(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112347320 B
(45)授权公告日 2024.08.06

(21)申请号 202011226149.3
(22)申请日 2020.11.05
(65)同一申请的已公布的文献号
申请公布号 CN 112347320 A
(43)申请公布日 2021.02.09
(73)专利权人 杭州数梦工场科技有限公司
地址 310024 浙江省杭州市转塘科技经济区块16号4幢326室
(72)发明人 魏良宵 徐鹏飞 周轶凡
(74)专利代理机构 北京博思佳知识产权代理有限公司 11415
专利代理师 刘秀玲
(51)Int.Cl.
G06F 16/903(2019.01)
G06F 16/9035(2019.01)
G06F 18/22(2023.01)
G06N 3/0442(2023.01)
(56)对比文件
CN 105488212 A,2016.04.13
CN 107609461 A,2018.01.19
审查员 李萌

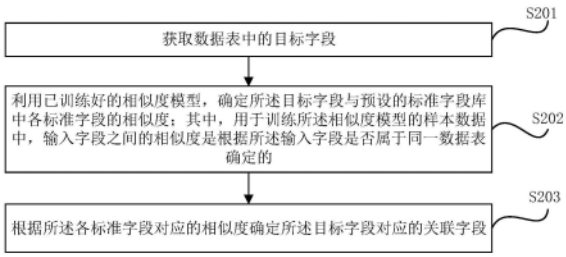
权利要求书2页 说明书8页 附图2页

(54)发明名称

数据表字段的关联字段推荐方法及装置

(57)摘要

本发明实施例提供一种数据表字段的关联字段推荐方法及装置。本发明实施例通过获取数据表中的目标字段,利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度,其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,利用数据表内字段的非冗余性构造训练的样本数据,提高了相似度模型的相似度计算结果的准确性,使得推荐的关联字段准确性更高,进而提高了数据表字段与标准字段的匹配准确性。



CN 112347320 B

1.一种数据表字段的关联字段推荐方法,其特征在于,包括:

获取数据表中的目标字段;

利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的,若所述输入字段属于同一数据表,则确定所述输入字段之间的相似度等于预设的相似度区间的最小值;

根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

2.根据权利要求1所述的方法,其特征在于,所述相似度模型的获取过程,包括:

设置机器学习模型;

构造样本数据,所述样本数据包括输入字段和标签相似度,所述标签相似度为所述输入字段之间的相似度;其中,若所述输入字段属于同一数据表,则确定所述标签相似度等于预设的相似度区间的最小值;

利用所述样本数据对所述机器学习模型进行训练,得到训练完毕的机器学习模型,以所述训练完毕的机器学习模型作为相似度模型。

3.根据权利要求1所述的方法,其特征在于,所述输入字段均为数据表中的字段;或者,所述输入字段包括数据表中的字段和标准字段库中的标准字段。

4.根据权利要求1所述的方法,其特征在于,若所述输入字段属于不同数据表,则根据预设相似度计算方式计算所述输入字段之间的相似度,作为标签相似度。

5.根据权利要求1所述的方法,其特征在于,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,包括:

将所述各标准字段对应的相似度按照数值进行排序;

根据排序结果,从所述标准字段库中提取相似度最大的设定数目个标准字段,作为所述目标字段对应的关联字段。

6.一种数据表字段的关联字段推荐装置,其特征在于,包括:

获取模块,用于获取数据表中的目标字段;

相似度确定模块,用于利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的,若所述输入字段属于同一数据表,则确定所述输入字段之间的相似度等于预设的相似度区间的最小值;

关联字段确定模块,用于根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

7.根据权利要求6所述的装置,其特征在于,所述相似度模型的获取过程,包括:

设置机器学习模型;

构造样本数据,所述样本数据包括输入字段和标签相似度,所述标签相似度为所述输入字段之间的相似度;其中,若所述输入字段属于同一数据表,则确定所述标签相似度等于预设的相似度区间的最小值;

