(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112559552 B
(45)授权公告日 2023.07.25

(21)申请号 202011410065.5

(22)申请日 2020.12.03

(65)同一申请的已公布的文献号
申请公布号 CN 112559552 A

(43)申请公布日 2021.03.26

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 王丽杰 张傲

(74)专利代理机构 北京鸿德海业知识产权代理有限公司 11412
专利代理师 谷春静

(51)Int.Cl.
G06F 16/242(2019.01)
G06F 16/2453(2019.01)
G06F 40/30(2020.01)

(56)对比文件
CN 111522839 A,2020.08.11
US 2014095469 A1,2014.04.03
CN 106649294 A,2017.05.10
US 2012078895 A1,2012.03.29
崔跃生;张勇;曾春;冯建华;邢春晓.数据库物理结构优化技术.软件学报.2013,(第04期),全文.
Qing Li;Lili Li;Qi Li;Jiang Zhong.A Comprehensive Exploration on Spider with Fuzzy Decision Text-to-SQL Model Publisher: IEEE Cite This PDF.IEEE.2019,全文.
孙理和.浅议成人英语教学中辅助课堂实践的句法分析.甘肃科技纵横.2007,(第02期),全文.

审查员 景京

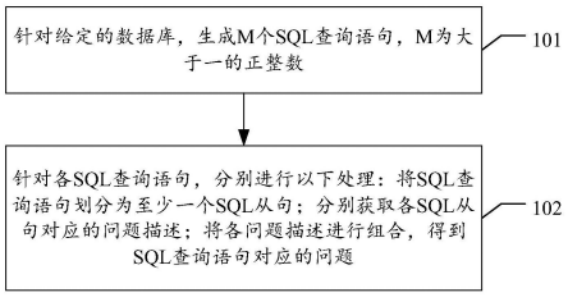
权利要求书2页 说明书9页 附图3页

(54)发明名称

数据对生成方法、装置、电子设备及存储介质

(57)摘要

本申请公开了数据对生成方法、装置、电子设备及存储介质,涉及自然语言处理及深度学习等人工智能领域,其中的方法可包括:针对给定的数据库,生成M个SQL查询语句,M为大于一的正整数;针对各SQL查询语句,分别进行以下处理:将SQL查询语句划分为至少一个SQL从句;分别获取各SQL从句对应的问题描述;将各问题描述进行组合,得到SQL查询语句对应的问题。应用本申请所述方案,可节省人力及时间成本等。



CN 112559552 B

1.一种数据对生成方法,包括:

针对给定的数据库,生成M个结构化查询语言SQL查询语句,M为大于一的正整数;

针对各SQL查询语句,分别进行以下处理:将所述SQL查询语句划分为至少一个SQL从句;利用预先训练得到的生成模型,分别生成各SQL从句对应的问题描述;将各问题描述进行组合,得到所述SQL查询语句对应的问题;

其中,所述生成模型为根据SQL从句-问题描述对训练得到的,所述SQL从句-问题描述对为根据已有的问题-SQL查询语句对构建出的,其中,针对对任一所述问题-SQL查询语句对中的所述SQL查询语句进行划分后得到的至少一个SQL从句,其对应的问题描述分别包括:通过字符串匹配的方式对齐所述问题-SQL查询语句对中的所述问题和所述SQL查询语句中的单元后、确定出的所述问题中覆盖所述SQL从句中的所有单元的问题片段,所述问题片段为覆盖所述SQL从句中的所有单元的最短问题片段,或者,所述问题片段为在所述最短问题片段的基础上、分别向左右扩展出部分没有匹配到任何单元的词后得到的问题片段。

2.根据权利要求1所述的方法,其中,所述生成M个SQL查询语句包括:根据基于SQL语法总结出的产生式规则,生成M个SQL查询语句。

3.根据权利要求1所述的方法,其中,所述将所述SQL查询语句划分为至少一个SQL从句包括:

根据所述SQL查询语句的结构,将所述SQL查询语句划分为至少一个SQL从句,其中,各SQL从句均为语义独立且完整的。

4.根据权利要求1所述的方法,其中,所述将各问题描述进行组合包括:

根据各SQL从句的执行顺序,将各SQL从句对应的问题描述进行组合。

5.根据权利要求1所述的方法,还包括:

将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中,训练语义解析模型,所述训练数据集中包括人工标注的训练数据。

6.根据权利要求5所述的方法,其中,所述将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中,训练语义解析模型包括:

在每一轮的训练中,从生成的问题-SQL查询语句对中随机采样与人工标注的训练数据同等规模的问题-SQL查询语句对,利用两种训练数据训练所述语义解析模型。

7.一种数据对生成装置,包括:第一生成模块以及第二生成模块;

所述第一生成模块,用于针对给定的数据库,生成M个结构化查询语言SQL查询语句,M为大于一的正整数;

所述第二生成模块,用于针对各SQL查询语句,分别进行以下处理:将所述SQL查询语句划分为至少一个SQL从句;利用预先训练得到的生成模型,分别生成各SQL从句对应的问题描述;将各问题描述进行组合,得到所述SQL查询语句对应的问题;

其中,所述生成模型为根据SQL从句-问题描述对训练得到的,所述SQL从句-问题描述对为根据已有的问题-SQL查询语句对构建出的,其中,针对对任一所述问题-SQL查询语句对中的所述SQL查询语句进行划分后得到的至少一个SQL从句,其对应的问题描述分别包括:通过字符串匹配的方式对齐所述问题-SQL查询语句对中的所述问题和所述SQL查询语句中的单元后、确定出的所述问题中覆盖所述SQL从句中的所有单元的问题片段,所述问题片段为覆盖所述SQL从句中的所有单元的最短问题片段,或者,所述问题片段为在所述最短

