逻辑规则执行的每一个变量的值或值的范围。输入数据字段以及对应于变量的值或值的范围被标识，并用于指定随后的由子集模块120选择更新的数据记录的子集时的附加子集规则。

[0058] 例如，如果所标识的变量是应用的直接对应于数据记录的一个数据字段中的一个的输入变量，则覆盖分析模块128标识相应的数据字段以及该字段的值或值的范围。例如，如果当变量x大于10时应用106中的逻辑规则就执行，并且变量x对应于包含关于客户交易的数量的数据的输入数据字段txn_amt，则覆盖分析模块确定输入数据应包括至少一个txn_amt>10的数据记录。该确定(例如，txn_amt>10)被提供给子集模块120，其指定附加子集规则使得随后提供给应用的106的数据记录的子集将包括足以导致x>10的逻辑规则执行的数据。

[0059] 例如，如果所识别的变量不是输入变量(也即，所标识的变量不直接对应数据记录的数据字段中的一个)，则覆盖分析模块128中的数据沿袭子模块130通过应用106的逻辑跟踪变量的派生，以标识(多个)输入变量，从其派生已标识的变量。然后覆盖分析模块128标识相应的(多个)数据字段以及该数据字段的值或值的范围。例如，如果当变量y的值是2时应用106的逻辑规则执行，则数据沿袭子模块130可确定y是经由应用中的逻辑步骤从对应于输入数据字段性别，inc_range和州的三个输入变量的逻辑组合派生的。通过跟随变量y的逻辑派生，可以确定导致y＝2的数据字段性别，inc_range和州的值。例如，当性别＝F，inc_range＝高，并且州＝ME、NH、VT、MA、RI、或CT时，可满足逻辑规则y＝2。将该确定提供给子集模块120，其指定使得随后提供给应用106的数据记录的子集将包括足以导致y＝2的逻辑规则执行的数据的附加子集规则。作为另一个例子，当两个变量的值具有特定关系时，诸如对应于数据字段名和姓的变量的值是相等的时，逻辑规则可以执行。

[0060] 在一些示例中，覆盖分析的结果也提供给用户。用户可以提供附加子集规则给子集模块120，或者可以修改先前提供的子集规则。用户还可以提供另外的输入到简档模块126以修改先前提供给简档模块的输入。

[0061] 在一些示例中，即使完整的(多个)数据记录集也不包括足以满足应用106中的逻辑规则的数据。例如，应用106可以包括仅当收入数据字段的值大于五百万美元时执行的逻辑规则。如果集合中不存在收入>$5,000,000的数据记录，那么数据记录没有将导致逻辑规则执行的子集。为识别这样的完整数据集中的缺陷，在一些示例中，可使用所有的数据记录作为输入执行应用一次或更多次。结果报告标识不论用于输入的选择数据记录的子集为何

也无法覆盖的规则。

[0062] 子集模块120和简档模块126的操作将参考图2A和图2B中示出的示例数据记录202和示例数据记录252进行描述。图2A是客户交易记录集200的一小部分的示例。每一个客户交易记录202有几个数据字段204，包括，例如客户标识符(cust_id)204a、交易类型(txn_type)204b、交易金额(txn_amt)204c、交易日期(日期)204d以及商店标识符(store_id)204e。也可以包括其它数据字段。图2B是人口统计数据记录集250的一小部分的示例。每一个人口统计数据记录252有几个数据字段254，包括，例如客户的标识符(cust_id)254a、客户地址(地址，州，邮政编码)254b、254c、254d、客户收入(收入)254e、以及客户的性别(性别)254f。也可以包括其它数据字段。简档模块126和子集模块120的操作并不限于这些示例数据集，并类似地应用于其它类型的数据集。

[0063] 子集模块120可以根据一个或多个类型的子集规则选择数据记录的子集。一些示例子集规则如下：

[0064] 过滤。在一些实例中，子集模块120根据过滤器选择数据记录的子集。例如，过滤器可以指定选择所有具有给定数据字段的特定值的数据记录。例如，过滤器可以指定选择集合250中所有具有州(数据字段254c)＝“MA”的人口统计数据记录用于子集。过滤器可以由用户、简档模块126、和/或覆盖分析模块128来指定。

[0065] 在一些示例中，子集模块120根据基于规则的过滤器选择数据记录的子集，其中基于给定的数据字段的值消除数据记录。例如，过滤器可以指定从该子集中消除store_id(数据字段204e)＝“在线”的数据记录。基于规则的过滤器可以由用户、简档模块126、和/或覆盖分析模块128指定。

[0066] 目标数据字段。在一些示例中，子集模块120基于一个或多个目标数据字段选择数据记录的子集。目标数据字段是诸如可能与应用的变量相关的数据字段。例如，如果操作客户交易记录的特定应用通过商店位置跟踪事务类型(即，购买或返回)，则应用的开发者可以标识数据字段txn_type(数据字段204c)和store_id(数据字段204e)作为目标数据字段。在一些情况下，子集模块120可以基于如在数据记录的简档中指示的诸如数据字段的基数这样的数据字段的特性，来标识目标数据字段。在一些情况下，覆盖分析模块128可以基于应用的变量和数据字段之间的关系标识目标数据字段。即使简档模块126只有很少或没有关于数据字段的内容的以及该内容可以如何与应用相关的其它信息，低基数的数据字段(例如，基数小于基数阈值的数据字段)也可被标识为目标数据字段。基数阈值可以由用户指定，或者可以由简档模块自动确定。例如，基于人口统计数据记录集250的简档，假如阈值基数设定为至少50，则可以标识数据字段州为目标数据字段。

