(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118069637 A
(43)申请公布日 2024.05.24

(21)申请号 202410033602.0
(22)申请日 2024.01.09
(71)申请人 太原罗克佳华工业有限公司
地址 030000 山西省太原市山西综改示范区太原学府园区佳华街8号(罗克佳华电子工业园)
(72)发明人 李玮 孙洪龙 朱德福 张晓岩 仇志伟 侯绍君 崔路凯 周江涛 柳行
(74)专利代理机构 北京超凡宏宇知识产权代理有限公司 11463
专利代理师 李翠
(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/2455(2019.01)
G06F 16/28(2019.01)

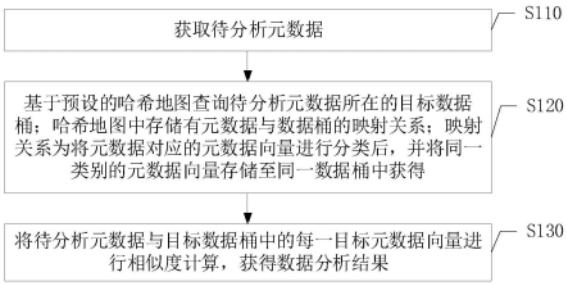
权利要求书2页 说明书10页 附图2页

(54)发明名称

一种元数据分析方法、装置、电子设备及存储介质

(57)摘要

本申请提供一种元数据分析方法、装置、电子设备及存储介质,该方法包括:获取待分析元数据;基于预设的哈希地图查询待分析元数据所在的目标数据桶;哈希地图中存储有元数据与数据桶的映射关系;映射关系为将元数据对应的元数据向量进行分类后,并将同一类别的元数据向量存储至同一数据桶中获得;将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果。在进行数据分析时,首先查找待分析元数据所在的目标数据桶,之后将待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,而无需将一个元数据与所有的元数据进行比对,减少运算量以及人工成本,提高数据分析的效率。



CN 118069637 A

1.一种元数据分析方法,其特征在于,包括:

获取待分析元数据;

基于预设的哈希地图查询所述待分析元数据所在的目标数据桶;所述哈希地图中存储有元数据与数据桶的映射关系;所述映射关系为将所述元数据对应的元数据向量进行分类后,并将同一类别的所述元数据向量存储至同一数据桶中获得;

将所述待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果。

2.根据权利要求1所述的方法,其特征在于,在所述基于预设的哈希地图查询所述待分析元数据所在的目标数据桶之前,所述方法还包括:

将预先采集的所述元数据进行向量转换,获得每一所述元数据对应的所述元数据向量;

获得预先构建的至少两个超平面;

利用所述超平面将所述元数据向量进行分类,获得向量分类结果;

基于所述向量分类结果,利用所述哈希地图存储所述元数据与所述数据桶的映射关系。

3.根据权利要求2所述的方法,其特征在于,所述利用所述超平面将所述元数据向量进行分类,获得向量分类结果,包括:

将每一所述元数据向量与每一所述超平面进行点积计算,获得每一所述元数据向量对应的哈希值;

基于所述元数据向量对应的所述哈希值,将所述元数据向量进行分类,获得所述向量分类结果。

4.根据权利要求2所述的方法,其特征在于,所述将每一所述元数据向量与每一所述超平面进行点积计算,获得每一所述元数据向量对应的哈希值,包括:

利用向量点积公式分别将所述元数据向量与每一所述超平面进行点积计算,获得所述元数据向量与每一所述超平面的点积结果;

将所述元数据向量对应的所述点积结果,按照预设的所述超平面顺序进行排序,获得所述元数据向量对应的所述哈希值。

5.根据权利要求4所述的方法,其特征在于,所述元数据向量包括向量横坐标、向量纵坐标和向量竖坐标;所述超平面包括超平面横坐标、超平面纵坐标和超平面竖坐标;所述向量点积公式包括:

$Vn \cdot Sn = X_1 \cdot X_2 + Y_1 \cdot Y_2 + Z_1 \cdot Z_2$

其中,Vn为元数据向量;Sn为超平面;Vn·Sn为点积结果;X_1为向量横坐标;Y_1为向量纵坐标;Z_1为向量竖坐标;X_2为超平面横坐标;Y_2为超平面纵坐标;Z_2为超平面竖坐标。

6.根据权利要求2所述的方法,其特征在于,所述基于所述向量分类结果,利用所述哈希地图存储所述元数据与所述数据桶的映射关系,包括:

基于所述向量分类结果,将所述同一类别的所述元数据向量存储至同一所述数据桶中;

利用所述哈希地图,存储所述元数据向量对应的所述元数据与所述数据桶的映射关系。

7.根据权利要求1-6任一所述的方法,其特征在于,所述将所述待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,包括:

获得所述目标数据桶中的所述目标元数据向量;

将所述待分析元数据与每一所述目标元数据向量进行相似度计算,获得每一所述目标元数据向量对应的相似度数据;

将所述相似度数据进行排序,基于相似度排序结果获得所述数据分析结果。

8.一种元数据分析装置,其特征在于,包括:

获取模块,用于获取待分析元数据;

查询数据桶模块,用于基于预设的哈希地图查询所述待分析元数据所在的目标数据桶;所述哈希地图中存储有元数据与数据桶的映射关系;所述映射关系为将所述元数据对应的元数据向量进行分类后,并将同一类别的所述元数据存储至同一数据桶中获得;