问题片段的基础上、分别向左右扩展出部分没有匹配到任何单元的词后得到的问题片段。

8.根据权利要求7所述的装置,其中,

所述第一生成模块根据基于SQL语法总结出的产生式规则,生成M个SQL查询语句。

9.根据权利要求7所述的装置,其中,

所述第二生成模块针对任一SQL查询语句,根据所述SQL查询语句的结构,将所述SQL查询语句划分为至少一个SQL从句,其中,各SQL从句均为语义独立且完整的。

10.根据权利要求7所述的装置,其中,

所述第二生成模块根据各SQL从句的执行顺序,将各SQL从句对应的问题描述进行组合。

11.根据权利要求7所述的装置,还包括:第二训练模块;

所述第二训练模块,用于将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中,训练语义解析模型,所述训练数据集中包括人工标注的训练数据。

12.根据权利要求11所述的装置,其中,

所述第二训练模块在每一轮的训练中,从生成的问题-SQL查询语句对中随机采样与人工标注的训练数据同等规模的问题-SQL查询语句对,利用两种训练数据训练所述语义解析模型。

13.一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-6中任一项所述的方法。

14.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使计算机执行权利要求1-6中任一项所述的方法。

数据对生成方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及人工智能技术领域，特别涉及自然语言处理及深度学习领域的数据对生成方法、装置、电子设备及存储介质。

背景技术

[0002] 语义解析(text-to-SQL)是语言理解的核心技术，旨在自动地将自然语言问题转化为可与数据库交互的结构化查询语言(SQL，Structured Query Language)查询语句。

[0003] 针对任一问题，可通过预先训练得到的语义解析模型来生成问题对应的SQL查询语句。语义解析模型通常都是基于标注的训练数据通过有指导方式训练得到的。在实际应用中，经常会遇到新数据库，即训练数据中未见过的数据库，当前的语义解析模型对于新数据库具有一定的泛化能力，但效果并不理想，因此最好有基于新数据库的训练数据。

[0004] 目前，通常采用人工标注的方式来构建训练数据，即问题-SQL查询语句对，但这种方式需要耗费大量的人力和时间成本，且效率低下。

发明内容

[0005] 本申请提供了数据对生成方法、装置、电子设备及存储介质。

[0006] 一种数据对生成方法，包括：

[0007] 针对给定的数据库，生成M个结构化查询语言SQL查询语句，M为大于一的正整数；

[0008] 针对各SQL查询语句，分别进行以下处理：将所述SQL查询语句划分为至少一个SQL从句；分别获取各SQL从句对应的问题描述；将各问题描述进行组合，得到所述SQL查询语句对应的问题。

[0009] 一种数据对生成装置，包括：第一生成模块以及第二生成模块；

[0010] 所述第一生成模块，用于针对给定的数据库，生成M个结构化查询语言SQL查询语句，M为大于一的正整数；

[0011] 所述第二生成模块，用于针对各SQL查询语句，分别进行以下处理：将所述SQL查询语句划分为至少一个SQL从句；分别获取各SQL从句对应的问题描述；将各问题描述进行组合，得到所述SQL查询语句对应的问题。

[0012] 一种电子设备，包括：

[0013] 至少一个处理器；以及

[0014] 与所述至少一个处理器通信连接的存储器；其中，

[0015] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行如以上所述的方法。

[0016] 一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使计算机执行如以上所述的方法。

[0017] 一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现如以上所述的方法。

[0018] 上述申请中的一个实施例具有如下优点或有益效果:针对给定的数据库,可首先生成多个SQL查询语句,之后可针对各SQL查询语句,分别通过一系列处理生成对应的问题,从而实现了自动地生成问题-SQL查询语句对,相比于现有方式节省了人力和时间成本,并提升了处理效率等。

[0019] 应当理解,本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征,也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0020] 附图用于更好地理解本方案,不构成对本申请的限定。其中:

[0021] 图1为本申请所述数据对生成方法实施例的流程图;

[0022] 图2为本申请所述SQL查询语句x对应的树的示意图;

[0023] 图3为本申请所述生成SQL查询语句y对应的问题的过程示意图;

[0024] 图4为本申请所述对齐方式示意图;

[0025] 图5为本申请所述数据对生成装置50实施例的组成结构示意图;

[0026] 图6为根据本申请实施例所述方法的电子设备的框图。

具体实施方式

[0027] 以下结合附图对本申请的示范性实施例做出说明,其中包括本申请实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本申请的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0028] 另外,应理解,本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0029] 图1为本申请所述数据对生成方法实施例的流程图。所述数据对即指问题-SQL查询语句对。如图1所示,包括以下具体实现方式。

[0030] 在步骤101中,针对给定的数据库,生成M个SQL查询语句,M为大于一的正整数。

[0031] 在步骤102中,针对各SQL查询语句,分别进行以下处理:将SQL查询语句划分为至少一个SQL从句(Clause);分别获取各SQL从句对应的问题描述;将各问题描述进行组合,得到SQL查询语句对应的问题。

[0032] 可以看出,上述方法实施例所述方案中,针对给定的数据库,可首先生成多个SQL查询语句,之后可针对各SQL查询语句,分别通过一系列处理生成对应的问题,从而实现了自动地生成问题-SQL查询语句对,相比于现有方式节省了人力和时间成本,并提升了处理效率等。

[0033] 以下分别对上述各部分内容的具体实现进行详细说明。

[0034] 一)生成SQL查询语句

[0035] 针对给定的数据库,可生成M个SQL查询语句。优选地,可根据基于SQL语法总结出的产生式规则,生成M个SQL查询语句,M的具体取值可根据实际需要而定。

[0036] SQL是一个基于自己语法可执行的语言,基于SQL的语法,可总结出一些产生式规

则,比如,可如下所示。

[0037] SQLs::=SQL|SQL交集(intersect)SQLs|SQL并集(union)SQLs|SQL差异(except)SQLs

[0038] SQL::=Select|Select Where|Select Group|Select Where Group|Select Order|Select Where Order|Select From SQL,SQL

