(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115203206 A
(43)申请公布日 2022.10.18

(21)申请号 202210823548.0

(22)申请日 2022.07.13

(71)申请人 树根互联股份有限公司
地址 510000 广东省广州市海珠区琶洲大道东路3号303-309房

(72)发明人 李开金 谭振海 刘伏桃 李建民

(74)专利代理机构 北京超凡宏宇专利代理事务所(特殊普通合伙) 11463
专利代理师 梁韬

(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/242(2019.01)
G06F 16/2457(2019.01)
G06F 16/28(2019.01)

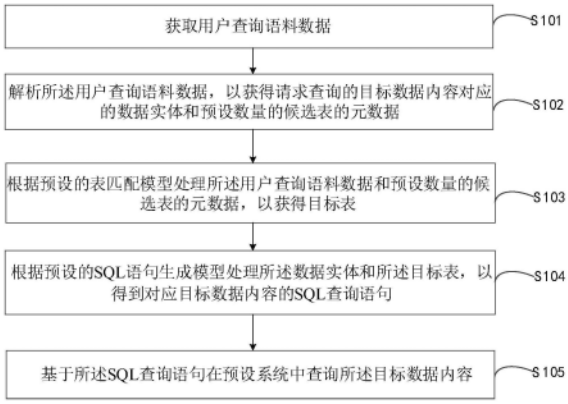
权利要求书2页 说明书10页 附图2页

(54)发明名称

数据内容搜索方法、装置、计算机设备及可读存储介质

(57)摘要

本发明实施例公开了一种数据内容搜索方法、装置、计算机设备及可读存储介质，所述方法包括：获取用户查询语料数据；解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和候选表的元数据；根据预设的表匹配模型处理所述用户查询语料数据和候选表的元数据，以获得目标表；根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句；基于所述SQL查询语句在预设系统中查询所述目标数据内容。通过表匹配模型和预设的SQL语句生成模型的处理，能够更加快速、准确的获取目标数据内容。



CN 115203206 A

1.一种数据内容搜索方法,其特征在于,所述方法包括:

获取用户查询语料数据;

解析所述用户查询语料数据,以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据;

根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据,以获得目标表;

根据预设的SQL语句生成模型处理所述数据实体和所述目标表,以得到对应目标数据内容的SQL查询语句;

基于所述SQL查询语句在预设系统中查询所述目标数据内容。

2.根据权利要求1所述的数据内容搜索方法,其特征在于,解析所述用户查询语料数据,以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据的步骤,包括:

利用命名实体识别模型处理所述用户查询语料数据,以得到对应的数据实体;

根据所述数据实体在码表库和表元数据库中进行弹性搜索,以得到所述预设数量的候选表的元数据。

3.根据权利要求2所述数据内容搜索方法,其特征在于,所述命名实体识别模型包括预设的BERT模型和条件随机场模型,基于命名实体识别模型处理所述用户查询语料数据,以得到对应的数据实体的步骤,包括:

基于所述预设的BERT模型处理所述用户查询语料数据,以得到具有语义标签的初始编码序列;

基于所述条件随机场模型解码所述初始编码序列,以获得对应的数据实体。

4.根据权利要求3所述的数据内容搜索方法,其特征在于,所述预设的BERT模型的编码层采用RoBERTa预训练模型。

5.根据权利要求1所述的数据内容搜索方法,其特征在于,根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据,以获得目标表的步骤,包括:

基于预设的BERT模型处理所述用户查询语料数据,以得到第一语义向量;

基于预设的BERT模型处理所述预设数量的候选表的元数据,以得到预设数量的第二语义向量;

基于深度匹配算法处理所述用户查询语料数据对应的第一语义向量和预设数量的候选表对应的所述第二语义向量,以计算所述用户查询语料数据与各候选表的匹配分数;

选择匹配分数最高的候选表作为所述目标表。

6.根据权利要求1所述的数据内容搜索方法,其特征在于,根据预设的SQL语句生成模型处理所述数据实体和所述目标表,以得到对应目标数据内容的SQL查询语句的步骤,包括:

按照所述目标表的列顺序将所述用户查询语料数据和所述目标表的列名进行拼接,以得到预设数量的待编码数据;

基于预设的BERT模型对各所述待编码数据进行编码处理,以得到预设数量的编码向量;

基于预设的二分类模型处理全部所述编码向量,以得到第一部分组合语句和第二部分

组合语句，其中，所述第一部分组合语句为对应所述目标数据内容的组合语句，所述第二部分组合语句为待删除的组合语句；

合并全部第一部分组合语句，以得到所述SQL查询语句。

7.一种数据内容搜索装置，其特征在于，所述数据内容搜索装置包括：

获取模块，用于获取用户查询语料数据；

解析模块，用于解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据；

第一处理模块，用于根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表；

第二处理模块，用于根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句；

查询模块，用于基于所述SQL查询语句在预设系统中查询所述目标数据内容。

8.根据权利要求7所述的数据内容搜索装置，其特征在于，所述解析模块，具体用于利用命名实体识别模型处理所述用户查询语料数据，以得到对应的数据实体；根据所述数据实体在码表库和表元数据库中进行弹性搜索，以得到所述预设数量的候选表的元数据。