分析模块,用于将所述待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果。

9.一种电子设备,其特征在于,包括:处理器和存储器,所述存储器存储有所述处理器可执行的机器可读指令,所述机器可读指令被所述处理器执行时执行如权利要求1至7任一所述的方法。

10.一种计算机可读存储介质,其特征在于,该计算机可读存储介质上存储有计算机程序,该计算机程序被处理器运行时执行如权利要求1至7任一所述的方法。

一种元数据分析方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及数据处理技术领域，具体而言，涉及一种元数据分析方法、装置、电子设备及存储介质。

背景技术

[0002] 元数据是定义和描述数据的数据，是业务和系统之间的翻译纽带，提供业务和系统双方一致的语义和逻辑。元数据分析应用于例如数据采集、数据开发、数据质量监控和数据集数据查询等多个领域。现有技术在对元数据进行分析时，通常需要人工对元数据进行梳理，根据元数据的定义找到相似表和字段，数据处理效率较低。

发明内容

[0003] 本申请实施例的目的在于一种元数据分析方法、装置、电子设备及存储介质，预先将元数据进行分类，并根据分类结果将元数据映射至对应的数据桶，每一数据桶中存储了一定相似程度的元数据。在进行数据分析时，首先查找待分析元数据所在的目标数据桶，将待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算，获得数据分析结果，减少运算量以及人工成本，提高数据分析的效率。

[0004] 第一方面，本申请实施例提供了一种元数据分析方法，包括：获取待分析元数据；基于预设的哈希地图查询待分析元数据所在的目标数据桶；哈希地图中存储有元数据与数据桶的映射关系；映射关系为将元数据对应的元数据向量进行分类后，并将同一类别的元数据向量存储至同一数据桶中获得；将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算，获得数据分析结果。

[0005] 在上述的实现过程中，预先将元数据进行分类，并根据分类结果将元数据映射至对应的数据桶，每一数据桶中存储了一定相似程度的元数据。在进行数据分析时，首先查找待分析元数据所在的目标数据桶，之后将待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算，获得数据分析结果，而无需将一个元数据与所有的元数据进行比对，减少运算量以及人工成本，提高数据分析的效率。

[0006] 可选的，在本申请实施例中，在基于预设的哈希地图查询待分析元数据所在的目标数据桶之前，方法还包括：将预先采集的元数据进行向量转换，获得每一元数据对应的元数据向量；获得预先构建的至少两个超平面；利用超平面将元数据向量进行分类，获得向量分类结果；基于向量分类结果，利用哈希地图存储元数据与数据桶的映射关系。

[0007] 在上述的实现过程中，将元数据转换为元数据向量，通过在数据空间中选择合适的超平面，可以将元数据向量划分为具有相似特征的类别，从而实现元数据向量的分类。以及利用哈希地图存储元数据与数据桶的映射关系，提高计算效率。

[0008] 可选的，在本申请实施例中，利用超平面将元数据向量进行分类，获得向量分类结果，包括：将每一元数据向量与每一超平面进行点积计算，获得每一元数据向量对应的哈希值；基于元数据向量对应的哈希值，将元数据向量进行分类，获得向量分类结果。

[0009] 在上述的实现过程中,利用点积求向量与二维超平面之间的夹角余弦值,从而确定它们之间的相似性,实现利用超平面将元数据向量进行分类,获得向量分类结果,提高数据分析的效率。

[0010] 可选的,在本申请实施例中,将每一元数据向量与每一超平面进行点积计算,获得每一元数据向量对应的哈希值,包括:利用向量点积公式分别将元数据向量与每一超平面进行点积计算,获得元数据向量与每一超平面的点积结果;将元数据向量对应的点积结果,按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值。

[0011] 在上述的实现过程中,利用向量点积公式分别将元数据向量与每一超平面进行点积计算,并且将点积结果按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值,为后续的数据分析提供更准确的支撑。

[0012] 可选的,在本申请实施例中,元数据向量包括向量横坐标、向量纵坐标和向量竖坐标;超平面包括超平面横坐标、超平面纵坐标和超平面竖坐标;向量点积公式包括:

[0013] $Vn \cdot Sn = X_1 \cdot X_2 + Y_1 \cdot Y_2 + Z_1 \cdot Z_2$

[0014] 其中,Vn为元数据向量;Sn为超平面;Vn·Sn为点积结果;X_1为向量横坐标;Y_1为向量纵坐标;Z_1为向量竖坐标;X_2为超平面横坐标;Y_2为超平面纵坐标;Z_2为超平面竖坐标。

[0015] 在上述的实现过程中,利用向量点积公式分别将元数据向量与每一超平面进行点积计算,并且将点积结果按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值,点积结果可以通过1或者0表示,超平面的数量可以根据实际情况设置,生成哈希值,为后续的数据分析提供更准确的支撑。

[0016] 可选的,在本申请实施例中,基于向量分类结果,利用哈希地图存储元数据与数据桶的映射关系,包括:基于向量分类结果,将同一类别的元数据向量存储至同一数据桶中;利用哈希地图,存储元数据向量对应的元数据与数据桶的映射关系。

[0017] 在上述的实现过程中,经过上述元数据向量分类之后,划分为同一类别的元数据向量为具有一定相似性的数据,因此,在一致性分析之前,很大程度上将毫无关联性的数据过滤,为进一步进行一致性分析时,提供了更精准的数据,提高了数据分析的准确性和效率。

