



(12)发明专利

(10)授权公告号 CN 105900057 B

(45)授权公告日 2018.10.02

(21)申请号 201480072395.4

(22)申请日 2014.02.19

(65)同一申请的已公布的文献号
申请公布号 CN 105900057 A

(43)申请公布日 2016.08.24

(30)优先权数据
50/DEL/2014 2014.01.07 IN

(85)PCT国际申请进入国家阶段日
2016.07.06

(86)PCT国际申请的申请数据
PCT/IB2014/059087 2014.02.19

(87)PCT国际申请的公布数据
W02015/104591 EN 2015.07.16

(73)专利权人 印度坎普尔理工学院
地址 印度北方邦坎普尔

(72)发明人 A·巴塔查里亚 S·米什拉

(74)专利代理机构 北京三友知识产权代理有限公司 11127

代理人 吕俊刚 杨薇

(51)Int.Cl.
G06F 7/00(2006.01)

(56)对比文件
US 8504581 B2, 2013.08.06,
CN 103150327 A, 2013.06.12,
CN 101895547 A, 2010.11.24,
CN 102323957 A, 2012.01.18,
CN 102254016 A, 2011.11.23,
CN 103150326 A, 2013.06.12,
CN 103150336 A, 2013.06.12,
CN 103308314 A, 2013.09.18,

审查员 於媛

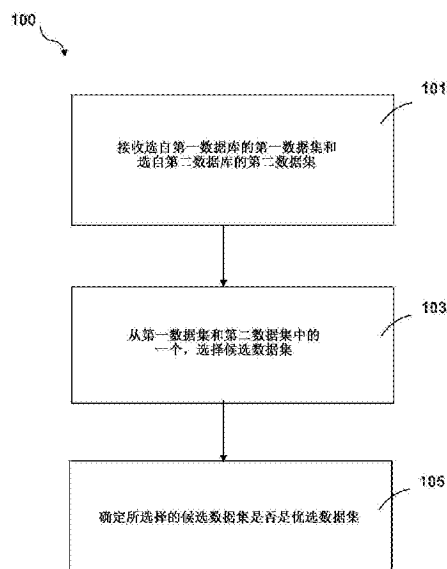
权利要求书5页 说明书8页 附图5页

(54)发明名称

分布式数据库中的多准则决策分析

(57)摘要

本公开的实施方式阐述了用于从分布式数据库选择优选数据集的方法,所述分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库。一个示例方法可包括接收来自所述第一数据库的第一数据集和来自所述第二数据库的第二数据集。所述第一数据集包括第一单调属性。所述第二数据集包括第二单调属性。示例方法还可包括基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集,以及确定所选择的候选数据集是否是优选数据集。



1. 一种用于从多个分布式数据库选择优选数据集的方法,所述多个分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库,该方法包括:

通过一个或更多个处理器接收选自所述第一数据库的第一数据集和选自所述第二数据库的第二数据集,所述第一数据集具有第一单调属性,所述第二数据集具有第二单调属性;

通过所述一个或更多个处理器,基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集;

通过所述一个或更多个处理器接收所述第一数据库中的数据的第一本地总结和所述第二数据库中的数据的第一本地总结;

通过所述一个或更多个处理器构造基于所述第一本地总结和所述第二本地总结的全局总结;以及

通过所述一个或更多个处理器,至少部分地基于根据所述全局总结确定的所选择的候选数据集是优选数据集的最小概率、根据所述全局总结确定的所选择的候选数据集是优选数据集的最大概率和阈值概率,确定所选择的候选数据集是否是优选数据集。

2. 根据权利要求1所述的方法,其中,所述第一数据集不受所述第一数据库中的其他数据集支配,所述第二数据集不受所述第二数据库中的其他数据集支配。

3. 根据权利要求1所述的方法,其中,所述第一单调属性小于所述第一数据库中的其他单调属性,所述第二单调属性小于所述第二数据库中的其他单调属性。

4. 根据权利要求1所述的方法,其中,选择所述候选数据集包括:当所述第一单调属性小于所述第二单调属性时,选择所述第一数据集作为所述候选数据集。

5. 根据权利要求1所述的方法,其中,所述全局总结与包括多个单元的网络结构关联,其中,所述多个单元中的每个单元存储与所述第一数据库和所述第二数据库中的数据的存在概率属性关联的概率。

6. 根据权利要求5所述的方法,其中,所述网络结构的单元是均匀的或非均匀的。

7. 根据权利要求5所述的方法,所述方法还包括:基于所述第一数据库或所述第二数据库中的、在所述全局总结中在针对所述候选数据集的支配区域内的其它数据集的相应存在概率属性来确定所述最大概率,并且其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最大概率小于所述阈值概率时,确定所选择的候选数据集不是所述优选数据集。

8. 根据权利要求5所述的方法,所述方法还包括:基于所述第一数据库或所述第二数据库中的、在所述全局总结中在针对所述候选数据集的互补支配区域内的其它数据集的相应存在概率属性来确定所述最小概率,并且其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最小概率大于所述阈值概率时,确定所选择的候选数据集是所述优选数据集。

9. 根据权利要求1所述的方法,所述方法还包括:当所述优选数据集选自所述第一数据库时,通过所述一个或更多个处理器按概率将所述优选数据集发送到所述第二数据库。

10. 根据权利要求1所述的方法,其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最大概率大于所述阈值概率并且所选择的候选数据集是所述优选数据集的最小概率小于所述阈值概率时,部分地基于与所选择

的候选数据集关联的分值,确定所选择的候选数据集不是所述优选数据集。

11.根据权利要求1所述的方法,所述方法还包括:当所选择的候选数据集是所述优选数据集的最大概率大于所述阈值概率并且所选择的候选数据集是所述优选数据集的最小概率小于所述阈值概率时,通过所述一个或更多个处理器广播所选择的候选数据集。

12.根据权利要求11所述的方法,所述方法还包括:通过所述一个或更多个处理器接收与所选择的候选数据集是所述第一数据库中的数据集中的支配者关联的支配概率。

13.根据权利要求12所述的方法,所述方法还包括:通过所述一个或更多个处理器,至少部分地基于所述支配概率,计算所选择的候选数据集是所述优选数据集的skyline概率。

14.根据权利要求1所述的方法,所述方法还包括:在确定所选择的候选数据集是否是所述优选数据集之后,通过所述一个或更多个处理器从所述第一数据库或所述第二数据库接收额外数据集。

15.一种用计算机可执行指令编码的非暂态计算机可读存储介质,所述计算机可执行指令用于从多个分布式数据库选择优选数据集,所述多个分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库,所述计算机可执行指令响应于被计算装置执行,致使所述计算装置:

通过一个或更多个处理器接收选自所述第一数据库的第一数据集和选自所述第二数据库的第二数据集,所述第一数据集具有第一单调属性,所述第二数据集具有第二单调属性;

通过所述一个或更多个处理器,基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集;

通过所述一个或更多个处理器接收所述第一数据库中的数据集中的第一本地总结和所述第二数据库中的数据集中的第二本地总结;

通过所述一个或更多个处理器构造基于所述第一本地总结和所述第二本地总结的全局总结;以及

通过所述一个或更多个处理器,至少部分地基于根据所述全局总结确定的所选择的候选数据集是优选数据集的最小概率、根据所述全局总结确定的所选择的候选数据集是优选数据集的最大概率和阈值概率,确定所选择的候选数据集是否是优选数据集。

16.根据权利要求15所述的非暂态计算机可读存储介质,其中,所述第一数据集不受所述第一数据库中的其他数据集支配,所述第二数据集不受所述第二数据库中的其他数据集支配。

17.根据权利要求15所述的非暂态计算机可读存储介质,其中,所述第一单调属性小于所述第一数据库中的其他单调属性,所述第二单调属性小于所述第二数据库中的其他单调属性。