9.一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器存储有计算机程序，所述计算机程序在所述处理器上运行时执行权利要求1至6任一项所述的数据内容搜索方法。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机程序，所述计算机程序在处理器上运行时执行权利要求1至6中任一项所述的数据内容搜索方法。

数据内容搜索方法、装置、计算机设备及可读存储介质

技术领域

[0001] 本发明涉及自然语言处理领域，尤其涉及一种数据内容搜索方法、装置、计算机设备及可读存储介质。

背景技术

[0002] 针对制造业的数据搜索过程，由于数据表数量较多、涉及的领域复杂、同领域的表数据很相似，多个表的数据具有相同的取值，表名、列名和数据名相近等原因，会使得数据搜索过程存在大量干扰。

[0003] 现有的数据搜索方案，无法实现快速、准确的搜索数据内容。

发明内容

[0004] 为了解决上述技术问题，本申请实施例提供了一种数据内容搜索方法、装置、计算机设备及可读存储介质，具体方案如下：

[0005] 第一方面，本申请实施例提供了一种数据内容搜索方法，所述方法包括：

[0006] 获取用户查询语料数据；

[0007] 解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据；

[0008] 根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表；

[0009] 根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句；

[0010] 基于所述SQL查询语句在预设系统中查询所述目标数据内容。

[0011] 根据本申请实施例的一种具体实施方式，解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据的步骤，包括：

[0012] 利用命名实体识别模型处理所述用户查询语料数据，以得到对应的数据实体；

[0013] 根据所述数据实体在码表库和表元数据库中进行弹性搜索，以得到所述预设数量的候选表的元数据。

[0014] 根据本申请实施例的一种具体实施方式，所述命名实体识别模型包括预设的BERT模型和条件随机场模型，基于命名实体识别模型处理所述用户查询语料数据，以得到对应的数据实体的步骤，包括：

[0015] 基于所述预设的BERT模型处理所述用户查询语料数据，以得到具有语义标签的初始编码序列；

[0016] 基于所述条件随机场模型解码所述初始编码序列，以获得对应的数据实体。

[0017] 根据本申请实施例的一种具体实施方式，所述预设的BERT模型的编码层采用RoBERTa预训练模型。

[0018] 根据本申请实施例的一种具体实施方式，根据预设的表匹配模型处理所述用户查

询语料数据和预设数量的候选表的元数据，以获得目标表的步骤，包括：

[0019] 基于预设的BERT模型处理所述用户查询语料数据，以得到第一语义向量；

[0020] 基于预设的BERT模型处理所述预设数量的候选表的元数据，以得到预设数量的第二语义向量；

[0021] 基于深度匹配算法处理所述用户查询语料数据对应的第一语义向量和预设数量的候选表对应的所述第二语义向量，以计算所述用户查询语料数据与各候选表的匹配分数；

[0022] 选择匹配分数最高的候选表作为所述目标表。

[0023] 根据本申请实施例的一种具体实施方式，根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句的步骤，包括：

[0024] 按照所述目标表的列顺序将所述用户查询语料数据和所述目标表的列名进行拼接，以得到预设数量的待编码数据；

[0025] 基于预设的BERT模型对各所述待编码数据进行编码处理，以得到预设数量的编码向量；

[0026] 基于预设的二分类模型处理全部所述编码向量，以得到第一部分组合语句和第二部分组合语句，其中，所述第一部分组合语句为对应所述目标数据内容的组合语句，所述第二部分组合语句为待删除的组合语句；

[0027] 合并全部第一部分组合语句，以得到所述SQL查询语句。

[0028] 第二方面，本申请实施例提供了一种数据内容搜索装置，所述数据内容搜索装置包括：

[0029] 获取模块，用于获取用户查询语料数据；

[0030] 解析模块，用于解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据；

[0031] 第一处理模块，用于根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表；

[0032] 第二处理模块，用于根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句；

[0033] 查询模块，用于基于所述SQL查询语句在预设系统中查询所述目标数据内容。

[0034] 根据本申请实施例的一种具体实施方式，所述解析模块，具体用于利用命名实体识别模型处理所述用户查询语料数据，以得到对应的数据实体；根据所述数据实体在码表库和表元数据库中进行弹性搜索，以得到所述预设数量的候选表的元数据。

[0035] 第三方面，本申请实施例提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器存储有计算机程序，所述计算机程序在所述处理器上运行时执行前述第一方面及第一方面任一实施方式所述的数据内容搜索方法。

[0036] 第四方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，所述计算机程序在处理器上运行时执行前述第一方面及第一方面任一实施方式所述的数据内容搜索方法。

[0037] 本申请实施例提供了一种数据内容搜索方法、装置、计算机设备及可读存储介质，所述方法包括：获取用户查询语料数据；解析所述用户查询语料数据，以获得请求查询的目

