



(12) 发明专利申请

(10) 申请公布号 CN 116303537 A

(43) 申请公布日 2023. 06. 23

(21) 申请号 202310257937.6

(22) 申请日 2023.03.10

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市大兴区北京经济技术
开发区科创十一街18号院2号楼6层
601

(72) 发明人 袁韶祖 刘瑞雪 祝天刚 陈蒙

(74) 专利代理机构 北京律智知识产权代理有限
公司 11438

专利代理师 李建忠

(51) Int. Cl.

G06F 16/242 (2019.01)

G06F 16/245 (2019.01)

G06F 16/22 (2019.01)

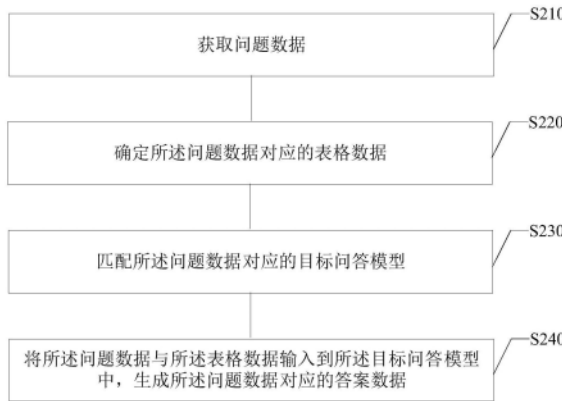
权利要求书2页 说明书13页 附图3页

(54) 发明名称

数据查询方法及装置、电子设备、存储介质

(57) 摘要

本公开提供了一种数据查询方法及装置、电子设备、存储介质,涉及人工智能技术领域。该数据查询方法包括:获取问题数据;确定所述问题数据对应的表格数据;匹配所述问题数据对应的目标问答模型;将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。本公开实施例的技术方案可以预训练的目标问答模型从表格数据中快速查询到问题的答案数据,有效提升答案数据的查询效率,提高答案数据的准确率。



1. 一种数据查询方法,其特征在于,包括:
获取问题数据;
确定所述问题数据对应的表格数据;
匹配所述问题数据对应的目标问答模型;
将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

2. 根据权利要求1所述的数据查询方法,其特征在于,所述匹配所述问题数据对应的目标问答模型,包括:

确定所述问题数据对应的业务领域类型;
根据所述业务领域类型匹配所述问题数据对应的目标问答模型,所述目标问答模型是通过对预训练问答模型进行定向训练得到的。

3. 根据权利要求2所述的数据查询方法,其特征在于,所述预训练问答模型是通过预训练过程得到的,所述预训练过程包括:

获取问题模板以及所述问题模板对应的结构化查询语句模板;
根据预设的训练表格数据、所述问题模板和所述结构化查询语句模板生成训练三元组数据;

通过所述训练三元组数据对预构建的初始预训练问答模型进行迭代训练,得到训练完成的预训练问答模型。

4. 根据权利要求3所述的数据查询方法,其特征在于,在获取问题模板以及所述问题模板对应的结构化查询语句模板之前,所述方法还包括:

从预设数据集提取样本问题数据以及所述样本问题数据对应的样本结构化查询语句;
对所述样本问题数据以及所述样本结构化查询语句中的关键词进行模糊匹配,确定模板填充位置,所述模板填充位置包括数据位置和数据类型;

基于所述模板填充位置以及所述样本问题数据和所述样本结构化查询语句,生成所述问题模板和所述问题模板对应的结构化查询语句模板。

5. 根据权利要求3所述的数据查询方法,其特征在于,所述根据预设的训练表格数据、所述问题模板和所述结构化查询语句模板生成训练三元组数据,包括:

获取预设的训练表格数据,并对所述训练表格数据进行拆解,得到训练子表格数据;
基于所述训练子表数据对所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句;

根据所述训练结构化查询语句在所述训练表格数据中查询,得到训练答案数据;
通过所述训练表格数据、所述训练问题数据和所述训练答案数据构成训练三元组数据。

6. 根据权利要求5所述的数据查询方法,其特征在于,所述基于所述训练子表数据对所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句,包括:

确定所述结构化查询语句模板的模板类型,并根据所述模板类型对所述问题模板和所述结构化查询语句模板进行分类;其中,所述模板类型是通过所述结构化查询语句模板的复杂程度划分的;

基于所述训练子表数据对不同分类结果下的所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句。

7.根据权利要求2所述的数据查询方法,其特征在于,所述目标问答模型是通过对预训练问答模型进行定向训练得到的,包括:

获取属于不同业务领域类型的样本三元组数据;

根据所述样本三元组数据对训练完成的预训练问答模型进行定向训练,得到属于不同业务领域类型的目标问答模型。

8.一种数据查询装置,其特征在于,包括:

问题获取模块,用于获取问题数据;

表格确定模块,用于确定所述问题数据对应的表格数据;

问答模型匹配模块,用于匹配所述问题数据对应的目标问答模型;

答案查询模块,用于将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

9.一种电子设备,包括:

处理器;以及

存储器,所述存储器上存储有计算机可读指令,所述计算机可读指令被所述处理器执行时实现如权利要求1至7中任一项所述的数据查询方法。

10.一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如权利要求1至7中任一项所述的数据查询方法。

数据查询方法及装置、电子设备、存储介质

技术领域

[0001] 本公开涉及人工智能技术领域,具体而言,涉及一种数据查询方法、数据查询装置、电子设备以及计算机可读存储介质。

背景技术

[0002] 表格问答是一种对话模型,是根据表格信息将人类的自然语言转换成可以执行的结构化查询语句(Structured Query Language,SQL),并从数据库执行该SQL语句并返回人们想要的答案。

[0003] 目前,相关技术中,在通过表格问答的方式生成答案时,需要先生成SQL语句,再通过执行SQL语句查询答案,而SQL语句的生成需要耗费时间,导致问题答案的查询效率较低。

[0004] 需要说明的是,在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解,因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

[0005] 本公开实施例的目的在于提供一种数据查询方法、数据查询装置、电子设备以及计算机可读存储介质,进而有效提升在表格数据中查询答案数据的查询效率。

[0006] 本公开的其他特性和优点将通过下面的详细描述变得显然,或部分地通过本公开的实践而习得。

[0007] 根据本公开实施例的第一方面,提供了一种数据查询方法,包括:

[0008] 获取问题数据;

[0009] 确定所述问题数据对应的表格数据;

[0010] 匹配所述问题数据对应的目标问答模型;

[0011] 将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

[0012] 在本公开的一些示例实施例中,基于前述方案,所述匹配所述问题数据对应的目标问答模型,包括:

[0013] 确定所述问题数据对应的业务领域类型;

[0014] 根据所述业务领域类型匹配所述问题数据对应的目标问答模型,所述目标问答模型是通过对预训练问答模型进行定向训练得到的。

[0015] 在本公开的一些示例实施例中,基于前述方案,所述预训练问答模型是通过预训练过程得到的,所述预训练过程包括:

[0016] 获取问题模板以及所述问题模板对应的结构化查询语句模板;

[0017] 根据预设的训练表格数据、所述问题模板和所述结构化查询语句模板生成训练三元组数据;

[0018] 通过所述训练三元组数据对预构建的初始预训练问答模型进行迭代训练,得到训练完成的预训练问答模型。

[0019] 在本公开的一些示例实施例中,基于前述方案,在获取问题模板以及所述问题模板对应的结构化查询语句模板之前,所述方法还包括:

[0020] 从预设数据集提取样本问题数据以及所述样本问题数据对应的样本结构化查询语句;

[0021] 对所述样本问题数据以及所述样本结构化查询语句中的关键词进行模糊匹配,确定模板填充位置,所述模板填充位置包括数据位置和数据类型;

[0022] 基于所述模板填充位置以及所述样本问题数据和所述样本结构化查询语句,生成所述问题模板和所述问题模板对应的结构化查询语句模板。

[0023] 在本公开的一些示例实施例中,基于前述方案,所述根据预设的训练表格数据、所述问题模板和所述结构化查询语句模板生成训练三元组数据,包括:

[0024] 获取预设的训练表格数据,并对所述训练表格数据进行拆解,得到训练子表格数据;

[0025] 基于所述训练子表数据对所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句;

[0026] 根据所述训练结构化查询语句在所述训练表格数据中查询,得到训练答案数据;

[0027] 通过所述训练表格数据、所述训练问题数据和所述训练答案数据构成训练三元组数据。

[0028] 在本公开的一些示例实施例中,基于前述方案,所述基于所述训练子表数据对所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句,包括:

[0029] 确定所述结构化查询语句模板的模板类型,并根据所述模板类型对所述问题模板和所述结构化查询语句模板进行分类;其中,所述模板类型是通过所述结构化查询语句模板的复杂程度划分的;

[0030] 基于所述训练子表数据对不同分类结果下的所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句。

[0031] 在本公开的一些示例实施例中,基于前述方案,所述目标问答模型是通过对预训练问答模型进行定向训练得到的,包括:

[0032] 获取属于不同业务领域类型的样本三元组数据;

[0033] 根据所述样本三元组数据对训练完成的预训练问答模型进行定向训练,得到属于不同业务领域类型的目标问答模型。

[0034] 根据本公开实施例的第二方面,提供了一种数据查询装置,包括:

[0035] 问题获取模块,用于获取问题数据;

[0036] 表格确定模块,用于确定所述问题数据对应的表格数据;

[0037] 问答模型匹配模块,用于匹配所述问题数据对应的目标问答模型;

[0038] 答案查询模块,用于将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

[0039] 根据本公开实施例的第三方面,提供了一种电子设备,包括:处理器;以及存储器,所述存储器上存储有计算机可读指令,所述计算机可读指令被所述处理器执行时实现上述任意一项所述的数据查询方法。

[0040] 根据本公开实施例的第四方面,提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现根据上述任意一项所述的数据查询方法。

[0041] 本公开实施例提供的技术方案可以包括以下有益效果:

[0042] 本公开的示例实施例中的数据查询方法,可以获取问题数据,并确定问题数据对应的表格数据,然后可以匹配问题数据对应的目标问答模型,进而可以将问题数据与表格数据输入到目标问答模型中,生成问题数据对应的答案数据。可以匹配问题数据的业务领域所对应的目标问答模型,进而可以通过目标问答模型从表格数据中快速查询到问题数据所对应的答案数据,可以有效提升答案数据的查询效率,并且由于不需要额外生成SQL语句进行查询,在一定程度上也可以提升答案数据的准确性。

[0043] 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,并不能限制本公开。

附图说明

[0044] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理。显而易见地,下面描述中的附图仅仅是本公开的一些实施例,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的附图。在附图中:

[0045] 图1示出了可以应用本公开实施例的一种数据查询方法及装置的示例性应用环境的系统架构的示意图;

[0046] 图2示意性示出了根据本公开的一些实施例的数据查询方法的流程示意图;

[0047] 图3示意性示出了根据本公开的一些实施例的对预训练问答模型进行预训练的流程图示意图;

[0048] 图4示意性示出了根据本公开的一些实施例的数据查询装置的示意图;

[0049] 图5示意性示出了根据本公开的一些实施例的电子设备的计算机系统的结构示意图;

[0050] 图6示意性示出了根据本公开的一些实施例的计算机可读存储介质的示意图。

[0051] 在附图中,相同或对应的标号表示相同或对应的部分。

具体实施方式

[0052] 现在将参考附图更全面地描述示例实施方式。然而,示例实施方式能够以多种形式实施,且不应被理解为限于在此阐述的范例;相反,提供这些实施方式使得本公开将更加全面和完整,并将示例实施方式的构思全面地传达给本领域的技术人员。

[0053] 此外,所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中,提供许多具体细节从而给出对本公开的实施例的充分理解。然而,本领域技术人员将意识到,可以实践本公开的技术方案而没有特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知方法、装置、实现或者操作以避免模糊本公开的各方面。

[0054] 此外,附图仅为示意性图解,并非一定是按比例绘制。附图中所示的方框图仅仅是功能实体,不一定必须与物理上独立的实体相对应。即,可以采用软件形式来实现这些功能

实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0055] 图1示出了可以应用本公开实施例的一种数据查询方法及装置的示例性应用环境的系统架构的示意图。