利用所述样本数据对所述机器学习模型进行训练,得到训练完毕的机器学习模型,以所述训练完毕的机器学习模型作为相似度模型。

8.根据权利要求6所述的装置,其特征在于,所述输入字段均为数据表中的字段;或者,

所述输入字段包括数据表中的字段和标准字段库中的标准字段。

9.根据权利要求6所述的装置,其特征在于,若所述输入字段属于不同数据表,则根据预设相似度计算方式计算所述输入字段之间的相似度,作为标签相似度。

10.根据权利要求6所述的装置,其特征在于,所述关联字段确定模块具体用于:

将所述各标准字段对应的相似度按照数值进行排序;

根据排序结果,从所述标准字段库中提取相似度最大的设定数目个标准字段,作为所述目标字段对应的关联字段。

数据表字段的关联字段推荐方法及装置

技术领域

[0001] 本发明涉及数据处理技术领域,尤其涉及一种数据表字段的关联字段推荐方法及装置。

背景技术

[0002] 在政务行业的数据标准化过程中,需要将数据表(也称为物理表)中的字段与给定的标准字段进行关联匹配。在现实场景中,数据表字段的数量是非常庞大的,用人工将每个数据表字段与标准字段进行匹配是不切实际的,人力投入将非常大。

[0003] 相关技术中,采用常规的机器学习方法,利用数据表字段的词向量生成特征向量,通过将该特征向量与特征库中的特征向量进行相似度计算,根据相似度计算结果确实是否匹配。该技术中,使用通用场景的语言模型生成的词向量作为特征向量进行相似度计算,在数据表字段的匹配场景中匹配准确性较低。

发明内容

[0004] 为克服相关技术中存在的问题,本发明提供了一种数据表字段的关联字段推荐方法及装置,提高数据表字段与标准字段的匹配准确性。

[0005] 根据本发明实施例的第一方面,提供一种数据表字段的关联字段推荐方法,包括:

[0006] 获取数据表中的目标字段;

[0007] 利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的;

[0008] 根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

[0009] 根据本发明实施例的第二方面,提供一种数据表字段的关联字段推荐装置,包括:

[0010] 获取模块,用于获取数据表中的目标字段;

[0011] 相似度确定模块,用于利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的;

[0012] 关联字段确定模块,用于根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

[0013] 本发明实施例提供的技术方案可以包括以下有益效果:

[0014] 本发明实施例,通过获取数据表中的目标字段,利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度,其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,利用数据表内字段的非冗余性构造训练的样本数据,提高了相似度模型的相似度计算结果的准确性,使得推荐的关联字段准确性更高,进而提高了数据表字段与标准字段的匹配准确性。

[0015] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本说明书。

附图说明

[0016] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本说明书的实施例,并与说明书一起用于解释本说明书的原理。

[0017] 图1是数据表字段的匹配场景示意图。

[0018] 图2是本发明实施例提供的数据表字段的关联字段推荐方法的流程示例图。

[0019] 图3是本发明实施例提供的数据表字段的关联字段推荐装置的功能方块图。

[0020] 图4是本发明实施例提供的电子设备的一个硬件结构图。

具体实施方式

[0021] 这里将详细地对示例性实施例进行说明,其示例表示在附图中。下面的描述涉及附图时,除非另有表示,不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反,它们仅是与如所附权利要求书中所详述的、本发明实施例的一些方面相一致的装置和方法的例子。

[0022] 在本发明实施例使用的术语是仅仅出于描述特定本发明实施例的目的,而非旨在限制本发明实施例。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式,除非上下文清楚地表示其他含义。还应当理解,本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0023] 应当理解,尽管在本发明实施例可能采用术语第一、第二、第三等来描述各种信息,但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如,在不脱离本发明实施例范围的情况下,第一信息也可以被称为第二信息,类似地,第二信息也可以被称为第一信息。取决于语境,如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0024] 在一些应用场景中,例如政府部门,各部门都有自己存储数据的数据库,不同部门的数据库中,数据表的字段名称定义各有不同。如果要打通各部门之间的数据壁垒,则需要将个部门之间的数据进行整合,这时需要将不同部门之间不同名称但含义相同的数据表字段进行融合。

[0025] 在对数据表进行标准化过程中,通过将数据表字段替换为与之匹配的标准字段,可以使不同数据表中名称不同但含义相同的数据表字段对应到同一个标准字段,为实现数据融合提供基础。