18.根据权利要求15所述的非暂态计算机可读存储介质,其中,选择所述候选数据集包括:当所述第一单调属性小于所述第二单调属性时,选择所述第一数据集作为所述候选数据集。

19.根据权利要求15所述的非暂态计算机可读存储介质,其中,所述全局总结与包括多个单元的网格结构关联,其中,所述多个单元中的每个单元存储与所述第一数据库或所述第二数据库中的数据集中的存在概率属性关联的概率。

20. 根据权利要求19所述的非暂态计算机可读存储介质,其中,所述网络结构的单元是均匀的或非均匀的。

21. 根据权利要求19所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置基于所述第一数据库或所述第二数据库中的、在所述全局总结中在针对所述候选数据集的支配区域内的其它数据集的相应存在概率属性来确定所述最大概率,并且其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最大概率小于所述阈值概率时,确定所选择的候选数据集不是所述优选数据集。

22. 根据权利要求19所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置基于所述第一数据库或所述第二数据库中的、在所述全局总结中在针对所述候选数据集的互补支配区域内的其它数据集的相应存在概率属性来确定所述最小概率,并且其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最小概率大于所述阈值概率时,确定所选择的候选数据集是所述优选数据集。

23. 根据权利要求15所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置在所述优选数据集选自所述第一数据库时,按概率将所述优选数据集发送到所述第二数据库。

24. 根据权利要求15所述的非暂态计算机可读存储介质,其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最大概率大于所述阈值概率并且所选择的候选数据集是所述优选数据集的最小概率小于所述阈值概率时,部分地基于与所选择的候选数据集关联的分值,确定所选择的候选数据集不是所述优选数据集。

25. 根据权利要求15所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置在所选择的候选数据集是所述优选数据集的最大概率大于所述阈值概率并且所选择的候选数据集是所述优选数据集的最小概率小于所述阈值概率时,广播所选择的候选数据集。

26. 根据权利要求25所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置接收与所选择的候选数据集是所述第一数据库中的数据集中的支配者关联的支配概率。

27. 根据权利要求26所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置至少部分地基于所述支配概率,计算所选择的候选数据集是所述优选数据集的skyline概率。

28. 根据权利要求15所述的非暂态计算机可读存储介质,所述非暂态计算机可读存储介质还包含额外指令,当所述计算装置执行所述额外指令时,致使所述计算装置在确定所选择的候选数据集是否是所述优选数据集之后,从所述第一数据库或所述第二数据库接收额外数据集。

29. 一种计算装置,所述计算装置被配置为从多个分布式数据库选择优选数据集,所述多个分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库,该计算装置包括:

处理单元,其中,所述处理单元被配置为:

通过一个或更多个处理器接收选自所述第一数据库的第一数据集和选自所述第二数据库的第二数据集,所述第一数据集具有第一单调属性,所述第二数据集具有第二单调属性;

通过所述一个或更多个处理器,基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集;

通过所述一个或更多个处理器接收所述第一数据库中的数据的第一本地总结和所述第二数据库中的数据的第一本地总结;

通过所述一个或更多个处理器构造基于所述第一本地总结和所述第二本地总结的全局总结;以及

通过所述一个或更多个处理器,至少部分地基于根据所述全局总结确定的所选择的候选数据集是优选数据集的最小概率、根据所述全局总结确定的所选择的候选数据集是优选数据集的最大概率和阈值概率,确定所选择的候选数据集是否是优选数据集。

30. 根据权利要求29所述的计算装置,其中,所述第一数据集不受所述第一数据库中的其他数据集支配,所述第二数据集不受所述第二数据库中的其他数据集支配。

31. 根据权利要求29所述的计算装置,其中,所述第一单调属性小于所述第一数据库中的其他单调属性,所述第二单调属性小于所述第二数据库中的其他单调属性。

32. 根据权利要求29所述的计算装置,其中,选择所述候选数据集包括:当所述第一单调属性小于所述第二单调属性时,选择所述第一数据集作为所述候选数据集。

33. 根据权利要求29所述的计算装置,其中,所述全局总结与包括多个单元的网格结构关联,其中,所述多个单元中的每个单元存储与所述第一数据库或所述第二数据库中的数据的数据集的存在概率属性关联的概率。

34. 根据权利要求33所述的计算装置,其中,所述网格结构的单元是均匀的或非均匀的。

35. 根据权利要求33所述的计算装置,其中,所述处理单元还被配置为:基于所述第一数据库或所述第二数据库中的、在所述全局总结中在针对所述候选数据集的支配区域内的其它数据集的相应存在概率属性来确定所述最大概率,并且其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最大概率小于所述阈值概率时,确定所选择的候选数据集不是所述优选数据集。

36. 根据权利要求33所述的计算装置,其中,所述处理单元还被配置为:基于所述第一数据库或所述第二数据库中的、在所述全局总结中在针对所述候选数据集的互补支配区域内的其它数据集的相应存在概率属性来确定所述最小概率,并且其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最小概率大于所述阈值概率时,确定所选择的候选数据集是所述优选数据集。

37. 根据权利要求29所述的计算装置,其中,所述处理单元还被配置为:当所述优选数据集选自所述第一数据库时,按概率将所述优选数据集发送到所述第二数据库。

38. 根据权利要求29所述的计算装置,其中,确定所选择的候选数据集是否是所述优选数据集包括:当所选择的候选数据集是所述优选数据集的最大概率大于所述阈值概率并且所选择的候选数据集是所述优选数据集的最小概率小于所述阈值概率时,部分地基于与所

选择的候选数据集关联的分值,确定所选择的候选数据集不是所述优选数据集。

39. 根据权利要求29所述的计算装置,其中,所述处理单元还被配置为:当所选择的候选数据集是所述优选数据集的最大概率大于所述阈值概率并且所选择的候选数据集是所述优选数据集的最小概率小于所述阈值概率时,广播所选择的候选数据集。

40. 根据权利要求39所述的计算装置,其中,所述处理单元还被配置为:接收与所选择的候选数据集是所述第一数据库中的数据集中的支配者关联的支配概率。

41. 根据权利要求40所述的计算装置,其中,所述处理单元还被配置为:至少部分地基于所述支配概率,计算所选择的候选数据集是所述优选数据集的skyline概率。

42. 根据权利要求29所述的计算装置,其中,所述处理单元还被配置为:在确定所选择的候选数据集是否是所述优选数据集之后,从所述第一数据库和所述第二数据库接收额外数据集。

分布式数据库中的多准则决策分析

[0001] 相关申请的交叉引用

[0002] 本申请是PCT申请,要求2014年1月7日提交的印度专利申请No. 50/DEL/2014的优先权,该申请的全部内容的全文以引用方式并入本文中。

背景技术

[0003] 多准则决策分析通常是指在具有各种不确定性的多种关系之间进行众多时常冲突的评价的方法。这种多准则决策分析具有各种应用。一个示例应用是通过接口从不同源发现旅行交易。该接口被配置为访问没有直接彼此访问的不同源。一些示例源可包括旅行搜索引擎(例如,kayak.com,fly.com,skyscanner.com)和酒店搜索引擎(例如,trivago.com,room77.com,google.com/hotelfinder)。这些源的数据通常没有保持在一个中央仓库,而是替代地保持在不同数据库。

发明内容

[0004] 本公开的一些实施方式可总体上涉及用于在分布式数据库之中选择优选数据集的方法,所述分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库。一个示例方法可包括接收选自所述第一数据库的第一数据集和选自所述第二数据库的第二数据集。所述第一数据集包括第一单调属性。所述第二数据集包括第二单调属性。示例方法还可包括基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集;以及至少部分地基于所选择的候选数据集是优选数据集的最小概率、所选择的候选数据集是优选数据集的最大概率和阈值概率,确定所选择的候选数据集是否是优选数据集。

