(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120162425 A
(43)申请公布日 2025.06.17

(21)申请号 202510372120.2

(22)申请日 2025.03.27

(71)申请人 浪潮云信息技术股份公司
地址 250101 山东省济南市高新区浪潮路1036号浪潮科技园T03号楼

(72)发明人 赵兴业 李廷 韩同

(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 崔清杨

(51)Int.Cl.
*G06F 16/3329*(2025.01)
*G06F 16/9535*(2019.01)
*G06F 16/334*(2025.01)
*G06F 16/335*(2019.01)
*G06F 40/30*(2020.01)

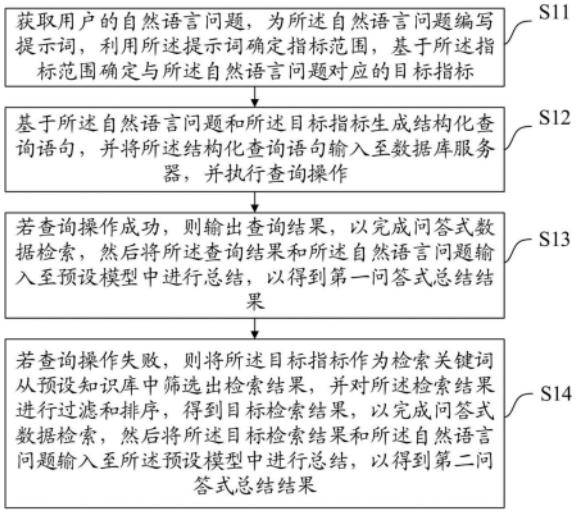
权利要求书2页 说明书10页 附图3页

(54)发明名称

# 问答式数据检索及大模型总结方法、装置、设备及介质

(57)摘要

本申请公开问答式数据检索及大模型总结方法、装置、设备及介质，涉及人工智能技术领域，为用户的自然语言问题编写提示词，确定与自然语言问题对应的目标指标；生成结构化查询语句，并输入至数据库服务器，执行查询操作；若查询操作成功，输出查询结果，将查询结果和自然语言问题输入至预设模型，得到第一问答式总结结果；若查询操作失败，将目标指标作为检索关键词筛选检索结果，进行过滤和排序，得到目标检索结果，将目标检索结果和自然语言问题输入至预设模型，得到第二问答式总结结果，在无需针对特定场景进行训练微调的情况下，实现问答式数据检索及大模型总结，降低模型部署复杂度，提高资源利用效率和问答式数据检索及大模型总结的准确率。




CN 120162425 A

1.一种问答式数据检索及大模型总结方法,其特征在于,包括:

获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;

基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;

若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;

若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。

2.根据权利要求1所述的问答式数据检索及大模型总结方法,其特征在于,所述为所述自然语言问题编写提示词,包括:

基于用户的自然语言问题并引入领域知识图谱确定提示词编写规则;

将所述编写规则、编写语句封装为模板,并利用所述自然语言问题对所述模板进行填充和调整,以生成提示词。

3.根据权利要求1所述的问答式数据检索及大模型总结方法,其特征在于,所述利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标,包括:

利用提示词从预设的数据表中筛选出与所述自然语言问题相关的目标数据表;

从所述目标数据表中确定指标范围,根据所述指标范围和所述自然语言问题编写目标提示词,利用所述目标提示词确定出与所述自然语言问题对应的目标指标。

4.根据权利要求1所述的问答式数据检索及大模型总结方法,其特征在于,所述基于所述自然语言问题和所述目标指标生成结构化查询语句,包括:

根据用户的自然语言问题编写结构化提示词;所述结构化提示词包括时间规则、地区规则、结构化查询语句生成规则、背景知识规则、语句及语句示例;

基于所述自然语言问题和所述目标指标,并利用NL2SQL技术和所述结构化提示词生成一个或多个结构化查询语句。

5.根据权利要求1所述的问答式数据检索及大模型总结方法,其特征在于,所述若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,包括:

若查询操作失败,则启动重试机制,记录当前重试次数,并跳转至生成结构化查询语句的流程,生成下一个结构化查询语句;

若所述下一个结构化查询语句对应的查询操作失败,则判断所述当前重试次数是否大于预设重试次数阈值;

若所述当前重试次数大于预设重试次数阈值,则将与自然语言问题对应的目标指标作为检索关键词从预设知识库中筛选出检索结果。

6.根据权利要求1所述的问答式数据检索及大模型总结方法,其特征在于,所述将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,包括:

获取与各指标相关的专项数据,利用检索增强生成技术构建专项知识库,将所述专项

数据以一比一复制的形式复制至所述专项知识库中；

获取与所有用户的自然语言问题相关的汇总数据，利用检索增强生成技术构建汇总知识库，将所述汇总数据以一比一复制的形式复制至所述汇总知识库中；所述预设知识库包括所述专项知识库和所述汇总知识库；

将所述目标指标作为检索关键词输入至所述专项知识库中进行检索，若所述专项知识库中存在相关数据，则输出检索结果；

若所述专项知识库中不存在相关数据，则将所述目标指标作为检索关键词输入至所述汇总知识库中进行检索，若所述汇总知识库中存在相关数据，则输出检索结果。

7.根据权利要求1至6任一项所述的问答式数据检索及大模型总结方法，其特征在于，所述对所述检索结果进行过滤和排序，包括：

根据用户的自然语言问题中的时间、地点以及指标对检索结果进行过滤，以得到过滤后的所述检索结果；

按照预设的排序规则对过滤后的所述检索结果进行排序；所述排序规则包括时效性规则、重要性规则以及相关性规则。

8.一种问答式数据检索及大模型总结装置，其特征在于，包括：

指标确定模块，用于获取用户的自然语言问题，为所述自然语言问题编写提示词，利用所述提示词确定指标范围，基于所述指标范围确定与所述自然语言问题对应的目标指标；

查询模块，用于基于所述自然语言问题和所述目标指标生成结构化查询语句，并将所述结构化查询语句输入至数据库服务器，并执行查询操作；

第一总结模块，用于若查询操作成功，则输出查询结果，以完成问答式数据检索，然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结，以得到第一问答式总结结果；

