

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)



[12] 发明专利说明书

专利号 ZL 200510088517.1

[45] 授权公告日 2009 年 7 月 29 日

[11] 授权公告号 CN 100520776C

[22] 申请日 2005.7.29

[21] 申请号 200510088517.1

[30] 优先权

[32] 2004. 8.30 [33] US [31] 10/929,514

[73] 专利权人 微软公司

地址 美国华盛顿州

[72] 发明人 R·莫特瓦尼 S·乔德里

V·甘提

[56] 参考文献

US2004/0003005A1 2004.1.1

Learning to Match and Cluster Large High -
Dimensional DataSets for Data Integration. Wil-
liam W. Cohen et al. ACM. 2002

审查员 王 亮

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 顾嘉运

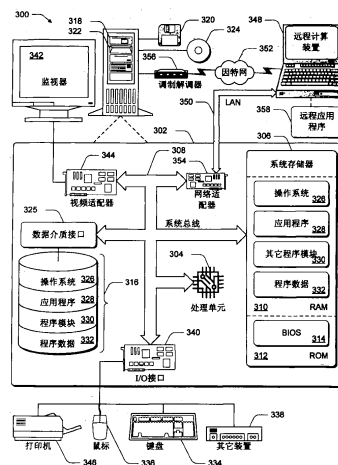
权利要求书 2 页 说明书 13 页 附图 3 页

[54] 发明名称

模糊重复的鲁棒检测器

[57] 摘要

在此所述的至少一个实现检测模糊重复并消除这样的重复。模糊重复是数据库中表示同一实际实体或现象的多个看起来不同的多元组(即记录)。



1. 一种模糊重复消除系统，其特征在于，包括：

—数据集获取装置，用于从数据库获取包括多个多元组的数据集；

—计算装置，用于对所述数据集的多个多元组的每一个，计算所述数据集的一个或多个最相近邻居多元组，所述计算装置进一步包括：

用于在所述数据集中计算已定义数量的最相近邻居多元组的装置；

用于将每个多元组的已定义数量的最相近邻居多元组集定义为一“邻域”的装置；

用于对每个邻域计算“邻域增长率”的装置；

—划分装置，用于基于“紧密集”和“稀疏邻域”的每个部分来定义多个多元组的多个不相交部分，其中每个部分中看起来不同的多元组包括彼此的模糊重复，以致一部分中的每个模糊重复的多元组表示一共同实际实体或现象；

—重复-消除装置，用于消除在多个模糊重复多元组的多个部分内的重复，从而使每一部分中都只剩下不重复的多元组。

2. 如权利要求1所述的系统，其特征在于，每个部分中的所述多元组还包括彼此的准确重复。

3. 如权利要求1所述的系统，其特征在于，所述计算装置还用于通过所述数据集中所述多元组的索引查找最相近的邻居。

4. 如权利要求1所述的系统，其特征在于，所述划分装置进一步包括用于基于“紧密集”、“稀疏邻域”和“小集势”的每个部分来定义多个多元组的多个部分的装置。

5. 一种用于消除数据集中模糊重复多元组的方法，其特征在于，所述方法包括：

对于所述数据集的多个多元组的每一个，计算一个或多个最相近邻居多元组，所述计算进一步包括：

在所述数据集中计算已定义数量的最相近邻居多元组；

将每个多元组的已定义数量的最相近邻居多元组集定义为一“邻域”；

对每个邻域计算“邻域增长率”；

基于“紧密集”和“稀疏邻域”的每个部分来划分多个多元组的多个不相

交部分，其中每个部分中看起来不同的多元组包括彼此的模糊重复，从而一部分中的每个模糊重复的多元组表示一共同实际实体或现象；

消除在多个模糊重复多元组的多个部分内的重复，从而使每一部分中都只剩下不重复的多元组。

6. 如权利要求 5 所述的方法，其特征在于，所述计算一个或多个最相近邻居多元组还包括通过所述数据集中所述多元组的索引查找最相近的邻居。

7. 如权利要求 5 所述的方法，其特征在于，所述划分包括基于是“紧密集”、“稀疏邻域”和“小集势”的每个部分来定义多个多元组的多个部分。

模糊重复的鲁棒检测器

技术领域

本发明一般涉及有关数据库和数据仓库的技术。

背景技术

对数据仓库的决策支持分析影响重要的商业决策；因此，这种分析的准确性就十分重要。然而，在数据仓库接收的来自外部源的数据常常包含错误（例如拼写错误、各数据源之间的不一致惯例）。这些错误常常导致多元组（tuple）的重复输入。因此，大量时间和金钱花费在作为检测和纠正数据中错误的任务的数据清理。

数据库中重复多元组的检测和消除的问题是数据清除和数据质量的广泛领域内的主要问题之一。它常常是同一逻辑实际实体在数据仓库内可具有多个表示的情形。

例如，当名为 Isabel 的顾客从 SuperMegaMarket 购物两次时，她的名字可能显现为两个不同记录：[Isabel Christie, Seattle, WA, USA, 98025] 和 [Christy Isabel, Seattle, WA, United States, 98025]。该差异可能是由于数据输入错误和/或输入数据的销售人员的喜好不同。

这样的重复信息可大大增加直接邮寄成本，因为若干像 Isabel 的客户会接收多份目录。在受紧张预算约束的直接邮寄活动中，这样的错误可能会导致活动成败之间的差异。此外，这种错误可导致错误的查询结果（例如，在西雅图有多少 SuperMegaMarket 客户？）以及错误的分析模型创建。

从数据库中去掉看起来不同，但实际重复的条目就是模糊重复消除（fuzzy duplicates elimination）问题。在这里，“模糊重复”是非准确匹配但表示同一实际实体或现象的看起来不同的多元组（即记录）。