[0018] 可选的,在本申请实施例中,将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,包括:获得目标数据桶中的目标元数据向量;将待分析元数据与每一目标元数据向量进行相似度计算,获得每一目标元数据向量对应的相似度数据;将相似度数据进行排序,基于相似度排序结果获得数据分析结果。

[0019] 在上述的实现过程中,将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,而无需将一个元数据与所有的元数据进行比对,减少运算量以及人工成本,提高数据分析的效率。

[0020] 第二方面,本申请实施例还提供了一种元数据分析装置,包括:获取模块,用于获取待分析元数据;查询数据桶模块,用于基于预设的哈希地图查询待分析元数据所在的目标数据桶;哈希地图中存储有元数据与数据桶的映射关系;映射关系为将元数据对应的元数据向量进行分类后,并将同一类别的元数据存储至同一数据桶中获得;分析模块,用于将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果。

[0021] 第三方面,本申请实施例还提供了一种电子设备,包括:处理器和存储器,存储器

存储有处理器可执行的机器可读指令,机器可读指令被处理器执行时执行如上面描述的方法。

[0022] 第四方面,本申请实施例还提供了一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,该计算机程序被处理器运行时执行上面描述的方法。

[0023] 采用本申请提供元数据分析方法、装置、电子设备及存储介质,预先将元数据进行分类,并根据分类结果将元数据映射至对应的数据桶,每一数据桶中存储了一定相似程度的元数据。在进行数据分析时,首先查找待分析元数据所在的目标数据桶,之后可以将待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,而无需将一个元数据与所有的元数据进行比对,减少运算量以及人工成本,提高数据分析的效率。

附图说明

[0024] 为了更清楚地说明本申请实施例的技术方案,下面将对本申请实施例中所需要使用的附图作简单地介绍,应当理解,以下附图仅示出了本申请的某些实施例,因此不应被看作是对范围的限定,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他相关的附图。

[0025] 图1为本申请实施例提供的一种元数据分析方法的流程示意图;

[0026] 图2为本申请实施例提供的超平面示意图;

[0027] 图3为本申请实施例提供的元数据分析装置的结构示意图;

[0028] 图4为本申请实施例提供的电子设备的结构示意图。

具体实施方式

[0029] 下面将结合附图对本申请技术方案的实施例进行详细的描述。以下实施例仅用于更加清楚地说明本申请的技术方案,因此只作为示例,而不能以此来限制本申请的保护范围。

[0030] 除非另有定义,本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同;本文中所使用的术语只是为了描述具体的实施例的目的,不是旨在于限制本申请。

[0031] 在本申请实施例的描述中,技术术语“第一”、“第二”等仅用于区别不同对象,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量、特定顺序或主次关系。在本申请实施例的描述中,“多个”的含义是两个及以上,除非另有明确具体的限定。

[0032] 数据治理中元数据管理成为不可或缺的重要组成部分,特别是在数据采集、数据开发、数据质量、数据应用各方面,元数据管理包括相似元数据推荐,可以帮助数据治理人员在数仓中快速找到不同表中定义一致的字段,从而便于做数据对比,数据一致性分析等应用场景。

[0033] 现有的元数据管理是通过数据采集将数据库中表信息、字段信息统一汇聚到数据仓库,经过盘点和编目,定义元数据的业务属性、技术属性和管理属性。在数据一致性分析时和数据对比时,需要根据元数据的定义找到相似表和字段。依靠人工梳理,在成千上万的数据表和字段中,根据元数据的定义描述找出一致的元数据。而数仓中的元数据往往涉及

到多个系统、部门和组织架构,各系统对相同的数据定义口径不统一,难以准确完全匹配。

[0034] 另一方面,现有技术在做数据一致性分析时,会将一个元数据与所有的元数据进行比对,造成运算量过大,资源消耗严重,且数据处理效率较低。

[0035] 本申请实施例提供一种元数据分析方法、装置、电子设备及存储介质,预先将元数据进行分类,并根据分类结果将元数据映射至对应的数据桶,每一数据桶中存储了一定相似程度的元数据。在进行数据分析时,首先查找待分析元数据所在的目标数据桶,之后将待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,而无需将一个元数据与所有的元数据进行比对,减少运算量以及人工成本,提高数据分析的效率。

[0036] 请参见图1示出的本申请实施例提供的一种元数据分析方法的流程示意图。本申请实施例提供的元数据分析方法可以应用于电子设备,该电子设备可以包括终端以及服务器;其中终端具体可以为智能手机、平板电脑、计算机、个人数字助理(Personal Digital Assitant,PDA)等;服务器具体可以为应用服务器,也可以为Web服务器。该元数据分析方法可以包括:

[0037] 步骤S110:获取待分析元数据。

[0038] 步骤S120:基于预设的哈希地图查询待分析元数据所在的目标数据桶;哈希地图中存储有元数据与数据桶的映射关系;映射关系为将元数据对应的元数据向量进行分类后,并将同一类别的元数据向量存储至同一数据桶中获得。

[0039] 步骤S130:将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果。

[0040] 在步骤S110中,待分析元数据是用户指定的需要进行相似性推荐的元数据。元数据是描述和定义数据的数据,可以提供关于数据的属性、特征和其他详细信息。举例来说,元数据可以包括数据的来源、格式、结构、内容、质量、使用规则以及加工过程等方面的信息。

