(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115617841 A
(43)申请公布日 2023.01.17

(21)申请号 202211403817.4

(22)申请日 2022.11.10

(71)申请人 北京商银微芯科技有限公司
地址 100055 北京市西城区朗琴国际A座17层

(72)发明人 卢瑶

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 高勇

(51)Int.Cl.
G06F 16/2452(2019.01)
G06F 16/242(2019.01)
G06F 40/253(2020.01)
G06F 40/289(2020.01)

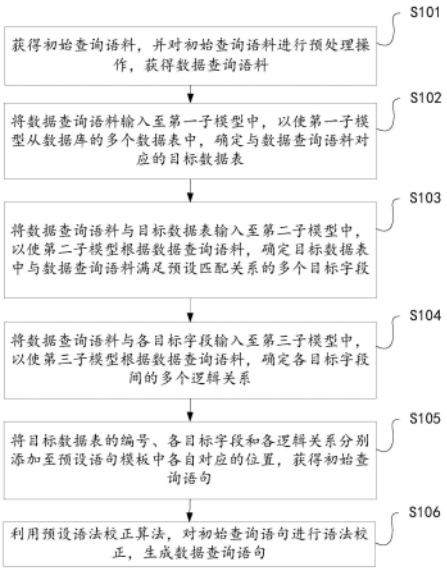
权利要求书3页 说明书13页 附图2页

(54)发明名称

一种数据查询语句的生成方法、系统、设备及存储介质

(57)摘要

本发明实施例提供了一种数据查询语句的生成方法、系统、设备及存储介质，其中，方法包括：对获得的初始查询语料进行预处理操作，获得数据查询语料，将数据查询语料输入至预设语句生成模型中，确定与数据查询语料对应的目标数据表，将数据查询语料与目标数据表输入至预设语句生成模型中，确定目标数据表中与数据查询语料满足预设匹配关系的多个目标字段，将数据查询语料与各目标字段输入至预设语句生成模型中，确定各目标字段间的多个逻辑关系，根据预设语句生成模型的输出结果对预设语句模板进行添加，获得初始查询语句，利用预设语法校正算法，对初始查询语句进行语法校正，生成数据查询语句。本发明提高了对数据的查询精度和效率。



CN 115617841 A

1.一种数据查询语句的生成方法,其特征在于,所述方法包括:

获得初始查询语料,并对所述初始查询语料进行预处理操作,获得数据查询语料,其中,所述初始查询语料是自然语言;

将所述数据查询语料输入至第一子模型中,以使所述第一子模型从数据库的多个数据表中,确定与所述数据查询语料对应的目标数据表,所述第一子模型是预设语句生成模型中的一个子模型;

将所述数据查询语料与所述目标数据表输入至第二子模型中,以使所述第二子模型根据所述数据查询语料,确定所述目标数据表中与所述数据查询语料满足预设匹配关系的多个目标字段,所述第二子模型是所述预设语句生成模型中的一个子模型;

将所述数据查询语料与各所述目标字段输入至第三子模型中,以使所述第三子模型根据所述数据查询语料,确定各所述目标字段间的多个逻辑关系,所述第三子模型是所述预设语句生成模型中的一个子模型;

将所述目标数据表的编号、各所述目标字段和各所述逻辑关系分别添加至预设语句模板中各自对应的位置,获得初始查询语句,其中,所述初始查询语句是基于程序语言编辑的语句;

利用预设语法校正算法,对所述初始查询语句进行语法校正,生成数据查询语句。

2.根据权利要求1所述的方法,其特征在于,所述预设语句生成模型的训练过程,包括:

获取多个初始训练样本数据,其中,所述初始训练样本数据由样本语料及其对应的样本数据表组成;

对各所述初始训练样本数据:对所述样本语料进行分词操作,并分别确定各分词结果与所述样本数据表中各字段的对应关系;基于所述样本语料、所述样本数据表的编号和各所述对应关系,生成与该初始训练样本数据对应的训练样本数据;

利用各所述训练样本数据,分别对初始第一子模型、初始第二子模型和初始第三子模型进行训练,获得由所述第一子模型、所述第二子模型和所述第三子模型组成的所述预设语句生成模型,所述预设语句生成模型的输入是所述数据查询语料,输出是所述目标数据表、各所述目标字段和各所述逻辑关系。

3.根据权利要求1所述的方法,其特征在于,所述将所述数据查询语料输入至第一子模型中,以使所述第一子模型从数据库的多个数据表中,确定与所述数据查询语料对应的目标数据表,包括:

将所述数据查询语料与数据库中的各数据表名称分别进行拼接,获得满足所述第一子模型输入格式的多个第一输入数据;

将各所述第一输入数据分别输入至所述第一子模型中,以使所述第一子模型计算各所述第一输入数据中,所述数据查询语料与所述数据表的匹配度,并将数值最大的所述匹配度对应的所述数据表,确定为所述目标数据表。

4.根据权利要求3所述的方法,其特征在于,所述将所述数据查询语料与所述目标数据表输入至第二子模型中,以使所述第二子模型根据所述数据查询语料,确定所述目标数据表中与所述数据查询语料满足预设匹配关系的多个目标字段,包括:

提取所述目标数据表的各字段名称;

将所述数据查询语料与各所述字段名称进行拼接,获得满足所述第二子模型输入格式

的第二输入数据；

将所述第二输入数据输入至所述第二子模型，以使所述第二子模型分别确定所述第二输入数据中，各所述字段名称与所述数据查询语料是否满足预设匹配关系组中的至少一个匹配关系，若是，则将满足至少一个所述匹配关系的所述字段名称对应的字段确定为所述目标字段，并将所述匹配关系作为标签添加至所述目标字段。

5.根据权利要求4所述的方法，其特征在于，所述将所述数据查询语料与各所述目标字段输入至第三子模型中，以使所述第三子模型根据所述数据查询语料，确定各所述目标字段间的多个逻辑关系，包括：

从所述数据查询语料中提取数据类型为关键词的至少一个关键词字符；

对各所述目标字段：判断该目标字段满足的至少一个所述匹配关系中，是否包括内容为该目标字段的数据类型为条件字段的所述匹配关系，若是，则将该目标字段确定为条件字段；

获得与各所述条件字段的字段类型分别对应的多个逻辑符号标识；

根据各所述字段的字段内容，分别确定各所述字段名称与各预设字段类型的对应关系，并基于所述对应关系，构建多个三元数据组；

将各所述三元数据组输入至所述第三子模型中，以使所述第三子模型对各所述三元数据组进行二分类，并基于分类结果，将匹配度大于预设阈值的至少一个所述三元数据组确定为目标三元数据组，其中，所述三元数据组中包括一个所述关键词字符，一个所述条件字段及其对应一个所述逻辑符号标识；

