(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118132762 A
(43)申请公布日 2024.06.04

(21)申请号 202410271746.X

(22)申请日 2024.03.11

(71)申请人 浙江阿里巴巴机器人有限公司
地址 310052 浙江省杭州市滨江区网商路699号1号楼3楼304室

(72)发明人 李鹤 余海洋 李永彬 黄非

(74)专利代理机构 北京同钧律师事务所 16037
专利代理师 柴海平 许怀远

(51)Int.Cl.
G06F 16/36(2019.01)
G06Q 30/014(2023.01)
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 16/35(2019.01)
G06N 3/088(2023.01)
G06F 18/22(2023.01)
G06F 40/30(2020.01)
G06F 40/186(2020.01)

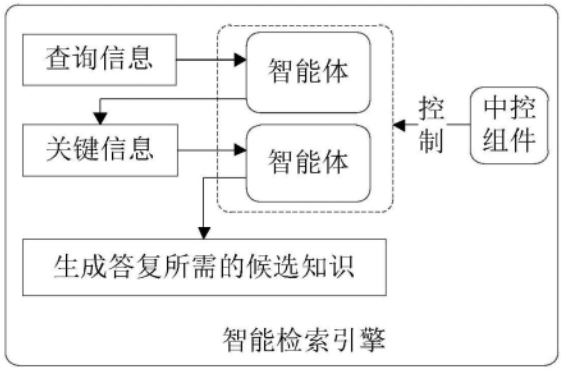
权利要求书4页 说明书23页 附图6页

(54)发明名称

多智能体的智能检索引擎、人机交互系统和方法

(57)摘要

本申请提供一种多智能体的智能检索引擎、人机交互系统和方法。本申请的方案应用于人机交互系统中,基于多智能体构建智能检索引擎,通过中控组件控制多智能体的协作提取查询信息的关键信息,并基于关键信息检索生成答复所需的候选知识,智能体借助人工智能模型对查询信息进行深度语义理解,可以精准、全面地提取查询信息的关键信息;智能体借助人工智能模型基于关键信息检索并召回生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度,进而可以提升基于候选知识生成的答复信息的质量,得到让用户满意的答复。



CN 118132762 A

1.一种多智能体的智能检索引擎,其特征在于,包括:中控组件和多个智能体,

所述中控组件接收人机交互系统发送的查询信息;

所述中控组件控制所述多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识;

所述中控组件向所述人机交互系统返回检索到的候选知识,所述候选知识用于生成所述查询信息的答复信息。

2.根据权利要求1所述的智能检索引擎,其特征在于,还包括:信息抽取智能体和检索智能体,

所述中控组件控制所述多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识,包括:

所述中控组件向所述信息抽取智能体发送所述查询信息;

所述信息抽取智能体提取所述查询信息的关键信息,并向所述中控组件返回所述关键信息;

所述中控组件向所述检索智能体发送所述关键信息;

所述检索智能体基于所述关键信息检索生成答复所需的候选知识,并向所述中控组件返回检索到的候选知识。

3.根据权利要求2所述的智能检索引擎,其特征在于,还包括:需求分析智能体,

所述中控组件向所述信息抽取智能体发送所述查询信息,包括:

所述中控组件获取所述人机交互系统的交互需求信息,向所述需求分析智能体发送所述查询信息和交互需求信息;

所述需求分析智能体基于所述查询信息和交互需求信息进行检索需求分析,确定答复所述查询信息是否需要进行知识检索,得到检索需求分析结果,并向所述中控组件返回检索需求分析结果;

所述中控组件根据所述检索需求分析结果,在确定需要检索的情况下,向所述信息抽取智能体发送所述查询信息。

4.根据权利要求3所述的智能检索引擎,其特征在于,所述交互需求信息包括检索规则,

所述需求分析智能体基于所述查询信息和交互需求信息进行检索需求分析,确定答复所述查询信息是否需要进行知识检索,得到检索需求分析结果,并向所述中控组件返回检索需求分析结果,包括:

所述需求分析智能体基于所述检索规则,判断答复所述查询信息是否需要进行知识检索,得到判断结果;

在所述判断结果为需要进行知识检索的情况下,向所述中控组件返回包含所述判断结果的检索需求分析结果;

在所述判断结果为不需要进行知识检索的情况下,确定所述查询信息对应的答复话术信息,向所述中控组件返回包含所述判断结果和所述答复话术信息的检索需求分析结果。

5.根据权利要求4所述的智能检索引擎,其特征在于,所述交互需求信息还包括配置的数据源,

所述需求分析智能体在所述判断结果为需要进行知识检索的情况下,向所述中控组件

返回包含所述判断结果的检索需求分析结果,包括:

所述需求分析智能体在所述判断结果为需要进行知识检索的情况下,向所述中控组件返回包含所述判断结果和所述数据源的检索需求分析结果。

6.根据权利要求5所述的智能检索引擎,其特征在于,所述中控组件根据所述检索需求分析结果,在确定需要检索的情况下,向所述信息抽取智能体发送所述查询信息,包括:

所述中控组件根据所述检索需求分析结果,在确定需要检索的情况下,向所述信息抽取智能体发送所述查询信息和所述数据源;

所述检索智能体基于所述关键信息检索生成答复所需的候选知识,包括:

所述检索智能体基于所述关键信息,在所述数据源中检索生成答复所需的候选知识。

7.根据权利要求2所述的智能检索引擎,其特征在于,还包括:引用监测智能体,

所述中控组件向所述人机交互系统返回检索到的候选知识,包括:

所述中控组件向所述引用监测智能体发送所述查询信息和检索到的候选知识;

所述引用监测智能体根据所述查询信息和检索到的候选知识,判断检索到的候选知识是否满足所述查询信息的答复需求,得到判断结果,并向所述中控组件返回所述判断结果;

所述中控组件根据所述判断结果,在检索到的候选知识满足所述答复需求的情况下,向所述人机交互系统返回检索到的候选知识。

8.根据权利要求7所述的智能检索引擎,其特征在于,还包括:问题识别智能体,

所述中控组件根据所述判断结果,在检索到的候选知识不满足所述答复需求的情况下,向所述问题识别智能体发送所述判断结果、所述查询信息和检索到的候选知识;

所述问题识别智能体在所述判断结果为检索到的候选知识不满足所述答复需求的情况下,根据所述查询信息和检索到的候选知识,识别检索到的候选知识不能满足所述答复需求的原因,得到原因识别结果,并向所述中控组件返回所述原因识别结果。

9.根据权利要求8所述的智能检索引擎,其特征在于,还包括:二次规划智能体,

所述中控组件向所述二次规划智能体发送所述查询信息、检索到的候选知识和所述原因识别结果;

所述二次规划智能体根据所述查询信息、检索到的候选知识和所述原因识别结果,生成补充检索信息,并向所述中控组件返回所述补充检索信息;

所述中控组件向所述检索智能体发送所述补充检索信息;

所述检索智能体基于所述补充检索信息进行候选知识的补充检索,并向所述中控组件返回补充检索到的候选知识。

10.根据权利要求9所述的智能检索引擎,其特征在于,所述中控组件向所述检索智能体发送所述补充检索信息,包括:

所述中控组件基于配置的补充检索次数阈值,在补充检索的次数小于或等于所述补充检索次数阈值的情况下,向所述检索智能体发送所述补充检索信息;

所述中控组件在所述补充检索的次数大于所述补充检索次数阈值的情况下,向所述人机交互系统返回已检索到的候选知识。

11.一种人机交互方法,其特征在于,包括:

接收查询信息;

根据所述查询信息,调用多智能体的智能检索引擎,通过所述智能检索引擎的中控组

件控制多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识;

将所述查询信息和所述候选知识输入人机交互模型,通过所述人机交互模型根据所述候选知识生成所述查询信息的答复信息;

输出所述答复信息。

12.一种多智能体的人机交互系统,其特征在于,包括:中控组件、用于知识检索的多个智能体和用于生成答复的问答智能体,

所述中控组件接收查询信息;

所述中控组件控制用于知识检索的多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识;

所述中控组件向所述问答智能体发送所述查询信息和检索到的候选知识;

所述问答智能体根据所述候选知识生成所述查询信息的答复信息,并向所述中控组件返回所述答复信息;

所述中控组件输出所述答复信息。

13.根据权利要求12所述的系统,其特征在于,还包括:信息抽取智能体和检索智能体,

所述中控组件控制所述多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识,包括:

所述中控组件向所述信息抽取智能体发送所述查询信息;

所述信息抽取智能体提取所述查询信息的关键信息,并向所述中控组件返回所述关键信息;

所述中控组件向所述检索智能体发送所述关键信息;

所述检索智能体基于所述关键信息检索生成答复所需的候选知识,并向所述中控组件返回检索到的候选知识。

14.根据权利要求13所述的系统,其特征在于,还包括:需求分析智能体,

所述中控组件向所述信息抽取智能体发送所述查询信息,包括:

所述中控组件获取所述人机交互系统的交互需求信息,向所述需求分析智能体发送所述查询信息和交互需求信息;

所述需求分析智能体基于所述查询信息和交互需求信息进行检索需求分析,确定答复所述查询信息是否需要进行知识检索,得到检索需求分析结果,并向所述中控组件返回检索需求分析结果;

所述中控组件根据所述检索需求分析结果,在确定需要检索的情况下,向所述信息抽取智能体发送所述查询信息。

15.根据权利要求13所述的系统,其特征在于,还包括:引用监测智能体,

所述中控组件向所述人机交互系统返回检索到的候选知识,包括:

所述中控组件向所述引用监测智能体发送所述查询信息和检索到的候选知识;

所述引用监测智能体根据所述查询信息和检索到的候选知识,判断检索到的候选知识是否满足所述查询信息的答复需求,得到判断结果,并向所述中控组件返回所述判断结果;

所述中控组件根据所述判断结果,在检索到的候选知识满足所述答复需求的情况下,向所述人机交互系统返回检索到的候选知识。

16.根据权利要求15所述的系统,其特征在于,还包括:问题识别智能体,

所述中控组件根据所述判断结果,在检索到的候选知识不满足所述答复需求的情况下,向所述问题识别智能体发送所述判断结果、所述查询信息和检索到的候选知识;

所述问题识别智能体在所述判断结果为检索到的候选知识不满足所述答复需求的情况下,根据所述查询信息和检索到的候选知识,识别检索到的候选知识不能满足所述答复需求的原因,得到原因识别结果,并向所述中控组件返回所述原因识别结果。

17.根据权利要求16所述的系统,其特征在于,还包括:二次规划智能体,

所述中控组件向所述二次规划智能体发送所述查询信息、检索到的候选知识和所述原因识别结果;

所述二次规划智能体根据所述查询信息、检索到的候选知识和所述原因识别结果,生成补充检索信息,并向所述中控组件返回所述补充检索信息;

所述中控组件向所述检索智能体发送所述补充检索信息;

所述检索智能体基于所述补充检索信息进行候选知识的补充检索,并向所述中控组件返回补充检索到的候选知识。

多智能体的智能检索引擎、人机交互系统和方法

技术领域

[0001] 本申请涉及计算机技术,尤其涉及一种多智能体的智能检索引擎、人机交互系统和方法。

背景技术

[0002] 随着人工智能的发展,大模型被广泛应用于自然语言处理领域的人机交互中。大模型是指具有大规模的模型参数的深度学习模型,例如大规模的语言模型、多模态模型等。

[0003] 在电商、工业、教育、金融、交通等领域的人机交互场景中,往往会有一些难度较高的复杂问题,例如多产品组合对比的问题、包含多个子问题的问题等。传统的基于机器学习的人机交互方案,大多采用模板匹配的方式,通过预先配置大量问题模板结合预置答案实现,很不灵活,泛化能力很差,很难理解复杂问题。

[0004] 大模型虽然具有强大的学习能力,目前基于大模型的人机交互方案中,基于查询信息在知识库中进行知识检索并生成答复,在面对这些复杂问题时,存在召回知识不全面(漏召回)、召回错误或召回失败的情况,导致答复质量差,无法给出让用户满意的答复。

发明内容

[0005] 本申请提供一种多智能体的智能检索引擎、人机交互系统和方法,用以解决现有的人机交互系统在面对复杂问题使答复质量差,无法给出用户满意答复的问题。

[0006] 第一方面,本申请提供一种多智能体的智能检索引擎,包括:中控组件和多个智能体,

[0007] 所述中控组件接收人机交互系统发送的查询信息;

[0008] 所述中控组件控制所述多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识;

[0009] 所述中控组件向所述人机交互系统返回检索到的候选知识,所述候选知识用于生成所述查询信息的答复信息。

[0010] 第二方面,本申请提供一种人机交互方法,包括:

[0011] 接收查询信息;根据所述查询信息,调用多智能体的智能检索引擎,通过所述智能检索引擎的中控组件控制所述多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知识;将所述查询信息和所述候选知识输入人机交互模型,通过所述人机交互模型根据所述候选知识生成所述查询信息的答复信息;输出所述答复信息。

[0012] 第三方面,本申请提供一种多智能体的人机交互系统,包括:中控组件、用于知识检索的多个智能体和用于生成答复的问答智能体,

[0013] 所述中控组件接收查询信息;

