



(12) 发明专利申请

(10) 申请公布号 CN 102460076 A

(43) 申请公布日 2012. 05. 16

(21) 申请号 201080035409. 7

(51) Int. Cl.

(22) 申请日 2010. 06. 09

G01D 3/00 (2006. 01)

(30) 优先权数据

61/185, 797 2009. 06. 10 US

(85) PCT申请进入国家阶段日

2012. 02. 10

(86) PCT申请的申请数据

PCT/US2010/038018 2010. 06. 09

(87) PCT申请的公布数据

W02010/144608 EN 2010. 12. 16

(71) 申请人 起元技术有限责任公司

地址 美国马萨诸塞州

(72) 发明人 C. R. 范曼

(74) 专利代理机构 北京市柳沈律师事务所

11105

代理人 邵亚丽

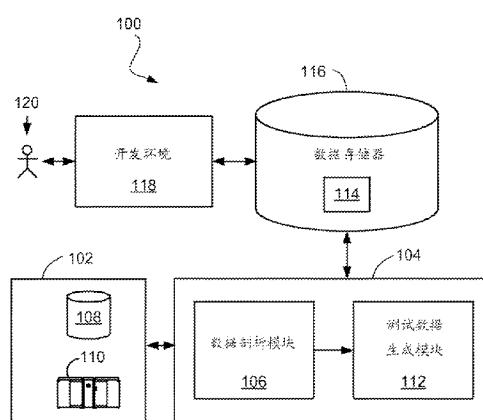
权利要求书 2 页 说明书 7 页 附图 3 页

(54) 发明名称

生成测试数据

(57) 摘要

生成测试数据包括：从数据源（102）读取在多个记录的至少一个字段中出现的值；存储包括对值的特征进行描述的统计信息的简档信息；基于统计信息生成（206）字段的概率分布的模型（300）；使用所生成的模型生成（206）多个测试数据值，使得给定值在测试数据值中出现的频率对应于由模型（300）分配给该给定值的概率；以及将包括测试数据值的测试数据（114）的集合存储（212）在数据存储系统（116）中。



1. 一种用于生成测试数据的方法,该方法包括 :

从数据源读取在多个记录的至少一个字段中出现的值 ;

存储包括对所述值的特征进行描述的统计信息的简档信息 ;

基于所述统计信息生成所述字段的概率分布的模型 ;

使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率 ;以及

将包括测试数据值的测试数据的集合存储在数据存储系统中。

2. 如权利要求 1 所述的方法,其中所生成的模型包括所述概率分布的估计的至少第一部分和所述概率分布的估计的至少第二部分,所述概率分布的估计的至少第一部分对应于值的连续范围上的概率密度,所述概率分布的估计的至少第二部分对应于离散值的一个或多个离散概率值。

3. 如权利要求 2 所述的方法,其中所述统计信息包括指示值的多个连续范围中的每个连续范围内落入多少值的信息。

4. 如权利要求 3 所述的方法,其中所述第一部分至少部分地从指示值的连续范围之一内落入多少值的信息导出。

5. 如权利要求 2 所述的方法,其中所述统计信息包括指示多个特定值中的每个值在所述记录中出现的频率的信息。

6. 如权利要求 5 所述的方法,其中所述第二部分中的离散概率值之一至少部分地从指示特定值之一在所述记录中出现的频率的信息中导出。

7. 如权利要求 1 所述的方法,还包括使用所生成的模型和指示从数据源读取的出现在字段中的值的特征的附加信息来生成多个测试数据值。

8. 如权利要求 7 所述的方法,其中所述附加信息指示出现在字段中的无效值的数量。

9. 如权利要求 7 所述的方法,其中所述附加信息指示出现在字段中的相异值的比率。

10. 如权利要求 1 所述的方法,还包括提供开发环境,用于使用所述数据存储系统中存储的测试数据的集合来开发用于处理来自所述数据源的记录的至少一个程序。

11. 如权利要求 10 所述的方法,其中所述开发环境不能访问来自所述数据源的记录。

12. 如权利要求 11 所述的方法,其中所述开发环境不能访问所述数据源。

13. 一种用于生成测试数据的系统,该系统包括 :

数据源,其提供在一个或多个字段中具有值的记录 ;

数据存储系统 ;以及

一个或多个处理器,其耦合到所述数据存储系统,用于提供执行环境以执行以下处理 :

从所述数据源读取在多个记录的至少一个字段中出现的值,

存储包括对所述值的特征进行描述的统计信息的简档信息,

基于所述统计信息生成所述字段的概率分布的模型,

使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率,以及

将包括测试数据值的测试数据的集合存储在所述数据存储系统中。

14. 一种用于生成测试数据的系统,该系统包括 :