[0056] 如图1所示,系统架构100可以包括终端设备101、102、103中的一个或多个,网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。终端设备101、102、103可以是各种具有图像处理功能的电子设备,包括但不限于台式计算机、便携式计算机、智能手机和平板电脑等等。应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的。比如服务器105可以是多个服务器组成的服务器集群等。

[0057] 本公开实施例所提供的数据查询方法一般由终端设备101、102、103中执行,相应地,数据查询装置一般设置于终端设备101、102、103中。但本领域技术人员容易理解的是,本公开实施例所提供的数据查询方法也可以由服务器105执行,相应的,数据查询装置也可以设置于服务器105中,本示例性实施例中对此不做特殊限定。举例而言,在一种示例性实施例中,可以是用户通过终端设备101、102、103获取以各种方式输入的问题数据,然后将问题数据上传至服务器105,服务器通过本公开实施例所提供的数据查询方法生成答案数据后,将答案数据传输给终端设备101、102、103进行后处理。

[0058] 在相关技术中,表格问答方式的通常做法是根据表格信息将人类的自然语言转换成可以执行的SQL语句,从数据库执行并返回人们想要的答案。这种技术通过计算设备可以自动生成SQL查询语句,用户只需要通过SQL查询语句就能够快速地查询到想要的答案数据。目前方案主要是采用预训练语言模型采用多个子任务,将SQL查询语句的各个部分生成出来,并组成一条SQL查询语句发送到数据库执行得到答案,这种将SQL查询语句通过多个子任务生成的经典模型主要包括SQL-NET和X-SQL等。例如,对于一条SQL查询语句,可以表示为“select agg sel_column from table where column1 op1 value1 conn where colnum2 op2value2”,其中该SQL查询语句由“agg”“sel_column”“column1”“op1”“value1”“conn”“colnum2”“op2”“value2”多个部分组成。直接将自然语言自动转为如此复杂的SQL查询语句是相当困难的,因此经典模型往往把一个SQL查询语句的生成问题分解成多个子任务来处理,分别对应以上各部分,每个部分都对应一个子任务。模型的主体部分为通常采用预训练语言模型BERT,并通过至少7个以上的子任务分别将SQL查询语句的各个部分生成,最后将各个部分拼接在一起就得输出的SQL查询语句,然后将SQL查询语句发到数据库运行得到答案。

[0059] 在该方案中,由于需要先得到SQL查询语句,在通过SQL查询语句查询答案,查询效率较低;其次,生成SQL查询语句时需要设计多个子任务,不仅极大的增加了模型的复杂度,同时让也让数据标注变得异常困难,因为数据集不仅需要标注问题,同时也要标注出问题对应的SQL语句作为监督信号,而SQL的标注必须是熟悉SQL并经过专业培训的人才行,标注费人费力且模型的高度复杂极大的限制了表格问答的应用范围。

[0060] 基于相关技术中的一个或者多个问题,在本示例实施例中,首先提供了一种数据查询方法,该数据查询方法可以应用于终端设备或者服务器,下面以服务器执行该方法

为例进行说明。图2示意性示出了根据本公开的一些实施例的数据查询方法流程的示意图。参考图2所示,该数据查询方法可以包括以下步骤:

[0061] 步骤S210,获取问题数据;

[0062] 步骤S220,确定所述问题数据对应的表格数据;

[0063] 步骤S230,匹配所述问题数据对应的目标问答模型;

[0064] 步骤S240,将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

[0065] 根据本示例实施例中的数据查询方法,可以匹配问题数据的业务领域所对应的目标问答模型,进而可以通过目标问答模型从表格数据中快速查询到问题数据所对应的答案数据,可以有效提升答案数据的查询效率,并且由于不需要额外生成SQL语句进行查询,在一定程度上也可以提升答案数据的准确性。

[0066] 下面,将对本示例实施例中的数据查询方法进行具体说明。

[0067] 在步骤S210中,获取问题数据。

[0068] 在一示例实施例中,问题数据是指用户通过任意方式输入的自然语言形式表征的问题,例如,问题数据可以是用户通过实体键盘、虚拟键盘等输入工具键入的自然语言表征的问题,也可以是通过具有语音采集功能的智能设备采集的问题语音,并将问题语音转换为自然语言表征的问题,当然,问题数据还可以是通过将问题手写在稿纸上并通过光学字符识别(Optical Character Recognition,OCR)识别得到的自然语言表征的问题,本示例实施例对此不做特殊限定。

[0069] 问题数据可以是用户通过任意方式实时输入的,也可以是预先存储在存储单元的,本实施例对于问题数据的来源不做任何特殊限定。

[0070] 在步骤S220中,确定所述问题数据对应的表格数据。

[0071] 在一示例实施例中,表格数据是指通过列数据和行数据组合实现数据存储以及数据查询的结构化表格,当然,本实施例中虽然以数据表格为例,但本领域技术人员容易理解的是,本实施例中的表格数据还可以是其他类型的结构化数据,或者可以转化为表格的结构化数据,并可以通过结构化数据快速定位查询相关数据,例如,本实施例中的表格数据还可以是日志文件数据、XML文档数据等,本示例实施例不以此为限。

[0072] 可以按照不同的业务领域类型预先将不同的表格数据进行划分,可以确定问题数据所对应的业务领域类型,进而可以根据业务领域类型确定问题数据对应的表格数据;当然,也可以提取问题数据对应的关键词,通过关键词确定问题数据对应的表格数据,本示例实施例对于确定问题数据对应的表格数据的方式不做任何特殊限定。

[0073] 通过确定问题数据的表格数据,可以有效减少数据匹配量,在一定程度上可以提升答案数据的查询效率。

[0074] 在步骤S230中,匹配所述问题数据对应的目标问答模型。

[0075] 在一示例实施例中,目标问答模型是指预先通过不同业务领域类型的训练数据对通用的预训练问答模型进行定向训练得到的问答模型,例如,假设问题数据是医疗业务领域的提问问题,那么可以匹配通过医疗业务领域的训练数据定向训练得到的目标问答模型;假设问题数据是客服业务领域的提问问题,那么可以匹配通过客服业务领域的训练数据定向训练得到的目标问答模型,当然,目标问答模型还可以是通过其他业务领域类型的