[0005] 本公开的额外实施方式可总体上涉及一种计算机可读介质,所述计算机可读介质包含用于在分布式数据库之中选择优选数据集的指令,所述分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库。一个示例计算机可读介质可包含指令,当计算装置执行所述指令时,致使所述计算装置接收选自所述第一数据库的第一数据集和选自所述第二数据库的第二数据集。所述第一数据集包括第一单调属性。所述第二数据集具有第二单调属性。示例计算机可读介质还可包含额外指令,当计算装置执行所述额外指令时,致使所述计算装置基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集;以及至少部分地基于所选择的候选数据集是优选数据集的最小概率、所选择的候选数据集是优选数据集的最大概率和阈值概率,确定所选择的候选数据集是否是优选数据集。

[0006] 本公开的其他实施方式可总体上涉及一种计算装置,所述计算装置被配置为在分布式数据库之中选择优选数据集的方法,所述分布式数据库包括具有第一概率属性的第一数据库和具有第二概率属性的第二数据库。一个示例计算装置包括处理单元。所述处理单元被配置为接收选自所述第一数据库的第一数据集和选自所述第二数据库的第二数据集。所述第一数据集包括第一单调属性。所述第二数据集包括第二单调属性。所述处理单元还

被配置为基于所述第一单调属性和所述第二单调属性的比较,从所述第一数据集和所述第二数据集中的选择一个选择候选数据集;以及至少部分地基于所选择的候选数据集是优选数据集的最小概率、所选择的候选数据集是优选数据集的最大概率和阈值概率,确定所选择的候选数据集是否是优选数据集。

[0007] 以上的发明内容仅仅是例证性的,并不旨在是任何方式的限制。除了上述的例证方面、实施方式和特征之外,通过参照附图和下面的具体实施方式,其他方面、实施方式和特征将变得清楚。

附图说明

[0008] 图1是用于在分布式数据库之中选择优选数据集的方法的例证性实施方式的流程图;

[0009] 图2A示出用于确定数据集是优选数据集的最大概率的表;

[0010] 图2B示出用于确定数据集是优选数据集的最小概率的表;

[0011] 图3示出为了在分布式数据库之中选择优选数据集而布置的计算机程序产品的例证性实施方式的框图;以及

[0012] 图4示出为了在均按照本公开的实施方式布置的分布式数据库之中选择优选数据集而布置的计算装置的例证性实施方式的框图。

具体实施方式

[0013] 在下面的具体描述中,参照形成其部分的附图。在附图中,类似符号通常标识类似组件,除非上下文另外指示。具体实施方式、附图和权利要求书中描述的例证性实施方式不旨在是限制性的。可利用其他实施方式,并且可在不脱离这里提出的主题的精神或范围的情况下进行其他改变。应该容易理解,本文中总体描述并且在附图中示出的本公开的方面可按各式各样不同构造进行布置、替代、组合和设计,所有这些构造是清楚预料的并且形成本公开的部分。

[0014] 本公开尤其涉及用于在分布数据库之中选择优选数据集的方法、设备、计算机程序和系统。

[0015] 在本公开中,术语“skyline计算”总体上是指应对多准则决策制定的问题的方法,其中没有优于不同数据集的属性的明确优选功能。替代地,此计算尝试基于一个或多个预设偏好来总体了解哪些对象支配其他对象。如果第一元组中的至少一个属性优于第二元组中的相应属性,同时第一元组中的属性的剩余部分被认为等于第二元组中的相应属性或者比第二元组中的相应属性更优选,则第一元组“支配”第二元组。术语“skyline概率”总体上指对象是通过skyline计算而得到的数据集的部分的概率。“存在概率属性”是关系中的属性。“存在概率”总体上是指与关系中的元组关联的概率。它指代是否存在元组的概率。

[0016] 在一些实施方式中,多准则决策分析可包括skyline计算。skyline计算具有广泛应用。例如,一个应用可涉及选择便宜且靠近海滩的酒店。在skyline计算之后,计算出的优选数据集中的酒店优于不在优选集中的酒店,因为它们在至少一种属性、成本或与海滩的接近度方面是优异的,同时所有其他属性方面是等同或更好的。另一个示例可包括选择总价格最低、相机的图像分辨率最高并且存储卡的容量最高的相机和存储卡的组合。当关

系中的一个元组的存在并不确定并且具有存在概率时,多准则决策分析会更复杂。当查询包括阈值概率时,存在概率会发挥作用。该元组是skyline的概率可以是存在概率乘以没有其他元组支配该元组的概率。在一些实施方式中该,如果元组具有小于阈值概率的skyline概率,则元组可不是skyline的一部分。如果元组具有大于或等于阈值概率的skyline概率,则元组可以是skyline的一部分。

[0017] 在一些实施方式中,数据集可以是关系数据库中的关系(或表)并且优选数据集可被计算为结构化查询语言(SQL)查询的结果集。在一些实施方式中,数据集可以是平面文件。在一些其他实施方式中,数据集可以是非关系数据库中的元组集。在其他实施方式中,数据集可以是半结构化数据存储中的元组集。今后讨论的示例实施方式涉及计算关系数据库系统中的优选数据集。然而,本领域的普通技术人员应该清楚,提出的技术可应用于除了存储在关系数据库系统中的数据集外的数据集。

[0018] 图1是按照本公开的一些实施方式的、用于在分布式数据库之中选择优选数据集的方法100的例证性实施方式的流程图。在一些实施方式中,分布式数据库可包括第一数据库和第二数据库。存储在第一数据库中的数据集和存储在第二数据库中的数据集可不保持在中央仓库。第一数据库和第二数据库可被配置为由被配置为接收查询的协调器访问。可响应于查询从存储在第一数据库和第二数据库中的数据集选择优选数据集。在一些实施方式中,查询可与skyline计算关联并且可包括阈值概率。

[0019] 在框101中,协调器可接收选自第一数据库的第一数据集和选自第二数据库的第二数据集。第一数据集可包括第一本地skyline概率属性和第一单调属性。第二数据集可包括第二本地skyline概率属性和第二单调属性。本地skyline概率总体上是对象是通过在单个数据库中执行的skyline计算而得到的数据集的部分的概率。如果数据集没有被同一数据库中的其他数据集支配,则数据集的本地skyline概率可以是该数据集的存在概率。数据集的本地skyline概率可以是该数据集的存在概率和同一数据库中支配该数据集的数据集的不存在概率的乘积。

[0020] 可使用单调函数来计算单调属性。一些示例单调属性包括但不限于曼哈顿距离(Manhattan distance)属性、欧氏距离(Euclidean distance)属性等。例如,在具有原点 $(0,0,\dots,0)$ 的坐标系中的 d 维点 $t=\langle t_1,t_2,\dots,t_d \rangle$ 的曼哈顿距离是 $t_1+t_2+\dots+t_d$ 。

[0021] 在一些实施方式中,第一数据库中的数据集和第二数据库中的数据集可被存储在技术可行索引结构中,例如, n 叉树索引结构中。 n 叉树是四叉树的 n 维泛化。基于 n 叉树结构,第一数据库可被配置为通过执行整个数据空间的网格划分(均匀或非均匀)来构造其自己的数据集的本地总结。当优选数据集有可能处于数据空间的特定区域时,获悉优选数据集的密集分布的非均匀网格划分可提供更高精度的总结。假设一组数据集位于网格的单元中,单元被配置为存储不存在位于单元中的一组数据集的概率。对于网格的空单元,空单元中的每个被配置为存储值1。第一数据库和第二数据库的本地总结可被发送到协调器。协调器可基于第一数据库的本地总结和第二数据库的本地总结来构造并且保持全局总结。在一些实施方式中,可在数据库的本地总结中的任一个改变时,更新全局总结。

