(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118511490 A
(43)申请公布日 2024.08.16

(21)申请号 202380015980.X

(22)申请日 2023.03.07

(30)优先权数据
63/318,170 2022.03.09 US

(85)PCT国际申请进入国家阶段日
2024.07.02

(86)PCT国际申请的申请数据
PCT/US2023/014691 2023.03.07

(87)PCT国际申请的公布数据
WO2023/172541 EN 2023.09.14

(71)申请人 西斯坦股份有限公司
地址 美国纽约州

(72)发明人 A·布莱 D·康

(74)专利代理机构 广州川墨知识产权代理事务所(普通合伙) 44485
专利代理师 陈梓斌 曾灿灿

(51)Int.Cl.
H04L 43/08(2006.01)
H04L 41/06(2006.01)
H04L 41/0681(2006.01)
H04L 41/14(2006.01)
H04L 67/14(2006.01)

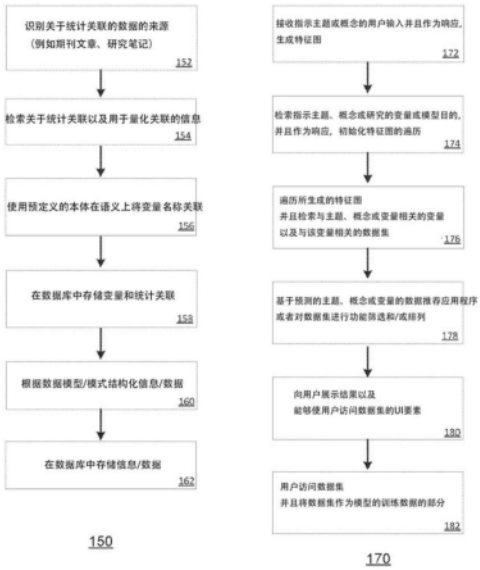
权利要求书3页 说明书38页 附图15页

(54)发明名称

监控相关指标的系统和方法

(57)摘要

一种系统和方法,其用于提高企业或其他实体监控业务相关指标(如KPI)的能力,以及评估用于生成这些指标的基础数据的质量。



CN 118511490 A

1.一种监控一个或多个指标的方法,其包括:

构建或访问特征图,所述特征图包括成组的节点和成组的边,其中成组的边中的每条边将成组的节点中的节点与一个或多个其他节点连接,并且其中每个节点代表被发现与主题有统计关联的变量,每条边代表节点与主题之间或者第一节点与第二节点之间的统计关联;

生成用户界面显示和用户界面工具,使用户能够执行以下一项或多项操作:

识别监控的指标;

定义规则,所述规则规定何时应生成有关已识别的指标的行为的警报;

定义如何在用户界面显示上指示应用所述规则的结果;

允许用户选择已生成警报的指标,并作为响应,提供有关指标的值随时间的一个或多个变化的信息、导致警报的规则、所述指标与其他指标的关系,以及

关于用于生成指标的数据集、机器学习模型、规则或其他因素的信息。

2.根据权利要求1所述的方法,其还包括为用户提供关于以下内容的一个或多个的推荐:待监控的不同的指标或成组的指标、可能有助于检查的数据集、可能与指标相关的元数据,或者基础数据或指标的一方面。

3.根据权利要求1所述的方法,其中构建特征图还包括:

访问一个或多个来源,其中每个来源包括关于来源中讨论的主题与讨论的主题中考虑的一个或多个变量之间的统计关联的信息;

处理从每个来源访问的信息,以识别所考虑的一个或多个变量,并针对每个变量,识别有关变量与主题之间的统计关联的信息;以及

将处理所访问的一个或多个来源的结果存储在数据库中,所存储的结果包括针对于每个来源,对一个或多个变量的每个变量的引用、对主题的引用以及关于每个变量与主题之间的统计关联的信息。

4.根据权利要求3所述的方法,其还包括存储能够访问数据集的要素,其中所述数据集包括用于证明每个变量与主题之间的统计关联的数据或者代表一个或多个变量的衡量的数据。

5.根据权利要求4所述的方法,其还包括:

遍历特征图,以识别与一个或多个变量相关的一个或多个数据集,所述一个或多个变量在统计上与用户感兴趣的主题相关,或在统计上与感兴趣的主题在语义上相关;

对识别的一个或多个数据集进行筛选和排序;并且

向用户展示已识别的数据集的筛选和排序结果。

6.根据权利要求3所述的方法,其中一个或多个来源包括至少一个包含专有数据的来源。

7.根据权利要求6所述的方法,其中专有数据来自业务、研究或实验。

8.根据权利要求1所述的方法,其中推荐由经过训练的模型或统计分析中的一个或多个生成。

9.一种系统,其包括:

一个或多个电子处理器,其配置为执行成组的计算机可执行的指令;以及

包含所述成组的计算机可执行的指令的一个或多个非暂存计算机可读介质,其中,当

所述指令被执行时,所述指令使一个或多个电子处理器或包括处理器的设备或装置执行:

构建或访问特征图,所述特征图包括成组的节点和成组的边,其中成组的边中的每条边将成组的节点中的节点与一个或多个其他节点连接,并且其中每个节点代表被发现与主题有统计关联的变量,每条边代表节点与主题之间或者第一节点与第二节点之间的统计关联;

生成用户界面显示和用户界面工具,使用户能够执行以下一项或多项操作:

识别监控的指标;

定义规则,所述规则规定何时应生成有关已识别的指标的行为的警报;

定义如何在用户界面显示上指示应用所述规则的结果;

允许用户选择已生成警报的指标,并作为响应,提供有关指标的值随时间的一个或多个变化的信息、导致警报的规则、所述指标与其他指标的关系,以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的信息。

10.根据权利要求9所述的系统,其中所述指令使一个或多个电子的处理器或包括所述处理器的设备或装置为用户提供关于以下一个或多个内容的一项或多项的推荐:待监控的不同的指标或成组的指标、可能有助于检查的数据集、可能与指标相关的元数据,或者基础数据或指标的一方面。

11.根据权利要求9所述的系统,其中构建特征图还包括:

访问一个或多个来源,其中每个来源包括关于来源中讨论的主题与讨论的主题中考虑的一个或多个变量之间的统计关联的信息;

处理从每个来源访问的信息,以识别所考虑的一个或多个变量,并针对每个变量,识别有关变量与主题之间的统计关联的信息;以及

将处理所访问的一个或多个来源的结果存储在数据库中,所存储的结果包括针对于每个来源,对一个或多个变量的每个变量的引用、对主题的引用以及关于每个变量与主题之间的统计关联的信息。

12.根据权利要求11所述的系统,其还包括存储能够访问数据集的要素,其中所述数据集包括用于证明每个变量与主题之间的统计关联的数据或者代表一个或多个变量的衡量的数据。

13.根据权利要求12所述的系统,其中所述指令使一个或多个电子处理器或包括处理器的装置或设备执行:

遍历特征图,以识别与一个或多个变量相关的一个或多个数据集,所述一个或多个变量在统计上与用户感兴趣的主题相关,或在统计上与感兴趣的主题在语义上相关;

对识别的一个或多个数据集进行筛选和排序;并且

向用户展示已识别的数据集的筛选和排序结果。

14.根据权利要求11所述的系统,其中一个或多个来源包括至少一个包含专有数据的来源,并且其中专有数据是从业务、研究或实验中获得的。

15.一个或多个非暂存计算机可读介质,其包含成组的计算机可执行的指令,当所述指令通过一个或多个编程的电子处理器执行时,所述指令使处理器或包括所述处理器的设备或装置执行:

构建或访问特征图,所述特征图包括成组的节点和成组的边,其中成组的边中的每条

边将成组的节点中的节点与一个或多个其他节点连接,并且其中每个节点代表被发现与主题有统计关联的变量,每条边代表节点与主题之间或者第一节点与第二节点之间的统计关联;并且

生成用户界面显示和用户界面工具,使用户能够执行以下一项或多项操作:

识别监控的指标;

定义规则,所述规则规定何时应生成有关已识别的指标的行为的警报;

定义如何在用户界面显示上指示应用所述规则的结果;

允许用户选择已生成警报的指标,并作为响应,提供有关指标的值随时间的一个或多个变化的信息、导致警报的规则、所述指标与其他指标的关系,以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的信息。

16.根据权利要求15所述的非暂存计算机可读介质,其中所述指令使一个或多个电子处理器或包括所述处理器的设备或装置为用户提供关于以下一个或多个内容的一项或多项的推荐:待监控的不同的指标或成组的指标、可能有助于检查的数据集、可能与指标相关的元数据,或者基础数据或指标的一方面。

17.根据权利要求15所述的非暂存计算机可读介质,其中构建特征图还包括:

访问一个或多个来源,其中每个来源包括关于来源中讨论的主题与讨论的主题中考虑的一个或多个变量之间的统计关联的信息;

处理从每个来源访问的信息,以识别所考虑的一个或多个变量,并针对每个变量,识别有关变量与主题之间的统计关联的信息;以及

将处理所访问的一个或多个来源的结果存储在数据库中,所存储的结果包括针对于每个来源,对一个或多个变量的每个变量的引用、对主题的引用以及关于每个变量与主题之间的统计关联的信息。

18.根据权利要求17所述的非暂存计算机可读介质,其还包括存储能够访问数据集的要素,其中所述数据集包括用于证明每个变量与主题之间的统计关联的数据或者代表一个或多个变量的衡量的数据。

19.根据权利要求18所述的非暂存计算机可读介质,其中所述指令使一个或多个电子处理器或包括所述处理器的装置或设备执行:

遍历特征图,以识别与一个或多个变量相关的一个或多个数据集,所述一个或多个变量在统计上与用户感兴趣的主题相关,或在统计上与感兴趣的主题在语义上相关;

对识别的一个或多个数据集进行筛选和排序;并且

向用户展示已识别的数据集的筛选和排序结果。

20.根据权利要求17所述的非暂存计算机可读介质,其中一个或多个来源包括至少一个包含专有数据的来源,并且其中专有数据是从业务、研究或实验中获得的。

监控相关指标的系统和方法

相关专利申请的交叉引用

[0001] 本申请要求2022年3月9日提交的申请号为63/318,170的美国临时申请的权益,该申请的名称为“System and Methods for Monitoring Related Metrics”,其内容全文通过引用并入本申请。

[0002] 请注意,本文在架构或系统架构或平台的语境下提及的“系统”是指2019年5月23日提交的名称为“Systems and Methods for Organizing and Finding Data”的美国专利申请16/421,249(现已于2022年6月7日公开为美国专利11,354,587)中描述的用于执行统计搜索和其他形式的数据组织的架构、平台和流程,该美国专利申请要求2019年2月1日提交的名称为“Systems and Methods for Organizing and Finding Data”的美国临时专利申请62/799,981的优先权,该美国专利申请的全部内容以引用的方式并入本申请。

背景技术

[0003] 数据驱动型组织追踪关键绩效指标(简称KPI)和其他指标,从而衡量组织状况并协助做出战略决策。KPI和衡量标准也越来越多地成为新闻报道的一部分(例如,道琼斯工业平均指数、标准普尔500指数、主要公司股价的水平和百分比变化,或每周新申请失业保险人数的水平和变化)。当前监控此类指标的方法依赖于指示板(dashboard)、数据目录和KPI追踪器,以便为用户提供有关特定KPI的信息。

[0004] 传统方法虽然有用,但也有其局限性和缺点。首先,传统方法提供的KPI信息与其他因素相对孤立。此外,传统方法无法在现代数据科学和分析团队所做的建模和统计关联工作中对关键指标进行追踪和监控。这就限制了用户对关键绩效指标的重要变化的了解,以及对这些变化与其他指标之间的关系或对其他指标的影响的了解。这使得用户无法更全面、更准确地了解各种指标、用于生成指标的数据以及生成基础数据的公司(或其他实体)的绩效之间的关系。

[0005] 开发工具需要专门的资源,而许多企业可能无法随时获得或负担得起这些资源,该开发工具用于评估数据集内部和数据集之间的统计关系,并自动进行生成基于这些数据集的指标和决策。本文所描述的系统和方法的实施例旨在单独或共同解决这些问题和相关问题。

发明内容

[0006] 本文中使用的术语“发明”、“本发明”、“此发明”、“本公开”或“此公开”意指文本、附图或图表以及权利要求书中公开的所有主题。包含这些术语的陈述并不限制所公开的主题或权利要求的含义或范围。本公开所涉及的实施例由权利要求书而不是本发明内容来限定。本发明内容是对本公开内容各个方面的高度概括,并介绍了下文详细说明部分将进一步阐述的一些概念。本发明内容无意确定所要求主题的关键、基本或必要特征,也无意单独用于确定所要求主题的范围。应参考整个说明书的适当部分、任何或所有图表或附图以及每项权利要求来理解本主题。

[0007] 本公开的实施方案涉及一种系统和方法,用于提高企业或其他实体监控业务相关指标(如KPI)的能力,以及评估用于生成这些指标的基础数据的质量。在一些实施例中,所公开的系统和方法可包括要素、组件或流程,这些要素、组件、功能、操作或流程的配置和运行可提供以下一个或多个功能:

●创建由节点和边组成的特征图,其中;

○节点代表一个或多个概念、主题、数据集、元数据、模型、指标、变量、可衡量的量、对象、特性、特征或因素(作为非限制性示例);

■在一些实施例中,基于发现(或获取利用)的数据集、元数据、模型,从已训练的模型生成输出、生成有关数据集的元数据或开发本体或其他形式的层次关系(作为非限制性示例),从而创建节点;

○边代表第一节点和第二节点之间的关系,例如统计意义上的关系、依赖关系或层级关系(作为非限制性示例);

■在一些实施例中,可以创建连接第一节点和第二节点的边,以表示由统计分析、机器学习模型或研究确定的两个节点之间的有效统计关系;

○与边相关联的标签可指示边所连接的两个节点之间关系的方面,如两个节点之间的关系所基于的元数据,或支持两个节点之间具有统计意义的关系的数据集(作为非限制性示例);

●为用户提供用户界面显示画面、工具、特征和可选要素,使用户能够执行以下一项或多项功能:

○确定监控或追踪的感兴趣的指标(如KPI);

■其中,相关指标可由经过训练的模型、公式、等式或规则集生成,并可进一步基于、生成自或衍生自基础数据,基础数据是时间的函数;

○定义规则,该规则规定何时应生成有关已识别的指标的行为的警报;

■这样的规则可以基于绝对值、值的变化、百分比变化、一段时间内的百分比变化或者超过或低于阈值(作为非限制性示例);

○定义如何在用户界面显示上识别或指示应用规则的结果;

■例如,这可能取决于用户的偏好和/或指标变化的值或类型;

○允许用户选择已生成警报的指标,并作为响应,提供有关指标值随时间变化的信息、已满足或已激活的导致警报的规则、指标与其他指标的关系(如果相关),以及关于用于生成指标的数据集、机器学习模型、规则、公式或其他因素的可用信息(作为非限制性示例);

●为用户提供关于以下内容中的推荐:可能具有监控价值的不同的指标或指标集、可能有助于检查的数据集、可能与确定的指标相关的元数据或用户可能感兴趣的基础数据或指标的其它方面;

○其中,推荐可能(至少部分)来自经过训练的机器学习模型所生成的输出、统计分析、研究、或其他形式的数据收集或评估。

[0008] 在一个实施例中,本公开内容涉及一种系统,用于提高企业或其他实体监控业务相关指标(如KPI)的能力,以及评估基础数据的质量(进而评估准确性和可靠性)。该系统可包括存储在一个或多个非暂存计算机可读介质(或其上)的成组的计算机可执行指令,以及

一个或多个电子处理器。当指令被处理器或协同处理器执行时,会导致处理器或协同处理器(或作为其一部分的设备或装置)执行一系列操作,以实现所公开的方法的一个或多个实施例。

[0009] 在一个实施例中,本公开涉及一个或多个非暂存计算机可读介质,其中包括成组计算机可执行指令,当指令被一个或多个电子处理器或协同处理器执行时,处理器或协同处理器(或作为其一部分的设备或装置)执行成组操作,这些操作实现本公开方法的实施例。

[0010] 在一些实施例中,本文所述的系统和方法可通过SaaS或多租户平台提供服务。该平台可访问多个实体,每个实体都有独立账户和相关数据存储。每个账户可对应用户、成组用户、提供数据集用于评估和生成业务相关指标的实体或组织等。每个账户可访问一个或多个服务,其中成组服务在其账户中实例化,并实现本文所述的一个或多个方法或功能。

[0011] 本领域一般技术人员在阅读详细说明和附图后,就会明白所述系统和方法的其他目的和优点。在全部附图中,相同的附图标记和描述表示相似但不一定相同的元件。虽然本文所描述的示例性实施方案可以有各种修改和替代形式,但具体的实施方案已在附图中举例说明,并将在本文中详细描述。然而,本文所描述的示例性或具体的实施例并不限于所描述的形式。相反,本公开涵盖了所附权利要求范围内的所有修改、等同物和替代品。

附图说明

[0012] 本发明的实施例将参照附图进行描述:

[0013] 图1(a)是框图,其展示了一组要素、组件、功能、流程或操作,它们可以是平台架构100的一部分,在平台架构100中可以实施所公开的指标监控系统和方法的实施例;

[0014] 图1(b)是流程图表或流程图,其示出了使用本文公开的系统和方法的实施例构建特征图150的流程、方法、功能或操作;

[0015] 图1(c)是流程图表或流程图,其示出了示例用例的流程、方法、功能或操作,在该用例中,特征图被遍历以识别潜在的相关数据集,该流程、方法、功能或操作可以在本文所公开的系统和方法的实施例中实现;

[0016] 图1(d)是特征图数据结构的部分的示例的示意图,该特征图数据结构可用于组织和访问数据和信息,并可使用本文所公开的系统和方法的实施例;

[0017] 图2(a)是框图,其展示了一组要素、组件、功能、流程或操作,它们可以是平台架构的一部分,在平台架构中可以实施所公开的指标监控系统和方法的实施例。具体来说,图2(a)描述了如何实施使用所公开的指标监控功能,从而监控存储在云数据库服务中的数据集的特征变化;

[0018] 图2(b)是流程图表或流程图,其展示了一组要素、组件、功能、流程或操作,它们可以是平台架构的一部分,在平台架构100中可以实施所公开的指标监控系统和方法的实施例。具体来说,图2(b)描述了图2(a)中的某些步骤,更侧重于不同用户的交互以及软件要素,该软件要素有助于实现指标监控功能并且供用户使用;

[0019] 图2(c)是用户界面显示的示例,其中显示了最近的数值、该数值的百分比变化以及变化最大的子群的识别(当指标被创造为表格中数值的集合,而该表格中的数据有多个子群/维度时,可以计算该最大变化);

[0020] 图2(d)是用户界面显示的示例,显示了每周活跃用户页面上的指标监控面板,每周活跃用户是指标。在左侧的平台特征图中,其他指标的指标监控已打开,图中节点之间的边包含描述指标之间统计关系的元数据;

[0021] 图2(e)是用户界面显示的一个示例,展示了指标监控的平台目录视图,其中该页面上的8个指标均已打开;

[0022] 图2(f)是用户界面显示的示例,说明了指标监控功能的一个或多个通知;

[0023] 图2(g)是用户界面显示示例,说明了简化的规则设置对话框。适用于这一指标的条件是,当变化百分比的绝对值严格大于4.5;

[0024] 图2(h)是示出要素、组件或流程的示意图,其说明了根据某些实施例,要素、组件或流程配置为可能存在于由实现方法、流程、功能或操作的计算设备、服务器、平台或系统中的一个或多个中,或者由这些计算设备、服务器、平台或系统中的一个或多个来执行;以及

[0025] 图3-5是示出多租户或SaaS平台的架构的示意图,该平台可用于实施本文描述的系统和方法的实施例。

[0026] 请注意,在整个公开和附图中,相同的数字用于指代类似的组件和功能。

具体实施方式

[0027] 本文对本公开的实施例的主题进行了具体描述,以满足法定要求,但这种描述无意限制权利要求的范围。要求保护的主题可以以其他方式体现,可以包括不同的元件或步骤,也可以与其他现有的或后来开发的技术结合使用。本说明不应被解释为暗示各步骤或元件之间有任何规定的顺序或布置,除非明确指出各步骤或元件的顺序或布置是必需的。

[0028] 本文将参照附图对本公开的实施例进行更全面的描述,附图是本公开的一部分,通过图示的方式展示了本公开可以通过其实施的示例性实施例。然而,本公开内容可以以不同的形式体现,不应被理解为仅限于本文所阐述的实施例;相反,提供这些实施例是为了使本公开内容满足法定要求,并向本领域技术人员传达本公开内容的范围。

[0029] 除此以外,本公开内容可全部或部分体现为系统、一种或多种方法或一种或多种设备。本公开的实施例可以采取硬件实施例、软件实施例或软件与硬件相结合的实施例的形式。例如,在一些实施例中,本文所述的一个或多个操作、功能、流程或方法可由一个或多个合适的处理元件(如处理器、微处理器、CPU、GPU、TPU或控制器,作为非限制性示例)实现,这些处理元件是客户端设备、服务器、网络元件、远程平台(如SaaS平台)、“云”服务或其他形式的计算或数据处理系统、设备或平台的一部分。

[0030] 一个或多个处理元件可使用一组可执行的指令(如软件指令)进行编程,其中指令可存储在一个或多个合适的非暂存的计算机可读的数据存储介质或元件上(或其中)。在一些实施例中,成组指令(如通过网络,如互联网)可通过指令传输传达给用户,或者通过执行成组指令的应用程序传达给用户。在一些实施例中,最终用户可通过访问SaaS平台或通过此类平台提供的服务来使用成组指令或应用程序。

[0031] 在一些实施例中,本文所述的一个或多个操作、功能、流程或方法可由专门形式的硬件实现,如可编程门阵列、专用集成电路(ASIC)或类似硬件。需要注意的是,本公开的实施例可以以应用程序、作为大型应用程序一部分的子程序、“插件”、数据处理系统或平台的

功能扩展或其他合适的形式来实现。因此,以下的详细描述并不具有限制性。

[0032] 如前所述,在一些实施例中,本文所述的系统和方法可通过SaaS或多租户平台提供服务。该平台可访问多个实体,每个实体都有独立账户和相关数据存储。例如,每个账户可以对应于用户、成组的用户、实体或组织。每个账户可访问一个或多个服务,其中成组服务在其账户中实例化,并实现本文所述的一个或多个方法或功能。

[0033] 本公开的实施方案涉及一种系统和方法,其用于提高企业或其他实体监控业务相关指标(如KPI)的能力,以及评估用于生成这些指标的基础数据的质量。

[0034] 作为一般原则,用于决策的数据期望与正在执行的任务或正在做出的决策相关(或在一些情况下"充分"相关)。要做出可靠的数据驱动型决策或预测,不仅需要有关于预期的决策结果或预测目标的数据,还需要有关于与该结果或目标在统计上相关的变量(优选是所有变量,但至少是最相关的变量)的数据。遗憾的是,使用传统方法很难发现哪些变量已被证明与结果或目标有统计学关联,也很难获取有关这些变量的数据,从而难以更好地评估基于这些变量所做决策的可靠性。

