(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120045750 A
(43)申请公布日 2025.05.27

(21)申请号 202411937157.7
(22)申请日 2024.12.26
(71)申请人 清华大学
地址 100084 北京市海淀区清华园
(72)发明人 陈舟 白玉琪
(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 江康宁
(51)Int.Cl.
G06F 16/9032(2019.01)
G06F 16/9035(2019.01)
G06F 16/903(2019.01)
G06F 16/906(2019.01)
G06F 16/9038(2019.01)
G06F 16/9535(2019.01)
G06F 16/9538(2019.01)
G06F 40/30(2020.01)

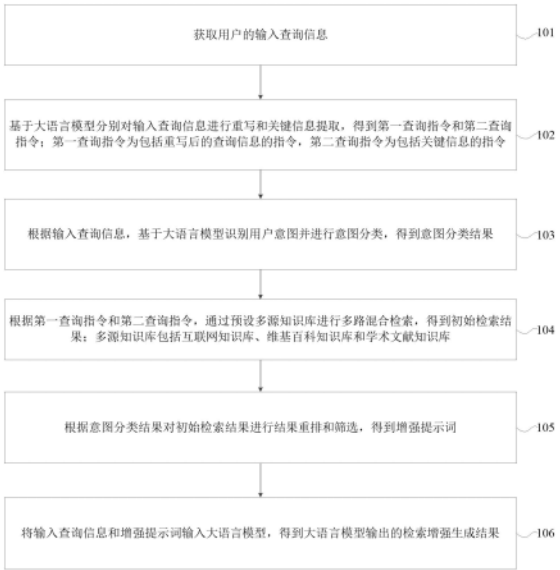
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种基于大语言模型的检索增强生成方法和系统

(57)摘要

本发明提供一种基于大语言模型的检索增强生成方法和系统,该方法包括:获取用户的输入查询信息;基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果;根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词;将输入查询信息和增强提示词输入大语言模型,得到检索增强生成结果。本发明能够精准识别用户查询的需求和意图,在开放域中高精度检索相关信息,提高了信息检索的即时性、准确性、专业性和回复质量。



CN 120045750 A

1.一种基于大语言模型的检索增强生成方法,其特征在于,包括:

获取用户的输入查询信息;

基于大语言模型分别对所述输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;所述第一查询指令为包括重写后的查询信息的指令,所述第二查询指令为包括关键信息的指令;

根据所述输入查询信息,基于所述大语言模型识别用户意图并进行意图分类,得到意图分类结果;

根据所述第一查询指令和所述第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;所述多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;

根据所述意图分类结果对所述初始检索结果进行结果重排和筛选,得到增强提示词;

将所述输入查询信息和所述增强提示词输入所述大语言模型,得到所述大语言模型输出的检索增强生成结果。

2.根据权利要求1所述的基于大语言模型的检索增强生成方法,其特征在于,所述基于大语言模型分别对所述输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令,包括:

对所述输入查询信息进行自然语言处理,以识别查询中的语法结构和语义内容;

根据识别出的所述语法结构和所述语义内容,使用所述大语言模型生成多个重写版本;

评估每个所述重写版本的语义相似度和信息保留度,选择语义相似度和所述信息保留度最高的重写版本整合到所述第一查询指令;

使用所述大语言模型从所述输入查询信息中提取查询关键信息,以根据所述查询信息生成所述第二查询指令;所述查询关键信息包括实体和概念。

3.根据权利要求1所述的基于大语言模型的检索增强生成方法,其特征在于,所述根据所述输入查询信息,基于所述大语言模型识别用户意图并进行意图分类,得到意图分类结果,包括:

使用所述大语言模型对所述输入查询信息进行深度语义分析,以理解用户的需求和查询目的,得到深度语义分析结果;

将所述深度语义分析结果与预设意图类别进行匹配,得到所述意图分类结果。

4.根据权利要求1所述的基于大语言模型的检索增强生成方法,其特征在于,所述根据所述第一查询指令和所述第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果,包括:

将所述预设多源知识库进行文本切分,得到多个文本片段;

基于预设向量模型将多个所述文本片段转换为向量表示,得到多个文本片段向量;