第二总结模块，用于若查询操作失败，则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果，并对所述检索结果进行过滤和排序，得到目标检索结果，以完成问答式数据检索，然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结，以得到第二问答式总结结果。

9.一种电子设备，其特征在于，包括：

存储器，用于保存计算机程序；

处理器，用于执行所述计算机程序，以实现如权利要求1至7任一项所述的问答式数据检索及大模型总结方法。

10.一种计算机可读存储介质，其特征在于，用于保存计算机程序；其中，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述的问答式数据检索及大模型总结方法。

# 问答式数据检索及大模型总结方法、装置、设备及介质

## 技术领域

[0001] 本发明涉及人工智能技术领域,特别涉及一种问答式数据检索及大模型总结方法、装置、设备及介质。

## 背景技术

[0002] 在当前的智能问答系统中,将用户的自然语言问题转换为可执行的SQL(Structured Query Language,结构化查询语句)是实现数据驱动回答的关键步骤,而现有技术中,在处理复杂问题或特定领域问题时,往往存在准确率不高的问题,尤其是在指标选择、条件构建等方面容易出现错误。此外,由于每个应用场景的数据结构和业务需求不同,对模型进行针对性的训练微调不仅需要大量的算力资源,还可能导致模型泛化能力下降,同时,当无法生成有效SQL时,系统往往缺乏有效的兜底策略,影响用户体验。

[0003] 由上可见,如何在无需针对特定场景进行训练微调的情况下,实现问答式数据检索及大模型总结,降低模型部署的复杂度,提高资源利用效率和问答式数据检索及大模型总结的准确率是本领域有待解决的问题。

## 发明内容

[0004] 有鉴于此,本发明的目的在于提供一种问答式数据检索及大模型总结方法、装置、设备及介质,能够在无需针对特定场景进行训练微调的情况下,实现问答式数据检索及大模型总结,降低模型部署的复杂度,提高资源利用效率和问答式数据检索及大模型总结的准确率。其具体方案如下:

[0005] 第一方面,本申请公开了一种问答式数据检索及大模型总结方法,包括:

[0006] 获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;

[0007] 基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;

[0008] 若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;

[0009] 若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。

[0010] 可选的,所述为所述自然语言问题编写提示词,包括:

[0011] 基于用户的自然语言问题并引入领域知识图谱确定提示词编写规则;

[0012] 将所述编写规则、编写语句封装为模板,并利用所述自然语言问题对所述模板进行填充和调整,以生成提示词。

[0013] 可选的,所述利用所述提示词确定指标范围,基于所述指标范围确定与所述自然

语言问题对应的目标指标,包括:

[0014] 利用提示词从预设的数据表中筛选出与所述自然语言问题相关的目标数据表;

[0015] 从所述目标数据表中确定指标范围,根据所述指标范围和所述自然语言问题编写目标提示词,利用所述目标提示词确定出与所述自然语言问题对应的目标指标。

[0016] 可选的,所述基于所述自然语言问题和所述目标指标生成结构化查询语句,包括:

[0017] 根据用户的自然语言问题编写结构化提示词;所述结构化提示词包括时间规则、地区规则、结构化查询语句生成规则、背景知识规则、语句及语句示例;

[0018] 基于所述自然语言问题和所述目标指标,并利用NL2SQL技术和所述结构化提示词生成一个或多个结构化查询语句。

[0019] 可选的,所述若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,包括:

[0020] 若查询操作失败,则启动重试机制,记录当前重试次数,并跳转至生成结构化查询语句的流程,生成下一个结构化查询语句;

[0021] 若所述下一个结构化查询语句对应的查询操作失败,则判断所述当前重试次数是否大于预设重试次数阈值;

[0022] 若所述当前重试次数大于预设重试次数阈值,则将与自然语言问题对应的目标指标作为检索关键词从预设知识库中筛选出检索结果。

[0023] 可选的,所述将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,包括:

[0024] 获取与各指标相关的专项数据,利用检索增强生成技术构建专项知识库,将所述专项数据以一比一复制的形式复制至所述专项知识库中;

[0025] 获取与所有用户的自然语言问题相关的汇总数据,利用检索增强生成技术构建汇总知识库,将所述汇总数据以一比一复制的形式复制至所述汇总知识库中;所述预设知识库包括所述专项知识库和所述汇总知识库;

[0026] 将所述目标指标作为检索关键词输入至所述专项知识库中进行检索,若所述专项知识库中存在相关数据,则输出检索结果;

[0027] 若所述专项知识库中不存在相关数据,则将所述目标指标作为检索关键词输入至所述汇总知识库中进行检索,若所述汇总知识库中存在相关数据,则输出检索结果。

[0028] 可选的,所述对所述检索结果进行过滤和排序,包括:

[0029] 根据用户的自然语言问题中的时间、地点以及指标对检索结果进行过滤,以得到过滤后的所述检索结果;

[0030] 按照预设的排序规则对过滤后的所述检索结果进行排序;所述排序规则包括时效性规则、重要性规则以及相关性规则。

[0031] 第二方面,本申请公开了一种问答式数据检索及大模型总结装置,包括:

[0032] 指标确定模块,用于获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;

[0033] 查询模块,用于基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;

[0034] 第一总结模块,用于若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;

[0035] 第二总结模块,用于若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。

[0036] 第三方面,本申请公开了一种电子设备,包括:

[0037] 存储器,用于保存计算机程序;

[0038] 处理器,用于执行所述计算机程序,以实现前述的问答式数据检索及大模型总结方法。

[0039] 第四方面,本申请公开了一种计算机存储介质,用于保存计算机程序;其中,所述计算机程序被处理器执行时实现前述公开的问答式数据检索及大模型总结方法的步骤。

[0040] 可见,本申请提供了一种问答式数据检索及大模型总结方法,包括获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。本申请为用户的自然语言问题编写提示词,从而确定目标指标,基于自然语言问题和目标指标生成结构化查询语句,能够在处理复杂或特定领域问题时,提升问数场景的整体准确率,将结构化查询语句输入至数据库服务器,并执行查询操作,若查询操作成功,则输出查询结果,完成问答式数据检索,将查询结果和自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果,利用大模型的原生能力,无需针对特定场景进行训练微调,降低了模型部署的成本和复杂度,若查询操作失败,则将目标指标作为检索关键词从预设知识库中筛选出检索结果,对检索结果进行过滤和排序,得到目标检索结果,完成问答式数据检索,将目标检索结果和自然语言问题输入至预设模型中进行总结,得到第二问答式总结结果,能够减少不必要的算力消耗,提高了资源利用效率,在数据库服务器失败后,利用预设知识库进行进一步检索,确保用户问题能够得到合理回应,提升系统的鲁棒性和用户体验,提高资源利用效率和问答式数据检索及大模型总结的准确率。

## 附图说明

[0041] 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,还可以根据提供的附图获得其他的附图。

[0042] 图1为本申请公开的一种问答式数据检索及大模型总结方法流程图;

[0043] 图2为本申请公开的一种问答式数据检索及大模型总结具体流程图;

[0044] 图3为本申请公开的一种问答式数据检索及大模型总结装置结构示意图;

[0045] 图4为本申请提供的一种电子设备结构图。

## 具体实施方式

[0046] 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。

[0047] 在当前的智能问答系统中,将用户的自然语言问题转换为可执行的SQL是实现数据驱动回答的关键步骤,而现有技术中,在处理复杂问题或特定领域问题时,往往存在准确率不高的问题,尤其是在指标选择、条件构建等方面容易出现错误。此外,由于每个应用场景的数据结构和业务需求不同,对模型进行针对性的训练微调不仅需要大量的算力资源,还可能导致模型泛化能力下降,同时,当无法生成有效SQL时,系统往往缺乏有效的兜底策略,影响用户体验。由上可见,如何在无需针对特定场景进行训练微调的情况下,实现问答式数据检索及大模型总结,降低模型部署的复杂度,提高资源利用效率和问答式数据检索及大模型总结的准确率是本领域有待解决的问题。

[0048] 参见图1所示,本发明实施例公开了一种问答式数据检索及大模型总结方法,具体可以包括:

[0049] 步骤S11:获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标。

[0050] 本实施例中,获取用户的自然语言问题,基于用户的自然语言问题并引入领域知识图谱确定提示词编写规则,将所述编写规则、编写语句封装为模板,并利用所述自然语言问题对所述模板进行填充和调整,以生成提示词,利用提示词从预设的数据表中筛选出与所述自然语言问题相关的目标数据表,从所述目标数据表中确定指标范围,根据所述指标范围和所述自然语言问题编写目标提示词,利用所述目标提示词确定出与所述自然语言问题对应的目标指标。

[0051] 具体的,在自然语言问题中,用户通常会询问关于某个或某些预设的数据表的信息。首先需要确定用户问题所涉及的预设的数据表,流程如下:

[0052] 编写Prompt(提示词):为了确定用户问题所涉及的数据表,首先需要编写一个prompt,这个prompt包含了所有可能的建表语句、用户问题以及选表规则等,选表规则可以根据用户问题中的关键词、短语或上下文来确定,基于编写的prompt,本申请利用通用大模型的语义理解能力,自动选择与用户问题最相关的目标数据表,如果问题仅涉及单表,则可以跳过此步骤,在确定了用户的自然语言问题所涉及的目标数据表后,需要进一步确定自然语言问题中所涉及的指标(即目标数据表中的列),根据选表结果框定一个可能的指标范围,并基于用户的自然语言问题及指标范围内指标再次编写一个目标提示词,这个目标提示词包含了选取的范围内指标、自然语言问题以及选取指标规则等,基于编写的目标提示词,利用通用大模型的语义理解能力,自动选择与用户的自然语言问题最相关的目标指标。

[0053] 此外,本申请采用一些优化策略来提高Prompt的编写效率和准确性,例如,可以利用模板化的方法来编写Prompt,将常用的建表语句、选表规则、SQL生成规则等封装成模板,并在需要时根据具体问题进行填充和调整。

[0054] 进一步的,在某些特定领域(如医疗、金融等)中,用户的自然语言问题可能涉及大量的专业术语和领域知识。为了处理这些问题,本申请将领域知识图谱引入技术方案中,领域知识图谱是一种结构化的知识表示方式,它可以将领域中的实体、属性和关系以图的形式进行表示。通过引入领域知识图谱,为技术方案提供更多的背景知识和上下文信息,从而提高查询的准确性和效率。