将所述目标三元数据组中的所述逻辑符号表示确定为所述逻辑关系。

6.根据权利要求5所述的方法，其特征在于，所述将所述目标数据表的编号、各所述目标字段和各所述逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，包括：

获得各所述目标字段的映射标识和所述目标数据表的编号；

将所述目标数据表的编号添加至预设语句模板中的数据表查表位置；

分别将各所述字段名分别添加至所述预设语句模板中各自对应的位置，其中，所述位置与所述映射标识具有对应关系；

将所述逻辑关系对应的所述逻辑符号添加至预设语句模板中的逻辑符号位置；

获得所述初始查询语句。

7.根据权利要求1所述的方法，其特征在于，所述对所述初始查询语料进行预处理操作，获得数据查询语料，包括：

利用预设正则匹配算法，查找所述初始查询语料中表征日期和数字的文字类型字符，并将所述文字类型字符转换为数字类型字符。

8.一种数据查询语句的生成系统，其特征在于，所述系统包括：

语料处理模块，用于获得初始查询语料，并对所述初始查询语料进行预处理操作，获得数据查询语料，其中，所述初始查询语料是自然语言；

第一数据确定模块，用于将所述数据查询语料输入至第一子模型中，以使所述第一子模型从数据库的多个数据表中，确定与所述数据查询语料对应的目标数据表，所述第一子模型是预设语句生成模型中的一个子模型；

第二数据确定模块，用于将所述数据查询语料与所述目标数据表输入至第二子模型中，以使所述第二子模型根据所述数据查询语料，确定所述目标数据表中与所述数据查询语料满足预设匹配关系的多个目标字段，所述第二子模型是所述预设语句生成模型中的一个子模型；

第三数据确定模块，用于将所述数据查询语料与各所述目标字段输入至第三子模型中，以使所述第三子模型根据所述数据查询语料，确定各所述目标字段间的多个逻辑关系，所述第三子模型是所述预设语句生成模型中的一个子模型；

数据填充模块，用于将所述目标数据表的编号、各所述目标字段和各所述逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，其中，所述初始查询语句是基于程序语言编辑的语句；

语句生成模块，用于利用预设语法校正算法，对所述初始查询语句进行语法校正，生成数据查询语句。

9.一种数据查询语句的生成设备，其特征在于，所述生成设备包括：

处理器；

用于存储所述处理器可执行指令的存储器；

其中，所述处理器被配置为执行所述指令，以实现如权利要求1至7中任一项所述的数据查询语句的生成方法。

10.一种计算机可读存储介质，其特征在于，当所述计算机可读存储介质中的指令由数据查询语句的生成设备的处理器执行时，使得所述生成设备能够执行如权利要求1至7中任一项所述的数据查询语句的生成方法。

一种数据查询语句的生成方法、系统、设备及存储介质

技术领域

[0001] 本发明涉及数据处理技术领域，特别是涉及一种数据查询语句的生成方法、系统、设备及存储介质。

背景技术

[0002] 数据库是依据数据结构来组织、存储和管理数据的仓库。伴随着互联网技术的发展，数据库中存储的数据表结构也愈发复杂。现有从数据库中查询数据的方式，是需要用户根据待查询数据梳理数据查询需求，并基于数据查询需求构建结构化查询语言(Structured Query Language，SQL)，从而完成数据查询。

[0003] 但是，上述基于数据查询需求构建SQL语句时，需要由掌握SQL语法结构的操作人员进行人工构建，这使得对数据进行查询的效率降低。同时，对于不了解SQL语法结构操作人员，进行SQL语句构建的难度过高，且会导致对数据进行查询的精度和效率降低。因此，如何提高对数据查询的精度和效率已成为亟待解决的问题。

发明内容

[0004] 本发明实施例的目的在于提供一种数据查询语句的生成方法、系统、设备及存储介质，以实现自动生成满足SQL语法结构的数据查询语句，提高对数据的查询精度和效率。具体技术方案如下：

[0005] 一种数据查询语句的生成方法，所述方法包括：

[0006] 获得初始查询语料，并对所述初始查询语料进行预处理操作，获得数据查询语料，其中，所述初始查询语料是自然语言；

[0007] 将所述数据查询语料输入至第一子模型中，以使所述第一子模型从数据库的多个数据表中，确定与所述数据查询语料对应的目标数据表，所述第一子模型是预设语句生成模型中的一个子模型；

[0008] 将所述数据查询语料与所述目标数据表输入至第二子模型中，以使所述第二子模型根据所述数据查询语料，确定所述目标数据表中与所述数据查询语料满足预设匹配关系的多个目标字段，所述第二子模型是所述预设语句生成模型中的一个子模型；

[0009] 将所述数据查询语料与各所述目标字段输入至第三子模型中，以使所述第三子模型根据所述数据查询语料，确定各所述目标字段间的多个逻辑关系，所述第三子模型是所述预设语句生成模型中的一个子模型；

[0010] 将所述目标数据表的编号、各所述目标字段和各所述逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，其中，所述初始查询语句是基于程序语言编辑的语句；

[0011] 利用预设语法校正算法，对所述初始查询语句进行语法校正，生成数据查询语句。

[0012] 可选的，所述预设语句生成模型的训练过程，包括：

[0013] 获取多个初始训练样本数据，其中，所述初始训练样本数据由样本语料及其对应

的样本数据表组成；

[0014] 对各所述初始训练样本数据：对所述样本语料进行分词操作，并分别确定各分词结果与所述样本数据表中各字段的对应关系；基于所述样本语料、所述样本数据表的编号和各所述对应关系，生成与该初始训练样本数据对应的训练样本数据；

[0015] 利用各所述训练样本数据，分别对初始第一子模型、初始第二子模型和初始第三子模型进行训练，获得由所述第一子模型、所述第二子模型和所述第三子模型组成的所述预设语句生成模型，所述预设语句生成模型的输入是所述数据查询语料，输出是所述目标数据表、各所述目标字段和各所述逻辑关系。

[0016] 可选的，所述将所述数据查询语料输入至第一子模型中，以使所述第一子模型从数据库的多个数据表中，确定与所述数据查询语料对应的目标数据表，包括：

[0017] 将所述数据查询语料与数据库中的各数据表名称分别进行拼接，获得满足所述第一子模型输入格式的多个第一输入数据；

[0018] 将各所述第一输入数据分别输入至所述第一子模型中，以使所述第一子模型计算各所述第一输入数据中，所述数据查询语料与所述数据表的匹配度，并将数值最大的所述匹配度对应的所述数据表，确定为所述目标数据表。

[0019] 可选的，所述将所述数据查询语料与所述目标数据表输入至第二子模型中，以使所述第二子模型根据所述数据查询语料，确定所述目标数据表中与所述数据查询语料满足预设匹配关系的多个目标字段，包括：