该问题不同于标准的准确重复消除问题，其中仅当两个多元组准确匹配所有属性才被认为是重复。除非上下文明确指出，否则此后将假设对重复检测和消除的引用集中在模糊重复消除问题上。

以前对模糊重复消除的解决方案可分成有监督方法和无监督方法。有监督方法从由已知重复组成的训练数据中学习特征化重复对的规则。此外，这些方法假设

训练数据展现了在实践中观察到的各种各样的错误。要获取这样的综合性训练数据是困难的（如果不是不可能的），在一定程度上，通过存在着需要交互式手动引导缺点的主动学习方法可解决一问题。在许多实际的数据集成情形中，要获取好的训练数据或交互式用户引导是不可能的。

无监督重复消除方法的问题与群集技术的问题相似，因为它们都尝试将一数据集分成不相交的组合。但是，在标准群集设计和重复消除问题之间还是有某些明显的差异。这些差异将在后面讨论。

当前的无监督方法旨在略去这些差异，而相反依赖于适用于检测重复对的多属性多元组和基于阈值约束之间的标准文本相似性函数（例如众所周知的诸如编辑距离和余弦度量的单链接群集算法）。然而，这种基于阈值方法导致大量的假阳性（不是真正的重复但被推测成是的多元组）或大量的假阴性（实际上是重复但不能识别出的多元组）。

发明内容

在此所述的至少一个实现可检测出模糊重复并消除这样的重复。模糊重复是在数据库中表示同一实际实体或现象的多个看起来不同的多元组（即记录）。

附图说明

相同的标号在附图中用来引用相似元件和特征。

图1是本文所述一个实现的框图。

图2是示出本文所述一方法实现的流程图。

图3是能够（全部或部分）实现本文所述的至少一实施例的计算操作环境的一个示例。

具体实施方式

以下描述陈述便于数据库中模糊重复多元组的检测和消除的技术。这些技术可用许多方法实现，包括（但不限于）程序模块、通用和专用计算系统、专用电器、以及一个或多个计算机网络的一部分。

这些技术的一示例性实现可被称为“示例性模糊重复检测器”，并如下所述。

示例性模糊重复检测器可解决模糊重复消除问题。在这里，“模糊重复”是非准确匹配但表示同一实际实体或现象的看起来不同的多元组（即记录）。检测和消

除模糊重复就是模糊重复消除问题。

标准特征化重复

在检测模糊重复时，示例性模糊重复检测器利用常规方法不使用的至少两个新的约束。特别地，这两个新标准被称为紧密集（CS）和稀疏邻域（SN）。这些标准明确地捕捉数据的本地结构性属性，用来特征化重复多元组的组合。

该 CS 和 SN 标准捕捉这些属性：

- 组合中的重复彼此比其它更相近；以及
- 重复多元组的“本地领域”是空的或稀疏的。

满足这些标准的多元组可组合在一起成为重复，即使它们相距遥远，而较相近但不满足这些标准的多元组则不能组合在一起。这些本地化的结构化属性将重复消除问题与标准群集设计区分开来。

ID	艺术家名字	乐曲名
1*	The Doors	LA Woman
2*	Doors	LA Woman
3*	The Beatles	A Little Help from My Friends
4*	Beatles, The	With A Little Help From My Friend
5*	Shania Twain	Im Holding on to Love
6*	Twain, Shania	I'm Holding On To Love
7	4th Elemynt	Ears/Eyes
8	4th Elemynt	Ears/Eyes - Part II
9	4th Elemynt	Ears/Eyes - Part III
10	4 th Elemynt	Ears/Eyes - Part IV
11	Aaliyah	Are You Ready
12	AC DC	Are You Ready
13	Bob Dylan	Are You Ready
14	Creed	Are You Ready

表格 1

来自媒体数据库的示例。标有星号的多元组为重复多元组。

表格 1 提供了典型音乐数据库的一个示例。前 6 个多元组（标有星号“*”）为重复多元组，而剩下的多元组（7-14）是唯一的。

紧密集标准

紧密集（CS）标准是重复集也是互相最相近邻居的紧密集。该标准的前提是重复多元组彼此比与其它不同多元组更相近。即，重复多元组常常互相是最相近的

邻居。对于表格 1 中的示例，多元组 1 是多元组 2 的最相近邻居，反之亦然。相反，多元组 8 可以是多元组 7 的最相近邻居，而多元组 9 则是多元组 8 的最相近邻居。

相反，基于单链接群集的常规的基于阈值方法则假设传递性（即如果‘a’是‘b’的重复，且‘b’是‘c’的重复，则‘a’是‘c’的重复）并在阈值图中标识相连组件。因此，它们更可能产生大量的假阳性。

稀疏邻域标准

稀疏邻域（SN）标准的前提是一组重复的本地邻域是稀疏的。例如。表格 1 中唯一多元组 7-14 都不满足该标准，其中出现了比重复集更大的组（自此例中为 4）。

从一个角度而言，多元组组合的本地邻域是根据周围区域大小所定义的紧邻，它取决于多元组本地分布。例如，它可以是半径为 $2 \cdot nn(v)$ 的球，其中 $nn(v)$ 是多元组 v 的最相近邻居的距离。

如果多元组周围的增长率—外部球中多元组的数量—较小，则本地邻域可被称为“稀疏”。该概念可扩展成一个多元组组合，而如果多元组的单个增长率集聚较小（例如限于阈值 c ）则它们的联合本地邻域被称为“稀疏”。例如，集聚函数 $maximum$ （最多）要求该组合中所有多元组的邻域值应小于阈值，而函数 $average$ （平均）则仅要求所有增长率的平均值较小。 $Maximum$ 函数比 $average$ 函数更为严格。