[0035] 在许多情况下,通过以特定格式或结构表示数据,可以提高发现和访问数据的效率。格式或结构可包括数据记录中一个或多个列、行或字段的标签。识别和发现感兴趣数据的传统方法通常基于将单词与数据集中(或指代数据集,或关于数据集)的标签语义匹配。虽然这种方法有助于发现和获取可能与主题(例如目标或结果)相关的数据,但它并不能解决发现和获取有关于导致、影响、预测或以其他方式与感兴趣的主题在统计上相关的变量的数据的问题。

[0036] 本文公开的系统和方法的实施例可包括构建或创建图数据库。在本公开的上下文中,图是成组的对象,如果这些对象有某种密切或相关的关系,它们就会一起展示。例如,两个展示为节点的数据由路径连接。一个节点可以连接到许多节点,许多节点也可以连接到特定节点。连接第一节点和第二节点的路径或线称为"边"。边可与一个或多个值相关联;作为非限制性示例,这些值可代表所连接节点的特性,或者代表节点或多个节点(如统计参数)之间关系的指标或衡量结果。图格式可能更容易识别一些类型的关系,如对成组的变量或关系更重要的关系,或不太重要的关系。图通常有两种主要类型:"无向",即图表示的关系是对称的;"有向",即图表示的关系不是对称的(在有向图形中,可以用箭头而不是直线来表示节点之间关系的某个方面)。

[0037] 在一些实施例中,信息和数据以本文中称为"特征图"的数据结构形式表示。特征图是包括节点和边的图或图表,其中边的作用是将节点与一个或多个其他节点"连接"起来。例如,特征图中的节点可以代表变量(即可测量的量)、对象、特性、特征或因素。特征图中的边可以代表节点与一个或多个其他节点之间的统计关联度。

[0038] 例如,关联度可以用数字和/或统计术语表示,可以是观察到的(或可能是听闻的)关系,也可以是测量到的相关性,还可以是因果关系。用于构建特征图的信息和数据可以从科学论文、实验、机器学习模型的结果、人为或机器观察、或两个变量之间关联的轶事证明(作为非限制性示例)中的一个或多个获得。

[0039] 例如,特征图可以通过访问成组的来源来构建,这些来源包括关于研究主题与研究中考虑的一个或多个变量之间的统计关联的信息。来源中包含的信息被用来构建数据结构或展示,数据结构或展示包括节点和连接节点的边。边可与关于两个节点之间的统计关

系的信息相关联。一个或多个节点可能有与之相关的数据集,数据集可通过链接或其他形式的地址或访问要素进行访问。实施例可包括允许用户描述和执行对数据结构的搜索的功能,以识别可能与训练机器学习模型相关的数据集,该模型用于做出特定决策或分类。

[0040] 因此,实施例可生成数据结构,其中包括节点、边和数据集链接。节点和边代表概念、感兴趣的主题或以前研究的主题。边代表关于节点之间的统计关系的信息。链接(或另一种形式的地址或访问要素)提供了访问数据集的途径,这些数据集建立(或支持、展示等)作为研究一部分的一个或多个变量之间的统计关系,或变量与概念或主题之间的统计关系。

[0041] 数据科学和数据工程团队的职责之一是管理“数据质量”。这是指将收集或获取的数据用于数据分析和机器学习(ML)建模的适当性和适用性。评估数据质量可包括收集有关数据的信息或事实,如数据来源、收集日期和收集过程的信息,以及核实数据的不同统计属性。这些统计特性可用于识别“更好”(即更准确或更可靠)的候选数据集,以用于训练模型或评估企业或其他实体的绩效。

[0042] 有一些传统工具可以为用户提供有关数据本身的详细信息,还有一些工具可以自动验证数据质量。不过,访问数据集的统计特征通常涉及编写自定义的计算机代码,以查询数据库或以其他方式访问数据,然后应用规则或探索式方法(使用额外的自定义代码)来确定访问的数据(或数据中包含的子集)是否在规则或探索式方法的范围内。这给许多实体带来了负担,这需要分配它们可能无法获得或负担不起的资源。

[0043] 数据质量也会影响机器学习模型的评估。机器学习(ML)包括对算法和统计模型的研究,计算机系统利用这些算法和模型来执行特定任务,而不使用明确的指令,而是依靠识别模式和应用推理过程。机器学习算法根据样本数据(称为“训练数据”)和数据所代表的信息(称为标签或注释)建立数学“模型”,从而进行预测、分类或决策,而无需明确编程来执行任务。

[0044] 机器学习算法被广泛应用于包括电子邮件筛选和计算机视觉在内的各种应用中,在这些应用中,开发一种传统算法来有效地执行任务是困难的或不可行的。由于用于任务的ML模型非常重要,因此基于机器学习的应用程序的研究人员和开发人员会花费时间和资源,为他们的用例建立最“准确”的预测模型。对模型性能的评估以及模型中每个特征的重要性,通常由用于描述模型及其性能的特定指标来表示。例如,这些指标可包括模型准确度、混淆矩阵、精确度(P)、召回率(R)、特异性、F1分数、精确度-召回率曲线、ROC(接收者工作特性)曲线或PR及ROC曲线。每个指标都可以提供略有不同的方法来评估模型或模型性能的一些方面。

[0045] 现代“数据驱动型”业务决策的一个重要因素是确定KPI(“关键绩效指标”或“关键指标”)。许多公司的领导团队都专注于保持KPI的增长,或将KPI作为衡量公司健康状况或业绩的主要“信号”或指标。KPI对业务决策的重要性与生成这些KPI所用数据的质量相关。这是因为,KPI的实用性以及将其作为公司或团队绩效指标的合理性,取决于其适用性以及用于计算KPI的基础数据的准确性和/或可靠性的统计(或其他)衡量。公司可能会投资分析师和工程师来建立“指示板”和其他分析工具,以突出公司KPI的水平和变化,并向决策者通报这些变化。

[0046] 由于用于确定KPI和/或训练模型的数据的重要性及其对模型性能的潜在影响,数

据集的特性可能是选择训练数据和解释训练模型结果的重要因素。这一点在商业情况中尤为重要,在这种情况中,企业生成的数据被用作训练数据或用作训练模型的输入,以生成公司感兴趣的指标。例如,训练有素的模型可用于生成代表企业运营某一方面的KPI,作为非限制性示例,例如收入增长、利润率、营销成本或销售转化率等。

[0047] 在一些实施例中,所描述的用户界面(UI)和用户体验(UX)可作为基础数据分析平台的一部分来实现,例如本文所提及的系统平台,以及在名称为"Systems and Methods for Organizing and Finding Data"的美国专利申请16/421,249(现已公布为美国专利11,354,587)中所描述的。所公开的平台可发现、存储数据、概念、变量或其他特征之间的统计关系,在某些情况下还可生成统计关系。这些关系可以由机器学习模型或程序运行的相关关系生成。

[0048] 所公开的指标监控功能提供了一种利用系统数据组织和分析平台来显示KPI的水平和变化的方法,类似于指示板、数据目录和KPI追踪器等传统方法。然而,该功能并不是孤立地执行的,而是可以显示有关指标"状态"(例如其水平以及其随时间的变化)的元数据,以及该指标与其他被衡量的或以其他方式被监控的指标之间的关系。指标监控功能可显示每个指标的水平和在这些水平的背景下的变化,以及其他指标的变化。不过,与传统方法不同的是,这种背景并非纯粹基于并发性(这可能导致指标之间的虚假关联和不正确的因果假设),而是基于平台对机器学习模型的基础编目和基于相关性的关联所驱动的统计关系。

[0049] 尽管指标监控功能被设计为所公开平台的一部分,但本领域的普通技术人员(例如,了解图数据库和HTTP请求的软件工程师)应该会发现本公开内容是可行的,并且能够用他们选择的编程语言实现指标监控功能。由于指标监控的目的是追踪重要KPI/指标的变化,因此指标监控假定有以事件驱动或其他自动化方式更新的来源(存储在云数据库服务中的数据集通常就是这种情况)。这些数据的更新频率并不那么重要;指标监控对金融服务部门的用户可能很有价值,因为金融服务部门的数据被认为几乎是持续更新的,但它也可能被从事科学研究和处理行政数据(通常由政府实体发布)的个人使用,因为这些数据可能按季度、年度甚至十年更新一次。

[0050] 图1(a)展示了一组要素、组件、功能、流程或操作的框图,这些要素、组件、功能、流程或操作可以是平台架构100的一部分,在平台架构100中可以实施所公开的用于指标监控的系统和方法的实施例。下文简要介绍了示例架构:

架构

●在一些实施例中,图1(a)所示的架构的要素或组件可根据其功能和/对这些元素和组件的访问方式来进行区分。在功能上,系统的架构100可区分以下几种情况:

○信息/数据访问和检索(如所示的应用程序112、增加/编辑118和开放科学103)--这些是实验、研究、机器学习模型或观察结果的信息和描述的来源,它们提供数据、变量、主题、概念和统计信息,作为生成特征图或类似数据结构的基础;

○数据库(图示为系统DB 108)--电子数据存储介质或元件,其采用适当的数据结构或模式以及数据检索协议/方法;以及

○应用程序(图示为应用程序112和网站116)--这些应用程序根据从公共用户(公众102)、客户104和/或管理员106接收到的指令或命令运行。应用程序可执行一个或多个流程、操作或功能,包括但不限于:

■搜索系统DB 108或特征图110,并检索与用户查询相关的变量、数据集和其他信息;

■识别特征图的特定节点或关系;

■将数据写入系统DB 108,以便公众102或者拥有或控制数据访问权限的客户或企业104以外的其他人访问数据(注意,在此意义上,客户104是信息或数据检索架构或来源的要素);

■从指定数据集生成特征图;

■将一个或多个复杂性、统计显著性的相对程度或其他方面或特征的指标或衡量标准,作为特定特征图的特征;和/或

■生成并获取用于训练机器学习模型的数据集推荐;

●从对系统100的访问及系统100的功能的角度来看,系统的架构区分为公众102可访问的要素或组件,限定客户、企业、组织或成组的企业或组织(如社会部门的行业联盟或"数据协作")104可访问的要素或组件,以及系统的管理员106可访问的要素或组件;

●可从多个来源检索(即访问和获取)关于或者表示主题、概念、因素或变量之间的统计关联的信息/数据。这些可能包括(但不限于或要求包括)期刊文章、技术和科学出版物和数据库、用于研究和数据科学的数字"笔记本"、实验平台(例如A/B测试)、数据科学和机器学习平台,和/或公共网站(要素/网站116),在该公共网站中用户可以输入观察到的变量与主题、概念或目标之间的统计(或经验)关系;

○例如,使用自然语言处理(NLP)、自然语言理解(NLU)和/或计算机视觉处理图像(如所示的输入/源处理元件120),信息和数据检索架构的组件可以扫描(例如通过使用光学字符识别,OCR)或"阅读"已出版或以其他方式可访问的科学期刊论文,并识别表明统计关联已衡量的单词和/或图像(例如,通过识别术语"增加"或识别其他相关术语或描述),并且作为响应,检索关于该关联的信息和数据以及检索关于衡量(例如,提供支持)该关联的数据集的信息和数据(例如,如图中所示的标注为"开放科学"的要素103和图1(a)中的步骤或阶段202);

○信息和数据检索架构的其他组件(未显示)可为用户提供将代码输入其数字"笔记本"(如Jupyter笔记本)的方法,以检索机器学习实验的元数据输出(如给定模型中使用的特征的"特征显著性"衡量值)以及实验中使用的数据集的相关信息;

○请注意,在一些实施例中,信息和数据检索通常是定期或持续进行的,这为系统100提供新的信息来存储和结构化,从而向用户公开;

●在一些实施例中,算法和模型类型(如逻辑回归)、模型参数、数值(如0.725)、单位(如对数损失)、统计属性(如p值=0.03)、特征重要性、特征等级、模型性能(如AUC分数)以及与关联相关的其他统计值在检索后被识别并存储;

○鉴于研究人员和数据科学家可能会使用不同的词语或术语来描述相同或近似的概念,因此变量名(如"有氧运动")可在检索时存储,然后与公共领域本体(如维基数据)进行语义关联(即链接或关联),以便根据常见或典型的同义或近似术语和概念对变量(及相关统计关联)进行聚类;

■例如,用户标记为"log_house_sale_price"的变量可能会被系统(并由用户进一步确认)与维基数据中具有唯一ID的主题"房地产价格"产生语义关联;

●如本文所述,中央数据库(图中的“系统DB”108)存储已检索的信息和数据及其相关数据结构(即节点、边、值)。通常以“特征图”110的形式,将中央数据库的实例或投影提供给特定客户、企业或组织104(或其团体)使用,该中央数据库包含系统DB中存储的全部或部分信息和数据;

○由于对特定特征图的访问可能仅限于与给定企业或组织相关的几个人,因此该特定特征图可用于表示视为给定企业或组织104所私有或专有的、关于变量和统计关联的信息和数据(作为非限制性示例,如就业数据、财务数据、产品开发数据、业务指标或研发数据);

○每个客户或用户都有自己的系统DB实例,其以特征图的形式提供。特征图通常从系统DB中并行地读取数据(大多数情况下是频繁读取),从而确保特征图的用户能够访问系统DB中存储的最新信息、数据和知识;

●应用程序112可以在特征图110的基础上开发(“构建”),以执行所需的功能、流程或操作;应用程序可以从中读取数据,向其写入数据,或同时执行这两种功能。应用程序的一个例子是数据集推荐系统(此处称为“数据推荐器”)。使用特征图110的客户104可以使用合适的应用程序112将信息和数据“写入”系统DB 108;如果他们希望与组织外的更多用户或公众共享一些信息和数据,这可能会很有帮助;

○应用程序112可与客户104的数据平台和/或机器学习(ML)平台114集成。谷歌云存储就是数据平台的例子。ML(或数据科学)平台可包括Jupyter Notebook等软件;

■例如,这种数据平台集成允许用户访问客户数据存储或其他数据存储库中的特征(如数据推荐器程序推荐的特征)。再比如,数据科学/ML平台集成可以让用户在笔记本中查询特征图;

○请注意,除与客户的数据平台和/或机器学习(ML)平台集成外,管理员还可使用合适的服务平台架构(如软件即服务(SaaS)或类似的多租户架构)向客户提供对应用程序的访问,或以此取代与客户数据平台和/或机器学习(ML)平台的集成。图3-5进一步描述了这种架构的主要要素或特征;

●在一些实施例中,公众102可以访问基于网络的应用程序。在网站(由www.xyz.com 116表示)上,用户可以从系统DB 108读取数据或向系统DB 108写入数据(如图中的添加/编辑功能118所示),其方式与维基百科等网站类似;以及

●在一些实施例中,存储在系统DB 108中并通过www.xyz.com 116向公众公开的数据可以类似于维基百科等网站的方式向公众提供。

[0051] 一旦信息和数据被访问和处理并存储到数据库(其中可能包含未处理的数据和信息、已处理的数据和信息以及以数据模型形式存储的数据和信息)中,就可以构建包含指定的成组变量、主题、目标或因素的特征图。特定用户的特征图可以包括平台数据库108中的所有数据和信息或其子集。例如,特定客户104的特征图(图1(a)中的110)可根据从系统DB 108中选择的数据和信息构建,这些数据和信息要满足特定领域(如公共卫生)对客户所关注的领域(如媒体)的适用性等条件。在为特定客户或用户部署、生成或构建特征图时,可对数据库108中的数据进行筛选,通过删除与正在调研的问题、概念或主题无关的数据来提高性能。

[0052] 在一些实施例或用途中,用于生成特征图的数据可能是组织或用户的专有数据。

例如,用于构建特征图的数据可以从实验、成组客户或用户、或者特定的受保护数据的数据库中获取,以上作为非限制性的例子。

[0053] 图1(b)是流程图表或流程图,其说明了使用本文公开的系统和方法的实施例构建特征图150的流程、方法、功能或操作。图1(c)是流程图表或流程图,其说明了示例用例的流程、方法、功能或操作,在该用例中,特征图被遍历以识别潜在的相关数据集和/或执行另一个感兴趣的功能(如执行特定应用程序所产生的功能,如图1(a)中要素112所建议的功能),该流程、方法、功能或操作可以在本文所公开的系统和方法的实施例中实现。

[0054] 如图所示(特别是图1(b)),通过识别和访问成组的来源构建或创建特征图,该成组的来源包含与研究中使用的变量或因素之间的统计关联相关的信息和数据(如所示的步骤或阶段152)。可定期或持续检索这类信息,以提供有关变量、统计关联以及用于支持这些关联的数据的信息(如154所示)。如本文所述,对这些信息和数据进行处理,以确定这些来源中使用或描述的变量,以及确定一个或多个变量与一个或多个其他变量之间的统计关联。

[0055] 继续看图1(b),在152处访问数据和信息来源。对访问的数据和信息进行处理,以确定在一个或多个来源中发现的变量和统计关联154。如上所述,此类处理可包括图像处理(如OCR)、自然语言处理(NLP)、自然语言理解(NLU)或其他形式的分析,以帮助理解期刊论文、研究笔记、实验日志或其他研究或调查记录的内容。

[0056] 进一步处理可包括将一些变量链接到本体(如国际疾病分类)或其他数据集,这些数据集提供了与变量所用术语的语义等价或语义相似的术语(如步骤或阶段156所示)。这有助于将特定研究中使用的变量名扩展到其他研究中可能使用的更多实质上等同或类似的实体或概念。一旦确定,变量(如前所述,可以使用不同的名称或标签)和统计关联就会存储在数据库中(158),例如存储在图1(a)中的系统DB 108中。

[0057] 然后,根据特定的数据模型对所访问的信息和数据的处理结果进行结构化或表示(如步骤或阶段160所示);该模型将在本文中进行更详细的描述,但该模型一般包括用于构建特征图的要素(即表示主题或变量的节点、表示统计关联的边,以及包括统计关联的指标或评估的度量)。然后将数据模型存储在数据库中(162);可对其进行访问,为特定用户或用户组构建或创建特征图。

[0058] 如前所述,参照图1(b)描述的过程或操作可以构建包含节点和连接一些节点的边的图(图1(d)是其中的一个示例)。节点代表研究或观察的主题、目标或变量,边代表节点与一个或多个其他节点之间的统计关联。每个统计关联可与数值、模型类型或算法以及统计属性中的一个或多个相关联,所关联内容描述了由边连接的节点(即变量、因素或主题)之间的统计关联的强度、置信度或可靠性。请注意,作为非限制性示例,与边相关的数值、模型类型或算法以及统计属性可能表示相关性、预测关系、因果关系或轶事观察。

[0059] 图1(c)是流程图表或流程图,其说明了根据所公开系统和方法的实施例,可用于为用户构建特征图的流程、方法、功能或操作190。在一个实施例中,这可能包括以下步骤或阶段(其中一些步骤或阶段与图1(b)中描述的步骤或阶段重复):

●确定并访问来源数据和信息(如步骤或阶段191所示);

○在一个实施例中,这可以是来自期刊、学术期刊或其他描述研究或调查的出版物的可用的数据和信息;

○在一个实施例中,这可能代表专有数据和信息,如组织生成的实验结果、组织感兴趣的研究课题或组织从客户或顾客处收集的数据;

●处理访问的数据和信息(如步骤或阶段192所示);

○在一个实施例中,这可能包括识别和提取关于一个或多个研究或调查的主题、研究或调查中考虑的变量或参数、用于在一个或多个变量之间和/或变量与主题之间建立统计关联的数据或数据集的信息,以及以指标、关系或类似量的形式的统计关联的衡量值;

○在一个实施例中,这种处理可通过使用训练过的模型自动或半自动执行,该模型利用语言模型或语言嵌入技术来识别感兴趣或相关的数据和信息;

●将处理过的数据和信息存储到数据库中(如步骤或阶段193所示);

○在一个实施例中,数据库可包括一个或多个分区,以便将从组织、成组的来源或成组的人群中获取的数据分离成单独的数据集,该数据集用于生成特征图;

■如果成组的数据是从专有研究、特定人群中获取的,或受到其他法规或限制(如隐私或安全法规),这可能是一种有用的方法;

○在一些实施例中,经过处理的数据和信息可以按照特定的数据模式进行存储,特定的数据模式包括特定的标签或字段;

●接收表示感兴趣主题的用户输入,并响应以生成特征图(如步骤或阶段194所示);

○在一个实施例中,用户输入可指定来源、日期、阈值或其他形式的限制,这些限制被用作用于生成特征图的数据和信息的筛选机制;

●遍历特征图,评估用于生成特征图的数据、信息和元数据(如步骤或阶段195所示);

○这可能包括在评估过程之前,根据规则、约束、阈值或其他条件筛选特征图所代表的数据和信息;

○这可能包括评估处理流程中的数据、信息和元数据,该评估由特定应用程序或成组的控制或指令决定;

■在一个实施例中,作为非限制性示例,这可能包括聚合统计数据和/或元数据,识别统计相关的或重要的关系,或生成关系或变量值的特定指标或指示;

■在一个实施例中,这可能包括使用规则集或条件评估聚合数据,以识别潜在的重要变量或关系,或提醒用户注意特定条件;

■在一个实施例中,这可能包括对层中的节点执行某种类型的网络分析,以确定网络特征;以及

●向用户展示图的遍历和评估的结果(如步骤或阶段196所示);

○在一个实施例中,这可能包括将用于生成特征图的主题、变量和数据分离成不同层的节点,并在节点和层之间建立连接的边;

○在一个实施例中,这可能包括向用户指示两个节点之间的关系,这两个节点具有一些特征(作为示例,例如强度、重复性、超过阈值或更可靠);

○在一个实施例中,这可能包括向用户提供列表或表格,在其中指明影响输入的概念或主题或受输入的概念或主题影响的概念或主题,并提供这种关系的属性的元数据;

○在一个实施例中,这可能包括将成组变量或主题与指标相关联,并向用户显示

指标的值和/或指标的变化；

○在一个实施例中，这可能包括使用一个或多个关于这些实体之间的统计关系的指标或指示符（如标志、警报或颜色）来表示两个变量之间、两个主题之间或变量与主题之间的关系。

[0060] 图1（d）是特征图的数据结构198的部分的示例的示意图，该特征图的数据结构可用于组织和访问数据和信息，并可使用本文所公开的系统和方法的实施例创建。下文介绍了特征图198的要素或组件以及相关的实施的数据模型。

[0061] 特征图

●如前所述，特征图[1]是一种结构化、表示和存储主题及其相关变量、因素或类别之间的统计关系的方法。特征图的核心要素或组件（即“构成要素”）是变量（在图1（d）中标识为V1、V2等）和统计关联（标识为变量之间的连接线或边）。变量可与“概念”（图中C1为其示例）链接或关联，“概念”是语义概念或主题，其本身通常无法直接衡量或以有效的方式衡量（例如，变量“抢劫案数量”可与概念“犯罪”关联）。变量是可衡量的经验对象或因素。在统计学中，关联被定义为“两个随机变量之间的统计关系，无论是否存在因果关系”。统计关联产生于通常所说的科学方法的一个或多个步骤或阶段，例如，统计关联可分为弱关联、强关联、观察关联、衡量关联、相关关联、因果关联或预测关联；