基于所述预设向量模型分别将所述第一查询指令和所述第二查询指令转换为向量表示,得到第一查询指令向量和第二查询指令向量;

根据多个所述文本片段向量、所述第一查询指令向量和所述第二查询指令向量进行多路混合检索,评估相似度和相关性,得到所述初始检索结果。

5.根据权利要求4所述的基于大语言模型的检索增强生成方法,其特征在于,所述根据

多个所述文本片段向量、所述第一查询指令向量和所述第二查询指令向量进行多路混合检索,评估相似度和相关性,得到所述初始检索结果,包括:

计算所述第一查询指令向量与所有所述文本片段向量的余弦距离,得到第一余弦距离;所述第一余弦距离用于评估在多路混合检索中所述第一查询指令向量与所有文本片段向量之间的相似度和相关性;

计算所述第二查询指令向量与所有所述文本片段向量的余弦向量,得到第二余弦距离;所述第二余弦距离用于评估在多路混合检索中所述第二查询指令向量与所有文本片段向量之间的相似度和相关性;

将所述第一余弦距离与所述第二余弦距离相加,得到第三余弦距离;

根据所述第三余弦距离,得到所述初始检索结果。

6.根据权利要求1至5任一项所述的基于大语言模型的检索增强生成方法,其特征在于,所述根据所述意图分类结果对所述初始检索结果进行结果重排和筛选,得到增强提示词,包括:

根据所述意图分类结果对所述初始检索结果的参数进行调整,得到调整后的检索结果;

根据调整后的检索结果的参数进行检索结果重新排序,并筛选排序在预设数量的结果作为所述增强提示词。

7.一种基于大语言模型的检索增强生成系统,其特征在于,包括:

获取模块,用于获取用户的输入查询信息;

查询指令模块,用于基于大语言模型分别对所述输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;所述第一查询指令为包括重写后的查询信息的指令,所述第二查询指令为包括关键信息的指令;

意图分类模块,用于根据所述输入查询信息,基于所述大语言模型识别用户意图并进行意图分类,得到意图分类结果;

检索模块,用于根据所述第一查询指令和所述第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;所述多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;

增强提示模块,用于根据所述意图分类结果对所述初始检索结果进行结果重排和筛选,得到增强提示词;

结果生成模块,用于将所述输入查询信息和所述增强提示词输入所述大语言模型,得到所述大语言模型输出的检索增强生成结果。

8.一种电子设备,包括存储器、处理器及存储在所述存储器上并在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现如权利要求1至6任一项所述基于大语言模型的检索增强生成方法。

9.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至6任一项所述基于大语言模型的检索增强生成方法。

10.一种计算机程序产品,包括计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至6任一项所述基于大语言模型的检索增强生成方法。

一种基于大语言模型的检索增强生成方法和系统

技术领域

[0001] 本发明涉及自然语言处理技术领域,尤其涉及一种基于大语言模型的检索增强生成方法和系统。

背景技术

[0002] 在数字化信息爆炸式增长的今天,构建高效的信息检索系统显得尤为关键,这些系统必须能够迅速且准确地从海量数据中筛选出所需的信息。然而,传统检索系统往往依赖于单一的知识源,例如互联网搜索引擎或专业数据库,这在很大程度上限制了它们处理多样化和跨领域查询的效率与准确性。此外,这些系统大多基于关键词匹配,难以深入理解用户复杂的查询意图,导致检索结果常常与用户的实际需求存在偏差。

[0003] 为了提升检索的准确性,近年来,基于深度学习的语言模型应运而生,例如GPT和BERT。这些模型利用语义理解技术,显著提高了检索的精确度。尽管如此,这些依赖单一模型的解决方案在整合多源数据和适应特定领域知识方面仍面临挑战,且在开放域中的检索结果往往包含大量噪声,这限制了知识服务的效率。

[0004] 因此,如何精确识别用户的查询需求和意图,在开放域中实现高精度的信息检索,提升信息检索系统回复的即时性、准确性和专业性,成为了当前亟待解决的技术问题。