数据源,其提供在一个或多个字段中具有值的记录;

数据存储系统;以及

用于处理所述记录以生成测试数据的装置,所述处理包括:

从所述数据源读取在多个记录的至少一个字段中出现的值,

存储包括对所述值的特征进行描述的统计信息的简档信息,

基于所述统计信息生成所述字段的概率分布的模型,

使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率,以及

将包括测试数据值的测试数据的集合存储在所述数据存储系统中。

15. 一种计算机可读介质,用于存储用于生成测试数据的计算机程序,所述计算机程序包括用于使计算机执行以下操作的指令,所述操作包括:

从所述数据源读取在多个记录的至少一个字段中出现的值;

存储包括对所述值的特征进行描述的统计信息的简档信息;

基于所述统计信息生成所述字段的概率分布的模型;

使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率;以及

将包括测试数据值的测试数据的集合存储在所述数据存储系统中。

生成测试数据

[0001] 相关申请的交叉引用

[0002] 本申请要求 2009 年 6 月 10 提交的美国申请第 61/185797 号的优先权，其通过引用合并于此。

技术领域

[0003] 本说明书涉及生成测试数据。

背景技术

[0004] 一些组织拥有它们希望保密的数据（例如，可能包括客户信息的生产数据）。当要通过程序处理保密数据时，出于安全的原因，可能需要开发者开发不对实际生产数据进行访问的程序。例如，使得生产数据保密的一种途径是使看到生产数据的人的数量最小化。然而，为了确保他们的应用在有生产数据的情况下正确运行，程序员可能需要真实的测试数据用于开发和测试，所述真实的测试数据展示生产数据的某些特征，但不暴露任何机密信息。

发明内容

[0005] 在一个方面，总体上，一种用于生成测试数据的方法包括：从数据源读取在多个记录 (record) 的至少一个字段 (field) 中出现的值；存储包括对所述值的特征进行描述的统计信息 (statistics) 的简档 (profile) 信息；基于所述统计信息生成所述字段的概率分布的模型；使用所生成的模型生成多个测试数据值，使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率；以及将包括测试数据值的测试数据的集合存储在数据存储系统中。

[0006] 多个方面可以包括下列特征中的一个或多个。

[0007] 所生成的模型包括所述概率分布的估计的至少第一部分和所述概率分布的估计的至少第二部分，所述概率分布的估计的至少第一部分对应于值的连续范围上的概率密度，所述概率分布的估计的至少第二部分对应于离散值的一个或多个离散概率值。

[0008] 所述统计信息包括指示值的多个连续范围中的每个连续范围内落入多少值的信息。

[0009] 所述第一部分至少部分地从指示值的连续范围之一内落入多少值的信息导出。

[0010] 所述统计信息包括指示多个特定值中的每个值在所述记录中出现的频率的信息。

[0011] 所述第二部分中的离散概率值之一至少部分地从指示特定值之一在所述记录中出现的频率的信息中导出。

[0012] 该方法还包括使用所生成的模型和指示从数据源读取的出现在字段中的值的特征的附加信息来生成多个测试数据值。

[0013] 所述附加信息指示出现在字段中的无效值的数量。

[0014] 所述附加信息指示出现在字段中的相异值 (distinct value) 的比率

(fraction)。

[0015] 该方法还包括提供开发环境,用于使用数据存储系统中存储的测试数据的集合来开发用于处理来自数据源的记录的至少一个程序。

[0016] 所述开发环境不能访问来自数据源的记录。

[0017] 所述方法还包括所述开发环境不能访问所述数据源。

[0018] 在另一方面,总体上,一种用于生成测试数据的系统包括:数据源,其提供在一个或多个字段中具有值的记录;数据存储系统;以及一个或多个处理器,其耦合到所述数据存储系统,用于提供执行环境以执行以下处理:从数据源读取在多个记录的至少一个字段中出现的值,存储包括对所述值的特征进行描述的统计信息的简档信息,基于所述统计信息生成所述字段的概率分布的模型,使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率,以及将包括测试数据值的测试数据的集合存储在数据存储系统中。

[0019] 在另一方面,总体上,一种用于生成测试数据的系统包括:数据源,其提供在一个或多个字段中具有值的记录;数据存储系统;以及用于处理所述记录以生成测试数据的装置,所述处理包括:从数据源读取在多个记录的至少一个字段中出现的值,存储包括对所述值的特征进行描述的统计信息的简档信息,基于所述统计信息生成所述字段的概率分布的模型,使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于由所述模型分配给该给定值的概率,以及将包括测试数据值的测试数据的集合存储在数据存储系统中。