○例如，参考图1（d），对输入变量V1进行统计搜索，可检索到：（i）与V1有统计关联的变量（如V6、V2）（在一些实施例中，只有当变量的统计关联值高于定义的阈值时，才会检索到该变量），（ii）与这些变量有统计关联的变量（例如，V5、V3、V4）（在一些实施例中，只有当变量的统计关联值高于定义的阈值时，才能检索到该变量），（iii）通过共同概念（如C1）与一个或多个变量（如V2）在语义上相关联的变量（如V7），一个或多个变量（如V2）与输入变量V1在统计上相关联，（iv）在统计上与这些变量相关联的变量（如V8）；以及数据集（如D6、D2、D5、D3、D4、D7、D8），其衡量相关联的变量或证明检索到的变量在统计上相关；

■请注意，与所公开的实施例不同的是，对输入变量V1的语义搜索会检索到以下内容：（1）变量V1，以及（2）衡量该变量的数据集（如D1）；

●特征图中包含了从期刊论文、科学和技术数据库、研究和数据科学的数字“笔记本”、实验日志、数据科学和机器学习平台、用户可在其中输入观察到或感知到的统计关系的公共网站、专有商业信息和/或其他可能来源获取的有关统计关联的信息和数据；

○如前所述，利用自然语言处理（NLP）、自然语言理解（NLU）和/或图像处理（OCR、视觉/图像处理和识别）技术，信息和数据检索架构的组件（图1（a）为其示例）可以扫描或“阅读”已发表的科学期刊论文，识别表明统计关联已衡量（例如“增加”）的单词或图像，并检索关于该关联的信息和数据以及关于衡量或确认该关联的数据集的信息和数据；

○信息和数据检索架构的其他组件为数据科学家和研究人员提供将代码输入其数字“笔记本”（如Jupyter笔记本）的方法，以检索机器学习实验的元数据输出（如模型中使用的特征的“特征显著性”衡量值）以及实验中使用的数据集的相关信息。请注意，信息和数据检索是定期进行的，在一些情况下是持续进行的，这就为系统提供了新的信息来进行存储和结构化，并向用户公开；

[1] 在本公开的内容中，使用术语“特征图”是因为本实施例将实体构成图，而实体是通过变量（在此称为特征）之间的统计关系（感兴趣的衡量结果）连接的，并不是通过语义共现（如在传统的“知识图谱”中）连接。

●在一个实施例中,数据集与特征图中的变量相关联,该关联通过与相关数据集/数据桶/渠道的URI或其他形式的路径或地址的链接实现;

○这样,特征图的用户就可以根据先前证明或确定的数据对指定目标或主题的预测能力来检索数据集(而不是像传统知识图那样,根据来源之间的语义共现,检索与指定目标或主题在语义上可能不太相关或不相关的数据集);

○例如,使用本文公开的系统和方法的一个实施例,如果数据科学家搜索“蓄意破坏”作为研究的目标主题或研究的目标,他们将检索已显示可预测该目标或主题的数据集,目标或主题例如“家庭收入”、“光照度”和“交通密度”(以及与目标统计关联的证明),而不是衡量蓄意破坏的实例的数据集;

●关联在被检索到时,其数值(如0.725)和统计属性(如p值=0.03)存储在系统DB 108中,并可作为构建的特征图的一部分提供。如前所述,鉴于研究人员和数据科学家可能会使用不同的词语来描述相同或类似的概念或主题,因此变量名称(如“有氧运动”)会在检索时存储,并可能在语义上与公共领域本体(如维基数据、字典、辞典或类似来源)相关联,以便根据共同或类似的概念(如同义词或业内人士认为可以互换的术语)对变量(以及相应的统计关联)进行聚类;

●从某种意义上说,系统100采用数学、基于语言和可视化的方法来表达可用数据和信息的认识论的和基本的属性,例如,表达支持给定统计关联的信息和/或数据的质量、严谨性、可信度、可重复性和完整性(作为非限制性示例);

○例如,给定的统计关联可与用户界面中的特定分数、标签和/或图标相关联,这些指示基于统计关联的科学质量(整体和/或特定参数,特定参数如“已通过同行评审”),从而向用户提示信息,该信息可用来决定是否进一步调查该关联的信息。在一些实施例中,通过搜索特征图检索到的统计关联可根据其“科学质量”分数进行筛选。在一些实施例中,质量分数的计算可将存储在特征图内的数据(例如,给定关联的统计意义或关联被记录的程度)与存储在特征图外的数据(例如,检索关联的期刊文章所获得的引用次数或文章作者的h指数)结合起来;

○例如,包括在具有高曲线下面积(AUC)分数的模型中衡量到的高且显著的“特征显著性”分数的特征、具有部分依赖图(PDP),以及被重复记录的统计关联,可能会在特征图中被视为“强”(可能更可靠)统计关联,并在图形用户界面中给出识别颜色或图标;

○请注意,除了检索主题或概念的变量和统计关联外,实施例还可以为用户检索在实验或研究中使用的其他变量,使得将统计关联置于背景中理解。例如,如果用户想知道实验中是否控制了一些变量,或者模型中包含了哪些其他变量(或特征),这可能会有所帮助。

[0062] 数据模型

特征图(或系统DB)中的主要对象通常包括以下一个或多个对象(注明了有助于定义该对象的信息):

●变量(或特征)--您要衡量什么,在哪些人群中衡量?

●概念--您正在研究的主题、假设、观点或理论是什么?

●邻域--您要衡量的主题是什么(这通常比概念更宽泛)?

●统计关联--这种关系的数学基础和价值是什么?

●模型(或实验)--衡量的来源是什么?

●数据集--用来提示或衡量关系的数据集(如模型训练数据)或衡量变量的数据集是什么?

这些对象是相互关联的,如图1(d)中的特征图的示例所示:

●变量通过统计关联与其他变量相连;

●统计关联来自模型,并得到数据集的支持;以及

●变量与概念相连,概念与邻域(或邻域的一部分)相连。

[0063] 参见图1(d),如前所述,特征图的一种用途是使用户能够在特征图中搜索一个或多个数据集,这些数据集包含的变量已被证明与研究的目标主题、变量或概念有统计学关联。使用示例如下:

●用户输入目标变量,并希望检索可用于训练模型以预测该目标变量的数据集,即那些与目标变量在统计上相关联的变量的数据集(如图1(b)中流程170所示);

○例如,参考图1(d),统计搜索输入V1(此处为变量)会导致算法(例如广度优先搜索(BFS))遍历特征图(如图1(b)的步骤或阶段174所示)并返回(如图1(b)的步骤或阶段176所示):

■在统计上与V1相关联的变量(如V6、V2);

●在一些实施例中,只有当统计关联值高于定义的阈值时,才能检索到该变量;

■在统计上与这些变量相关联的变量(如V5、V3);

●在一些实施例中,只有当统计关联值高于定义的阈值时,才能检索到该变量;

■由共同概念(如C1)与一个或多个变量(如V2)语义相关的变量(如V7),一个或多个变量(如V2)与输入变量V1在统计上相关;

■在统计上与这些变量相关联的变量(如V8);以及

■衡量或证明所检索的变量的统计意义的数据集(如D6、D2、D5、D3、D4、D7、D8);

●在遍历特征图并检索可能相关的数据集后,可根据应用或用例对这些数据集进行“筛选”、排序或者排列(如图1(b)的步骤或阶段178所示):

○通过所述遍历过程检索到的数据集随后可根据用户输入的搜索的标准和/或软件实例管理员输入的标准进行筛选。搜索数据集筛选器示例可包括以下一个或多个:

■群体与关键字:所关注的变量是否在用户感兴趣的群体和关键字中进行衡量(例如,用户、物种、城市或公司的唯一标识符)?这影响了用户将数据加入训练集以用于机器学习算法的能力;

■符合规范:数据集是否符合适用的监管规定(如符合GDPR规范或HIPAA规定)?

■可解释性/可说明性:变量是否可以被人类解释或理解?

■可操作性:模型的用户是否可以操作该变量?

[0064] 在一个实施例中,用户可以输入概念(图1(d)的198中的C1所示),如“犯罪”、“财富”或“高血压”。作为回应,本文公开的系统和方法可使用语义和/或统计搜索技术的组合来识别以下一项或多项内容:

●与C1语义相关的概念(C2)(注意,此步骤可以是可选的);

●与C1和/或C2语义相关的变量(V_X);

●与每个变量V_X在统计上相关的变量;

●已确定的统计关联的一种或多种衡量方法;以及

●数据集,其衡量每个变量V_X和/或证明或支持与每个变量V_X在统计上相关的变量。

[0065] 图2(a)是框图,其展示了一组要素、组件、功能、流程或操作,它们可以是平台架构的一部分,在平台架构中可以实施所公开的指标监控系统和方法的实施例。图2(a)是流程图表或流程图,其展示了一组要素、组件、功能、流程或操作,它们可以是平台架构的一部分,在平台架构100中可以实施所公开的指标监控系统和方法的实施例。具体来说,图2(b)描述了图2(a)中的一些步骤,更侧重于不同用户的交互以及软件要素,该软件要素有助于实现指标监控功能并且供用户使用。

[0066] 图2(a)描述了如何实施使用所公开的指标监控功能,从而监控存储在云数据库服务(或"数据仓库"204)中的数据集的特征变化。左列(由要素202表示)中展示要素、功能或操作的区块(例如,数据集元数据206)是系统平台展示特征和指标(以及特征之间的衡量统计关系)的示例,而右侧(由要素203表示)展示要素、功能或操作的区块则示出了用户交互、用户输入以及软件计算或其他执行代码,平台可使用上述信息来处理和存储有关数据集及其特征的元数据。

[0067] 在一些实施例中,图2(a)所示的步骤、阶段、功能、操作或处理流程可包括处理步骤,通过该处理步骤,平台的数据仓库检索集成计算相关元数据并将其发送(通常通过HTTP请求)到平台的后端API。后端服务将元数据存储到平台的图数据库(如图1(a)中的要素108),在其中包含支持特征图的功能的数据。特征图是用户在使用平台的前端和生成的用户界面时看到并与之互动的内容。

[0068] 用户可以与平台的前端用户界面互动,识别感兴趣的特征,并且当特征具有所期望的形式(即具有与时间戳关联的数值)时,用户就可以定义用于监控的指标,将指标与这些特征连接起来,并激活指标监控功能。指标监控可根据指标的值或者指标的值的变化(以及平台的基础数据)为用户提供可视化指示(在特征图上),并可通过电子邮件或平台应用程序本身生成警报和通知。

[0069] 如前所述,指标监控功能或能力将在指标彼此关联的情况下显示指标的变化(例如,如图2(a)所示,平台用户将能够看到指标一(208)的变化以及指标二(210)的变化),并描述这些指标之间所衡量的统计关系(分别如数据209和211所示)。用于显示两个指标的变化的平台背景信息不仅显示指标的当前水平和变化,还可以使用来自机器学习模型的输出以及与指标相连的基础特征之间的其他统计关系来生成数据和信息并向用户显示数据和信息。

[0070] 图2(b)描述了图2(a)中的一些步骤,这些步骤更侧重于用户的交互以及软件要素,该软件要素有助于实现指标监控功能并且供用户使用。图中的每个步骤、阶段、要素、功能或操作都与所公开平台的软件组件(或软件服务)相对应,有助于用户使用指标监控功能。在图2(b)所示的示例中,显示的组件为(按图中从上到下的顺序排列):

●用户可根据步骤、阶段、操作、流程或功能250所示,通过与数据库服务(数据仓库)的集成,在平台上添加用于追踪的数据集;

●如步骤、阶段、操作、流程或功能252所示,平台的检索服务计算相关数据集和特征的元数据,并且向平台的后端API提交HTTP请求;

●如步骤、阶段、操作、流程或功能254所示,平台的后端API处理这些请求中包含的数据负载,以准备用于存储的数据集和/或特征的元数据;

●如步骤、阶段、操作、流程或功能256所示,平台的后端服务将数据集和/或特征的元数据和统计关系存储到图形数据库中;

●如步骤、阶段、操作、流程或功能258所示,平台的后端服务会将检索流程中的新元数据与图数据库中的现有元数据连接起来,从而在适用时将数据集和特征与现有对象连接起来(注意,这是一个可选步骤,取决于现有图数据库的内容);

●平台的元数据在平台前端提供,用户可以通过平台前端查看作为特征图一部分的对象(例如数据集和特征)之间的联系。如步骤、阶段、操作、流程或功能260所示,用户还可以在特征和指标之间建立联系,该指标用于追踪KPI或关键指标;

●如步骤、阶段、操作、流程或功能262所示,当特征具有正确形式时(例如,如要素264所示,数据具有相关的时间索引),平台会显示特征和指标以及它们的最新值和最近的变化,并提示用户打开指标监控;

○如果这些对象与当前正在监控的指标有重要关系的话,平台或系统还可能会提示用户打开指标监控,并建议用户监控重要的特征和指标;

●如步骤、阶段、操作、流程或功能266所示,用户可为指标监控设置规则,这些规则管理用于所监控的指标的可视化指示/区分,并且通过电子邮件和平台生成警报和通知--这些规则将写入平台后端并存储在特征图中;

●然后,如步骤、阶段、操作、流程或功能268所示,对用户设定的条件进行评估,以生成显示的可视化区分、警报和/或通知。如上所述,平台的后端还可以追踪指标监控的状态,以发现指标之间的重要关系并提出建议;

●如步骤、阶段、操作、流程或功能270以及与步骤、阶段、操作、流程或功能254相连的控制回路所示,这些步骤或流程是迭代进行的,从而使检索到的新信息或数据产生用户所感兴趣的监控的数据的变化。

[0071] 在一些实施例中,作为其架构的一部分,所公开的平台包括软件,该软件自动从远程数据库检索和处理数据,并将计算出的元数据(包括数据集中的特征之间统计关系的元数据)写入平台数据存储器。这种架构基于微服务,微服务旨在按计划和/或以事件驱动的方式运行。不过,如果更新的数据是从来源“检索”而来并且写入指标监控软件和功能可以访问的存储位置,则可能不需要以这种形式实施。如前所述,为了实现指标监控功能,期望是通过将数据的感兴趣的值与特定时间段或其他形式的索引相关联的方式检索数据。

[0072] 例如,JavaScript中的关联数组可用于将数据的值与特定的时间戳的对象关联:{“2010-01-01 00:00:00Z”:10.4,“2010-01-02 00:00:00Z”:11.2},其中,关联数组的“键(keys)”代表“UTC”时间标准中的时间戳,键后面的数字代表与这些时间戳相关联的数据值。这是一个可以保存数值并将其与特定时间戳关联的数据结构的非限制性示例。

[0073] 实施例可包括以下具体方法:对不同时间段的数据进行内插和聚合,以及指定应与时间段相关联的数据值。无论使用哪种方法来“决定”与每个值相关的时间段或索引,本文所公开的指标监控功能都能为用户提供帮助;不过,由于用户通常会依赖数据来了解所关注的指标是如何随时间变化的,因此应向用户明确说明这样做的方法。

[0074] 如果数据以电子方式存储,并带有与数据值相关联的时间戳,那么在一个实施例

中,实现指标监控功能的软件可包括以下数据组织操作或流程:

●“当前”或“最新”值是时间戳按“降序”时间顺序排序时与第一个时间戳相关的值。“前项”值是按“降序”时间顺序排序时与倒数第二个时间戳的相关的值(参见图2(a)中的要素209和211);

●当只有一个值时,“前项”值将显示为“不可用”、“不适用(N/A)”或“非数字”,而百分比变化则显示为“不可用”(或“不适用”或“非数字”)。如果这两个值都不是数值,那么这两个值都会显示为“不可用”或“不适用”或“非数字”,百分比变化也是如此;

●否则,百分比变化的计算方法是当前值减去前项值,再除以前项值。如果前一个值为零,平台可以用表示“无限”的“Inf”来表示百分比变化;

●在平台上,上述值存储在图数据库中,并可通过HTTP请求访问后端API获取。用户可以使用“前端”技术计算百分比变化,但在一些实施例中,指标监控会将百分比变化值写入图数据库中的指标对象。这是可取的,也是推荐的,因为用户可能希望通过查询后端API来获取关于指标监控进程或状态的信息;

●实施指标监控功能的另一个方面是设定和评估监控“规则”(如图2(a)中的功能、操作或流程212和213所示)。在一个实施例中,作为平台架构的一部分,包含了比较/警报规则的参数化,其中监控规则由“字段”、“运算符”和“值”组成的“三元组”表示;

○“字段”指的是存储在图数据库中的指标监控对象的字段。该字段可以是“最新值”、“百分比变化”或其他可用于指标监控功能的元数据,供用户监控KPI或指标。该字段的设计非常灵活--最新值和百分比变化是常用的追踪值,但用户可能希望追踪“历史最高值(价格)”或“52周最低值(价格)”,以这两个常用的财务指标为例;

○“值”字段是用户可以指定的某个值(可能有一个默认值),作为规则中用于比较的基础。由于指标监控的本质是数字化的,因此用户应该指定这个“值”为数字;

○“运算符”字段表示如何在监控指标的“字段”的值和用户指定的“值”(如前所述,指标监控功能可能会向用户建议该值)之间进行数学比较。例如,运算符可以指定为“大于绝对值”,这意味着将“字段”中所指值的绝对值与所提供的“值”进行比较,看其是否大于该“值”。

■“运算符”的定义优选是足够灵活的,以包含可能涉及存储在“字段”中的值的计算或“聚合”的监控规则。该功能的实现可包括运算符枚举,其中预定义的软件功能(如果所使用的编程语言允许)实现每个运算符;

●指标监控功能包括可视化要素,使用户能够快速查看其监控的指标的水平和变化。在指标监控的一种实现方式中,需要关注的指标或处于“警报”阶段的指标会以用户选择的非默认颜色、指定的格式(如斜体或粗体)或图标的形式(对于不喜欢用颜色或格式区分用户界面要素的用户而言)显示。颜色或格式的选择将作为监控规则的一部分保存;

●指标监控功能可包括用户界面,用户可在此指定所需的监控规则。在一个实施例中,这是一种基于语言的“下拉菜单”功能,用户可以从一组可用的“字段”、“运算符”中进行选择,然后设置“值”来指定规则。这些已定义的三元组(基于用户输入)作为与相关指标相关的属性保存在图数据库中;

●指标监控的一种实现方式还可以让用户在指定或定义规则时看到监控结果的情况。例如,如果监控规则是在最新值大于0时,将可视化要素设置为绿色,那么如果指标的

最新值实际上大于0,监控数据上的最新值的字段就会设置为绿色。如果监控规则是在百分比变化小于10%时,将可视化要素设置为蓝色,那么如果满足条件,监控数据上的百分比变化值将为蓝色。如果用户将规则中的值更改为条件不再成立的比较值,其将变回默认颜色或外观;

●本文所公开的指标监控功能与其他编目、指示板或分析工具的区别在于,用户可以在看到建模结果或表明统计关系的其他数据源的同时,看到其所监控信息的完整背景信息。这是所公开平台的一个特点,并且涉及监控指标的关系的展示的实施细节与所公开平台的设计和实施方式有关;

○在这方面,所公开的平台基于图数据库建立,因此每个被监控的指标对象都有可能与其他对象建立丰富的连接网络或"边"。当图中有许多关系,并且许多关系都在被监控时,指标监控可视化要素对用户特别有用。在这种情况下,用户可以看到不同的联系,并了解他们所选的指标如何以及为何具有所显示的统计变化"模式";

○在一个实施例中,实施指标监控功能包括指定的可以应用监控规则的数据结构,同时还需要一种存储技术,在这种存储技术中,相关指标可跨不同的元数据进行关联;

●除了上述特征或功能外,在一些实施例中,指标监控功能的实施还可包括让用户发现或获知最优(或更优)规则的能力,并因此更多地了解其数据所代表的系统和关系;

●需要注意的是,如果没有预定义的业务规则或已发布的KPI/指标(作为示例),

用户可能不知道如何最好地定义指标监控规则。在一个实施例中,可通过推荐功能提供这种帮助,推荐功能根据收集的所关注的相关特征和指标的元数据,推荐监控的值/指标;

○作为非限制性的例子,当某个特征或指标的值勉强超过或低于某个数值界限时,就可能会建议使用临界值,在这种情况下,用户只能在一定比例的时间内收到警报或通知。或者,所关注的特征和指标可能与另一个特征或指标相似,建议的规则可能是以相同的方式监控这两个指标;

■所公开的平台、图数据库(系统DB)和后端基础设施使用户能够查看来自大量来源的形成为系统的数据和元数据。这种设计使开发人员和用户能够快速查询具有相似统计特征和/或相似元数据属性的特征、变量和关系(图中的节点和边);

■即使在缺少由用户定义的指标监控规则或者其他预定义业务规则的情况下,也可利用所公开的平台独有的这些信息来发现指标监控的自然候选对象。例如,"内置"推荐功能可以考虑许多此类统计特征或属性,从而提出监控规则建议;

■推荐功能的实施可包括查询和代码,以确定实际的KPI,如活跃用户量的衡量值(通常可预测销量和收入)。在一些实施例中,这些指标可基于以下一个或多个:(1)统计特征(如对其他特征具有高度预测性,或与公司的其他重要指标密切相关);(2)元数据,包括特征或变量名,其作为特征存在于多个数据集中,或被追踪的时间相对较长;或者(3)使用量,如相对于其他变量或特征,用户访问该变量或特征页面的次数;

■推荐功能可根据指标的统计特征或元数据提出"智能"监控规则。如何实施这些规则的训练数据也可以从该平台的公共版本中获取--在那里,用户可以针对各种来源的数据设置指标监控规则,而这些规则的有效性(触发频率以及用户如何响应这些警报)可以推动推荐规则性能的迭代改进;

●在一个实施例中,推荐功能的“构成模块”包括衡量不同特征和指标中元数据的相似性,以及对统计特征中的相似性进行索引。相比之下,为典型数据仓库中的每个特征生成跨特征的统计关系往往非常困难,而且计算成本高昂;

●这种推荐功能可以使用建议的基于规则的相似性表达式或关系来实施;

○作为非限制性的例子,为任何语义相似的指标设置相同的规则可能是推荐的第一规则。实现这一点的一种方法是在搜索服务中索引指标名称的值(可能还有关于指标的其他元数据),当用户为某个不同的指标设置监控规则时,为其他每个受监控的指标计算相似度分数--然后推荐与最相似指标相关的规则以及现有的默认规则;

○另一个可能的实施功能是建议对不在数据集检索/更新流程中的指标进行监控;

■作为非限制性示例,模型性能指标(如果定期更新)可能与用于指标监控功能的时间戳索引值数组相似(如上所述,可能由时间戳索引值数组表示)。这些数据可存储为与模型对象相关的元数据,供所公开的平台的用户使用。平台的用户界面可将这些以时间为索引的模型性能指标作为附加功能呈现,这些附加功能可与其他指标连接并被监控;

■当模型性能指标与时间戳相关联时,可使用单独的软件服务或功能来查找具有相同时间戳索引的其他数据数组(必要时,这可能是由于使用了时间实例之间的内插或外推方法),并计算时间序列分析值,以建立以时间为索引的特征之间的稳定关系。

[0075] 所公开的指标监控功能旨在为用户提供其监控的KPI或其他指标的完整统计背景和关系。为此,平台前端描绘了利用平台架构及其收集和识别的元数据构建的特征图。指标监控功能的可视化提示与特征图的可视化提示相结合,可帮助用户更深入、更全面地了解图中数据的关联。