标数据内容对应的数据实体和候选表的元数据;根据预设的表匹配模型处理所述用户查询语料数据和候选表的元数据,以获得目标表;根据预设的SQL语句生成模型处理所述数据实体和所述目标表,以得到对应目标数据内容的SQL查询语句;基于所述SQL查询语句在预设系统中查询所述目标数据内容。通过表匹配模型和预设的SQL语句生成模型的处理,能够更加快速、准确的获取目标数据内容。

附图说明

[0038] 为了更清楚地说明本发明的技术方案,下面将对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发明的某些实施例,因此不应被看作是对本发明保护范围的限定。在各个附图中,类似的构成部分采用类似的编号。

[0039] 图1示出了本申请实施例提供的一种数据内容搜索方法的方法流程示意图;

[0040] 图2示出了本申请实施例提供的一种数据内容搜索方法的命名实体识别模型的工作交互示意图;

[0041] 图3示出了本申请实施例提供的一种数据内容搜索方法的表匹配模型的工作交互示意图;

[0042] 图4示出了本申请实施例提供的一种数据内容搜索装置的装置模块示意图。

具体实施方式

[0043] 下面将结合本发明实施例中附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。

[0044] 通常在此处附图中描述和示出的本发明实施例的组件可以以各种不同的配置来布置和设计。因此,以下对在附图中提供的本发明的实施例的详细描述并非旨在限制要求保护的本发明的范围,而是仅仅表示本发明的选定实施例。基于本发明的实施例,本领域技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0045] 在下文中,可在本发明的各种实施例中使用的术语“包括”、“具有”及其同源词仅意在表示特定特征、数字、步骤、操作、元件、组件或前述项的组合,并且不应被理解为首先排除一个或更多个其它特征、数字、步骤、操作、元件、组件或前述项的组合的存在或增加一个或更多个特征、数字、步骤、操作、元件、组件或前述项的组合的可能性。

[0046] 此外,术语“第一”、“第二”、“第三”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0047] 除非另有限定,否则在这里使用的所有术语(包括技术术语和科学术语)具有与本发明的各种实施例所属领域普通技术人员通常理解的含义相同的含义。所述术语(诸如在一般使用的词典中限定的术语)将被解释为具有与在相关技术领域中的语境含义相同的含义并且将不被解释为具有理想化的含义或过于正式的含义,除非在本发明的各种实施例中被清楚地限定。

[0048] 参考图1,为本申请实施例提供的一种数据内容搜索方法的方法流程示意图,本申请实施例提供的数据内容搜索方法,如图1所示,所述数据内容搜索方法包括:

[0049] 步骤S101,获取用户查询语料数据;

[0050] 在具体实施例中,所述用户查询语料数据可以为用户在前端界面输入的query查询语句。

[0051] 所述用户查询语料数据也可以为存储在历史语料数据库中的query查询语句。具体的,所述用户查询语料数据的来源可以根据实际应用场景进行自适应替换,此处不作限定。

[0052] 用户在前端界面输入query查询语句后,会将query查询语句自动存储至历史语料数据库中,以供预设的查询系统进行更新学习。

[0053] 步骤S102,解析所述用户查询语料数据,以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据;

[0054] 在具体实施方式中,后端界面在获取query查询语句后,即开始对用户query查询语句进行解析,分析语句中每个字的语义,并以字为单位进行标签分类,根据标签分类结果可以分析得到初步的数据实体内容。

[0055] 具体的,所述数据实体包括产品类型、产品型号、产品地域和搜索标签等与目标数据内容相关的数据信息。

[0056] 在获取得到所述初步的数据实体内容后,在元码表库和表元数据库中进行弹性搜索(Elastic Search,简称ES),即能够得到对应候选表的元数据。

[0057] 其中,候选表为与数据实体相关的目标数据内容的表数据,通过后续的匹配模型能够从多个候选表中确定与目标数据内容最匹配度目标表,候选表的元数据包括表名/表别名以及码表数据等。

[0058] 根据本申请实施例的一种具体实施方式,解析所述用户查询语料数据,以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据的步骤,包括:

[0059] 利用命名实体识别模型处理所述用户查询语料数据,以得到对应的数据实体;

[0060] 根据所述数据实体在码表库和表元数据库中进行弹性搜索,以得到所述预设数量的候选表的元数据。

[0061] 在具体实施方式中,如图2所示,通过所述命名实体识别(Named Entity Recognition,简称NER)模型用于获取用户查询语料数据中的数据实体。

[0062] NER模型通常采用序列标注的方式实现,即对输入的中文以字为单位进行BIO标签分类,其中,B为实体起始位置,I为实体中间位置,O为实体外部的字符,即无关字符。

[0063] 在本实施例中,采用BERT/BiLSTM+CRF模型结构作为NER模型的基础结构,即通过BERT模型(Bidirectional Encoder Representation from Transformers,预训练的语言表征模型)或BiLSTM预训练模型对输入的语句序列进行语义编码,再通过CRF模型计算序列的最优解码结果。