[0020] 提取所述目标数据表的各字段名称；

[0021] 将所述数据查询语料与各所述字段名称进行拼接，获得满足所述第二子模型输入格式的第二输入数据；

[0022] 将所述第二输入数据输入至所述第二子模型，以使所述第二子模型分别确定所述第二输入数据中，各所述字段名称与所述数据查询语料是否满足预设匹配关系组中的至少一个匹配关系，若是，则将满足至少一个所述匹配关系的所述字段名称对应的字段确定为所述目标字段，并将所述匹配关系作为标签添加至所述目标字段。

[0023] 可选的，所述将所述数据查询语料与各所述目标字段输入至第三子模型中，以使所述第三子模型根据所述数据查询语料，确定各所述目标字段间的多个逻辑关系，包括：

[0024] 从所述数据查询语料中提取数据类型为关键词的至少一个关键词字符；

[0025] 对各所述目标字段：判断该目标字段满足的至少一个所述匹配关系中，是否包括内容为该目标字段的数据类型为条件字段的所述匹配关系，若是，则将该目标字段确定为条件字段；

[0026] 获得与各所述条件字段的字段类型分别对应的多个逻辑符号标识；

[0027] 根据各所述字段的字段内容，分别确定各所述字段名称与各预设字段类型的对应关系，并基于所述对应关系，构建多个三元数据组；

[0028] 将各所述三元数据组输入至所述第三子模型中，以使所述第三子模型对各所述三元数据组进行二分类，并基于分类结果，将匹配度大于预设阈值的至少一个所述三元数据组确定为目标三元数据组，其中，所述三元数据组中包括一个所述关键词字符，一个所述条件字段及其对应一个所述逻辑符号标识；

[0029] 将所述目标三元数据组中的所述逻辑符号表示确定为所述逻辑关系。

[0030] 可选的，所述将所述目标数据表的编号、各所述目标字段和各所述逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，包括：

[0031] 获得各所述目标字段的映射标识和所述目标数据表的编号；

[0032] 将所述目标数据表的编号添加至预设语句模板中的数据表查表位置；

[0033] 分别将各所述字段名分别添加至所述预设语句模板中各自对应的位置，其中，所述位置与所述映射标识具有对应关系；

[0034] 将所述逻辑关系对应的所述逻辑符号添加至预设语句模板中的逻辑符号位置；

[0035] 获得所述初始查询语句。

[0036] 可选的，所述对所述初始查询语料进行预处理操作，获得数据查询语料，包括：

[0037] 利用预设正则匹配算法，查找所述初始查询语料中表征日期和数字的文字类型字符，并将所述文字类型字符转换为数字类型字符。

[0038] 一种数据查询语句的生成系统，所述系统包括：

[0039] 语料处理模块，用于获得初始查询语料，并对所述初始查询语料进行预处理操作，获得数据查询语料，其中，所述初始查询语料是自然语言；

[0040] 第一数据确定模块，用于将所述数据查询语料输入至第一子模型中，以使所述第一子模型从数据库的多个数据表中，确定与所述数据查询语料对应的目标数据表，所述第一子模型是预设语句生成模型中的一个子模型；

[0041] 第二数据确定模块，用于将所述数据查询语料与所述目标数据表输入至第二子模型中，以使所述第二子模型根据所述数据查询语料，确定所述目标数据表中与所述数据查询语料满足预设匹配关系的多个目标字段，所述第二子模型是所述预设语句生成模型中的一个子模型；

[0042] 第三数据确定模块，用于将所述数据查询语料与各所述目标字段输入至第三子模型中，以使所述第三子模型根据所述数据查询语料，确定各所述目标字段间的多个逻辑关系，所述第三子模型是所述预设语句生成模型中的一个子模型；

[0043] 数据填充模块，用于将所述目标数据表的编号、各所述目标字段和各所述逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，其中，所述初始查询语句是基于程序语言编辑的语句；

[0044] 语句生成模块，用于利用预设语法校正算法，对所述初始查询语句进行语法校正，生成数据查询语句。

[0045] 可选的，所述系统还配置有模型训练模块，所述模型训练模块在对所述预设语句生成模型进行训练时被设置为：

[0046] 获取多个初始训练样本数据，其中，所述初始训练样本数据由样本语料及其对应的样本数据表组成；

[0047] 对各所述初始训练样本数据：对所述样本语料进行分词操作，并分别确定各分词结果与所述样本数据表中各字段的对应关系；基于所述样本语料、所述样本数据表的编号和各所述对应关系，生成与该初始训练样本数据对应的训练样本数据；

[0048] 利用各所述训练样本数据，分别对初始第一子模型、初始第二子模型和初始第三子模型进行训练，获得由所述第一子模型、所述第二子模型和所述第三子模型组成的所述预设语句生成模型，所述预设语句生成模型的输入是所述数据查询语料，输出是所述目标

数据表、各所述目标字段和各所述逻辑关系。

[0049] 可选的，所述第一数据确定模块被设置为：

[0050] 将所述数据查询语料与数据库中的各数据表名称分别进行拼接，获得满足所述第一子模型输入格式的多个第一输入数据；

[0051] 将各所述第一输入数据分别输入至所述第一子模型中，以使所述第一子模型计算各所述第一输入数据中，所述数据查询语料与所述数据表的匹配度，并将数值最大的所述匹配度对应的所述数据表，确定为所述目标数据表。

[0052] 可选的，所述第二数据确定模块被设置为：

[0053] 提取所述目标数据表的各字段名称；

[0054] 将所述数据查询语料与各所述字段名称进行拼接，获得满足所述第二子模型输入格式的第二输入数据；

[0055] 将所述第二输入数据输入至所述第二子模型，以使所述第二子模型分别确定所述第二输入数据中，各所述字段名称与所述数据查询语料是否满足预设匹配关系组中的至少一个匹配关系，若是，则将满足至少一个所述匹配关系的所述字段名称对应的字段确定为所述目标字段，并将所述匹配关系作为标签添加至所述目标字段。

[0056] 可选的，所述第三数据确定模块被设置为：

[0057] 从所述数据查询语料中提取数据类型为关键词的至少一个关键词字符；

[0058] 对各所述目标字段：判断该目标字段满足的至少一个所述匹配关系中，是否包括内容为该目标字段的数据类型为条件字段的所述匹配关系，若是，则将该目标字段确定为条件字段；

[0059] 获得与各所述条件字段的字段类型分别对应的多个逻辑符号标识；

[0060] 根据各所述字段的字段内容，分别确定各所述字段名称与各预设字段类型的对应关系，并基于所述对应关系，构建多个三元数据组；

[0061] 将各所述三元数据组输入至所述第三子模型中，以使所述第三子模型对各所述三元数据组进行二分类，并基于分类结果，将匹配度大于预设阈值的至少一个所述三元数据组确定为目标三元数据组，其中，所述三元数据组中包括一个所述关键词字符，一个所述条件字段及其对应一个所述逻辑符号标识；