[0041] 在步骤S120中,在预设的哈希地图查询待分析元数据所在的目标数据桶,哈希地图中存储有元数据与数据桶的映射关系。哈希地图是一种数据结构,可以将键(key)与值(value)建立关联,并通过哈希函数将键映射到唯一的索引位置,从而实现高效的数据存取和查找操作。例如,可以将元数据作为键(key),将元数据所在的数据桶作为值(value),实现将元数据与数据桶建立关联。

[0042] 数据桶(或称为“Bucket”)是数据库中存储数据的基本元素,可以理解为数据库中的一个存储单元,例如它是一块连续的内存空间。

[0043] 下面对获得元数据与数据桶的映射关系的过程进行描述,首先对元数据进行采集,将采集到的元数据进行向量转换,获得每一元数据对应的元数据向量,将元数据向量进行分类,获得元数据向量的分类结果,并将同一类别的元数据向量存储至同一数据桶中。这样就得到了元数据向量和数据桶的关联关系,可以利用哈希地图,存储元数据向量对应的元数据与数据桶的映射关系。

[0044] 在步骤S130中,同一类别的元数据向量存储至同一数据桶中,那么待分析元数据所在的目标数据桶中包括至少一个目标元数据向量,将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得待分析元数据与每一目标元数据向量的相似度数据,

可以基于相似度数据,获得数据分析结果。

[0045] 数据分析结果可以表征待分析元数据的一致性分析结果或相似性推荐数据等。示例性的,可以将相似度数据大于预设阈值的目标元数据向量作为数据分析结果;还可以将相似度数据进行排序,将排序结果前预设位次的目标元数据向量作为数据分析结果。

[0046] 在上述的实现过程中,预先将元数据进行分类,并根据分类结果将元数据映射至对应的数据桶,每一数据桶中存储了一定相似程度的元数据。在进行数据分析时,首先查找待分析元数据所在的目标数据桶,之后将待分析元数据与所述目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,而无需将一个元数据与所有的元数据进行比对,减少运算量以及人工成本,提高数据分析的效率。

[0047] 可选的,在本申请实施例中,在基于预设的哈希地图查询待分析元数据所在的目标数据桶之前,方法还包括:将预先采集的元数据进行向量转换,获得每一元数据对应的元数据向量;获得预先构建的至少两个超平面;利用超平面将元数据向量进行分类,获得向量分类结果;基于向量分类结果,利用哈希地图存储元数据与数据桶的映射关系。

[0048] 在具体的实现过程中:在基于预设的哈希地图查询待分析元数据所在的目标数据桶之前,先利用哈希地图存储元数据与数据桶的映射关系。

[0049] 具体例如,预先利用元数据采集工具采集元数据,元数据采集工具可以通过扫描数据源(如数据库、文件系统、API等)来提取元数据。也可以直接连接到数据库系统,并获取数据库的元数据信息。

[0050] 将采集到的元数据进行向量转换的过程例如:对采集到的元数据进行数据解析,获得元数据属性,将元数据属性进行组合生成元数据特征,其中,组合的方式可以为文本拼接。将获取到元数据特征进行向量转换,获得每一元数据对应的元数据向量。向量转换可以利用word2vec技术实现,也可以根据需求利用其他方式实现。

[0051] 请参见图2示出的本申请实施例提供的超平面示意图。

[0052] 构建至少两个超平面,超平面是高维空间中的一个平面,其维度比支撑这个平面的空间的维度少一维。在二维空间中,超平面是一个直线;在三维空间中,超平面是一个平面。

[0053] 本申请实施例中,超平面可以是二维的embedding(嵌入)。如图2所示,在二维中,超平面可以看作是一个线性分类器,超平面能够将不同的元数据向量映射到平面上的不同区域,并在该区域中对它们进行分类,获得向量分类结果。

[0054] 可以理解的是,如果二维超平面无法将元数据向量分开,则可能需要使用一些非线性方法,如Kernel Trick,将元数据向量映射到更高维的空间中,以使对元数据向量进行分类。

[0055] 超平面的数量可以根据元数据向量的数量确定,也可以根据分类精度确定。超平面的数量越大,表示分类精度越高,同时相似的元数据概率越低;反之超平面的数量越小,表示精度越低,相似的元数据概率越高。示例性的,假设元数据向量有700条,预计每100条数据构建一个超平面,则超平面的数量可以为7个。如图2所示,可以在二维平台直角坐标系下表示超平面,超平面的数量可以为7,分别为S1-S7。

[0056] 获得向量分类结果之后,将同一类别的元数据向量存储至同一数据桶中,基于元数据向量与数据桶的关联关系,利用哈希地图存储元数据与数据桶的映射关系。

[0057] 在上述的实现过程中,将元数据转换为元数据向量,通过在数据空间中选择合适的超平面,可以将元数据向量划分为具有相似特征的类别,从而实现元数据向量的分类。以及利用哈希地图存储元数据与数据桶的映射关系,提高计算效率。

[0058] 可选的,在本申请实施例中,利用超平面将元数据向量进行分类,获得向量分类结果,包括:将每一元数据向量与每一超平面进行点积计算,获得每一元数据向量对应的哈希值;基于元数据向量对应的哈希值,将元数据向量进行分类,获得向量分类结果。

[0059] 在具体的实现过程中:利用超平面将元数据向量进行分类具体例如,将元数据向量分别与每一超平面进行点积计算,获得元数据向量与每一超平面的点积结果,基于点积结果获得元数据向量对应的哈希值。