[0039] 选择(Select)::=SELECT A|SELECT A A|SELECT A A A|SELECT A A A A

[0040] 过滤(Where)::=WHERE Conditions

[0041] 分组(Group)::=GROUP BY C|GROUP BY C HAVING Conditions|GROUP BY C Order

[0042] 排序(Order)::=ORDER BY C Dir|ORDER BY C Dir筛选值(LIMIT value)|ORDER BY A Dir LIMIT value

[0043] Dir::=升序(ASC)|降序(DESC)

[0044] 条件(Conditions)::=Condition|Condition AND Conditions|Condition OR Conditions

[0045] Condition::=A op value|A op SQL

[0046] A::=C|最小值(min)C|最大值(max)C|平均值(avg)C|数量(count)C|和(sum)C

[0047] C::=表格.列(table.column)|table.column mathop table.column

[0048] mathop::=+|-|*|/

[0049] op::=|==|!=|>|>=|<=|like|in|not in|exist|between

[0050] 上述加粗的内容可用来产生复杂语法,如嵌套查询、多子句查询等。

[0051] 基于所述产生式规则可以将任意一个SQL查询语句表示为一棵树。比如,SQL查询语句x为“将员工按年龄升序排序后选择姓名(SELECT name FROM employee ORDER BY age ASC)”,如图2所示,图2为本申请所述SQL查询语句x对应的树的示意图。对应的产生式序列为:{SQLs=SQL,SQL=Select Order,Select=SELECT A,Order=ORDER BY C Dir,A=C,C=table.column,Dir=ASC,C=table.column}等,该产生式序列也展示了SQL查询语句x的生成过程。

[0052] 基于所述产生式规则,可方便准确地生成各种可能的SQL查询语句,生成的SQL查询语句需要尽可能多地覆盖各种SQL形式,以提升后续的语义解析模型训练效果等。

[0053] 二)生成SQL查询语句对应的问题

[0054] 对于任一SQL查询语句,需要生成一个高质量的问题来描述它。其中,高质量的问题需要满足以下两点:1)符合SQL查询语句的结构性,即需要表达出SQL查询语句的嵌套等结构;2)保证SQL查询语句的语义正确性,尤其是保证其包含的数据库元素的语义。

[0055] 通过对数据的分析发现,复杂的SQL查询语句都可以划分为由简单且常见的SQL从句(或称为SQL子句)构成。比如,上述SQL查询语句x即由两个常见的SQL从句构成,即Select从句和Order从句,即图2中所示的从句1和从句2。其中,每一个SQL从句都是语义独立且完整的。

[0056] 因此,本申请中提出了一种基于SQL结构性分层生成对应问题的机制。针对每个SQL查询语句,可分别进行以下处理:将SQL查询语句划分为至少一个SQL从句;分别获取各SQL从句对应的问题描述;将各问题描述进行组合,得到SQL查询语句对应的问题。

[0057] 优选地，针对每个SQL查询语句，可首先根据SQL查询语句的结构，将SQL查询语句划分为至少一个SQL从句，其中，各SQL从句均为语义独立且完整的。

[0058] 比如，SQL结构(Structure)可包括：WHERE A_1 op SELECT A_2，它来自嵌套查询{WHERE A_1 op SQL}，其中op来自{>，≥，<，≤，=，≠}的集合，A_2是SQL的select从句中的组件(It is from the nested query{WHERE A_1op SQL}，where op is from the set of{>，≥，<，≤，=，≠}and A_2 is the component in the select clause of SQL)；WHERE $table_1$ op $table_2$，表示在表1和表2上执行了{和，或，否}集合中的操作，它来自嵌套查询{WHERE A op SQL}op来自{存在于，不存在于}集合，或多SQL查询，如{SQL交集SQLs}(It means an operation from the set of{and，or，not}is performed on $table_1$ and $table_2$，It is from the nested query{WHERE A op SQL}with the op from{in，not in}，or multi-SQL queries，such as{SQL intersect SQLs})。

[0059] 参照上述产生式规则，划分得到的SQL从句可包括以下形式：SELECT AFROM table、SELECT A GROUP BY C、GROUP BY C HAVING Conditions、GROUP BY C ORDER BY A Dir、GROUP BY C ORDER BY A Dir LIMIT value、ORDER BY C Dir、ORDER BY C Dir LIMIT value、WHERE C op value等。

[0060] 针对划分得到的各SQL从句，可分别获取对应的问题描述。优选地，针对任一SQL从句，可分别利用预先训练得到的生成模型，生成该SQL从句对应的问题描述。

[0061] 进一步地，可将各SQL从句对应的问题描述进行组合，从而得到最终所需的问题，即SQL查询语句对应的问题。优选地，可根据各SQL从句的执行顺序，即SQL查询语句的结构，将各SQL从句对应的问题描述进行组合。

[0062] 通过上述处理，可使得得到的问题与对应的SQL查询语句表达相同的语义，且符合SQL查询语句的结构等，即可确保生成高质量的问题。

[0063] 以SQL查询语句y“将员工按年龄升序排序后选择前三个(SELECT name FROM employee ORDER BY age ASC LIMIT 3)”为例，图3为本申请所述生成SQL查询语句y对应的问题的过程示意图。

[0064] 如图3所示，首先，可根据SQL查询语句y的结构，将SQL查询语句y划分为两个SQL从句，分别为“从员工中选择姓名(select name from employee)”以及“按年龄升序排序后前三个(order by age asc limit 3)”，针对每个SQL从句，可分别利用生成模型生成图3中所示的对应的问题描述，进而可根据两个SQL从句的执行顺序，将两个SQL从句对应的问题描述进行组合，从而得到SQL查询语句y对应的问题“列出最年轻的三名员工姓名(list the employee name of 3youngest)”。

[0065] 如图3所示，生成模型可为基于编码-解码(Encoder-Decoder)结构的序列到序列(Seq2Seq)生成模型。优选地，生成模型可为带复制的生成模型，即将复制机制整合到序列到序列学习中(Incorporating Copying Mechanism in Sequence-to-Sequence Learning)，输入为SQL从句，输出为对应的问题描述。