[0067] 图3是基于目标数据字段选择数据记录的子集的示例过程的流程图。标识一个或多个目标数据字段(300)，例如，基于包括在数据记录的简档的信息、来自用户的信息、来自覆盖分析模块128的信息，等等。标识了记录集中的每一个目标数据字段的不同的值的集合(302)。为子集选择数据记录(304)使得每一个目标数据字段的每一个不同的值被包括在子集中的至少一个数据记录中。在一个示例中，数据字段州和数据字段性别被标识为人口统计数据记录集250的目标数据字段。分析人口统计数据记录集250以标识州的50个不同的值和性别的两个不同的值。选择数据记录使得州的五十个值的每一个以及性别的两个值的每一个都包括在子集中的至少一个数据记录中。在一些示例中，子集规则可以指定每个目标

数据字段的每个不同的值被包括在该子集中的次数(如,一次、十次、五十次等)。

[0068] 基于目标数据字段的子集并不一定意味着每个数据字段的每个值的每一种组合都在子集中有表示。例如,包括了州的五十个值中每一个以及性别的两个值的每一个的数据记录的子集可以仅包含50个数据记录。在一些实例中,目标数据字段是诸如伪字段(例如,如下所述的通过简档模块构造)的构造字段,并且取决于相同的记录内的或跨不同的记录的一个或多个数据字段。

[0069] 数据分类。在一些示例中,基于数据记录的一个或多个目标数据字段中的数据的分类选择数据记录的子集。例如,子集规则可以标识目标数据字段并且指定值的不同范围("容器(bin)"),基于其可以对目标数据字段的值分类。基于目标数据字段的容器而不是目标数据字段的精确值选择用于子集的数据记录。在一个示例中,人口统计数据记录集250的收入数据字段被标识为目标数据字段。指定三个容器:"低"(收入<$50,000)、"中等"(收入在$50,000和$150,000美元之间)和"高"(收入>$150,000)。由子集模块120考虑包括在子集中的每个数据记录的收入数据字段的值被分类为低、中、或高;选择数据记录使得收入的三个容器的每一个包括在子集中的至少一个数据记录中。在一些示例中,对数据字段的值进行分类(例如,由简档模块)并且每一个数据记录的伪字段使用相应的分类的值填充(例如,数据字段inc_range 256)。在这些示例中,伪字段被视为目标数据字段,并且选择数据记录使得伪字段的每一个不同的值包括在子集中的至少一个数据记录中。将要分类的数据字段、容器的数目和/或每个容器的值的范围可以由用户指定或由简档模块126及/或覆盖分析模块128自动标识。

[0070] 组合。在一些示例中,根据可以指定两个或更多的其它子集规则组合的组合规则选择数据记录的子集。例如,组合规则可以标识两个目标数据字段并且指定这两个目标数据字段的每一个的所有值的所有可能的组合被包括在子集中的至少一个数据记录中。一个示例组合规则可以标识字段inc_range和性别作为目标数据字段并且指定这两个数据字段的所有可能组合被包括在子集中。满足这个组合规则的子集将包括六个数据记录(即,低+女性、低+男性、中+女性、中+男性、高+女性、高+男性)。相反,如果没有组合规则,则少至仅3个记录(例如,低+女性、中+男性、高+女性)就可以满足inc_range和性别作为目标数据字段的规范。在一些示例中,子集规则可以指定两个或更多目标数据字段的组合(combinatoric combination)以及一个或多个组合之外的其它的目标数据字段。例如,子集规则可以指定inc_range和性别作为组合将采取的目标数据字段,并且还可以指定该组合以外的州作为目标数据字段。更复杂的组合也是可能的。目标数据字段和组合的具体类型可以由用户指定或由简档模块126及/或覆盖分析模块128自动标识。

[0071] 数据记录之间的关系。在一些示例中,根据数据记录集内的或跨不同的数据记录集的数据记录之间的关系,选择数据记录的子集。子集规则可以指定联接关键字(joint key),这样如果选择一个数据记录用于数据记录的子集,则经由联接关键字与该数据记录相关的其它数据记录用于该子集也被选择。例如,子集规则可以标识数据字段cust_id为联接关键字,其关联客户交易记录集200内的数据记录以及集合200和人口统计数据记录集250之间的数据记录。对于来自选择用于子集的一个集合中的每一个数据记录来说(例如,根据另一个子集规则),与所选的数据记录中的cust_id值相同的其它数据记录也被选择用于子集。通过根据关系选择数据记录,子集将包含,例如,特定的客户的所有交易的数据记

录，以及该客户的人口统计数据记录。关系可以由用户指定或由简档模块126及/或覆盖分析模块128自动标识。

[0072] 在一些示例中，数据记录之间的关系可基于数据记录中的一个或多个特性。例如，可以标识感兴趣的数据记录(例如，对应于欺诈的信用卡交易数据记录)。然后相应的子集规则可以指定子集将包括具有类似于所标识的感兴趣的数据记录的特性的50个其它的数据记录，例如，以帮助标识数据记录中的其它的欺诈实例。