[0060] 点积(Dot Product)计算,又称为向量积、数量积,是指两个向量在数学上的一种运算。在欧几里德空间中,点积是指两个长度相等的向量在相应位置的对应元素的乘积之和。点积可以用于求两个向量或向量与二维超平面之间的夹角余弦值,从而确定它们之间的相似性。

[0061] 举例来说,共有6个超平面,元数据向量分别与6个超平面进行点积计算,获得6个点积结果;将这6个点积结果进行组合、排序或其他运算,生成该元数据向量对应的哈希值。将每一个元数据向量分别进行上述计算,即可以生成每一元数据向量对应的哈希值。

[0062] 基于元数据向量对应的哈希值,将元数据向量进行分类,获得向量分类结果,例如,可以将哈希值相同的元数据向量作为一个类别,也可以先对哈希值进行数据段划分,将每一数据段的哈希值对应的元数据向量作为一个类别。

[0063] 在上述的实现过程中,利用点积求向量与二维超平面之间的夹角余弦值,从而确定它们之间的相似性,实现利用超平面将元数据向量进行分类,获得向量分类结果,提高数据分析的效率。

[0064] 可选的,在本申请实施例中,将每一元数据向量与每一超平面进行点积计算,获得每一元数据向量对应的哈希值,包括:利用向量点积公式分别将元数据向量与每一超平面进行点积计算,获得元数据向量与每一超平面的点积结果;将元数据向量对应的点积结果,按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值。

[0065] 在具体的实现过程中:利用向量点积公式分别将元数据向量与每一超平面进行点积计算,获得元数据向量与每一超平面的点积结果。作为一种实施方式,元数据向量包括向量横坐标、向量纵坐标和向量竖坐标;超平面包括超平面横坐标、超平面纵坐标和超平面竖坐标;向量点积公式包括:

[0066] $Vn \cdot Sn = X_1 \cdot X_2 + Y_1 \cdot Y_2 + Z_1 \cdot Z_2$

[0067] 其中,Vn为元数据向量;Sn为超平面;Vn·Sn为点积结果;X_1为向量横坐标;Y_1为向量纵坐标;Z_1为向量竖坐标;X_2为超平面横坐标;Y_2为超平面纵坐标;Z_2为超平面竖坐标。

[0068] 利用上述公式可以获得元数据向量与每一超平面的点积结果,这里点积结果Vn·Sn可以代表元数据向量Vn在超平面Sn上的投影与Sn长度的乘积。若点积结果大于0的,代表两者的角度差不超过90,也就是说,元数据向量可以投影在超平面上,可以在点积结果大于0时,将点积结果记录为1,代表相似度为1。

[0069] 同理,点积结果等于0,表示两者的角度互相垂直,而点积结果小于零,代表两者的角度差大于90度。这两种情况,元数据向量无法或者难以投影在超平面上,可以在点积结果

不大于0时,将点积结果记录为0,代表相似度为0。

[0070] 将元数据向量对应的点积结果,按照预设的超平面顺序进行排序,例如,共有6个超平面,编号顺序为S1-S6,根据上述点积公式,可以分别获得元数据向量与S1-S6对应的6个点积结果,按照超平面编号顺序S1-S6,将6个点积结果进行排序,获得元数据向量对应的哈希值。

[0071] 为了便于记录,可以将每一元数据向量与每一超平面的点积结果和哈希值用表格表示。如表1示出的元数据向量与超平面的点积结果和哈希值。

[0072]

	S1	S2	S3	S4	S5	S6	哈希值
V1	1	1	1	0	0	0	111000
V2	1	1	1	0	1	0	111010
V3	1	1	1	0	0	0	111000
...	...	...	...	...	...	...	....
Vn	0	1	1	1	0	1	011101

[0073] 表1元数据向量与超平面的点积结果和哈希值

[0074] 表1中,第一行数据分别为6个超平面编号和哈希值,第二列数据为n个元数据向量。第二行数据为元数据向量V1分别与6个超平面的点积结果,以及根据这6个点积结果获得的元数据向量V1的哈希值。第三行数据为元数据向量V2分别与6个超平面的点积结果,以及根据这6个点积结果获得的元数据向量V2的哈希值。

[0075] 通过表1可以得出,元数据向量V1和元数据向量V3的哈希值一致,若类别划分规则为将哈希值一致的元数据向量划分为一个类别,则可以将元数据向量V1和元数据向量V3划分为同一个类别的元数据向量。

[0076] 在上述的实现过程中,利用向量点积公式分别将元数据向量与每一超平面进行点积计算,并且将点积结果按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值,点积结果可以通过1或者0表示,超平面的数量可以根据实际情况设置,生成哈希值,为后续的数据分析提供更准确的支撑。

[0077] 可选的,在本申请实施例中,基于向量分类结果,利用哈希地图存储元数据与数据桶的映射关系,包括:基于向量分类结果,将同一类别的元数据向量存储至同一数据桶中;利用哈希地图,存储元数据向量对应的元数据与数据桶的映射关系。

[0078] 在具体的实现过程中:接上述实施例,获得每一元数据向量的哈希值之后,若将哈希值一致的元数据向量划分为一个类别,则获得了向量分类结果。或者将每一数据段的哈希值对应的元数据向量作为一个类别,获得了向量分类结果。

[0079] 将同一类别的元数据向量存储至同一数据桶中,则获得了元数据向量和数据桶的关联关系,每一元数据向量对应一个元数据,因此,基于元数据向量和数据桶的关联关系,可以获得元数据与数据桶的映射关系,利用哈希地图,存储元数据向量对应的元数据与数据桶的映射关系。例如,可以将元数据作为哈希地图的键(key),将元数据所在的数据桶作为哈希地图的值(value)进行存储。