[0062] 将所述目标三元数据组中的所述逻辑符号表示确定为所述逻辑关系。

[0063] 可选的，所述数据填充模块被设置为：

[0064] 获得各所述目标字段的映射标识和所述目标数据表的编号；

[0065] 将所述目标数据表的编号添加至预设语句模板中的数据表查表位置；

[0066] 分别将各所述字段名分别添加至所述预设语句模板中各自对应的位置，其中，所述位置与所述映射标识具有对应关系；

[0067] 将所述逻辑关系对应的所述逻辑符号添加至预设语句模板中的逻辑符号位置；

[0068] 获得所述初始查询语句。

[0069] 可选的，所述语料处理模块在对所述初始查询语料进行预处理操作，获得数据查询语料时被设置为：

[0070] 利用预设正则匹配算法，查找所述初始查询语料中表征日期和数字的文字类型字符，并将所述文字类型字符转换为数字类型字符。

[0071] 一种数据查询语句的生成设备，所述生成设备包括：

[0072] 处理器；

[0073] 用于存储所述处理器可执行指令的存储器；

[0074] 其中，所述处理器被配置为执行所述指令，以实现如上述任一种所述的数据查询语句的生成方法。

[0075] 一种计算机可读存储介质，当所述计算机可读存储介质中的指令由数据查询语句的生成设备的处理器执行时，使得所述生成设备能够执行如上述任一种所述的数据查询语句的生成方法。

[0076] 本发明实施例提供的一种数据查询语句的生成方法、系统、设备及存储介质，可以通过预处理操作对初始查询预料中进行处理，避免了由于口语化字符串无法被准确识别的风险。并通过设置预设语句生成模型中的第一子模型，实现了对数据查询语料中冗余数据和查询内容的区分，筛选出存储有查询内容的目标数据表。同时，又通过设置第二子模型和第三子模型，实现了对数据表中表征查询内容的数据和逻辑关系的准确提取，避免了由于存在其他冗余数据导致数据查询语句生成效率和精度降低的风险。最后，通过对构建的预设语句模板进行添加，并利用预设语法校正算法对获得的初始查询语句进行语法校正，提高了最终获得的数据查询语句的精度和效率。

[0077] 当然，实施本发明的任一产品或方法必不一定需要同时达到以上所述的所有优点。

附图说明

[0078] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，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。

[0079] 图1为本发明实施例提供的一种数据查询语句的生成方法的流程图；

[0080] 图2为本发明的一个可选实施例提供的一种数据查询语句的生成系统的框图；

[0081] 图3为本发明的另一个可选实施例提供的一种数据查询语句的生成设备的框图。

具体实施方式

[0082] 下面将结合本发明实施例中的附图，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。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动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。

[0083] 本发明实施例提供了一种数据查询语句的生成方法，如图1所示，该生成方法包括：

[0084] S101、获得初始查询语料，并对初始查询语料进行预处理操作，获得数据查询语料，其中，初始查询语料是自然语言。

[0085] 可选的，在本发明的一个可选实施例中，上述初始查询语料可以是用户通过人机交互界面输入的表征查询内容的自然语言的字符串，也可以是通过声音采集设备，并经过

语音识别算法转换后的自然语言的字符串。

[0086] 可选的，在本发明的另一个可选实施例中，上述预处理操作可以是对初始查询语料中的字符进行的类型转换操作。由于在实际应用场景下，上述初始查询语料的字符内容偏向口语化，例如用上半年表征一月到六月。不利于后续的模型处理。因此，本发明通过上述预处理操作对初始查询预料中进行处理，避免了由于口语化字符串无法被准确识别的风险，提高了后续模型的处理精度，从而提高数据查询语句的生成精度。

[0087] S102、将数据查询语料输入至第一子模型中，以使第一子模型从数据库的多个数据表中，确定与数据查询语料对应的目标数据表，第一子模型是预设语句生成模型中的一个子模型。

[0088] 可选的，在本发明的一个可选实施例中，上述预设语句生成模型可以是基于深度双向表示(BidirectionalEncoder Representation fromTransformers，BERT)模型构建的。由于现有的单向语言模型只能识别单方向的上下文特征信息，限制了对字符串的表征能力，导致现有的数据查询语句生成精度低。而本发明通过构建上述语句生成模型，可以提高对数据查询语料中上下文特征信息的利用率，从而提高最终生成的数据查询语句的精度。

[0089] 可选的，在本发明的另一个可选实施例中，上述目标数据表可以是存储有上述数据查询语料中查询内容的数据表。由于上述数据查询语料是用户编辑的，其结构不满足结构化查询语言(Structured Query Language，SQL)的语法结构，除包括上述查询内容外，还包括部分冗余数据。因此，若要提高最终生成的数据查询语句精度，需要明确上述数据查询语料中的查询内容，与数据库中存储的数据表间的对应关系。因此，本发明通过设置上述预设语句生成模型中的第一子模型，实现了对上述数据查询语料中冗余数据和查询内容的区分，并筛选出存储有上述查询内容的目标数据表，从而提高了最终生成的数据查询语句的精度和效率。

[0090] S103、将数据查询语料与目标数据表输入至第二子模型中，以使第二子模型根据数据查询语料，确定目标数据表中与数据查询语料满足预设匹配关系的多个目标字段，第二子模型是预设语句生成模型中的一个子模型。

[0091] 需要说明的是，在实际应用场景下，由于数据表除了包括上述数据查询语料中的查询内容数据外，还包括了其他数据。因此，本发明通过设置上述第二子模型，实现了对数据表中表征查询内容的数据的提取，从而避免了由于存在其他冗余数据导致数据查询语句生成效率和精度降低的风险。

[0092] 可选的，在本发明的一个可选实施例中，上述预设匹配关系可以是用于确定不同类型的查询内容与目标数据表中不同字段间对应关系的至少一个筛选条件。例如，若上述数据查询语料的查询内容是“甲电影的票房总和”，则上述预设匹配关系可以有两个，分别是第一预设匹配关系和第二预设匹配关系。其中，上述第一预设匹配关系为：当前字段是否为电影名称字段，且该字段内容是否为甲。上述第一预设匹配关系为：当前字段内容是否为日票房。

[0093] S104、将数据查询语料与各目标字段输入至第三子模型中，以使第三子模型根据数据查询语料，确定各目标字段间的多个逻辑关系，第三子模型是预设语句生成模型中的一个子模型。

[0094] 需要说明的是，由于在实际应用场景下，上述数据查询语料的查询内容，可以是对目标数据表中的多个字段内容进行运算或比较后产生的结果。因此，本发明通过上述第三子模型确定上述逻辑关系，可以明确数据查询语料中包含的运算符号或比较符号，从而提高最终生成的数据查询语句的精度。