[0022] 在一些实施方式中,同一数据库中的数据集可从原点起按其单调属性的升序来布置。如果第一数据库中的数据集没有被检验为优选数据集,则第一数据集可具有数据集之中的最小单调属性。如果第二数据库中的数据集没有被检验为优选数据集,则第二数据集

可具有这些数据集中的最小单调属性。

[0023] 在框103中,协调器可从第一数据集和第二数据集中的选择一个选择候选数据集。选择可以是基于第一单调属性和第二单调属性的比较。在一些实施方式中,候选数据集可具有比其他数据集的单调属性小的单调属性。

[0024] 在框105中,协调器可确定所选择的候选数据集是否是优选数据集。在一些实施方式中,确定可以是至少部分地基于所选择的候选数据集是优选数据集的最小概率、所选择的候选数据集是优选数据集的最大概率和阈值概率。

[0025] 在一些实施方式中,当所选择的候选数据集是优选数据集的最大概率小于阈值概率时,所选择的候选数据集可不是优选数据集。在一些实施方式中,当所选择的候选数据集是优选数据集的最小概率大于或等于阈值概率时,所选择的候选数据集可以是优选数据集。

[0026] 在一些实施方式中,如果从第一数据库选择优选数据集,则优选数据集可按反馈概率被发送到第二数据库。将优选数据集发送到除了优选数据集所驻留的节点外的节点可减少用于根据协调器接收的查询为了优选数据集而检验存储在第一数据库和第二数据库中的所有数据集的总处理时间。在一些实施方式中,当(1)所选择的候选数据集是优选数据集的最大概率大于或等于阈值概率并且(2)所选择的候选数据集是优选数据集的最小概率小于阈值概率时,如果所选择的候选的分值小于阈值,则所选择的候选数据集可被丢弃并且被确定不是优选数据集。所选择的候选数据集可不在第一数据库和第二数据库之间广播,以减小第一数据库和第二数据库之间的通信成本,但是代价是优选数据集的完整性。

[0027] 在一些其他实施方式中,当(1)所选择的候选数据集是优选数据集的最大概率大于或等于阈值概率并且(2)所选择的候选数据集是优选数据集的最小概率小于阈值概率时,如果所选择的候选数据集选自第一数据库,则所选择的候选数据集可被发送到第二数据库。在接收到所选择的候选数据集之后,第二数据库可被配置为计算第二数据库中不存在支配所选择的候选数据集的元组的支配概率。仅仅为了例证,如果在第二数据库中存在其存在概率分别是0.3和0.6的两个元组并且第一数据库中的两个元组都支配所选择的候选数据集,则在第二数据库中不存在支配所选择的候选数据集的元组的支配概率将是 $(1-0.3) \times (1-0.6) = 0.28$ 。

[0028] 遵循以上阐述的例证示例,第二数据库可将0.28的支配概率发送回协调器。协调器随后可被配置为计算所选择的候选数据集的skyline概率。所选择的候选数据集的skyline概率可以是0.28和(所选择的候选数据集没有被第一数据库中的其他数据集支配的概率)的乘积。如果所选择的候选数据集的skyline概率小于阈值概率,则所选择的候选数据集可不是优选数据集。如果所选择的候选数据集的skyline概率等于或大于阈值概率,则所选择的候选数据集可以是优选数据集。

[0029] 在一些实施方式中,协调器可被配置为接收选自第一数据库的第三数据集。第三数据集可对应于根据第一数据库中的单调属性的升序布置的第一数据库中的下一个数据集。在一些实施方式中,可在用第三数据集取代第三数据集的情况下重复方法100。

[0030] 图2A示出按照本公开的一些实施方式的用于确定数据集是优选数据集的最大概率的表200。仅仅出于例证,五个数据集(即, $t_{1,1}$ 、 $t_{1,2}$ 、 $t_{1,3}$ 、 $t_{2,1}$ 和 $t_{2,2}$)位于表200中。三个数据集(即, $t_{1,1}$ 、 $t_{1,2}$ 、 $t_{1,3}$)被存储在第一数据库中,另两个数据集(即, $t_{2,1}$ 和 $t_{2,2}$)被存储在第二

二数据库中。数据集中的任一个包括其自己的存在概率,例如, $t_{1,1}$ 、 $t_{1,2}$ 、 $t_{1,3}$ 、 $t_{2,1}$ 和 $t_{2,2}$ 的存在概率分别可以是0.2、0.2、0.7、0.8和0.9。如以上阐述的,在一些实施方式中,单元被配置为存储不存在位于该单元中的一组数据集的概率。因此,单元201、203、205、207和209可分别存储值 $(1-0.2)$ 、 $(1-0.2)$ 、 $(1-0.7)$ 、 $(1-0.8)$ 和 $(1-0.9)$ 。

[0031] 在均匀网格分隔(例如,表200)中,数据集的支配区域可对应于使得位于区域内的任何点都支配数据集的最大空间区域。在表200中,可使用均匀网格分隔将针对 $t_{2,2}$ 的支配区域近似为表200中的点线框所示的正交范围空间。在一些实施方式中,数据集是优选数据集的最大概率是位于针对该数据集的正交范围空间中的每个单元的值的连续乘积。例如,对于 $t_{2,2}$, $t_{2,2}$ 是优选数据集的最大概率是 $(1-0.2) \times (1-0.2) = 0.64$ 。

[0032] 图2B示出按照本公开的一些实施方式的用于确定数据集是优选数据集的最小概率的表210。仅仅出于例证,五个数据集(即, $t_{1,1}$ 、 $t_{1,2}$ 、 $t_{1,3}$ 、 $t_{2,1}$ 和 $t_{2,2}$)位于表210中。三个数据集(即, $t_{1,1}$ 、 $t_{1,2}$ 、 $t_{1,3}$)被存储在第一数据库中,另两个数据集(即, $t_{2,1}$ 和 $t_{2,2}$)被存储在第二数据库中。数据集中的任一个包括其自己的存在概率,例如, $t_{1,1}$ 、 $t_{1,2}$ 、 $t_{1,3}$ 、 $t_{2,1}$ 和 $t_{2,2}$ 的存在概率分别可以是0.2、0.2、0.7、0.8和0.9。如以上阐述的,在一些实施方式中,单元可被配置为存储不存在位于该单元中的一组数据集的概率。因此,单元211、213、215、217和219可分别存储 $(1-0.2)$ 、 $(1-0.2)$ 、 $(1-0.7)$ 、 $(1-0.8)$ 和 $(1-0.9)$ 。

[0033] 在均匀网格分隔(例如,表210)中,数据集的互补支配区域可对应于使得位于区域外的任何点都不可控制数据集的最小空间区域。在表210中,可使用均匀网格分隔将针对 $t_{2,2}$ 的互补支配区域近似为作为表210中的双点虚线区域的互补正交范围空间。在一些实施方式中,数据集是优选数据集的最小概率可以是位于针对该数据集的互补正交范围空间中的每个单元的值的连续乘积。例如,对于 $t_{2,2}$, $t_{2,2}$ 是优选数据集的最小概率是 $(1-0.2) \times (1-0.2) \times (1-0.9) = 0.064$ 。

[0034] 图3示出为了在分布式数据库之中选择优选数据集而布置的计算机程序产品的框图。计算机程序产品300可包括信号承载介质304,信号承载介质304可包括可执行指令302的一个或更多个集合,当例如计算装置的处理器的处理器执行可执行指令302时,可提供至少上述在图1中示出的功能。

[0035] 在一些实现方式中,信号承载介质304可涵盖非暂态计算机可读介质308,诸如但不限于硬盘驱动、紧凑盘(CD)、数字通用盘(DVD)、数字带、存储器等。在一些实现方式中,信号承载介质304可涵盖可记录介质310,诸如但不限于存储器、读/写(R/W) CD、R/W DVD等。在一些实现方式中,信号承载介质304可涵盖通信介质306,诸如但不限于数字和/或模拟通信介质(例如,光纤电缆、波导、有线通信链路、无线通信链路等)。计算机程序产品300还可被记录在非暂态计算机可读介质308或另一个类似的可记录介质310中。