[0066] 上述生成模型可为预先训练得到的。优选地，可根据已有的问题-SQL查询语句对构建SQL从句-问题描述对，根据构建的SQL从句-问题描述对训练得到生成模型。

[0067] 其中，针对任一问题-SQL查询语句对，可分别进行以下处理：将问题-SQL查询语句对中的SQL查询语句划分为至少一个SQL从句；分别获取各SQL从句对应的问题描述；其中，

对于任一SQL从句,其对应的问题描述包括:问题-SQL查询语句对中的问题中覆盖该SQL从句中的所有单元的问题片段;相应地,可将各SQL从句及对应的问题描述分别作为一个构建出的SQL从句-问题描述对。

[0068] 比如,某一问题-SQL查询语句对中的问题为“显示2014年或之后举办最多音乐会的体育场名称和容量(Show the stadium name and capacity with most number of concerts in year 2014or after)”,对应的SQL查询语句为“SELECT T2.name, T2.capacity FROM concert AS T1 JOIN stadium AS T2 ON T1.stadium_id＝T2.stadium_id WHERE T1.year>＝2014GROUP BY T1.stadium_id ORDER BY count(*) DESC LIMIT 1”,可对齐问题和SQL查询语句中的单元(unit),单元可定义为表格名、列名、值、聚合操作等,可采用字符串匹配的方式进行对齐,如图4所示,图4为本申请所述对齐方式示意图。

[0069] 针对SQL查询语句“SELECT T2.name,T2.capacity FROM concert AS T1 JOIN stadium AS T2 ON T1.stadium_id＝T2.stadium_id WHERE T1.year>＝2014 GROUP BY T1.stadium_id ORDER BY count(*)DESC LIMIT 1”,可将其划分为三个SQL从句,分别为“选择体育场名称和容量(SELECT name,capacity FROM stadium)”、“音乐会年份大于或等于2014(WHERE concert year>＝2014)”以及“按举办音乐会次数降序排序后处于第一位的体育场GROUP BY stadium id ORDER BY count concert DESC LIMIT 1”,可分别获取这三个SQL从句对应的问题描述,其中,SQL从句“SELECT name,capacity FROM stadium”对应的问题描述可为“显示体育场名称和容量(show the stadium name and capacity with)”,SQL从句“WHERE concert year>＝2014”对应的问题描述可为“2014年或之后的音乐会(of concerts in year 2014or after)”,SQL从句“GROUP BY stadium id ORDER BY count concert DESC LIMIT 1”对应的问题描述可为“举办了最多次数的音乐会(with most number of concerts in)”,这样,可得到三个SQL从句-问题描述对,分别为“SELECT name, capacity FROM stadium”-“show the stadium name and capacity with”、“WHERE concert year>＝2014”-“of concerts in year 2014or after”以及“GROUP BY stadium id ORDER BY count concert DESC LIMIT 1”-“with most number of concerts in”。

[0070] 如前所述,对于上述任一SQL从句,其对应的问题描述可为问题-SQL查询语句对中的问题中覆盖该SQL从句中的所有单元的问题片段,所述问题片段可以是指覆盖该SQL从句中的所有单元的最短问题片段,也可以是在最短问题片段的基础上,进一步左右扩展出一些没有匹配到任何单元的词到该片段中,以使得得到的问题片段的语义更为完整等。

[0071] 可以看出,按照本申请所述方式得到的问题描述和SQL从句并不一定是严格对齐的,但分析显示,这样的数据是比较稀疏的,对最终的结果影响不大。

[0072] 针对不同的问题-SQL查询语句对,可分别按照上述方式生成多个SQL从句-问题描述对,用于进行生成模型的训练。

[0073] 在实际应用中,可能会出现以下情况:同一个SQL从句对应不同的问题描述,比如,对于SQL从句“按年龄升序排序(ORDER BY age ASC)”,对应“按年龄升序排序(in ascending order of the age)”、“按年龄升序排序(sort them by age in ascending order)”以及“按年龄升序排序(from youngest to oldest)”等问题描述,可针对各问题描述,按照出现频次降序排序,并选出排序后处于前P位的问题描述,P为正整数,具体取值可

根据实际需要而定，比如，可选出排序后处于前三位的问题描述，假设分别为“in ascending order of the age”、“sort them by age in ascending order”以及“from youngest to oldest”，并分别与对应的SQL从句“ORDER BY age ASC”组成SQL从句-问题描述对，用于进行生成模型的训练。

[0074] 通过上述方式，可快速准确地构建出各个SQL从句-问题描述对，进而可训练得到生成模型，并确保了训练得到的生成模型具有很好的准确性等。

[0075] 进一步地，按照本申请所述方式自动生成多个问题-SQL查询语句对后，可将生成的问题-SQL查询语句对作为训练数据，训练语义解析模型。

[0076] 优选地，可将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中，训练语义解析模型，训练数据集中包括人工标注的训练数据。

[0077] 由于自动生成的训练数据的质量通常没有人工标注的训练数据的质量高，而且，自动生成的训练数据的分布与实际应用中的分布可能也不是很一致，因此，为了最大程度地发挥自动生成的训练数据的作用，可将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中，训练语义解析模型。

[0078] 优选地，可采用动态采样的方法，在每一轮的训练中，从生成的问题-SQL查询语句对中随机采样与人工标注的训练数据同等规模的问题-SQL查询语句对，利用两种训练数据训练语义解析模型，从而最大程度地发挥了自动生成的训练数据的作用，并提升了模型训练效果等。

[0079] 需要说明的是，对于前述的方法实施例，为了简单描述，将其表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

[0080] 以上是关于方法实施例的介绍，以下通过装置实施例，对本申请所述方案进行进一步说明。

[0081] 图5为本申请所述数据对生成装置50实施例的组成结构示意图。如图5所示，包括：第一生成模块501以及第二生成模块502。

[0082] 第一生成模块501，用于针对给定的数据库，生成M个SQL查询语句，M为大于一的正整数。

[0083] 第二生成模块502，用于针对各SQL查询语句，分别进行以下处理：将SQL查询语句划分为至少一个SQL从句；分别获取各SQL从句对应的问题描述；将各问题描述进行组合，得到SQL查询语句对应的问题。