[0026] 在处理这类问题时,通常预先设置一个标准字段库,标准字段库中包括多个标准字段,标准字段是人工标注的字段,标准字段也可以称为数据元,相应地,标准字段库也可以称为数据元库。然后将数据表字段与标准字段库中的各个标准字段一一进行比对,以便找到与数据表字段匹配的标准字段。

[0027] 图1是数据表字段的匹配场景示意图。如图1所示,数据表字段需要与标准字段库中的每个标准字段一一计算相似度,然后根据相似度计算结果确定与数据表字段匹配的标准字段。

[0028] 相关技术中,获取数据表字段与标准字段的相似度的过程是:使用通用场景的语言模型生成数据表字段的词向量,作为数据表字段对应的特征向量,以及生成标准字段的词向量,作为标准字段对应的特征向量,然后计算该两个特征向量的相似度。接着,根据相似度从标准字段库中获得与数据表字段匹配的标准字段。

[0029] 由于通用场景的语言模型是基于通用场景训练的,其训练的样本数据是通用场景中的数据,既包括数据表字段,也包括非数据表字段(即不属于数据表字段的字段),数据表字段在样本数据中只占一部分。因此通用场景的语言模型对于数据表字段的匹配针对性弱,从而使得相关技术在数据表字段的匹配场景中匹配准确性较低。

[0030] 例如,数据表1中包括字段“电话号码”,数据表2中包括字段“手机号码”,该两个字段内容均为手机号码,即含义相同,标准字段库中与之对应的标准字段为“号码”,但字段“电话号码”的特征向量(电话,号码)与标准字段“号码”的特征向量(号码)的相似度数值却不大,从而使得字段“电话号码”无法准确匹配到标准字段“号码”。同理,字段“手机号码”也无法准确匹配到标准字段“号码”。

[0031] 再比如,数据表3中包括字段“住宅电话号码”和“办公电话号码”,按照相关技术,该两个字段均会匹配到标准字段“电话号码”,但是这两个字段位于同一数据表内,其实际含义显然不同,是不应该匹配到同一个标准字段的。

[0032] 本发明实施例针对数据表字段的匹配场景,基于同一数据表内部字段的差异性和不同数据表间的字段的相关性构造训练相似度模型的样本数据,获得专门针对数据表字段匹配场景的相似度模型,利用该相似度模型计算数据表字段与标准字段的相似度,以提高匹配准确性。

[0033] 下面通过实施例对本发明提供的数据表字段的关联字段推荐方法进行详细说明。

[0034] 图2是本发明实施例提供的数据表字段的关联字段推荐方法的流程示例图。如图2所示,数据表字段的关联字段推荐方法可以包括:

[0035] S201,获取数据表中的目标字段。

[0036] S202,利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的。

[0037] S203,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

[0038] 其中,目标字段是数据表中需要与标准字段匹配的字段。

[0039] 在应用过程中,可以对数据表中的所有字段一一按照对目标字段的处理方式进行处理。例如,假设一张数据表中有m个字段(字段1、字段2……字段m),标准字段库中有d个标准字段(标准字段1、标准字段2……标准字段d),则将m个字段中的每个字段,分别与d个标准字段进行相似度计算。以字段1为例,利用相似度模型,分别获取字段1与标准字段1、标准字段2……标准字段m的相似度,得到m个相似度。

[0040] 其中,利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度,是将目标字段与标准字段输入相似度模型,相似度模型输出的数据即为目标字段与该标准字段的相似度。

[0041] 例如,将前述的字段1和标准字段1输入相似度模型,相似度模型的输出为字段1和标准字段1的相似度,……将前述的字段1和标准字段d输入相似度模型,相似度模型的输出