[0014] 所述中控组件控制用于知识检索的多个智能体协作完成对所述查询信息的如下检索处理:提取所述查询信息的关键信息,基于所述关键信息检索生成答复所需的候选知

识；

[0015] 所述中控组件向所述问答智能体发送所述查询信息和检索到的候选知识；

[0016] 所述问答智能体根据所述候选知识生成所述查询信息的答复信息，并向所述中控组件返回所述答复信息；

[0017] 所述中控组件输出所述答复信息。本申请提供的多智能体的智能检索引擎、人机交互系统和方法，应用于人机交互系统中，基于多智能体构建智能检索引擎，通过中控组件控制多智能体的协作提取查询信息的关键信息，并基于关键信息检索生成答复所需的候选知识，智能体借助人工智能模型对查询信息进行深度语义理解，可以精准、全面地提取查询信息的关键信息；智能体借助人工智能模型基于关键信息检索并召回生成答复所需的候选知识，可以更加全面和精准地召回生成答复所需的候选知识，减少候选知识的漏召回、召回错误或失败的情况发生，提升了召回候选知识的精准度，进而可以提升基于候选知识生成的答复信息的质量，得到让用户满意的答复。

附图说明

[0018] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

[0019] 图1为本申请所适用的一示例系统架构的示意图；

[0020] 图2为本申请所适用的另一示例系统架构的示意图；

[0021] 图3为本申请一示例性实施例提供的基于多智能体的智能检索引擎的框架图；

[0022] 图4为本申请一示例性实施例提供的基于多智能体的智能检索引擎的检索流程图；

[0023] 图5为本申请另一示例性实施例提供的多智能体的智能检索引擎的框架图；

[0024] 图6为本申请另一示例性实施例提供的基于多智能体的智能检索引擎的检索流程图；

[0025] 图7为本申请另一示例性实施例提供的多智能体的智能检索引擎的框架图；

[0026] 图8为本申请另一示例性实施例提供的多智能体的智能检索引擎的框架图；

[0027] 图9为本申请一示例性实施例提供的人机交互方法的流程图；

[0028] 图10为本申请一示例性实施例提供的多智能体的人机交互系统的流程图；

[0029] 图11为本申请实施例提供的一种服务器的结构示意图。

[0030] 通过上述附图，已示出本申请明确的实施例，后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本申请构思的范围，而是通过参考特定实施例为本领域技术人员说明本申请的概念。

具体实施方式

[0031] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[0032] 需要说明的是，本申请所涉及的用户信息(包括但不限于用户设备信息、用户属性

信息等)和数据(包括但不限于用于分析的数据、存储的数据、展示的数据等),均为经用户授权或者经过各方充分授权的信息和数据,并且相关数据的收集、使用和处理需要遵守相关法律法规和标准,并提供有相应的操作入口,供用户选择授权或者拒绝。

[0033] 首先对本申请所涉及的名词进行解释:

[0034] 预训练:指使用大规模无监督数据对模型进行训练,获得具有强大泛化能力的预训练模型的过程。

[0035] 微调训练:也即微调,是指使用特定任务的监督数据对预训练模型进行训练,使得预训练模型适应特定的任务和场景。

[0036] 提示模板(Prompt):也即Prompt工程,中文一般称为提示模板,指在大模型相关的指令设计中使用的一套精心设计的提问范式,流程上算法工程师通过意图理解等方式把用户的提问填充进其中,得到输入大模型的指令,再送入大模型。精妙设计Prompt往往能激发大模型更强的潜在能力,从而取得相比不使用Prompt更好的效果。

[0037] 思维链(Chain Of Thought,简称COT):广义上指的是一系列有关逻辑推理的思考步骤,最终形成一个完整的语义理解过程。在大模型中特指在Prompt工程中融入针对某类指令的语义理解过程,如应该分哪几个步骤进行回答,注意哪几个关键点。思维链如同做“因式分解”,把复杂的问题进行拆解,最终获得高质量的答案。

[0038] 多路检索:是指通过多种路径/方式进行检索,获得多路的检索结果。本实施例中,多路检索包括通过文本匹配和向量匹配等多种不同的检索方式分别进行检索,每一种检索方式的检索称为一路检索,最终获得多路检索结果。

[0039] 视觉问答任务:根据输入的图像和问题,从输入图像的视觉信息中确定问题的答案。

[0040] 图像描述任务:生成输入图像的描述文本。

[0041] 视觉蕴涵任务:预测输入图像和文本在语义上的相关性,即蕴涵、中性或矛盾。

[0042] 指代表达与理解任务:根据输入文本定位输入图像中与输入文本对应的图像区域。

[0043] 图像生成任务:基于输入的描述文本生成图像。

[0044] 基于文本的情感分类任务:预测输入文本的情感分类信息。

[0045] 文本摘要任务:生成输入文本的摘要信息。

[0046] 多模态任务:是指输入输出数据涉及图像和文本等多种模态数据的下游任务,例如视觉问答任务、图像描述任务、视觉蕴涵任务、指代表达与理解任务、图像生成任务等。

[0047] 多模态预训练模型:是指输入输出数据涉及图像和文本等多种模态数据的预训练模型,经过微调训练后可以应用于多模态任务处理。

[0048] 预训练语言模型:对大规模语言模型(Large Language Model,简称LLM)进行预训练后得到的预训练模型。

[0049] 组件(Component):简而言之组件就是对象,是对数据和方法的简单封装。

[0050] 智能体(Agent):是指具有智能的实体,由大脑、感知、行动三部分组成。作为智能体的控制器,智能体的大脑部分,承担记忆、思考和决策等基本任务,通常可以是人工智能模型(如LLM等大模型)、工具等。感知部分使得智能体能够感知和处理来自外部环境的多模态信息。行动部分使得智能体可以使用工具执行处理逻辑并输出结果(影响周围环境)。本

实施例中,智能体是封装了一定的处理逻辑、并可以调用人工智能模型和工具配合执行处理逻辑的软件实体。

[0051] 大模型是指具有大规模模型参数的深度学习模型,通常包含上亿、上百亿、甚至上千亿的模型参数。大模型又可以称为基石模型/基础模型(Foundation Model,简称FM),通过大规模无标注的语料进行大模型的预训练,产出亿级以上参数的预训练模型,这种模型能适应广泛的下游任务,模型具有较好的泛化能力,例如大规模语言模型(Large Language Model,简称LLM)、多模态预训练模型(Multi-modal Pre-training Model)等。

[0052] 大模型在实际应用时,仅需少量样本对预训练模型进行微调即可应用于不同的任务中,大模型可以广泛应用于自然语言处理(Natural Language Processing,简称NLP)、计算机视觉等领域,具体可以应用于如视觉问答(Visual Question Answering,简称VQA)、图像描述(Image Caption,简称IC)、图像生成等计算机视觉领域任务,以及基于文本的情感分类、文本摘要生成、机器翻译等自然语言处理领域任务,大模型主要的应用场景包括数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等。

[0053] 本申请实施例提供一种多智能体的智能检索引擎,包括中控组件和多个智能体,中控组件接收人机交互系统发送的交互需求信息和查询信息;中控组件控制多个智能体协作完成对查询信息的如下检索处理:基于查询信息和交互需求信息进行检索需求分析,在确定需要检索的情况下提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识。进一步地,中控组件向人机交互系统返回检索到的候选知识,候选知识用于生成查询信息的答复信息。

[0054] 本申请基于多智能体构建智能检索引擎,用于人机交互系统中,通过中控组件控制多智能体的协作,先提取查询信息的关键信息,再基于关键信息检索生成答复所需的候选知识,能够通过智能体借助人工智能模型对查询信息进行深度语义理解,可以精准、全面地提取查询信息的关键信息;进一步地,智能体借助人工智能模型基于关键信息检索并召回生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度,进而可以提升基于候选知识生成的答复信息的质量,得到让用户满意的答复。

[0055] 本申请提供一种基于前述多智能体的智能检索引擎实现的人机交互方法,应用于人机交互系统/设备,通过接收查询信息,并获取人机交互系统配置的交互需求信息;根据查询信息和交互需求信息,调用多智能体的智能检索引擎,通过智能检索引擎的中控组件控制多个智能体协作完成对查询信息的如下检索处理:基于查询信息和交互需求信息进行检索需求分析,在确定需要检索的情况下提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识;并返回检索到的候选知识;进一步地,将查询信息和候选知识输入人机交互模型,通过人机交互模型根据候选知识生成查询信息的答复信息,输出答复信息。

[0056] 本申请的人机交互方法,可以借助多智能体的智能检索引擎,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度,进而可以提升基于候选知识生成的答复信息的质量,得到让用户满意的答复。

[0057] 本申请还提供一种多智能体的人机交互系统,包括:中控组件、用于知识检索的多个智能体、用于生成答复的问答智能体;中控组件用于:接收查询信息,并获取人机交互系

统配置的交互需求信息;中控组件还用于控制用于知识检索的多个智能体协作完成对查询信息的如下检索处理:基于查询信息和交互需求信息进行检索需求分析,在确定需要检索的情况下提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识;中控组件还用于向问答智能体发送查询信息和检索到的候选知识;问答智能体用于:根据候选知识生成查询信息的答复信息,并向中控组件返回答复信息;中控组件还用于输出答复信息。

[0058] 基于多智能体构建人机交互系统,在中控组件的协调控制下,通过多智能体的协作,基于查询信息和交互需求信息进行检索需求分析,在确定需要检索的情况下提取查询信息的关键信息,并基于关键信息检索生成答复所需的候选知识,能够利用智能体中的人工智能模型自动规划是否进行检索候选知识的增强检索阶段,并在确定需要增强检索时,智能体借助人工智能模型对查询信息进行深度语义理解,可以精准、全面地提取查询信息的关键信息;进一步地,智能体借助人工智能模型基于关键信息检索并召回生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度;进一步地,智能体借助人工智能模型基于查询信息和检索到的候选知识生成答复信息,可以提升生成答复信息的质量,得到让用户满意的答复。

[0059] 其中,智能体的大脑部分使用的人工智能模型,可以是预训练的深度学习模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言理解(Natural Language Understanding,简称NLU)能力或自然语言处理(Natural Language Processing,简称NLP)能力的大模型。另外,智能体的大脑部分使用的人工智能模型,可以是针对智能体所需执行的处理逻辑,对预训练模型进行微调后得到的模型。

[0060] 需要说明的是,不同智能体中充当大脑部分的人工智能模型可以使用同一个具有强大能力的通用人工智能模型。但是由于不同智能体的功能不同,所需执行的处理逻辑不同,因此不同智能体中充当大脑部分的人工智能模型,也可以使用不同的模型。例如,分别针对不同智能体所需执行的任务,对基础大模型进行微调,获得各智能体的人工智能模型。

[0061] 图1为本申请所适用的一示例系统架构的示意图。如图1所示,该系统架构包括端侧设备、第一服务器和智能检索引擎。其中,第一服务器与端侧设备之间具有可通信的通信链路,能够实现第一服务器与端侧设备间的通信连接。第一服务器与智能检索引擎之间具有可通信的通信链路,能够实现第一服务器与智能检索引擎间的通信连接。

[0062] 其中,端侧设备是指使用人机交互系统的用户所使用的电子设备,具体可以为具有网络通信功能、运算功能以及信息显示功能的硬件设备,其包括但不限于智能手机、平板电脑、台式电脑、本地服务器、云端服务器等。用户通过端侧设备向第一服务器发送查询信息。

[0063] 第一服务器是实现人机交互功能的设备,具体可以是部署在云端或本地的具有计算能力的设备,例如各类场景下的数字助理、智能机器人等各类人机对话场景下的人机交互系统。第一服务器运行有人机交互系统,负责接收查询信息和人机交互系统配置的交互需求信息,向智能检索引擎发送查询信息和交互需求信息,使得智能检索引擎检索并返回生成答复所需的候选知识。

[0064] 智能检索引擎是基于多智能体构建的智能检索引擎,可以部署在人机交互系统本地或云端的另一服务器上。在一些场景中,智能检索引擎还可以与人机交互系统一起部署

在第一服务器上。智能检索引擎负责基于人机交互系统的第一服务器提供的查询信息和交互需求信息,通过多个智能体协作完成对查询信息的增强检索,检索到生成答复所需的候选知识,并向第一服务器返回检索到的候选知识。

[0065] 第一服务器还负责根据智能检索引擎返回的候选知识生成查询信息的答复信息,并向端侧设备返回答复信息。

[0066] 端侧设备还负责接收第一服务器返回的答复信息,并向用户输出答复信息,实现人机交互。

[0067] 图2为本申请所适用的另一示例系统架构的示意图。如图2所示,该系统架构包括端侧设备和第二服务器。其中,第二服务器与端侧设备之间具有可通信的通信链路,能够实现第二服务器与端侧设备间的通信连接。

[0068] 其中,端侧设备是指使用人机交互系统的用户所使用的电子设备,具体可以为具有网络通信功能、运算功能以及信息显示功能的硬件设备,其包括但不限于智能手机、平板电脑、台式电脑、本地服务器、云端服务器等。用户通过端侧设备向第二服务器发送查询信息。