发明内容

[0005] 鉴于上述问题,本发明提供一种基于大语言模型的检索增强生成方法和系统,通过部署多源知识库,使用大语言模型重写用户的输入查询信息、抽取用户的输入查询信息的关键信息、识别用户的意图并进行分类,然后进行多路混合检索、检索结果重排,最后生成回复。本发明能够精准识别用户查询的需求和意图,在开放域中高精度检索相关信息,提高信息检索系统获取的信息的即时性、准确性和专业性,提高系统的回复质量。

[0006] 本发明提供一种基于大语言模型的检索增强生成方法,包括:获取用户的输入查询信息;基于大语言模型分别对所述输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;所述第一查询指令为包括重写后的查询信息的指令,所述第二查询指令为包括关键信息的指令;根据所述输入查询信息,基于所述大语言模型识别用户意图并进行意图分类,得到意图分类结果;根据所述第一查询指令和所述第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;所述多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;根据所述意图分类结果对所述初始检索结果进行结果重排和筛选,得到增强提示词;将所述输入查询信息和所述增强提示词输入所述大语言模型,得到所述大语言模型输出的检索增强生成结果。

[0007] 根据本发明提供的一种基于大语言模型的检索增强生成方法,所述基于大语言模型分别对所述输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令,包括:对所述输入查询信息进行自然语言处理,以识别查询中的语法结构和语义内容;根据识别出的所述语法结构和所述语义内容,使用所述大语言模型生成多个重写版本;评

估每个所述重写版本的语义相似度和信息保留度,选择语义相似度和所述信息保留度最高的重写版本整合到所述第一查询指令;使用所述大语言模型从所述输入查询信息中提取查询关键信息,以根据所述查询信息生成所述第二查询指令;所述查询关键信息包括实体和概念。

[0008] 根据本发明提供的一种基于大语言模型的检索增强生成方法,所述根据所述输入查询信息,基于所述大语言模型识别用户意图并进行意图分类,得到意图分类结果,包括:使用所述大语言模型对所述输入查询信息进行深度语义分析,以理解用户的需求和查询目的,得到深度语义分析结果;将所述深度语义分析结果与预设意图类别进行匹配,得到所述意图分类结果。

[0009] 根据本发明提供的一种基于大语言模型的检索增强生成方法,所述根据所述第一查询指令和所述第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果,包括:将所述预设多源知识库进行文本切分,得到多个文本片段;基于预设向量模型将多个所述文本片段转换为向量表示,得到多个文本片段向量;基于所述预设向量模型分别将所述第一查询指令和所述第二查询指令转换为向量表示,得到第一查询指令向量和第二查询指令向量;根据多个所述文本片段向量、所述第一查询指令向量和所述第二查询指令向量进行多路混合检索,评估相似度和相关性,得到所述初始检索结果。

[0010] 根据本发明提供的一种基于大语言模型的检索增强生成方法,所述根据多个所述文本片段向量、所述第一查询指令向量和所述第二查询指令向量进行多路混合检索,评估相似度和相关性,得到所述初始检索结果,包括:计算所述第一查询指令向量与所有所述文本片段向量的余弦距离,得到第一余弦距离;所述第一余弦距离用于评估在多路混合检索中所述第一查询指令向量与所有文本片段向量之间的相似度和相关性;计算所述第二查询指令向量与所有所述文本片段向量的余弦向量,得到第二余弦距离;所述第二余弦距离用于评估在多路混合检索中所述第二查询指令向量与所有文本片段向量之间的相似度和相关性;将所述第一余弦距离与所述第二余弦距离相加,得到第三余弦距离;根据所述第三余弦距离,得到所述初始检索结果。

[0011] 根据本发明提供的一种基于大语言模型的检索增强生成方法,所述根据所述意图分类结果对所述初始检索结果进行结果重排和筛选,得到增强提示词,包括:根据所述意图分类结果对所述初始检索结果的参数进行调整,得到调整后的检索结果;根据调整后的检索结果的参数进行检索结果重新排序,并筛选排序在预设数量的结果作为所述增强提示词。