[0064] 根据本申请实施例的一种具体实施方式,所述命名实体识别模型包括预设的BERT模型和条件随机场模型,基于命名实体识别模型处理所述用户查询语料数据,以得到对应的数据实体的步骤,包括:

[0065] 基于所述预设的BERT模型处理所述用户查询语料数据,以得到具有语义标签的初始编码序列;

[0066] 基于所述条件随机场模型解码所述初始编码序列,以获得对应的数据实体。

[0067] 在具体实施方式中,如图2所示,当用户基于预训练好的BERT模型处理用户输入的

query语句时，会得到具有BIO语义标签的初始编码序列。

[0068] 举例来说，当用户输入的query语句为“查询一下重机近期的缺件物料?”，预设的BERT模型会对所述query语句进行语义编码动作，以得到为“OOOOBIOOOBIBIO”的初始编码序列。

[0069] 将所述初始编码序列发送至所述条件随机场(Conditional Random Field，简称CRF)模型，可以通过所述CRF模型计算最优的解码结果，将所述解码结果发送至神经网络中的全连接层Dense，即能够导出对应所述query的数据实体“重机、缺件、物料”。

[0070] 根据本申请实施例的一种具体实施方式，所述预设的BERT模型的编码层采用RoBERTa预训练模型。

[0071] 在本实施例中，BERT编码层被替换为RoBERTa预训练模型，从而能够使得本实施例的NER模型的语义识别和语义编码性能优于传统的BERT模型，且在计算量上没有为处理器增加更多负担。

[0072] 在具体实施方式中，在得到用户查询语料数据对应的数据实体后，处理器会继续使用所述数据实体进行多路候选表召回动作。

[0073] 具体的，例如获取数据实体为“重机、缺件、物料”时，根据数据实体的标签类型对候选表进行召回。“重机”属于产品类型，在进行候选表召回时，会将产品类型限制为事业部，在码表库中进行弹性搜索。“缺件”属于表名，在进行候选表召回时，在表元数据库中的表名、表别名存储区域进行弹性搜索。“物料”属于列名，在进行候选表召回时，在表元数据库中的列名、列别名存储区域进行弹性搜索。

[0074] 具体的，在码表库进行弹性搜索时，会记录码表对应的code值、code类型、code所在表和code所在列等。

[0075] 在表元数据库中进行弹性搜索时，会记录表的所有字段定义，如表名、列名、表别名、列别名等。需知的，在表元数据库中进行弹性搜索仅搜索得到表元数据，并不记录表中存储的数据内容。

[0076] 通过先获取数据实体，再在码表数据库和表元数据库进行多路候选表召回的动作，能够极大的减少搜索数据内容时的索引范围，能够更加精准和快速的定位数据内容，避免对于数据库的全面搜索，减少处理器的消耗量。

[0077] 步骤S103，根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表；

[0078] 在具体实施例中，预设的表匹配模型采用编码-解码(encoder-decoder)结构，即编码器输入的query语句，将所述query语句转换为定长的一个向量，再通过解码器将所述向量转换为对应的目标文字。

[0079] 在实际运行过程中，所述表匹配模型的输入参数包括两部分，第一部分为用户query语句的数据实体，第二部分为候选表的元数据，其中，候选表的元数据包括表名/表别名、列名/列别名和码表数据。

[0080] 所述码表数据是指，根据NER模型解析的数据实体，在码表库中进行弹性搜索获得的数据。若候选表无码表库搜索结果，则候选表由1个句子组成，即表名/表别名。若有码表库检索结果，则候选表由两个句子组成，即表名/表别名，如图3中的“摄像头宽表”；和列名加码表数据，如图3中“公司名称:泵送”。

[0081] 所述表匹配模型在被输入所述数据实体和预设数量的候选表元数据后，会计算所述数据实体和每一候选表的元数据之间的匹配分数，并根据所述匹配分数确定与所述数据实体最相关的候选表的元数据。

[0082] 根据本申请实施例的一种具体实施方式，根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表的步骤，包括：

[0083] 基于预设的BERT模型处理所述用户查询语料数据，以得到第一语义向量；

[0084] 基于预设的BERT模型处理所述预设数量的候选表的元数据，以得到预设数量的第二语义向量；

[0085] 基于深度匹配算法处理所述用户查询语料数据对应的第一语义向量和预设数量的候选表对应的所述第二语义向量，以计算所述用户查询语料数据与各候选表的匹配分数；

[0086] 选择匹配分数最高的候选表作为所述目标表。

[0087] 在具体实施方式中，如图3所示，所述表匹配模型在获得所述用户查询语料数据对应的数据实体，以及候选表的元数据后，会通过预先训练好的BERT模型对各输入参数进行语义编码，以获取对应的第一语义向量和预设数量的第二语义向量。

[0088] 所述表匹配模型通过语义匹配算法实现精排，具体原理为将所述query语句和表元数据分别编码为语义向量，然后通过深度匹配模型，将两个向量映射到同一空间中，具有语义匹配关系的，在向量空间中距离越近。