[0076] 与指标监控功能相关的用户界面(UI)显示是根据存储在平台后台的数据生成的。当指标监控能力或功能被激活时,平台前端会对指标的最新值和任何相关的前项值应用已定义的监控规则,平台提供给用户的视图可能会因此而改变。

[0077] 在一个实施例中,前端JavaScript代码(在渲染指标节点的可视化显示之前,无论是在作为平台一部分的特征图中,还是在平台生成的特定指标页面中)用于处理所定义的规则,该规则通常存储在指标对象本身中。如前所述,规则可以表示为以下内容的集合:

●值(即指标值要与之比较的临界值或阈值);

●字段(应作为规则的一部分进行比较的指标的值的来源,例如最近值的水平,或最近值与紧接着的前项值之间的百分比变化);以及

●操作符(相关字段应如何与规则的值进行比较,例如“大于或等于”或“严格小于”)。

[0078] 可以在平台架构中的一个或多个位置选择或定义规则,在这些位置可以编辑有关指标的元数据。在一个实施例中,这包括指标页面、指标“卡片”(其中指标作为例如在模型或数据集中的其他对象的一部分被引用)以及“匹配控制台”,其中用户可以将指标与特征进行匹配。在一个实施例中,规则制定可包括三个步骤:

●设置“规则”,即选择阈值或条件,确定应向用户发出警报时指标的水平或变化;

●规定如何直观地显示任何规则“违反”或警报(例如通过颜色、格式或图标);以及

●如何向用户发送警报(例如,用户可以选择通知的方式,例如电子邮件或平台通知,以及选择这些警报发送的频率)。

一旦定义了规则,规则的定义就会显示在指标页面上。

[0079] 在一个实施例中,无论是否设置了规则,都可以执行指标监控功能。如果未设置规则,则指标的表示不会触发警报(通过通知或在平台上直观地显示),但最新值、紧邻的前项值以及两个值之间的百分比变化可能会在显示指标的任何地方显示出来(例如,在平台图中、在指标页面上和/或正在追踪的指标目录中)。

[0080] 指标值由平台前端通过图查询生成,该图查询可找到用于衡量所选指标的特征的适当值。当只有一个具有特定时间(索引)数据的特征与指标连接/相关联时,该特征将用于指标监控值。如果有多个具有特定时间数据的特征与指标连接,那么默认情况下,第一个与指标连接的特征就是用作指标监控值的特征(尽管用户可以将此默认值更改为其他特征)。在一个实施例中,为指标监控提供数值的特征可显示在指标页面的顶部,并提供指向该特征的链接,以便用户检查用于生成指标监控数据的每个特征。

[0081] 所公开的平台和数据模型捕获有关数据集和模型的信息,帮助用户管理、发现和使用由机器学习模型的相关性和关联性生成的统计关系。平台数据模型将特征、数据集、模型和其他对象指定为节点,并使用图架构构建平台,从而存储这些对象和平台所创建的对象之间的边,这些边编码有关这些关系的信息。

[0082] 平台基于数据集和模型的统计属性追踪(并可计算)关系强度。在一个实施例中,平台可以定期更新如何评估关系强度的科学标准,首先是统计显著性的标准衡量方式(如计算置信区间和各种形式的统计假设检验)、统计的“经验法则”(如科恩(1962)定义的传统公认的效应大小水平),以及编码到平台后端和机器学习渠道中的其他特定领域知识的来源。

[0083] 该平台发现和学习的统计关系来自由平台计算的相关性和机器学习模型,对这些统计关系的处理会产生特征图,该特征图是指标监控能力和功能的基础。所公开的指标监控能力和功能可为用户提供来自不同数据源的定期更新的指标值,并可告知用户指标水平或指标增长率的重要或显著变化。因此,特征图可用来告知用户可以或应该预期的KPI/指标的变化。添加到平台中的相关性和机器学习模型包括当前时间段的数据,并且可纳入统计关系的衡量中;这可使平台不断“学习”和改进用户可利用和用于决策的知识和数据。

[0084] 正如所公开的,用于为平台生成用户界面显示的数据存储在图数据库中。图数据库包括特征节点(可连接到汇总每个特征的统计信息的节点)、特征之间的边和“关联”节点(聚合和汇总特征之间的统计关系)。特征节点与指标节点之间还可能具有边,用户(和平台)在边中存储有关指标的元数据以及指标的追踪或支持信息。

[0085] 在一些实施例中,所公开的系统和方法使用户能够监控与业务相关的指标(如KPI),并更有效地评估用于生成这些指标的基础数据的质量。这一功能有望使用户在企业运营方面做出更明智的决策。在一些实施例中,这可能包括实现以下一个或多个功能或能力:

●创建由节点和边组成的特征图,其中;

○节点代表一个或多个概念、主题、数据集、元数据、模型、指标、变量、可测量的量、对象、特性、特征或因素(作为非限制性示例);

■在一些实施例中,基于发现(或获取利用)的数据集、元数据、模型,从已训练的模型生成输出、生成有关数据集的元数据或开发本体或其他形式的层次关系(作为非限制性示例),从而创建节点;

○边代表第一节点和第二节点之间的关系,例如统计意义上的关系、依赖关系或层级关系(作为非限制性示例);

■在一些实施例中,可以创建连接第一节点和第二节点的边,以表示由机器学习模型或其他评估方式确定的两个节点之间的有效统计关系;

○与边相关联的标签可指示边所连接的两个节点之间关系的一个方面,如两个节点之间的关系所基于的元数据,或支持两个节点之间具有统计意义的关系的数据集(作为非限制性示例);

●为用户提供用户界面显示、工具、特征和可选要素,使用户能够执行以下一项或多项功能或操作:

○确定监控或追踪的感兴趣的指标(如KPI);

■其中,相关指标可由经过训练的模型、公式、等式或规则集(作为非限制性示例)生成,并可进一步基于、生成自或衍生自基础数据,基础数据是时间的函数(即时间索引);

○定义规则,该规则说明何时应生成有关已确定的指标的行为的警报或通知;

■这可以基于绝对值、值的变化、百分比变化、一段时间内的百分比变化或阈值(作为非限制性示例);

○定义如何在用户界面显示上识别或指示应用规则的结果,例如通过颜色、图标或格式来进行显示(作为非限制性示例);

○允许用户选择已生成警报的指标,并作为响应,提供有关指标值随时间的一个或多个变化的信息、已满足或已激活的导致警报或通知的规则、指标与其他指标的关系(如果相关),以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的可用信息(作为非限制性示例);

●为用户提供关于以下内容中的一项或多项的推荐:可能具有监控价值的不同的指标或指标集、可能有助于检查的数据集、可能与确定的指标相关的元数据或用户可能感兴趣的基础数据或指标的另一个方面;

○其中,推荐可能(至少部分)来自经过训练的机器学习模型所生成的输出、统计分析、研究、与其他指标或数据集的比较或其他形式的评估。

[0086] 所公开的指标监控能力和功能以集成的方式改进了KPI(或其他指标)监控和数据质量分析流程。指标监控功能可提供数据质量监控,数据质量监控衡量数据集的统计属性,例如(但不限于)数据中观察结果的缺失率,或汇总统计数据(如最小值、最大值或平均值)的变化,并允许用户可视化和理解在背景环境中的数据的变化。

[0087] 在一些实施例中,用户可能会收到提示或通知,这表明数据发生了变化,这些变化会在不同来源的数据集之间进行比较,并与数据来源和/或受监控指标的相关元数据一起显示。与以孤立方式显示KPI的传统指示板不同,所公开的系统和方法还以图格式或图表现方式来展示受监控的指标,即作为特征图的一部分(或与特征图结合)。这样就能识别指标之间的重要统计关系,使用户能够识别重要指标的“共同运动”。这项功能为用户提供了一种高效、有效的方法,用于评估某个指标的当前水平和/或增长率,并预测相关指标的未来

水平和增长率。

[0088] 如上所述,所公开的用于监控指标和评估基础数据集的统计关联的系统和方法的实施例可与参考平台结合使用,该参考平台由代理人运营。该平台可用于向用户揭示驱动任务、团队、公司和社区发展的潜在关系。从某种意义上说,数据团队的任务就是通过收集和分析数据来创造理解。所公开的平台可用于汇总这些信息,并向用户显示所产生知识的环境和背景。同样,团队可以衡量KPI或其他指标,以衡量团队、公司或社区特定部分的相对健康状况。所公开的指标监控功能可让这些团队更好、更全面地了解团队(或公司或社区)的健康状况,正如成组的指标所反映或表明的那样。

[0089] 在一个实施例中,本文提及的或美国专利申请16/421,249“Systems and Methods for Organizing and Finding Data”(现已分布的美国专利11,354,587)中描述的“系统”平台包括(作为与数据库服务集成的软件的一部分)“检索”工具,该“检索”工具可从数据集中自动检索元数据和统计属性。这种自动检索功能允许平台存储以时间为索引的统计元数据。在一个实施例中,当存在以时间为索引的特征(如变量或参数)时,用户可以通过用户界面指出这是他们想要监控的指标。如果对某项指标进行监控,那么除了前项值之外,还可以向用户显示用于衡量或确定指标值的数据的当前“水平”,以及(在一些实施例中)前项值与当前值之间的百分比变化。

[0090] 在一个实施例中,指标监控功能不依赖于自动检索功能。相反,当功能存在时间索引时,用户可获得相同的工具,并且可以“监控”指标。这可能包括实际上未存储在数据库中的指标,如机器学习模型的性能指标值,或模型中不同重要特征的值。用户也可以设置这些值,以便进行监控。

[0091] 正如所公开的,用户可以(例如)根据指标的水平(指标值)和/或指标当前值与前项值之间的百分比变化,指定用于监控指标的“规则”。当提示用户指定规则时,指标监控功能还可以(或作为代替)根据类似的被监控指标推荐规则,其中相似性可通过指标的统计属性、指标名称的语义分析或用户先前指定的指标监控规则中的一个或多个来确定(作为非限制性示例)。

[0092] 这种“推荐”可能包括这样的提示:“平均值变化的建议阈值为2.2%(在5%的观测中会出现这种情况)”。由用户定义或由平台建议的规则的形式取决于数据的结构和值,但通常包括基于以下内容(作为示例)的规则:

●数据值(例如,数据为正、至少为零、为负、大于/大于或等于特定值、或小于/小于或等于特定值);

●数据值的“绝对”变化(例如,数值变化完全为零、数值变化小于/小于或等于特定值,或数值变化的绝对值小于/小于或等于特定值);或

●数据与前项值相比的变化百分比(例如,变化百分比为零,或变化百分比大于特定值)。

[0093] 在一个实施例中,用户可以指定多个规则,并可指定是在“违反”特定规则时通知/警示,还是在“违反”所有规则时通知/警示,其中“违反”规则是指出现或满足规则所指定的条件。也就是说,如果用户设定了规则,要求在某个指标值为负值时对其进行监控,那么只要该指标值为负值,该规则就会被“违反”,即满足了规则中设定的条件。

[0094] 根据规则,平台可显示数值(如果规则基于数值)或数值变化(如果规则基于最近

的数值变化)是否“违反”设定的规则。这种“违反”代表了一种“警报”或通知生成状态,作为回应,平台可以按照用户指定的方式改变值(或值的变化)的显示。如前所述,用户可以选择显示方式的变化,例如为警报状态设置颜色和/或选择与数值或数值变化一同显示的图标。

[0095] 在一个实施例中,指标显示的默认变化是,当规则处于警报状态时(当规则被“违反”时),将值显示(或值的变化,取决于所应用的规则)为红色;当规则未处于警报状态时,将值显示为绿色。没有应用规则时,监控可能会显示默认颜色,默认颜色可能是黑色。用户可以更改这些设置,以及更改用户在平台显示上设置的可访问参数。

[0096] 在一些实施例中,指标监控功能可为用户提供对其尚未熟悉的对象的监控。作为非限制性的例子,团队可能专注于KPI,并为指标监控功能设置了特定规则。由于平台正在采集元数据以及指标之间的关系,因此可能会出现这样的情况:某个不同的指标(或成组的指标),或某个已添加到平台的机器学习模型的性能指标,是受监控指标的“良好”预测信息或主要指示信息。在这种情况下,平台的指标监控功能可能会建议对该指标进行监控,并可根据添加到平台的元数据中的机器学习关系,提供更全面、更完善的监控建议。

[0097] 该功能建立在已公开的平台的内置功能之上。作为通过数据检索(例如,定期查询云数据库服务的元数据检索服务)构建特征图的一部分,平台的软件流程可自动计算不同特征之间的统计关系,并根据校准流程衡量这些关系的相对强度。作为校准过程的一部分,可通过查询确定密切相关的指标,当新添加的指标与当前正在监控的指标密切相关时,可将此信息存储在图中。然后,平台会提示具有相应权限的用户(该权限基于身份),建议他们打开监控模型,并将监控规则应用于新添加的指标。随着时间的推移,校准过程将继续以同样的方式识别新的指标,也可以识别与已监控的成组指标高度相关的现有指标。

[0098] 作为使用案例的非限制性示例,可以考虑以下场景:

“企业”用户可以使用平台追踪由公司领导团队定义并确定为对公司运营和业务战略非常重要的成组的16项核心KPI/指标。该平台与数据库和数据仓库服务的集成可用于更新有关数据集和特征的统计元数据,因此16项核心指标可与定期更新的数据源相连接。公司数据团队的成员可以设置适当的指标监控规则,以便在所追踪的指标达到临界水平或临界增长率时进行追踪并向用户发出警报。

[0099] 利用与这些指标相关的数据计算出的确定的相关性或机器学习模型输出结果可在平台生成的特征图上查看和引导,因此公司核心指标的“地图”将是可查看、可引导和可共享的。企业用户可以定期访问平台,检查核心指标的水平,和/或查看数据团队的工作如何在公司核心指标之间建立额外的(或改进现有的)统计关系。

[0100] 指标监控功能允许用户追踪用来衡量公司运营状况的重要指标,平台特征图允许用户查找指标之间的联系和/或关系。例如,用户可能会选择连接两个指标的UI要素,以发现同行的模型,同行的模型探讨了如何用一个指标来“预测”另一个指标,了解上述关系可以更准确、更可靠地了解运行状况。例如,来自模型和相关性的元数据可以量化订单的平均等待时间与客户向公司再次订购的可能性之间的预测关系,从而改进公司在多个领域(如营销、履行处理或库存管理)的决策。

[0101] 平台公共版本(如通过www.system.com获取)的用户可能会通过浏览他们感兴趣的平台特征图的某一部分来使用指标监控功能。例如,平台的公共版本可能将指标定义为“全球二氧化氮排放量”。该指标可能与NASA发布的测量全球大气排放水平的数据集的一部

分的特征相关联,用户可能将该特征用作监控全球二氧化氮排放量指标的基础。

[0102] 然后,公共平台用户界面将全球二氧化氮排放量作为指标显示,用户可以访问该指标的页面,获取从NASA公布的数据集检索到的元数据所报告的水平或增长变化的信息。当平台建立、创建或发现与其他指标之间的联系时(无论是通过特定的机器学习建模,还是基于数据集的特征与平台上长期追踪的其他特征之间计算出的统计相关性),这些联系将显示在图中。这将使用户能够了解其他指标是否与二氧化氮排放量有关。利用用户界面,用户将能够看到这些相关指标的水平和近期变化,并可使用平台特征图中提供的链接,获取图中显示的关系的统计和/或科学依据(如果需要,还可观察这些关系随时间推移变强或变弱的程度)。

[0103] 在一些实施例中,这些信息可通过HTTP API请求(如gRPC、REST和/或GraphQL请求)提供给其他应用程序。例如,对指标端点的调用将返回平台关于指标的元数据,而对指标/关联端点的调用将返回关于哪些指标与给定指标相关联的元数据(以及有关统计关系的详细信息,例如证实这种关系的证据以及促成这种关系的模型或关联类型)。

[0104] 在一个实施例中,为与指标监控功能相关的指标提供的元数据可包括以下一个或多个元数据:

- ●名称,说明;
- ●创建时间;
- ●更新时间;
- ●创建者;
- ●更新者;
- ●衡量特征;
- ●指标监控状态;
- ●指标监控规则;以及
- ●包括该指标的关联。

当平台配置为提供元数据时,也可提供其他(或更少)元数据。

[0105] 另一个用例是,平台所提供的生成视图或显示的数据可被报道金融市场的数据记者利用。在这种用例中,数据记者可能会查询水平或最近变化超过预定义阈值的指标,然后使用查询来查找相关指标。在对这些查询的回复中所包含的信息将提供统计的背景信息,以说明为什么某个感兴趣的指标处于某个水平(或有特定幅度的变化),并提供统计依据,以说明为什么其他历史相关的指标可能会朝着某个方向移动。例如,数据记者可能会发现,在某一商品市场交易的白银价格出现了大幅下跌--利用白银价格计算出的模型或相关性就会告诉记者,近期(或历史上)有哪些其他市场力量与白银价格的变化相关联,以及市场可能会出现哪些进一步的变化。

[0106] 关于平台的实施和功能的进一步说明如下:

●平台存储具有与特定时间相关的值的特征,例如每周/每月的销售或收入数据、不同国家的GDP年值或不同公开交易股票的每日收盘价。将此类数据添加到平台时,可存储一系列索引值以及值本身,这些索引值与所记录的每个值的具体时间相对应(即存储为时间戳)。当这些值是数字时,可以追踪其水平和变化,因为平台了解如何按时间顺序排列数据,并能计算特定值之间的增长率;

●该平台的数据模型区分了“特征”(数据的集合或成组的衡量值)和“指标”(用户定义的感兴趣的对象,用户希望对其进行衡量和追踪)。例如,如果用户有兴趣衡量一家公司的销售额,可以将“月销售总额”定义为感兴趣的指标;

该指标的值是由公司存储的电子数据记录生成的特征(或特征的变换);

●该平台的架构和功能包括将指标与特征连接成特征图的方法。该平台允许用户指定某个(或多个)特征提供用于确定给定指标的值,从而让其他用户了解该指标是使用所连接的特征进行衡量或评估的。然后,该平台架构可利用从机器学习模型推断出的关系和/或直接从数据中计算出的统计关系(如度量之间的相关性)推断出的关系,在指标和特征之间建立联系;

●已公开的指标监控功能利用平台的这些方面为用户提供指标监控功能和背景信息。监控功能基于从各种来源检索的数据,该数据按照通常存储的基于时间戳的索引进行排列。当可以将此索引运用于来自平台的数据集的特征时,用户将该特征与数值连接/关联到指标,指标的可视化界面将(在一些实施例中)

显示最新值和紧邻的前项值,以及这些值之间的变化百分比;

●指标监控可提供指标的背景信息,因为当模型和数据集添加到平台时,平台会建立指标之间的关系。此外,通用时间戳索引还允许平台自动计算时间序列分析,以生成所追踪的指标之间按照时间维度在统计上稳定的关系。

[0107] 指标监控功能可用于从不同类型来源收集的数据,包括从平台本身生成的数据。举例来说,对于用户定期更新(例如,通过手动更新模型、使用在线机器学习工具或服务自动有计划地更新模型,或从已部署的机器学习模型服务(如AWS Sagemaker)定期更新)的添加到平台的模型,可根据固定的时间间隔收集模型性能指标。这类数据还可以附加到指标中进行监控,并建立追踪模型性能指标与平台上其他衡量指标之间的统计关系(通过相关性分析或显式建模)。这样,平台用户就可以使用指标监控功能,在其他收集到的数据的背景下,管理其模型的性能和指标(因为这些指标通常是数据科学团队的KPI或关键指标)。

[0108] 在一个实施例中,当具有基于时间的索引的数据集或另一个数据块的特征可利用指标监控时,可视化界面的变化或指示(显示数据中最近的水平和百分比变化)可用于通知用户这是可以追踪或监控的数据。可视化界面还可以让用户设置特定的规则,从而以更直观的方式监控这些变化,并接收有关指标值变化的警报和通知。用户可以通过设置规则来配置指标监控功能,这些规则根据使用成组的预定义的比较运算符来比较指标值的最近水平或最近值之间的变化来定义,规则还根据关于在指标“违反”或满足规则所表达的条件时如何以可视化方式显示(以及如何通知用户发生了“违反”)的选项来定义。一旦设置了规则,特征图上的可视化指标就会设置为反映所选颜色或格式(或利用图标为有色觉问题的用户进行标记),从而将受监控的指标与可监控但未设置规则的指标(仍为默认颜色或格式)区分开来。

[0109] 在一个实施例中,作为指标监控的一部分或独立于指标监控,平台可生成可视化,其显示基础特征图随时间的变化或者显示不同来源集之间发生的变化。这可能有助于确定:以前确定的统计关系是否被后来的工作所证实,或者以前认为是有效的关系现在是否应作不同的解释。

[0110] 通过突出显示在用户确定的时间段内发生变化的关系值,该功能对指标监控进行

补充。用户可以利用指标监控快速识别重要的指标及其数值随时间的变化情况,并利用此类功能(例如以可视化形式呈现)来识别关键指标的数值发生变化是否是因为(统计上)密切相关的指标的数值发生了变化,或者潜在的统计关系是否比曾经认为的更强或更弱。这一功能可自动提供给平台用户,取代数据分析师或科学家为应对关键指标变化而进行的探索性建模。

[0111] 在设置规则的流程的一个示例实施例中,根据指标中用于设置监控规则的字段(如当前值、前项值、变化百分比),为用户预先填充默认规则。默认规则可针对使用平台的不同团队进行配置,因为每个企业或团队账户通常都有各自的数据和模型的工作区。这样就可以为每个独立的企业或团队账户分别存储配置设置(包括指标监控规则)。对于企业和团队账户,监控规则通常是根据经验规则的水平设置的(例如,指标的标准规则可能是当某个值的百分比变化绝对值大于或等于5%时发出红色警报)。当账户已为不同的指标设置了指标监控时,平台会建议根据语义相似(即名称、描述或类型相同或十分相似)的指标已有的设置来设置未来的警报。例如,团队可能设置了指标监控规则,当“产品X的库存”的值小于100时显示“黄色”警报--针对该用户或团队的“产品Y的库存”或“产品X的生产”的建议规则可能是设置与“产品X的库存”相同的规则。

[0112] 当指标在统计上相似时,也可建议采用规则。例如,如果由于机器学习模型或其他确定的统计关联,已知“产品X的产量”与“产品X的库存”在统计上相关,则针对“产品X的产量”建议的规则可以与针对相关指标所建议的规则相同,也可以配置为建议与“产品X的库存”的警报集出现的可能性相似的规则。指标监控功能可用于发现或“学习”和应用监控规则,与传统解决方案相比,这种功能更具优势,因为传统解决方案要求孤立地设置规则,而不考虑同一系统中不同指标的背景信息。

[0113] 如前所述,目前用于监控指标或者管理机器学习模型的元数据的解决方案集中在孤立的数据集和模型上。相比之下,本公开的平台架构及其重点是将数据集、模型和其他面向数据的工作中的元数据连接到一处并且连接在特征图中,这意味着指标监控功能并不局限于特定类型的元数据。此外,尽管指标监控是参照数据集中的实际特征的水平或百分比变化进行描述的,但监控功能可应用于平台上所收集的与相应时间要素相关联的其他元数据。