[0012] 本发明还提供一种基于大语言模型的检索增强生成系统,包括:获取模块,用于获取用户的输入查询信息;查询指令模块,用于基于大语言模型分别对所述输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;所述第一查询指令为包括重写后的查询信息的指令,所述第二查询指令为包括关键信息的指令;意图分类模块,用于根据所述输入查询信息,基于所述大语言模型识别用户意图并进行意图分类,得到意图分类结果;检索模块,用于根据所述第一查询指令和所述第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;所述多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;增强提示模块,用于根据所述意图分类结果对所述初始检索结果进行结果重排和筛选,得到增强提示词;结果生成模块,用于将所述输入查询信息和所述增强提

示词输入所述大语言模型,得到所述大语言模型输出的检索增强生成结果。

[0013] 本发明还提供一种电子设备,包括存储器、处理器及存储在存储器上并在处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现如上述任一种所述基于大语言模型的检索增强生成方法。

[0014] 本发明还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如上述任一种所述基于大语言模型的检索增强生成方法。

[0015] 本发明还提供一种计算机程序产品,包括计算机程序,所述计算机程序被处理器执行时实现如上述任一种所述基于大语言模型的检索增强生成方法。

[0016] 本发明提供的一种基于大语言模型的检索增强生成方法和系统,该方法包括:获取用户的输入查询信息;基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果;根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词;将输入查询信息和增强提示词输入大语言模型,得到检索增强生成结果。本发明能够精准识别用户查询的需求和意图,在开放域中高精度检索相关信息,提高了信息检索的即时性、准确性、专业性和回复质量。

附图说明

[0017] 为了更清楚地说明本发明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一简单地介绍,显而易见地,下面描述中的附图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。

[0018] 图1是本发明提供的一种基于大语言模型的检索增强生成方法的流程示意图。

[0019] 图2是本发明提供的一种基于大语言模型的检索增强生成系统的结构示意图。

[0020] 图3是本发明提供的电子设备的结构示意图。

具体实施方式

[0021] 为使本发明的目的、技术方案和优点更加清楚,下面将结合本发明中的附图,对本发明中的技术方案进行清楚、完整地描述,显然,所描述的实施例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作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。

[0022] 请参考图1,图1为本发明提供的一种基于大语言模型的检索增强生成方法的流程示意图。

[0023] 本发明提供一种基于大语言模型的检索增强生成方法,包括:

101:获取用户的输入查询信息。

[0024] 102:基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;第一查询指令为包括重写后的查询信息的指令,第二查询指令为包括关键信息的指令。

[0025] 作为一种优选的实施例,基于大语言模型分别对输入查询信息进行重写和关键信

息提取,得到第一查询指令和第二查询指令,包括:对输入查询信息进行自然语言处理,以识别查询中的语法结构和语义内容;根据识别出的语法结构和语义内容,使用大语言模型生成多个重写版本;评估每个重写版本的语义相似度和信息保留度,选择语义相似度和信息保留度最高的重写版本整合到第一查询指令;使用大语言模型从输入查询信息中提取查询关键信息,以根据查询信息生成第二查询指令;查询关键信息包括实体和概念。

[0026] 在本实施例中,大语言模型首先接收并分析用户以口语化形式的输入查询信息,通过深度学习和自然语言处理能力,识别出查询中的语法结构和语义内容。根据识别出的语法结构和语义内容提炼出关键信息。大语言模型根据这些关键信息自动修改查询,得到多个重写版本。这些重写版本旨在从不同角度表达相同的查询意图,以增加检索的覆盖面。然后可以使用余弦相似度来计算每个重写版本的语义相似度。同时,信息保留度可以通过比较重写版本与原始查询中的关键词和概念的匹配程度来评估。为了生成一个结构更清晰、表达准确且无歧义的查询指令,选择语义相似度和信息保留度最高的重写版本,整合到第一查询指令中。这一转换不仅增强了查询的明确性,也优化了后续检索过程的相关性和效率。通过这种方式,确保从用户的非正式表达中抽取出精确的需求,为混合检索过程提供坚实的基础。

[0027] 利用大语言模型对用户的输入查询信息进行深入分析,以抽取包含的直接和隐含的查询关键信息,从而生成第二查询指令。此过程涉及到识别查询文本中的重要实体,如人名、地点、事件和专有名词等。大语言模型通过其高级语义理解能力,能够不仅捕捉到明显的关键词,还能洞察到那些隐含的、可能未直接表达但对查询意图至关重要的信息。生成的第二查询指令因此包括了所有关键实体和概念,这为后续的混合检索和数据处理提供了精准的导向,确保了检索结果的相关性和全面性。