[0089] 由于表元数据会先通过BERT模型编码为N个基础语义向量，例如，当所述表元数据只包括表名/表别名时，所述表元数据会通过BERT模型编码为1个基础语义向量，所述基础语义向量为所述第二语义向量。当所述表元数据包括表名/表别名时，所述表元数据会通过BERT模型编码为2个基础语义向量，此时需要对所述基础语义向量进行合并处理，以得到对应的第二语义向量。

[0090] 为通过第二语义向量表示所述表元数据，算法通过重要性Attention机制，计算每个基础语义向量的重要性得分，然后加权得到一个第二语义向量，作为表的语义向量表示。

[0091] 在本示例中，如图3所示，用户搜索语句中的“泵送”可能会匹配到很多表，“摄像头”可能是一个更为关键的信息，算法利用深度学习中的注意力机制，自动学习匹配过程中的关键信息，使得重要的字段在表最终的语义表示中有较高权重，提升匹配的精准程度。

[0092] 具体的，所述Attention权重分配可以根据实际应用过程中query语句中的数据实体进行自适应分配，图3中“摄像头信息宽表”被分配权重为0.8，“公司名称:泵送”被分配权重为0.2。

[0093] 在经过BERT模型的语义编码处理后，即能够得到对应每一组query语句和候选表的元数据的第一语义向量和第二语义向量。计算表匹配模型输出的匹配得分时，直接计算一组第一语义向量和第二语义向量的内积，再通过Softmax对所述内积进行归一化处理，即能够得到对应一组query语句和候选表的元数据的匹配得分。

[0094] 所述表匹配模型中Attention机制的具体算法如下：

[0095] 本算法模型采用encoder-decoder结构，即通过编码器encoder读取输入的句子将其转换为定长的语义向量，然后通过解码器decoder再将这个向量转换成对应的目标文字。

[0096] (1)首先我们利用RNN结构得到编码器encoder中的隐藏状态hidden state(h_1，

$h_2,\cdots,h_t$)；

[0097] (2)假设当前解码器decoder的隐藏状态hidden state是s_{t-1},我们可以计算每一个输入位置j与当前输出位置的关联性$e_{tj}=a(s_{t-1},h_j)$,写成相应的向量形式即为$E_t=(a(s_{t-1},h_1),\cdots,a(s_{t-1},h_T))$,其中a是一种相关性的算符,例如常见的有点乘形式的$E_t=S_{t-1}^T H$,加权点乘$E_t=S_{t-1}^T WH$等等。

[0098] (3)对于E_t进行softmax归一化操作将其归一化函数normalize转换得到重要性Attention的分布,$A_t=\mathrm{softmax}(E_t)$,展开形式为$A_{tj}=\frac{\exp(E_{tk})}{\sum_{k=1}^{T}\exp(E_{tk})}$;

[0099] (4)利用A_t我们可以进行加权求和得到相应的语义向量context vector $C_t=\sum_{j=1}^{T}A_{tj}*h_j$;

[0100] (5)由此,我们可以计算解码器decoder的下一个隐藏状态hidden state $S_t=f(s_{t-1},y_{t-1},c_t)$以及该位置的输出$p(y_t|y_1,y_2,\cdots,y_{t-1},x)=g(y_{i-1},s_i,c_i)$。

[0101] 通过计算编码器encoder与解码器状态decoder state之间的关联性的权重,就可以得到重要性Attention的权重分布,从而能够自动确定数据实体中占据更加重要位置的部分。

[0102] 由于存在多个候选表的元数据,所述表匹配模型每一次处理时仅处理一个所述query语句向量与一个候选表的原数据的语义向量。在得到所述用户查询语料数据与各候选表的元数据之间的匹配得分后,选取匹配得分最高的候选表的元数据作为目标表的元数据。

[0103] 并根据所述目标表元数据,从预设的表数据库中提取对应的目标表,以进行后续目标数据内容的获取步骤。

[0104] 步骤S104,根据预设的SQL语句生成模型处理所述数据实体和所述目标表,以得到对应目标数据内容的SQL查询语句;

[0105] 在具体实施例中,所述预设的SQL语句生成模型为基于BERT的中文NL2SQL模型。

[0106] 在具体实施方式中,基于BERT的中文NL2SQL模型,模型的输入为query语句的数据实体和数据表,模型输出为一个SQL结构,该SQL结构对应一条SQL语句。

[0107] 具体的,在所述NL2SQL模型中,sel字段格式为列表形式,表示select语句所选取的列。agg字段格式为列表形式,与sel字段进行一一对应,表示对select语句所选取的各列做何种聚合操作(例如count,sum等)。conds字段格式为列表形式,表示where子句中的各个条件,每个条件形式为(条件列,条件运算符,条件值)的形式。cond_conn_op字段格式为整型数字形式,表示条件之间的关系(例如and、or)。