[0114] 虽然元数据管理或数据编目方面的传统方案可能是追踪特定数据集中的所观测的数量,并在该数量发生变化时提供警报或通知,但现有方案并不收集和存储所追踪的不同元数据之间的统计关系。例如,团队可能会追踪“生产中”部署的模型的每日模型性能,同时使用指标监控主动监控(设置适当规则后)5个KPI指标。平台的特征图将显示这5个指标的变化情况,并根据指标的值(或变化)与在指标监控规则中设置的阈值的比较,进行背景的高亮显示(或其他指示)。

[0115] 传统的指标监控方法没有提供足够灵活的监控框架,无法将不同来源的指标(如已部署的机器学习模型生成的模型性能数据和从不同数据源追踪的指标)的变化联系起来。所公开的平台旨在作为整个数据堆栈的知识管理工具,而平台上的指标监控则是监控、警报和背景驱动工具,以用于了解重要指标的变动情况,而这些指标的来源是分布式的。

[0116] 如上所述,在一些实施例中,该平台可对平台可用的元数据进行自动机器学习建模。由于平台上指标的元数据可以索引到相同的时间跨度,因此平台可以“了解”或“学习”

每日模型性能(存储在特征图中)与平台上其他指标之间的统计关系,这些指标是从数据库服务中检索的(或由用户添加的),并且具有时间索引。

[0117] 这一功能可以发现团队目前尚未监控的新的重要指标,和/或提出更有效的指标监控规则,该指标监控规则突出模型成功的关键拐点(例如,通过追踪模型性能指标),或者突出预测其他指标的已知临界值的指标水平/变化。这可以通过在规则设置面板中提出推荐来实现(例如,提出"更好的"规则,并向用户解释平台通过自动机器学习"学习"了什么)。

[0118] 作为这一功能及其对用户的好处的例子,平台可用于获取指标监控数据(其中包含指标是否处于"警报"状态的由时间索引的指示信息),并执行分类模型,在分类模型中其他指标的前项值("滞后"值)用于"预测"给定指标是否处于警报状态。该模型的结果可用于为受监控的指标确定"更好的"阈值(当某个指标的特定水平或变化可以很好地预测其他指标处于"通知"或"警报"状态时,就会出现这种情况),或者如果模型性能指标的水平/变化可以预测其他指标的警报状态(这表明用户可能希望为该模型性能指标设置指标监控)。

[0119] 在一些实施例中,为提高计算效率,平台自动执行的统计比较次数可能会受到限制,以避免突出不准确的相关性。由于平台的元数据包含了正在监控的指标以及平台上使用率较高的指标(无论是在模型中还是在用户的浏览行为中)的相关知识,因此自动规则生成和推荐功能可以专注于平台上关注度相对较高和统计重要性较高的指标和对象。

[0120] 如前所述,在为特定用户或用户组构建特征图之后,可以遍历该图,以确定与研究、模型或调查的主题或目标相关的变量,如果需要,还可以检索支持或证明这些变量相关性的数据集,或衡量相关变量的数据集。请注意,特征图的遍历过程可由两种方法之一控制:(a)用户明确调整搜索参数,或(b)基于算法调整变量/数据检索参数。

[0121] 回到图2(a)和图2(b),如上所述,图2(a)描述了如何实施使用所公开的指标监控功能,从而监控存储在云数据库服务(或"数据仓库"204)中的数据集的特征变化。在图中显示的示例中,数据集元数据206显示了两个统计相关的特征,分别为特征一和特征二。定义第一指标(指标一208),并显示其最新值(209)。显示了管理警报或通知的显示的规则(212),以及由此产生的与"指标一"相关的信息(在显示部分214中展示)。同样,定义第二指标(指标二210),显示其最新值(211),显示管理警报或通知的显示的规则(213),并在显示部分215中展示与指标二有关的结果信息。

[0122] 继续描述支持生成要素或部分202所示显示内容的平台后台处理过程,如要素或部分203所示,数据仓库集成流程220运行以从数据仓库204中"检索"数据集和特征,并计算或访问相关元数据。这一检索过程会向平台的后台API发送http请求以及数据集和特征元数据。元数据包括特征之间的统计关系(如流程222所示)。

[0123] 平台后台会将数据集、特征和关系元数据写入平台的图数据库(如流程224所示)。用户可以在可用网站上查看数据集、特征和关系。当特征具有与数值相关联的时间索引时(如206所示的特征一和特征二示例),且用户将特征一和特征二与指标一(208)和指标二(210)相关联时,用户就可以激活或选择指标监控功能(如流程226所示)。

[0124] 用户可以激活或选择指标监控功能,然后定义监控规则,该监控规则指定(除其他方面外)可视化的警报和设置电子邮件/应用程序通知(如流程228所示)。对此,平台前端提供的指标反映了特征之间的统计关系。根据流程230所述,用户可以看到带有详细元数据和完整统计背景信息(如级别、百分比变化、特征历史、警报和关系)的监控指标。

[0125] 图2(c)至图2(g)是可由平台或系统生成的用户界面显示的示例,根据所公开的平台和系统的实施例,该平台或系统配置为发现或确定并表示指定指标、数据集和机器学习模型之间具有统计意义的关系。

[0126] 图2(c)是用户界面显示的示例,其中显示了最近的数值(314,779)、该数值的百分比变化(-4%)以及变化最大的子群的识别(当指标被定义为表格中数值的集合,而该表格中的数据有多个子群/维度时,可以计算该最大变化)。

[0127] 图2(d)是用户界面显示的示例,其显示了每周活跃用户页面上的指标监控面板,每周活跃用户是已定义的指标。周平均用户数(wau)的数据源已连接并有时间索引,因此可以进行监控。通过选择[+监控]按钮,用户可以设置/定义监控规则,然后指定监控的颜色和电子邮件警报的频率。在图中左侧的平台特征图中,其他指标的指标监控已打开,图中节点之间的边包含描述指标之间统计关系的元数据。知道哪些指标处于警报状态,了解指标之间的关系,用户就能在数据集的背景下了解KPI/关键指标的统计驱动因素。

[0128] 图2(e)是用户界面显示的一个示例,其展示了指标监控的平台目录视图,其中显示页面上的8个指标均已打开。虽然其他数据监控方案的视图在一些方面(或指示板工具的其他图表视图)可能与之类似,但指标监控功能的方法的优势在于每个“卡片”或部分的底部都收集了特定指标的证明依据。每个指标都用于不同的模型(有些是模型的预测结果),点击任何一张卡片都可以查看每个指标的元数据,以及包含在同一机器学习模型或包含在由用户或自动机器学习建立的其他统计关系中的任何指标之间关系的元数据。

[0129] 图2(f)是用户界面显示的示例,其说明了指标监控功能的一个或多个通知。显示最新和最近的数值(以及百分比变化),以及相关指标的数值。这些关系的创建来源于:从添加到平台的机器学习模型中获取的元数据、用户直接添加的关系、应用于特征元数据的自动机器学习,该特征元数据由用户添加,从数据库服务中获取,或从部署在生产中的追踪模型的定期更新中生成。

[0130] 图2(g)是用户界面显示示例,其说明了简化的规则设置对话框。适用于这一指标的条件是,当变化百分比的绝对值严格大于4.5。在本例中,有一个默认颜色区别-百分比变化(73.10%)的绝对值大于4.5%,因此颜色指示为红色。

[0131] 图2(h)是要素、组件或流程的示意图,其示出了根据一些实施例,要素、组件或流程配置为可能存在于由实现方法、流程、功能或操作的计算设备、服务器、平台或系统280中的一个或多个中或者由这些执行。在一些实施例中,所公开的系统和方法可以以一个或多个设备(如作为系统或平台一部分的服务器,或者客户端设备)的形式实现,该设备包括处理元件和成组的可执行的指令。可执行的指令可以是一个(或多个)软件应用程序的一部分,并被布置在软件架构中。

[0132] 一般来说,本公开的实施例可以使用成组的软件指令来实现,这些软件指令旨在由适当编程的处理元件(如GPU、TPU、CPU、微处理器、处理器、控制器或计算设备,作为非限制性示例)执行。在复杂的应用程序或系统中,这些指令通常被排列成“模块”,每个模块通常执行特定的任务、流程、功能或操作。整套模块在运行时可由操作系统(OS)或其他形式的组织平台进行控制或协调。

[0133] 模块和/或子模块可包括适当的计算机可执行的代码或成组指令,例如与编程语言相对应的计算机可执行代码。例如,编程语言源代码可编译成计算机可执行代码。替代地

或附加地,编程语言也可以是解释型编程语言,如脚本语言。

[0134] 如图2(h)所示,系统280可以代表服务器、客户端设备、平台或其他形式的计算或数据处理设备中的一种或多种。模块282各包含成组可执行指令,当这组指令被合适的电子处理器(如图中用"物理处理器298"表示的处理器)执行时,系统(或服务器或设备)280就会运行,以执行特定的流程、操作、功能或方法。

[0135] 模块282可包含一组或多组指令,用于执行参照附图描述的方法或功能,以及说明书中提供的功能和操作说明。这些模块可能包括图示模块,但也可能比图示模块数量更多或更少。此外,模块和模块中包含的成组的计算机可执行指令(全部或部分)可由同一处理器或多个处理器执行。如果由一个以上的处理器执行,协同处理器可以包含在不同的设备中,例如客户端设备中的处理器和服务器中的处理器。

[0136] 模块282存储在内存281中,内存220通常包括操作系统模块284,操作系统模块包含指令,该指令(除其他功能外)用于访问和控制其他模块所含指令的执行。通过"总线"或通信线路290可以访问存储器281中的模块282,以传输数据和执行指令,"总线"或通信线路216还允许处理器298与模块通信,以便访问和执行指令。总线或通信线路290还允许处理器298与系统280的其他元件进行交互,例如输入或输出设备292、用于与系统280外部的设备交换数据和信息的通信元件294以及附加存储设备296。

[0137] 每个模块或子模块可以对应特定的功能、方法、流程或操作,通过执行模块或子模块中的(全部或部分)指令来实现。每个模块或子模块可包含成组的计算机可执行指令,当这些计算机可执行指令被已编程的处理器或协同处理器执行时,会导致处理器或协同处理器(或包含这些指令的设备、装置、服务器)执行特定的功能、方法、流程或操作。如前所述,包含处理器或协同处理器的设备可以是客户端设备或远程服务器或平台中的一种或两种。因此,模块可能包含由客户端设备、服务器或平台或以上组合(全部或部分)执行的指令。这些功能、方法、流程或操作可包括用于实现所公开系统和方法的一个或多个方面的功能、方法、流程或操作,例如:

●创建由节点和边组成的特征图(如模块284所示),其中;

○节点代表一个或多个概念、主题、数据集、元数据、模型、指标、变量、可衡量的量、对象、特性、特征或因素(作为非限制性示例);

○边代表第一节点和第二节点之间的关系,例如统计意义上的关系、依赖关系或层级关系(作为非限制性示例);以及

○与边相关联的标签,该标签可指示边所连接的两个节点之间关系的一方面,如两个节点之间的关系所基于的元数据,或支持两个节点之间具有统计意义的关系的数据集(作为非限制性示例);

●为用户提供用户界面显示、工具、特征和可选要素,使用户能够执行以下一项或多项功能或操作(如模块286所示):

○确定监控或追踪的感兴趣的指标(如KPI);

○定义规则,该规则规定何时应生成有关已确定的指标的行为的警报;

○定义如何在用户界面显示上识别或指示应用该规则的结果;

○允许用户选择已生成警报的指标,并作为响应,提供有关指标值随时间变化的信息、已满足或已激活的导致警报的规则、指标与其他指标的关系(如果相关),以及关于用

于生成指标的数据集、机器学习模型、规则或其他因素的可用信息(作为非限制性示例);

●为用户提供关于以下内容的推荐:可能具有监控价值的不同的指标或指标集、可能有助于检查的数据集、可能与确定的指标相关的元数据或用户可能感兴趣的基础数据或指标的其它方面;

○其中,推荐可能(至少部分)来自经过训练的机器学习模型所生成的输出、统计分析、研究,或其他形式的评估。

[0138] 在一些实施例中,本文所公开的系统和方法提供的功能和服务可通过访问账户提供给多个用户,该账户由服务器或服务平台维护。这种服务器或服务平台可以是称为软件即服务(SaaS)的一种形式。图3是示出SaaS系统的示意图,实施例可以在该系统中实施。图4是示意图,其说明了可实施实施例的示例操作环境的元件或组件。图5是说明图4多租户分布式计算服务平台的元件或组件的其他细节的示意图,在其中可以实施实施例。

[0139] 在一些实施例中,本文所描述或公开的系统或服务可实施为响应用户的回应而的提交而执行的微服务、流程、工作流或功能。微服务、流程、工作流或功能可由服务器、数据处理元件、平台或系统执行。在一些实施例中,数据分析和其他服务可由位于“云”中的服务平台提供。在这种情况下,可通过API和SDK访问平台。功能、流程和能力可作为平台内的微服务提供。微服务的接口可由REST和GraphQL端点定义。管理控制台可允许用户或管理员安全地访问基础请求和响应数据、管理账户和访问,并在一些情况下,修改处理的工作流程或配置。

[0140] 需要注意的是,虽然图3至图5展示的多租户或SaaS架构可用于向多个账户/用户提供业务相关的应用和服务或其他应用和服务,但这种架构也可用于提供其他类型的数据处理服务,并提供对其他应用的访问。虽然在一些实施例中,图3至图5中所示类型的平台或系统可由第三方提供商运营,以提供成组特定的业务相关应用,但在其他实施例中,平台可由提供商运营,不同的企业可通过平台为用户提供应用或服务。

[0141] 图3是示出系统300的示意图,在该系统中可以实施实施例,或者通过该系统可以访问所公开或所描述的服务的实施例。根据由应用服务提供商(ASP)运营的业务服务系统(如多租户数据处理平台)的优势,本文所述服务的用户可包括个人、企业、商店、组织等。用户可以使用任何合适的客户端访问服务,例如包括但不限于台式电脑、笔记本电脑、平板电脑、扫描仪、智能手机或专用VR头戴设备等。用户通过互联网308或其他合适的通信网络或网络组合与服务平台进行交互。合适的客户端设备包括台式电脑303、智能手机304、平板电脑或笔记本电脑305。

[0142] 系统310可以由第三方运营,系统310可以包括成组服务,帮助用户访问本文所述的数据处理和指标监控服务312,以及包括网络接口服务器314,如图3所示耦合。可以理解的是,服务312和网络接口服务器314可以在一个或多个不同的硬件系统和组件上实现,即使在图3中硬件系统和组件表示为单个单元。服务312可包括一个或多个功能或操作,使用户能够访问特征图并执行本文所公开的指标监控功能。

[0143] 举例来说,在一些实施例中,通过平台310提供的成组的功能、操作或服务可包括:

●账户管理服务318,例如

○认证用户的流程或服务(与使用客户端设备提交用户凭据一起);

○用于生成服务或应用程序的载体或实例的流程或服务,该服务或应用程序将提

供给用户;

●特征图生成服务320,例如

○用于生成或访问所公开的特征图的流程或服务,特征图由成组节点和连接一些节点的边组成;

●用户界面显示和工具生成服务322,如

○流程或服务,其用于生成一个或多个用户界面显示以及用户界面工具和要素,使用户能够:

■确定监控或追踪的感兴趣的指标(如KPI);

■定义规则,该规则规定何时应生成有关已确定的指标的行为的警报;

■定义如何在用户界面显示上识别或指示应用该规则的结果;

○允许用户选择已生成警报的指标,并作为响应,提供有关指标值随时间变化的信息、已满足或已激活的导致警报的规则、指标与其他指标的关系(如果相关),以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的可用信息(作为非限制性示例);

●推荐生成服务324,例如

○流程或服务,其为用户提供关于以下内容的推荐:可能具有监控价值的不同的指标或指标集、可能有助于检查的数据集、可能与确定的指标相关的元数据或用户可能感兴趣的基础数据或指标的其它方面;

●管理服务326,如

○流程或服务,该流程或服务使服务和/或平台的提供者能够管理和配置向用户提供的流程和服务,例如,作为非限制性示例,通过改变用户数据的建模方式、指标的计算方式,或由此产生的指标和建议给特定用户的呈现方式。

[0144] 请注意,除所列操作或功能外,应用模块或子模块还可包含计算机可执行指令,当这些指令被编程处理器执行时,会导致系统或设备执行与服务平台操作相关的功能。此类功能可包括但不限于与以下内容相关的功能:用户注册、用户账户管理、账户间数据安全、数据处理和/或存储能力分配、提供对SystemDB以外数据源(如本体论或参考资料)的访问。

[0145] 图3所示的平台或系统可以运营在分布式计算系统上,该分布式计算系统由至少一台,但很可能是多台“服务器”组成。服务器是物理计算机,专门为一个或多个软件应用程序或服务提供数据存储和执行环境,以满足与服务器进行数据通信(例如通过互联网等公共网络)的其他计算机用户的需求。服务器及其提供的服务可称为“主机”,远程计算机和远程计算机上运行的软件应用程序可称为“客户端”。根据服务器提供的计算服务,服务器可以被称为数据库服务器、数据存储服务器、文件服务器、邮件服务器、打印服务器或网络服务器等。网络服务器通常是硬件和软件的组合,通常通过运营网站向通过互联网访问网络服务器的客户端网络浏览器提供内容。

[0146] 图4是示意图,其说明了可实施实施例的示例操作环境400的元件或组件。如图所示,集成和/或并入各种计算设备的各种客户端402可以通过一个或多个网络414与多租户服务平台408通信。例如,客户端可以包含和/或并入至少部分由一个或多个计算设备实现的客户端应用程序(即软件)。合适的计算设备包括个人电脑、服务器电脑404、台式电脑406、笔记本电脑407、笔记本电脑、平板电脑或个人数字助理(PDA)410、智能电话412、手机以及包含一个或多个计算设备组件(如一个或多个电子处理器、微处理器、中央处理器

(CPU)或控制器)的消费电子设备。合适的网络414包括利用有线和/或无线通信技术的网络,以及按照任何合适的网络和/或通信协议(如互联网)运行的网络。

[0147] 分布式计算服务/平台(也可称为多租户数据处理平台)408可包括多个处理层,其中包括用户界面层416、应用服务器层420和数据存储层424。用户界面层416可以维护多个用户界面417,用户界面417包括图形用户界面和/或基于网络的界面。用户界面可包括服务的默认用户界面,用于为服务的用户或"租户"提供对应用程序和数据的访问(在图中表示为"服务UI"),以及一个或多个根据用户特定需求专门化/定制的用户界面(例如,在图中表示为"租户AUI"、...、"租户ZUI",其可通过一个或多个API访问)。

[0148] 默认用户界面可包括用户界面组件,用户界面组件使租户能够管理租户对由服务平台提供的功能和能力的访问和使用。这可能包括访问租户数据、启动特定应用程序的实例化、执行特定数据处理操作等。图中所示的每个应用服务器或处理层422可通过成组的计算机和/或组件(包括计算机服务器和处理器)来实现,并可执行各种功能、方法、流程或操作,各种功能、方法、流程或操作由软件应用程序或成组的指令的执行来决定。数据存储层424可包括一个或多个数据存储,其中可包括服务数据存储425和一个或多个租户数据存储426。数据存储可采用任何合适的数据存储技术,包括基于结构化查询语言(SQL)的关系数据库管理系统(RDBMS)。

[0149] 服务平台408可以是多租户的,可以由实体运营,为多个租户提供一系列业务相关的或其他的数据处理应用、数据存储和功能。例如,应用程序和功能可包括提供基于网络的对功能的访问,以便企业向终端用户提供服务,从而允许用户使用浏览器和互联网或内部网连接来查看、输入、处理或修改一些类型的信息。这些功能或应用程序通常由一个或多个软件代码/指令模块实现,这些模块由一个或多个服务器422维护和执行,这些服务器是平台应用服务器层420的一部分。图4所示的平台或系统可以运营在分布式计算系统上,该分布式计算系统由至少一台,但很可能是多台"服务器"组成。

[0150] 如前所述,企业可以利用第三方提供的系统,而不是自己建立和维护这样一个平台或系统。第三方可在多租户平台的背景下实施如上所述的业务系统/平台,向用户提供企业数据处理工作流的各个实例,每个企业代表平台的一个租户。这种多租户平台的一个优势是,每个租户都能根据自己的具体业务需求或操作方法定制数据处理工作流的实例。每个租户可以是使用多租户平台向多个用户提供业务服务和功能的企业或实体。

[0151] 图5是示出图4多租户分布式计算服务平台的元件或组件的其他细节的示意图,在其中可以实施实施例。图5所示的软件架构是架构的一个示例,可用来实现本公开的实施例。一般来说,本申请的实施例可以使用成组的软件指令来实现,这些软件指令旨在由适当编程的处理元件(如CPU、GPU、微处理器、处理器、控制器或计算设备)执行。在复杂的系统中,这些指令通常被排列成"模块",每个模块执行特定的任务、流程、功能或操作。整套模块在运行时可由操作系统(OS)或其他形式的组织平台进行控制或协调。

[0152] 如上所述,图5是示出多租户分布式计算服务平台的元件或组件500的其他细节的示意图,在其中可以实施实施例。示例架构包括具有一个或多个用户界面503的用户界面层502。这类用户界面的例子包括图形用户界面和应用编程界面(API)。每个用户界面可包括一个或多个界面要素504。例如,用户可以与界面要素交互,访问示例架构的应用层和/或数据存储层提供的功能和/或数据。图形用户界面要素包括按钮、菜单、复选框、下拉列表、滚

动条、滑块、旋转器、文本框、图标、标签、进度条、状态条、工具栏、窗口、超链接和对话框。应用程序接口可以是本地的,也可以是远程的,可包括接口要素,接口要素例如各种控件、参数化过程调用、程序对象和消息传递协议。

[0153] 应用程序层510可包括一个或多个应用程序模块511,每个应用程序模块有一个或多个子模块512。每个应用程序模块511或子模块512可对应由模块或子模块实现的功能、方法、流程或操作(例如,与向平台用户提供数据处理和服务有关的功能或流程)。这些功能、方法、流程或操作可包括用于实现所公开系统和方法的一个或多个方面的功能、方法、流程或操作,例如用于本文公开或描述的一个或多个过程、功能或操作:

[0154] 应用程序模块和/或子模块可包括任何合适的计算机可执行代码或成组指令(例如,由适当编程的处理器、微处理器、GPU、TPU、或CPU),如与编程语言相对应的计算机可执行代码。例如,编程语言源代码可编译成计算机可执行代码。替代地或附加地,编程语言也可以是解释型编程语言,如脚本语言。每个应用程序服务器(例如,由图4中的元素422表示)可包括每个应用模块。另外,不同的应用程序服务器可以包含不同的成组的应用程序模块。这些集合可能是不相交的,也可能是重叠的。

[0155] 数据存储层520可包括一个或多个数据对象522,每个数据对象具有一个或多个数据对象部分521,如属性和/或行为。例如,数据对象可以与关系数据库的表相对应,数据对象部分可以与这些表的列或栏相对应。替代地或附加地,数据对象也可以对应于具有栏和相关服务的数据记录。替代地或附加地,数据对象也可以是程序数据对象的持久实例,如结构和类。数据存储层中的每个数据存储可包括每个数据对象。另外,不同的数据存储可能包括不同的成组的数据对象。这些集合可能是不相交的,也可能是重叠的。