[0036] 图4示出为了在分布式数据库之中选择优选数据集而布置的计算装置的例证性实施方式的框图。在非常基本的构造401中,计算装置400通常包括一个或更多个处理器410和系统存储器420。存储器总线430可用于在处理器410和系统存储器420之间通信。

[0037] 根据期望的构造,处理器410可以是任何类型,包括但不限于微处理器(μ P)、微控制器(μ C)、数字信号处理器(DSP)或其任何组合。处理器404可包括一级或多级缓存(诸如,一级缓存411和二级缓存412、处理器核413和寄存器414)。示例处理器核413可包括算术逻辑单元(ALU)、浮点单元(FPU)、数字信号处理核(DSP核)、或其任何组合。示例存储器控制器

415还可与处理器404一起使用,或者在一些实现方式中,存储器控制器418可以是处理器410的内部部分。

[0038] 根据期望的构造,系统存储器420可以是任何类型,包括但不限于易失性存储器(诸如,RAM)、非易失性存储器(诸如,ROM、闪存存储器等)、或其任何组合。系统存储器420可包括操作系统421、一个或更多个应用422和程序数据424。在一些实施方式中,应用422可包括数据选择算法423,数据选择算法423被布置成执行如本文中描述的功能,包括相对于图1的方法100的步骤101至105描述的功能。程序数据424可包括如以下进一步描述的可用于操作数据选择算法423的数据集425。在一些实施方式中,数据集425可包括而限于图1的框101的第一数据集和第二数据集。在一些实施方式中,应用422可被布置成用操作系统421上的程序数据424进行操作,使得可如本文中描述地提供实现在分布式数据库中选择优选数据集。在图4中通过内部虚线内的那些组件来示出描述的这个基本构造401。

[0039] 在一些其他实施方式中,应用422可包括数据选择算法423,数据选择算法423被布置成执行如本文中描述的功能,包括相对于图1的方法100的步骤101至105描述的功能。

[0040] 计算装置400可具有额外特征或功能和促成基本构造401和任何所需装置和接口之间通信的额外接口。例如,总线/接口控制器440可用于促成经由存储接口总线441进行的基本构造401和一个或更多个数据存储装置450之间通信。数据存储装置450可以是可移除存储装置451、不可移除存储装置452、或其组合。可移除存储装置和不可移除存储装置的示例包括诸如软盘驱动器和硬盘驱动器(HDD)的磁盘装置、诸如CD驱动器或DVD驱动器的光盘驱动器、固态驱动器(SSD)、和带驱动器,这只是几个例子。

[0041] 系统存储器420、可移除存储装置451和不可移除存储装置452是计算机存储介质的示例。示例计算机存储介质可包括用存储信息(诸如,计算机可读指令、数据结构、程序模块或其他数据)的任何方法或技术实现的易失性和非易失性、可移除和不可移除的介质。可使用任何这种计算机存储介质来存储期望的信息并且可由计算装置400来访问并且可以是计算装置400的部分。

[0042] 计算装置400还可包括接口总线442,接口总线442用于促成经由总线/接口控制器440的从各种接口装置(例如,输出装置460、外围接口470和通信装置480)到基本构造401的通信。示例输出装置460包括图形处理单元461和音频处理单元462,图形处理单元461和音频处理单元462可被配置为经由一个或更多个A/V端口463与诸如显示器或扬声器的各种外部装置进行通信。示例外围接口470包括串行接口控制器471或并行接口控制器472,串行接口控制器471或并行接口控制器472可被配置为经由一个或更多个I/O端口473与诸如输入装置(例如,键盘、鼠标、笔、语音输入装置、触摸输入装置等)或其它外围装置(例如,打印机、扫描仪等)的外部装置通信。示例通信装置4480包括网络控制器481,网络控制器481可被配置为经由一个或更多个通信端口482通过网络通信链路促成与一个或更多个其他计算装置490通信。在一些实施方式中,其他计算装置490可包括可基于应用422的结果进行操作的其他应用。

[0043] 网络通信链路可以是通信介质的一个示例。通信介质通常可用计算机可读指令、数据结构、程序模块或调制后的数据信号(诸如,载波或其它传输机制)中的其他数据来实施,并且可包括任何信息传递介质。“调制后的数据信号”可以是其特征中的一个或更多个以编码信号中的信息这样的方式被设置或改变的信号。举例来说,而并非限制地,通信介质

可包括诸如有线网络或直接连线连接的有线介质和诸如声学、射频 (RF)、微波、红外 (IR) 和其他无线介质的无线介质。本文中使用的术语“计算机可读介质”可既包括存储介质又包括通信介质。

[0044] 计算装置400可被实现为诸如手机、个人数字助理 (PDA)、个人媒体播放器装置、无线网络手表装置、个人头戴式装置、应用专用装置或包括以上功能中的任一种的混合装置的小形状因子便携式(或移动)电子装置的一部分。计算装置400还可被实现为既包括便携式计算机构造又包括非便携式计算机构造的个人计算机。

[0045] 系统的一些方面的硬件实现和软件实现之间的区别极小;硬件或软件的使用一般是(但不是一直是,在某些背景下,硬件和软件之间的选择可变得非常大)代表成本与效率的权衡的设计选择。可用各种媒介物(例如,硬件、软件和/或固件)实现本文中描述的处理和/或系统和/或其它技术,并且优选的媒介物将随着其中有效利用处理和/或系统和/或其它技术的背景而变化。例如,如果实现方确定速度和准确性是最重要的,则实现方可选择主要是硬件和/或固件媒介物;如果灵活性是最重要的,则实现方可选择主要是软件实现;或者,另外另选地,实现方可选择硬件、软件和/或固件的某种组合。

[0046] 以上的详细描述已经借助使用框图、流程图和/或示例阐述了装置和/或处理的各种实施方式。在这些框图、流程图和/或示例包含一个或更多个功能和/或操作的范围内,本领域内的技术人员将理解,可通过大范围的硬件、软件、固件或实际上其任何组合,单独地和/或一齐地实现这些框图、流程图或示例内的各功能和/或操作。在一个实施方式中,可借助专用集成电路 (ASIC)、现场可编程门阵列 (FPGA)、数字信号处理器 (DSP) 或其它集成格式来实现本文中描述的主题的一些部分。然而,本领域的技术人员将认识到,本文中公开的实施方式的一些方面全部地或部分地可等同地在集成电路中被实现为在一个或更多个计算机上运行的一个或更多个计算机程序(例如,在一个或更多个计算机系统上运行的一个或更多个程序)、在一个或更多个处理器上运行的一个或更多个程序(例如,在一个或更多个微处理器上运行的一个或更多个程序)、固件、或实际上其任何组合,并且本领域的技术人员依照本公开将适宜地针对软件或固件来设计电路和/或写入代码。另外,本领域的技术人员将理解,本文中描述的主题的机制能够被分配为各种形式的程序产品,并且适用本文中描述的主题的例证实施方式,而不顾及用于实际执行分配的信号承载介质的特定类型。