[0069] 第二服务器是实现人机交互功能的设备,具体可以是部署在云端或本地的具有计算能力的设备,例如各类场景下的数字助理、智能机器人等各类人机对话场景下的人机交互系统。第二服务器运行有多智能体的人机交互系统,负责接收查询信息,通过多智能体协作完成对查询信息的关键信息提取,并基于关键信息检索到生成答复所需的候选知识,根据候选知识和查询信息生成答复信息。第二服务器向端侧设备返回答复信息。

[0070] 端侧设备还负责接收第二服务器返回的答复信息,并向用户输出答复信息,实现人机交互。

[0071] 下面以具体地实施例对本申请的技术方案以及本申请的技术方案如何解决上述技术问题进行详细说明。下面这几个具体的实施例可以相互结合,对于相同或相似的概念或过程可能在某些实施例中不再赘述。下面将结合附图,对本申请的实施例进行描述。

[0072] 图3为本申请一示例性实施例提供的基于多智能体的智能检索引擎的框架图;图4为本申请一示例性实施例提供的基于多智能体的智能检索引擎的检索流程图。如图3所示,本实施例提供的多智能体的智能检索引擎包括中控组件和多个智能体。基于多智能体的智能检索引擎可以用于人机交互系统,实现对用户输入的查询信息的增强检索,获得用于生成答复信息的候选知识。如图4所示,基于多智能体的智能检索引擎进行智能检索的步骤如下:

[0073] 步骤S401、中控组件接收人机交互系统发送的查询信息。

[0074] 本实施例中,中控组件负责与人机交互系统间的交互,接收用户通过端侧设备向人机交互系统提交的查询信息(query)。人机交互系统将查询信息发送至智能检索引擎。

[0075] 其中,查询信息可以是各类复杂问题,包括但不限于多目标对比类问题,包含多个子问题的问题,多轮交互中包含指代、省略信息的问题。现有的基于单一人工智能模型的人机交互方案往往无法深度理解复杂问题的语义信息,召回能力差,在面对这些复杂问题时,存在召回知识不全面(漏召回)、召回错误或召回失败的情况,导致答复质量差,无法给出让用户满意的答复。

[0076] 在实际的各类应用场景中,尤其是电商领域智能客服场景中,产品的对比问答是

非常常见的一类复杂问题,这类的查询信息通过较难理解,比如产品的别名模糊识别、多产品属性对比的语义准确理解等,这些都是传统机器学习算法不容易解决的问题,即使是在大模型时代,多产品对比的复杂问题的知识检索也是个难题,目前的方法无法存在漏召回的问题,知识召回不全面直接决定了基于召回知识生成的答复无法较好解决用户问题,生成答复信息的质量低。

[0077] 步骤S402、中控组件控制多个智能体协作完成对查询信息的如下检索处理:提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识。

[0078] 本实施例中,通过中控组件控制多个智能体,通过多个智能体协作完成对查询信息的检索处理。具体地,通过智能体利用人工智能模型对查询信息进行深度语义理解,来提取查询信息的关键信息,可以精准、全面地提取查询信息的关键信息;进一步地,通过智能体利用人工智能模型基于关键信息检索生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度。

[0079] 步骤S403、中控组件向人机交互系统返回检索到的候选知识,候选知识用于生成查询信息的答复信息。

[0080] 在通过智能检索获得生成查询信息的答复信息所需的候选知识之后,中控组件将通过多智能体检索到的候选知识返回至人机交互系统。人机交互系统接收到候选知识后,根据候选知识生成查询信息的答复信息。

[0081] 本实施例中,基于多智能体构建智能检索引擎,基于人机交互系统发送的查询信息,通过智能体利用人工智能模型对查询信息进行深度语义理解,来提取查询信息的关键信息,可以精准、全面地提取查询信息的关键信息;进一步地,通过智能体利用人工智能模型基于关键信息检索生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度,进而可以提升基于候选知识生成的答复信息的质量,得到让用户满意的答复。

[0082] 图5为本申请一示例性实施例提供的多智能体的智能检索引擎的框架图。多智能体的智能检索引擎包括中控组件和多个智能体。如图5所示,本实施例中,智能检索引擎可以包括信息抽取智能体和检索智能体。其中,信息抽取智能体用于提取查询信息的关键信息,并向中控组件返回关键信息。检索智能体用于基于关键信息检索生成答复所需的候选知识,并向中控组件返回检索到的候选知识。

[0083] 图6为本实施例提供的基于多智能体的智能检索引擎的检索流程图。基于图5所示框架的智能检索引擎,如图6所示,前述步骤S402具体可以采用如下步骤实现:

[0084] 步骤S4021、中控组件向信息抽取智能体发送查询信息。

[0085] 步骤S4022、信息抽取智能体提取查询信息的关键信息,并向中控组件返回关键信息。

[0086] 其中,关键信息是指查询信息中与用户提问意图相关的信息,是查询信息中的重要信息,决定待生成的答复信息中要包含哪些方面的内容。

[0087] 本实施例中,信息抽取智能体的大脑部分可以使用预先训练的语义理解模型实现。信息抽取智能体使用预先训练的语义理解模型,对查询信息进行深度地语义理解,并精准地提取查询信息的关键信息。

[0088] 其中,语义理解模型是预先训练获得的深度学习模型,通过使用第一数据集(包括查询信息和标注的关键信息)对基础模型进行微调训练获得。语义理解模型专门用于问题的深度语义理解,并精准地提取问题中的关键信息。

[0089] 训练语义理解模型使用的基础模型可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)或其他的具有强大自然语言理解能力的大模型,此处不做具体限定。经过微调训练后的语义理解模型,具有对问题进行深度语义理解的强大能力,能够精准地提取查询信息的关键信息。

[0090] 在一可选实施例中,该步骤具体可以采用如下方式实现:通过语义理解模型识别查询信息的用户意图,并基于用户意图提取查询信息的关键信息。

[0091] 语义理解模型可以识别查询信息的用户意图,并基于用户意图,针对不同类型用户意图的查询信息,按照不同的语义理解过程进行分析和推理,以实现对查询信息的语义理解,并提取查询信息的关键信息。

[0092] 其中,不同类型的用户意图对应的语义理解过程不同,语义理解过程中定义了语义理解的流程和步骤,将对查询信息的语义理解并提取关键信息的过程按流程拆解,并强调关注哪些方面,可以激发语义理解模型对这类用户意图的查询信息的深度语义理解能力。

[0093] 具体地,对于多目标对比类的查询信息,通过语义理解模型识别查询信息的多目标对比意图,基于多目标对比意图,识别查询信息包含的多个对比目标和待对比的对比属性,并提取查询信息包含的多个对比目标和对比属性的信息。

[0094] 对于包含多个子问题类的查询信息,通过语义理解模型识别查询信息的多子问题意图,拆分出查询信息包含的多个子问题,分别提取多个子问题的关键信息。

[0095] 对于多轮交互中的查询信息,通过语义理解模型识别查询信息的多轮交互意图,获取查询信息的上下文信息,并提取查询信息及上下文信息的关键信息。或者,通过语义理解模型识别查询信息的多轮交互意图,获取查询信息的上下文信息,根据查询信息的上下文信息改写查询信息得到改写问题,提取改写问题的关键信息。

[0096] 在语义理解模型的微调训练阶段,使用包含思维链CoT的提示模板(Prompt)和第一数据集(包括查询信息和标注的关键信息)进行微调训练。具体地,该思维链CoT包含语义理解的总体流程,以及不同类型的用户意图对应的语义理解过程。例如,思维链CoT中定义了语义理解的总体流程为:1.意图识别:识别查询信息的用户意图;2.理解推理:基于用户意图结合不同意图的语义理解过程,对查询信息进行分析和推理,得到查询信息包含的关键信息;3.输出结果:输出上一步得到的关键信息。

[0097] 示例性地,对于多目标对比意图的查询信息,其语义理解过程可以包括:识别查询信息中待对比的多个目标、对比属性,从查询信息中抽取多个目标的标识信息和对比属性的信息,作为关键信息。对于这类查询信息,语义理解模型基于思维链CoT的语义理解过程包括:识别出查询信息的多目标对比意图,基于多目标对比意图识别查询信息包含的多个对比目标和待对比的对比属性,并提取查询信息包含的多个对比目标和对比属性的信息。其中,对于多目标对比意图的查询信息,查询信息中用户指定了将多个目标在特定属性方面进行对比。对比属性是指待对比的特定属性,具体可以是待对比目标的任意一种属性。所提取的信息可以包括查询信息中出现的多个对比目标的标识信息和对比属性的标识信息。

其中对比目标的标识信息可以包括对比目标的名称、标识码等标识性信息,可以确定对应的目标。对比属性的标识信息可以对比属性的名称等,可以唯一标识一种属性。

[0098] 例如,对于多个产品对比的查询信息,语义理解模型基于思维链CoT的语义理解过程包括:识别出查询信息的多产品对比意图,识别查询信息包含的多个产品、对比属性,从查询信息中提取多个产品的标识信息和对比属性的信息。其中,产品的标识信息可以是查询信息中的产品名称、产品识别码等信息。对比属性的信息可以是产品的任意一种属性的名称,如商品的产地、保质期、性能参数等。

[0099] 示例性地,对于包含多个子问题的查询信息,其语义理解过程可以包括:识别查询信息包含的子问题的数量,拆分出查询信息包含的多个子问题,分别提取每个子问题的关键信息。对于这类查询信息,语义理解模型基于思维链CoT的语义理解过程包括:识别出查询信息的多子问题意图,基于多子问题意图识别查询信息包含的子问题的数量,拆分出查询信息包含的多个子问题,分别提取每个子问题的关键信息。

[0100] 示例性地,对于多轮交互中的查询信息,除了用户首轮提出的查询信息,后续各轮查询信息具有上下文信息(通常包含之前各轮的查询信息和答复)。这类查询信息的语义理解过程可以包括:获取查询信息的上下文信息,并提取查询信息及上下文信息的关键信息。对于这类查询信息,语义理解模型基于思维链CoT的语义理解过程包括:识别查询信息的多轮交互意图,获取查询信息的上下文信息,并提取查询信息及上下文信息的关键信息。可选地,这类查询信息的语义理解过程还可以包括:获取查询信息的上下文信息,根据查询信息的上下文信息改写查询信息得到改写问题,提取改写问题的关键信息。对于这类查询信息,语义理解模型基于思维链CoT的语义理解过程包括:识别查询信息的多轮交互意图,获取查询信息的上下文信息,根据查询信息的上下文信息改写查询信息得到改写问题,提取改写问题的关键信息。

[0101] 本实施例中,语义理解模型的微调训练阶段使用的提示模板(Prompt)中的思维链CoT,可以由相关技术人员根据实际应用需求进行配置和调整,此处不做具体限定。

[0102] 对语义理解模型进行微调训练时,将第一数据集中的查询信息插入提示模板中,获得包含查询信息和思维链CoT的提示信息,将提示信息输入语义理解模型,通过语义理解模型按照思维链CoT对查询信息进行语义理解,提取查询信息的关键信息,并输出查询信息的关键信息。基于语义理解模型输出的关键信息和第一数据集中预先标注的关键信息,计算损失函数值,并根据损失函数值更新语义理解模型的参数。本实施例中,对语义理解模型进行微调训练的具体实现方式,与现有的对预训练模型进行基于包含思维链CoT的提示模板(Prompt)的监督微调过程类似,此处不再赘述。

[0103] 本实施例中,通过使用包含思维链CoT的提示模板微调训练获得语义理解模型,将思维链CoT中对各类复杂问题进行深度语义理解的分析和推理能力赋予语义理解模型,使得语义理解模型具有对各类复杂问题进行深度语义理解并精准地提取关键信息的能力。基于微调训练得到的语义理解大模型对查询信息进行语义理解并提取关键信息,在面对各类复杂问题时,可以实现对各类复杂问题的深度语义理解,并精准地提取查询信息包含的关键信息。

[0104] 步骤S4023、中控组件向检索智能体发送关键信息。

[0105] 在接收到查询信息的关键信息之后,中控组件向检索智能体发送关键信息,以触

发检索智能体进行知识检索。

[0106] 步骤S4024、检索智能体基于关键信息检索生成答复所需的候选知识，并向中控组件返回检索到的候选知识。

[0107] 本实施例中，检索智能体的大脑可以为检索工具或搜索引擎。检索智能体通过调用检索工具或搜索引擎，将查询信息的关键信息与知识库中的知识进行相似度匹配，在知识库中检索到与关键信息匹配的候选知识。

[0108] 示例性地，若查询信息的关键信息包含多项，可以将多项关键信息进行拼接，将拼接结果与知识库中的知识进行相似度匹配，在知识库中检索到与关键信息的拼接结果相匹配的候选知识。

[0109] 在一示例场景中，例如电商领域，用户往往需要将多个产品进行对比，也即输入的查询信息是多产品对比类的问题。用户通过端侧设备向人机交互系统提交查询信息，人机交互系统向智能检索引擎发送查询信息。智能检索引擎的中控组件接收人机交互系统发送的查询信息。中控组件向信息抽取智能体发送查询信息。

[0110] 信息抽取智能体，通过语义理解模型对查询信息进行语义理解，并提取查询信息的关键信息，该关键信息包括：待对比的多个产品和对比属性的信息。信息抽取智能体向中控组件返回查询信息的关键信息。