[0156] 请注意,图3至图5中描述的示例计算环境并非限制性示例。可以全部或部分实施本公开实施例的其他环境包括设备(包括移动设备)、软件应用程序、系统、装置、网络、SaaS平台、IaaS(基础设施即服务)平台或其他可配置组件,这些组件可被多个用户用于数据录入、数据处理、应用程序执行或数据审查。

[0157] 本公开包括以下条款和实施例:

1.一种监控一个或多个指标的方法,其包括:

构建或访问特征图,所述特征图包括成组的节点和成组的边,其中成组的边中的每条边将成组的节点中的节点与一个或多个其他节点连接,并且其中每个节点代表被发现与主题有统计关联的变量,每条边代表节点与主题之间或者第一节点与第二节点之间的统计关联;

生成用户界面显示和用户界面工具,使用户能够执行以下一项或多项操作:

识别监控的指标;

定义规则,所述规则规定何时应生成有关已识别的指标的行为的警报;

定义如何在用户界面显示上指示应用所述规则的结果;

允许用户选择已生成警报的指标,并作为响应,提供有关指标的值随时间的一个或多个变化的信息、导致警报或通知的规则、所述指标与其他指标的关系,以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的信息。

2.根据条款1所述的方法,其还包括为用户提供关于以下内容的一项或多项的推荐:待监控的不同的指标或成组的指标、可能有助于检查的数据集、可能与指标相关的元数

据,或者基础数据或指标的某个方面。

3.根据条款1所述的方法,其中构建特征图还包括:

访问一个或多个来源,其中每个来源包括关于来源中讨论的主题与讨论的主题中考虑的一个或多个变量之间的统计关联的信息;

处理从每个来源访问的信息,以识别所考虑的一个或多个变量,并针对每个变量,识别有关变量与主题之间的统计关联的信息;以及

将处理所访问的一个或多个来源的结果存储在数据库中,所存储的结果包括针对于每个来源,对一个或多个变量中每一个变量的引用、对主题的引用以及关于每个变量与主题之间的统计关联的信息。

4.根据条款3所述的方法,其还包括存储能够访问数据集的要素,其中数据集包括用于证明每个变量与主题之间的统计关联的数据或者代表一个或多个变量的衡量的数据。

5.根据权利要求4所述的条款,其还包括:

遍历特征图,以识别与一个或多个变量相关的一个或多个数据集,所述一个或多个变量在统计上与用户感兴趣的主题相关,或在统计上与感兴趣的主题在语义上相关;

对识别的一个或多个数据集进行筛选和排序;并且

向用户展示已识别的数据集的筛选和排序结果。

6.根据条款3所述的方法,其中一个或多个来源包括至少一个包含专有数据的来源。

7.根据条款6所述的方法,其中所述专有数据来自业务、研究或实验。

8.根据条款1所述的方法,其中所述推荐由经过训练的模型或统计分析中的一个或多个生成。

9.一种系统,其包括:

一个或多个电子处理器,其配置为执行成组的计算机可执行的指令;以及

包含所述成组的计算机可执行的指令的一个或多个非暂存计算机可读介质,其中,当所述指令被执行时,所述指令使一个或多个电子处理器或包括处理器的设备或装置执行:

构建或访问特征图,所述特征图包括成组的节点和成组的边,其中成组的边中的每条边将成组的节点中的节点与一个或多个其他节点连接,并且其中每个节点代表被发现与主题有统计关联的变量,每条边代表节点与主题之间或者第一节点与第二节点之间的统计关联;

生成用户界面显示和用户界面工具,使用户能够执行以下一项或多项操作:

识别监控的指标;

定义规则,所述规则规定何时应生成有关已识别的指标的行为的警报;

定义如何在用户界面显示上指示应用所述规则的结果;

允许用户选择已生成警报的指标,并作为响应,提供有关指标的值随时间的一个或多个变化的信息、导致警报或通知的规则、指标与其他指标的关系,以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的信息。

10.根据条款9所述的系统,其中所述指令使一个或多个电子处理器或包括处理器的设备或装置为用户提供关于以下一个或多个内容的一项或多项的推荐:待监控的不同的

指标或成组的指标、可能有助于检查的数据集、可能与指标相关的元数据,或者基础数据或指标的某个方面。

11.根据条款9所述的系统,其中构建特征图还包括:

访问一个或多个来源,其中每个来源包括关于来源中讨论的主题与讨论的主题中考虑的一个或多个变量之间的统计关联的信息;

处理从每个来源访问的信息,以识别所考虑的一个或多个变量,并且针对每个变量,识别有关变量与主题之间的统计关联的信息;以及

将处理所访问的一个或多个来源的结果存储在数据库中,所存储的结果包括针对于每个来源,对一个或多个变量的每一个变量的引用、对主题的引用以及关于每个变量与主题之间的统计关联的信息。

12.根据条款11所述的系统,其还包括存储能够访问数据集的要素,其中所述数据集包括用于证明每个变量与主题之间的统计关联的数据或者代表一个或多个变量的衡量的数据。

13.根据条款12所述的系统,其中所述指令使一个或多个电子处理器或包括处理器的装置或设备执行:

遍历特征图,以识别与一个或多个变量相关的一个或多个数据集,所述一个或多个变量在统计上与用户感兴趣的主题相关,或在统计上与感兴趣的主题在语义上相关;

对识别的一个或多个数据集进行筛选和排序;并且

向用户展示已识别的数据集的筛选和排序结果。

14.根据条款11所述的系统,其中一个或多个来源包括至少一个包含专有数据的来源,而且,其中专有数据是从业务、研究或实验中获得的。

15.一个或多个非暂存计算机可读介质,其包含成组的计算机可执行的指令,当所述指令通过一个或多个编程的电子处理器执行时,所述指令使处理器或包括处理器的设备或装置执行:

构建或访问特征图,所述特征图包括成组的节点和成组的边,其中成组的边中的每条边将成组的节点中的节点与一个或多个其他节点连接,并且其中每个节点代表被发现与主题有统计关联的变量,每条边代表节点与主题之间或者第一节点与第二节点之间的统计关联;并且

生成用户界面显示和用户界面工具,使用户能够执行以下一项或多项操作:

识别监控的指标;

定义规则,所述规则规定何时应生成有关已识别的指标的行为的警报;

定义如何在用户界面显示上指示应用所述规则的结果;

允许用户选择已生成警报的指标,并作为响应,提供有关指标的值随时间的一个或多个变化的信息、导致警报或通知的规则、指标与其他指标的关系,以及关于用于生成指标的数据集、机器学习模型、规则或其他因素的信息。

16.根据条款15所述的非暂存计算机可读介质,其中所述指令使一个或多个电子处理器或包括处理器的设备或装置为用户提供关于以下一个或多个内容的一项或多项的推荐:待监控的不同的指标或成组的指标、可能有助于检查的数据集、可能与指标相关的元数据,或者基础数据或指标的某个方面。

17.根据条款15所述的非暂存计算机可读介质,其中构建特征图还包括:

访问一个或多个来源,其中每个来源包括关于来源中讨论的主题与讨论的主题中考虑的一个或多个变量之间的统计关联的信息;

处理从每个来源访问的信息,以识别所考虑的一个或多个变量,并针对每个变量,识别有关变量与主题之间的统计关联的信息;以及

将处理所访问的一个或多个来源的结果存储在数据库中,所存储的结果包括针对于每个来源,对一个或多个变量中每个变量的引用、对主题的引用以及关于每个变量与主题之间的统计关联的信息。

18.根据条款17所述的非暂存计算机可读介质,其还包括存储能够访问数据集的要素,其中所述数据集包括用于证明每个变量与主题之间的统计关联的数据或者代表一个或多个变量的衡量的数据。

19.根据条款18所述的非暂存计算机可读介质,其中所述指令使一个或多个电子处理器或包括处理器的装置或设备执行:

遍历特征图,以识别与一个或多个变量相关的一个或多个数据集,所述一个或多个变量在统计上与用户感兴趣的主题相关,或在统计上与感兴趣的主题在语义上相关;

对识别的一个或多个数据集进行筛选和排序;并且

向用户展示已识别的数据集的筛选和排序结果。

20.根据条款17所述的非暂存计算机可读介质,其中一个或多个来源包括至少一个包含专有数据的来源,而且,其中专有数据是从业务、研究或实验中获得的。

[0158] 可以使用以模块化或集成化的方式的计算机软件,以控制逻辑的形式实施所公开的系统和方法。根据本文的公开和教导,本领域一般技术人员将知道并理解使用硬件以及软件和硬件的结合实现本发明的其他方式和/或方法。

[0159] 机器学习(ML)正被越来越多地用于多个行业的数据分析和辅助决策。为了从机器学习中获益,需要将机器学习算法应用于成组的训练数据和标签,以生成“模型”,“模型”代表算法应用从训练数据中“学习”到的内容。成组的训练数据中的每个要素(或实例或示例,以一个或多个参数、变量、特性或“特征”的形式)都与标签或注释相关联,该标签或注释定义了训练模型应如何对该要素进行分类。神经网络形式的机器学习模型是成组由多层神经元连接而成的网络,这些神经元通过运作对输入数据样本做出决策(如分类)。训练完成后(即关联神经元的权重已趋于稳定或在可接受的变化范围内),模型将对新的输入数据要素进行运算,生成正确的标签或分类作为输出。

[0160] 在一些实施例中,本文所述的一些方法、模型或功能可以以训练过的神经网络的形式体现,其中网络是通过执行成组的计算机可执行指令或通过表现数据结构来实现的。这些指令可存储在非暂存计算机可读介质中(或其上),并由编程的处理器或处理元件执行。成组指令(如通过网络,如互联网)可通过指令传输传达给用户,或者通过执行成组指令的应用程序传达给用户。最终用户可通过访问SaaS平台或通过此类平台提供的服务来使用成组指令或应用程序。可以使用训练过的神经网络、训练过的机器学习模型或任何其他形式的决策或分类过程来实现本文所述的一种或多种方法、功能、流程或操作。请注意,神经网络或深度学习模型能够以数据结构的形式表征,其中存储的数据表现为成组包含节点的层,在不同层的节点之间建立(或形成)连接,连接对输入进行处理,以提供决策或值作为输

出。

[0161] 一般来说,神经网络可被视为由相互连接的人工“神经元”或节点组成的系统,这些“神经元”或节点可相互交换信息。这些连接具有数字权重,在训练过程中对其进行“调整”,这样,经过适当训练的网络就能(例如)在出现需要识别的图像或模式时做出正确反应。在这一特性下,网络由多层特征检测“神经元”组成;每一层的神经元都对在先前层输入的不同组合做出响应。网络的训练是通过使用“标注的”的输入的数据集进行的,该输入在各种具有代表性的输入模式下,与其预期的输出响应相关联。训练使用通用目的方法迭代确定中间和最终特征神经元的权重。就计算模型而言,每个神经元计算输入和权重的点积,加上偏差,并应用非线性触发或激活函数(例如,使用sigmoid响应函数)。

[0162] 本申请中描述的任何软件组件、流程或功能都可以作为由处理器执行的软件代码来实现,使用任何合适的计算机语言(如Python、Java、JavaScript、C、C++或Perl)的传统或面向对象技术。软件代码可作为一系列指令或命令存储在非暂存计算机可读介质中(或其上),非暂存计算机可读介质如随机存取存储器(RAM)、只读存储器(ROM)、磁性介质(如硬盘)或光学介质(如光盘)。在这种情况下,非暂存计算机可读介质几乎是指除暂态波形之外的任何适于存储数据或指令集的介质。任何此类计算机可读介质都可以存在于单个计算设备上或设备内,也可以存在于系统或网络内的不同计算设备上或设备内。

[0163] 根据一个实施示例,本文使用的术语处理元件或处理器,可以是中央处理器(CPU),或者是符合CPU的概念(如虚拟机)。在本实施例中,CPU或CPU集成在其中的设备可与一个或多个外设(如显示器)耦合、连接和/或通信。在另一个实施示例中,处理元件或处理器可以集成到移动计算设备中,例如智能手机或平板电脑。

[0164] 本文提及的非暂态计算机可读存储介质可包括多个物理驱动器单元,如独立硬盘冗余阵列(RAID)、闪存、USB闪存驱动器、外置硬盘驱动器、拇指驱动器、笔式驱动器、钥匙驱动器、高密度数字多功能光盘(HD-DVD)光驱、内置硬盘驱动器、蓝光光盘驱动器或全息数字数据存储(HDDS)光驱、同步动态随机存取存储器(SDRAM)或类似设备或基于类似技术的其他形式的存储器。这种计算机可读存储介质允许处理元件或处理器访问存储在可移动和不可移动的存储介质上的计算机可执行的流程步骤、应用程序等,以便从设备卸载数据或将数据上传到设备。如前所述,就本文所述的实施例而言,非暂存计算机可读介质几乎可以包括除暂态波形或类似介质之外的任何结构、技术或方法。

[0165] 本文参照系统框图和/或功能、操作、流程或方法的流程图或流程图表,对所公开技术的一些实施方案进行了描述。可以理解的是,框图中的一个或多个块,或流程图或流程图表中的一个或多个阶段或步骤,以及框图中的块和流程图或流程图表中的阶段或步骤的组合,都可以分别通过计算机可执行程序指令来实现。请注意,在一些实施例中,一个或多个块或阶段或步骤不一定需要按照所提出的顺序执行,或者根本不一定需要执行。

[0166] 这些计算机可执行程序指令可加载到通用计算机、专用计算机、处理器或其他可编程数据处理设备上,以生成机器的具体示例,从而使由计算机、处理器或其他可编程数据处理设备执行的指令产生用于实现本文所述的一种或多种功能、操作、流程或方法的手段。这些计算机程序指令也可存储在计算机可读存储器中,该计算机程序指令可引导计算机或其他可编程数据处理设备以特定方式运行,从而使存储在计算机可读存储器中的指令产生制造品,该制造品包括实现本文所述的一种或多种功能、操作、流程或方法的指令装置。

[0167] 虽然已结合目前被认为是最实用和最多样的实施例对所公开技术的一些实施例进行了描述,但应该理解的是,所公开的技术并不局限于所公开的实施例。相反,所公开的实施例旨在涵盖所附权利要求范围内的各种修改和等效布置。尽管本文使用了一些特定术语,但它们只是一般的和描述性的,并不用于限制目的。

[0168] 本书面说明使用示例公开了所公开技术的一些实施例,并使本领域的任何技术人员能够实践所公开技术的一些实例,包括制造和使用任何装置或系统以及执行任何并入的方法。所公开技术的一些实施例的专利范围已在权利要求书中限定,并可包括本领域技术人员想到的其他示例。如果这些示例的结构和/或功能要素与权利要求书的文字表述没有差异,或者如果这些示例包括的结构和/或功能要素与权利要求书的文字表述没有实质性差异,那么这些示例就属于权利要求书的范围。

[0169] 本文引用的所有参考文献,包括出版物、专利申请和专利,均以引用的方式并入本文,相当于各参考文献均单独并特别注明作为参考并入本文和/或全文载入本文。

[0170] 在本说明书和以下权利要求书中使用的术语“一”、“一个”和“这个”以及类似的指代,应解释为包括单数和复数,除非本文另有说明或与上下文明显矛盾。除非另有说明,本说明书和以下权利要求书中的术语“具有”、“包括”、“含有”及类似指代应理解为开放式术语(例如,意为“包括但不限于”)。除非本文另有说明,否则本文中对数值范围的叙述仅仅是作为一种简记方法,用于单独提及落入该范围内的每个单独的数值,并且每个单独的数值都被纳入说明书中,如同在本文中单独叙述一样。除非本文另有说明或与上下文明显矛盾,否则本文所述的所有方法可以任何适当的顺序执行。本文提供的任何及所有示例或示例性语言(如“如”)的使用,只是为了更好地阐明本发明的实施例,并不构成对本发明范围的限制,除非另有要求。说明书中的任何表述都不应被解释为表明任何非要求的要素对于本发明的每个实施例都是必不可少的。

[0171] 本文(即权利要求书、附图和说明书)中使用的术语“或”包括备选项目和组合项目。

[0172] 图中描绘或本文描述的组件的不同布置,以及未显示或未描述的组件和步骤都是可能的。同样,一些特征和子组合是有用的,可以在不参考其他特征和子组合的情况下使用。所描述的实施例是为了说明而非处于限制性的目的,替代的实施例对于本说明书的读者来说是显而易见的。因此,本发明的实施例并不局限于上述或附图中描述的实施例,在不脱离下述权利要求范围的情况下,还可以做出各种实施例和修改。

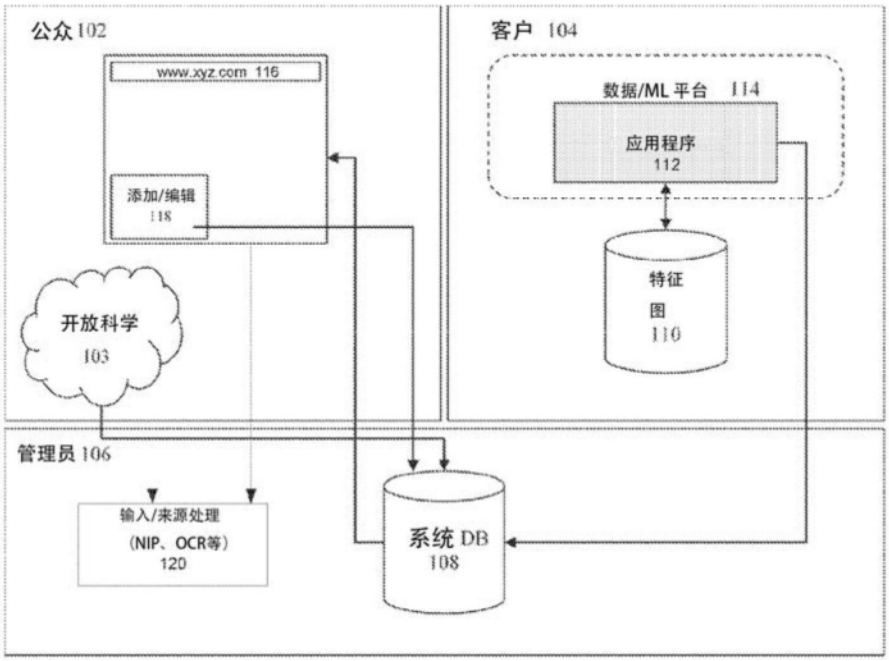

公众 102
www.xyz.com 116
添加/编辑
118
开放科学
103
客户 104
数据/ML 平台 114
应用程序
112
特征
图
110
管理员 106
输入/来源处理
(NIP、OCR等)
120
系统 DB
108
100

图1(a)

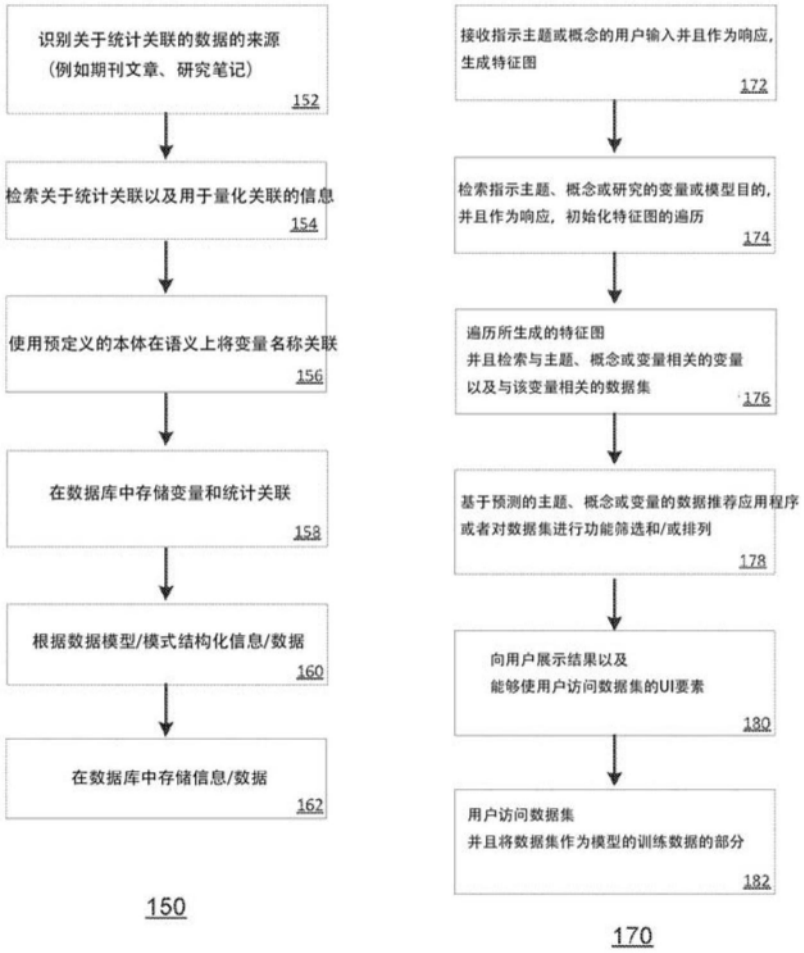

识别关于统计关联的数据的来源
（例如期刊文章、研究笔记）
152
检索关于统计关联以及用于量化关联的信息
154
使用预定义的本体在语义上将变量名称关联
156
在数据库中存储变量和统计关联
158
根据数据模型/模式结构化信息/数据
160
在数据库中存储信息/数据
162
150
接收指示主题或概念的用户输入并且作为响应，
生成特征图
172
检索指示主题、概念或研究的变量或模型目的，
并且作为响应，初始化特征图的遍历
174
遍历所生成的特征图
并且检索与主题、概念或变量相关的变量
以及与该变量相关的数据集
176
基于预测的主题、概念或变量的数据推荐应用程序
或者对数据集进行功能筛选和/或排列
178
向用户展示结果以及
能够使用户访问数据集的UI要素
180
用户访问数据集
并且将数据集作为模型的训练数据的部分
182
170

图1(b)