标准的定形

在以下定义中，设 R 为一关系（即数据集）且 $d: R \times R \rightarrow [0, 1]$ 是 R 中多元组之间的对称距离函数。为了清楚地展示，因此假设(i) 仅当多元组完全相同时两个多元组之间的距离为零；以及(ii) R 中没有两个多元组是彼此相同的。该假设的有效性可通过更改 d 以当多元组完全相同时返回 0 否则返回 $d(v_1, v_2) + \epsilon$ ，其中 ϵ 为 >0 的较小值。

CS 标准：一来自 R 的多元组集 S 是紧密集，如果对于 S 中的每个多元组 v ， v 和 S 中任何其它多元组 v' 之间的距离 $d(v, v')$ 都小于 v 和 $R-S$ 中任何其它 v'' 之间的距离 $d(v, v'')$ 。

SN 标准：对于一多元组 v ，考虑两个同心球：较小球具有半径 $nn(v)$ 为 v 与其最相近邻居之间距离，且较大球具有半径 $g(nn(v))$ ($>nn(v)$)。这里，使用 $g(x) = 2x$ 。邻域增长率 $ng(v)$ 是在 v 周围较大球中的点的数量。

设 $AGG: 2^R \rightarrow R$ 是一集聚函数且 c (>0) 是一常数。如果(i) $|S| = 1$ 或(ii) S 中所有

多元组的邻域增长率的集聚值小于 c (即 $AGG(\{ng(v): v \text{ in } S\}) < c$)，则多元组集 S 是一 $SN(AGG, c)$ 组合。

SG (小组合) 标准：多元组组合的另一个特征可被视为是它们常常非常小。重复组合 G 如果 $|G| \leq K$ ， K 为某大于 1 的预定义常数，则该组合较小。这也可被称为“小集势”标准，因为组合的集势（即成员的数量）较小。

示例性模糊重复检测器

一般而言，示例性模糊重复检测器将输入关系 R （例如数据库的数据集）分成最少量的“有效”组合，其中如果组合较小并满足 CS 和 SN 标准则该组合有效。

在示例性模糊重复检测器的环境中，这是重复消除 (DE) 问题：给定一关系 R ，距离函数 d ，正整数 $K (>1)$ ，集聚函数 AGG ，以及一正实数 c ，示例性模糊重复检测器将 R 分成最少量的组合 $\{G_1, \dots, G_m\}$ ，使得对于所有的 $1 \leq i \leq m$ ：

- $|G_i| \leq K$,
- G_i 是一紧密集，以及
- G_i 是一 $SN(AGG, c)$ 组合。

图 1 示出一个模糊重复检测和消除 (FDDE) 系统 100。该示例性模糊重复检测器可由在此所述的 FDDE 系统 100 实现。该 FDDE 系统 100 的组件可用软件、硬件、或其组合来实现。一般而言，FDDE 系统 100 确定最相近邻居集相等的多元组对，然后将对的等同性扩展到多元组组合中。

FDDE 系统 100 可以是服务器或客户机的一部分。该 FDDE 系统 100 可以是能够管理数据库的通用计算机、专用计算机、便携式计算机、个人数字助理 (PDA)、电子设备、专用电器、或其它这种计算和电器装置上实现。

FDDE 系统 100 的组件在图 1 虚线内示出。所示系统 100 包括数据集获取子系统 110、最相近邻居计算子系统 120、划分子系统 130、以及重复消除子系统 140。

系统 100 与数据仓库服务器 160 耦合（例如，通过直接连接或诸如因特网的网络）。在这样的服务器 160 上是数据库 170。系统 100 和数据库 170 可驻留在同一计算装置上（诸如服务器 160）或不同装置上。

使用该系统 100，获取器 110 从数据库 170 获取一数据集。该数据集是系统 100 的操作的主题。该数据集还可被称为“关系”和/或简称为“ R ”。这样的获取示例包括接收带有多元组集的数据包、访问包含多元组的已存储文件、和/或发送查询并接收该查询的结果。

最相近邻居计算子系统 120 确定在已获取关系中每个多元组的 K 个最相近邻

居。该子系统 120 生成索引或利用现有索引，这些索引允许对以下查询进行有效回应：对于 R 中任何给定多元组 v ，取得其 K 个最相近邻居。

对于关系 R 中的每个多元组，最相近邻居计算子系统 120 确定其最相近的 K 个邻居及其邻域增长率。该子系统的输出是一关系 $NN_Reln[ID, ID1, \dots, IDK, NG]$ ，其中 $ID1, \dots, IDK$ 是多元组 v 的带有标识符 D 和邻域增长率 NG 的 K 个最相近邻居的标识符。给定可用来取得 K 个最相近邻居并计算邻域增长率的一个索引，该子系统对 R 中的每个多元组 v 扫描输入关系 R ，查找索引并将多元组 $[v, v1, \dots, vK, ng(v)]$ 写入输出。

划分子系统 130 将该关系划分成最少量的有效重复组合。在至少一个实现中，该子系统利用标准数据库（例如 SQL）进行查询。这样，划分子系统 130 采用数据库系统的查询处理能力并避免在客户机和服务器之间移动大量数据。

划分子系统 130 使用第一阶段的输出将输入关系划分成最少量的紧密 SN 集。最终所产生的划分是对 DE 问题的解决方案。

基于由划分子系统对模糊重复的标识，重复消除子系统 140 通过去除过量的重复多元组来清理数据库。该子系统更改数据库 170 以去除这些标识的重复。

索引查找顺序

在这里，讨论有关最相近邻居计算子系统 120 的操作的其它细节。特别地，访问索引的顺序是此处的焦点。

通常用来获取 K 个最相近邻居的索引结构是基于盘的。换言之，它们访问存储在较慢次要存储系统（像硬盘）上的数据，而不是存储在可较快访问的主要存储器上的数据（例如随机存取存储器）。