[0073] 也可以指定其它子集规则。例如，可以指定数据记录的计数(例如，子集将包括至少100个txn_type＝“购买”的记录)。可以指定统计参数(例如，子集将包括所有txn_type＝“购买”的数据记录，以及15％的txn_type＝“退货”的数据记录)。可以指定数值参数(例如，数据记录集中的每百万的数据记录中，子集将包括至少指定数目的数据记录)。这些子集规则可以由用户指定和/或基于对简档(由简档模块126生成)的分析和/或对执行的分析的结果由子集模块120指定。

[0074] 在一些示例中，可以施加多个子集规则到一个数据记录集。在某些情况下，施加这些多个子集规则可能会导致多次选中某些数据记录用于子集。可以施加重复数据删除的规则到所选的数据记录，以删除子集中出现一次以上的任何数据记录。

[0075] 在一些示例中，基于对由简档模块126生成的简档的分析制定子集规则。简档模块126可以分析没有来自外部源的输入、或具有来自用户的和/或覆盖分析模块128的输入的数据记录。简档分析的一些例子如下：

[0076] 基数。在一些示例中，简档模块126标识数据字段的基数(即，跨一个集合的所有数据记录的不同的数据记录的值的数量)。例如，当简档客户交易记录集200时，简档模块可以标识txn_type为低基数(集合200的全部数据记录中仅有两个不同的值)的数据字段。当简档人口统计数据记录集250时，可以标识数据字段州为基数为50的数据字段——假定基数阈值被设定为至少50。子集模块120可以使用一些或所有的数据字段的基数，以指定子集规则。

[0077] 分类。在一些示例中，简档模块126对数据字段中的数据进行分类。例如，简档模块可以标识值的不同范围(“容器”)，基于其可以对高基数的数据字段的值进行分类。根据分类，数据字段具有较低的基数并且如上所述可被标识为目标数据字段。在一些情况下，简档模块根据它对记录的分析对每一个记录的数据记录的值进行分类，但不存储该分类。在一些情况下，简档模块为每一个存储对应于数据字段的值的容器的记录生成伪字段。作为例子，人口统计数据记录集250中的数据字段收入是高基数。简档模块将每个记录的收入值分类成三个容器(高、中、或低)之一，并生成伪字段inc_range356来存储分类的数据。该伪字段356的基数为三，因此可能由子集模块120识别为目标数据字段，而高基数数据字段收入可能不会被标识为目标数据字段。在一些示例中，简档模块识别可以被自动分类的高基数数据字段。在一些示例中，用户标识进行分类的数据字段，也可以指定容器的数量和落入每个容器内的值的范围。在一些示例中，用户指定将进行分类的数据字段的特征而不标识特定的数据字段(例如，用户可以指定值为数的以及基数在10到100之间的所有数据字段可以被四分位分类)。

[0078] 数据字段之间的关系。在一些示例中，简档模块126确定单个数据记录内的数据字段之间的关系。举例来说，如果数据记录中的第一数据字段取决于每个数据记录中的第二

数据字段，那么仅需要考虑第一数据字段和第二数据字段中的一个作为目标数据字段。例如，数据字段州和数据字段邮政编码相关(即，邮政编码的值取决于州的值)。根据简档中对这种关系的指示，子集模块120可以只考虑两个相关数据字段中的一个作为可能的目标数据字段。可以标识数据字段之间更复杂的关系，并由子集模块120在标识目标数据字段时使用。简档模块可由用户的输入指导，例如，由用户指定很可能是相关的数据字段。

[0079] 数据记录之间的关系。在一些示例中，简档模块126确定在一个数据记录集内的或跨不同的数据集的不同的数据记录之间的关系。例如，简档模块可以识别集合内的一些数据记录经由数据字段的共同的值链接。例如，客户交易记录集200可以包括对应相同的客户的交易的多个数据记录。这些数据记录通过cust_id的共同的值(即，联接关键字)链接。简档模块还可以识别第一集合内的第一数据记录经由数据字段的共同的值与第二集合内的第二数据记录相关。例如，在客户交易记录集200中的数据记录可经由数据字段cust_id链接到人口统计数据记录集250(也即，特定客户的交易记录可以链接到该客户的人口统计数据记录)。简档模块可由用户的输入指导，例如，由用户指定可能链接数据记录的数据字段。也可以指导简档模块，经由与(多个)数据记录集相关的关系数据库的模式的分析以标识联接关键字或其它关系。在一些示例中，简档模块126确定数据记录之间的关系，并将关系展示给用户，用户之后可以使用有关关系的信息为子集模块120指定子集规则。

[0080] 基于简档中对数据记录之间的这样关系的指示，子集模块120可以指定作为子集规则的一部分的联接关键字。在这样的子集规则下，如果选择了子集的一个数据记录用于子集，则经由联接关键字与该数据记录关联的其它数据记录(例如，如果选择了具有给定的cust_id的一个数据记录用于子集，则具有相同的cust_id的其它数据记录也被选择)也被选择。