训练数据对通用的预训练问答模型进行定向训练得到的问答模型,本示例实施例对此不做特殊限定。

[0076] 可以在数据库中存储目标问答模型以及该目标问答模型所对应的业务领域类型,那么可以先确定问题数据的业务领域类型,进而可以通过问题数据的业务领域类型在数据库中匹配并调用相应业务领域类型的目标问答模型;当然,也可以训练得到用于任意业务领域类型的目标问答模型,本示例实施例对此不做特殊限定。

[0077] 在步骤S240中,将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

[0078] 在一示例实施例中,可以将问题数据与相应的表格数据进行组合,共同作为目标问答模型的输入数据,经过目标问答模型的处理,可以直接输出该问题数据对应的答案数据。

[0079] 可以匹配问题数据的业务领域所对应的目标问答模型,进而可以通过目标问答模型从表格数据中快速查询到问题数据所对应的答案数据,可以有效提升答案数据的查询效率,并且由于不需要额外生成SQL语句进行查询,在一定程度上也可以提升答案数据的准确性。

[0080] 下面,对步骤S210至步骤S240中的数据查询方法进行详细说明。

[0081] 在本公开的一种示例性实施例中,可以通过以下方式匹配问题数据对应的目标问答模型:可以确定问题数据对应的业务领域类型,并根据业务领域类型匹配问题数据对应的目标问答模型,目标问答模型是通过预训练问答模型进行定向训练得到的。

[0082] 其中,业务领域类型是指用于对不同业务领域的数据进行划分时设置的类型数据,例如,业务领域类型可以包括但不限于医疗业务领域、客服业务领域、仓储业务领域;具体的,可以对不同的业务领域类型设置不同的标识进行区分,例如,可以通过唯一编码的形式对业务领域类型进行区分,也可以建议的数字编码对业务领域类型进行区分,当然,还可以通过字符串等对业务领域类型进行区分,本示例实施例对此不做特殊限定。

[0083] 定向训练是指按照属于不同的业务领域类型的训练数据对预训练问答模型进行训练的过程,其中,预训练问答模型是指预选通过大量的数据训练得到的具有表格问答能力的预训练模型,是一种通用模型,通过属于不同的业务领域类型的训练数据对预训练问答模型进行定向训练后,得到的目标问答模型对于相应业务领域类型的答案查询效率更高,准确率更高。

[0084] 通过问题数据的业务领域类型匹配相应业务领域类型的目标问答模型,不仅不需要根据问题数据生成SQL查询语句,节省大量人力物力,而且通过匹配相应业务领域类型的目标问答模型,可以有效提升问题数据所对应的答案数据的查询效率,并且由于目标问答模型是基于业务领域类型的训练数据定向训练得到的,能够有效提高答案数据的准确性。

[0085] 在本公开的一种示例性实施例中,预训练问答模型可以通过预训练过程得到的,可以通过图3中的步骤实现预训练过程,参考图3所示,具体可以包括:

[0086] 步骤S310,获取问题模板以及所述问题模板对应的结构化查询语句模板;

[0087] 步骤S320,根据预设的训练表格数据、所述问题模板和所述结构化查询语句模板生成训练三元组数据;

[0088] 步骤S330,通过所述训练三元组数据对预构建的初始预训练问答模型进行迭代训

练,得到训练完成的预训练问答模型。

[0089] 其中,问题模板是指预先通过训练数据集中的大量问题数据进行通用部分提取得到的模板,结构化查询语句模板是指对训练数据集中的大量SQL查询语句进行通用部分提取得到的模板,问题模板和结构化查询语句模板中包含可以用于数据填充的数据槽位,用于填充数据内容,并且每个数据槽位可以包括可以填写的数据类型。

[0090] 举例而言,对于问题数据“‘人大’的‘2020年’的‘招生’是多少?”以及相应的SQL查询语句“Select ‘招生人数’ where ‘大学==人民大学and年份==2020年’”,其中单引号所引内容为可替换内容,通过对该问题数据以及SQL查询语句进行通用部分提取,可以得到问题模板“<cond_value1>的<cond_value2>的<sql_col1>是多少?”以及结构化查询语句模板“Select<sql_col1>where<cond_col1>==<cond_value1>and<cond_col2>==<cond_value2>”,模板中的“<cond_value1>”、“<cond_value2>”、“<sql_col1>”、“<cond_col1>”、“<cond_col2>”为数据槽位以及填充的数据内容所对应的数据类型,可以通过数据槽位以及数据类型对问题模板和相应的结构化查询语句模板进行填充。

[0091] 训练表格数据是指预先收集的用于训练预训练问答模型的样本数据,训练表格数据可以是预先收集的任意业务领域类型的表格数据;可以将训练表格数据中的数据内容随机填写到问题模板和结构化查询语句模板中,填充得到大量的训练问题数据和训练问题数据相应的训练结构化查询语句,并将训练结构化查询语句模板在训练表格数据中执行得到训练问题数据相应的训练答案数据,进而可以根据训练问题数据、训练结构化查询语句、训练答案数据得到训练三元组数据。

[0092] 可选的,若填充得到的训练结构化查询语句在训练表格数据中执行后无法得到相应的训练答案数据,则可以将该训练结构化查询语句以及相应的训练问题数据进行删除。

[0093] 预构建的初始预训练问答模型是指预先构建的尚未开始模型训练的预训练问答模型,例如,初始预训练问答模型可以是基于中文预训练模型T5构建的,也可以是基于预训练模型BERT构建的,当然,还可以是基于其他类型的预训练模型构建的,本实施例对此不做特殊限定。

[0094] 可选的,本实施例中的预训练问答模型的损失函数可以采用交叉熵损失函数进行训练,当然,预训练问答模型还可以采用其他类型的损失函数,本实施例不以此为限举例而言,初始预训练问答模型的损失函数可以表示为关系式(1):

$$[0095] \quad Loss = \sum loss_t = -\log P(w_t^*)$$

[0096] 其中,Loss可以表示预训练问答模型的总损失,loss_t可以表示预训练问答模型输出的答案数据中每个字符的损失,w_t^{*}可以表示第t个时间步生成的文字。