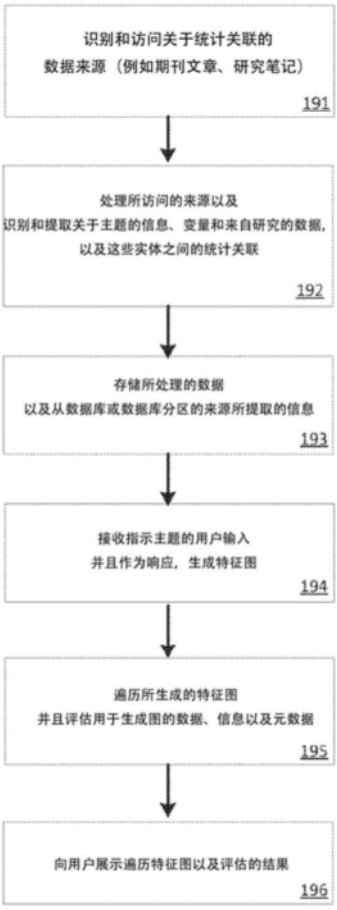
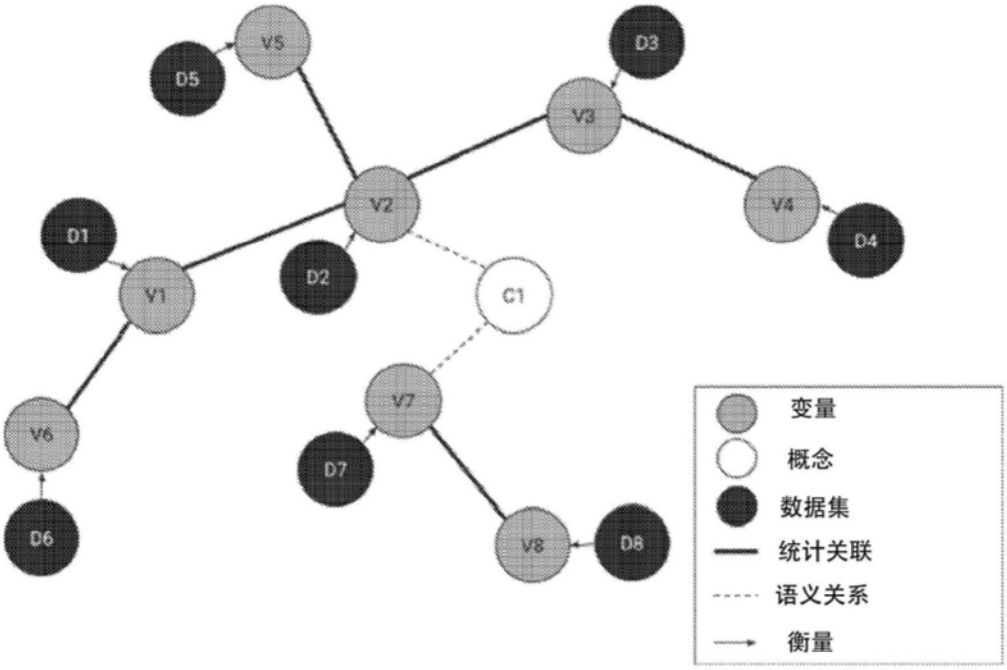

识别和访问关于统计关联的
数据来源（例如期刊文章、研究笔记）
191
处理所访问的来源以及
识别和提取关于主题的信息、变量和来自研究的数据，
以及这些实体之间的统计关联
192
存储所处理的数据
以及从数据库或数据库分区的来源所提取的信息
193
接收指示主题的用户输入
并且作为响应，生成特征图
194
遍历所生成的特征图
并且评估用于生成图的数据、信息以及元数据
195
向用户展示遍历特征图以及评估的结果
196
190

图1(c)


V5
D3
D5
V3
V2
V4
D1
D4
D2
C1
V1
V7
V6
D7
D6
V8
D8
变量
概念
数据集
统计关联
语义关系
衡量

198

图1(d)

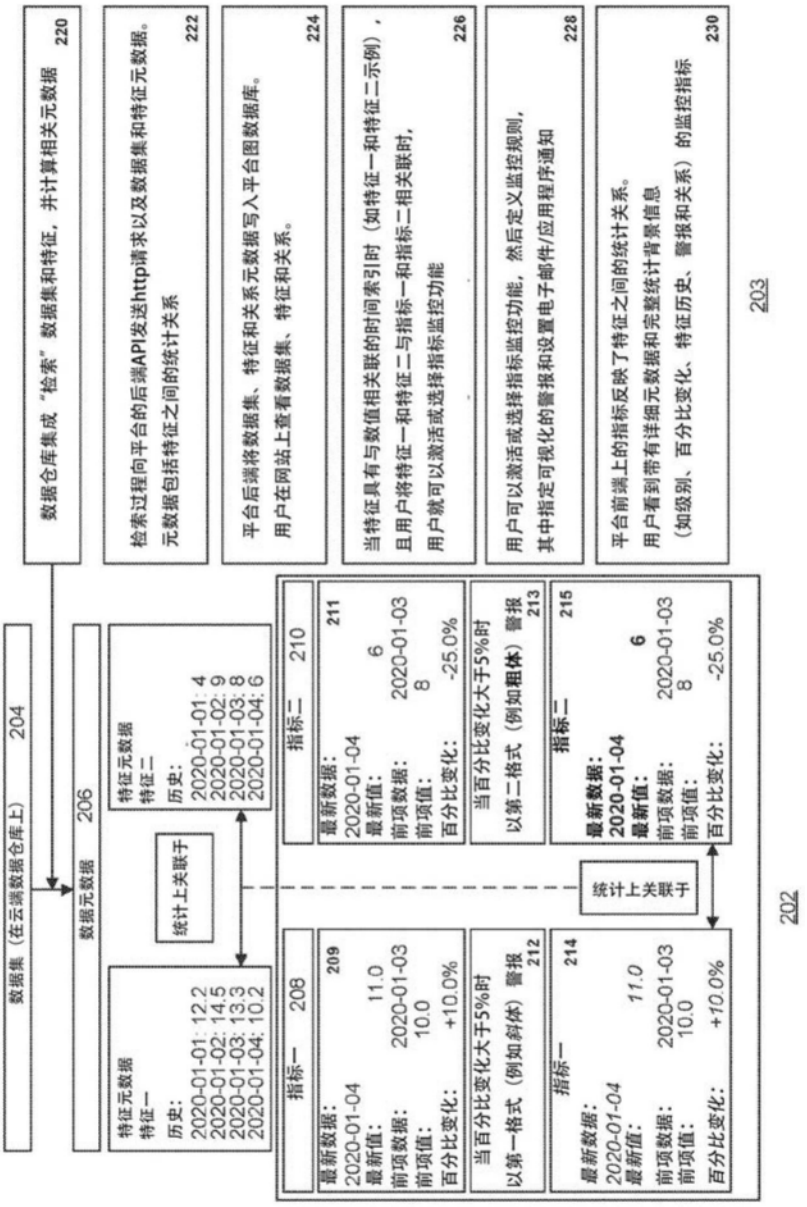

数据集（在云端数据仓库上） 204
数据仓库集成“检索”数据集和特征，并计算相关元数据 220
检索过程向平台的后端API发送http请求以及数据集和特征元数据。
元数据包括特征之间的统计关系 222
平台后端将数据集、特征和关系元数据写入平台图数据库。
用户在网站上查看数据集、特征和关系。 224
当特征具有与数值相关联的时间索引时（如特征一和特征二示例），
且用户将特征一和特征二与指标一和指标二相关联时，
用户就可以激活或选择指标监控功能 226
用户可以激活或选择指标监控功能，然后定义监控规则，
其中指定可视化的警报和设置电子邮件/应用程序通知 228
平台前端上的指标反映了特征之间的统计关系。
用户看到带有详细元数据和完整统计背景信息
（如级别、百分比变化、特征历史、警报和关系）的监控指标 230
203
数据元数据 206
特征元数据
特征一
历史：
2020-01-01: 12.2
2020-01-02: 14.5
2020-01-03: 13.3
2020-01-04: 10.2
统计上关联于
特征元数据
特征二
历史：
2020-01-01: 4
2020-01-02: 9
2020-01-03: 8
2020-01-04: 6
指标一 208
最新数据： 209
2020-01-04
最新值： 11.0
前项数据： 2020-01-03
前项值： 10.0
百分比变化： +10.0%
当百分比变化大于5%时
以第一格式（例如斜体）警报 212
指标二 210
最新数据： 211
2020-01-04
最新值： 6
前项数据： 2020-01-03
前项值： 8
百分比变化： -25.0%
当百分比变化大于5%时
以第二格式（例如粗体）警报 213
统计上关联于
指标一 214
最新数据：
2020-01-04
最新值： 11.0
前项数据： 2020-01-03
前项值： 10.0
百分比变化： +10.0%
指标二 215
最新数据：
2020-01-04
最新值： 6
前项数据： 2020-01-03
前项值： 8
百分比变化： -25.0%
202

图2(a)

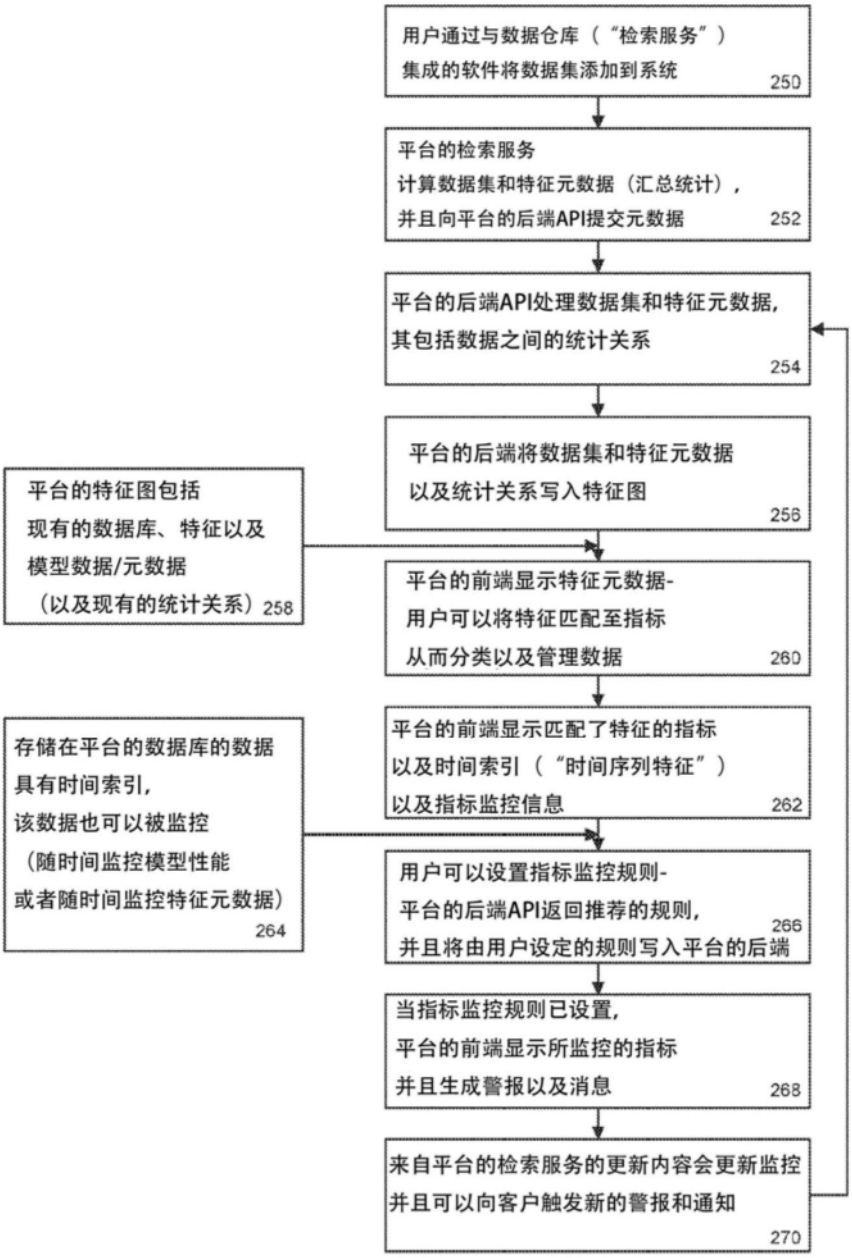

用户通过与数据仓库（“检索服务”）
集成的软件将数据集添加到系统
250
平台的检索服务
计算数据集和特征元数据（汇总统计），
并且向平台的后端API提交元数据
252
平台的后端API处理数据集和特征元数据，
其包括数据之间的统计关系
254
平台的后端将数据集和特征元数据
以及统计关系写入特征图
256
平台的特征图包括
现有的数据库、特征以及
模型数据/元数据
（以及现有的统计关系）258
平台的前端显示特征元数据-
用户可以将特征匹配至指标
从而分类以及管理数据
260
平台的前端显示匹配了特征的指标
以及时间索引（“时间序列特征”）
以及指标监控信息
262
存储在平台的数据库的数据
具有时间索引，
该数据也可以被监控
（随时间监控模型性能
或者随时间监控特征元数据）
264
用户可以设置指标监控规则-
平台的后端API返回推荐的规则，
并且将由用户设定的规则写入平台的后端
266
当指标监控规则已设置，
平台的前端显示所监控的指标
并且生成警报以及消息
268
来自平台的检索服务的更新内容会更新监控
并且可以向客户触发新的警报和通知
270

图2(b)

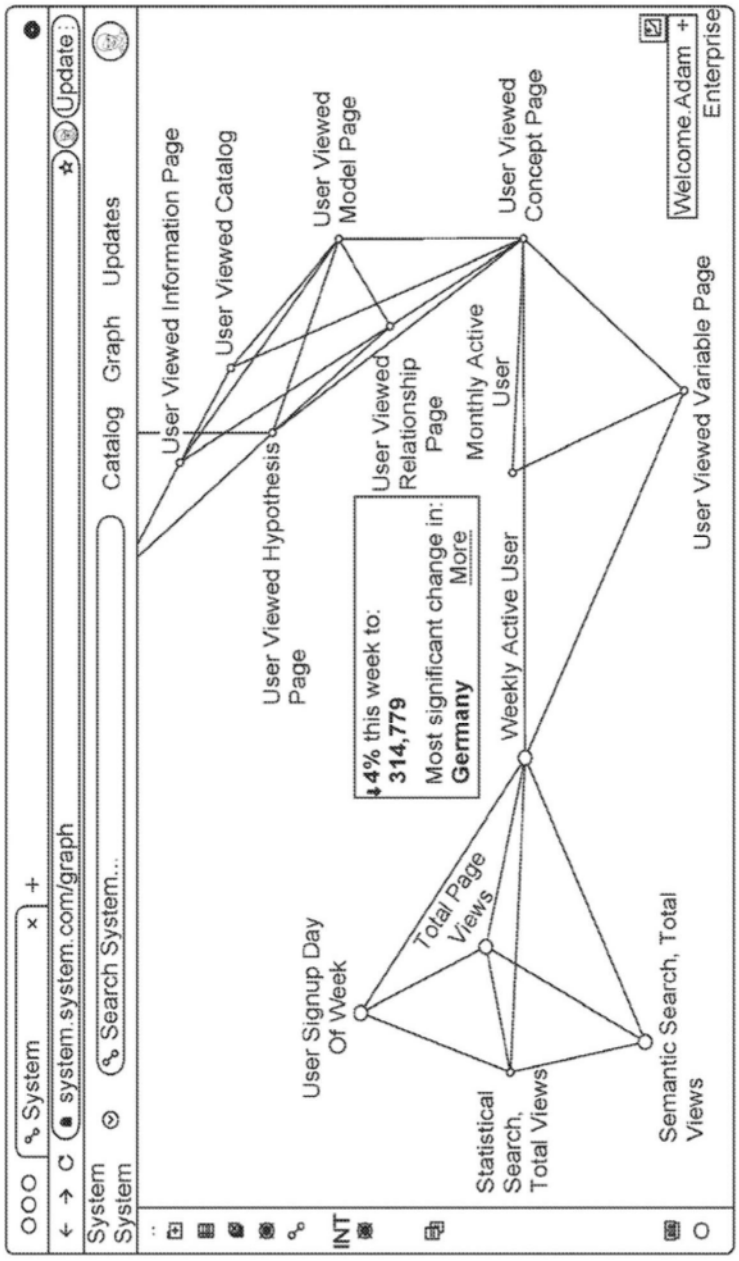

System
system.system.com/graph
Update
System
System
Search System...
Catalog
Graph
Updates
User Viewed Information Page
User Viewed Catalog
User Viewed Model Page
User Viewed Concept Page
Welcome.Adam +
Enterprise
User Viewed Hypothesis Page
User Viewed Relationship Page
Monthly Active User
User Viewed Variable Page
4% this week to:
314,779
Most significant change in:
Germany
More
Weekly Active User
User Signup Day Of Week
Total Page Views
Statistical Search, Total Views
Semantic Search, Total Views
INT

图2C

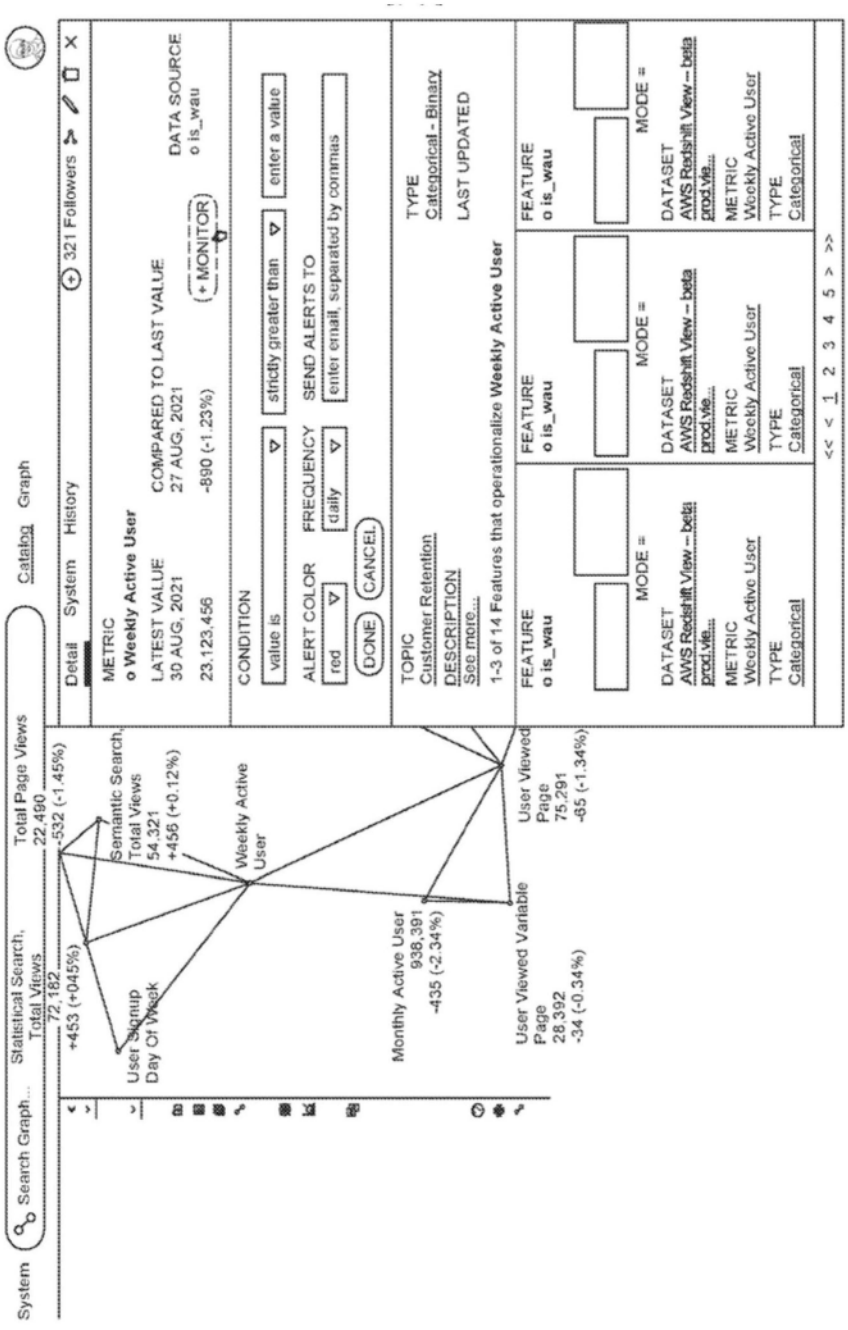

System
Search Graph...
Statistical Search, Total Views 72,182 +453 (+045%)
Total Page Views 22,490 -532 (-1.45%)
User Signup Day Of Week
Semantic Search, Total Views 54,321 +456 (+0.12%)
Weekly Active User
Monthly Active User 938,391 -435 (-2.34%)
User Viewed Variable Page 28,392 -34 (-0.34%)
User Viewed Page 75,291 -65 (-1.34%)
Catalog
Graph
Detail
System
History
321 Followers
METRIC
o Weekly Active User
LATEST VALUE
30 AUG, 2021
23,123,456
COMPARED TO LAST VALUE
27 AUG, 2021
-890 (-1.23%)
+ MONITOR
DATA SOURCE
o is_wau
CONDITION
value is
strictly greater than
enter a value
ALERT COLOR
red
FREQUENCY
daily
SEND ALERTS TO
enter email, separated by commas
DONE
CANCEL
TOPIC
Customer Retention
DESCRIPTION
See more...
1-3 of 14 Features that operationalize Weekly Active User
TYPE
Categorical - Binary
LAST UPDATED
FEATURE
o is_wau
MODE =
DATASET
AWS Redshift View -- beta prod.vie...
METRIC
Weekly Active User
TYPE
Categorical
<< < 1 2 3 4 5 > >>