[0020] 在另一方面,总体上,一种计算机可读介质存储用于生成测试数据的计算机程序。所述计算机程序包括用于使计算机执行以下操作的指令,所述操作包括:从数据源读取在多个记录的至少一个字段中出现的值;存储包括对所述值的特征进行描述的统计信息的简档信息;基于所述统计信息生成所述字段的概率分布的模型;使用所生成的模型生成多个测试数据值,使得给定值在测试数据值中出现的频率对应于所述模型分配给该给定值的概率;以及将包括测试数据值的测试数据的集合存储在数据存储系统中。

[0021] 多个方面可以包括以下优点中的一个或多个优点。

[0022] 测试数据可以随机地生成,但是要以这样的方式进行,即,生产数据的原始集合的各种特征可以被复制以提供更真实的测试。来自原始数据的诸如统计特性的特征以及一些实际值可以被包括在测试数据中,同时确保机密信息不被包括在测试数据中。初始剖析过程使得统计特性和实际值被提取出来并且在简档信息内进行概述,所述简档信息随后用于生成测试数据。只要机密信息没有泄露到简档信息中,测试数据中就不会有机密信息。在简档信息中表示的实际值是各个字段的最常见的值,其不可能对应于机密信息(机密信息一般对于给定客户是唯一的,因此在原始数据内很少见)。

[0023] 本发明的其它特征和优点将从后面的说明书以及权利要求中变得清楚。

附图说明

[0024] 图 1 是使用所生成的测试数据来开发程序的示例性系统。

[0025] 图 2 是示例性测试数据生成过程的流程图。

[0026] 图 3A-3B 是统计信息的示例。

[0027] 图 3C 是概率分布模型的示例。

具体实施方式

[0028] 参考图 1, 使用测试数据来开发程序的系统 100 包括数据源 102, 数据源 102 可以包括数据的一个或多个源, 比如存储设备或到在线数据流的连接, 它们中的每一个可能存储各种存储格式 (比如, 数据库表、电子数据表文件、纯文本文件 (flat text file)、或者大型机所使用的原生格式) 中的任何一种存储格式的数据。用于生成测试数据的执行环境 104 包括数据剖析 (profiling) 模块 106 和测试数据生成模块 112。执行环境 104 可以被托管 (host) 在处于适合的操作系统 (例如 UNIX 操作系统) 控制下的一个或多个通用计算机上。例如, 执行环境 104 可以包括多节点并行计算环境, 该多节点并行计算环境包括使用多个中央处理单元 (CPU) 的计算机系统的配置, 所述计算机系统的配置要么是本地的 (例如, 诸如 SMP 计算机的多处理器系统)、要么是本地分布式的 (例如, 髮合成簇或 MPP 的多个处理器)、要么是远程的、要么是远程分布式的 (例如, 经由 LAN 网络或 WAN 网络耦合的多个处理器) 或者是它们的任意组合。

[0029] 数据剖析模块 106 从数据源 102 读取数据并且存储对数据源 102 中出现的数据值的各种特征进行描述的简档信息。提供数据源 102 的存储设备可以位于执行环境 104 本地, 例如, 存储在连接到运行执行环境 104 的计算机的存储介质 (例如, 硬驱 108) 上, 或者可以远离执行环境 104, 例如, 托管在通过局域或广域数据网络与运行执行环境 104 的计算机通信的远程系统 (例如, 大型机 110) 上。

[0030] 测试数据生成模块 112 使用数据剖析模型 106 生成的简档信息来生成存储在执行环境 104 可访问的数据存储系统 116 中的测试数据 114 的集合 (collection)。数据存储系统 116 也可以被开发环境 118 访问, 在开发环境 118 中, 开发者 120 能够使用测试数据 114 来开发和测试程序。然而, 可以通过使数据源 102 中的原始生产数据不能被开发者 120 访问, 来保持数据源 102 中的原始生产数据的安全。在一些实现方案中, 开发环境 118 是用于开发作为数据流图的应用的系统, 所述数据流图包括顶点 (组件或数据集), 所述顶点由顶点之间的有向链接 (directed link) (代表工作元素的流) 连接。例如, 在美国公开第 2007/0011668 号 (名称为“Managing Parameters for Graph-Based Applications”, 通过引用合并且此) 中更加详细地描述了这样的环境。

[0031] 数据剖析模块 106 能够剖析 (profile) 来自包括不同形式数据库系统的各种类型系统的数据。所述数据可以被组织为记录, 所述记录具有针对各个字段 (也称为“属性”或“列”) 的值, 包括可能为空的值。可以组织简档信息以便为不同的字段提供单独的简档, 称为“字段简档”, 其描述在那些字段中出现的值。当从数据源第一次读取数据时, 数据剖析模块 106 一般以关于那个数据源中的记录的一些初始格式信息开始。(注意, 在一些情况下, 即使最初可能不知道数据源的记录结构, 也可以在分析该数据源之后确定该数据源的记录结构)。关于记录的初始信息可以包括表示相异值 (distinct value) 的比特的数量、记录内的字段的顺序、以及所述比特所表示的值的类型 (例如, 字符串 (string)、有符号 / 无符号整数)。当数据剖析模块 106 从数据源读取记录时, 其计算反映给定字段内的值的统计信息和其它描述信息。然后, 数据剖析模块 106 以字段简档的形式存储那些统计信息和描述信息, 以供测试数据生成模块 112 访问。针对给定字段的字段简档中的统计信息可以包括