为字段1和标准字段d的相似度。

[0042] 相似度模型有两个输入,一个输出。

[0043] 在训练过程中,相似度模型的两个输入可以是任意两个数据表字段,还可以是一个数据表字段和一个标准字段。

[0044] 在训练好之后的应用过程中,相似度模型的两个输入中一个是数据表字段,另一个是标准字段库中的标准字段。

[0045] 其中,用于训练所述相似度模型的样本数据包括两个输入字段和一个标签相似度,标签相似度是两个输入字段的已知的相似度。

[0046] 标签相似度是根据两个输入字段是否属于同一数据表确定的。即当两个输入字段属于同一数据表时,认为两个输入字段最不相似,此时可以确定标签相似度等于预设的相似度区间的最小值。这是因为在设计理念上,数据表内部的字段需要避免冗余,即每张数据表内部的不同字段的含义是相差很大的,或者说含义是不同的。当两个输入字段属于不同数据表时,认为两个输入字段有可能相似。

[0047] 举例说明。假设相似度区间为[0,1],相似度为0表示最不相似,相似度为1表示最相似。则前述的数据表3中的字段“住宅电话号码”和“办公电话号码”对应的标签相似度设置为0,前述的数据表1中的字段“电话号码”和数据表2中的字段“手机号码”对应的标签相似度可以设置为字段“电话号码”和“手机号码”的余弦相似度(还可以是用其他相似度计算方式计算出的“电话号码”和“手机号码”的相似度,此处仅为举例,本实施例对标签相似度的计算方式不作限制)。

[0048] 这样,训练之后的相似度模型就不会将同一数据表内的不同字段匹配到相同的标准字段,例如前述数据表3中的字段“住宅电话号码”和“办公电话号码”,提高匹配的准确性。

[0049] 在一个示例性的实现过程中,所述相似度模型的获取过程,可以包括:

[0050] 设置机器学习模型;

[0051] 构造样本数据,所述样本数据包括输入字段和标签相似度,所述标签相似度为所述输入字段之间的相似度;其中,若所述输入字段属于同一数据表,则确定所述标签相似度等于预设的相似度区间的最小值;

[0052] 利用所述样本数据对所述机器学习模型进行训练,得到训练完毕的机器学习模型,以所述训练完毕的机器学习模型作为相似度模型。

[0053] 其中,机器学习模型中可以包括LSTM(Long-Short Term Memory,长短记忆)神经网络模型和相似度算法模型,LSTM神经网络模型用于计算输入字段对应的向量。其中,相似度算法模型可以采用余弦相似度算法,还可以采用其他的用于计算文本相似度的算法,例如欧几里得距离算法,曼哈顿距离算法等。本实施例对相似度算法模型采用的相似度算法不作限制。

[0054] 其中,输入字段均为数据表中的字段;或者,所述输入字段包括数据表中的字段和标准字段库中的标准字段。

[0055] 其中,若所述输入字段属于不同数据表,则根据预设相似度计算方式计算所述输入字段之间的相似度,作为所述标签相似度。

[0056] 例如,预设相似度计算方式可以是前述的余弦相似度算法、欧几里得距离算法、曼

哈顿距离算法等等。

[0057] 假设相似度区间为[0,1],样本数据中的两个输入字段为字段a和字段b。其中,字段a来自数据表A,字段b来自数据表B,如果数据表A和数据表B为不同数据表,则字段a和字段b的相似度sim(a,b)为:

[0058] sim(a,b)＝cos(vec(a),vec(b))

[0059] 其中,vec(a)为字段a的向量,vec(b)为字段b的向量。cos(vec(a),vec(b))表示向量vec(a)和vec(b)的余弦相似度。

[0060] 如果数据表A和数据表B为不同数据表,则sim(a,b)＝0。

[0061] 则(字段a,字段b,sim(a,b))为样本数据。利用该样本数据,训练相似度模型的过程可以如下:

[0062] 训练过程中机器学习模型包括LSTM神经网络模型和余弦相似度算法模型,第一组样本数据对应的参数值为初始参数值,经第j组样本数据训练调整后的参数值为第j+1组样本数据对应的参数值,j为自然数,且j≥1;在每组样本数据的训练中执行如下操作:

[0063] 利用LSTM神经网络模型对字段a和字段b分别进行编码,得到向量vec(a)和vec(b);

[0064] 利用余弦相似度算法模型计算cos(vec(a),vec(b));

[0065] 计算cos(vec(a),vec(b))与样本数据中sim(a,b)之间的差值;