[0095] S105、将目标数据表的编号、各目标字段和各逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，其中，初始查询语句是基于程序语言编辑的语句。

[0096] 需要说明的是，在实际应用场景下，基于SQL语言的数据查询语句是具有固定结构的，需要包括查找字段(含聚合函数)、目标数据表查找字段、查找条件字段和分组字段。因此，本发明通过构建包括上述四个字段的预设语句模板，并基于预设映射关系将获得的目标数据表的编号、目标字段的内容及逻辑关系进行自动添加。从而提高了生成的数据查询语句的精度和效率。其中，上述预设映射关系是指上述字段与预设语句模板中字段填充位置间的映射关系。

[0097] S106、利用预设语法校正算法，对初始查询语句进行语法校正，生成数据查询语句。

[0098] 需要说明的是，在实际用场景下，上述初始查询语句随意具备可读性。但是，由于上述初始查询语句中填充的字段是由数据表中提取的，导致初始查询语句不满足SQL执行语句的结构。因此，为了提高最终生成的数据查询语句的可执行性，还需要通过上述预设语法校正算法进行语法校正，以便用户可以根据数据查询语句直接查询数据库。其中，上述预设语法校正算法可以是基于SQL语言的语法结构构建的算法。其具体功能包括但不限于：在多个查询条件间添加连接符，如AND、OR等；加入时间转换函数，如TO-DATE函数；对时间范围类型的查询条件，进行边界校正；对模糊匹配字符串，添加Like函数。

[0099] 本发明通过预处理操作对初始查询预料中进行处理，避免了由于口语化字符串无法被准确识别的风险。并通过设置预设语句生成模型中的第一子模型，实现了对数据查询语料中冗余数据和查询内容的区分，筛选出存储有查询内容的目标数据表。同时，又通过设置第二子模型和第三子模型，实现了对数据表中表征查询内容的数据和逻辑关系的准确提取，避免了由于存在其他冗余数据导致数据查询语句生成效率和精度降低的风险。最后，通过对构建的预设语句模板进行添加，并利用预设语法校正算法对获得的初始查询语句进行语法校正，提高了最终获得的数据查询语句的精度和效率。

[0100] 可选的，预设语句生成模型的训练过程，包括：

[0101] 获取多个初始训练样本数据，其中，初始训练样本数据由样本语料及其对应的样本数据表组成；

[0102] 对各初始训练样本数据：对样本语料进行分词操作，并分别确定各分词结果与样本数据表中各字段的对应关系；基于样本语料、样本数据表的编号和各对应关系，生成与该初始训练样本数据对应的训练样本数据；

[0103] 利用各训练样本数据，分别对初始第一子模型、初始第二子模型和初始第三子模型进行训练，获得由第一子模型、第二子模型和第三子模型组成的预设语句生成模型，预设语句生成模型的输入是数据查询语料，输出是目标数据表、各目标字段和各逻辑关系。

[0104] 需要说明的是，在实际应用场景下，上述样本数据表可以是经过结构调整和类型分类调整后获得的。由于数据表中的表结构通常采用英文进行编辑的。若样本语料是语言

类型为汉语或其他类型的自然语言，则会导致初始语句生成模型无法识别。并且，由于数据表中的字段类型繁多，不利于提高模型效率。因此，上述样本数据表是表结构经过人工标注，并且将字段类型重新划分后获得数据表。

[0105] 其中，上述字段类型重新划分的实施方式包括：

[0106] 将数据表中的存储时间或日期的字段类型，统一为日期(Date)类型；将存储数据类型为整数型和浮点型的字段类型，统一为数值(Number)类型；将数据表中除上述日期(Date)类型和数值(Number)类型外的其他字段类型，统一为字符串(Text)类型。

[0107] 需要说明的是，在实际应用场景下，上述基于样本语料、样本数据表的编号和各对应关系，生成与该初始训练样本数据对应的训练样本数据的实施方式有多种，在此示例性地提供一种：

[0108] 设定当前应用场景下的样本语料为“甲影片的总票房是多少”。上述样本数据表包括电影名称、日票房、放映日期、放映影院和电影时长五个字段。其中，电影名称字段的编号为1，日票房字段的编号为2，放映日期字段的编号为3，放映影院字段的编号为4，电影时长字段的编号为5。

[0109] 获得训练样本模板。该训练样本模板由多个字段构成，包括：语料(Question)字段、表编号(table_id)字段、查找(select)字段、聚合函数(agg)字段、查找条件(conds)字段和分组(group)字段。

[0110] 则在获得上述样本语料、样本数据表的编号和各对应关系后，将样本语料添加至上述语料字段，将样本数据表的编号添加至上述表编号字段。

[0111] 根据上述对应关系可知，样本语料中的“甲”对应电影名称字段，“票房”对应日票房字段。则将电影名称字段的编号1和日票房字段的编号2添加至查找字段。将电影名称字段的编号1添加至分组字段。

[0112] 根据上述对应关系可知，样本语料中的“总”对应求和函数，且求和对象是日票房字段。对电影名称字段或运算。则将无运算标识和求和函数标识添加至函数字段。

[0113] 根据上述样本语料可确定，查找条件为“电影名称＝甲”。则程序语言为将“甲”赋值给“电影名称”。因而，将电影名称字段的编号1、赋值函数标识和“甲”字符依次添加至查找条件字段。

[0114] 需要说明的是，在实际应用场景下，若查找条件有多个，可以通过逻辑连接符将多个查找条件进行拼接。例如，查找条件1AND查找条件2。

[0115] 根据上述样本语料可确定，分组内容是根据电影名称设定的。因此，将电影名称字段的编号1添加至分组(group)字段。

[0116] 将经过上述添加步骤后的训练样本模板经过编译后，确定为训练样本数据。

[0117] 需要说明的是，在实际应用场景下，由于不同公司具有不同的业务方向，其针对的业务场景相对固定。且受限于公司规模导致样本规模不同。因此为了提高训练质量和普适性，可以通过预训练(pre-train)和微调(fine-tune)方式对上述模型进行训练。具体的，利用上述各样本数据预训练方式对初始语句BERT模型进行训练。在预训练完成的情况下，再利用上述各样本数据通过微调方式，分别进行上述初始第一模型、初始第二子模型和初始第三子模型的训练。

[0118] 本领域技术人员可以理解的是，在实际应用场景下，语料中会包含较多的专业术

语。因此，为了提高模型对语料的识别精度，可以通过构建业务词典的方式，将特定业务场景下的专业词汇，如产品类型、对应代码、业务流转状态标志码等，整理至业务词典中，以便模型进行调用。本发明对上述业务词典的具体构建过程及模型对业务词典的调用过程不作过多限定和赘述。