例如该字段中的值的直方图、该字段中出现的最大值、最小值和平均值、以及该字段中出现的最不常见和最常见的值的样本。简档信息也可以包括与数据源 102 中的记录的多个字段相关联的信息，比如，记录的总数、以及有效记录的总数或无效记录的总数。例如，在美国公开第 2005/0114369 号（名称为“Data Profiling”，通过引用合并于此）中描述了剖析数据源的字段的过程。

[0032] 图 2 示出示例性测试数据生成过程 200 的流程图。测试数据生成模块 112 针对将为其生成测试数据的第一字段检索 (202) 所存储的字段简档。在一些实现方案中，可以在所述字段简档被加载之后计算并且存储能够从所述字段简档中的信息中导出的附加信息。可选地，模块 112 接收 (204) 关于所生成的测试数据值应当具有的特征的用户输入（例如，指示测试数据值中应当出现的相异值的数量的信息，如以下更详细地描述的）。模块 112 基于针对该字段所检索的字段简档中的统计信息生成 (206) 该字段的概率分布的模型。该模型包括针对允许出现在该字段中的、值的范围的概率估计，如以下参照图 3 的示例所描述的。

[0033] 模块 112 对于每个将要生成的测试数据值都调用 (208) 一次测试数据生成器函数。该测试数据生成器函数将概率分布的模型以及可选的自变量 (argument) “索引 (index)” 和“限制 (limit)” 作为输入自变量，如下面更详细地描述的。测试数据生成器函数根据作为输入提供的模型和下面描述的其它特征，提供测试数据值作为输出。在一些情况下，可以选择将为给定字段生成的测试数据值的数量，使其与所剖析的来自原始数据源的数据集中的记录的总数相匹配，所述记录的总数包含在简档信息中。在一些情况下，用户可能想要生成特定数量的值，这可以由用户直接提供（例如，在步骤 204 中）。从所述函数输出的值可以被插入到将要被提供为测试数据 114 的集合的记录的适当字段中。模块 112 确定 (210) 是否将处理附加字段，如果是则执行测试数据生成的另一次迭代。模块 112 将所生成的测试数据 114 的集合存储 (212) 在数据存储系统 116 中。

[0034] 在一些实现方案中，将要为给定字段生成的测试数据值的数量可以通过对于在输入记录流中所接收的每个记录都调用一次测试数据生成器函数来隐式地确定。例如，当开发环境 118 支持作为数据流图的应用的开发和执行时，执行环境 104 也可以将数据剖析模块 106 和测试数据生成模块 112 实现为数据流图本身。在一些实现方案中，数据剖析模块将从数据源 112 读取作为 N 个单独记录的流的生产数据，并且将这些记录作为 N 个单独记录的流提供给模块 112。模块 112 将能够利用针对那些字段生成的测试数据值来替换原始记录的每个字段中的值。模块 112 能够对 M 个字段中的每个字段执行单独的处理 N 个记录的流的迭代，从而在 M 次迭代中的每次迭代中生成 N 个测试数据值。不同迭代的测试数据值接连地顺序生成，或者同时并行地生成。在一些实现方案中，被替换的原始值的某些特性可以保留在所生成的测试数据值中（比如，例如州和邮政编码等字段之间的函数依赖 (functional dependence)）。此外，在一些实现方案中，可以选择原始记录的字段的子集，用于利用根据各自模型生成的测试数据值对其进行替换，并且剩余字段可以保持它们的原始值或者根据不同的技术对剩余字段进行处理，比如利用常数值或者根据均匀概率分布生成的伪随机值对剩余字段进行替换。

[0035] 测试数据生成器函数返回特定数据类型（例如，字符串、十进制数 (decimal)、日期、或者包括日期和时间的日期 - 时间类型）的测试数据值，其中基于给定字段的字段简档

来确定所述数据类型。测试数据值作为记录内的给定字段的值被收集在测试数据 114 的集合中。使用该模型至少部分地基于随机选择（例如，利用伪随机数生成技术）来生成测试数据值，使得给定值在测试数据 114 中出现的频率对应于由作为到测试数据生成器函数的输入提供的模型分配给该值的概率。测试数据生成器函数生成测试数据，使得对测试数据 114 的集合进行处理的数据剖析模块 106 将会产生与用来生成测试数据 114 的集合的简档信息类似的简档信息。