[0066] 判断所述差值是否小于预设阈值,若小于则停止训练,将本组样本数据对应的参数值作为训练好的机器学习模型的参数值;否则,根据所述差值调整机器学习模型的参数值,转入下一组样本数据的训练。

[0067] 在一个示例性的实现过程中,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,可以包括:

[0068] 将所述各标准字段对应的相似度按照数值进行排序;

[0069] 根据排序结果,从所述标准字段库中提取相似度最大的设定数目个标准字段,作为所述目标字段对应的关联字段。

[0070] 例如,假设标准字段库中共有d个标准字段,将目标字段与d个标准字段对应的相似度S1、S2……S_d按照从大到小的顺序排列,取前k(k为自然数)个相似度对应的标准字段作为目标字段对应的关联字段。后续开发人员可以从该k个关联字段中人工确定与目标字段匹配的标准字段。

[0071] 本发明实施例提供的数据表字段的关联字段推荐方法,通过获取数据表中的目标字段,利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度,其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,利用数据表内字段的非冗余性构造训练的样本数据,提高了相似度模型的相似度计算结果的准确性,使得推荐的关联字段准确性更高,进而提高了数据表字段与标准字段的匹配准确性。

[0072] 基于上述的方法实施例,本发明实施例还提供了相应的装置、设备及存储介质实施例。关于本发明实施例的装置、设备及存储介质实施例的详细实现方式,请参见前述方法实施例部分的相应说明。

[0073] 图3是本发明实施例提供的数据表字段的关联字段推荐装置的功能方块图。如图3所示,本实施例中,数据表字段的关联字段推荐装置可以包括:

[0074] 获取模块310,用于获取数据表中的目标字段;

[0075] 相似度确定模块320,用于利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的;

[0076] 关联字段确定模块330,用于根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

[0077] 在一个示例性的实现过程中,所述相似度模型的获取过程,包括:

[0078] 设置机器学习模型;

[0079] 构造样本数据,所述样本数据包括输入字段和标签相似度,所述标签相似度为所述输入字段之间的相似度;其中,若所述输入字段属于同一数据表,则确定所述标签相似度等于预设的相似度区间的最小值;

[0080] 利用所述样本数据对所述机器学习模型进行训练,得到训练完毕的机器学习模型,以所述训练完毕的机器学习模型作为相似度模型。

[0081] 在一个示例性的实现过程中,所述输入字段均为数据表中的字段;或者,所述输入字段包括数据表中的字段和标准字段库中的标准字段。

[0082] 在一个示例性的实现过程中,若所述输入字段属于不同数据表,则根据预设相似度计算方式计算所述输入字段之间的相似度,作为所述标签相似度。

[0083] 在一个示例性的实现过程中,所述关联字段确定模块330可以具体用于:

[0084] 将所述各标准字段对应的相似度按照数值进行排序;

[0085] 根据排序结果,从所述标准字段库中提取相似度最大的设定数目个标准字段,作为所述目标字段对应的关联字段。

[0086] 本发明实施例还提供了一种电子设备。图4是本发明实施例提供的电子设备的一个硬件结构图。如图4所示,电子设备包括:内部总线401,以及通过内部总线连接的存储器402,处理器403和外部接口404。

[0087] 所述处理器403,用于读取存储器402上的机器可读指令,并执行所述指令以实现如下操作:

[0088] 获取数据表中的目标字段;

[0089] 利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的;

[0090] 根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

[0091] 在一个示例性的实现过程中,所述相似度模型的获取过程,包括:

[0092] 设置机器学习模型;

[0093] 构造样本数据,所述样本数据包括输入字段和标签相似度,所述标签相似度为所述输入字段之间的相似度;其中,若所述输入字段属于同一数据表,则确定所述标签相似度等于预设的相似度区间的最小值;

[0094] 利用所述样本数据对所述机器学习模型进行训练,得到训练完毕的机器学习模

型,以所述训练完毕的机器学习模型作为相似度模型。

[0095] 在一个示例性的实现过程中,所述输入字段均为数据表中的字段;或者,所述输入字段包括数据表中的字段和标准字段库中的标准字段。

[0096] 在一个示例性的实现过程中,若所述输入字段属于不同数据表,则根据预设相似度计算方式计算所述输入字段之间的相似度,作为所述标签相似度。

[0097] 在一个示例性的实现过程中,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,包括:

[0098] 将所述各标准字段对应的相似度按照数值进行排序;

[0099] 根据排序结果,从所述标准字段库中提取相似度最大的设定数目个标准字段,作为所述目标字段对应的关联字段。

[0100] 本发明实施例还提供一种计算机可读存储介质,所述计算机可读存储介质上存储有若干计算机指令,所述计算机指令被执行时进行如下处理:

[0101] 获取数据表中的目标字段;

[0102] 利用已训练好的相似度模型,确定所述目标字段与预设的标准字段库中各标准字段的相似度;其中,用于训练所述相似度模型的样本数据中,输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的;

[0103] 根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段。

[0104] 在一个示例性的实现过程中,所述相似度模型的获取过程,包括:

[0105] 设置机器学习模型;

[0106] 构造样本数据,所述样本数据包括输入字段和标签相似度,所述标签相似度为所述输入字段之间的相似度;其中,若所述输入字段属于同一数据表,则确定所述标签相似度等于预设的相似度区间的最小值;

[0107] 利用所述样本数据对所述机器学习模型进行训练,得到训练完毕的机器学习模型,以所述训练完毕的机器学习模型作为相似度模型。

[0108] 在一个示例性的实现过程中,所述输入字段均为数据表中的字段;或者,所述输入字段包括数据表中的字段和为标准字段库中的标准字段。

[0109] 在一个示例性的实现过程中,若所述输入字段属于不同数据表,则根据预设相似度计算方式计算所述输入字段之间的相似度,作为所述标签相似度。

[0110] 在一个示例性的实现过程中,根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段,包括:

[0111] 将所述各标准字段对应的相似度按照数值进行排序;

[0112] 根据排序结果,从所述标准字段库中提取相似度最大的设定数目个标准字段,作为所述目标字段对应的关联字段。

[0113] 对于装置和设备实施例而言,由于其基本对应于方法实施例,所以相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的模块可以是或者也可以不是物理上分开的,作为模块显示的部件可以是或者也可以不是物理模块,即可以位于一个地方,或者也可以分布到多个网络模块上。可以根据实际的需要选择其中的部分或者全部模块来实现本说明书方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0114] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[0115] 本领域技术人员在考虑说明书及实践这里申请的发明后,将容易想到本说明书的其它实施方案。本说明书旨在涵盖本说明书的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本说明书的一般性原理并包括本说明书未申请的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本说明书的真正范围和精神由下面的权利要求指出。

[0116] 应当理解的是,本说明书并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本说明书的范围仅由所附的权利要求来限制。

[0117] 以上所述仅为本说明书的较佳实施例而已,并不用以限制本说明书,凡在本说明书的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本说明书保护的范围之内。