[0097] 在经过训练表格数据对问题模板和结构化查询语句模板进行随机填充,以及执行填充得到的训练结构化查询语句,可以得到相应的训练答案数据,进而构成多个训练三元组数据,例如,训练三元组数据可以表示为{table,question,answer},其中,table可以表示训练表格数据、question可以表示训练问题数据,answer可以表示训练答案数据,进而可以根据训练三元组数据对预构建的初始预训练问答模型进行迭代训练。举例而言,可以将训练表格数据和训练问题数据作为预构建的初始预训练问答模型的输入数据,如可以将训练表格数据和训练问题数据按照以下方式进行组织得到#question@1:t-h1|t-c1|t-c2@2:

t-h2|t-c1|t-c2@3...,其中,首开头的“#question”可以表示训练问题数据,@可以表示训练表格数据的表格内容,具体的,@1可以表示训练表格数据的第一行数据,@2可以表示训练表格数据的第二行数据,依次类推,t-h可以表示训练表格数据的表头(table header),t-c可以表示训练表格数据的表格内容(table content),这样所有需要的信息都被组织到一起,并将表格数据进行了展平,作为初始预训练问答模型的输入数据;然后可以将训练答案数据作为训练标签,对预构建的初始预训练问答模型进行迭代训练,直到得到训练完成的预训练问答模型。

[0098] 在本公开的一种示例性实施例中,可以通过以下方式获取问题模板和结构化查询语句模板:

[0099] 可以从预设数据集提取样本问题数据以及样本问题数据对应的样本结构化查询语句;对样本问题数据以及样本结构化查询语句中的关键词进行模糊匹配,确定模板填充位置;可以基于模板填充位置以及样本问题数据和样本结构化查询语句,生成问题模板和问题模板对应的结构化查询语句模板。

[0100] 预设数据集是指已经公开的SQL相关数据集,例如,预设数据集可以是NL2SQL数据集、WikiSQL数据集、Spider数据集和追一数据集等,当然,预设数据集还可以是其他类型的SQL相关数据集,本示例实施例不以此为限。

[0101] 样本问题数据是指从预设数据集中提取到的问题数据,样本结构化查询语句是指从预设数据集中提取到的与样本问题数据所对应的结构化查询语句。

[0102] 可以对样本问题数据以及样本结构化查询语句中的关键词进行模糊匹配,确定模板填充位置;其中,模板填充位置是指样本问题数据以及样本结构化查询语句中可以被替换掉的数据内容,模板填充位置就是问题模板和结构化查询语句模板中用于填充数据内容的数据槽位,该模板填充位置可以包括数据位置和数据类型,其中数据类型是指模板中填充到该数据位置的数据内容的类型,例如,数据类型可以包括数值型和字符型,通过将模板中的模板填充位置的数据类型,可以有效限制填充到模板中的数据,保证填充后得到的样本问题数据和样本结构化查询语句的准确性,保证样本数据质量。

[0103] 由于SQL中的部分在问题描述可能会发生一些语义转化,如“A市大学”与“A大”,在生成问题模板和结构化查询语句模板时,通过模糊匹配的方式可以准确确定出样本问题数据以及样本结构化查询语句中的模板填充位置,将之用不同的字符标识出来,能够有效提升问题模板和结构化查询语句模板的准确性以及通用性,保证模板质量。

[0104] 可选的,如果有除了表格数据中的列类型之外的(因为列类型在语言中经常被省略)任何数据无法从问题中被标识出来,则直接丢弃当前的样本实例,进一步保证问题模板和结构化查询语句模板的质量。

[0105] 可选的,可以通过以下步骤实现根据预设的训练表格数据、问题模板和结构化查询语句模板生成训练三元组数据:

[0106] 可以获取预设的训练表格数据,并对训练表格数据进行拆解,得到训练子表格数据;基于训练子表数据对问题模板和结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句;根据训练结构化查询语句在训练表格数据中查询,得到训练答案数据;通过训练表格数据、训练问题数据和训练答案数据构成训练三元组数据。

[0107] 其中,训练子表格数据是指对训练表格数据进行拆解得到的数据量较小的表格,

例如,可以按照训练表格数据中的列类型对表格进行拆解得到训练子表格数据,也可以按照训练表格数据的尺寸以及预先设置的表格尺寸对训练表格数据进行拆解,得到训练子表格数据,当然,还可以通过其他方式对训练表格数据进行拆解,本示例实施例不以此为限。

[0108] 可以通过训练子表数据对问题模板和结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句,进而可以将训练结构化查询语句在训练表格数据进行执行,查询得到训练答案数据,通过训练表格数据、训练问题数据和训练答案数据构成训练三元组数据。通过问题模板和结构化查询语句模板可以有效提升训练三元组数据的数量,同时,由于并不需要手动标注,可以节省大量的人力物力,模板的随机填充也可以有效提升训练三元组数据的覆盖范围,提升预训练问答模型的泛化能力,提升预训练问答模型的准确性。

[0109] 可选的,可以通过以下步骤实现对问题模板和结构化查询语句模板进行随机填充:

[0110] 可以确定结构化查询语句模板的模板类型,并根据模板类型对问题模板和结构化查询语句模板进行分类;基于训练子表数据对不同分类结果下的问题模板和结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句。

[0111] 其中,模板类型可以是通过结构化查询语句模板的复杂程度划分的,具体可以通过结构化查询语句模板的子句数量确定结构化查询语句模板的复杂程度,例如,只有一个子句数量的结构化查询语句模板可以划分到一个模板类型,两个子句数量的结构化查询语句模板可以划分到一个模板类型,以此类推,可以得到不同模板类型的结构化查询语句模板。

[0112] 可以按照模板类型对结构化查询语句模板以及相应的问题模板进行分类,基于训练子表数据对不同分类结果下的问题模板和结构化查询语句模板进行随机填充,可以得到不同复杂程度的样本结构化查询语句,可以使得到的训练三元组数据尽可能覆盖所有场景下的问题语句以及结构化查询语句,保证训练三元组数据的丰富性,进一步提升预训练问答模型的泛化能力,提高预训练问答模型的准确性。

[0113] 在本公开的一种示例性实施例中,可以通过以下步骤对预训练问答模型进行定向训练:可以获取属于不同业务领域类型的样本三元组数据,进而可以根据样本三元组数据对训练完成的预训练问答模型进行定向训练,得到属于不同业务领域类型的预训练问答模型。