[0081] 伪字段。在一些示例中，简档模块126使用通过操纵相关数据记录的一个或多个数据字段的值确定的值生成新的伪字段并标识该伪字段为目标数据字段。伪字段的值可以是对于经由联接关键字相关的数据记录的一个或多个数据字段的值的组合。例如，伪字段的值可以是累计的值，例如，诸如经由第二数据字段的公共值相关的数据记录的第一数据字段所有值的和、计数或其它累计值的累计。伪字段的值也可以是累计值的分类。例如，为了处理执行取决于给定客户的总交易量的动作的应用中的逻辑，在客户交易记录集200中生成伪字段total_amt 306。具有给定的cust_id值的数据记录中的伪字段total_amt的值，是通过对具有那个cust_id值的所有数据记录的txn_amt字段求和并将总和分至三个容器中的一个(高、中、或低)来确定的。所述伪字段然后可以由子集模块标识为目标数据字段。

[0082] 参照图4，在一个示例过程中，访问多个数据记录(400)。每个数据记录具有多个数据字段。分析多个数据记录中的至少一些中的一个或多个数据字段的值(402)。基于该分析生成多个数据记录的简档(404)。多个数据记录的简档包括表征数据记录集中的数据的信息。基于该简档(406)制定至少一个子集规则。子集规则是对规则的规范，基于其从多个数据记录中选择数据记录的子集。基于至少一个子集规则选择数据记录的子集(408)。例如，可以基于目标数据字段的值和/或基于经由数据字段的值相关的数据记录之间的关系选择数据记录的子集。

[0083] 参照图5，在另一个示例性过程中，访问多个数据记录(500)。每个数据记录具有多个数据字段。从多个数据记录中选择数据记录的第一子集(502)。将数据记录的第一子集提

供给如正在测试的应用的数据处理应用(504)。所述应用实现多种规则。在数据处理应用中的规则是其执行取决于(例如,被触发)一个或多个变量的值的应用的可执行部分。接收指示至少一个规则被数据处理应用执行的次数的报告(506)。基于该报告,从所述多条数据记录中选择数据记录的第二子集(508)。将数据记录的第二子集提供给数据处理应用(510)。例如,选择第二子集使得先前未执行的规则可被执行,或者使得可以执行某些规则。

[0084] 在一些示例中,可以基于由简档模块126进行的对简档的分析生成新的数据记录。例如,该简档的分析揭示了数据记录内和数据记录之间的数据字段之间的关系,现有的数据记录集的数据字段的可能值的范围。可以构造新的数据记录,其中数据字段的至少一些使用从现有的数据记录相关的信息计算的或确定的值填充。可以使用测试数据生成,例如,当在源数据集中没有导致应用中的特定的逻辑规则执行的数据记录,例如,逻辑规则要求收入>$10,000,000;或逻辑规则需要多个数据字段的特定值的复杂组合,其中并非所有的所需要的值在数据记录集都有表示。测试数据生成也可以用于生成简档与原始数据集的简档相匹配的新的数据集。例如,可以通过随机化原始数据集的数据来生成新的数据集,以保留原始数据记录的隐私。

[0085] 在一些示例中,以上描述的方法在可以在诸如UNIX操作系统的合适的操作系统的控制下的一个或多个通用计算机中托管的执行环境中实现。例如,执行环境可以包括多节点并行计算环境,其包括使用多个中央处理单元(CPU)的计算机系统的配置,无论是本地的(例如,诸如SMP计算机的多处理器系统)、或本地分布(例如,耦合为群集或MPP的多个处理器)、或远程的、或远程分布(例如,多个通过局域网里(LAN)和/或广域网(WAN)耦合的处理器),或它们的任何组合。

[0086] 在一些情况下,以上描述的方法由作为数据流图开发应用的系统实现,数据流图包括通过顶点(表示组件或数据集)之间的有向链接(表示工作元素的流)连接的顶点。例如,这样的环境在题为“Managing parameters for graph-based application”的美国公开号2007/0011668中描述的更加详细,以引用方式将其并入本文。用于执行这样的基于图的计算的系统在美国专利5,566,072“EXECUTING COMPUTATIONS”中描述,通过引用将其并入本文。根据本系统制定的数据流图提供用于将信息传入或移出由图组件表示的单个进程、用于在进程之间传递消息、以及用于规定进程的运行顺序的方法。该系统包括选择进程间通信的方法的算法(例如,根据图的链接的通信路径可以使用TCP/IP或UNIX域套接字,或使用在过程之间传递数据的共享内存)。

[0087] 以上描述的方法可以使用用于在计算机上执行的软件实现。例如,该软件形成一个或多个计算机程序的过程,所述程序在一个或多个已编程的或可编程的计算机系统(其可以是诸如分布式、客户端/服务器、或网格的各种架构)上运行,计算机系统的每一个包括至少一个处理器、至少一个数据存储系统(包括易失性和非易失性存储器和/或存储元件)、至少一个输入设备或端口,以及至少一个输出设备或端口。该软件可以形成较大程序的一个或多个模块,例如,提供与数据流图的设计和配置相关的其它服务。图的节点和元素可以实现为存储在计算机可读介质的数据结构、或符合在数据仓库中存储的数据模型的其它组织数据。

[0088] 该软件可以被提供在通用或专用目的的可编程计算机可读取的诸如CD-ROM的存储介质上,或通过网络的通信介质向运行它的计算机传递(在传播信号中编码)。所有的功