[0084] 针对给定的数据库，第一生成模块501可生成M个SQL查询语句。优选地，可根据基于SQL语法总结出的产生式规则，生成M个SQL查询语句，M的具体取值可根据实际需要而定。

[0085] 第二生成模块502可分别生成各SQL查询语句对应的问题，从而得到所需的问题-SQL查询语句对。其中，针对任一SQL查询语句，第二生成模块502可将SQL查询语句划分为至少一个SQL从句，并可分别获取各SQL从句对应的问题描述，进而可将各问题描述进行组合，得到SQL查询语句对应的问题。

[0086] 具体地，第二生成模块502可针对任一SQL查询语句，根据SQL查询语句的结构，将

SQL查询语句划分为至少一个SQL从句,各SQL从句均为语义独立且完整的。

[0087] 针对划分得到的各SQL从句,第二生成模块502可分别获取对应的问题描述。优选地,针对任一SQL从句,可分别利用预先训练得到的生成模型,生成该SQL从句对应的问题描述。

[0088] 进一步地,第二生成模块502还可将各SQL从句对应的问题描述进行组合,从而得到最终所需的问题,即SQL查询语句对应的问题。优选地,可根据各SQL从句的执行顺序,即SQL查询语句的结构,将各SQL从句对应的问题描述进行组合。

[0089] 上述生成模块可为预先训练得到的,相应地,如图5所示,所述装置中还可包括:第一训练模块500,用于根据已有的问题-SQL查询语句对构建SQL从句-问题描述对,根据SQL从句-问题描述对训练得到生成模型。

[0090] 其中,针对任一问题-SQL查询语句对,第一训练模块500可分别进行以下处理:将问题-SQL查询语句对中的SQL查询语句划分为至少一个SQL从句;分别获取各SQL从句对应的问题描述;其中,对于任一SQL从句,其对应的问题描述包括:问题-SQL查询语句对中的问题中覆盖该SQL从句中的所有单元的问题片段;相应地,可将各SQL从句及对应的问题描述分别作为一个构建出的SQL从句-问题描述对。

[0091] 如图5所示,所述装置中还可包括:第二训练模块503,用于将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中,训练语义解析模型,训练数据集中包括人工标注的训练数据。

[0092] 由于自动生成的训练数据的质量通常没有人工标注的训练数据的质量高,而且,自动生成的训练数据的分布与实际应用中的分布可能也不是很一致,因此,为了最大程度地发挥自动生成的训练数据的作用,可将生成的问题-SQL查询语句对通过数据增强的方式加入到训练数据集中,训练语义解析模型。

[0093] 优选地,第二训练模块503可采用动态采样的方法,在每一轮的训练中,从生成的问题-SQL查询语句对中随机采样与人工标注的训练数据同等规模的问题-SQL查询语句对,利用两种训练数据训练语义解析模型。

[0094] 图5所示装置实施例的具体工作流程请参照前述方法实施例中的相关说明,不再赘述。

[0095] 总之,采用本申请装置实施例所述方案,针对给定的数据库,可首先生成多个SQL查询语句,之后可针对各SQL查询语句,分别通过一系列处理生成对应的问题,从而实现了自动地生成问题-SQL查询语句对,相比于现有方式节省了人力和时间成本,并提升了处理效率等。

[0096] 本申请所述方案可应用于人工智能领域,特别涉及自然语言处理及深度学习等领域。

[0097] 人工智能是研究使计算机来模拟人的某些思维过程和智能行为(如学习、推理、思考、规划等)的学科,既有硬件层面的技术也有软件层面的技术,人工智能硬件技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理等技术,人工智能软件技术主要包括计算机视觉技术、语音识别技术、自然语言处理技术以及机器学习/深度学习、大数据处理技术、知识图谱技术等几大方向。

[0098] 根据本申请的实施例,本申请还提供了一种电子设备和一种可读存储介质。

[0099] 如图6所示，是根据本申请实施例所述方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0100] 如图6所示，该电子设备包括：一个或多个处理器Y01、存储器Y02，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如，耦合至接口的显示设备)上显示图形用户界面的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作(例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图6中以一个处理器Y01为例。

[0101] 存储器Y02即为本申请所提供的非瞬时计算机可读存储介质。其中，所述存储器存储有可由至少一个处理器执行的指令，以使所述至少一个处理器执行本申请所提供的方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的方法。

[0102] 存储器Y02作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块，如本申请实施例中的方法对应的程序指令/模块。处理器Y01通过运行存储在存储器Y02中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的方法。

[0103] 存储器Y02可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据电子设备的使用所创建的数据等。此外，存储器Y02可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器Y02可选包括相对于处理器Y01远程设置的存储器，这些远程存储器可以通过网络连接至电子设备。上述网络的实例包括但不限于互联网、企业内部网、区块链网络、局域网、移动通信网及其组合。

[0104] 电子设备还可以包括：输入装置Y03和输出装置Y04。处理器Y01、存储器Y02、输入装置Y03和输出装置Y04可以通过总线或者其他方式连接，图6中以通过总线连接为例。

[0105] 输入装置Y03可接收输入的数字或字符信息，以及产生与电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置Y04可以包括显示设备、辅助照明装置和触觉反馈装置(例如，振动电机)等。该显示设备可以包括但不限于，液晶显示器、发光二极管显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

[0106] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用集成电路、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一

个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0107] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令，并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的，术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如，磁盘、光盘、存储器、可编程逻辑装置)，包括，接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0108] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置(例如，阴极射线管或者液晶显示器监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0109] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)、或者包括中间件部件的计算系统(例如，应用服务器)、或者包括前端部件的计算系统(例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如，通信网络)来将系统的部件相互连接。通信网络的示例包括：局域网、广域网、区块链网络和互联网。