[0055] 步骤S12:基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作。

[0056] 本实施例中,根据用户的自然语言问题编写结构化提示词;所述结构化提示词包括时间规则、地区规则、结构化查询语句生成规则、背景知识规则、语句及语句示例;基于所述自然语言问题和所述目标指标,并利用NL2SQL技术和所述结构化提示词生成一个或多个结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作。

[0057] 本申请在确定了用户的自然语言问题所涉及的目标数据表和目标指标后,下一步就是生成SQL查询语句,具体流程如下:

[0058] 编写结构化提示词:为了生成SQL查询语句,首先需要编写一个包含生成SQL规则的结构化提示词,这个结构化提示词包含了筛选时间规则、筛选地区规则等SQL生成规则,以及选取指标的相关背景知识规则、建表语句、选取指标相关insert语句示例和用户问题等;

[0059] 生成SQL:基于编写的结构化提示词,利用通用大模型的生成能力,自动生成一个或多个SQL查询语句,这些SQL查询语句能够准确地反映用户问题的意图。

[0060] 在生成SQL查询语句后,将SQL查询语句输入至数据库服务器,并执行查询操作,数据库服务器会返回查询结果,这些结果将用于后续的分析和总结。

[0061] 步骤S13:若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果。

[0062] 步骤S14:若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。

[0063] 本实施例中,若查询操作失败,则启动重试机制,记录当前重试次数,并跳转至生成结构化查询语句的流程,生成下一个结构化查询语句;若所述下一个结构化查询语句对应的查询操作失败,则判断所述当前重试次数是否大于预设重试次数阈值;若所述当前重试次数大于预设重试次数阈值,则获取与各指标相关的专项数据,利用检索增强生成技术构建专项知识库,将所述专项数据以一比一复制的形式复制至所述专项知识库中;获取与所有用户的自然语言问题相关的汇总数据,利用检索增强生成技术构建汇总知识库,将所述汇总数据以一比一复制的形式复制至所述汇总知识库中;所述预设知识库包括所述专项知识库和所述汇总知识库;将所述目标指标作为检索关键词输入至所述专项知识库中进行检索,若所述专项知识库中存在相关数据,则输出检索结果;若所述专项知识库中不存在相

关数据,则将所述目标指标作为检索关键词输入至所述汇总知识库中进行检索,若所述汇总知识库中存在相关数据,则输出检索结果,然后根据用户的自然语言问题中的时间、地点以及指标对检索结果进行过滤,以得到过滤后的所述检索结果;按照预设的排序规则对过滤后的所述检索结果进行排序,得到目标检索结果,以完成问答式数据检索;所述排序规则包括时效性规则、重要性规则以及相关性规则。

[0064] 具体的,将SQL查询语句输入至数据库服务器,并执行查询操作,数据库服务器会返回查询结果,若查询操作失败(例如由于语法错误或数据库连接问题),则启动重试机制,重试机制可以包括调整提示词并重新生成SQL查询语句、修改数据库连接参数等。此外,还可以设置一个预设重试次数阈值,以确保在多次尝试后仍然无法成功查询时,能够及时地转向其他处理策略。

[0065] 若当前重试次数大于预设重试次数阈值,则表明此时无法成功生成SQL查询语句或查询结果不符合用户期望,此时则需要利用预设知识库进行检索,以便为用户提供准确的回答,具体流程为:

[0066] (1)构建预设知识库。构建专项知识库:针对具体的指标,利用检索增强生成技术构建专项知识库,将专项数据以一比一复制的形式复制至专项知识库中,以确保检索结果的针对性和准确性,同时,为了支持高效的检索操作,需要对知识库中的数据进行合理的索引和存储;构建汇总知识库:除了专项知识库外,还可以构建一个包含所有数据的汇总知识库,这个知识库可以作为兜底方案,在专项知识库无法找到匹配数据时提供额外的检索机会。

[0067] (2)知识库检索。在构建了预设知识库后,下一步就是实现知识库检索功能,提取指标:提取与自然语言问题对应的目标指标,将这些目标指标将作为检索关键词,用于在预设知识库中查找相关的数据。

[0068] 优先检索专项知识库:基于提取的目标指标,优先在专项知识库中进行检索。由于专项知识库只包含与当前指标相关的数据,因此检索结果通常更加准确和有针对性。

[0069] 其次检索汇总知识库:如果专项知识库无法找到匹配的数据,则转向汇总知识库进行检索,由于汇总知识库包含了所有数据,因此它可以作为兜底方案,为用户提供额外的检索机会。

[0070] 此外,还可以调整检索阈值:为了提高检索的召回率,可以适当降低检索阈值。这意味着在检索过程中,会放宽对匹配度的要求,以找到更多可能相关的数据,然而,这也会增加检索结果的噪声和冗余度,因此需要在后续的步骤中对检索结果进行过滤和排序。

[0071] (3)结果过滤与排序。在得到相关的检索结果后,需要对检索结果进行过滤和排序,以确保用户获得准确、有条理的回答。

[0072] 过滤:根据用户的自然语言问题中的时间、地点、指标等条件,可以对检索到的结果进行过滤。过滤操作可以去除与问题无关或不符合条件的数据,从而提高结果的准确性和相关性。

[0073] 排序:为了使用户更容易理解和接受检索结果,需要对过滤后的检索结果进行排序。排序可以根据数据的时效性、重要性、相关性等因素进行。通过排序操作,可以使结果更加规律、有条理地呈现给用户。

[0074] 最后,将目标检索结果和自然语言问题输入至预设模型中进行总结,以得到第二

问答式总结结果。为了生成准确的总结结果,我们需要将SQL语句、目标检索结果和自然语言问题作为输入信息提供给预设模型,这些信息可以帮助预设模型理解查询的上下文和目的,并生成符合用户期望的回答(即第二问答式总结结果)。基于输入的信息,预设模型可以利用其强大的语义理解和生成能力,自动生成一个或多个总结结果。这些总结结果应该能够准确地反映查询结果的内容,并以简洁明了的方式呈现给用户。