[0036] 图 3A 和图 3B 示出字段简档中的示例性统计信息，并且图 3C 示出字段的概率分布的示例性模型 300。本示例中的测试数据值是从具有定义的顺序的可能值的范围中选择出来的（例如，按照数字值排序的数、或按照定义的字母顺序排序的字符串）。本示例中的统计信息包括十分位数的图表（plot of deciles）（图 3A），其指示值的几分之几落在范围的最小值和最大值之间的 10 个等份（10deciles）中的每一份中（在本示例中为 0 和 10 之间的实数）。图 3A 中示出的图表的水平轴被标记成示出与每个等份相对应的值的范围（0-1、1-2、等等）。在其他示例中，水平轴将覆盖与被建模的任何字段相对应的值的范围。潜在值的任何域（domain）（包括字符串）都可以被映射到这样的图表的数值范围（例如，通过将字符串中的字符解释为适当基数（base）中的数）。统计信息也包括频率值的列表（图 3B），在本示例中，该列表包括前 5 个最频繁的值以及每个值出现的次数。这个字段的概率分布的模型 300 既考虑十分位数的图表提供的连续信息，也考虑频率值列表提供的离散信息。

[0037] 例如，模型 300 包括与十分位数的图表给出的比率成比例的、每个连续等份范围内的值的连续概率密度、以及频率值列表中值的离散概率，该离散概率对应于与列表中的出现次数成比例的、那些离散值处的离散概率（例如， δ （delta）函数）。相对于 δ 函数的高度的概率分布的连续部分的高度取决于等份所表示的原始数据中值的数量。通过减去频率值列表中落在每个等份中的任何值所表示的出现的总次数，相对于等份的高度降低概率分布的连续部分的高度（这样，那些频率值不会被进行两次计数）。如果频率值落在等份之间的边界上（例如，图 3C 中的值 4.0），那么该值的出现次数被从包括该值的等份中减去。例如，在一个实现方案中，该值的出现次数被从其左侧的等份中减去，在该实现方案中，值基于等于或小于测试被分配给等份。概率的绝对值被确定为使得分布的积分为 1（即，所有概率的和等于 1）。其它类型的统计信息可以用来导出该模型，其它类型的统计信息比如是在所剖析的数据的字段中出现的值的直方图。指示多少值落在某个范围中的统计信息促成该模型的连续部分，而指示特定值出现频率的统计信息促成该模型的离散部分（例如， δ 函数）。

[0038] 通过将该模型与所剖析的数据的概率分布相匹配，通过概率分布确定所生成的测试数据的特征，比如均值、标准差、公共值（例如，该模型中的 δ 函数处的值）、以及最大和最小容许值，被自动与所剖析的数据的那些特征相匹配。

[0039] 除了通过概率分布确定的特征之外，测试数据生成器函数能够考虑附加特征以获得与字段的简档信息更接近的匹配。下面是测试数据生成器函数在基于所提供的模型执行随机选择时能够考虑的一些示例性特征。

[0040] • 无效值比率：测试数据生成器函数能够以与所剖析的数据中的比率近似相同的比率在字段中生成无效值。在一些情况中，字段简档可以包括可从中进行选择的公共无效值列表。在一些情况中，如果字段简档包括规定什么值构成字段的有效值的有效性规范，则

测试数据生成器函数能够选择违反该规范的无效值。或者,如果不存在样本无效值或有效性规范,则该函数可以选择对于字段的数据类型来说无效的值(例如,十进制的字母、或不正确格式的日期)。

[0041] • 相异值比率 : 测试数据生成器函数能够以与所剖析的数据中的比率近似相同的比率(例如,相异值的数量 / 值的总数)在字段中生成相异值。为此,该函数接收可选的“索引”和“限制”自变量作为输入。索引自变量是非负整数,其对于每次函数调用都不同(例如,对于每个记录递增的整数),而限制自变量等于或大于已经或将要作为索引自变量提供的任何值。限制自变量可以在开始生成字段的测试数据值时确定一次,并且可以基于用户输入来提供(例如,步骤 204)。例如,限制自变量的值可以被设置为用户将要请求的记录的数量,用户可以事先知道该数量,但是运行测试数据生成器函数的处理器(多个处理器)事先不会知道该数量。

[0042] • 最大和最小长度 : 测试数据生成器函数能够生成与所剖析的数据具有相同最大和最小长度的值。例如,对于字符串来说,长度对应于字符的数量,而对于数(number)来说,长度可以被定义为数字(digit)的数量,包括小数点后的数字。