能可以在专用计算机上执行，或者使用诸如协处理器的专用硬件执行。该软件可以在其中由软件指定的计算的不同部分由不同的计算机执行的分布式方式来实现。每一个这样的计算机程序优选存储在或下载到通用或专用可编程计算机可读取的有形的、非暂时性的存储介质或设备(例如，固态存储器或介质、或者磁或光学介质)，用于当存储介质或设备由计算机系统读取以执行其中描述的过程时，配置和操作计算机。本发明的系统也可以被认为实现为配置有计算机程序的计算机可读存储介质，其中如此配置的存储介质使得计算机系统按照特定和预定义的方式操作以执行这里描述的功能。

[0089] 已经描述了本发明的多个实施例。然而，将理解的是，可以进行各种修改而不脱离本发明的精神和范围。例如，上述一些步骤可以是顺序无关的，并且由此可以按照与描述的顺序不同的顺序来执行。

[0090] 应当理解的是，前述描述旨在说明而不是限制由所附权利要求的范围限定的本发明的范围。例如，上面描述的多个功能步骤可以按照不同顺序执行，而基本上不影响总体处理。其它实施例在以下权利要求的范围中。

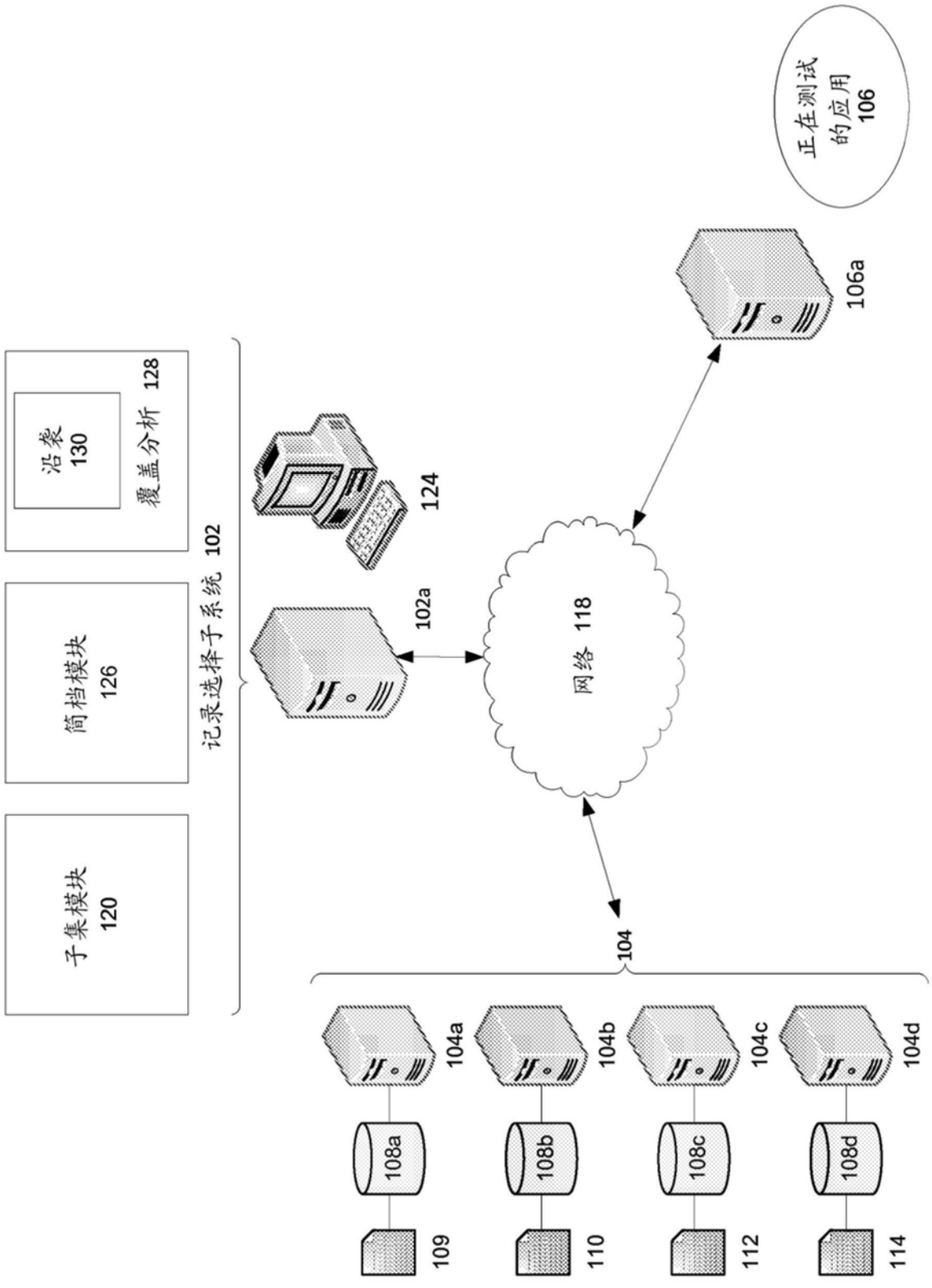

正在测试的应用 106
106a
网络 118
102a
124
记录选择子系统 102
覆盖分析 128
沿袭 130
简档模块 126
子集模块 120
104
104a
104b
104c
104d
108a
108b
108c
108d
109
110
112
114