[0047] 本领域的技术人员将认识到,一般在本领域内以本文中阐述的方式描述装置和/或过程,此后使用工程实践将描述的这些装置和/或过程集成为数据处理系统。也就是说,本文中描述的装置和/或过程的至少一部分可借助合理量的实验被集成在数据处理系统中。也就是说,可借助合理量的实验将本文中描述的装置和/或处理的至少一部分集成在数据处理系统中。本领域的技术人员将认识到,典型的数据处理系统一般包括系统单元外壳、视频显示装置、诸如易失性和非易失性存储器的存储器、诸如微处理器和数字信号处理器的处理器、诸如操作系统、驱动器、图形用户界面和应用程序的计算实体、诸如触摸板或触摸屏的一个或多个交互装置、和/或包括反馈回路和控制电机(例如,感测位置和/或速率的反馈;移动和/或调节组件和/或数量的控制电机)的控制系统中的一个或更多个。可利用任何合适的商购组件(诸如,通常见于数据计算/通信和/或网络计算/通信系统中的那些)来实现典型的数据处理系统。

[0048] 本文中描述的主题有时示出包含在其他不同组件内或与其他不同组件连接的不

同组件。要理解,描绘的这些构造仅仅是示例性的,并且事实上,可实现用于实现相同功能的许多其它构造。在感官认识上,用于实现相同功能的组件的任何布置被有效“关联”,使得所需功能得以实现。因此,本文中被组合用于实现特定功能的任何两个组件可被视为彼此“关联”,使得所需功能得以实现,而不顾及构造或中间组件。同样地,如此关联的任何两个组件还可被视为彼此“可操作地连接”或“可操作地联接”以实现所需功能,并且能够如此关联的任何两个组件还可被视为彼此“可操作地联接”以实现所需功能。可操作联接的特定示例包括但不限于物理可配合和/或物理交互的组件和/或无线可交互和/或无线交互的组件和/或逻辑上交互和/或逻辑上可交互的组件。

[0049] 相对于本文中基本上任何多个和/或单个术语,本领域的技术人员可从复数转化成单数和/或从单数转换成复数,使之适于背景和/或应用。为了清晰的缘故,在本文中明确地阐述各种单数/复数排列。

[0050] 本领域的技术人员应该理解,一般,本文中使用的,尤其在随附权利要求书(例如,随附权利要求书的主体)中使用的术语一般旨在是“开放性”术语(例如,术语“包括”应该被解释为“包括但不限于”,术语“具有”应该被解释为“具有至少”,术语“包括”应该被解释为“包括但不限于”,等等)。本领域的技术人员还将理解,如果旨在有特定数量的引入的权利要求陈述,则在权利要求中将明确地陈述此意图,并且在缺乏此陈述时,不存在此意图。例如,为了辅助理解,下面随附的权利要求书可包含使用引入短语“至少一个”和“一个或更多个”来引入权利要求陈述。然而,这种短语的使用不应该被理解成意味着,用不定冠词引入权利要求陈述将包含引入的此权利要求陈述的任何特定权利要求限于只包含此陈述的实施方式,即使当同一权利要求包括引入性短语“一个或更多个”或“至少一个”和诸如“一”或“一个”的不定冠词(例如,“一”和/或“一个”应该被解释为意味着“至少一个”或“一个或更多个”)时;对于使用用于引入权利要求陈述的定冠词,同样如此。另外,即使明确陈述了特定数量的引入的权利要求陈述,本领域的技术人员将认识到,这种陈述应该被解释为意味着至少所陈述的数量(例如,不带其它修辞的“两个陈述物”的单纯陈述意味着至少两个陈述物、或两个或更多个陈述物)。例如,在使用与“A、B和C等中的至少一个”类似的约定的那些情形下,一般就意义而言旨在是使本领域的技术人员将理解惯例的这种构造(例如,“具有A、B和C中的至少一个的系统”将包括但不限于只具有A、只具有B、只具有C、同时具有A和B、同时具有A和C、同时具有B和C、和/或同时具有A、B和C等的系统)。在使用与“A、B或C等中的至少一个”类似的惯例的这些情形下,一般就意义而言旨在是使本领域的技术人员将理解惯例的这种构造(例如,“具有A、B和C中的至少一个的系统”将包括(但不限于)只具有A、只具有B、只具有C、同时具有A和B、同时具有A和C、同时具有B和C、和/或同时具有A、B和C等的系统)。本领域的技术人员将进一步理解,表现两个或更多个替代术语的几乎任何连词和/或短语(无论是在描述中、权利要求书中还是附图中都)应该被理解为料想到包括术语中的一个、术语中的任一个、或这两个术语的可能性。例如,短语“A或B”将被理解为包括“A”或“B”或“A和B”的可能性。

[0051] 虽然在本文中已经公开了各种方面和实施方式,但本领域的技术人员将清楚其他方面和实施方式。本文中公开的各种方面和实施方式是出于例证的目的,不旨在是限制,其真实范围和精神由所附的权利要求书指示。