[0043] • 观察字符(observed character) : 测试数据生成器函数能够生成仅仅由所剖析的数据中出现的字符组成的值,比如,字符串。

[0044] 可以用于测试数据生成器函数以对于给定的限制自变量 L、当索引自变量 I 从 0 变化到 L-1 时确定输出测试数据值的技术的一个示例涉及生成 I 从 0 到 L-1 的所有值的伪随机排列(permutation)以及对这些值进行换算使得它们落在 0 到 1 的范围内(例如,通过将它们除以 L)。这样, I 的每个输入值可以被映射到对模型概率分布的相应部分进行限定的数的不同范围。例如,对于 L = 10, 使用所述排列可以将从 0 到 9 的 10 个索引值伪随机地映射到 0 和 1 之间的不同范围:0 到 0.1、0.1 到 0.2、等等。概率分布被划分成相等概率的相应数量的段(slice)(在本例中为 10 段),并且所选择的段内某处的输出值被选为输出测试数据值。该函数通过相应地减少段的数量并且不止一次地输出不同段内的值来考虑要输出的唯一值的数量。例如,如果存在全部值的一半那么多的唯一值,那么每个值被输出两次。各种技术中的任何一种都可以用于生成所述排列(例如,在 Berlin :Springer Verlag 的“Advances in Cryptology-EUROCRYPT’ 92(Lecture Notes in Computer Science),” 的 1992 年第 658 卷、第 239–255 页中的 Ueli Maurer 的“A simplified and generalized treatment of Luby-Rackoff pseudorandom permutation generators” 中描述了基于 Luby-Rackoff 类型的算法的技术,其通过引用合于此)。

[0045] 上述的测试数据生成方法可以使用在计算机上执行的软件来实现。例如,该软件形成在一个或多个已编程或可编程的计算机系统(其可以是各种架构,比如分布式、客户机/服务器、或网格)上执行的一个或多个计算机程序中的过程,所述计算机系统每个都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,举例来说,所述较大程序提供与计算图的设计和配置有关的其它服务。图的节点和元素可以被实现为计算机可读介质中存储的数据结构或者符合数据存储库中存储的数据模型的其它经组织的数据。

[0046] 该软件可以被提供在可以由通用或专用可编程计算机读取的、诸如 CD-ROM 的存

储介质上,或者该软件可以通过连接到执行该软件的计算机的网络的通信介质来递送(被编码成传播信号)。所有功能都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来执行。该软件可以以分布式方式来实现,其中该软件所规定的计算的不同部分由不同计算机来执行。每个这样的计算机程序优选地存储在或下载到通用或专用可编程计算机可读取的存储介质或设备(例如,固态存储器或介质、或者磁或光介质),用于当计算机系统读取该存储介质或设备时配置和操作该计算机,以执行这里描述的过程。也可以考虑将本发明的系统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以特定和预先定义的方式操作,以执行这里描述功能。

[0047] 已经描述了本发明的一些实施例。尽管如此,还是要理解可以在不脱离本发明的精神和范围的情况下进行各种修改。例如,上述的一些步骤可以是与顺序无关的,因此可以与所描述的顺序不同的顺序来执行这些步骤。

[0048] 要理解,前面的描述旨在说明而不是限制本发明的范围,本发明的范围由所附权利要求的范围来限定。例如,上述的一些功能步骤可以以不同的顺序来执行,而基本上不影响总体处理。其它实施例落在所附权利要求的范围内。