[0108] 根据本申请实施例的一种具体实施方式,根据预设的SQL语句生成模型处理所述数据实体和所述目标表,以得到对应目标数据内容的SQL查询语句的步骤,包括:

[0109] 按照所述目标表的列顺序将所述用户查询语料数据和所述目标表的列名进行拼接,以得到预设数量的待编码数据;

[0110] 基于预设的BERT模型对各所述待编码数据进行编码处理,以得到预设数量的编码向量;

[0111] 基于预设的二分类模型处理全部所述编码向量,以得到第一部分组合语句和第二

部分组合语句,其中,所述第一部分组合语句为对应所述目标数据内容的组合语句,所述第二部分组合语句为待删除的组合语句;

[0112] 合并全部第一部分组合语句,以得到所述SQL查询语句。

[0113] 在具体实施方式中,所述SQL语句生成模型在获取所述query语句目标表后,将query语句与目标表的列名按预设顺序进行拼接,以得到待输入至SQL语句生成模型输入参数。

[0114] 具体的,在每个列名之前添加TEXT或REAL这两个BERT模型没有的Token令牌进行区分,其中TEXT和REAL的Token令牌,可以通过任选两个BERT原本预留的未经训练的Token令牌进行替换即可。

[0115] 所述SQL语句生成模型在预测SQL语句时,需要预测数据表中哪些列会被选取,由于数据表的每一列的含义均不一样,因此我们需要把query语句与数据表中每一列的header按照预设顺序拼接起来,然后一起输入BERT模型中进行实时编码。

[0116] 具体的,所述顺序的选择可以为从上到下选取数据表的每一列,也可以预先对数据表的每一列进行编号,按照编号顺序选取所述数据表的每一列。本实施例不对所述顺序做具体限定,可以根据实际应用中具体情况进行自适应替换。

[0117] 经过BERT模型进行编码之后,可以得到一系列编码向量,然后我们用得到的编码向量去对SQL各个子句进行预测。

[0118] 所述SQL语句生成模型在对cond_val(条件值)子句进行预测时,具体方式是根据前述实施例中所选择的编码向量,即cond_col(条件列)子句结果,对cond_op(连接符)与cond_val(条件值)的组合进行枚举来生成一系列的候选组合,将这些候选组合的选择转化为多个二分类问题。

[0119] 将生成的多组[cond_col,cond_op,cond_val]组合与query按照之前所述,拼接起来依次输入BERT模型,然后经BERT编码后的向量输入一层Dense Layer进行二分类,判断每个组合是否与问题对应,最后将输出为1(即与目标数据内容对应的cond组合)所有cond组合进行合并,作为模型最终的输出。

[0120] 其中,所述第一部分组合语句即经过二分类处理后结果为1的所有cond组合,将所有第一部分组合语句进行预设的拼接和转化步骤后,就可以得到对应目标数据内容的SQL查询语句。

[0121] 具体的,对于经过二分类处理后结果为0的所有第二部分组合语句,均进行删除处理。

[0122] 步骤S105,基于所述SQL查询语句在预设系统中查询所述目标数据内容。

[0123] 在具体实施例中,在得到对应的SQL查询语句后,则可以在任意通过SQL语句进行数据搜索的查询系统搜索所述目标数据内容。

[0124] 具体的,所述预设系统为现有技术中任意一种可以使用SQL语句进行数据查询的数据存储系统,本实施例不作赘述。

[0125] 综上所述,本实施例提出了一种数据内容搜索方法,通过改进的NER模型对用户输入的查询语料数据进行实时分析,以得到对应的数据实体和多个候选表的元数据。通过表匹配模型对所述候选表和数据实体进行筛选,确定最关联的数据表。通过SQL语句生成模型进行SQL语句判断,最终能快速、准确的得到用于查询目标数据内容的SQL语句。有效解决的

现有技术中无法准确、快速的处理搜索大量数据内容的问题。

[0126] 参考图4，为本申请实施例提供的一种数据内容搜索装置400的装置模块示意图，本申请实施例提供的数据内容搜索装置400，如图4所示，所述数据内容搜索装置400包括：

[0127] 获取模块401，用于获取用户查询语料数据；

[0128] 解析模块402，用于解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据；

[0129] 第一处理模块403，用于根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表；

[0130] 第二处理模块404，用于根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句；

[0131] 查询模块405，用于基于所述SQL查询语句在预设系统中查询所述目标数据内容。

[0132] 根据本申请实施例的一种具体实施方式，所述解析模块402，具体用于利用命名实体识别模型处理所述用户查询语料数据，以得到对应的数据实体；根据所述数据实体在码表库和表元数据库中进行弹性搜索，以得到所述预设数量的候选表的元数据。

[0133] 另外，本申请实施例还提供的一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器存储有计算机程序，所述计算机程序在所述处理器上运行时执行前述实施例中的数据内容搜索方法。

[0134] 本申请实施例还提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，所述计算机程序在处理器上运行时执行前述实施例中的数据内容搜索方法。