[0111] 中控组件向检索智能体发送查询信息的关键信息。检索智能体根据查询信息的关键信息(待对比的多个产品和对比属性的信息)，在知识库中检索与关键信息匹配的候选知识(包括各个产品的对比属性的具体属性信息)。检索智能体向中控组件返回与关键信息匹配的候选知识。

[0112] 进一步地，中控组件将查询信息和与关键信息匹配的候选知识返回至人机交互系统。人机交互系统根据通过问答模型根据关键信息匹配的候选知识生成查询信息的答复信息。人机交互系统将查询信息的答复信息返回给端侧设备。端侧设备通过用户交互界面向用户输出答复信息。

[0113] 在一示例场景中，用户一次输入的查询信息包括多个问题，其中每一问题称为查询信息的子问题，也即查询信息包含多个子问题。人机交互系统接收端侧设备发送的查询信息，向智能检索引擎发送查询信息。智能检索引擎的中控组件接收人机交互系统发送的查询信息。中控组件向信息抽取智能体发送查询信息。

[0114] 信息抽取智能体，通过语义理解模型对查询信息进行语义理解，并提取查询信息的关键信息，该关键信息包括：每个子问题的关键信息。信息抽取智能体向中控组件返回查询信息的关键信息。

[0115] 中控组件向检索智能体发送查询信息的关键信息。检索智能体根据查询信息的关键信息(包括每个子问题的关键信息)，在知识库中检索与关键信息匹配的候选知识(包括与多个子问题的关键信息匹配的候选知识)。检索智能体向中控组件返回与关键信息匹配的候选知识。

[0116] 进一步地，中控组件将查询信息(包含多个子问题)和与关键信息匹配的候选知识(包括与多个子问题的关键信息匹配的候选知识)返回至人机交互系统。人机交互系统，通过问答模型根据关键信息匹配的候选知识生成查询信息的答复信息，该答复信息包含各个子问题的答复。进一步地，人机交互系统将查询信息的答复信息(包含各个子问题的答复)

返回给端侧设备。端侧设备通过用户交互界面向用户输出答复信息(包含各个子问题的答复)。

[0117] 在一可选实施例中,知识库中存储有知识文本。该步骤中,第一服务器将查询信息的关键信息的文本与知识库中的知识文本进行文本相似度匹配,根据关键信息的文本与知识库中的知识文本间的文本相似度,确定与关键信息的文本相似度较高的知识文本,作为与关键信息匹配的候选知识。

[0118] 可选地,根据关键信息的文本与知识库中的知识文本间的文本相似度,从知识库中召回与关键信息文本的文本相似度大于或等于文本相似度阈值的知识文本,得到与关键信息匹配的候选知识。其中,文本相似度阈值可以根据实际应用场景和经验值进行设置和调整,此处不做具体限定。可选地,还可以根据关键信息的文本与知识库中的知识文本间的文本相似度,召回第一预设数量的候选知识。其中,第一预设数量为正整数,第一预设数量的值可以根据实际应用场景和经验值进行设置和调整,例如,第一预设数量可以为1、3、5等,此处不做具体限定。

[0119] 其中,关键信息的文本与知识库中的知识文本间的文本相似度,具体可以采用基于词频的文本匹配算法实现,例如基于词频-逆文件频率(Term Frequency-Inverse Document Frequency,简称TF-IDF)的文本匹配算法、或者BM25算法实现。TF-IDF算法和BM25算法是较主流的计算用户查询信息(query)与文档相似度得分的算法。

[0120] 在一可选实施例中,知识库中存储有知识文本的向量表示。该步骤中,第一服务器获取查询信息的关键信息的向量表示,将查询信息的关键信息的向量表示与知识库中知识文本的向量表示进行向量相似度匹配,计算关键信息的向量表示与知识库中的知识文本的向量表示间的相似度,作为关键信息与知识库中的知识文本间的向量相似度,确定与关键信息的向量相似度较高的知识文本,作为与关键信息匹配的候选知识。

[0121] 其中,知识库中知识文本的向量表示,可以使用预先训练的文本表征模型表征获得。使用预先训练的文本表征模型,将知识库中的各个知识文本表征为向量表示,并存储到知识库中。获取查询信息的关键信息的向量表示时,使用预先训练的文本表征模型表征,将查询信息的关键信息表征为向量表示。可选地,还可以基于知识库中的各个知识文本的向量表示,构建另一知识向量库。该知识向量库包括知识文本的标识和向量表示。该步骤中,将查询信息的关键信息的向量表示与所构建的知识向量库中的向量表示进行向量相似度匹配,确定与关键信息匹配的候选知识。

[0122] 本实施例中,文本表征模型是预先训练获得的深度学习模型,通过使用微调数据集对基础模型进行微调训练获得。文本表征模型专门用于将输入文本表征为向量表示。其中,训练文本表征模型使用的基础模型可以是各类预训练大模型、文本表征模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大文本表征能力的模型,此处不做具体限定。

[0123] 可选地,根据关键信息与知识库中的知识文本间的向量相似度,召回向量相似度大于或等于向量相似度阈值的知识文本,得到与关键信息匹配的候选知识。其中,向量相似度阈值可以根据实际应用场景和经验值进行设置和调整,此处不做具体限定。可选地,还可以根据关键信息与知识库中的知识文本间的向量相似度,召回第二预设数量的候选知识。其中,第二预设数量为正整数,第二预设数量的值可以根据实际应用场景和经验值进行设

置和调整,例如,第二预设数量可以为1、3、5等,此处不做具体限定。

[0124] 其中,关键信息与知识库中的知识文本间的向量相似度,可以是关键信息的向量表示与知识文本的向量表示间的余弦相似度。另外,关键信息与知识库中的知识文本间的向量相似度,还可以是其他常用于衡量两个文本的向量表示间的语义相关性的其他的向量间相似度或距离指标,如欧式距离、曼哈顿距离等,本实施例此处不做具体限定。

[0125] 在一优选实施例中,知识库中既存储有知识文本,也存储有知识文本的向量表示。该步骤中,获取查询信息的关键信息的向量表示,基于关键信息的文本和向量表示,在知识库中进行多路检索,获得与关键信息匹配的候选知识。

[0126] 具体地,第一服务器基于关键信息的文本进行第一路检索,并基于关键信息的向量表示进行第二路检索。第一路检索的过程如下:第一服务器将查询信息的关键信息的文本与知识库中的知识文本进行文本相似度匹配,根据关键信息的文本与知识库中的知识文本间的文本相似度,确定与关键信息的文本相似度较高的知识文本,作为与关键信息匹配的候选知识。第一路检索的具体实现方式参见前述实施例中基于文本相似度进行检索的相关内容,此处不再赘述。第二路检索的过程如下:第一服务器将查询信息的关键信息的向量表示与知识库中知识文本的向量表示进行向量相似度匹配,计算关键信息的向量表示与知识库中的知识文本的向量表示间的相似度,作为关键信息与知识库中的知识文本间的向量相似度,确定与关键信息的向量相似度较高的知识文本,作为与关键信息匹配的候选知识。第二路检索的具体实现方式参见前述实施例中基于向量相似度进行检索的相关内容,此处不再赘述。

[0127] 本实施例中,在知识库中进行多路检索之前,首先通过预先训练的文本表征模型,将知识库中的知识表征为向量表示;将知识库中知识的向量表示存储到知识库中。可选地,还可以基于知识库中的各个知识文本的向量表示,构建另一知识向量库。该知识向量库包括知识文本的标识和向量表示。该步骤中,将查询信息的关键信息的向量表示与所构建的知识向量库中的向量表示进行向量相似度匹配,确定与关键信息匹配的候选知识。

[0128] 相应地,在获取查询信息的关键信息的向量表示时,使用预先训练的文本表征模型表征,将查询信息的关键信息表征为向量表示。

[0129] 其中,文本表征模型是预先训练获得的深度学习模型,通过使用微调数据集对基础模型进行微调训练获得。文本表征模型专门用于将输入文本表征为向量表示。其中,训练文本表征模型使用的基础模型可以是各类预训练大模型、文本表征模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大文本表征能力的模型,此处不做具体限定。

[0130] 本实施例中,基于关键信息的文本和向量表示,在知识库中进行多路检索,获得与关键信息匹配的候选知识,可以提升候选知识召回能力,更加全面地召回与查询信息的关键信息匹配的候选知识,减少了召回知识不全面(漏召回)的情况,使得送入问答大模型的候选知识更加完整全面,从而提升问答大模型生成答复信息的质量。

[0131] 图7为本申请另一示例性实施例提供的多智能体的智能检索引擎的框架图,在图5所示框架的基础上,如图7所示,智能检索引擎可以包括需求分析智能体。

[0132] 本实施例中,中控组件可以获取人机交互系统的交互需求信息,向需求分析智能体发送查询信息和交互需求信息。其中,交互需求信息是通过人机交互系统配置的信息,包含对人机交互过程中所需遵循的一些规则、要求等。例如,交互需求信息可以包括判断是否

需要检索所依据的规则(称为检索规则)、检索使用的数据源、补充检索的次数限制等等。通过交互需求信息可以灵活地对检索过程进行干预和控制,使得方案更具扩展性。

[0133] 需求分析智能体基于查询信息和交互需求信息进行检索需求分析,确定答复查询信息是否需要进行知识检索,得到检索需求分析结果,并向中控组件返回检索需求分析结果。

[0134] 本实施例中,需求分析智能体的大脑部分可以使用人工智能模型(如LLM等),通过人工智能模型基于人机交互系统配置的交互需求信息,进行是否需要后续检索处理流程的自动规划和决策。检索需求分析结果至少包括针对查询信息是否需要进行知识检索的判断结果。

[0135] 其中,需求分析智能体使用的人工智能模型(如LLM等),可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言处理(NLP)能力的大模型;也可以是针对此处的需求分析任务进行有针对性的微调训练获得的,专门用于进行需求分析的人工智能模型。

[0136] 中控组件接收需求分析智能体返回的检索需求分析结果,根据检索需求分析结果在确定需要检索的情况下,向信息抽取智能体发送查询信息,以触发后续的检索处理流程。另外,根据检索需求分析结果,在确定不需要检索的情况下,中控组件则不会触发后续的检索处理流程。

[0137] 需要说明的是,智能检索引擎中的中控组件可以是预先开发的程序模块,也可以使用智能体实现。在中控组件的处理逻辑是清晰且固定的处理逻辑的场景中,中控组件可以是预先开发的执行对应处理逻辑的程序模块。在中控组件的处理逻辑不固定,需要智能地基于外部信息进行规划和决策的场景中,中控组件可以使用智能体实现。

[0138] 在一可选实施例中,交互需求信息包括检索规则。该检索规则中说明了判断针对查询信息是否需要进行知识检索的规则。示例性地,检索规则中可以指定何种查询信息需要进行知识检索、和/或、何种查询信息不需要进行知识检索。例如,对于闲聊问候的查询信息(如“你好”、“在吗”),不需要进行知识检索,按照配置的话术模板进行答复即可。

[0139] 在接收到中控组件发送的查询信息之后,需求分析智能体可以使用人工智能模型,基于检索规则,判断答复查询信息是否需要进行知识检索,得到判断结果;在判断结果为需要进行知识检索的情况下,向中控组件返回包含判断结果的检索需求分析结果;在判断结果为不需要进行知识检索的情况下,确定查询信息对应的答复话术信息,向中控组件返回包含判断结果和答复话术信息的检索需求分析结果。

[0140] 其中,需求分析智能体使用的人工智能模型(如LLM等)被训练为具有基于检索规则判断针对查询信息是否进行知识检索能力,以及对于不需要进行知识检索的查询信息生成答复话术信息的能力。

[0141] 在一可选实施例中,交互需求信息还包括配置的数据源。该数据源是通过人机交互系统配置的后续进行知识检索所使用的知识库,包括但不限于人机交互系统自己构建的知识库、第三方知识库、开放域或所属领域的公开知识库等。

[0142] 需求分析智能体在判断结果为需要进行知识检索的情况下,向中控组件返回包含判断结果的检索需求分析结果时,向中控组件返回包含判断结果和配置的数据源的检索需求分析结果。进一步地,中控组件根据检索需求分析结果,在确定需要检索的情况下向信息

抽取智能体发送查询信息和数据源。进一步地,检索智能体接收到中控组件发送的查询信息和数据源之后,基于关键信息,在配置的数据源中检索生成答复所需的候选知识。

[0143] 本实施例的方法,支持人机交互系统灵活地配置自己进行知识检索使用的数据源(知识库)。

[0144] 在一可选实施例中,在图5所示框架的基础上,智能检索引擎可以包括引用监测智能体。引用监测智能体用于根据查询信息和检索到的候选知识,判断检索到的候选知识是否满足查询信息的答复需求,得到判断结果,并向中控组件返回判断结果。