[0114] 在得到训练完成的预训练问答模型后,由于此时的预训练问答模型仅具备表格问答的能力,即是一个通用模型,进而可以根据属于不同业务领域类型的样本三元组数据,对训练完成的预训练问答模型进行进一步训练,可以得到应用于不同业务领域类型的目标问答模型,可以有效保证各个业务领域类型的目标问答模型的准确性。

[0115] 通过已有的公开数据集对本实施例中提供的预训练方式进行评测,相比于传统的生成SQL查询语句查询答案数据,准确率得到了有效提升,例如,通过实现可以得到表1的数据:

[0116] 表1准确率评测

[0117]

评测数据集	是否使用本实施例的预训练方式	准确率
追一	否	68.8%

追一	是	78.9%
SDCUP	否	69.7%
SDCUP	是	81.8%

[0118] 由此可见,采用本实施例中所提供的预训练方式,即先通过问题模板和结构化查询语句模板生成的训练三元组数据训练得到预训练问答模型,进而根据不同业务领域类型的样本数据对预训练问答模型进行定向训练得到目标问答模型,将问题数据和表格数据输入到目标问答模型中输出答案数据,相比于相关技术中先生成SQL查询语句,在通过SQL查询语句在数据库中执行得到答案数据的方式,不仅提高了答案数据的查询效率,而且得到的答案数据的准确率得到有效提升。

[0119] 需要说明的是,尽管在附图中以特定顺序描述了本公开中方法的各个步骤,但是,这并非要求或者暗示必须按照该特定顺序来执行这些步骤,或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的,可以省略某些步骤,将多个步骤合并为一个步骤执行,以及/或者将一个步骤分解为多个步骤执行等。

[0120] 此外,在本示例实施例中,还提供了一种数据查询装置。参照图4所示,该数据查询装置400包括:

[0121] 问题获取模块410用于获取问题数据;

[0122] 表格确定模块420用于确定所述问题数据对应的表格数据;

[0123] 问答模型匹配模块430用于匹配所述问题数据对应的目标问答模型;

[0124] 答案查询模块440用于将所述问题数据与所述表格数据输入到所述目标问答模型中,生成所述问题数据对应的答案数据。

[0125] 在本公开的一种示例性实施例中,基于前述方案,问答模型匹配模块可以用于:

[0126] 确定所述问题数据对应的业务领域类型;

[0127] 根据所述业务领域类型匹配所述问题数据对应的目标问答模型,所述目标问答模型是通过对预训练问答模型进行定向训练得到的。

[0128] 在本公开的一种示例性实施例中,基于前述方案,数据查询装置可以包括预训练模块,该预训练模块可以用于:

[0129] 获取问题模板以及所述问题模板对应的结构化查询语句模板;

[0130] 根据预设的训练表格数据、所述问题模板和所述结构化查询语句模板生成训练三元组数据;

[0131] 通过所述训练三元组数据对预构建的初始预训练问答模型进行迭代训练,得到训练完成的预训练问答模型。

[0132] 在本公开的一种示例性实施例中,基于前述方案,预训练模块可以用于:

[0133] 从预设数据集提取样本问题数据以及所述样本问题数据对应的样本结构化查询语句;

[0134] 对所述样本问题数据以及所述样本结构化查询语句中的关键词进行模糊匹配,确定模板填充位置,所述模板填充位置包括数据位置和数据类型;

[0135] 基于所述模板填充位置以及所述样本问题数据和所述样本结构化查询语句,生成所述问题模板和所述问题模板对应的结构化查询语句模板。

[0136] 在本公开的一种示例性实施例中,基于前述方案,预训练模块可以用于:

[0137] 获取预设的训练表格数据,并对所述训练表格数据进行拆解,得到训练子表格数据;

[0138] 基于所述训练子表数据对所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句;

[0139] 根据所述训练结构化查询语句在所述训练表格数据中查询,得到训练答案数据;

[0140] 通过所述训练表格数据、所述训练问题数据和所述训练答案数据构成训练三元组数据。

[0141] 在本公开的一种示例性实施例中,基于前述方案,预训练模块可以用于:

[0142] 确定所述结构化查询语句模板的模板类型,并根据所述模板类型对所述问题模板和所述结构化查询语句模板进行分类;其中,所述模板类型是通过所述结构化查询语句模板的复杂程度划分的;

[0143] 基于所述训练子表数据对不同分类结果下的所述问题模板和所述结构化查询语句模板进行随机填充,得到训练问题数据和训练结构化查询语句。

[0144] 在本公开的一种示例性实施例中,基于前述方案,预训练模块可以用于:

[0145] 获取属于不同业务领域类型的样本三元组数据;

[0146] 根据所述样本三元组数据对训练完成的预训练问答模型进行定向训练,得到属于不同业务领域类型的目标问答模型。

[0147] 上述中数据查询装置各模块的具体细节已经在对应的数据查询方法中进行了详细的描述,因此此处不再赘述。

[0148] 应当注意,尽管在上文详细描述中提及了数据查询装置的若干模块或者单元,但是这种划分并非强制性的。实际上,根据本公开的实施方式,上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之,上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0149] 此外,在本公开的示例性实施例中,还提供了一种能够实现上述数据查询方法的电子设备。

[0150] 所属技术领域的技术人员能够理解,本公开的各个方面可以实现为系统、方法或程序产品。因此,本公开的各个方面可以具体实现为以下形式,即:完全的硬件实施例、完全的软件实施例(包括固件、微代码等),或硬件和软件方面结合的实施例,这里可以统称为“电路”、“模块”或“系统”。

[0151] 下面参照图5来描述根据本公开的这种实施例的电子设备500。图5所示的电子设备500仅仅是一个示例,不应对本公开实施例的功能和使用范围带来任何限制。

[0152] 如图5所示,电子设备500以通用计算设备的形式表现。电子设备500的组件可以包括但不限于:上述至少一个处理单元510、上述至少一个存储单元520、连接不同系统组件(包括存储单元520和处理单元510)的总线530、显示单元540。