[0119] 可选的，将数据查询语料输入至第一子模型中，以使第一子模型从数据库的多个数据表中，确定与数据查询语料对应的目标数据表，包括：

[0120] 将数据查询语料与数据库中的各数据表名称分别进行拼接，获得满足第一子模型输入格式的多个第一输入数据；

[0121] 将各第一输入数据分别输入至第一子模型中，以使第一子模型计算各第一输入数据中，数据查询语料与数据表的匹配度，并将数值最大的匹配度对应的数据表，确定为目标数据表。

[0122] 可选的，在本发明的一个可选实施例中，上述在将数据查询语料与数据库中的各数据表名称分别进行拼接时，其中的各数据表可以是上述数据库中的每一个数据表，也可以是经过筛选后的某一类型的多个数据表。例如：A部门在生成数据查询语句时，仅对数据库中部门标识符为A的数据表进行筛选。或根据数据查询语句对应的业务场景的业务标识，查找数据库中存在该业务标识的数据表并进行筛选。

[0123] 可选的，在本发明的另一个可选实施例中，上述计算数据查询语料与数据表的匹配度，可以是通过计算数据查询语料中的几个连续字符构成的字符串，与数据表名称几个连续字符构成的字符串间的匹配度来实现的。因此，在计算匹配度前，需要上述第一子模型需要对数据查询语料和数据表名称进行分词。

[0124] 本领域技术人员可以理解的是，在实际应用场景下，上述分词可以通过BERT模型自带的分词器(BasicTokenizer)来实现。本发明对利用上述分词器进分词的具体实施方式不作过多赘述和限制。

[0125] 可选的，将数据查询语料与目标数据表输入至第二子模型中，以使第二子模型根据数据查询语料，确定目标数据表中与数据查询语料满足预设匹配关系的多个目标字段，包括：

[0126] 提取目标数据表的各字段名称；

[0127] 将数据查询语料与各字段名称进行拼接，获得满足第二子模型输入格式的第二输入数据；

[0128] 将第二输入数据输入至第二子模型，以使第二子模型分别确定第二输入数据中，各字段名称与数据查询语料是否满足预设匹配关系组中的至少一个匹配关系，若是，则将满足至少一个匹配关系的字段名称对应的字段确定为目标字段，并将匹配关系作为标签添加至目标字段。

[0129] 可选的，在本发明的一个可选实施例中，上述以使第二子模型确定各第二输入数据中，各字段名称与数据查询语料是否满足预设匹配关系组中的至少一个匹配关系的具体实施方式，可以是：

[0130] 将上述第二子模型配置为以序列标注的方式，将字段名称与数据查询语料进行比对和预测，并判断该字段名称是否满足预设匹配关系组中的至少一个匹配关系。其中，上述预设匹配关系组可以包括内容不同多个匹配关系，例如：根据数据查询语料，判断该字段名

称是否被选中，若是，则将该字段对应的数据类型确定为查找字段；判断该字段名称是否是分组字段的内容，若是，则将该字段名称对应的数据类型确定为分组字段；判断该字段名称是否是查找条件中的字段，若是则将该字段名称对应的数据类型确定为条件字段。

[0131] 可选的，将数据查询语料与各目标字段输入至第三子模型中，以使第三子模型根据数据查询语料，确定各目标字段间的多个逻辑关系，包括：

[0132] 从数据查询语料中提取数据类型为关键词的至少一个关键词字符；

[0133] 对各目标字段：判断该目标字段满足的至少一个匹配关系中，是否包括内容为该目标字段的数据类型为条件字段的匹配关系，若是，则将该目标字段确定为条件字段；

[0134] 获得与各条件字段的字段类型分别对应的多个逻辑符号标识；

[0135] 根据各字段的字段内容，分别确定各字段名称与各预设字段类型的对应关系，并基于对应关系，构建多个三元数据组；

[0136] 将各三元数据组输入至第三子模型中，以使第三子模型对各三元数据组进行二分类，并基于分类结果，将匹配度大于预设阈值的至少一个三元数据组确定为目标三元数据组，其中，三元数据组中包括一个关键词字符，一个条件字段及其对应一个逻辑符号标识；

[0137] 将目标三元数据组中的逻辑符号表示确定为逻辑关系。

[0138] 可选的，在本发明的一个可选实施例中，上述关键词字符可以是数据查询语料中用于构建查找条件字段的字符串。例如：数据查询语料为：A股份有限公司今年全部的融资放款明细。其中，“A股份有限公司”就是上述数据类型为关键词的关键词字符。

[0139] 需要说明的是，在实际应用场景下，由于查找条件中的内容通常以企业名称、产品类型、形态代码等专有名词的形式出现。为了提高对条件字段的确定精度，本发明可以通过命名实体识别(Name Entity Recognition，NER)技术，结合上述业务词典和正则匹配算法实现确定。

[0140] 可选的，在本发明的另一个可选实施例中，上述预设字段类型可以是用于避免三元数据组中错误选取逻辑符号所构建的。例如，假设当前两个字段名称分别是日期和票房，且这两个字段名称对应的字段的字段内容，其数据类型均是整数型。显然，日期与票房并不存在逻辑关系。但是由于两个字段内容的数据类型均是整数型，则易出现由于两者具有相同数据类型，导致模型建立错误逻辑关系的风险。因此，本发明通过构建预设字段类型，并根据字段内容确定字段名称与预设字段类型的对应关系，实现避免错误逻辑关系的构建。

[0141] 其中，上述预设字段类型与上述本发明的一个可选实施例中提供的字段类型重新划分的划分结果相同。即上述各预设字段类型分别是日期(Date)类型、数值(Number)类型和字符串(Text)类型。

[0142] 可选的，将目标数据表的编号、各目标字段和各逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，包括：

[0143] 获得各目标字段的映射标识和目标数据表的编号；

[0144] 将目标数据表的编号添加至预设语句模板中的数据表查表位置；

[0145] 分别将各字段名分别添加至预设语句模板中各自对应的位置，其中，位置与映射标识具有对应关系；

[0146] 将逻辑关系对应的逻辑符号添加至预设语句模板中的逻辑符号位置；

[0147] 获得初始查询语句。

[0148] 可选的，对初始查询语料进行预处理操作，获得数据查询语料，包括：

[0149] 利用预设正则匹配算法，查找初始查询语料中表征日期和数字的文字类型字符，并将文字类型字符转换为数字类型字符。

[0150] 与上述方法实施例相对应地，本发明还提供了一种数据查询语句的生成系统，如图2所示，该生成系统包括：

[0151] 语料处理模块201，用于获得初始查询语料，并对初始查询语料进行预处理操作，获得数据查询语料，其中，初始查询语料是自然语言；

[0152] 第一数据确定模块202，用于将数据查询语料输入至第一子模型中，以使第一子模型从数据库的多个数据表中，确定与数据查询语料对应的目标数据表，第一子模型是预设语句生成模型中的一个子模型；