[0145] 本实施例中,在向人机交互系统返回检索到的候选知识之前,中控组件向引用监测智能体发送查询信息和检索到的候选知识。

[0146] 引用监测智能体根据查询信息和检索到的候选知识,判断检索到的候选知识是否满足查询信息的答复需求,得到判断结果,并向中控组件返回判断结果。

[0147] 其中,引用监测智能体的大脑部分可以使用人工智能模型(如LLM等)。引用监测智能体利用人工智能模型对检索到的候选知识的有效性和完整性进行分析,判断检索到的候选知识是否满足查询信息的答复需求,也即检索到的候选知识是否足以覆盖查询信息中所有意图,能否完全回答查询信息包含的所有问题,并给出判断结果。

[0148] 引用监测智能体使用的人工智能模型(如LLM等),可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言处理(NLP)能力的大模型;也可以是针对此处的判断检索到的候选知识是否满足查询信息的答复需求的监测任务进行有针对性的微调训练获得的,专门用于进行判断检索到的候选知识是否满足查询信息的答复需求。

[0149] 中控组件根据判断结果,在检索到的候选知识满足答复需求的情况下,向人机交互系统返回检索到的候选知识。

[0150] 可选地,在检索到的候选知识不满足答复需求的情况下,中控组件可以向人机交互系统返回已检索到的候选知识,并向人机交互系统发送候选知识可能不满足查询信息的答复需求的提示。人机交互模型可以根据检索到的候选知识生成答复信息;或者,在候选知识可能不满足查询信息的答复需求的情况下,拒绝生成答复信息。

[0151] 在一可选实施例中,智能检索引擎还可以包括问题识别智能体。问题识别智能体用于在判断结果为检索到的候选知识不满足答复需求的情况下,根据查询信息和检索到的候选知识,识别检索到的候选知识不能满足答复需求的原因,得到原因识别结果,并向中控组件返回原因识别结果。

[0152] 本实施例中,在接收到引用监测智能体返回的判断结果之后,中控组件根据判断结果,在检索到的候选知识不满足答复需求的情况下,向问题识别智能体发送判断结果、查询信息和检索到的候选知识。

[0153] 问题识别智能体在确定判断结果为检索到的候选知识不满足答复需求的情况下,根据查询信息和检索到的候选知识,识别检索到的候选知识不能满足答复需求的原因,得到原因识别结果,并向中控组件返回原因识别结果。

[0154] 其中,问题识别智能体的大脑部分可以使用人工智能模型(如LLM等)。问题识别智能体利用人工智能模型根据查询信息和检索到的候选知识,识别检索到的候选知识不能满足答复需求的原因,得到原因识别结果。问题识别智能体使用的人工智能模型(如LLM等),

可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言处理(NLP)能力的大模型;也可以是针对此处的识别检索到的候选知识不能满足答复需求的原因的任务进行有针对性的微调训练获得的,专门用于识别检索到的候选知识不能满足答复需求的原因。

[0155] 示例性地,问题识别智能体使用的人工智能模型,可以在检索到的候选知识不满足答复需求的情况下,根据查询信息和检索到的候选知识,详细评估候选知识不满足答复需求的问题分类,包括但不限于冗余知识、完整性不足、语义模糊等类别。

[0156] 其中,冗余知识是指答复查询信息用不到的候选知识。语义模糊是指候选知识本身语义表达不清楚,无法作为生成答复使用的知识。冗余知识和语义模糊类别的候选知识是对回答查询信息无效的知识。完整性不足是指候选知识回答了查询信息的部分问题,但不足以生成能完整回答查询信息的答复信息。

[0157] 需要说明的是,完整性不足的候选知识可能有多条,这种情况下,表示这多条候选知识的集合也不足以生成能完整回答查询信息的答复信息。

[0158] 具体地,问题识别智能体使用的人工智能模型,识别每一条候选知识对应的类别,获得原因识别结果。

[0159] 进一步地,中控组件接收问题识别智能体返回的原因识别结果。根据原有识别结果中每一条候选知识对应的类别,过滤掉对于回答查询信息无效的知识,如冗余知识或语义模糊等类别的候选知识,仅保留对于回答查询信息有效的候选知识,如完整性不足等类别的候选知识。

[0160] 可选地,中控组件向人机交互系统返回候选知识时,仅向人机交互系统返回经过滤后的对回答查询信息有效的候选知识。可选地,中控组件可以向人机交互系统返回检索到的所有候选知识以及原因识别结果,以向人机交互系统提供生成答复信息时选择使用候选知识提供参考,或拒绝生成答复信息。

[0161] 本实施例的方案,可以在检索到的候选知识不满足答复需求的情况下,根据查询信息和检索到的候选知识,识别检索到的候选知识不能满足答复需求的原因,过滤掉对回答查询信息无效的候选知识,仅保留对回答查询信息有效的候选知识,可以避免无效的候选知识向生成答复引入噪声,避免无效的候选知识影响生成答复的质量,可以提升生成答复信息的质量。

[0162] 在一可选实施例中,智能检索引擎还可以包括二次规划智能体。二次规划智能体用于根据查询信息、检索到的候选知识和原因识别结果,生成补充检索信息,并向中控组件返回补充检索信息。其中,补充检索信息用于检索生成答复所缺失的候选知识。

[0163] 本实施例中,在接收到问题识别智能体返回的原因识别结果之后,中控组件向二次规划智能体发送查询信息、检索到的候选知识和原因识别结果。二次规划智能体根据查询信息、检索到的候选知识和原因识别结果,生成补充检索信息,并向中控组件返回补充检索信息。其中,中控组件向二次规划智能体发送的检索到的候选知识,可以是已经检索到的全部候选知识,也可以是基于原始识别结果过滤后的对回答查询信息有效的候选知识。

[0164] 其中,二次规划智能体的大脑部分可以使用人工智能模型(如LLM等)。二次规划智能体利用人工智能模型,根据查询信息、检索到的候选知识和原因识别结果,分析答复查询信息还缺失哪些信息,并基于所缺失的信息生成补充检索信息。该补充检索信息可以是一

个新的查询信息,基于该新的查询信息进行检索可以获得生成答复所缺失的候选知识。

[0165] 二次规划智能体使用的人工智能模型(如LLM等),可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言处理(NLP)能力的大模型;也可以是针对此处根据查询信息、检索到的候选知识和原因识别结果,分析答复查询信息还缺失哪些信息,并基于所缺失的信息生成补充检索信息的二次规划任务进行有针对性的微调训练获得的,专门用于根据查询信息、检索到的候选知识和原因识别结果,分析答复查询信息还缺失哪些信息,并基于所缺失的信息生成补充检索信息。

[0166] 进一步地,中控组件向检索智能体发送补充检索信息。检索智能体基于补充检索信息进行候选知识的补充检索,并向中控组件返回补充检索到的候选知识。

[0167] 在接收到补充检索到的候选知识之后,中控组件可以基于当前已经检索(包括第一次检索和补充检索)到的候选知识,向引用监测智能体发送查询信息和当前检索到的候选知识。引用监测智能体判断当前检索到的候选知识是否满足查询信息的答复需求,得到新的判断结果。

[0168] 进一步地,中控组件根据新的判断结果,在当前检索到的候选知识不满足答复需求的情况下,向问题识别智能体发送新的判断结果、查询信息和当前检索到的候选知识。问题识别智能体在新的判断结果为当前检索到的候选知识不满足答复需求的情况下,根据查询信息和当前检索到的候选知识,识别当前检索到的候选知识不能满足答复需求的原因,得到新的原因识别结果。

[0169] 进一步地,中控组件向二次规划智能体发送查询信息、当前检索到的候选知识和新的原因识别结果。二次规划智能体根据查询信息、当前检索到的候选知识和新的原因识别结果,生成新的补充检索信息。

[0170] 进一步地,中控组件向检索智能体发送新的补充检索信息。检索智能体基于新的补充检索信息进行候选知识的第二次补充检索,并向中控组件返回第二次补充检索到的候选知识。

[0171] 同理,还可以进行第三次补充检索,直至检索到的候选知识满足答复需求时,中控组件向人机交互系统返回检索到的候选知识。

[0172] 可选地,中控组件可以获取人机交互系统配置的补充检索次数阈值。中控组件基于配置的补充检索次数阈值,在补充检索的次数小于或等于补充检索次数阈值的情况下,向检索智能体发送补充检索信息。在补充检索的次数大于补充检索次数阈值的情况下,中控组件向人机交互系统返回已检索到的候选知识。其中,补充检索次数阈值可以根据实际应用场景的需要进行灵活地配置和调整,此处不做具体限定。

[0173] 本实施例的方案,可以在检索到的候选知识不满足答复需求的情况下,通过二次规划智能体根据查询信息、检索到的候选知识和原因识别结果,生成补充检索信息,并通过检索智能体基于补充检索信息进行候选知识的补充检索,获得生成答复所缺失的候选知识,得到生成完整答复所需的候选知识,减少了候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度,进而可以提升基于候选知识生成的答复信息的质量,得到让用户满意的答复。

[0174] 图8为本申请另一示例性实施例提供的多智能体的智能检索引擎的框架图。在图5所示框架的基础上,如图8所示,智能检索引擎可以包括中控组件,信息抽取智能体、检索智

能体、需求分析智能体、引用监测智能体、问题识别智能体和二次规划智能体。基于图8所示的智能检索引擎，基于多智能体的智能检索引擎的检索流程如下：

[0175] S1、中控组件接收人机交互系统发送的查询信息和交互需求信息。

[0176] S2、中控组件向需求分析智能体发送查询信息和交互需求信息。

[0177] S3、需求分析智能体基于查询信息和交互需求信息进行检索需求分析，确定答复查询信息是否需要进行知识检索，得到检索需求分析结果，并向中控组件返回检索需求分析结果。

[0178] S4、中控组件根据检索需求分析结果，在确定需要检索的情况下，向信息抽取智能体发送查询信息。

[0179] S5、信息抽取智能体提取查询信息的关键信息，并向中控组件返回关键信息。

[0180] S6、中控组件向检索智能体发送关键信息。

[0181] S7、检索智能体基于关键信息检索生成答复所需的候选知识，并向中控组件返回检索到的候选知识。

[0182] S8、中控组件向引用监测智能体发送查询信息和检索到的候选知识。

[0183] S9、引用监测智能体根据查询信息和检索到的候选知识，判断检索到的候选知识是否满足查询信息的答复需求，得到判断结果，并向中控组件返回判断结果。

[0184] S10、中控组件根据判断结果，在检索到的候选知识满足答复需求的情况下，向人机交互系统返回检索到的候选知识。

[0185] S11、中控组件根据判断结果，在检索到的候选知识不满足答复需求的情况下，向问题识别智能体发送判断结果、查询信息和检索到的候选知识。

[0186] S12、问题识别智能体在判断结果为检索到的候选知识不满足答复需求的情况下，根据查询信息和检索到的候选知识，识别检索到的候选知识不能满足答复需求的原因，得到原因识别结果，并向中控组件返回原因识别结果。

[0187] S13、中控组件向二次规划智能体发送查询信息、检索到的候选知识和原因识别结果。

[0188] S14、二次规划智能体根据查询信息、检索到的候选知识和原因识别结果，生成补充检索信息，并向中控组件返回补充检索信息。

[0189] S15、中控组件向检索智能体发送补充检索信息。

[0190] S16、检索智能体基于补充检索信息进行候选知识的补充检索，并向中控组件返回补充检索到的候选知识。

[0191] 中控组件还可以根据当前检索到的候选知识，执行步骤S8，触发执行步骤S8-S16的循环执行，直至执行步骤S10后结束检索。

[0192] 本实施例中中控组件及各个智能体的具体实现和效果参见前述实施例，此处不再赘述。

[0193] 图9为本申请一示例性实施例提供的人机交互方法的流程图。本实施例的执行主体可以是人机交互系统所在的服务器，基于多智能体的智能检索引擎可以对外提供应用程序编程接口(Application Program Interface，简称API)，人机交互系统可以通过该API调用基于多智能体的智能检索引擎，以更加全面和精准地获取到答复查询信息所需的候选知识。

[0194] 如图9所示,基于前述任一实施例提供的基于多智能体的智能检索引擎实现的人机交互的具体步骤如下:

[0195] 步骤S901、接收查询信息。

[0196] 其中,查询信息是用户通过端侧设备向人机交互系统提交的信息(query)。查询信息可以是各类复杂问题,包括但不限于多目标对比类问题,包含多个子问题的问题,多轮交互中包含指代、省略信息的问题。

[0197] 步骤S902、根据查询信息,调用多智能体的智能检索引擎,通过智能检索引擎的中控组件控制多个智能体协作完成对查询信息的如下检索处理:提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识。

[0198] 本实施例中,人机交互系统可以根据多智能体的智能检索引擎的API,调用多智能体的智能检索引擎,并向智能检索引擎发送查询信息。

[0199] 本实施例中的调用的多智能体的智能检索引擎,可以是前述任一实施例提供的多智能体的智能检索引擎,具体实现和技术效果参见前述实施例,此处不再赘述。

[0200] 需要说明的是,本实施例中人机交互系统可以调用前述任一实施例提供的智能检索引擎,并向智能检索引擎提供需要从人机交互系统获得的查询信息、交互需求信息、以及其他的配置信息等等,人机交互系统具体需要向智能检索引擎提供哪些信息,参见前述实施例中智能检索引擎所需从人机交互系统获取的信息,此处不再赘述。