[0075] 生成第一问答式总结结果的方式与生成第二问答式总结结果的方式相同。同时,由于检索结果可能包含多个相关的数据点或信息片段,因此预设模型还需要具备将这些信息整合成一个连贯、有条理的回答的能力。

[0076] 本申请通过NL2SQL分步分解技术与RAG(Retrieval Augmented Generation)技术相结合,构建了问答式数据检索及大模型总结系统,包括NL2SQL模块、RAG模块、模型总结模块、用户接口模块,具体流程如图2所示。

[0077] NL2SQL模块,用于根据用户的自然语言问题生成SQL查询语句,并尝试执行查询;主要包括选表步骤、指标确定步骤、SQL生成步骤、数据库查询步骤和重试机制步骤;

[0078] RAG模块,用于当NL2SQL模块查询失败时,基于问题检索预设知识库中的相关数据,并进行过滤和排序;主要包括数据知识库构建步骤、知识库检索步骤和检索结果过滤及排序步骤;

[0079] 模型总结模块,用于将NL2SQL模块或RAG模块的输出结果与用户的自然语言问题一起输入预设模型中进行总结,得到第一问答式总结结果或第二问答式总结结果;

[0080] 用户接口模块,用于接收用户输入的自然语言问题,并展示模型总结的第一问答式总结结果或第二问答式总结结果。

[0081] 本申请利用通用大模型的原生能力,无需针对特定场景进行训练微调,降低了模型部署的成本和复杂度,显著提升问数场景的整体准确率,特别是在处理复杂或特定领域问题时;在不牺牲模型性能的前提下,尽可能降低训练微调所需的算力资源,提高资源利用效率;结合NL2SQL和RAG的双重保障,显著提升了问数场景的准确率,特别是在处理复杂或特定领域问题时表现更为突出;通过高效的NL2SQL分步分解和RAG兜底策略,减少了不必要的算力消耗,提高了资源利用效率,确保用户问题能够得到合理回应,提升系统的鲁棒性和用户体验。

[0082] 本实施例中,获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。本申请为用户的自然语言问题编写提示词,从而确定目标指标,基于自然语言问题和目标指标生成结构化查询语句,能够在处理复杂或特定领域问题时,提升问数场景的整体准确率,将结构化查询语句输入至数据库服务器,并执行查询操作,若查询操作成功,则输出查询结果,完成问答式数据检索,将查询结果和自然语言问题输入至

预设模型中进行总结,以得到第一问答式总结结果,利用大模型的原生能力,无需针对特定场景进行训练微调,降低了模型部署的成本和复杂度,若查询操作失败,则将目标指标作为检索关键词从预设知识库中筛选出检索结果,对检索结果进行过滤和排序,得到目标检索结果,完成问答式数据检索,将目标检索结果和自然语言问题输入至预设模型中进行总结,得到第二问答式总结结果,能够减少不必要的算力消耗,提高了资源利用效率,在数据库服务器失败后,利用预设知识库进行进一步检索,确保用户问题能够得到合理回应,提升系统的鲁棒性和用户体验,提高资源利用效率和问答式数据检索及大模型总结的准确率。

[0083] 参见图3所示,本发明实施例公开了一种问答式数据检索及大模型总结装置,具体可以包括:

[0084] 指标确定模块11,用于获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;

[0085] 查询模块12,用于基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;

[0086] 第一总结模块13,用于若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;

[0087] 第二总结模块14,用于若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。

[0088] 本实施例中,获取用户的自然语言问题,为所述自然语言问题编写提示词,利用所述提示词确定指标范围,基于所述指标范围确定与所述自然语言问题对应的目标指标;基于所述自然语言问题和所述目标指标生成结构化查询语句,并将所述结构化查询语句输入至数据库服务器,并执行查询操作;若查询操作成功,则输出查询结果,以完成问答式数据检索,然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果;若查询操作失败,则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果,并对所述检索结果进行过滤和排序,得到目标检索结果,以完成问答式数据检索,然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结,以得到第二问答式总结结果。本申请为用户的自然语言问题编写提示词,从而确定目标指标,基于自然语言问题和目标指标生成结构化查询语句,能够在处理复杂或特定领域问题时,提升问数场景的整体准确率,将结构化查询语句输入至数据库服务器,并执行查询操作,若查询操作成功,则输出查询结果,完成问答式数据检索,将查询结果和自然语言问题输入至预设模型中进行总结,以得到第一问答式总结结果,利用大模型的原生能力,无需针对特定场景进行训练微调,降低了模型部署的成本和复杂度,若查询操作失败,则将目标指标作为检索关键词从预设知识库中筛选出检索结果,对检索结果进行过滤和排序,得到目标检索结果,完成问答式数据检索,将目标检索结果和自然语言问题输入至预设模型中进行总结,得到第二问答式总结结果,能够减少不必要的算力消耗,提高了资源利用效率,在数据库服务器失败后,利用预设知识库进行进一步检索,确保用户问题能够得到合理回应,提升系统

的鲁棒性和用户体验,提高资源利用效率和问答式数据检索及大模型总结的准确率。

[0089] 在一些具体实施例中,所述指标确定模块11,具体可以包括:

[0090] 规则确定模块,用于基于用户的自然语言问题并引入领域知识图谱确定提示词编写规则;

[0091] 提示词生成模块,用于将所述编写规则、编写语句封装为模板,并利用所述自然语言问题对所述模板进行填充和调整,以生成提示词。

[0092] 在一些具体实施例中,所述指标确定模块11,具体可以包括:

[0093] 目标数据表筛选模块,用于利用提示词从预设的数据表中筛选出与所述自然语言问题相关的目标数据表;

[0094] 目标指标确定模块,用于从所述目标数据表中确定指标范围,根据所述指标范围和所述自然语言问题编写目标提示词,利用所述目标提示词确定出与所述自然语言问题对应的目标指标。

[0095] 在一些具体实施例中,所述查询模块12,具体可以包括:

[0096] 规则生成模块,用于根据用户的自然语言问题编写结构化提示词;所述结构化提示词包括时间规则、地区规则、结构化查询语句生成规则、背景知识规则、语句及语句示例;

[0097] 结构化查询语句生成模块,用于基于所述自然语言问题和所述目标指标,并利用NL2SQL技术和所述结构化提示词生成一个或多个结构化查询语句。

[0098] 在一些具体实施例中,所述第二总结模块14,具体可以包括:

[0099] 重试机制启动模块,用于若查询操作失败,则启动重试机制,记录当前重试次数,并跳转至生成结构化查询语句的流程,生成下一个结构化查询语句;

[0100] 判断模块,用于若所述下一个结构化查询语句对应的查询操作失败,则判断所述当前重试次数是否大于预设重试次数阈值;

[0101] 第一检索模块,用于若所述当前重试次数大于预设重试次数阈值,则将与自然语言问题对应的目标指标作为检索关键词从预设知识库中筛选出检索结果。

[0102] 在一些具体实施例中,所述第二总结模块14,具体可以包括:

[0103] 专项知识库构建模块,用于获取与各指标相关的专项数据,利用检索增强生成技术构建专项知识库,将所述专项数据以一比一复制的形式复制至所述专项知识库中;

[0104] 汇总知识库构建模块,用于获取与所有用户的自然语言问题相关的汇总数据,利用检索增强生成技术构建汇总知识库,将所述汇总数据以一比一复制的形式复制至所述汇总知识库中;所述预设知识库包括所述专项知识库和所述汇总知识库;

[0105] 第二检索模块,用于将所述目标指标作为检索关键词输入至所述专项知识库中进行检索,若所述专项知识库中存在相关数据,则输出检索结果;

[0106] 第三检索模块,用于若所述专项知识库中不存在相关数据,则将所述目标指标作为检索关键词输入至所述汇总知识库中进行检索,若所述汇总知识库中存在相关数据,则输出检索结果。

[0107] 在一些具体实施例中,所述第二总结模块14,具体可以包括:

[0108] 过滤模块,用于根据用户的自然语言问题中的时间、地点以及指标对检索结果进行过滤,以得到过滤后的所述检索结果;

[0109] 排序模块,用于按照预设的排序规则对过滤后的所述检索结果进行排序;所述排

序规则包括时效性规则、重要性规则以及相关性规则。

[0110] 图4为本申请实施例提供的一种电子设备的结构示意图。该电子设备20,具体可以包括:至少一个处理器21、至少一个存储器22、电源23、通信接口24、输入输出接口25和通信总线26。其中,所述存储器22用于存储计算机程序,所述计算机程序由所述处理器21加载并执行,以实现前述任一实施例公开的由电子设备执行的问答式数据检索及大模型总结方法中的相关步骤。

[0111] 本实施例中,电源23用于为电子设备20上的各硬件设备提供工作电压;通信接口24能够为电子设备20创建与外界设备之间的数据传输通道,其所遵循的通信协议是能够适用于本申请技术方案的任意通信协议,在此不对其进行具体限定;输入输出接口25,用于获取外界输入数据或向外界输出数据,其具体的接口类型可以根据具体应用需要进行选取,在此不进行具体限定。

[0112] 另外,存储器22作为资源存储的载体,可以是只读存储器、随机存储器、磁盘或者光盘等,其上所存储的资源包括操作系统221、计算机程序222及数据223等,存储方式可以是短暂存储或者永久存储。

[0113] 其中,操作系统221用于管理与控制电子设备20上的各硬件设备以及计算机程序222,以实现处理器21对存储器22中数据223的运算与处理,其可以是Windows、Unix、Linux等。计算机程序222除了包括能够用于完成前述任一实施例公开的由电子设备20执行的问答式数据检索及大模型总结方法的计算机程序之外,还可以进一步包括能够用于完成其他特定工作的计算机程序。数据223除了可以包括问答式数据检索及大模型总结设备接收到的由外部设备传输进来的数据,也可以包括由自身输入输出接口25采集到的数据等。

[0114] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块,或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0115] 进一步的,本申请实施例还公开了一种计算机可读存储介质,所述存储介质中存储有计算机程序,所述计算机程序被处理器加载并执行时,实现前述任一实施例公开的问答式数据检索及大模型总结方法步骤。

[0116] 最后,还需要说明的是,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。而且,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、物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、物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,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0117] 以上对本发明所提供的一种问答式数据检索及大模型总结方法、装置、设备及存储介质进行了详细介绍,本文中应用了具体个例对本发明的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本发明的限制。