100

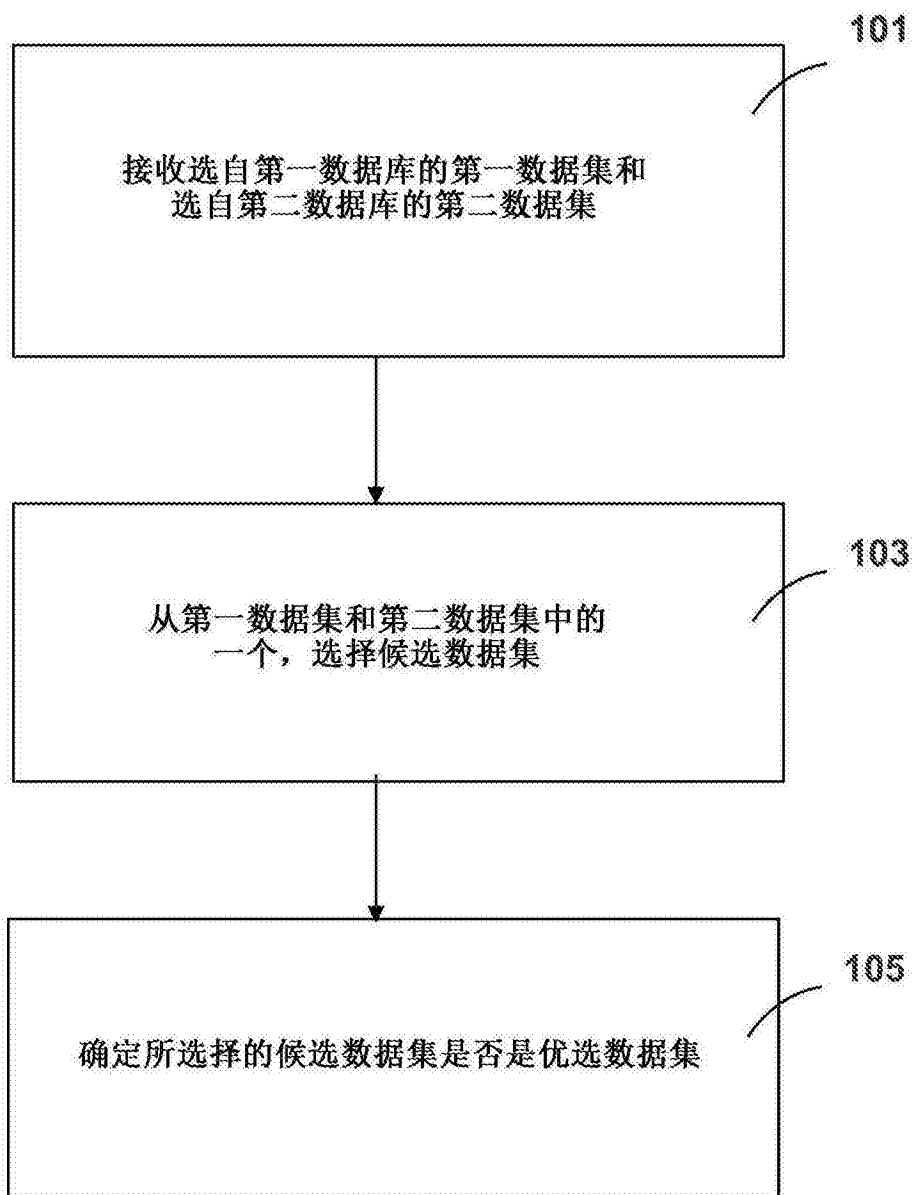


图1

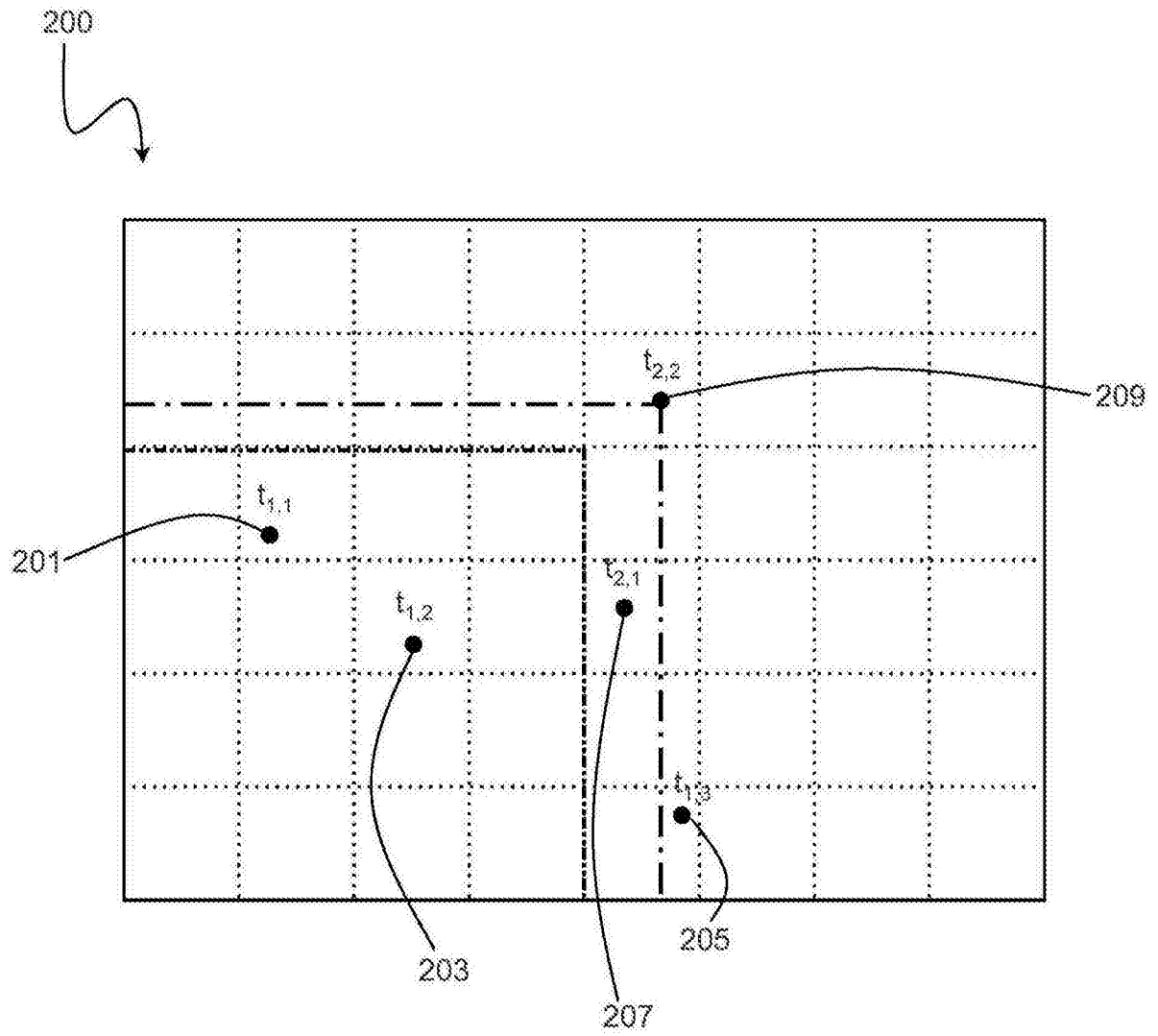


图2A