[0080] 在上述的实现过程中,经过上述元数据向量分类之后,划分为同一类别的元数据向量为具有一定相似性的数据,因此,在一致性分析之前,很大程度上将毫无关联性的数据过滤,为进一步进行一致性分析时,提供了更精准的数据,提高了数据分析的准确性和效

率。

[0081] 可选的,在本申请实施例中,将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果,包括:获得目标数据桶中的目标元数据向量;将待分析元数据与每一目标元数据向量进行相似度计算,获得每一目标元数据向量对应的相似度数据;将相似度数据进行排序,基于相似度排序结果获得数据分析结果。

[0082] 在具体的实现过程中:在对待分析元数据进行数据分析时,可以先利用待分析元数据,在哈希地图中查找待分析元数据所在的目标数据桶,之后获得目标数据桶中的目标元数据向量。将待分析元数据与每一目标元数据向量进行相似度计算,获得每一目标元数据向量对应的相似度数据。相似度计算是用于衡量两个对象之间相似性的一种方法,相似度计算可以包括欧氏距离、余弦相似度、皮尔逊相关系数或曼哈顿距离等。本申请实施例对此不作限定。

[0083] 获得待分析元数据每一目标元数据向量对应的相似度数据之后,将相似度数据进行排序,基于相似度排序结果获得数据分析结果,例如取相似度排序结果的前预设位的数据作为待分析元数据的一致性分析结果,或者相似数据推荐。

[0084] 在上述的实现过程中,将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,而无需将一个元数据与所有的元数据进行比对,减少运算量以及人工成本,提高数据分析的效率。

[0085] 在一个可选的实施例中,通过元数据采集工具采集元数据,依据元数据管理标准,元数据的特征包括但不限于:字段编码、字段名称、数据类型,对象类、特性等属性构成。将元数据的属性进行组合构成元数据特征,组合的方式可以是文本拼接。利用word2vec技术将对元数据特征进行向量转换,则每个元数据对应唯一一个向量。

[0086] 构建至少两个超平面,每个超平面也是一个二维的embedding,超平面根据元数据向量的个数确定,超平面可以是随机构成,例如从一定的范围内随机选择一个数值作为斜率或截距,根据斜率或截距构建超平面。

[0087] 将元数据向量V1-Vn分别与m个超平面进行向量点积,获得点积结果,若点积结果大于0,将点积结果记录为1,代表相似度为1;若点积结果不大于0,将点积结果记录为0,代表相似度为0。

[0088] 将元数据向量对应的点积结果,按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值。若将哈希值一致的元数据向量划分为一个类别,则获得了向量分类结果。

[0089] 将同一类别的元数据向量存储至同一数据桶中,则获得了元数据向量和数据桶的关联关系,利用哈希地图,存储元数据向量对应的元数据与数据桶的映射关系。

[0090] 在需要对待分析元数据进行一致性分析时,可以在哈希地图中查找待分析元数据所在的目标数据桶,之后获得目标数据桶中的目标元数据向量。将待分析元数据与每一目标元数据向量进行相似度计算,获得每一目标元数据向量对应的相似度数据。将相似度数据进行排序,获取从相似度数据中选择排名前N个的目标元数据向量作为待分析元数据的一致性分析结果,或者相似数据推荐。

[0091] 请参见图3示出的本申请实施例提供的元数据分析装置的结构示意图;本申请实施例提供了一种元数据分析装置200,包括:

[0092] 获取模块210,用于获取待分析元数据;

[0093] 查询数据桶模块220,用于基于预设的哈希地图查询待分析元数据所在的目标数据桶;哈希地图中存储有元数据与数据桶的映射关系;映射关系为将元数据对应的元数据向量进行分类后,并将同一类别的元数据存储至同一数据桶中获得;

[0094] 分析模块230,用于将待分析元数据与目标数据桶中的目标元数据向量进行相似度计算,获得数据分析结果。

[0095] 可选地,在本申请实施例中,元数据分析装置,还包括,数据映射模块,用于将预先采集的元数据进行向量转换,获得每一元数据对应的元数据向量;获得预先构建的至少两个超平面;利用超平面将元数据向量进行分类,获得向量分类结果;基于向量分类结果,利用哈希地图存储元数据与数据桶的映射关系。

[0096] 可选地,在本申请实施例中,元数据分析装置,数据映射模块,还用于将每一元数据向量与每一超平面进行点积计算,获得每一元数据向量对应的哈希值;基于元数据向量对应的哈希值,将元数据向量进行分类,获得向量分类结果。

[0097] 可选地,在本申请实施例中,元数据分析装置,数据映射模块,还用于利用向量点积公式分别将元数据向量与每一超平面进行点积计算,获得元数据向量与每一超平面的点积结果;将元数据向量对应的点积结果,按照预设的超平面顺序进行排序,获得元数据向量对应的哈希值。

[0098] 可选地,在本申请实施例中,元数据分析装置,元数据向量包括向量横坐标、向量纵坐标和向量竖坐标;超平面包括超平面横坐标、超平面纵坐标和超平面竖坐标;向量点积公式包括:

[0099] $Vn \cdot Sn = X_1 \cdot X_2 + Y_1 \cdot Y_2 + Z_1 \cdot Z_2$

[0100] 其中,Vn为元数据向量;Sn为超平面;$Vn \cdot Sn$为点积结果;X_1为向量横坐标;Y_1为向量纵坐标;Z_1为向量竖坐标;X_2为超平面横坐标;Y_2为超平面纵坐标;Z_2为超平面竖坐标。

[0101] 可选地,在本申请实施例中,元数据分析装置,数据映射模块,还用于基于向量分类结果,将同一类别的元数据向量存储至同一数据桶中;利用哈希地图,存储元数据向量对应的元数据与数据桶的映射关系。

[0102] 可选地,在本申请实施例中,元数据分析装置,分析模块230,具体用于获得目标数据桶中的目标元数据向量;将待分析元数据与每一目标元数据向量进行相似度计算,获得每一目标元数据向量对应的相似度数据;将相似度数据进行排序,基于相似度排序结果获得数据分析结果。

[0103] 应理解的是,该装置与上述的元数据分析方法实施例对应,能够执行上述方法实施例涉及的各个步骤,该装置具体的功能可以参见上文中的描述,为避免重复,此处适当省略详细描述。该装置包括至少一个能以软件或固件(firmware)的形式存储于存储器中或固化在装置的操作系统(operating system,OS)中的软件功能模块。

[0104] 请参见图4示出的本申请实施例提供的电子设备的结构示意图。本申请实施例提供的一种电子设备300,包括:处理器310和存储器320,存储器320存储有处理器310可执行的机器可读指令,机器可读指令被处理器310执行时执行如上的方法。

[0105] 本申请实施例还提供了一种存储介质,该存储介质上存储有计算机程序,该计算机程序被处理器运行时执行如上的方法。

[0106] 其中,存储介质可以由任何类型的易失性或非易失性存储设备或者它们的组合实

现,如静态随机存取存储器(Static Random Access Memory,简称SRAM),电可擦除可编程只读存储器(Electrically Erasable Programmable Read-Only Memory,简称EEPROM),可擦除可编程只读存储器(Erasable Programmable Read Only Memory,简称EPROM),可编程只读存储器(Programmable Red-Only Memory,简称PROM),只读存储器(Read-Only Memory,简称ROM),磁存储器,快闪存储器,磁盘或光盘。

[0107] 本申请实施例所提供的几个实施例中,应该理解到,所揭露的装置和方法,也可以通过其他的方式实现。以上所描述的装置实施例仅仅是示意性的,例如,附图中的流程图和框图显示了根据本申请实施例的多个实施例的装置、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分,模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现方式中,方框中所标注的功能也可以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0108] 另外,在本申请实施例各个实施例中的各功能模块可以集成在一起形成一个独立的部分,也可以是各个模块单独存在,也可以两个或两个以上模块集成形成一个独立的部分。

[0109] 以上的描述,仅为本申请实施例的可选实施方式,但本申请实施例的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请实施例揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请实施例的保护范围之内。

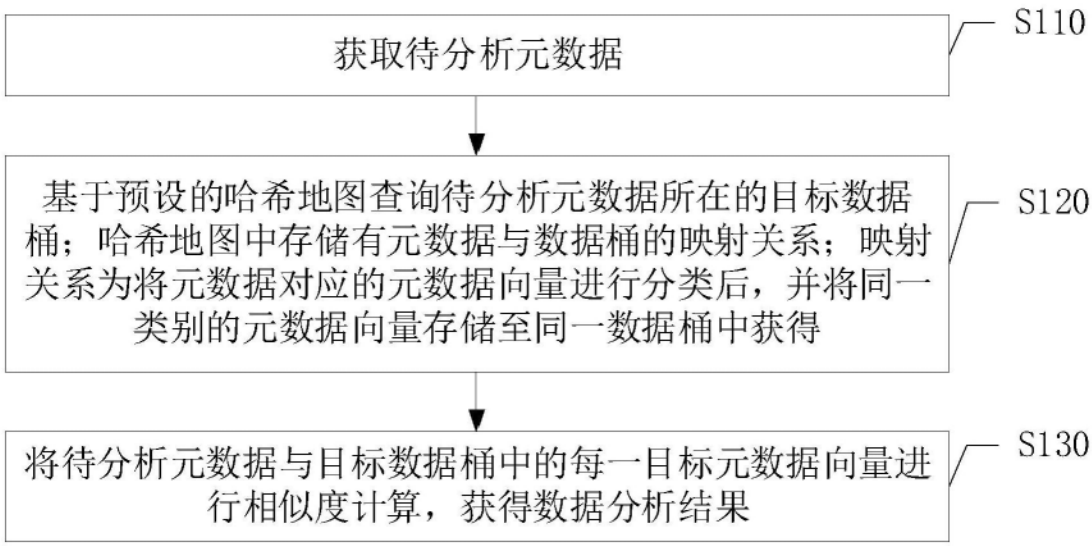

获取待分析元数据
S110
基于预设的哈希地图查询待分析元数据所在的目标数据桶；哈希地图中存储有元数据与数据桶的映射关系；映射关系为将元数据对应的元数据向量进行分类后，并将同一类别的元数据向量存储至同一数据桶中获得
S120
将待分析元数据与目标数据桶中的每一目标元数据向量进行相似度计算，获得数据分析结果
S130