[0028] 103:根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果。

[0029] 作为一种优选的实施例,根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果,包括:使用大语言模型对输入查询信息进行深度语义分析,以理解用户的需求和查询目的,得到深度语义分析结果;将深度语义分析结果与预设意图类别进行匹配,得到意图分类结果。

[0030] 在本实施例中,利用大语言模型对输入查询信息进行深度语义分析。这一分析过程涉及到对用户输入的自然语言文本进行解析,以识别用户的真正需求和查询目的。深度语义分析的过程可以包括文本预处理、特征提取和语义理解。随后,大语言模型将深度语义分析结果(理解到的信息)与预设意图类别进行匹配,将查询归类到预设的三个类别之一:互联网类、维基百科类或学术文献类,得到意图分类结果。这种分类是根据查询内容的特征和用户可能寻求的信息类型进行的。例如,对当前事件的查询可能归类为互联网类,对历史或定义性内容的查询可能归类为维基百科类,而具有学术或研究深度的查询则归类为学术文献类。这一步骤不仅加快了检索流程,还确保了检索结果的精确性和相关性,有效地匹配了用户的检索需求与最合适的数据源。

[0031] 104:根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;多源知识库包括互联网知识库、维基百科知识库和学术文献知识库。

[0032] 作为一种优选的实施例,根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果,包括:将预设多源知识库进行文本切分,得到多个文本片段;基于预设向量模型将多个文本片段转换为向量表示,得到多个文本片段向量;基于预设向量模型分别将第一查询指令和第二查询指令转换为向量表示,得到第一查询指令向量和第二查询指令向量;根据多个文本片段向量、第一查询指令向量和第二查询指令向量进行多路混合检索,评估相似度和相关性,得到初始检索结果。

[0033] 作为一种优选的实施例,根据多个文本片段向量、第一查询指令向量和第二查询指令向量进行多路混合检索,评估相似度和相关性,得到初始检索结果,包括:计算第一查询指令向量与所有文本片段向量的余弦距离,得到第一余弦距离;第一余弦距离用于评估在多路混合检索中第一查询指令向量与所有文本片段向量之间的相似度和相关性;计算第二查询指令向量与所有文本片段向量的余弦向量,得到第二余弦距离;第二余弦距离用于评估在多路混合检索中第二查询指令向量与所有文本片段向量之间的相似度和相关性;将第一余弦距离与第二余弦距离相加,得到第三余弦距离;根据第三余弦距离,得到初始检索结果。

[0034] 在本实施例中,多源知识库包括互联网知识库、维基百科知识库和学术文献知识库。

[0035] 互联网知识库是指将必应搜索引擎的API(Application Programming Interface,应用程序编程接口)作为接口的知识库。使用必应搜索引擎提供的开放接口,以实时查询互联网上的数据。此部署方式允许访问最广泛的信息源,包括最新的网页内容和动态数据。通过API调用,可以直接将用户的查询发送至必应搜索引擎,并接收搜索结果,这些结果随后被用于后续的处理和分析。

[0036] 维基百科知识库是指爬取维基百科所有词条得到的本地知识库。设有一个定期运行的爬虫程序,该程序负责从维基百科网站下载新的和更新的词条内容,将其存储在本地服务器上。这种方式的优点是可以无需实时互联网连接即可访问维基百科的广泛信息,同时还减少了对维基百科服务器的请求负担。本地知识库的存在保证了检索操作的快速响应和高可靠性。

[0037] 学术文献知识库是指学术文献知识库指将Semantic Scholar的API作为接口的知识库。Semantic Scholar是一个广泛使用的学术论文搜索引擎,提供对大量学术文献的访问。通过API,可以直接查询Semantic Scholar的数据库,获取关于学术论文的详细信息,包括论文摘要、引用信息和下载链接。这种部署方式能够利用最新的学术研究成果,支持对科学问题的深入分析和理解。