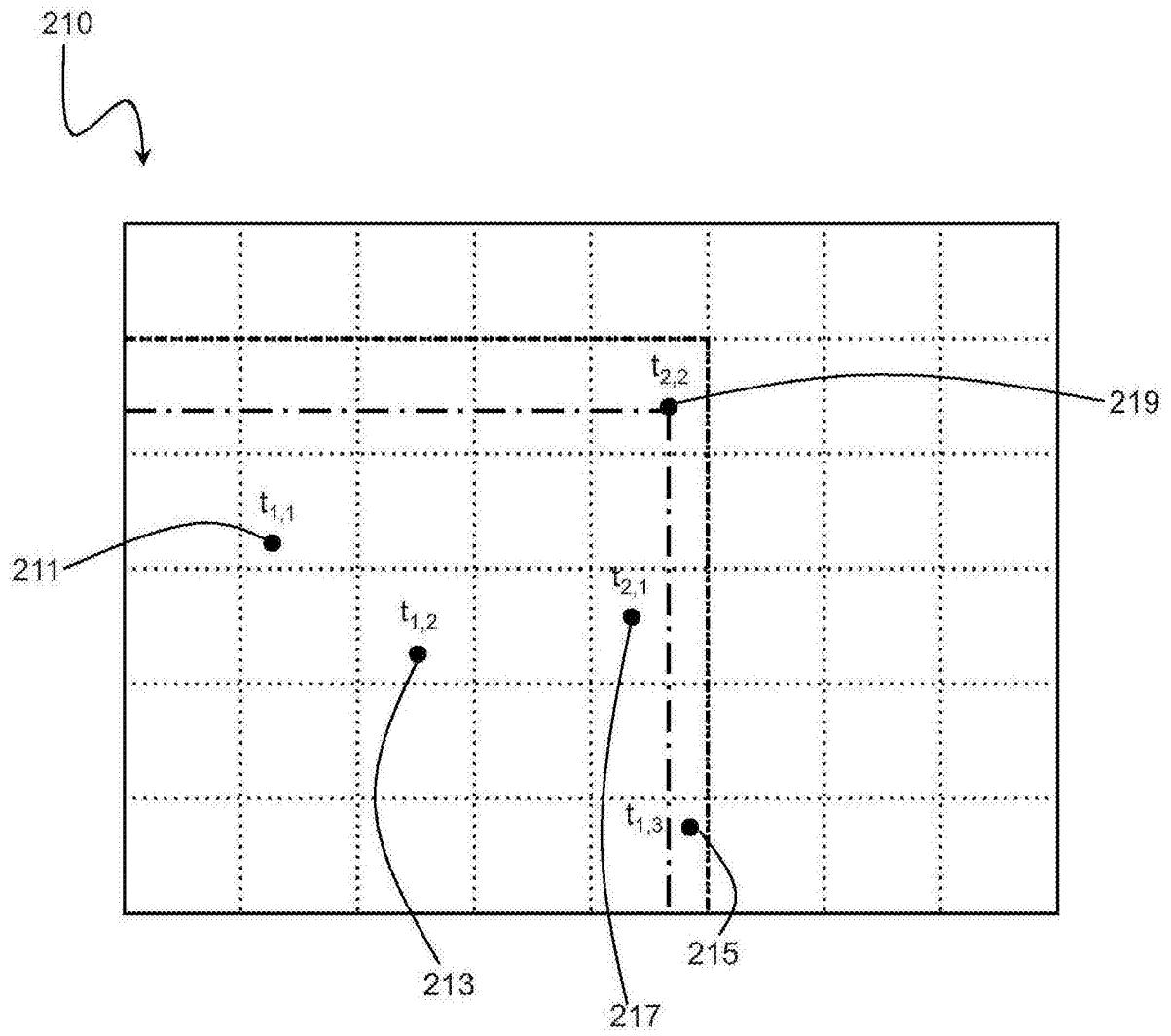


图2B

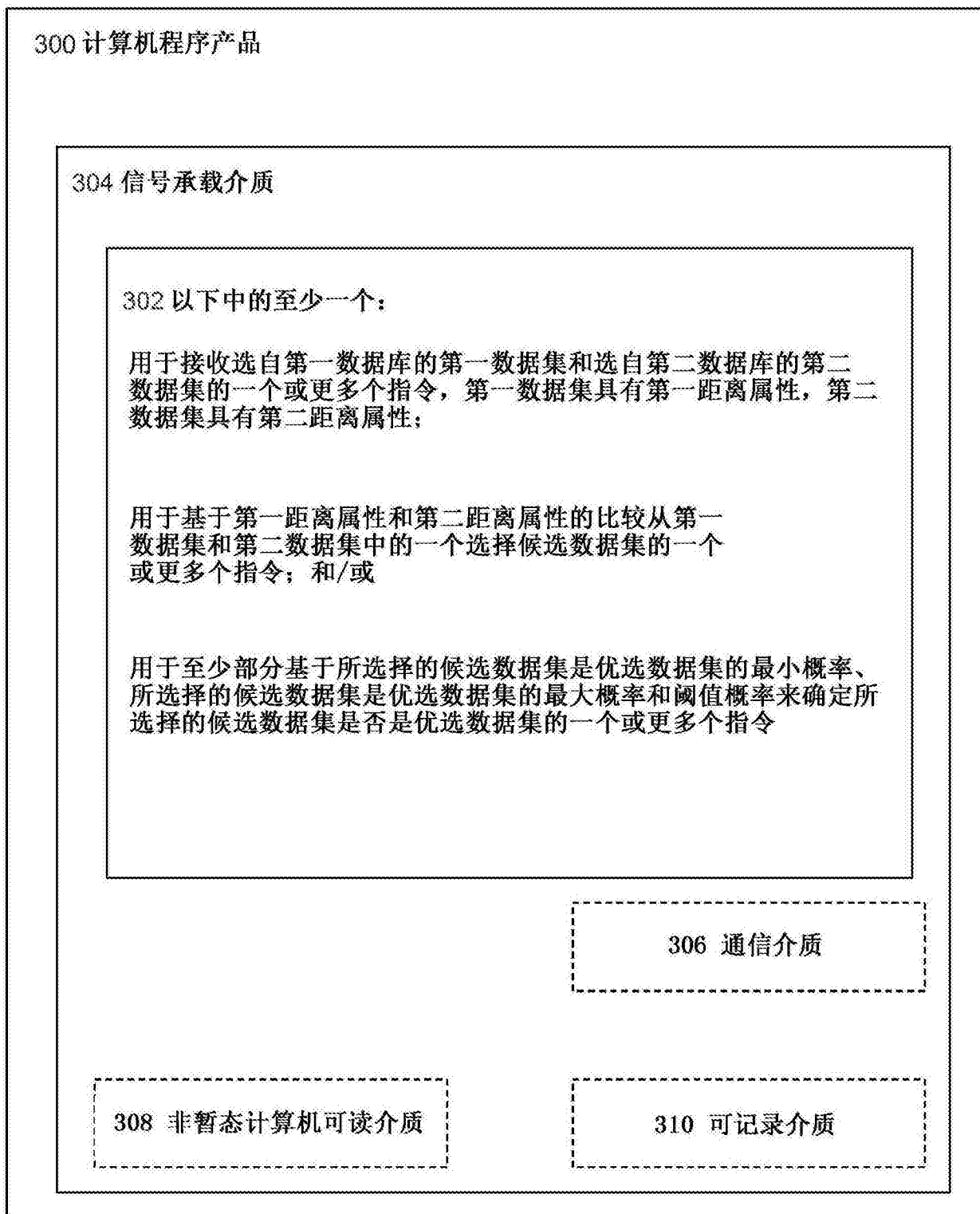


图3

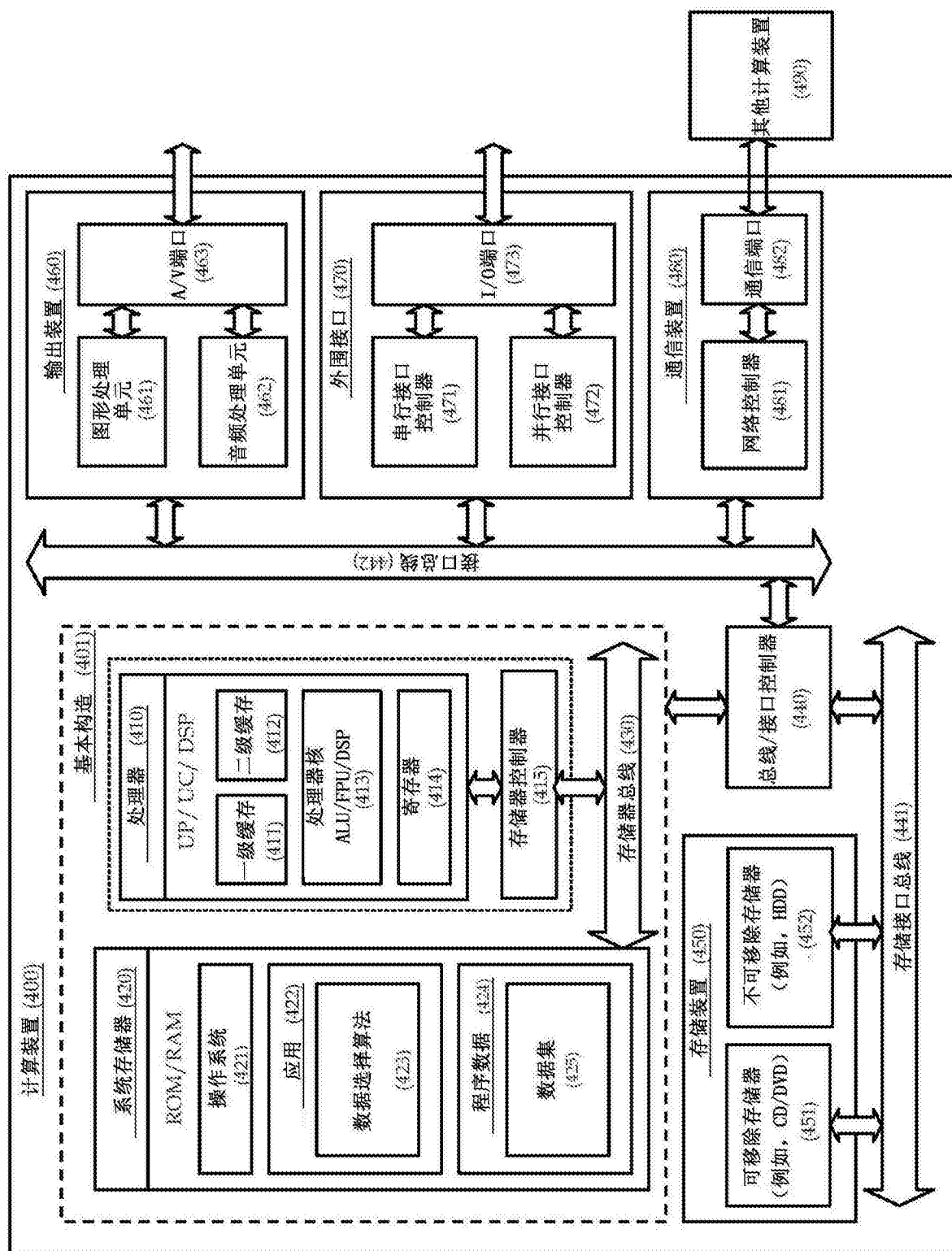


图4