图1

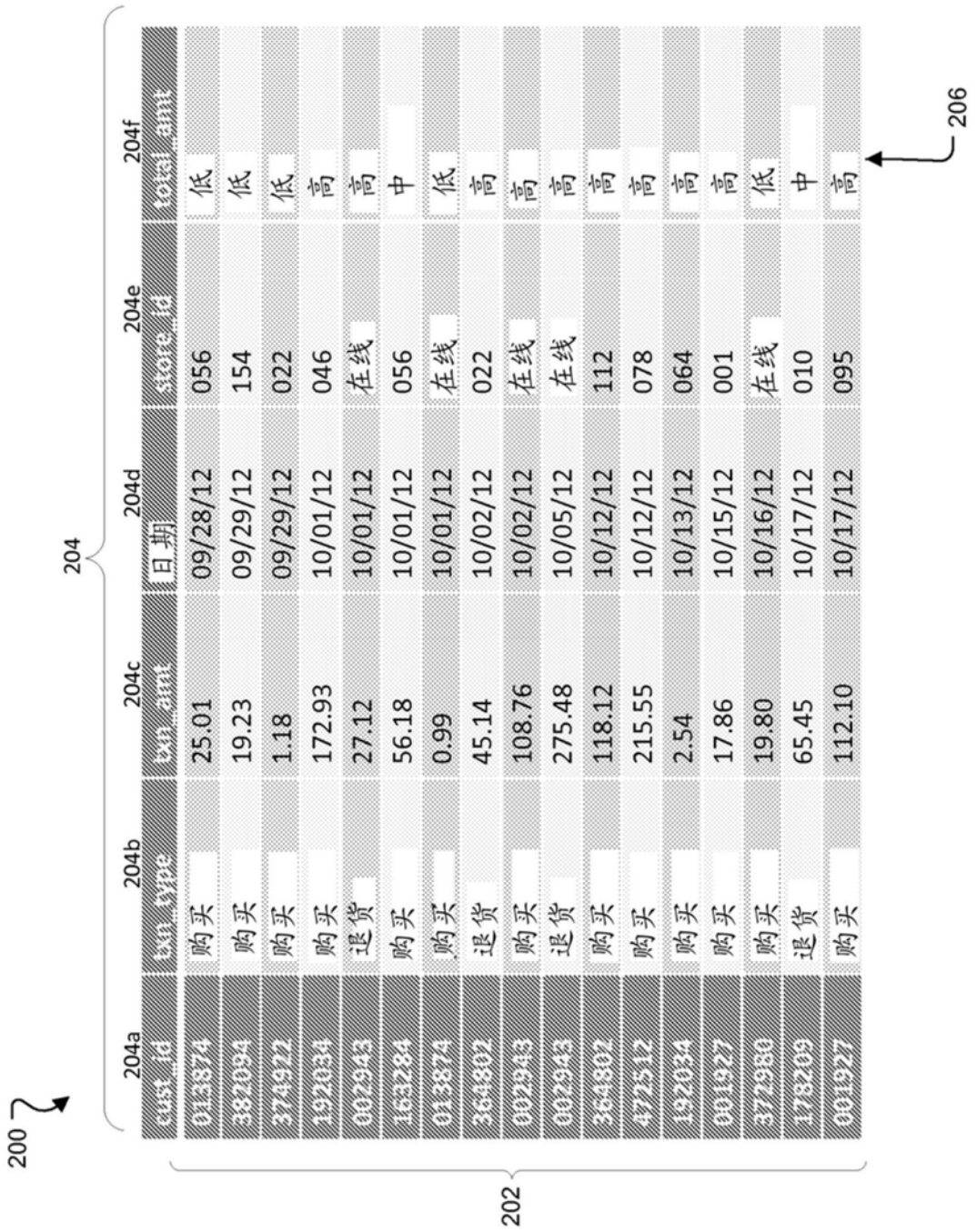

200
204
202
206
204a cust_id
204b txn_type
204c txn_amt
204d 日期
204e store_id
204f total_amt
013874 购买 25.01 09/28/12 056 低
382094 购买 19.23 09/29/12 154 低
374922 购买 1.18 09/29/12 022 低
192034 购买 172.93 10/01/12 046 高
002943 退货 27.12 10/01/12 在线 高
163284 购买 56.18 10/01/12 056 中
013874 购买 0.99 10/01/12 在线 低
364802 退货 45.14 10/02/12 022 高
002943 购买 108.76 10/02/12 在线 高
002943 退货 275.48 10/05/12 在线 高
364802 购买 118.12 10/12/12 112 高
472512 购买 215.55 10/12/12 078 高
192034 购买 2.54 10/13/12 064 高
001827 购买 17.86 10/15/12 001 高
372980 购买 19.80 10/16/12 在线 低
178209 退货 65.45 10/17/12 010 中
001927 购买 112.10 10/17/12 095 高