[0110] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器，又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决了传统物理主机与VPS服务中，存在的管理难度大，业务扩展性弱的缺陷。

[0111] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

[0112] 上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

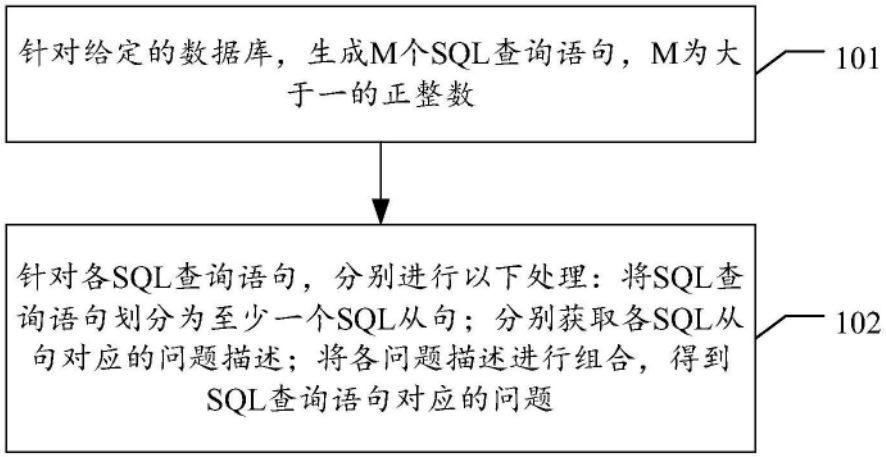

针对给定的数据库，生成M个SQL查询语句，M为大于一的正整数
101
针对各SQL查询语句，分别进行以下处理：将SQL查询语句划分为至少一个SQL从句；分别获取各SQL从句对应的问题描述；将各问题描述进行组合，得到SQL查询语句对应的问题
102

图1

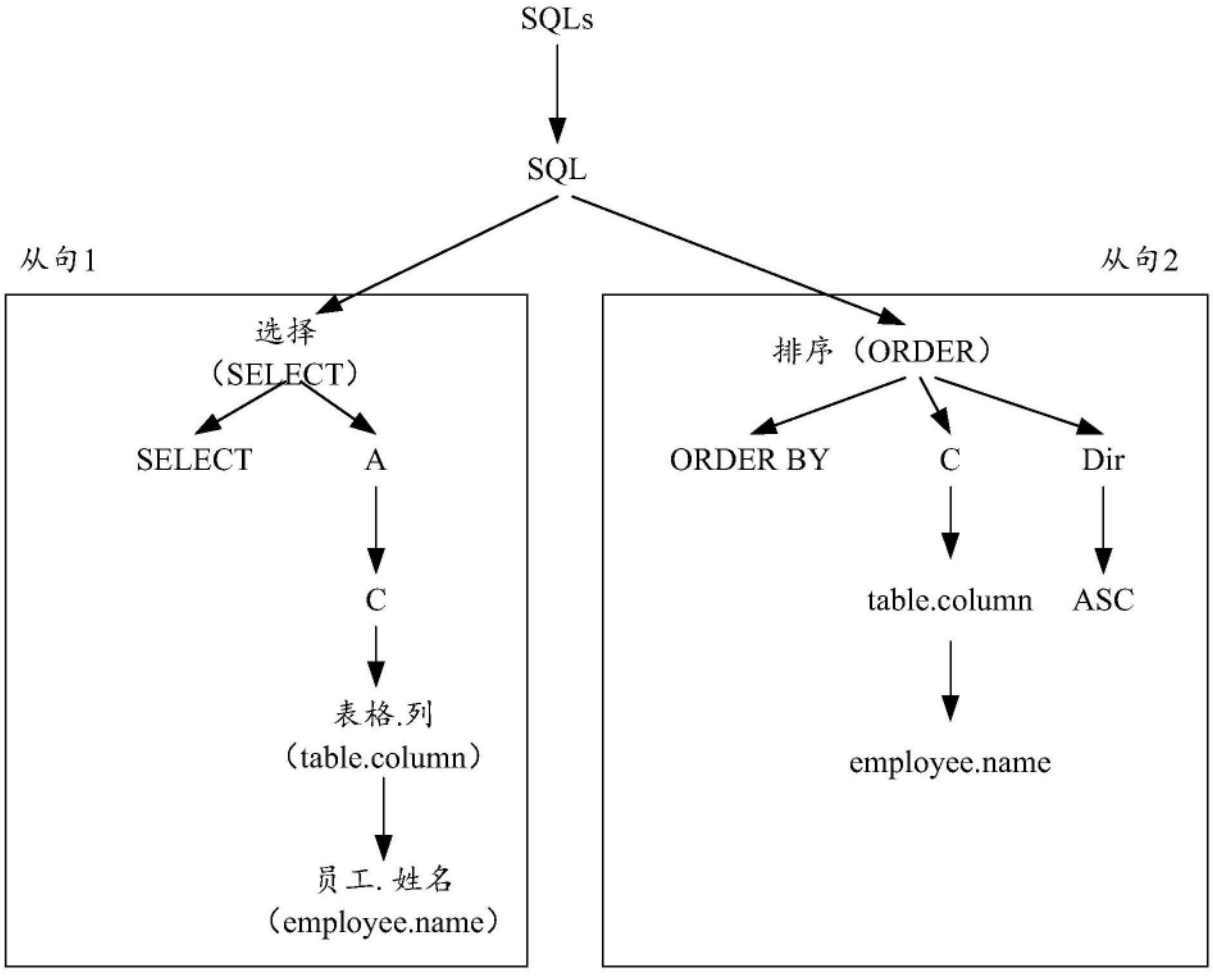

SQLs
SQL
从句1
从句2
选择
（SELECT）
SELECT
A
C
表格.列
（table.column）
员工.姓名
（employee.name）
排序（ORDER）
ORDER BY
C
Dir
table.column
ASC
employee.name