如果以常用的基于盘的最相近邻居索引查找的连续多元组彼此相近，则查找过程可能访问索引的同一部分。因而，第二次查找获益于第一次查找。这大大改进了缓冲器命中率和整体运行时间。该示例性模糊重复检测器利用可有效实现的查找顺序。

考虑表格 1 中的示例多元组。假设系统查找 R 中多元组的最相近邻居的顺序是 1、12、5 等。为了获取多元组 1（“The Doors, LA Woman”）的最相近邻居，索引过程将访问索引的一部分，且在该过程中将其缓存在数据库缓冲器中。对多元组 12（“Aliyah, Are you ready”）的最相近邻居的相似查找将访问索引的完全不同部分，因为多元组 12 距离多元组 1 非常遥远。或者，如果我们在处理多元组 1 之后查找多元组 2（“Doors, LA Woman”）的最相近邻居，系统将使用索引的几乎同一部分。

因此，系统将利用它已在数据库缓冲器中的优点。

该示例性模糊重复检测器采用满足良好查找顺序的两个属性的宽度优先(BF)顺序：首先，在顺序上紧接任何多元组之前的多元组必须与其相近。其次，用于排序输入多元组的过程必须有效。

该顺序对应于对构建如下的树 T 的宽度优先遍历：任何输入多元组被选为树的根。树中任何节点的子节点是其还未插入树中的 K 个最相近邻居。注意，该示例性模糊重复检测器实际上并不需要建立这样的树而只需获取适当顺序的输入多元组。

BF 顺序中的每个多元组（除根外）总被其兄弟节点或其父节点的兄弟节点的子节点领先。这些多元组彼此之间比任意的多元组对更为相近。因此，该查找顺序中的所有多元组被与它们非常相近的多元组领先。从而，查找算法形成本地化的索引访问。

对于输入多元组的每次查找，都可获取其 K 个最相近多元组。因此，当遇到一多元组时，系统在查找其父多元组时已经按 BF 顺序获得了它。因此，数据库缓冲器已缓存该多元组。系统可显式地缓存这些多元组（如果存储器允许）或依赖数据库系统来保护最近的访问。

在这里，对最相近邻居索引的引用预期对准确或概率性的最相近邻居索引的使用。本邻域技术人员熟悉基于标准距离函数（例如余弦度量、编辑距离和模糊匹配相似性）利用近似和概率的许多常规索引。

方法论实现

图 2 示出 FDDE 系统 100 的方法实现。该方法论实现可用软件、硬件或其组合来实现。为便于理解，方法步骤可被视为单独步骤；然而，这些单独示出的步骤不必被解释为是其执行中决定的必须顺序。

为作说明，该方法论实现的描述讨论了其在一特定示例中的应用。在该示例中，组合{10, 50, 100, 150}形成一紧密 SN 集。除每个多元组的邻域增长(NG)值之外，足以知道{10, 50}, {10, 100}, {10, 150}对的 4 个最相近邻居集都是相等的。可从对的等同性和集等同性的传递性中推断，该组合{10, 50, 100, 150}是一密集组合。

在图 2 的 212，FDDE 系统 100 可从数据库 170 中获取一数据集（即关系）。

在 214，FDDE 系统 100 计算在获取关系中每个多元组的 K 个最相近邻居。为此，它可利用现有索引或生成它自己的索引。

作为框 214 的一部分,FDDE 系统 100 计算多元组对之间变化尺寸的邻居集的等同性。即,对于一多元组对 (10,100),该系统确定 K 个最相近邻居集中它们的 2 个-最相近邻居集、3 个-最相近邻居集等都相等。在多元组 v1 和 v2 之间的这种比较产生布尔结果 CS2, ..., CSK, 以及它们的邻域增长率 $ng(v1)$ 和 $ng(v2)$ 。数值 CS_i ($2 \leq i \leq K$)表示 v1 和 v2 的 I-邻居集是否相等。该结果被存储在一临时的关系 CS 对 (CSPair) 中。

仍作为框 214 的一部分,FDDE 系统 100 可根据第一阶段的输出(NN_ReIn)发出一数据库查询(例如 SQL (select into)查询)来计算 CSPair 关系。该查询涉及在多元组 NN_ReIn1.ID 小于 NN_ReIn2.ID 且它在 NN_ReIn2.ID 的 K 个-最相近邻居集中的谓词上 NN_ReIn 的自结合(例如 FROM NN_ReIn1, NN_ReIn2),反之亦然。

选定列列表具有以下两部分:(i) NN_ReIn1.ID, NN_ReIn2.ID, NN_ReIn1.NG, NN_ReIn2.NG, 以及(ii)对于 2 和 K 之间的每个 j, 如果 NN_ReIn1.ID 的 j 个-最相近邻居集等于 NN_ReIn2.ID 的 j 个-最相近邻居集, case 语句返回 1。注意,该查询可使用标准 SQL 来表达,并在数据库服务器上执行所有的计算。

在 216, FDDE 系统 100 将关系划分成最少量的有效重复组合。所产生的有效重复组合满足以下标准的每一项:

- $|G_i| \leq K$,
- G_i 是一紧密集, 以及
- G_i 是一 $SN(AGG, c)$ 组合。

<u>ID</u> : NN1, NN2, NN3, ..., NG(TID)	
10	: 100, 50, 150, ..., 2.0
50	: 10, 150, 100, ..., 2.0
100	: 50, 10, 150, ..., 3.0
150	: 10, 100, 100, ..., 2.0
...	

<u>ID1, ID2</u> : CS2, CS3, CS4, ..., NG(ID1), NG(ID2)	
10, 50	: 0, 0, 1, ..., 2.0, 2.0
10, 100	: 0, 1, 1, ..., 2.0, 3.0
10, 150	: 0, 0, 1, ..., 2.0, 2.0
...	