[0153] 其中,所述存储单元存储有程序代码,所述程序代码可以被所述处理单元510执行,使得所述处理单元510执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施例的步骤。例如,所述处理单元510可以执行如图1中所示的步骤S210,获取问题数据;步骤S220,确定所述问题数据对应的表格数据;步骤S230,匹配所述问题数据对应的目标问答模型;步骤S240,将所述问题数据与所述表格数据输入到所述目标问答模型中,生

成所述问题数据对应的答案数据。

[0154] 存储单元520可以包括易失性存储单元形式的可读介质,例如随机存取存储单元(RAM) 521和/或高速缓存存储单元522,还可以进一步包括只读存储单元(ROM) 523。

[0155] 存储单元520还可以包括具有一组(至少一个)程序模块525的程序/实用工具524,这样的程序模块525包括但不限于:操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0156] 总线530可以为表示几类总线结构中的一种或多种,包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0157] 电子设备500也可以与一个或多个外部设备570(例如键盘、指向设备、蓝牙设备等)通信,还可与一个或者多个使得用户能与该电子设备500交互的设备通信,和/或与使得该电子设备500能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口550进行。并且,电子设备500还可以通过网络适配器560与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器560通过总线530与电子设备500的其它模块通信。应当明白,尽管图中未示出,可以结合电子设备500使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0158] 通过以上的实施例的描述,本领域的技术人员易于理解,这里描述的示例实施例可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本公开实施例的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开实施例的方法。

[0159] 在本公开的示例性实施例中,还提供了一种计算机可读存储介质,其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施例中,本公开的各个方面还可以实现为一种程序产品的形式,其包括程序代码,当所述程序产品在终端设备上运行时,所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施例的步骤。

[0160] 参考图6所示,描述了根据本公开的实施例的用于实现上述数据查询方法的程序产品600,其可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码,并可以在终端设备,例如个人电脑上运行。然而,本公开的程序产品不限于此,在本文件中,可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0161] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括:具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0162] 计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了可读程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质,该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0163] 可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于无线、有线、光缆、RF等等,或者上述的任意合适的组合。

[0164] 可以以一种或多种程序设计语言的任意组合来编写用于执行本公开操作的程序代码,所述程序设计语言包括面向对象的设计语言—诸如Java、C++等,还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中,远程计算设备可以通过任意种类的网络,包括局域网(LAN)或广域网(WAN),连接到用户计算设备,或者,可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0165] 此外,上述附图仅是根据本公开示例性实施例的方法所包括的处理的示意性说明,而不是限制目的。易于理解,上述附图所示的处理并不表明或限制这些处理的时间顺序。另外,也易于理解,这些处理可以是例如在多个模块中同步或异步执行的。

[0166] 通过以上的实施例的描述,本领域的技术人员易于理解,这里描述的示例实施例可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本公开实施例的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、触控终端、或者网络设备等)执行根据本公开实施例的方法。

[0167] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其它实施例。本申请旨在涵盖本公开的任何变型、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公开的一般性原理并包括本公开未公开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,本公开的真正范围和精神由权利要求指出。

[0168] 应当理解的是,本公开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。本公开的范围仅由所附的权利要求来限制。