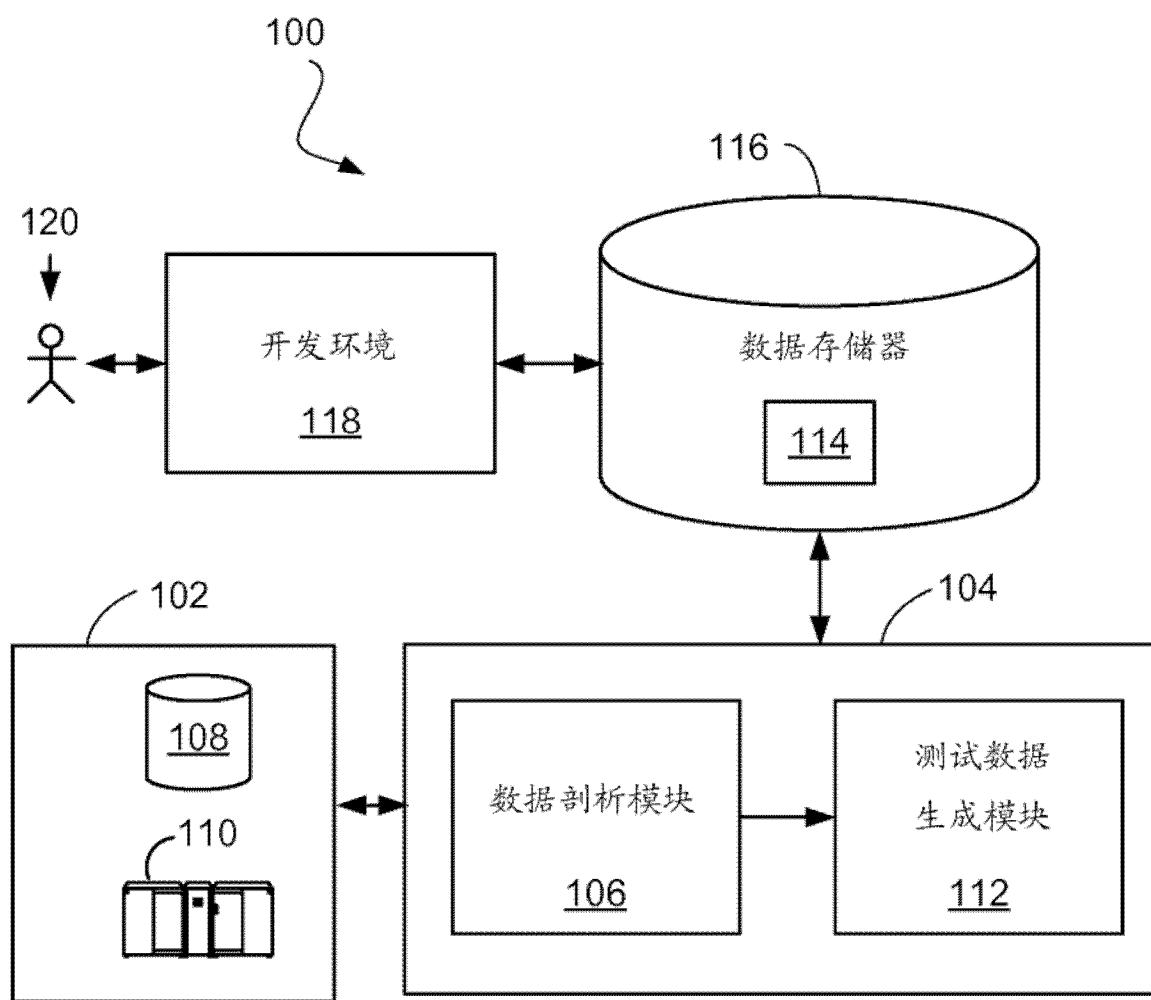


图 1

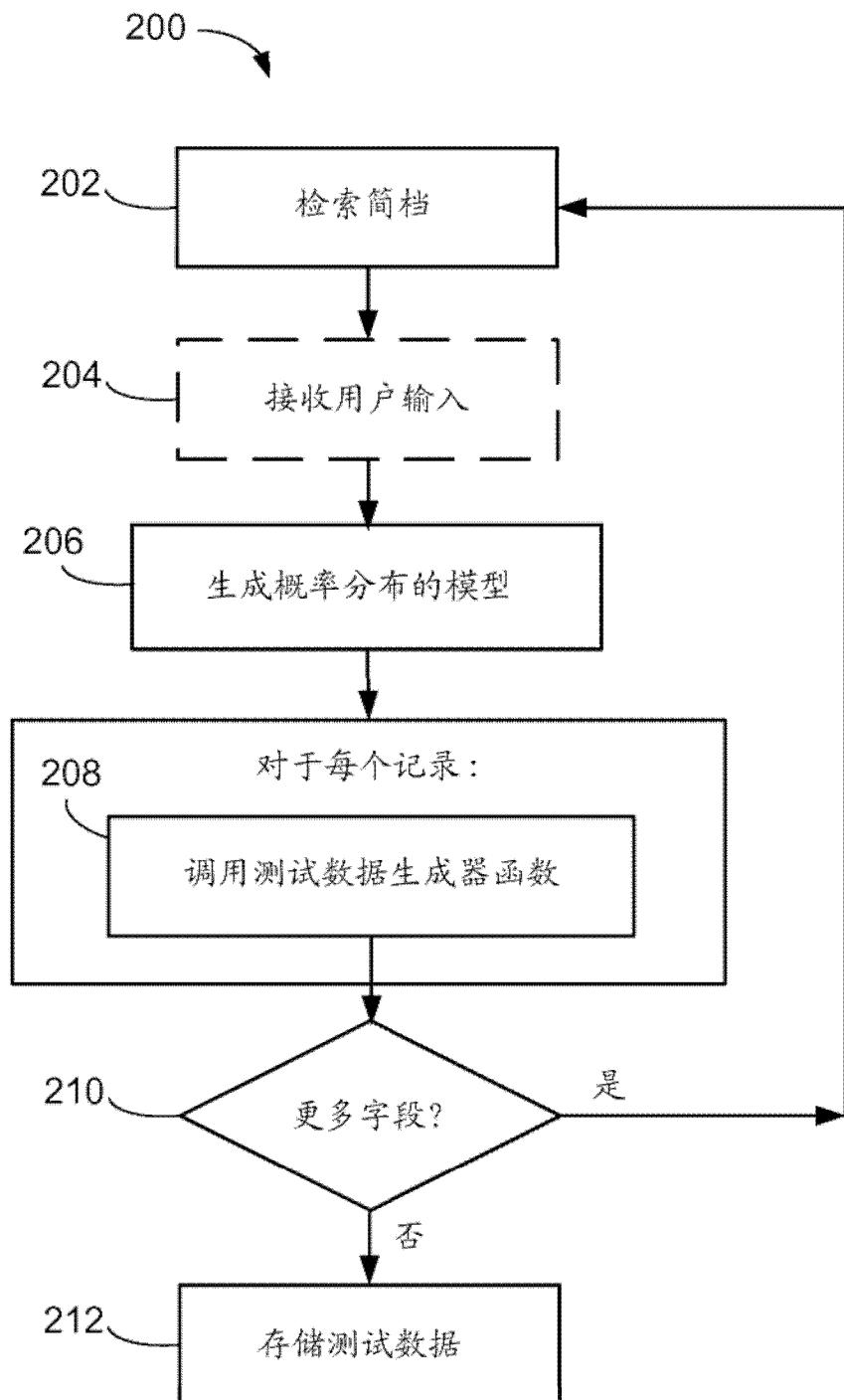


图 2

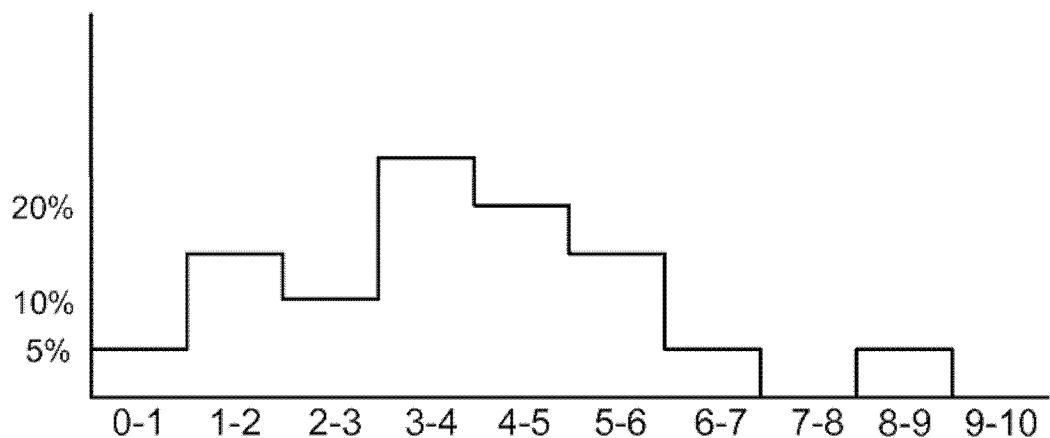