[0153] 第二数据确定模块203，用于将数据查询语料与目标数据表输入至第二子模型中，以使第二子模型根据数据查询语料，确定目标数据表中与数据查询语料满足预设匹配关系的多个目标字段，第二子模型是预设语句生成模型中的一个子模型；

[0154] 第三数据确定模块204，用于将数据查询语料与各目标字段输入至第三子模型中，以使第三子模型根据数据查询语料，确定各目标字段间的多个逻辑关系，第三子模型是预设语句生成模型中的一个子模型；

[0155] 数据填充模块205，用于将目标数据表的编号、各目标字段和各逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句，其中，初始查询语句是基于程序语言编辑的语句；

[0156] 语句生成模块206，用于利用预设语法校正算法，对初始查询语句进行语法校正，生成数据查询语句。

[0157] 可选的，上述如图2所示的生成系统还配置有模型训练模块，该模型训练模块在对预设语句生成模型进行训练时被设置为：

[0158] 获取多个初始训练样本数据，其中，初始训练样本数据由样本语料及其对应的样本数据表组成；

[0159] 对各初始训练样本数据：对样本语料进行分词操作，并分别确定各分词结果与样本数据表中各字段的对应关系；基于样本语料、样本数据表的编号和各对应关系，生成与该初始训练样本数据对应的训练样本数据；

[0160] 利用各训练样本数据，分别对初始第一子模型、初始第二子模型和初始第三子模型进行训练，获得由第一子模型、第二子模型和第三子模型组成的预设语句生成模型，预设语句生成模型的输入是数据查询语料，输出是目标数据表、各目标字段和各逻辑关系。

[0161] 可选的，上述第一数据确定模块202被设置为：

[0162] 将数据查询语料与数据库中的各数据表名称分别进行拼接，获得满足第一子模型输入格式的多个第一输入数据；

[0163] 将各第一输入数据分别输入至第一子模型中，以使第一子模型计算各第一输入数据中，数据查询语料与数据表的匹配度，并将数值最大的匹配度对应的数据表，确定为目标数据表。

[0164] 可选的，上述第二数据确定模块203被设置为：

[0165] 提取目标数据表的各字段名称；

[0166] 将数据查询语料与各字段名称进行拼接，获得满足第二子模型输入格式的第二输入数据；

[0167] 将第二输入数据输入至第二子模型，以使第二子模型分别确定第二输入数据中，各字段名称与数据查询语料是否满足预设匹配关系组中的至少一个匹配关系，若是，则将满足至少一个匹配关系的字段名称对应的字段确定为目标字段，并将匹配关系作为标签添加至目标字段。

[0168] 可选的，上述第三数据确定模块204被设置为：

[0169] 从数据查询语料中提取数据类型为关键词的至少一个关键词字符；

[0170] 对各目标字段：判断该目标字段满足的至少一个匹配关系中，是否包括内容为该目标字段的数据类型为条件字段的匹配关系，若是，则将该目标字段确定为条件字段；

[0171] 获得与各条件字段的字段类型分别对应的多个逻辑符号标识；

[0172] 根据各字段的字段内容，分别确定各字段名称与各预设字段类型的对应关系，并基于对应关系，构建多个三元数据组；

[0173] 将各三元数据组输入至第三子模型中，以使第三子模型对各三元数据组进行二分类，并基于分类结果，将匹配度大于预设阈值的至少一个三元数据组确定为目标三元数据组，其中，三元数据组中包括一个关键词字符，一个条件字段及其对应一个逻辑符号标识；

[0174] 将目标三元数据组中的逻辑符号表示确定为逻辑关系。

[0175] 可选的，上述数据填充模块205被设置为：

[0176] 获得各目标字段的映射标识和目标数据表的编号；

[0177] 将目标数据表的编号添加至预设语句模板中的数据表查表位置；

[0178] 分别将各字段名分别添加至预设语句模板中各自对应的位置，其中，位置与映射标识具有对应关系；

[0179] 将逻辑关系对应的逻辑符号添加至预设语句模板中的逻辑符号位置；

[0180] 获得初始查询语句。

[0181] 可选的，上述语料处理模块201在对初始查询语料进行预处理操作，获得数据查询语料时被设置为：

[0182] 利用预设正则匹配算法，查找初始查询语料中表征日期和数字的文字类型字符，并将文字类型字符转换为数字类型字符。

[0183] 本发明实施例还提供了一种数据查询语句的生成设备，如图3所示，该生成设备包括：

[0184] 处理器301；

[0185] 用于存储处理器301可执行指令的存储器302；

[0186] 其中，处理器302被配置为执行指令，以实现如上述任一种的数据查询语句的生成方法。

[0187] 本发明实施例还提供了一种计算机可读存储介质，当计算机可读存储介质中的指令由数据查询语句的生成设备的处理器执行时，使得生成设备能够执行如上述任一种的数据查询语句的生成方法。

[0188] 在一个典型的配置中，设备包括一个或多个处理器(CPU)、存储器和总线。设备还可以包括输入/输出接口、网络接口等。

[0189] 存储器可能包括计算机可读介质中的非永久性存储器,随机存取存储器(RAM)和/或非易失性内存等形式,如只读存储器(ROM)或闪存(flash RAM),存储器包括至少一个存储芯片。存储器是计算机可读介质的示例。

[0190] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。按照本文中的界定,计算机可读介质不包括暂存电脑可读媒体(transitory media),如调制的数据信号和载波。

[0191] 本领域技术人员应明白,本申请的实施例可提供为方法、系统或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0192] 需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。还需要说明的是,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0193] 本说明书中的各个实施例均采用相关的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0194] 以上仅为本申请的实施例而已,并不用于限制本申请。对于本领域技术人员来说,本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本申请的权利要求范围之内。