图2A

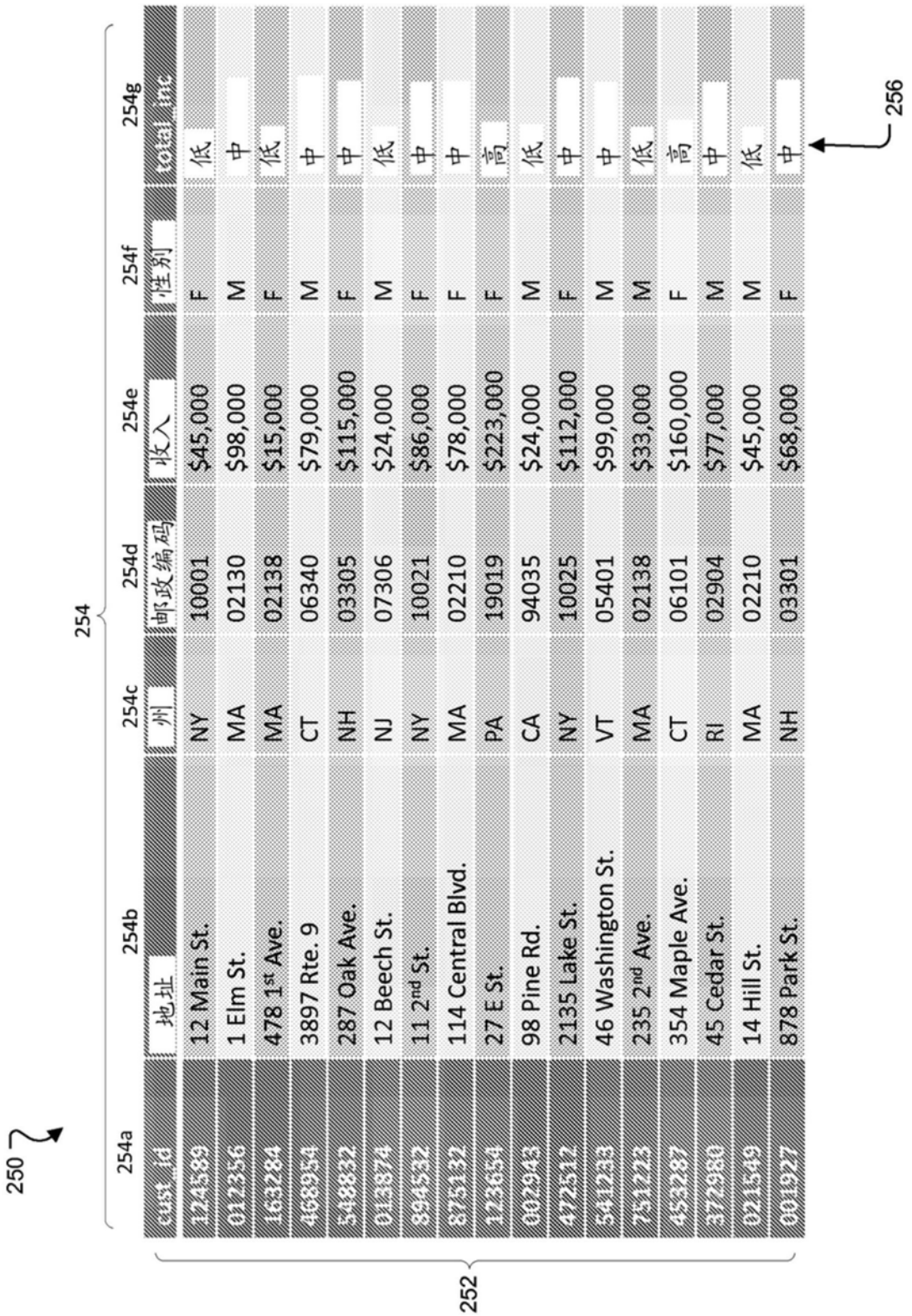

250
254
252
256
254a
254b
254c
254d
254e
254f
254g
cust_id
地址
州
邮政编码
收入
性别
total_inc
123589 12 Main St. NY 10001 $45,000 F 低
017356 1 Elm St. MA 02130 $98,000 M 中
163284 478 1st Ave. MA 02138 $15,000 F 低
468954 3897 Rte. 9 CT 06340 $79,000 M 中
548832 287 Oak Ave. NH 03305 $115,000 F 中
013874 12 Beech St. NJ 07306 $24,000 M 低
894532 11 2nd St. NY 10021 $86,000 F 中
875132 114 Central Blvd. MA 02210 $78,000 F 中
123654 27 E St. PA 19019 $223,000 F 高
002943 98 Pine Rd. CA 94035 $24,000 M 低
472512 2135 Lake St. NY 10025 $112,000 F 中
541233 46 Washington St. VT 05401 $99,000 M 中
751223 235 2nd Ave. MA 02138 $33,000 M 低
453287 354 Maple Ave. CT 06101 $160,000 F 高
372980 45 Cedar St. RI 02904 $77,000 M 中
021549 14 Hill St. MA 02210 $45,000 M 低
001927 878 Park St. NH 03301 $68,000 F 中