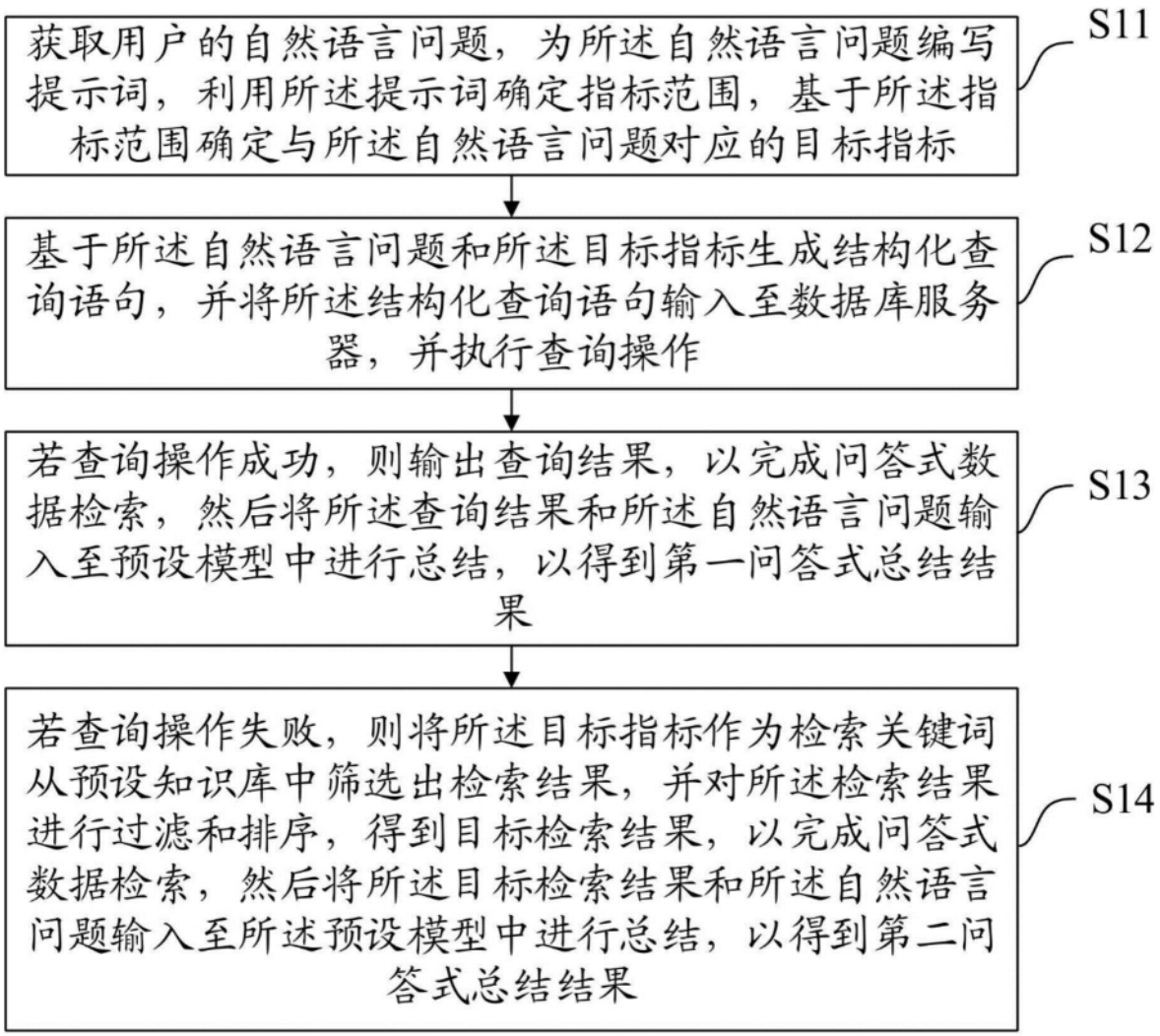

获取用户的自然语言问题，为所述自然语言问题编写提示词，利用所述提示词确定指标范围，基于所述指标范围确定与所述自然语言问题对应的目标指标
S11
基于所述自然语言问题和所述目标指标生成结构化查询语句，并将所述结构化查询语句输入至数据库服务器，并执行查询操作
S12
若查询操作成功，则输出查询结果，以完成问答式数据检索，然后将所述查询结果和所述自然语言问题输入至预设模型中进行总结，以得到第一问答式总结结果
S13
若查询操作失败，则将所述目标指标作为检索关键词从预设知识库中筛选出检索结果，并对所述检索结果进行过滤和排序，得到目标检索结果，以完成问答式数据检索，然后将所述目标检索结果和所述自然语言问题输入至所述预设模型中进行总结，以得到第二问答式总结结果
S14