[0201] 步骤S903、将查询信息和候选知识输入人机交互模型,通过人机交互模型根据候选知识生成查询信息的答复信息。

[0202] 在接收到智能检索引擎返回的候选知识后,人机交互系统可以根据将查询信息和候选知识输入人机交互模型,通过人机交互模型根据候选知识生成查询信息的答复信息。

[0203] 其中,人机交互模型是预先训练获得的,用于基于查询信息和候选知识生成答复信息。人机交互模型可以使用第二数据集(包括查询信息、候选知识和标注的答复信息)对基础模型进行微调训练获得。其中,训练人机交互模型使用基础模型可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言处理(NLP)能力的大模型,与训练语义理解模型使用的基础模型可以是同一基础模型,也可以是不同的基础模型,此处不做具体限定。

[0204] 步骤S904、输出答复信息。

[0205] 在获得查询信息的答复信息之后,人机交互系统可以向端侧设备输出查询信息的答复信息。端侧设备通过用户交互界面向用户输出答复信息。

[0206] 需要说明的是,本实施例中人机交互模型可以根据查询信息和检索到的与查询信息的关键信息匹配的候选知识,进行精准问答,生成答复信息。同时该人机交互模型还具有在候选知识不满足答复需求的情况下的拒识能力。人机交互模型在确定基于检索到的候选知识不满足答复需求,也即无法生成能够完整回答查询信息的答复信息的情况下,拒绝生成答复信息。这种情况下,人机交互系统可以提示用户更加详细地描述自己的需求。

[0207] 人机交互模型在候选知识不满足答复需求的情况下的拒识能力,可以通过在微调训练阶段,在第二数据集中增加候选知识不完整的训练数据,并标注模型拒绝答复的结果,通过微调训练使得人机交互模型具有在候选知识不完整的情况下的拒识能力。

[0208] 本实施例的方案,人机交互系统可以通过API调用多智能体的智能检索引擎,基于

多智能体构建智能检索引擎,通过智能体利用人工智能模型对查询信息进行深度语义理解,来提取查询信息的关键信息,可以精准、全面地提取查询信息的关键信息;进一步地,通过智能体借助人工智能模型基于关键信息检索生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度,进而可以提升基于候选知识生成的答复信息的质量,得到让用户满意的答复。

[0209] 图10为本申请一示例性实施例提供的多智能体的人机交互系统的流程图。本实施例提供的人机交互系统包括中控组件、用于知识检索的多个智能体和用于生成答复的问答智能体。如图10所示,基于多智能体的人机交互系统的人机交互流程如下:

[0210] 步骤S1001、中控组件接收查询信息。

[0211] 中控组件为人机交互系统中负责与用户端侧设备进行交互的组件,可以接收用户通过端侧设备向人机交互系统提交的查询信息(query)。

[0212] 步骤S1002、中控组件控制用于知识检索的多个智能体协作完成对查询信息的如下检索处理:提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识。

[0213] 本实施例中,中控组件控制用于知识检索的多个智能体协作完成对查询信息的检索处理,来获取生成答复所需的候选知识,参见前述实施例提供的多智能体的智能检索引擎中的处理过程,不同之处在于中控组件为人机交互系统内部的组件,中控组件直接从人机交互系统内部获取所需的信息即可。

[0214] 本实施例中,人机交互系统的中控组件在获取到生成答复所需的候选知识之后,还负责控制问答智能体根据查询信息和检索到的候选知识生成答复信息。

[0215] 步骤S1003、中控组件向问答智能体发送查询信息和检索到的候选知识。

[0216] 步骤S1004、问答智能体根据候选知识生成查询信息的答复信息,并向中控组件返回答复信息。

[0217] 其中,问答智能体的大脑部分使用的人机交互模型(如LLM等)。人机交互模型是预先训练获得的,用于基于查询信息和候选知识生成答复信息。人机交互模型可以使用第二数据集(包括查询信息、候选知识和标注的答复信息)对基础模型进行微调训练获得。其中,训练人机交互模型使用基础模型可以是各类预训练大模型,包括但不限于大规模预训练语言模型(LLM)、或其他的具有强大自然语言处理(NLP)能力的大模型,与训练语义理解模型使用的基础模型可以是同一基础模型,也可以是不同的基础模型,此处不做具体限定。

[0218] 需要说明的是,本实施例中人机交互模型可以根据查询信息和检索到的与查询信息的关键信息匹配的候选知识,进行精准问答,生成答复信息。同时该人机交互模型还具有在候选知识不满足答复需求的情况下的拒识能力。人机交互模型在确定基于检索到的候选知识不满足答复需求,也即无法生成能够完整回答查询信息的答复信息的情况下,拒绝生成答复信息。这种情况下,人机交互系统可以提示用户更加详细地描述自己的需求。

[0219] 人机交互模型在候选知识不满足答复需求的情况下的拒识能力,可以通过在微调训练阶段,在第二数据集中增加候选知识不完整的训练数据,并标注模型拒绝答复的结果,通过微调训练使得人机交互模型具有在候选知识不完整的情况下的拒识能力。

[0220] 步骤S1005、中控组件输出答复信息。

[0221] 在获得查询信息的答复信息之后,人机交互系统的中控组件还负责向端侧设备输

出查询信息的答复信息。端侧设备通过用户交互界面向用户输出答复信息。

[0222] 本实施例的方案,多智能体的人机交互系统,通过中控组件接收查询信息,通过智能体利用人工智能模型对查询信息进行深度语义理解,来提取查询信息的关键信息,可以精准、全面地提取查询信息的关键信息;通过智能体利用人工智能模型基于关键信息检索生成答复所需的候选知识,可以更加全面和精准地召回生成答复所需的候选知识,减少候选知识的漏召回、召回错误或失败的情况发生,提升了召回候选知识的精准度和完整性;进一步地,通过问答智能体根据候选知识生成查询信息的答复信息,可以提升基于生成答复信息的质量,得到让用户满意的答复。

[0223] 在一可选实施例中,多智能体的人机交互系统还包括信息抽取智能体和检索智能体。前述步骤S1002中,中控组件控制多个智能体协作完成对查询信息的如下检索处理:提取查询信息的关键信息,基于关键信息检索生成答复所需的候选知识,具体可以采用如下方式实现:中控组件向信息抽取智能体发送查询信息。信息抽取智能体提取查询信息的关键信息,并向中控组件返回关键信息。中控组件向检索智能体发送关键信息。检索智能体基于关键信息检索生成答复所需的候选知识,并向中控组件返回检索到的候选知识。

[0224] 在一可选实施例中,多智能体的人机交互系统还包括需求分析智能体。本实施例中,中控组件获取人机交互系统的交互需求信息,向需求分析智能体发送查询信息和交互需求信息。需求分析智能体基于查询信息和交互需求信息进行检索需求分析,确定答复查询信息是否需要进行知识检索,得到检索需求分析结果,并向中控组件返回检索需求分析结果。中控组件根据检索需求分析结果,在确定需要检索的情况下,向信息抽取智能体发送查询信息。

[0225] 可选地,交互需求信息包括检索规则。需求分析智能体基于检索规则,判断答复查询信息是否需要进行知识检索,得到判断结果。在判断结果为需要进行知识检索的情况下,向中控组件返回包含判断结果的检索需求分析结果。在判断结果为不需要进行知识检索的情况下,确定查询信息对应的答复话术信息,向中控组件返回包含判断结果和答复话术信息的检索需求分析结果。

[0226] 可选地,交互需求信息还包括配置的数据源。需求分析智能体在判断结果为需要进行知识检索的情况下,向中控组件返回包含判断结果和配置的数据源的检索需求分析结果。

[0227] 进一步地,中控组件根据检索需求分析结果,在确定需要检索的情况下,向信息抽取智能体发送查询信息配置的数据源。检索智能体基于关键信息,在数据源中检索生成答复所需的候选知识。

[0228] 在一可选实施例中,多智能体的人机交互系统还包括引用监测智能体。本实施例中,中控组件向人机交互系统返回检索到的候选知识,可以通过如下方式实现:中控组件向引用监测智能体发送查询信息和检索到的候选知识。引用监测智能体根据查询信息和检索到的候选知识,判断检索到的候选知识是否满足查询信息的答复需求,得到判断结果,并向中控组件返回判断结果。中控组件根据判断结果,在检索到的候选知识满足答复需求的情况下,向人机交互系统返回检索到的候选知识。

[0229] 进一步地,多智能体的人机交互系统还可以包括问题识别智能体。本实施例中,中控组件根据判断结果,在检索到的候选知识不满足答复需求的情况下,向问题识别智能体

发送判断结果、查询信息和检索到的候选知识。

[0230] 问题识别智能体在判断结果为检索到的候选知识不满足答复需求的情况下,根据查询信息和检索到的候选知识,识别检索到的候选知识不能满足答复需求的原因,得到原因识别结果,并向中控组件返回原因识别结果。

[0231] 进一步地,多智能体的人机交互系统还可以包括二次规划智能体。本实施例中,中控组件向二次规划智能体发送查询信息、检索到的候选知识和原因识别结果。二次规划智能体根据查询信息、检索到的候选知识和原因识别结果,生成补充检索信息,并向中控组件返回补充检索信息。中控组件向检索智能体发送补充检索信息。检索智能体基于补充检索信息进行候选知识的补充检索,并向中控组件返回补充检索到的候选知识。

[0232] 可选地,中控组件可以基于配置的补充检索次数阈值,在补充检索的次数小于或等于补充检索次数阈值的情况下,向检索智能体发送补充检索信息;中控组件在补充检索的次数大于补充检索次数阈值的情况下,向人机交互系统返回已检索到的候选知识。

[0233] 需要说明的是,多智能体的人机交互系统中,用于知识检索的多个智能体(包括信息抽取智能体、检索智能体、需求分析智能体、引用监测智能体、问题识别智能体、二次规划智能体)的实现原理和技术效果,以及用于知识检索的多个智能体与中控组件间交互过程,参见前述多智能体的智能检索引擎实施例,此处不再赘述。

[0234] 图11为本申请实施例提供的一种服务器的结构示意图。如图11所示,该服务器包括:存储器1101和处理器1102。存储器1101,用于存储计算机执行指令,并可被配置为存储其它各种数据以支持在服务器上的操作。处理器1102,与存储器1101通信连接,用于执行存储器1101存储的计算机执行指令,以实现上述任一方法实施例中提供的多智能体的智能检索引擎、多智能体的人机交互系统、基于多智能体的智能检索引擎的人机交互方法、基于多智能体的人机交互系统的人机交互方法,其具体功能和所能实现的技术效果类似,此处不再赘述。

[0235] 可选的,如图11所示,该服务器还包括:防火墙1103、负载均衡器1104、通信组件1105、电源组件1106等其它组件。图11中仅示意性给出部分组件,并不意味着服务器只包括图11所示组件。图11中仅以服务器为部署在云端的云服务器为例进行示例性地说明,服务器还可以部署在本地,本实施例此处不做具体限定。

[0236] 本申请实施例还提供一种计算机可读存储介质,计算机可读存储介质中存储有计算机执行指令,当处理器执行计算机执行指令时,实现前述任一实施例中多智能体的智能检索引擎、多智能体的人机交互系统、基于多智能体的智能检索引擎的人机交互方法、基于多智能体的人机交互系统的人机交互方法,具体功能和所能实现的技术效果此处不再赘述。

[0237] 本申请实施例还提供一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现前述任一实施例的方法。计算机程序存储在可读存储介质中,服务器的至少一个处理器可以从可读存储介质读取计算机程序,至少一个处理器执行计算机程序使得服务器执行上述任一方法实施例中多智能体的智能检索引擎、多智能体的人机交互系统、基于多智能体的智能检索引擎的人机交互方法、基于多智能体的人机交互系统的人机交互方法,具体功能和所能实现的技术效果此处不再赘述。

[0238] 本申请实施例提供一种芯片,包括:处理模块与通信接口,该处理模块能执行前述

方法实施例中第一服务器的技术方案。可选的,该芯片还包括存储模块(如,存储器),存储模块用于存储指令,处理模块用于执行存储模块存储的指令,并且对存储模块中存储的指令的执行使得处理模块执行前述任一方法实施例中多智能体的智能检索引擎、多智能体的人机交互系统、基于多智能体的智能检索引擎的人机交互方法、基于多智能体的人机交互系统的人机交互方法。

[0239] 上述以软件功能模块的形式实现的集成的模块,可以存储在一个计算机可读取存储介质中。上述软件功能模块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(可以是个人计算机,服务器,或者网络设备等)或处理器执行本申请各个实施例方法的部分步骤。