[0135] 综上所述，本申请实施例提供了一种数据内容搜索方法、装置、计算机设备及可读存储介质，首先使用NER模型得到初步的实体解析结果后，通过元码表过滤，初步得到候选表的元数据。然后通过第二阶段的表匹配模型，对用户查询语料数据和所述候选表的元数据，进行第二次过滤，根据计算内积后得到的匹配得分，得到目标表的元数据。最后根据第一阶段得到的query实体和第二阶段得到的表元数据定义信息，通过NL2SQL模型生成对应的查询语句，得到最终的查询结果。通过本方案能够在海量数据、结构复杂等场景下，快速、准确地对数据进行召回。另外，上述实施例中提到的数据内容搜索装置、计算机设备及计算机可读存储介质的具体实施过程，可以参见上述方法实施例的具体实施过程，在此不再一一赘述。

[0136] 在本申请所提供的几个实施例中，应该理解到，所揭露的装置和方法，也可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的，例如，附图中的流程图和结构图显示了根据本发明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，所述模块、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，在作为替换的实现方式中，方框中所标注的功能也可以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，结构图和/或流程图中的每个方框、以及结构图和/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。

[0137] 另外,在本发明各个实施例中的各功能模块或单元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,也可以两个或更多个模块集成形成一个独立的部分。

[0138] 所述功能如果以软件功能模块的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是智能手机、个人计算机、服务器、或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0139] 以上所述,仅为本发明的具体实施方式,但本发明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发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。

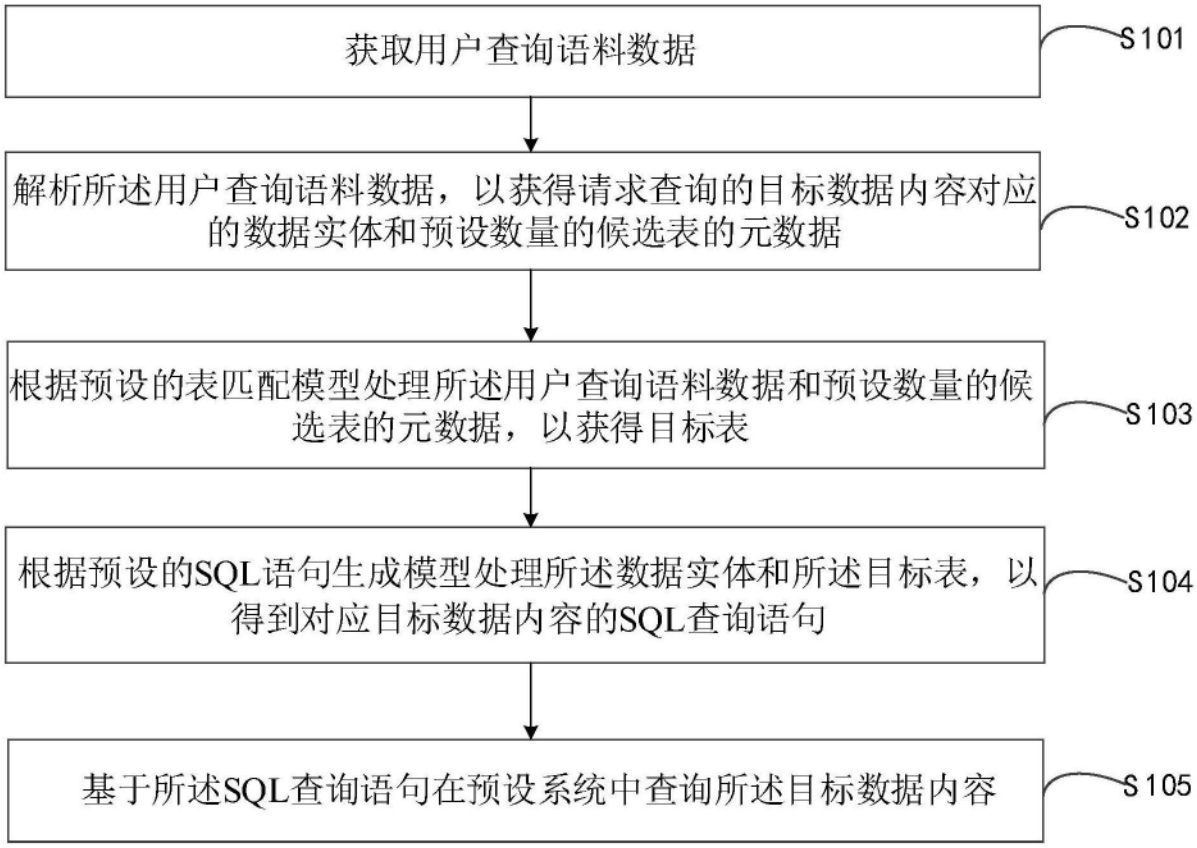

获取用户查询语料数据
S101
解析所述用户查询语料数据，以获得请求查询的目标数据内容对应的数据实体和预设数量的候选表的元数据
S102
根据预设的表匹配模型处理所述用户查询语料数据和预设数量的候选表的元数据，以获得目标表
S103
根据预设的SQL语句生成模型处理所述数据实体和所述目标表，以得到对应目标数据内容的SQL查询语句
S104
基于所述SQL查询语句在预设系统中查询所述目标数据内容
S105