图2B

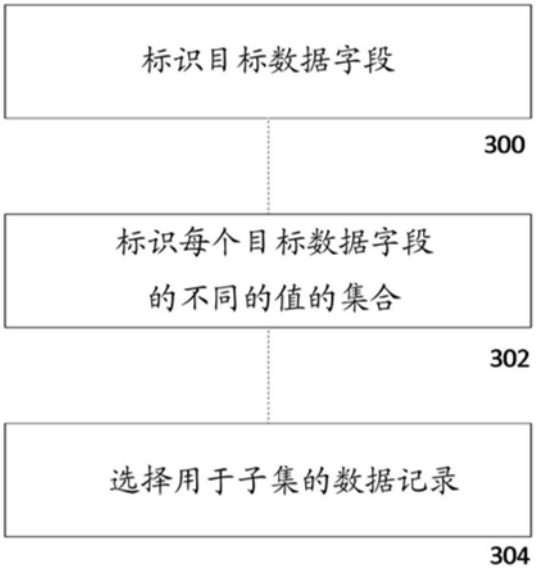

标识目标数据字段
300
标识每个目标数据字段
的不同的值的集合
302
选择用于子集的数据记录
304

图3

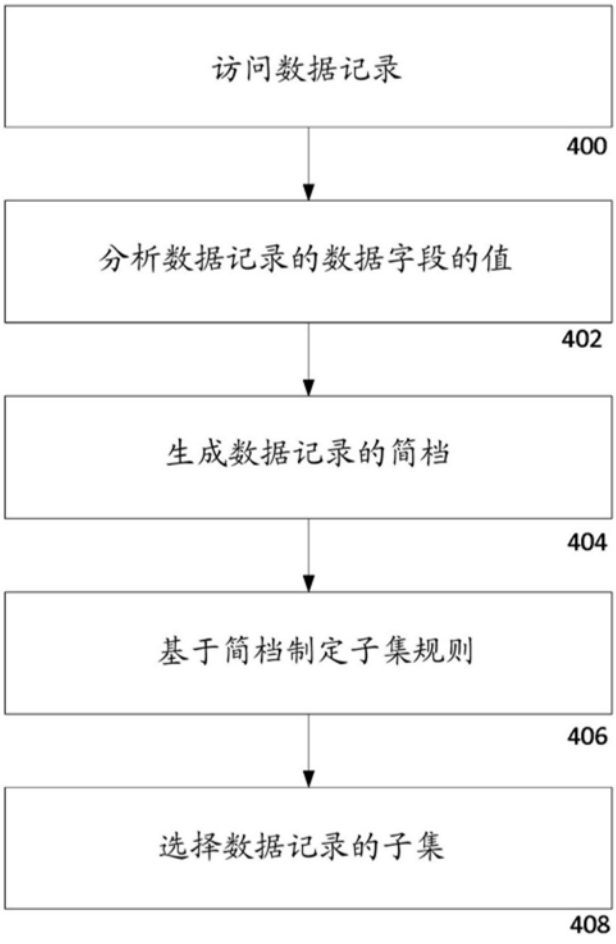

访问数据记录
400
分析数据记录的数据字段的值
402
生成数据记录的简档
404
基于简档制定子集规则
406
选择数据记录的子集
408

图4

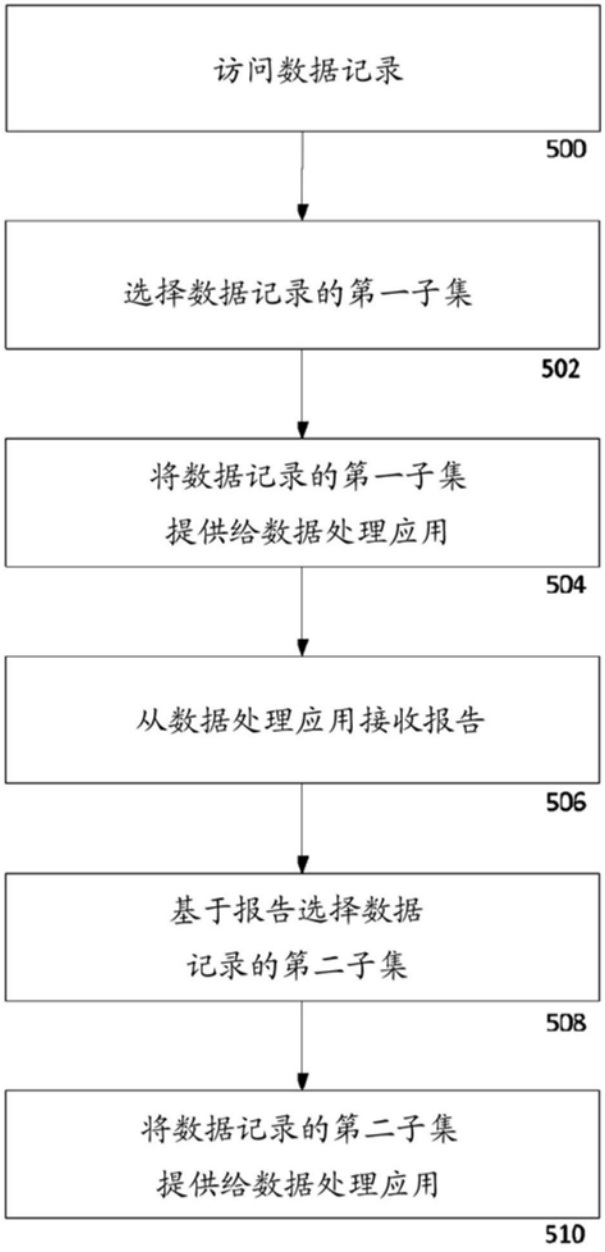

访问数据记录
500
选择数据记录的第一子集
502
将数据记录的第一子集
提供给数据处理应用
504
从数据处理应用接收报告
506
基于报告选择数据
记录的第二子集
508
将数据记录的第二子集
提供给数据处理应用
510

图5