<u>ID</u> : 紧密 SN 组合	
10	: 50, 100, 150
...	

表格 2: 示出划分的示例

作为框 216 的一部分,FDDE 系统 100 将多元组对的邻居集之间的等同性扩展到多元组集,并确定邻居集是否紧密并满足 SN 标准。在表格 2, 多元组对(10, 50), (10, 100)和(10, 150)的 4 个-邻居集相等,且因此形成大小为 4 的紧密集。集{10, 50,

100, 150}可输出一重复组合, 假如(i)该组合的集聚 SN 值小于阈值 c , 以及(ii)它不能扩展为较大的密集 SN 组合。对(50, 100), (50, 150)对的 4 个-邻居集是否因集等同性是可传递的而相等未作显式检查。

FDDE 系统 100 通过发出以下 CS-组合查询: “select * from CSPairs order by ID” 来处理 CSPair 关系(框 214 的输出)以组合多元组 v 及其邻居 v' (其中 $v.ID < v'.ID$) 之间所有邻居集的比较结果。注意, 在 CS-组合查询的结果中, 每个紧密 SN 集 G 将在 G 中带有最小 ID 的多元组下组合在一起。

FDDE 系统 100 处理 CS-组合查询的结果中多元组的每个组合 $Q[ID=v]$ (由同一 ID 标识)。对于多元组的一个组合 $Q[ID=v]$, 如果 v 未被标识为属于一紧密 SN 集, 则系统确定 v 所属的最大的非琐碎 (尺寸大于 1) 紧密 SN 集。该集可仅从组合 $Q[ID=v]$ 中标识。系统输出集 G_v 并标记 G_v 中每个多元组以指示其向紧密集的分配。例如, 对多元组 10 的集是 {10, 50, 100, 150}。集 {10, 50, 100, 150} 可输出, 并将多元组标识符 10, 50, 100, 150 标记为属于一有效紧密 SN 集, 从而不必再次处理它们。

在 218, FDDE 系统 100 通过去移除过多的重复多元组来清理数据库 170, 其中这些重复多元组的标识是基于框 216 的结果。

其它实现细节

由示例性模糊重复检测器提供的模糊重复消除问题的解决方案比现有的常规方法更准确有效。它是比例不变的, 使得距离函数的比例的确影响多元组的本地结构化属性。它是分/合一致的, 因为重复组合中多元组之间的距离缩小和组合中多元组之间的距离扩大可仅在有限方面改变划分。它具有受限丰富性, 从而重复消除函数的范围允许在实践中有用的所有组合。

在至少一个实现中, 用户可指定稀疏邻域阈值 c 。然而, 计算可自动完成或可在阈值 c 的计算中提供对用户的帮助。

示例性计算系统和环境

图 3 示出其中可 (全部或部分) 实现本文所述的示例性模糊重复检测器的适当计算环境 300 的一个示例。计算环境 300 可在本文所述的计算机和网络体系结构中利用。

该计算机环境 300 仅是适当计算环境的一个示例, 并非旨在提出对本发明使用或功能性范围作任何限制。计算环境 300 也不应被解释为对示例性操作环境 300 中所示的任一组件或其组合有任何依赖性 or 任何需求。

示例性模糊重复检测器可在很多其它通用或专用计算系统环境或配置中实现。适于本发明使用的众所周知的计算系统、环境、和/或配置的示例包括，但不限于，个人计算机、服务器计算机、瘦客户机、手持式或膝上型装置、多处理器系统、基于微处理器的系统、机顶盒、个人数字助理（PDA）、电器、专用电器（例如 DVD 播放器）、可编程消费电器、网络 PC、小型计算机、大型机、包括任一种以上系统或设备的分布式计算环境等等。

示例性模糊重复检测器可在处理器可执行指令的一般环境中进行说明，诸如由计算机执行的程序模块。一般而言，程序模块包括执行具体任务或实现具体抽象数据类型的例程、程序、对象、组件、数据结构等等。示例性模糊重复检测器还可在任务由经通信网络或其它数据传输介质连接的远程处理装置执行的分布式计算环境中实践。在分布式计算环境中，程序模块和其它数据可置于包括存储器存储设备的本地和远程计算机存储介质中。

计算机环境 300 包括以计算机 320 形式的通用计算装置。计算机 320 的组件可包括，但不限于，一个或多个处理器或处理单元 304、系统存储器 306 以及把包括处理器 304 在内的各种系统组件耦合到系统存储器 306 的系统总线 308。

系统总线 308 代表任一若干总线结构类型中的一种或多种，包括存储器总线或存储器控制器、外围总线、加速图形端口、以及使用各种总线架构的任一种的处理器或本地总线。作为示例，这些架构可包括卡总线、个人计算机存储卡国际协会（PCMCIA）、加速图形端口（AGP）、小型计算机系统接口（SCSI）、通用串行总线（USB）、IEEE 1394、视频电子标准协会（VESA）局部总线和也称为 Mezzanine 总线的外围部件互连（PCI）总线。

计算机 302 通常包括各种处理器可读介质。这些介质可以是能被计算机 302 访问的任何可用介质，并包括易失性和非易失性介质、可移动和不可移动介质。

系统存储器 306 包括诸如只读存储器（ROM）312 和随机存取存储器（RAM）310 的易失性和/或非易失性存储器形式的计算机可读介质。包含有助于如起动时在计算机 302 内元件间传送信息的基本例程的基本输入/输出系统（BIOS）314 通常存储在 ROM 312 中。RAM 310 通常包含可被处理单元 304 立即访问和/或当时正被操作的数据和/或程序模块。