图1

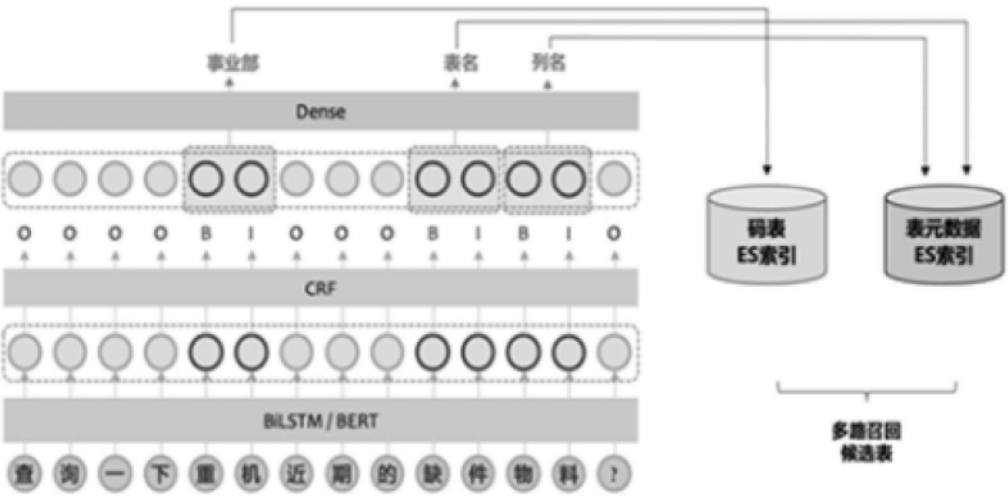

事业部
表名
列名
Dense
O O O O B I O O O B I B I O
CRF
BILSTM / BERT
查 询 一 下 重 机 近 期 的 缺 件 物 料 ?
码表 ES索引
表元数据 ES索引
多路召回 候选表

图2

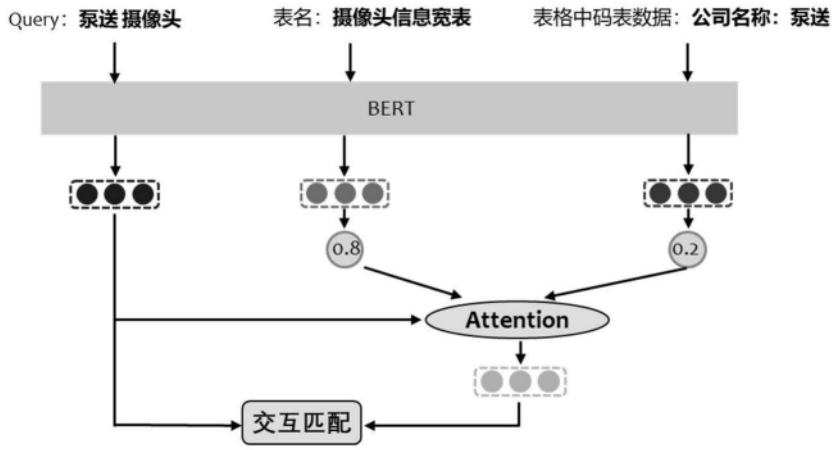

Query：泵送 摄像头
表名：摄像头信息宽表
表格中码表数据：公司名称：泵送
BERT
0.8
0.2
Attention
交互匹配

图3

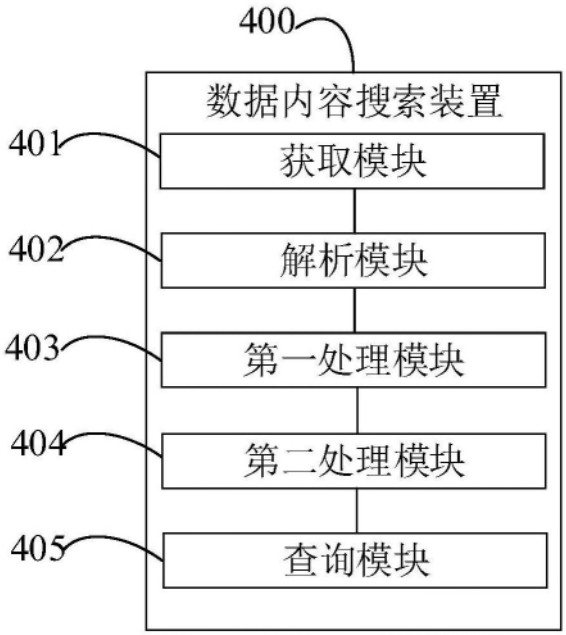

400
数据内容搜索装置
401
获取模块
402
解析模块
403
第一处理模块
404
第二处理模块
405
查询模块

图4