图 3A

值	出现次数
4.0	140
5.25	70
6.5	60
1.5	30
8.75	20

图 3B

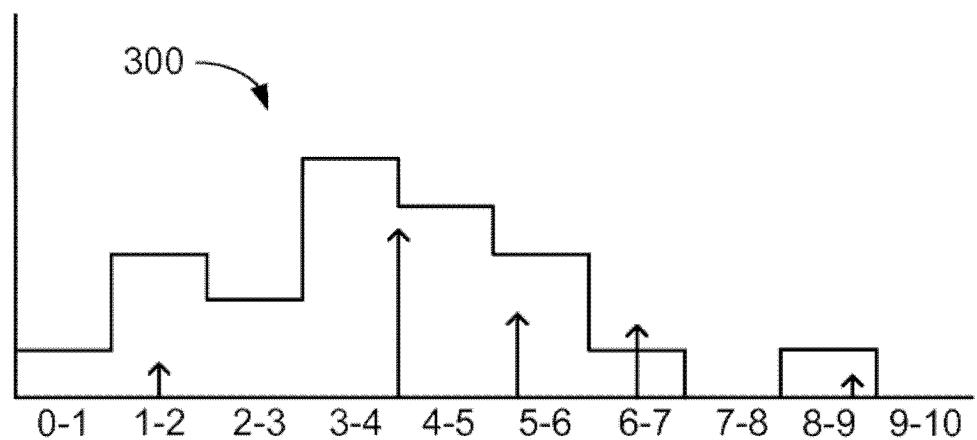


图 3C