[0240] 应理解,上述处理器可以是处理单元(Central Processing Unit,简称CPU),还可以是其它通用处理器、数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合申请所公开的方法的步骤可以直接体现为硬件处理器执行完成,或者用处理器中的硬件及软件模块组合执行完成。存储器可能包含高速随机存取存储器(Random Access Memory,简称RAM),也可能还包括非易失性存储,例如至少一个磁盘存储器,还可以为U盘、移动硬盘、只读存储器、磁盘或光盘等。上述存储器可以是对象存储(Object Storage Service,简称OSS)。上述存储器可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(Static Random Access Memory,简称SRAM),电可擦除可编程只读存储器(Electrically Erasable Programmable Read Only Memory,简称EEPROM),可擦除可编程只读存储器(Erasable Programmable Read Only Memory,简称EPROM),可编程只读存储器(Programmable Read Only Memory,简称PROM),只读存储器(Read Only Memory,简称ROM),磁存储器,快闪存储器,磁盘或光盘。上述通信组件被配置为便于通信组件所在设备和其他设备之间有线或无线方式的通信。通信组件所在设备可以接入基于通信标准的无线网络,如移动热点(WiFi),第二代移动通信系统(2G)、第三代移动通信系统(3G)、第四代移动通信系统(4G)/长期演进(Long Term Evolution,简称LTE)、第五代移动通信系统(5G)等移动通信网络,或它们的组合。在一个示例性实施例中,通信组件经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,通信组件还包括近场通信(Near Field Communication,简称NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(Radio Frequency Identification,简称RFID)技术,红外技术,超宽带(Ultra Wide Band,简称UWB)技术,蓝牙技术和其他技术来实现。上述电源组件,为电源组件所在设备的各种组件提供电力。电源组件可以包括电源管理系统,一个或多个电源,及其他与为电源组件所在设备生成、管理和分配电力相关联的组件。上述存储介质可以是由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。存储介质可以是通用或专用计算机能够存取的任何可用介质。

[0241] 一种示例性的存储介质耦合至处理器,从而使处理器能够从该存储介质读取信息,且可向该存储介质写入信息。当然,存储介质也可以是处理器的组成部分。处理器和存

储介质可以位于专用集成电路中。当然,处理器和存储介质也可以作为分立组件存在于电子设备或主控设备中。

[0242] 需要说明的是,在本文中,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、物品或者装置不仅包括那些要素,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、物品或者装置所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0243] 上述本申请实施例的顺序仅仅为了描述,不代表实施例的优劣。另外,在上述实施例及附图中的描述的一些流程中,包含了按照特定顺序出现的多个操作,但是应该清楚了解,这些操作可以不按照其在本文中出现的顺序来执行或并行执行,仅仅是用于区分开各个不同的操作,序号本身不代表任何的执行顺序。另外,这些流程可以包括更多或更少的操作,并且这些操作可以按顺序执行或并行执行。需要说明的是,本文中的“第一”、“第二”等描述,是用于区分不同的消息、设备、模块等,不代表先后顺序,也不限定“第一”和“第二”是不同的类型。“多个”的含义是两个以上,除非另有明确具体的限定。

[0244] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献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(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,服务器,空调器,或者网络设备等)执行本申请各个实施例的方法。

[0245] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、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申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。

[0246] 以上仅为本申请的优选实施例,并非因此限制本申请的专利范围,凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本申请的专利保护范围内。

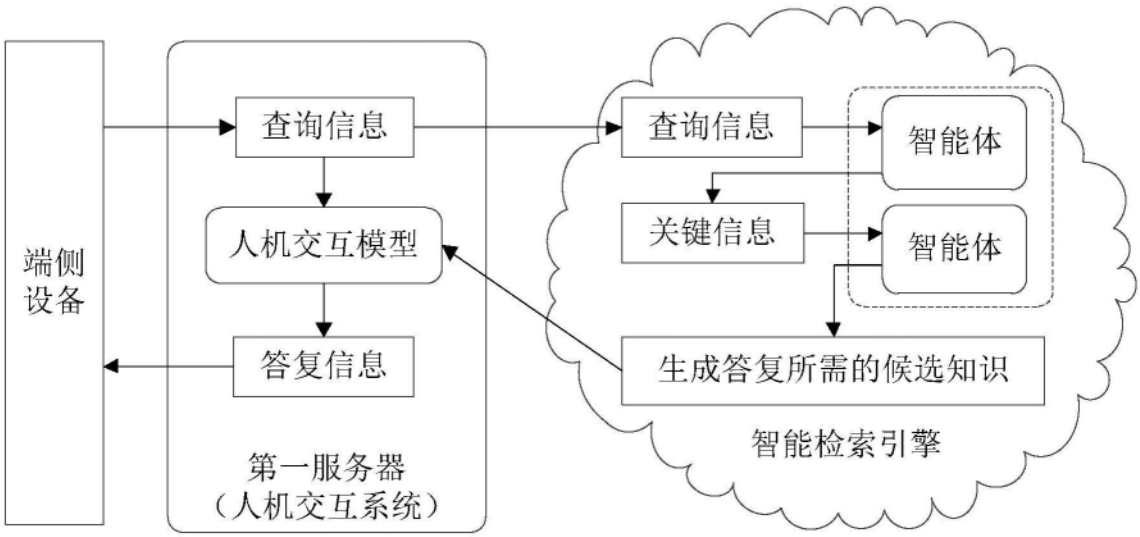

端侧
设备
查询信息
人机交互模型
答复信息
第一服务器
（人机交互系统）
查询信息
关键信息
生成答复所需的候选知识
智能体
智能体
智能检索引擎

图1

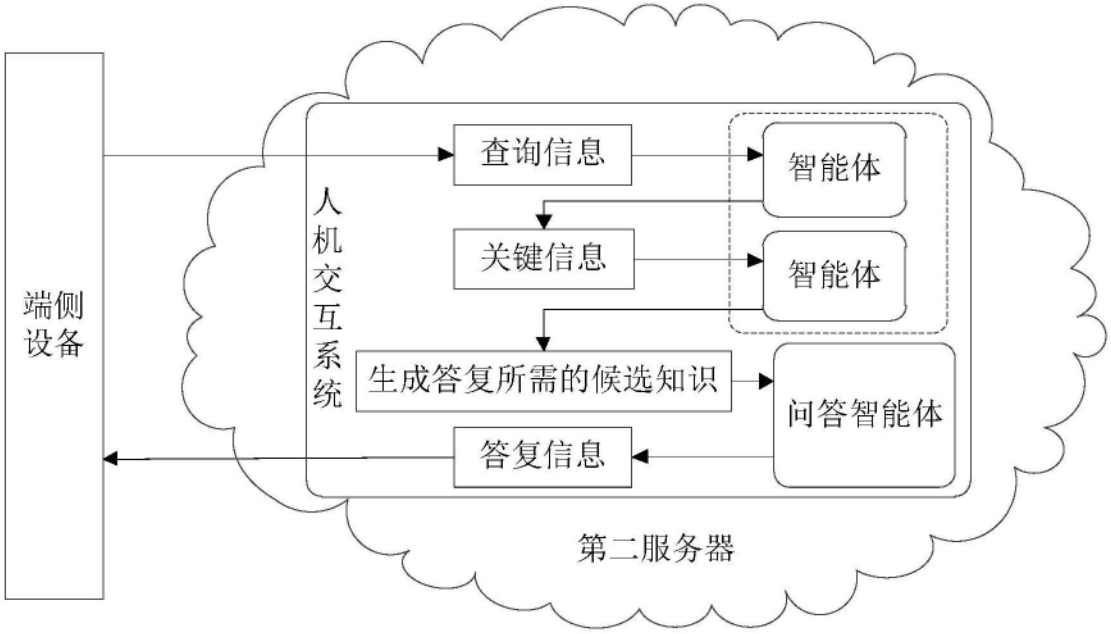

端侧
设备
人机交互系统
查询信息
关键信息
生成答复所需的候选知识
答复信息
智能体
智能体
问答智能体
第二服务器

图2

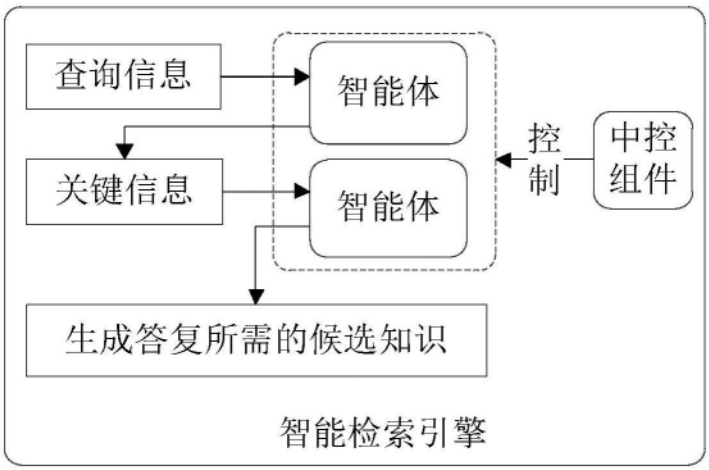

查询信息
智能体
关键信息
智能体
控制
中控组件
生成答复所需的候选知识
智能检索引擎

图3

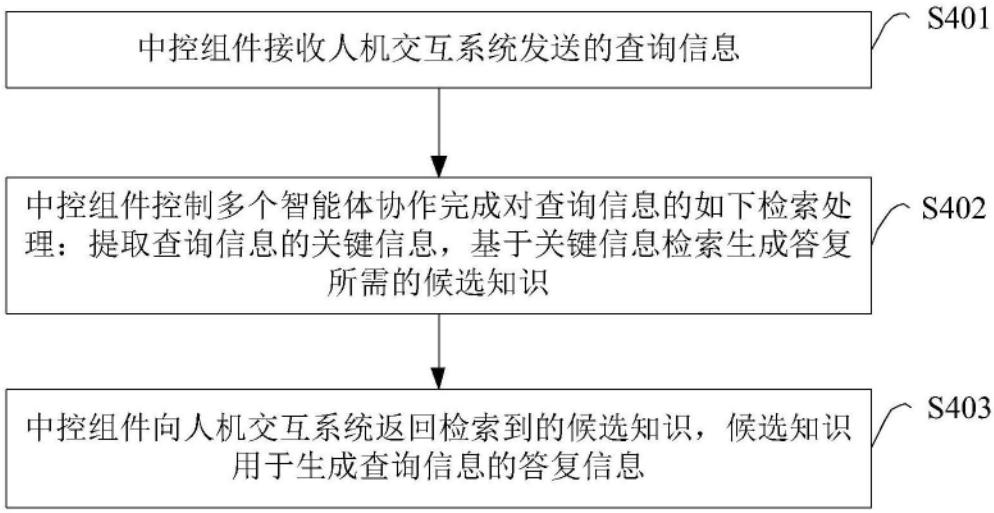

中控组件接收人机交互系统发送的查询信息
S401
中控组件控制多个智能体协作完成对查询信息的如下检索处理：提取查询信息的关键信息，基于关键信息检索生成答复所需的候选知识
S402
中控组件向人机交互系统返回检索到的候选知识，候选知识用于生成查询信息的答复信息
S403

图4

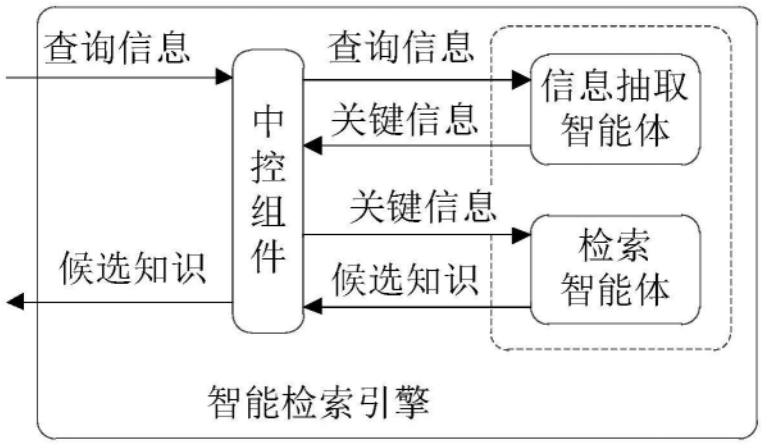

查询信息
查询信息
信息抽取
智能体
关键信息
中
控
组
件
关键信息
检索
智能体
候选知识
候选知识
智能检索引擎

图5

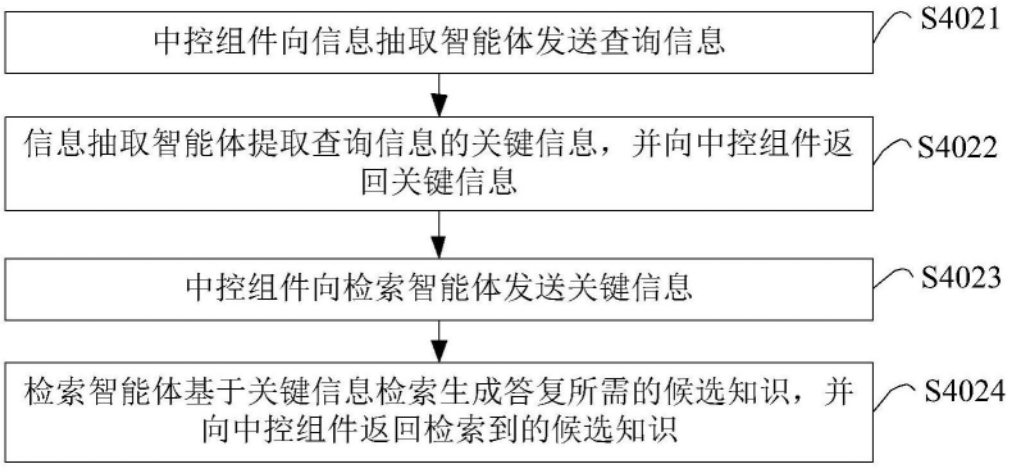

中控组件向信息抽取智能体发送查询信息
S4021
信息抽取智能体提取查询信息的关键信息，并向中控组件返回关键信息
S4022
中控组件向检索智能体发送关键信息
S4023
检索智能体基于关键信息检索生成答复所需的候选知识，并向中控组件返回检索到的候选知识
S4024