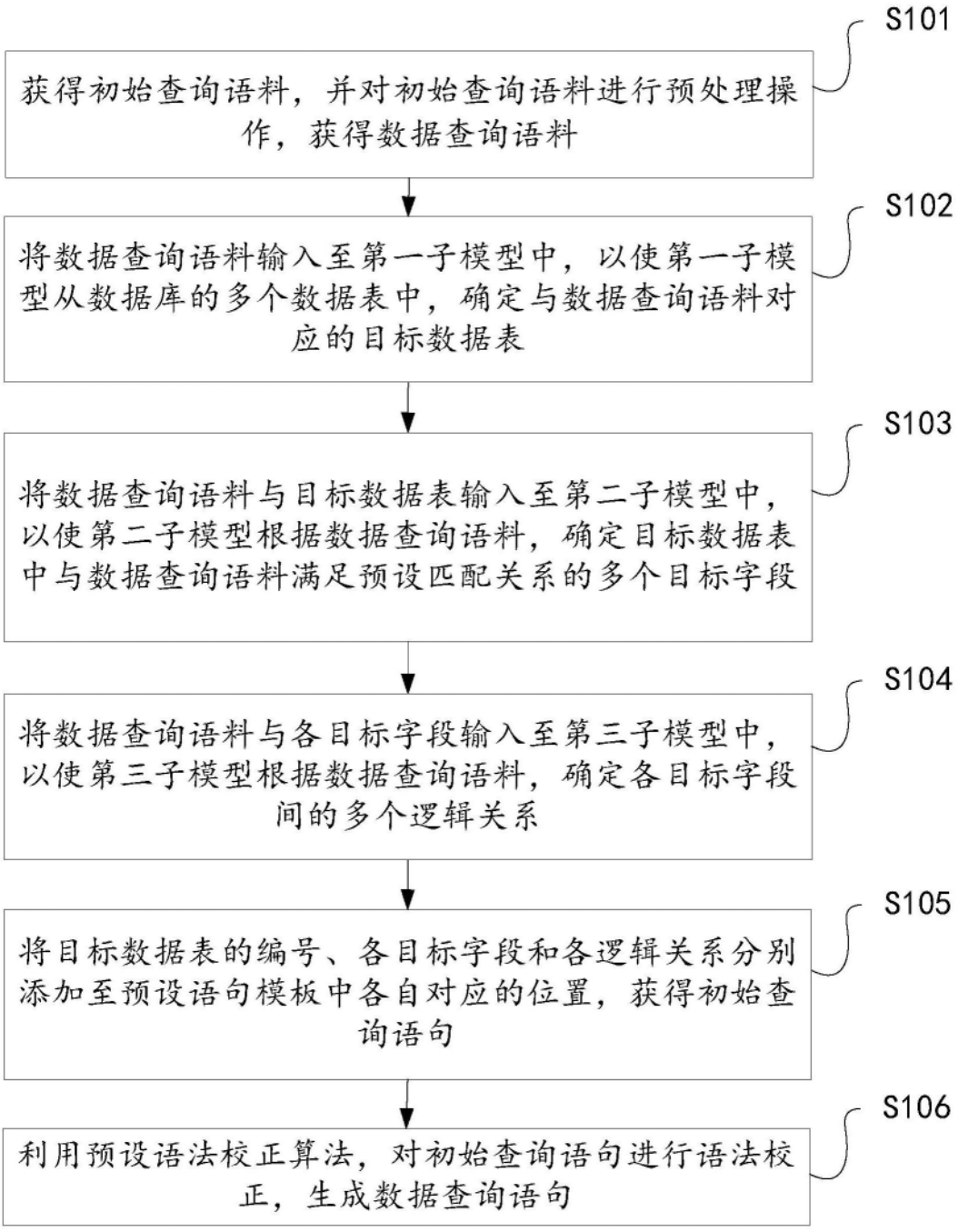

S101
获得初始查询语料，并对初始查询语料进行预处理操作，获得数据查询语料
S102
将数据查询语料输入至第一子模型中，以使第一子模型从数据库的多个数据表中，确定与数据查询语料对应的目标数据表
S103
将数据查询语料与目标数据表输入至第二子模型中，以使第二子模型根据数据查询语料，确定目标数据表中与数据查询语料满足预设匹配关系的多个目标字段
S104
将数据查询语料与各目标字段输入至第三子模型中，以使第三子模型根据数据查询语料，确定各目标字段间的多个逻辑关系
S105
将目标数据表的编号、各目标字段和各逻辑关系分别添加至预设语句模板中各自对应的位置，获得初始查询语句
S106
利用预设语法校正算法，对初始查询语句进行语法校正，生成数据查询语句

图1

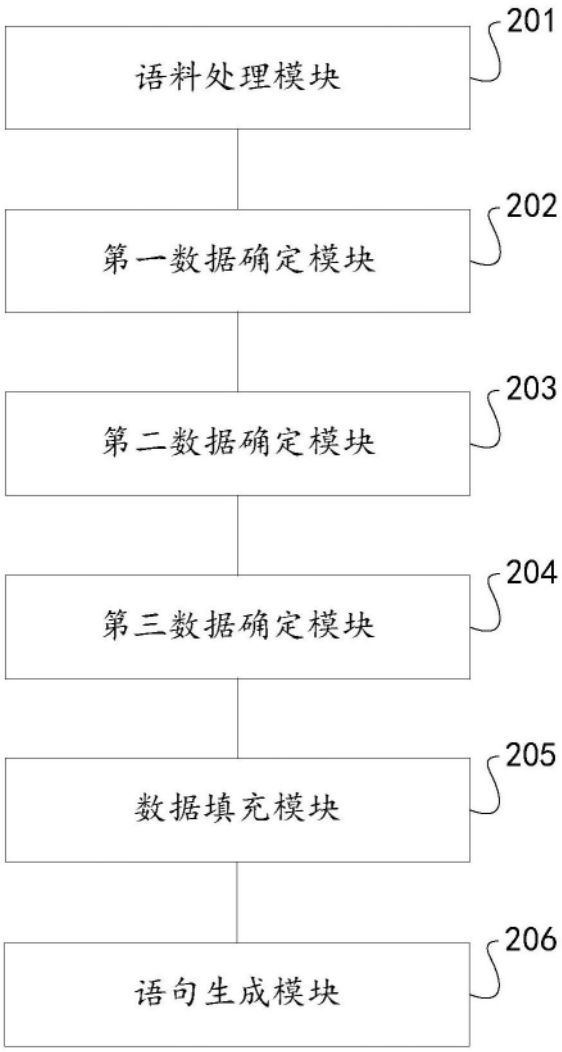

201
语料处理模块
202
第一数据确定模块
203
第二数据确定模块
204
第三数据确定模块
205
数据填充模块
206
语句生成模块

图2

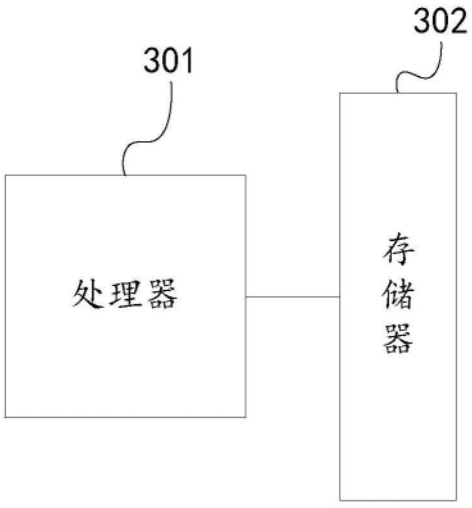

301
302
处理器
存储器

图3