计算机 302 还可包括其它可移动/不可移动、易失/非易失计算机存储介质。作为示例，图 3 示出了读取和写入不可移动、非易失性介质（未示出）的硬盘驱动器 316，读取和写入可移动、非易失磁盘 320（例如“软盘”）的磁盘驱动器 318，读

取和写入可移动、非易失光盘 324, 诸如 CD-ROM、DVD-ROM、或其它光学介质的光盘驱动器 322。硬盘驱动器 316、磁盘驱动器 318 和光盘驱动器 322 分别通过一个或多个数据介质接口 325 与系统总线 308 连接。或者, 硬盘驱动器 316、磁盘驱动器 318 和光盘驱动器 322 分别通过一个或多个接口(未示出)与系统总线 308 连接。

盘驱动器及其相关联的计算机可读介质为计算机 302 提供计算机可读指令、数据结构、程序模块、和其它数据的非易失性存储。尽管示例示出硬盘 316、可移动磁盘 320 和可移动光盘 324, 可以理解可利用其它类型的可存储数据可由计算机访问的计算机可读介质来实现示例性计算系统和环境, 诸如磁卡或其它磁性存储装置、闪存卡、CD-ROM、数字多功能盘(DVD)或其它光学存储、随机存取存储器(RAM)、只读存储器(ROM)、电可擦除只读存储器(EEPROM)等。

包括作为示例的操作系统 326、一个或多个应用程序 328、其它程序模块 330 和程序数据 332 的任何数量的程序模块可被存储在硬盘 316、磁盘 320、光盘 324、ROM 312、和/或 RAM 310 中。

用户可通过输入装置如键盘 334 和定位装置 336(例如“鼠标”)向计算机 302 输入命令和信息。其它输入装置 338(未具体示出)可包括话筒、游戏杆、游戏垫、卫星接收器、串行端口、扫描仪等等。这些和其它输入装置常常通过与系统总线 308 耦合的输入/输出接口 340 与处理单元 304 相连, 但也可通过诸如并行端口、游戏端口或通用串行总线(USB)的其它接口和总线结构连接。

监视器 342 或其它类型的显示装置也可通过诸如视频接口 344 的接口与系统总线 308 相连。除了监视器 342, 其它输出外围装置还可包括通过输入/输出接口 340 与计算机 302 相连的诸如扬声器(未示出)和打印机 346 的组件。

计算机 302 可以在使用与一台或多台远程计算机, 诸如远程计算装置 348 的逻辑连接的网络化环境中运行。作为示例, 远程计算装置 348 可以是个人计算机、服务器、路由器、网络计算机、对等装置或其它公共网络节点等。远程计算装置 348 被示为包括本文所述与计算机 302 相关的许多或全部元件的便携式计算机。

计算机 302 和远程计算机 348 之间的逻辑连接被示为局域网(LAN) 350 和广域网(WAN) 352。这样的网络化环境在家庭、办公室、企业范围计算机网络、企业内部互联网和因特网上是常见的。这样的网络化环境可以有或无线的。

当在 LAN 网络化环境中实现时, 计算机 302 通过网络接口或适配器 354 与局域网 350 连接。当在 WAN 网络化环境中实现时, 计算机 302 通常包括调制解调器

356 或其它用于在广域网 352 上建立通信的装置。可以在计算机 302 内部或外部的调制解调器 356 与系统总线 308 通过输入/输出接口 340 或其它适当机制连接。应当理解,所示网络连接是示例性的,且也可以使用其它在计算机 302 和 348 之间建立通信链接的技术。

在诸如计算环境 300 所示的网络化环境中,所示与计算机 302 相关的程序模块或其一部分可存储在远程存储器存储装置中。作为示例,远程应用程序 358 驻留于远程计算机 348 的存储装置上。为进行说明,应用程序和诸如操作系统的其它可执行程序组件在此被示为单独的框,尽管可以理解这种程序和组件在不同时间驻留在计算装置 302 的不同存储组件上,并由该计算机的数据处理器执行。

处理器-可执行指令

示例性模糊重复检测器可在由一个或多个计算机或其它装置执行的诸如程序模块的计算机可执行指令的一般环境中描述。一般而言,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。通常,程序模块的功能可在各个实施例中按需组合或分布。

示例性操作环境

图 3 示出其中可实现示例性模糊重复检测器的适当计算环境 300 的一个示例。特别地,本文所述的示例性模糊重复检测器可(全部或部分地)由图 3 中的任何程序模块 328-330 和/或操作系统 326 或其一部分来实现。

该操作环境仅是适当操作环境的一个示例,并非旨在提出对示例性模糊重复检测器的范围或功能性使用作任何限制。其它适于使用的众所周知的计算系统、环境、和/或配置的示例包括,但不限于,个人计算机、服务器计算机、手持式或膝上型装置、多处理器系统、基于微处理器的系统、可编程消费电器、无线电话或装置、通用或专用电器、应用程序专用集成电路、网络 PC、小型计算机、大型计算机、包括任一种以上系统或设备的分布式计算环境等等。

处理器-可读介质

示例性模糊重复检测器的实现可存储于某些形式的计算机可读介质上或在其上传输。处理器可读介质可以是处理器可访问的任何可用介质。作为示例,处理器可读介质可包括但不限于“计算机存储介质”和“通信介质”。

“计算机存储介质”包括以任何方法或技术实现的用于存储诸如计算机可读指令、数据结构、程序模块、或其它数据的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括,但不限于, RAM、ROM、EEPROM、闪存或其它存