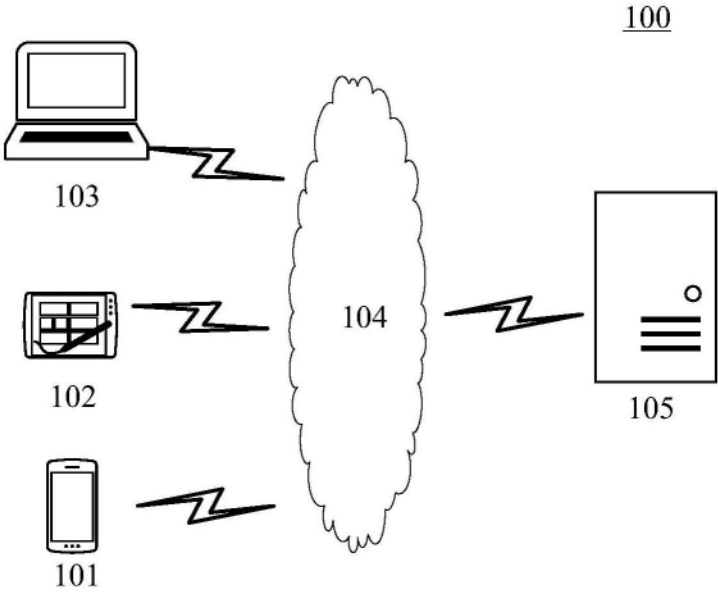


图1

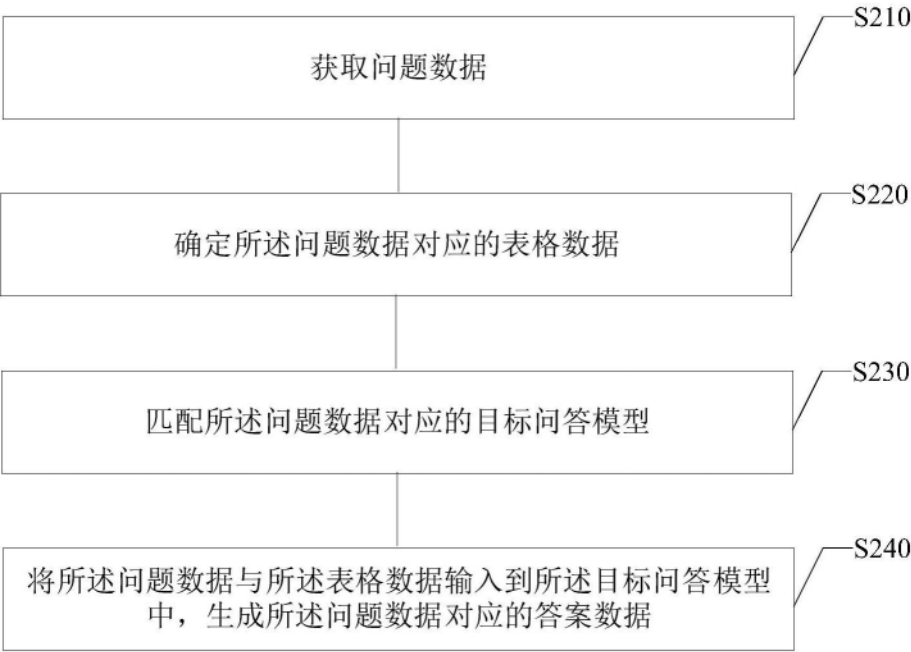


图2

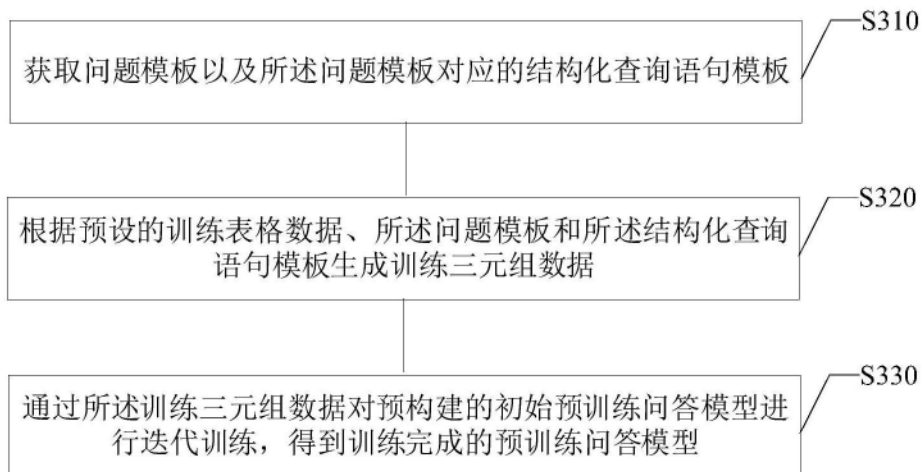


图3

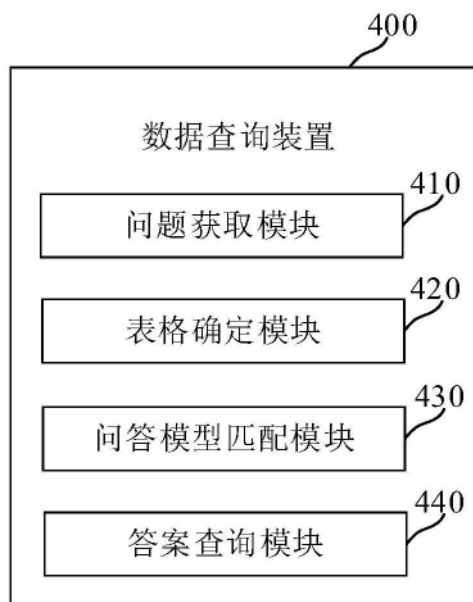


图4

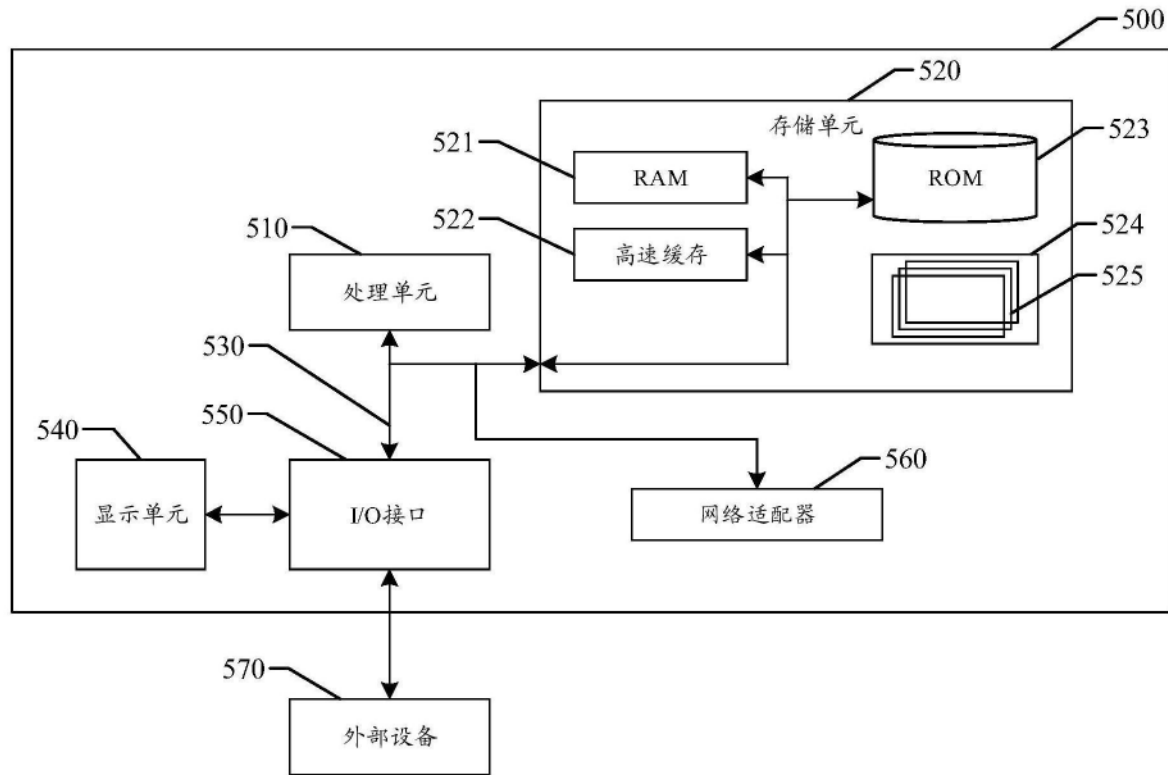


图5

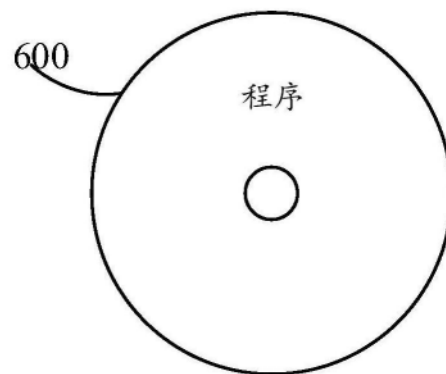


图6