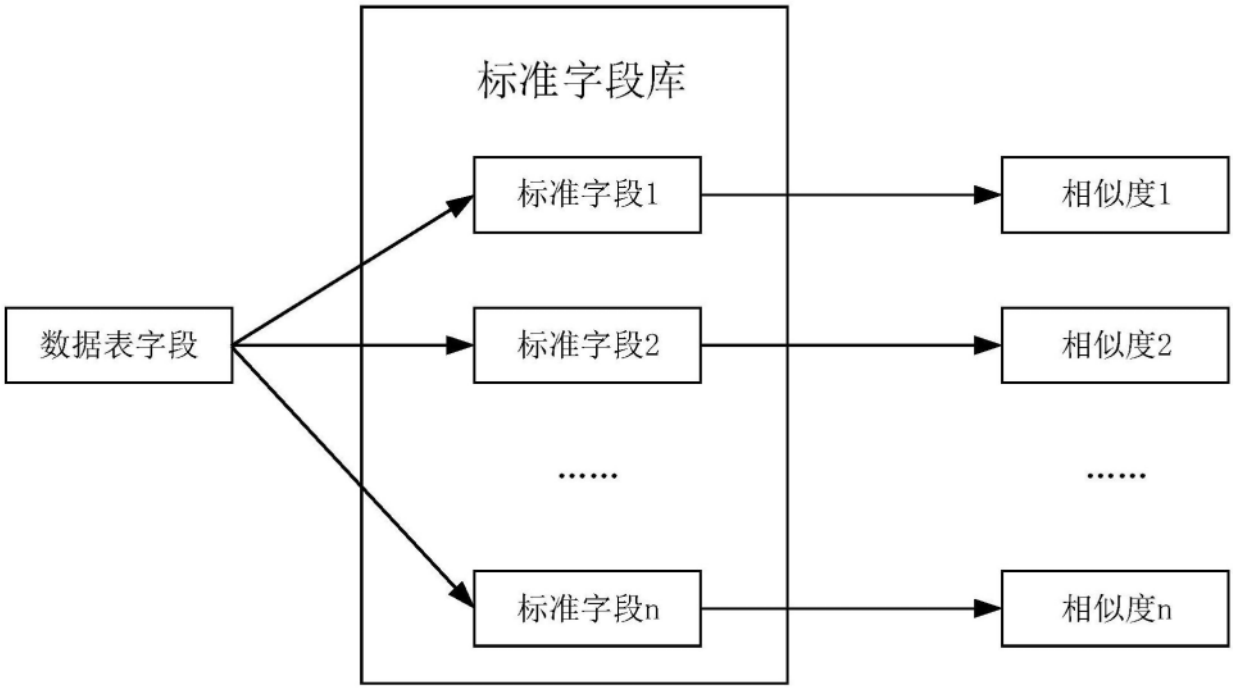

标准字段库
数据表字段
标准字段1
标准字段2
.....
标准字段n
相似度1
相似度2
.....
相似度n

图1

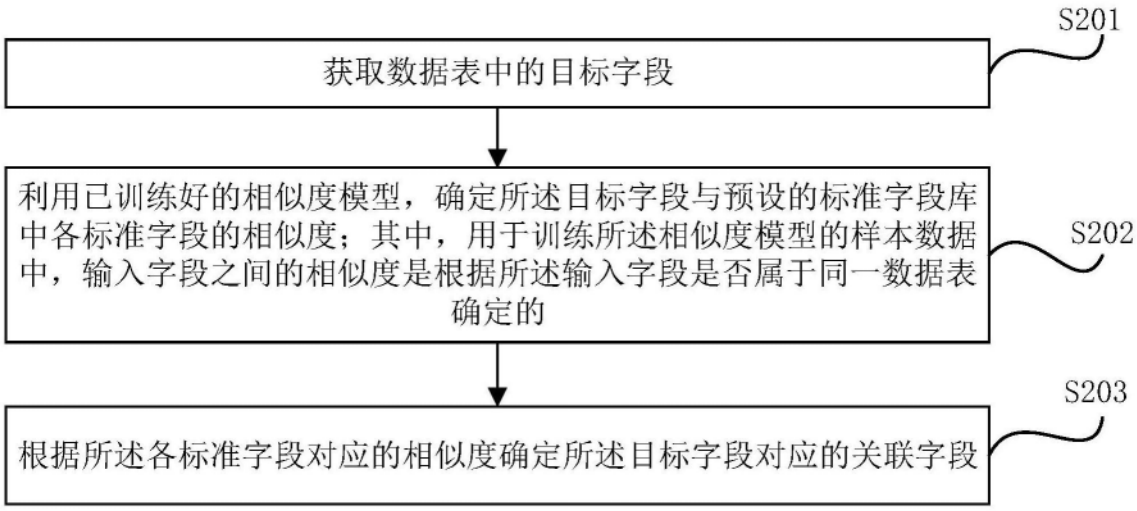

S201
获取数据表中的目标字段
利用已训练好的相似度模型，确定所述目标字段与预设的标准字段库中各标准字段的相似度；其中，用于训练所述相似度模型的样本数据中，输入字段之间的相似度是根据所述输入字段是否属于同一数据表确定的
S202
S203
根据所述各标准字段对应的相似度确定所述目标字段对应的关联字段