图2

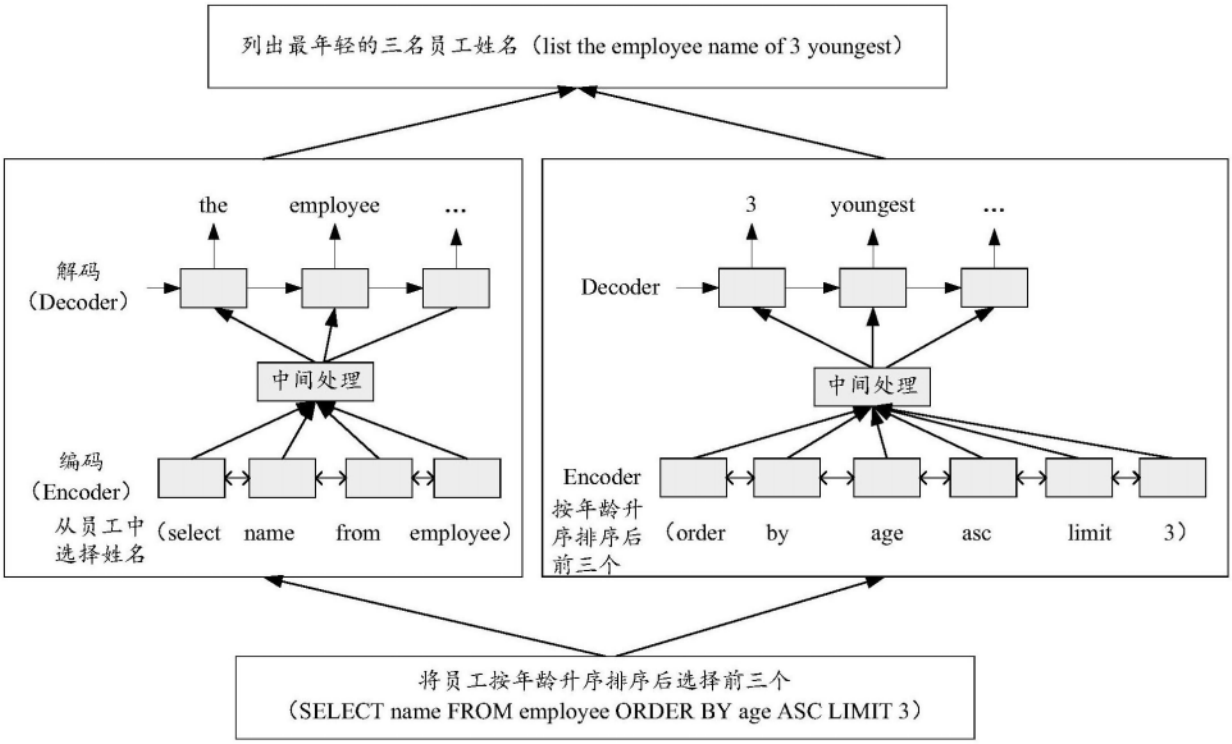

列出最年轻的三名员工姓名（list the employee name of 3 youngest）
the
employee
...
解码
（Decoder）
中间处理
编码
（Encoder）
从员工中选择姓名
（select name from employee）
3
youngest
...
Decoder
中间处理
Encoder
按年龄升序排序后前三个
（order by age asc limit 3）
将员工按年龄升序排序后选择前三个
（SELECT name FROM employee ORDER BY age ASC LIMIT 3）

图3

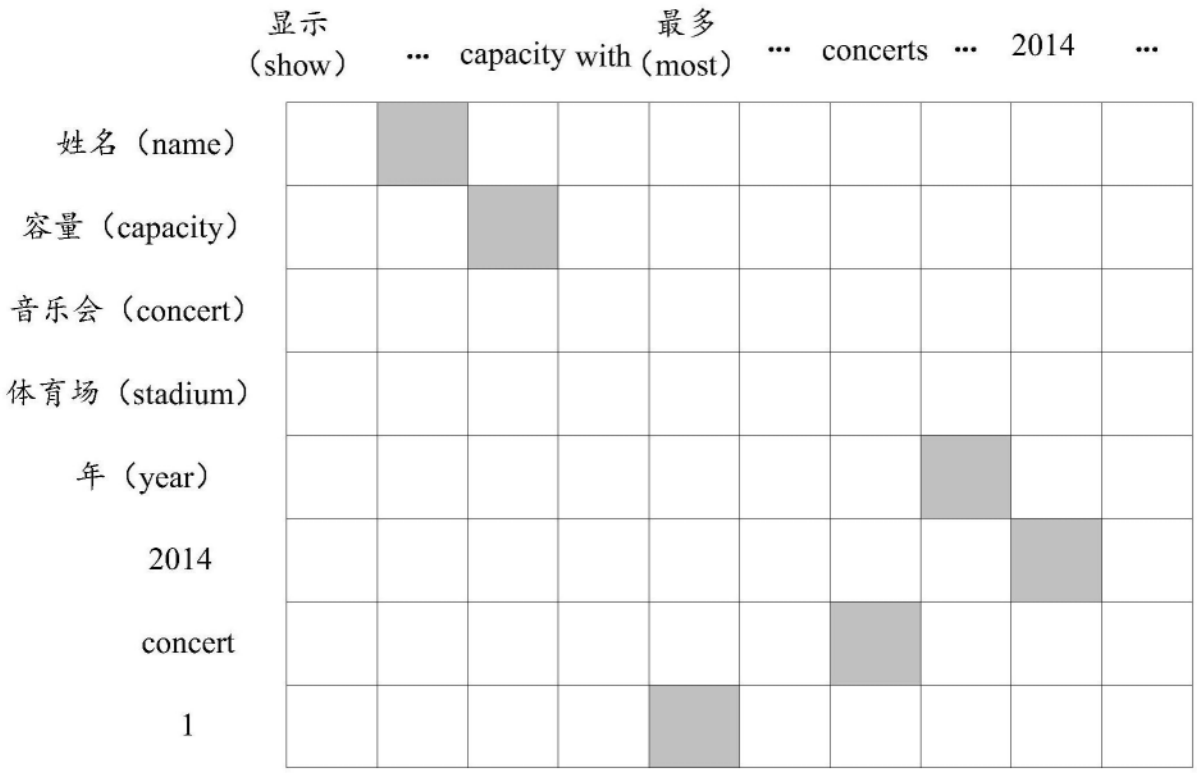

显示
（show）
...
capacity
with
最多
（most）
...
concerts
...
2014
...
姓名（name）
容量（capacity）
音乐会（concert）
体育场（stadium）
年（year）
2014
concert
1