图6

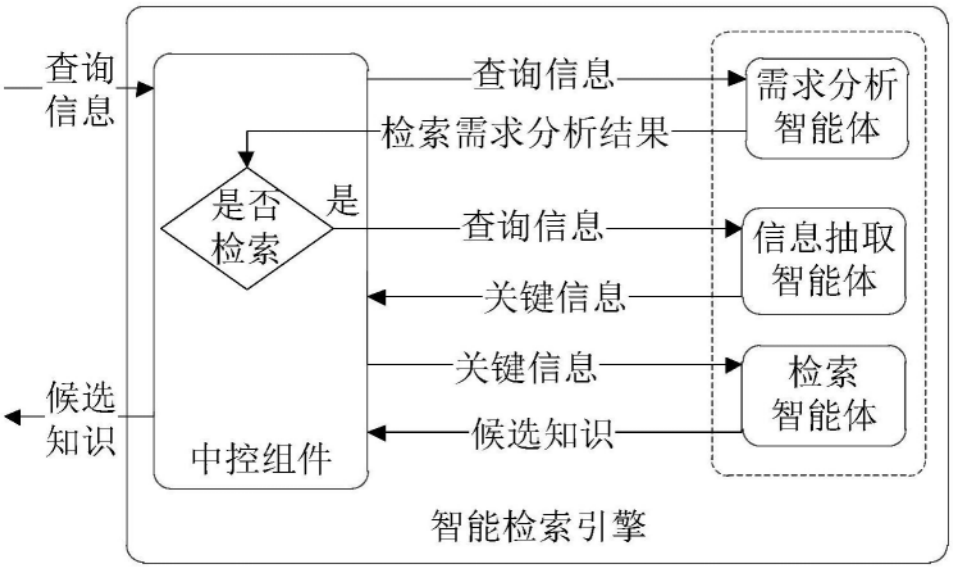

查询信息
查询信息
需求分析智能体
检索需求分析结果
是否检索
是
查询信息
信息抽取智能体
关键信息
关键信息
检索智能体
候选知识
候选知识
中控组件
智能检索引擎

图7

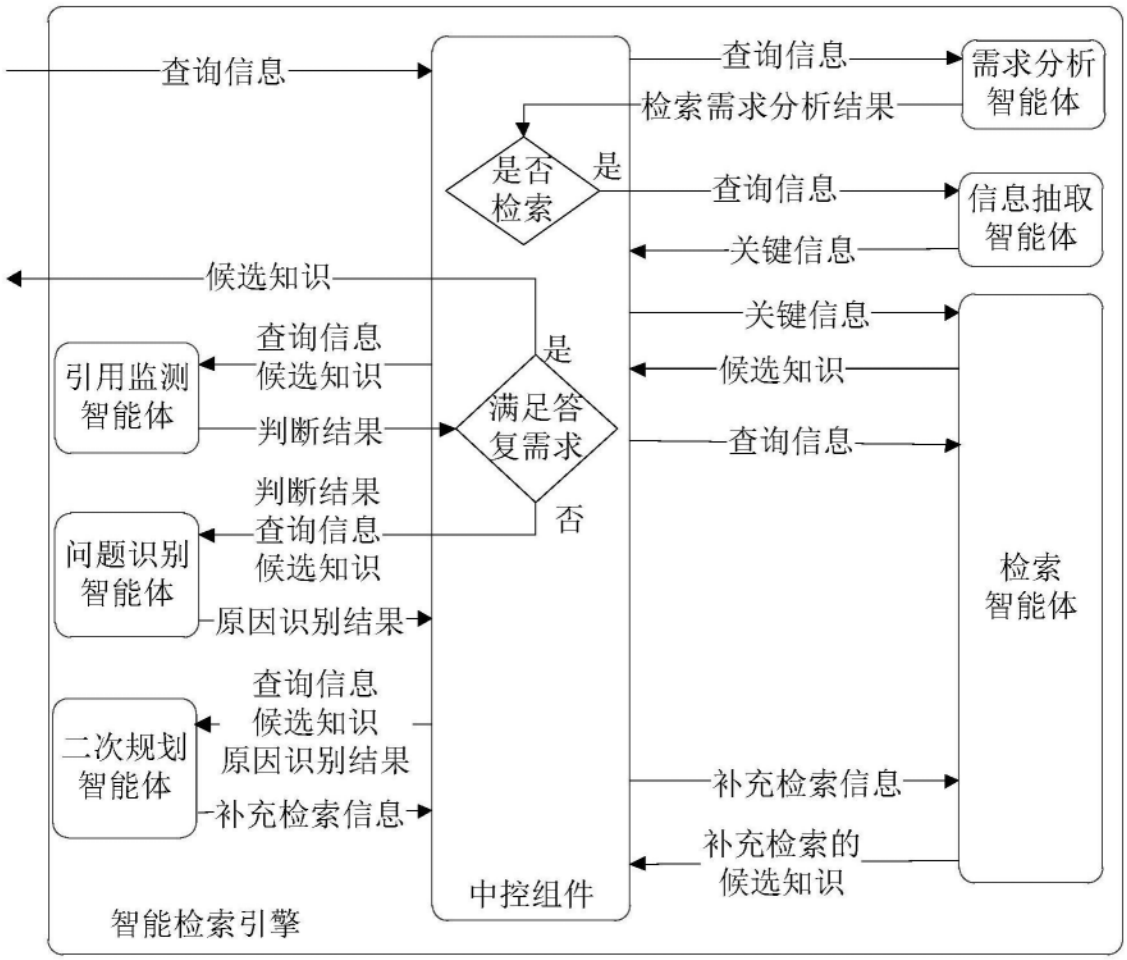

查询信息
查询信息
需求分析智能体
检索需求分析结果
是否检索
是
查询信息
信息抽取智能体
关键信息
候选知识
关键信息
查询信息
候选知识
引用监测智能体
是
满足答复需求
候选知识
判断结果
查询信息
查询信息
否
判断结果
查询信息
候选知识
问题识别智能体
检索智能体
原因识别结果
查询信息
候选知识
二次规划智能体
原因识别结果
补充检索信息
补充检索信息
补充检索的候选知识
中控组件
智能检索引擎

图8

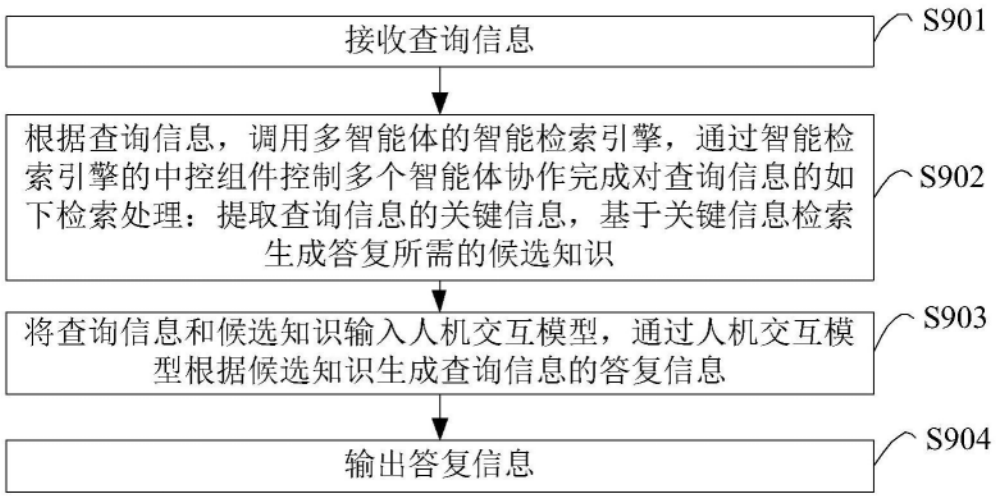

接收查询信息
S901
根据查询信息，调用多智能体的智能检索引擎，通过智能检索引擎的中控组件控制多个智能体协作完成对查询信息的如下检索处理：提取查询信息的关键信息，基于关键信息检索生成答复所需的候选知识
S902
将查询信息和候选知识输入人机交互模型，通过人机交互模型根据候选知识生成查询信息的答复信息
S903
输出答复信息
S904

图9

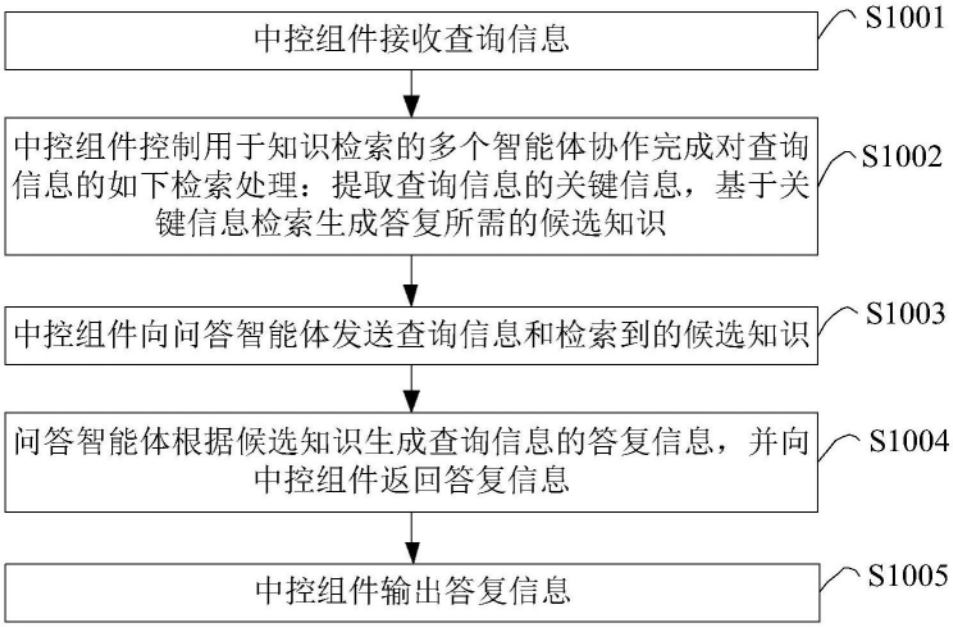

中控组件接收查询信息
S1001
中控组件控制用于知识检索的多个智能体协作完成对查询信息的如下检索处理：提取查询信息的关键信息，基于关键信息检索生成答复所需的候选知识
S1002
中控组件向问答智能体发送查询信息和检索到的候选知识
S1003
问答智能体根据候选知识生成查询信息的答复信息，并向中控组件返回答复信息
S1004
中控组件输出答复信息
S1005

图10

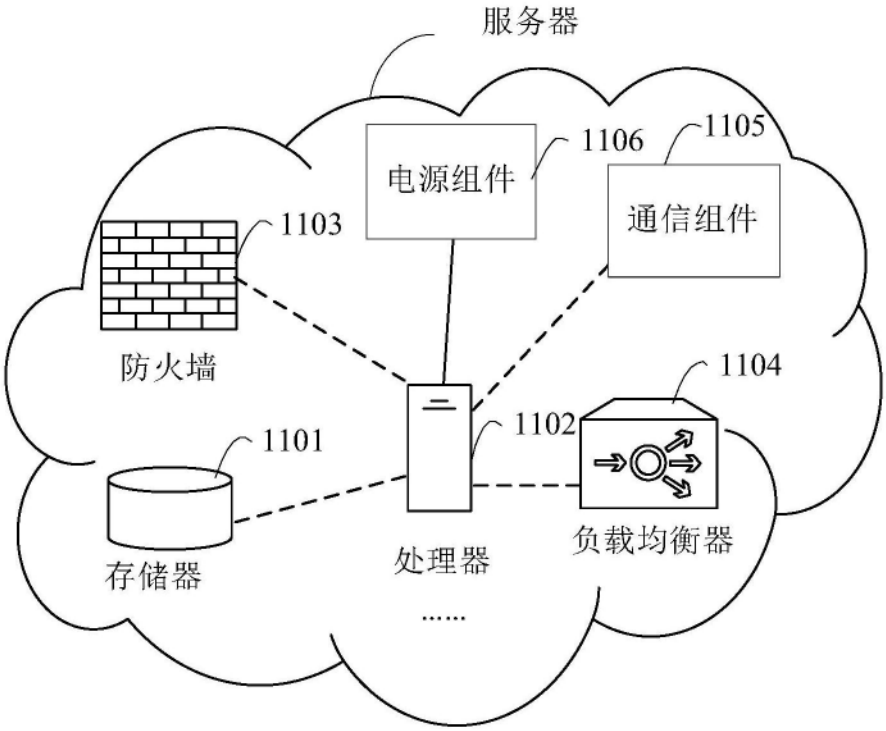


图11