图1

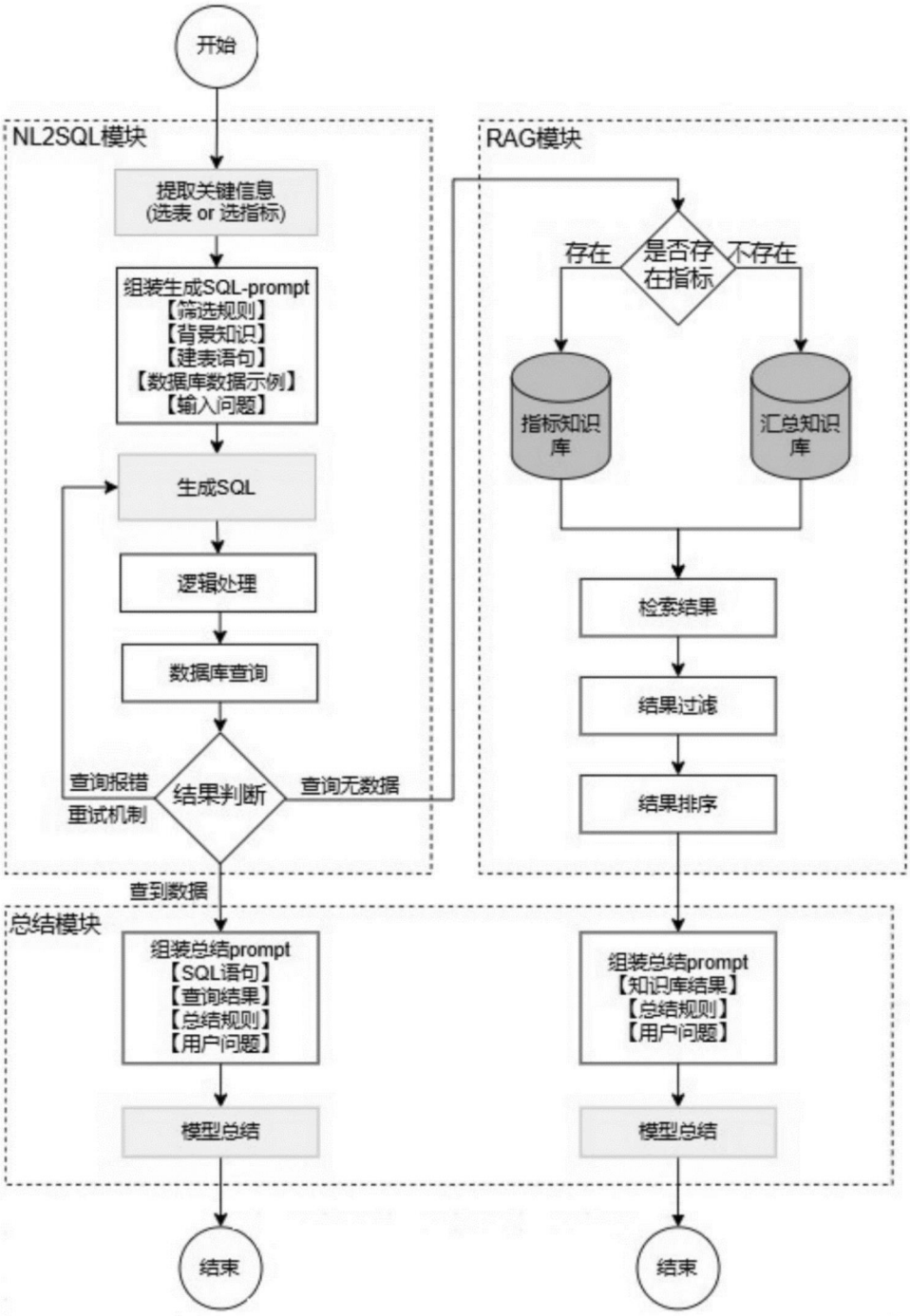

开始
NL2SQL模块
提取关键信息
(选表 or 选指标)
组装生成SQL-prompt
【筛选规则】
【背景知识】
【建表语句】
【数据库数据示例】
【输入问题】
生成SQL
逻辑处理
数据库查询
结果判断
查询报错
重试机制
查询无数据
查到数据
RAG模块
是否存在指标
存在
不存在
指标知识库
汇总知识库
检索结果
结果过滤
结果排序
总结模块
组装总结prompt
【SQL语句】
【查询结果】
【总结规则】
【用户问题】
模型总结
组装总结prompt
【知识库结果】
【总结规则】
【用户问题】
模型总结
结束
结束

图2


指标确定模块11
查询模块12
第一总结模块13
第二总结模块14

图3

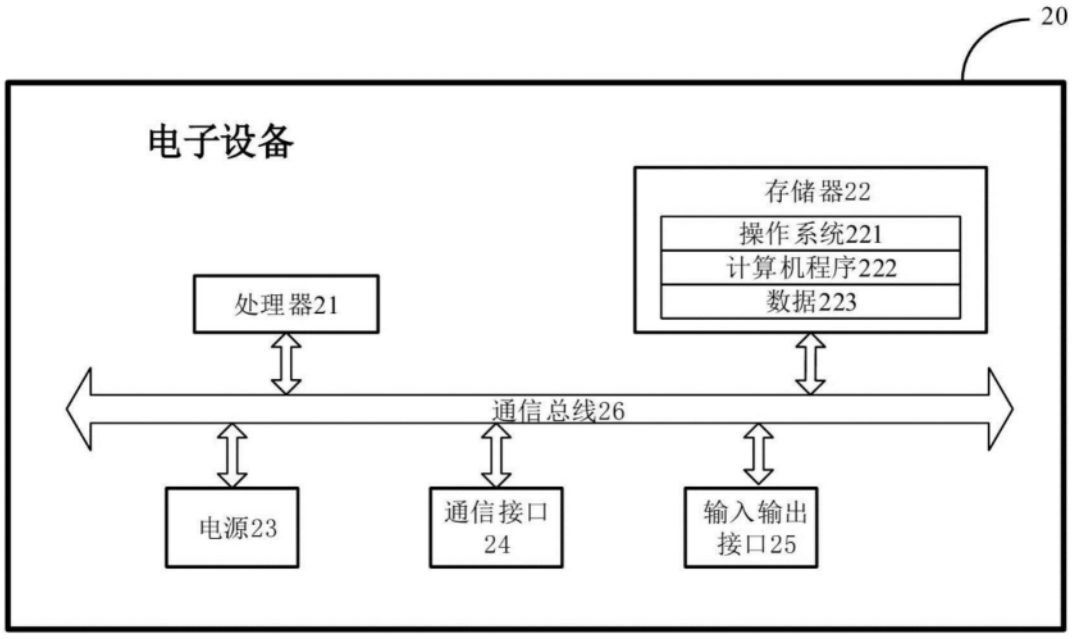

20
电子设备
存储器22
操作系统221
计算机程序222
数据223
处理器21
通信总线26
电源23
通信接口24
输入输出接口25

图4