图4

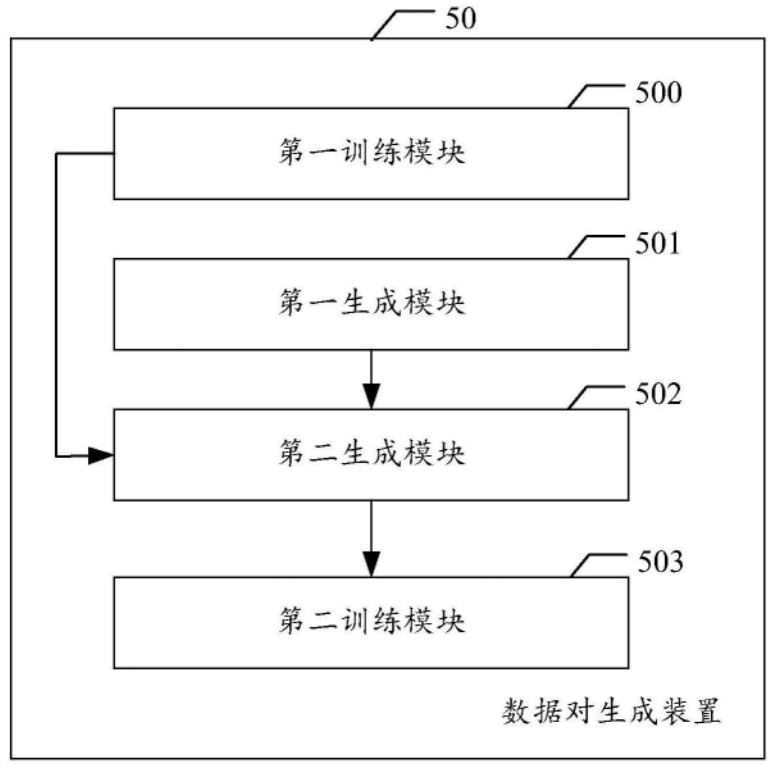

50
500
第一训练模块
501
第一生成模块
502
第二生成模块
503
第二训练模块
数据对生成装置

图5

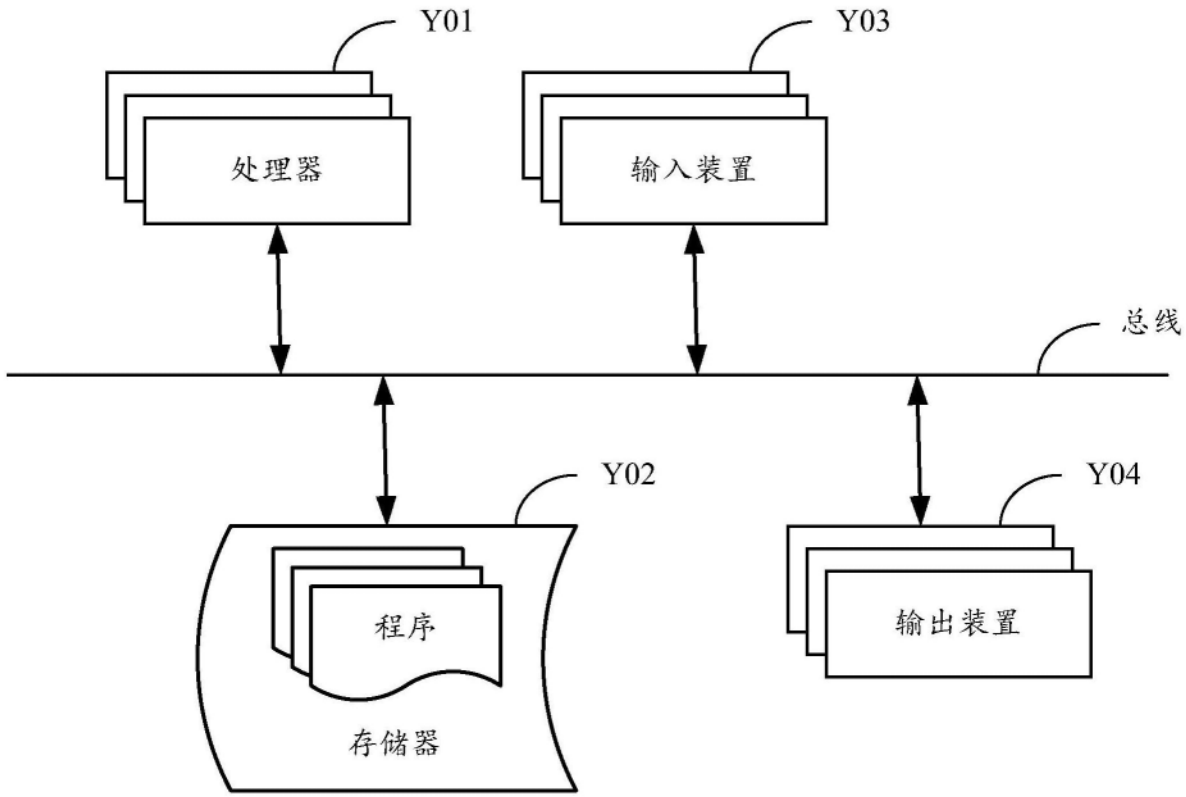

Y01
处理器
Y03
输入装置
总线
Y02
程序
存储器
Y04
输出装置

图6