图1

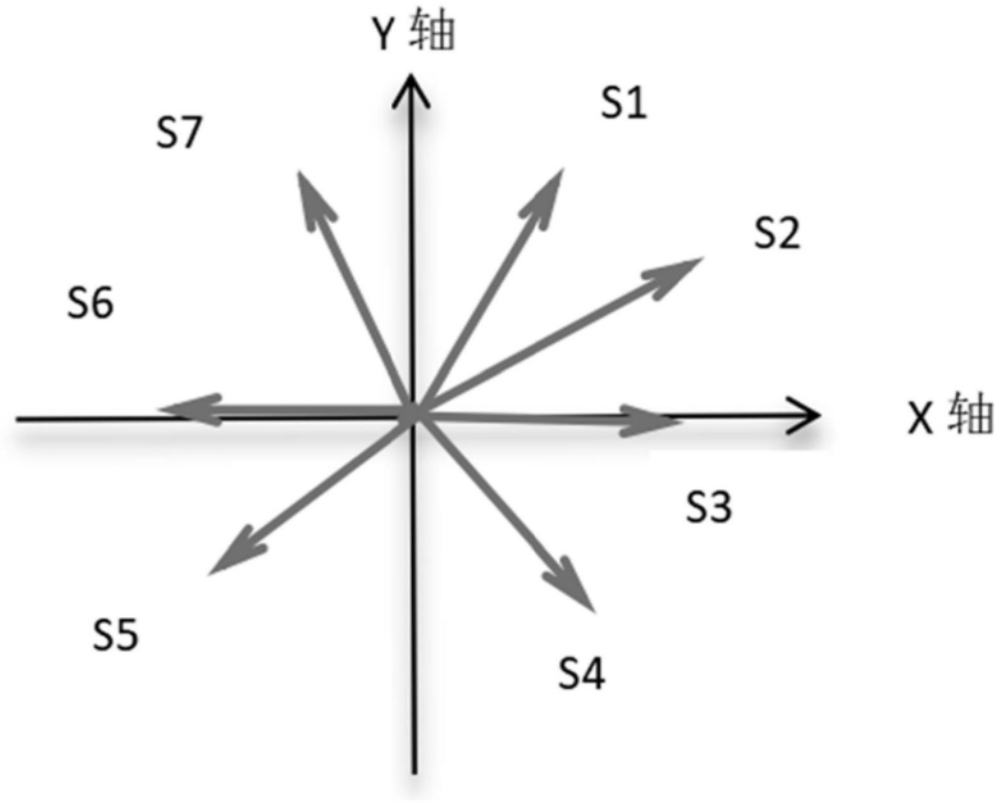

Y 轴
X 轴
S1
S2
S3
S4
S5
S6
S7

图2

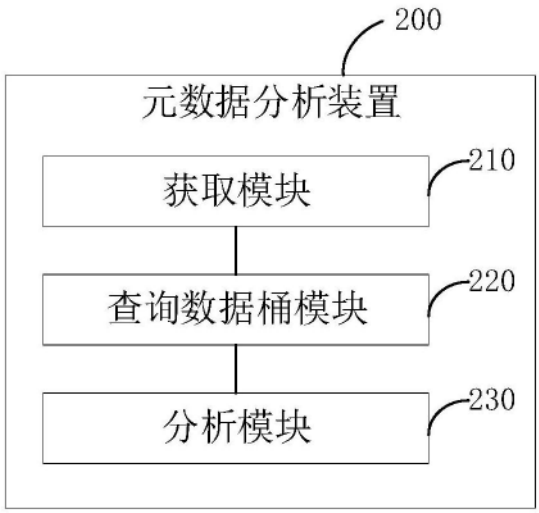

200
元数据分析装置
210
获取模块
220
查询数据桶模块
230
分析模块

图3

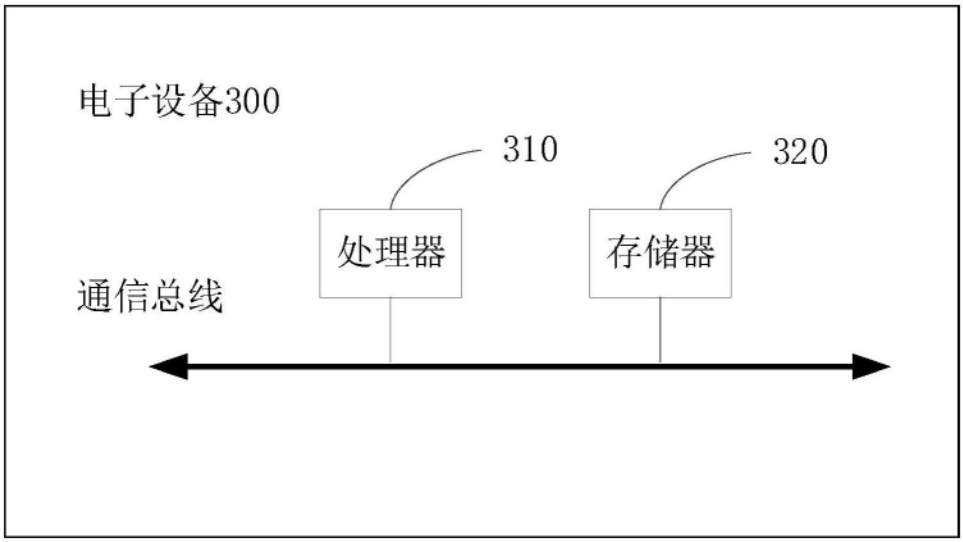

电子设备300
310
320
处理器
存储器
通信总线

图4