图2

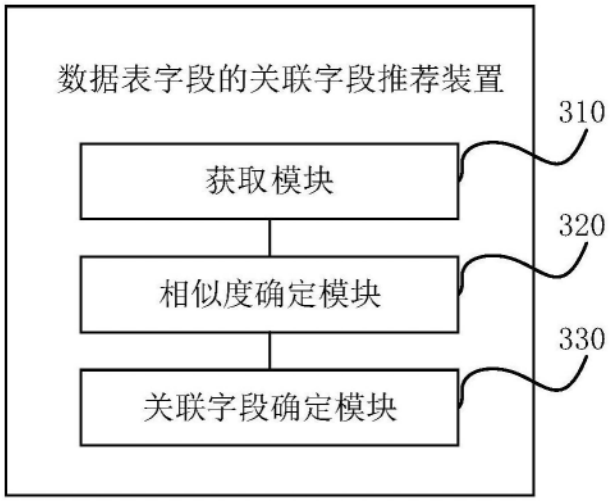

数据表字段的关联字段推荐装置
310
获取模块
320
相似度确定模块
330
关联字段确定模块

图3

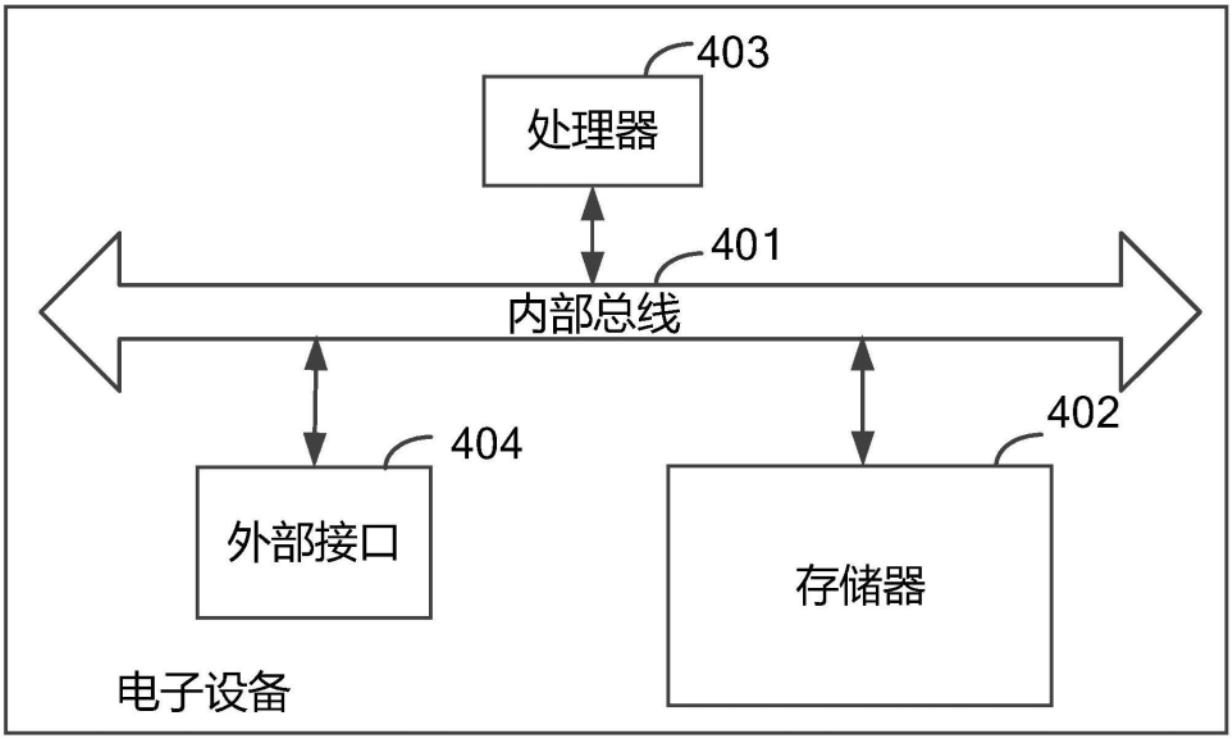

403
处理器
401
内部总线
404
外部接口
402
存储器
电子设备

图4