储器技术、CD-ROM、数字多功能盘（DVD）或其它光学存储、磁卡、磁带、磁盘存储或其它磁性存储装置、或任何其它可用来存储所需信息并可由计算机访问的介质。

“通信介质”通常包括计算机可读指令、程序模块、或在诸如载波或其它传输介质的已调制数据信号中的其它数据。通信介质还包括任何信息传送介质。

术语“已调制数据信号”指其特征的一个或多个以在信号中编码信息的方式设置或改变的信号。作为示例，而非限制，通信介质包括诸如有线网络或直接接线连接的有线介质，以及诸如声学、RF、红外线或其它无线介质的无线介质。以上的任意组合也被包括在计算机可读介质的范围内。

结论

尽管一个或多个上述实现使用了结构化特征和/或方法论动作的专用语言进行描述，可以理解其它实现无需所述特定特征或动作就可实践。相反，特定特征和动作被揭示为一个或多个实现的优选形式。

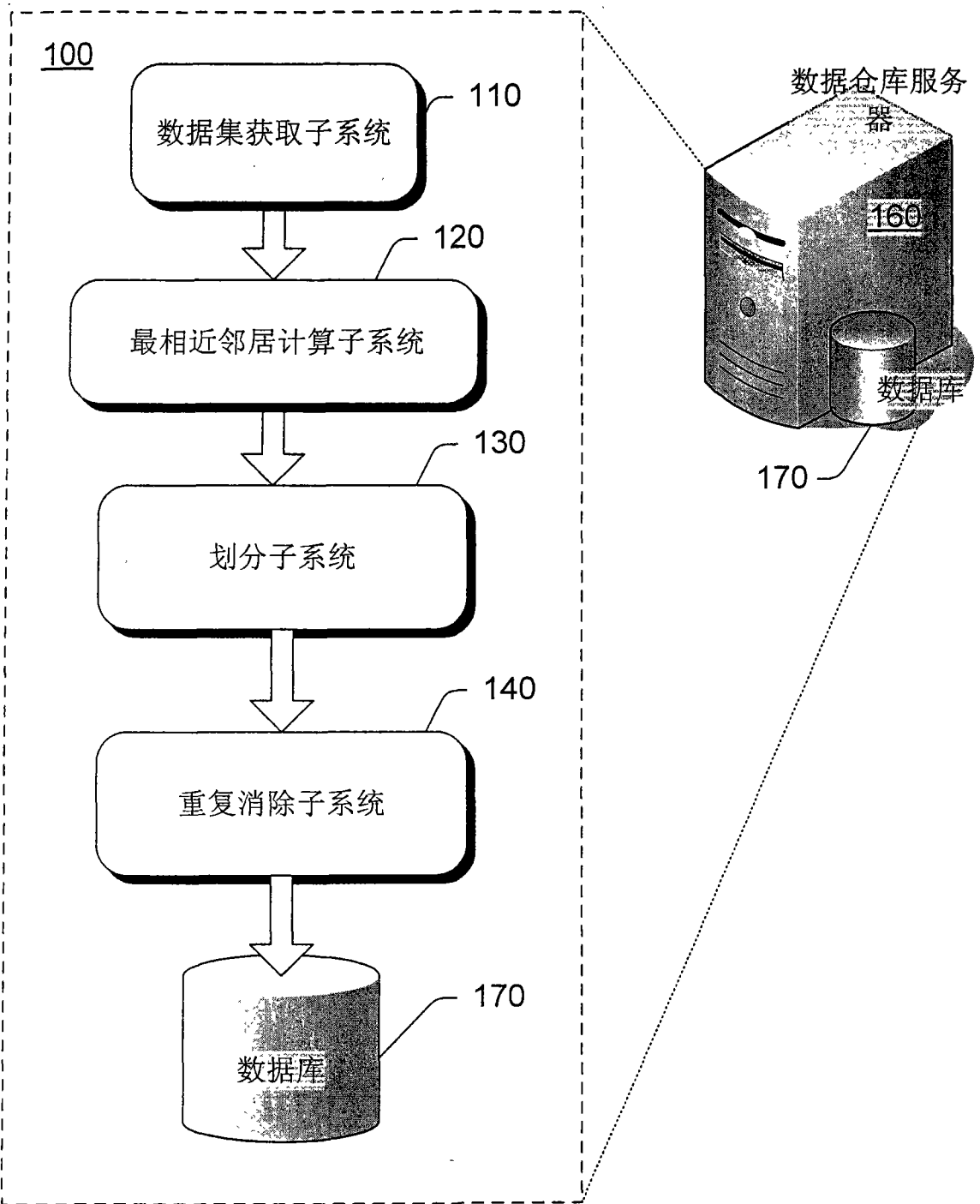


图 1

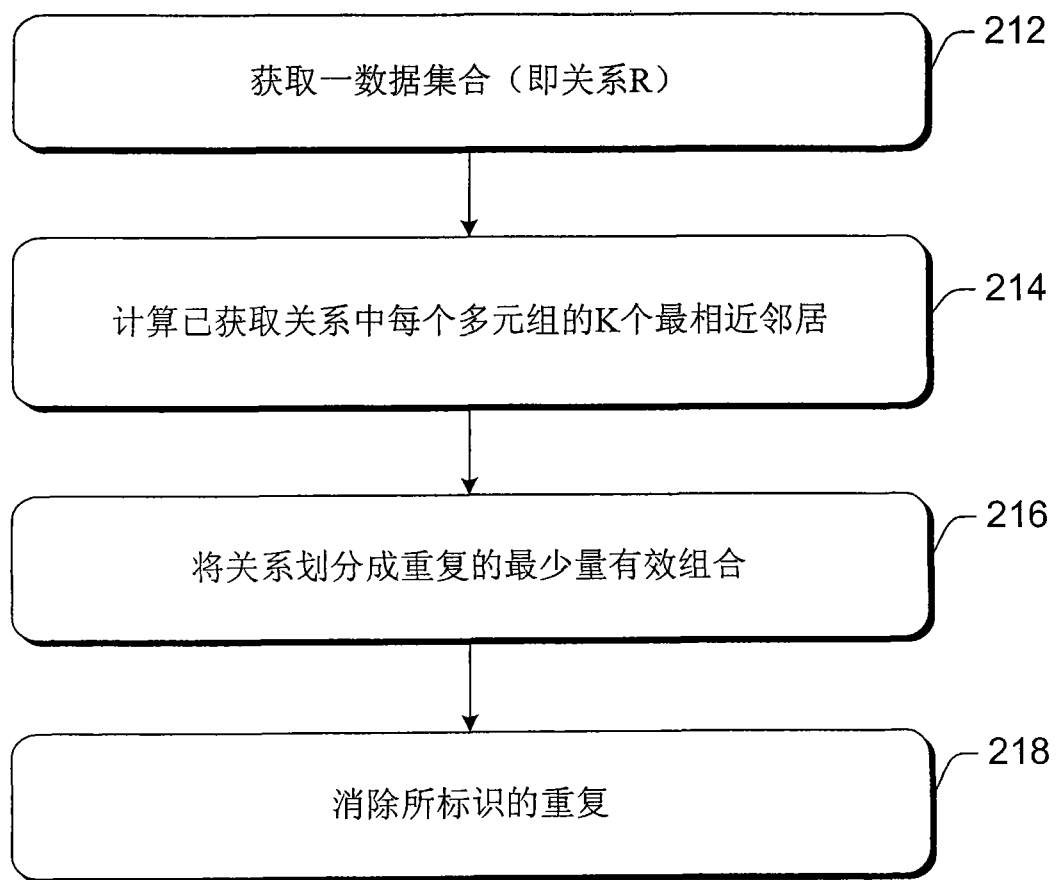


图 2

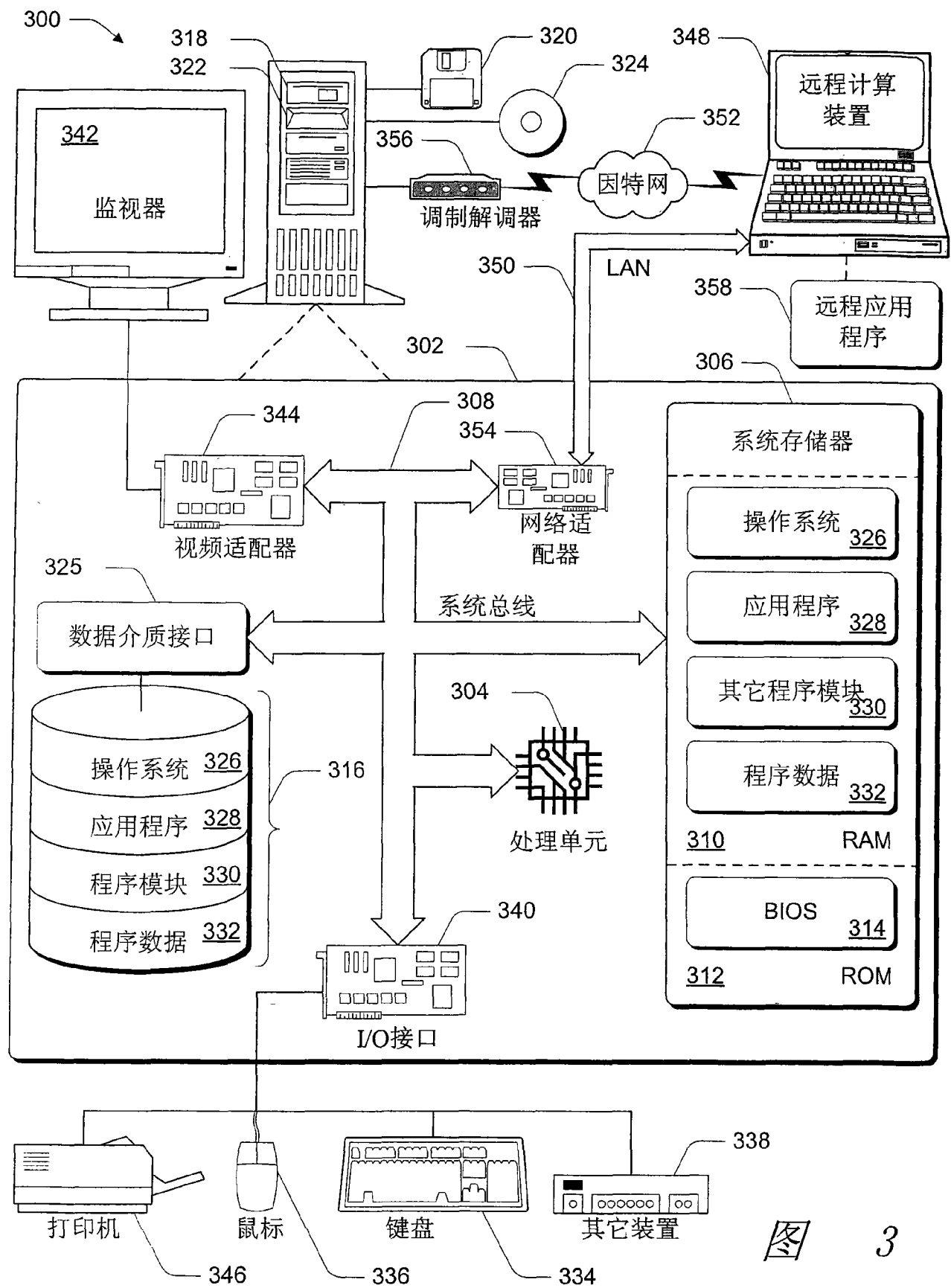


图 3