图2D

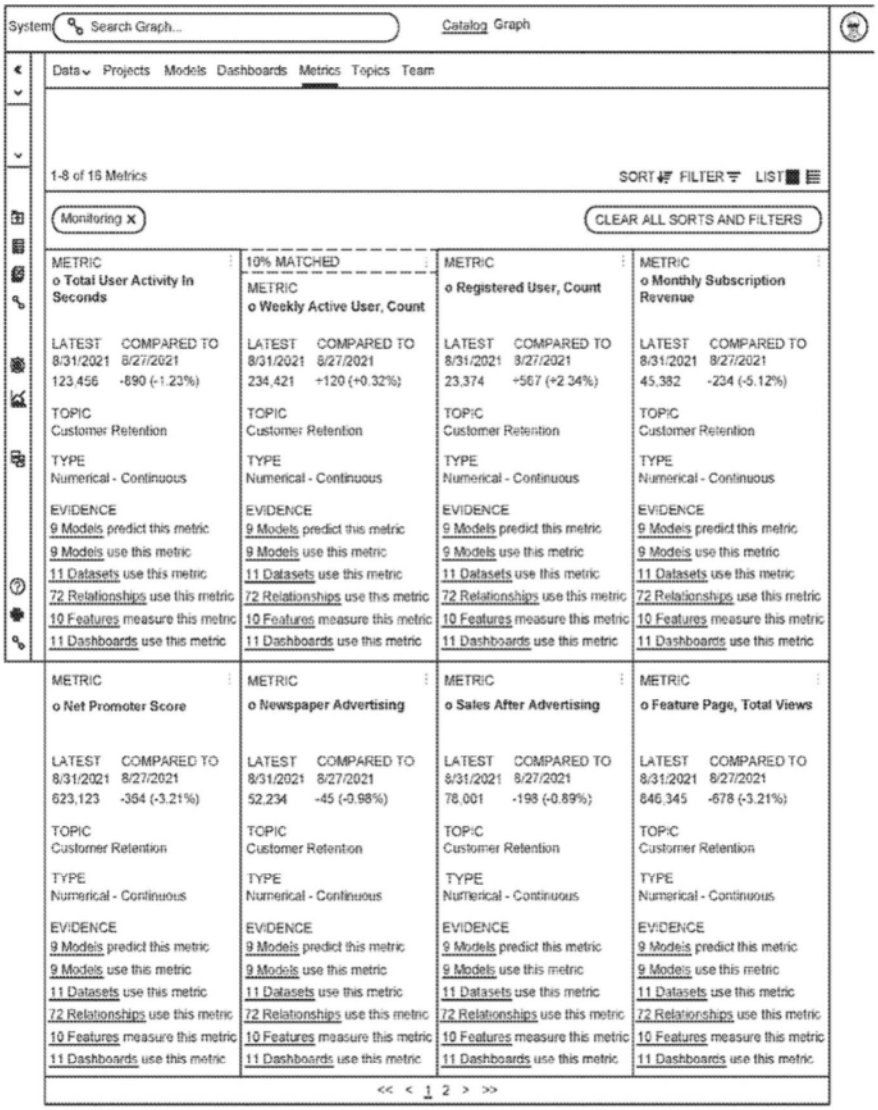
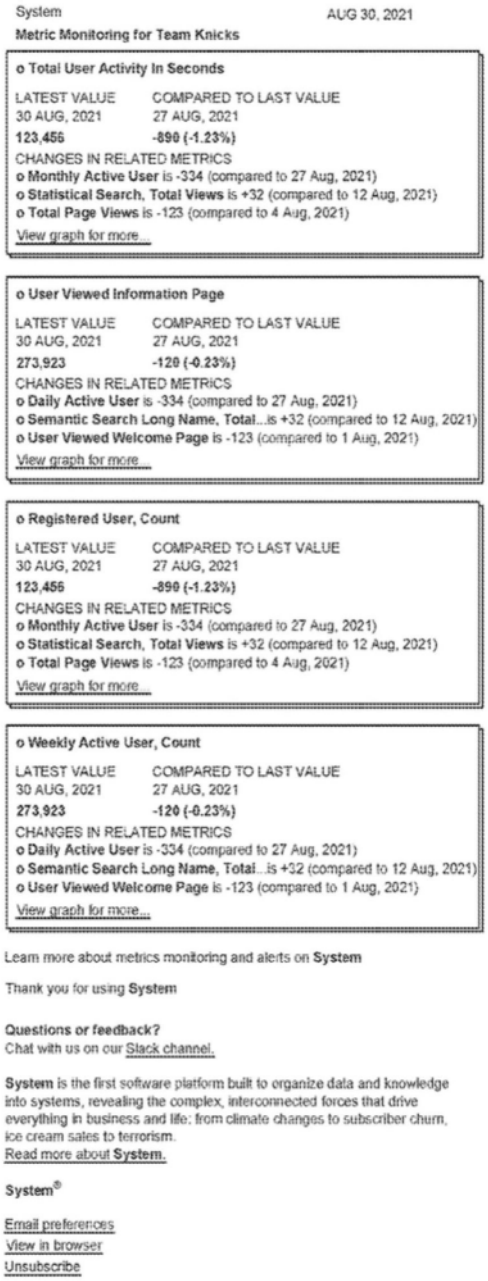

System
Search Graph...
Catalog Graph
Data
Projects
Models
Dashboards
Metrics
Topics
Team
1-8 of 16 Metrics
SORT
FILTER
LIST
Monitoring
CLEAR ALL SORTS AND FILTERS
METRIC
o Total User Activity In Seconds
LATEST 8/31/2021 123,456
COMPARED TO 8/27/2021 -890 (-1.23%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
10% MATCHED
METRIC
o Weekly Active User, Count
LATEST 8/31/2021 234,421
COMPARED TO 8/27/2021 +120 (+0.32%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
METRIC
o Registered User, Count
LATEST 8/31/2021 23,374
COMPARED TO 8/27/2021 +567 (+2.34%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
METRIC
o Monthly Subscription Revenue
LATEST 8/31/2021 45,382
COMPARED TO 8/27/2021 -234 (-5.12%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
METRIC
o Net Promoter Score
LATEST 8/31/2021 623,123
COMPARED TO 8/27/2021 -364 (-3.21%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
METRIC
o Newspaper Advertising
LATEST 8/31/2021 52,234
COMPARED TO 8/27/2021 -45 (-0.98%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
METRIC
o Sales After Advertising
LATEST 8/31/2021 78,001
COMPARED TO 8/27/2021 -198 (-0.89%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
METRIC
o Feature Page, Total Views
LATEST 8/31/2021 846,345
COMPARED TO 8/27/2021 -678 (-3.21%)
TOPIC Customer Retention
TYPE Numerical - Continuous
EVIDENCE
9 Models predict this metric
9 Models use this metric
11 Datasets use this metric
72 Relationships use this metric
10 Features measure this metric
11 Dashboards use this metric
<< < 1 2 > >>

图2E



System AUG 30, 2021

Metric Monitoring for Team Knicks

o Total User Activity In Seconds

LATEST VALUE	COMPARED TO LAST VALUE
30 AUG, 2021	27 AUG, 2021
123,456	**-890 (-1.23%)**

CHANGES IN RELATED METRICS
- **o Monthly Active User** is -334 (compared to 27 Aug, 2021)
- **o Statistical Search, Total Views** is +32 (compared to 12 Aug, 2021)
- **o Total Page Views** is -123 (compared to 4 Aug, 2021)

View graph for more...

o User Viewed Information Page

LATEST VALUE	COMPARED TO LAST VALUE
30 AUG, 2021	27 AUG, 2021
273,923	**-120 (-0.23%)**

CHANGES IN RELATED METRICS
- **o Daily Active User** is -334 (compared to 27 Aug, 2021)
- **o Semantic Search Long Name, Total**...is +32 (compared to 12 Aug, 2021)
- **o User Viewed Welcome Page** is -123 (compared to 1 Aug, 2021)

View graph for more...

o Registered User, Count

LATEST VALUE	COMPARED TO LAST VALUE
30 AUG, 2021	27 AUG, 2021
123,456	**-890 (-1.23%)**

CHANGES IN RELATED METRICS
- **o Monthly Active User** is -334 (compared to 27 Aug, 2021)
- **o Statistical Search, Total Views** is +32 (compared to 12 Aug, 2021)
- **o Total Page Views** is -123 (compared to 4 Aug, 2021)

View graph for more...

o Weekly Active User, Count

LATEST VALUE	COMPARED TO LAST VALUE
30 AUG, 2021	27 AUG, 2021
273,923	**-120 (-0.23%)**

CHANGES IN RELATED METRICS
- **o Daily Active User** is -334 (compared to 27 Aug, 2021)
- **o Semantic Search Long Name, Total**...is +32 (compared to 12 Aug, 2021)
- **o User Viewed Welcome Page** is -123 (compared to 1 Aug, 2021)

View graph for more...

Learn more about metrics monitoring and alerts on **System**

Thank you for using **System**

Questions or feedback?
Chat with us on our Slack channel.

System is the first software platform built to organize data and knowledge into systems, revealing the complex, interconnected forces that drive everything in business and life: from climate changes to subscriber churn, ice cream sales to terrorism.
Read more about **System.**

System®

Email preferences
View in browser
Unsubscribe

图2F

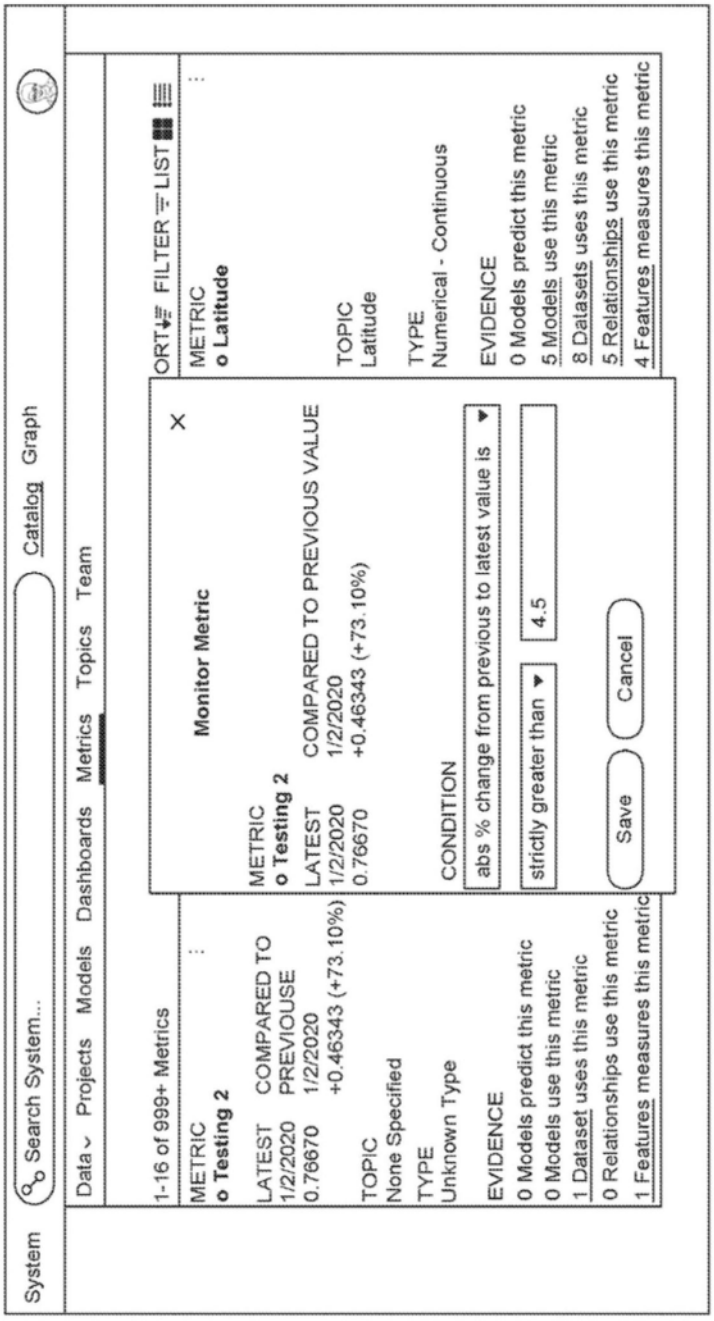

System
Search System...
Catalog
Graph
Data
Projects
Models
Dashboards
Metrics
Topics
Team
1-16 of 999+ Metrics
METRIC
o Testing 2
LATEST COMPARED TO PREVIOUSE
1/2/2020 1/2/2020
0.76670 +0.46343 (+73.10%)
TOPIC
None Specified
TYPE
Unknown Type
EVIDENCE
0 Models predict this metric
0 Models use this metric
1 Dataset uses this metric
0 Relationships use this metric
1 Features measures this metric
Monitor Metric
METRIC
o Testing 2
LATEST COMPARED TO PREVIOUS VALUE
1/2/2020 1/2/2020
0.76670 +0.46343 (+73.10%)
CONDITION
abs % change from previous to latest value is
strictly greater than
4.5
Save
Cancel
ORT
FILTER
LIST
METRIC
o Latitude
TOPIC
Latitude
TYPE
Numerical - Continuous
EVIDENCE
0 Models predict this metric
5 Models use this metric
8 Datasets uses this metric
5 Relationships use this metric
4 Features measures this metric

图2G

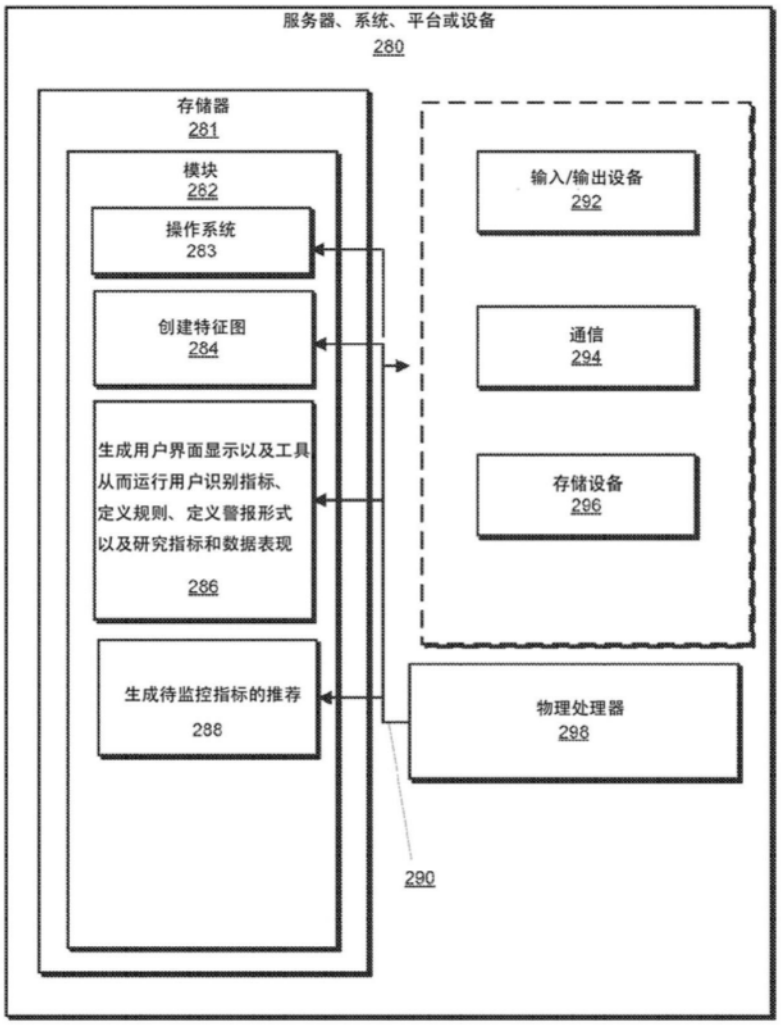

服务器、系统、平台或设备
280
存储器
281
模块
282
操作系统
283
创建特征图
284
生成用户界面显示以及工具
从而运行用户识别指标、
定义规则、定义警报形式
以及研究指标和数据表现
286
生成待监控指标的推荐
288
输入/输出设备
292
通信
294
存储设备
296
物理处理器
298
290

图2(h)

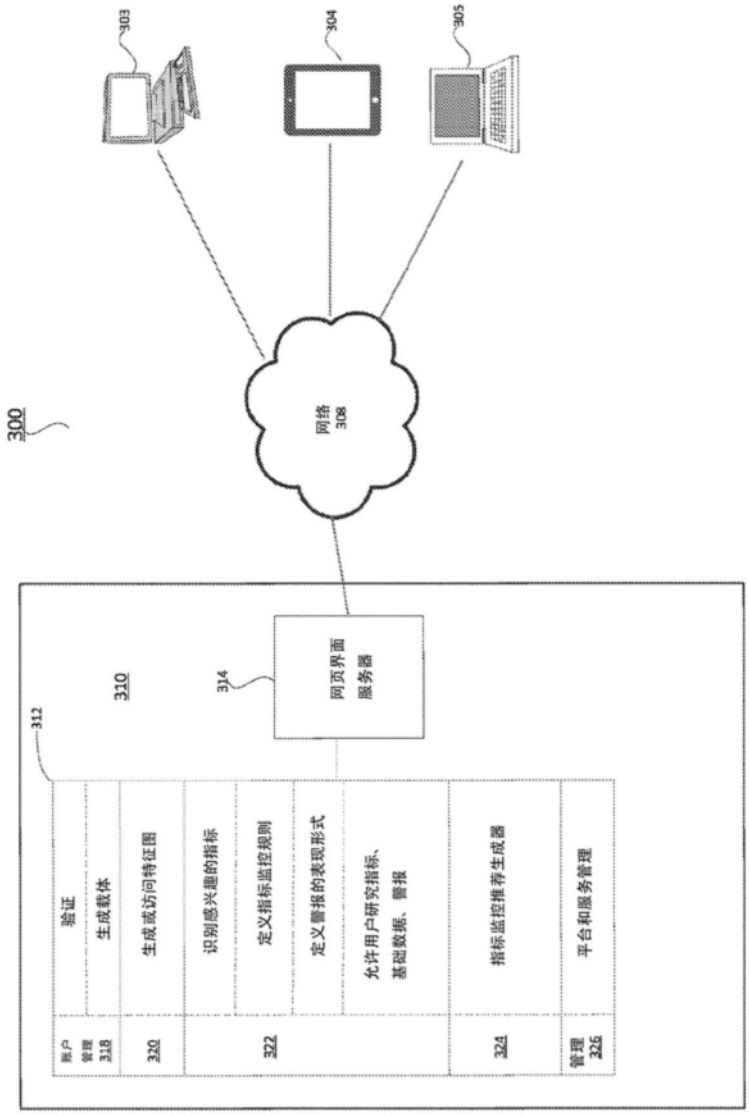

303
304
305
300
网络
308
310
312
314
网页界面
服务器
验证
生成载体
生成或访问特征图
识别感兴趣的指标
定义指标监控规则
定义警报的表现形式
允许用户研究指标、
基础数据、警报
指标监控推荐生成器
平台和服务管理
账户
管理
318
320
322
324
管理
326

图3

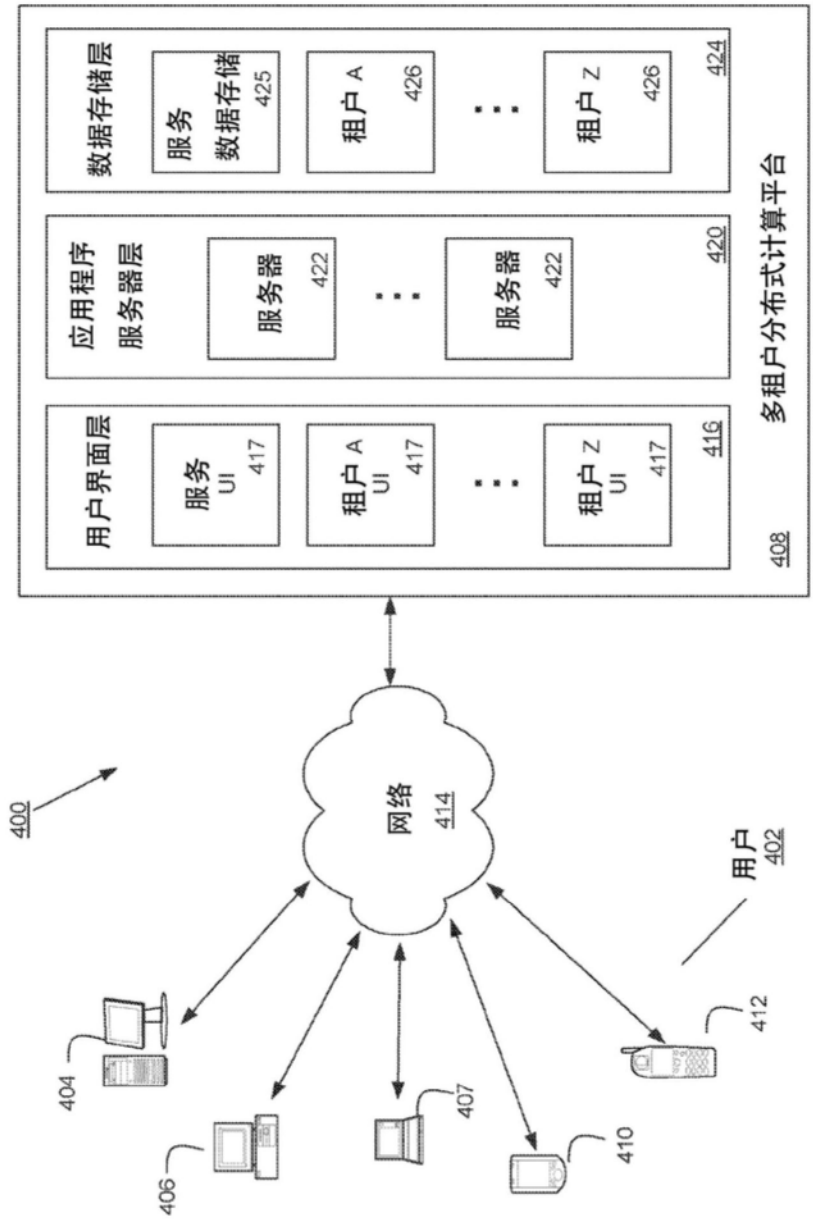

400
多租户分布式计算平台
408
用户界面层
服务 UI 417
租户 A UI 417
租户 Z UI 417
416
应用程序服务器层
服务器 422
服务器 422
420
数据存储层
服务数据存储 425
租户 A 426
租户 Z 426
424
网络 414
用户 402
404
406
407
410
412

图4

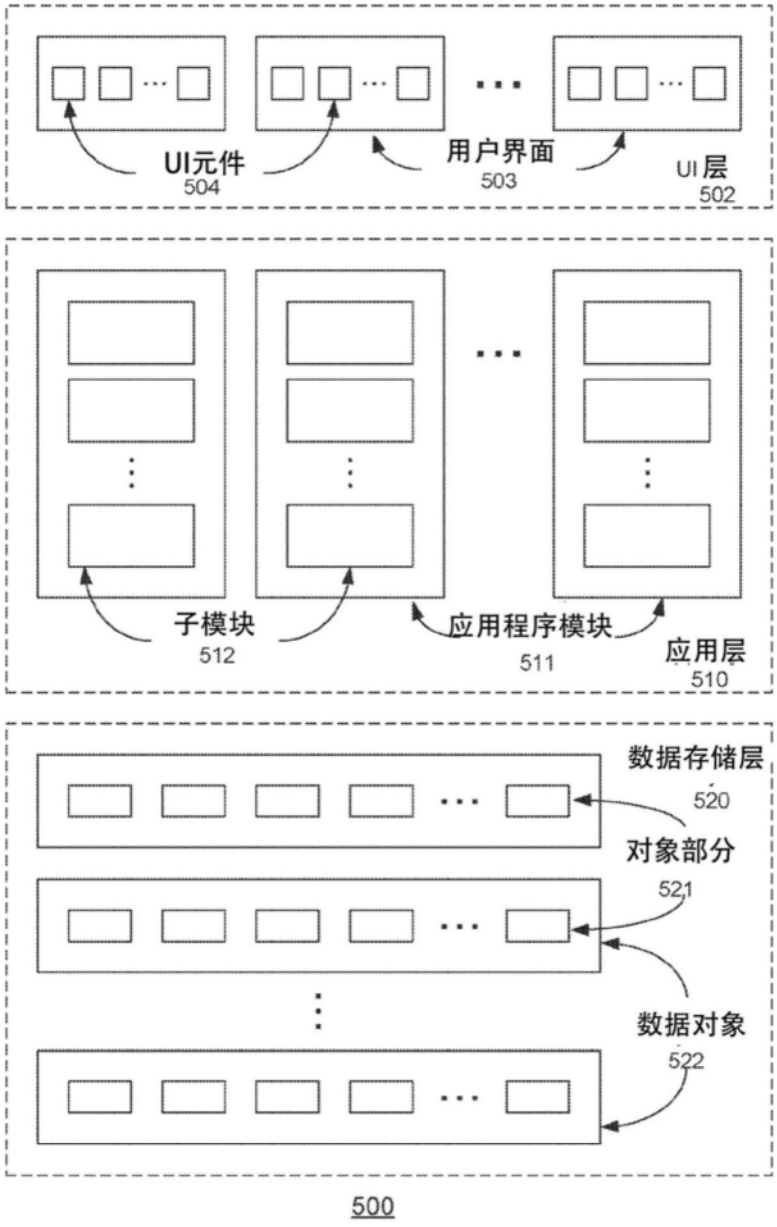

UI元件
504
用户界面
503
UI层
502
子模块
512
应用程序模块
511
应用层
510
数据存储层
520
对象部分
521
数据对象
522
500

图5