[0038] 设置最大字符数为1024,将互联网知识库、维基百科知识库和学术文献知识库中的内容进行文本切分,得到多个文本片段。这一过程确保每个文本片段都是容易管理和处理的大小,有助于提高后续处理的效率和效果。然后基于预设向量模型(如词嵌入或BERT模型),将互联网知识库、维基百科知识库和学术文献知识库的文本片段转换为向量表示,得到多个文本片段向量,使得每个文本片段都能在向量空间中有其数学表达,便于进行数学上的比较和计算。基于预设向量模型,将第一查询指令转换为向量表示,得到第一查询指令向量。使用相同的向量模型来确保查询指令和文本片段在同一向量空间内,从而使得其后的比较操作具有可行性和准确性。计算第一查询指令向量和所有文本片段向量的余弦距

离,得到第一余弦距离。余弦距离用于衡量向量之间的相似度,这里用来确定每个文本片段与查询指令的相关性程度。基于向量模型,将第二查询指令转换为向量表示,得到第二查询指令向量。第二查询指令也被转换成向量表示,确保这一查询指令可以有效地与文本片段进行相似度比较。计算第二查询指令向量和所有文本片段向量的余弦距离,得到第二余弦距离,同样用以评估相似度和相关性。将第一余弦距离与第二余弦距离相加,以得到一个综合的余弦距离度量(第三余弦距离),这个度量考虑了两个不同维度的查询指令对相同文本片段的相关性评估。通过取第三余弦距离的倒数作为数据片段的得分(初始检索结果)。这样的得分方法使得距离越小(即越相关)的文本片段得分越高,从而优先被选用。

[0039] 将第一余弦距离与第二余弦距离相加,具体为第一余弦距离的权值和第二余弦距离的权值之和为1,一般取值分别为0.7和0.3,具体取值需要通过具体实验确定。

[0040] 105:根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词。

[0041] 作为一种优选的实施例,根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词,包括:根据意图分类结果对初始检索结果的参数进行调整,得到调整后的检索结果;根据调整后的检索结果的参数进行检索结果重新排序,并筛选排序在预设数量的结果作为增强提示词。

[0042] 在本实施例中,首先根据意图分类结果对初始检索结果的参数进行调整,得到调整后的检索结果。具体来说,如果某个数据片段属于被识别出的查询类别对应的知识库(互联网类、维基百科类或学术文献类),则该数据片段的得分将被扩大两倍。这种得分调整是为了优先考虑与用户查询意图更为相关的数据源。随后,根据调整后的检索结果的参数进行检索结果重新排序,并筛选排序在预设数量的结果作为增强提示词。具体的,对所有调整后的得分进行重新排序,并选择得分最高的前N个数据片段作为增强提示词。N的默认值设定为5,但这个参数可以根据实验或实际应用的需要进行调整。这一步骤确保了检索结果的高相关性和精确性,从而更好地满足用户的信息需求。

[0043] 106:将输入查询信息和增强提示词输入大语言模型,得到大语言模型输出的检索增强生成结果。

[0044] 在本实施例中,将用户的输入查询信息和N个片段(增强提示词)组装为输入提示词。输入提示词设计得简洁而无歧义,包含所有关键信息,确保能够有效地引导大语言模型理解和响应用户的查询需求。然后,输入提示词输入到大语言模型中。大语言模型利用这些信息进行深度语义处理,并生成一个针对用户查询的详尽回答或提供相关信息的输出,得到检索增强生成结果。这一过程充分利用了混合检索得到的高质量数据片段,以期达到最优的回答质量和用户满意度。

[0045] 大语言模型为开源大语言模型,大语言模型(Large Language Model,LLM)也称大型语言模型,是一种人工智能模型,旨在理解和生成人类语言或者说自然语言。大语言模型在大量的文本数据上进行训练,可以执行广泛的任务,包括进行对话、问答、文本分类、文本总结、翻译、情感分析等等。大语言模型的特点是规模庞大,包含数十亿的参数,帮助它们学习语言数据中的复杂模式。大语言模型通常基于深度学习架构,如转化器,这有助于它们执行各种自然语言处理任务。

[0046] 大规模语言模型也可以为任意开源的ChatGLM-6B、ChatGPT系列、StableVicuna、PaLM、Galactica或者LLaMA系列的模型,根据具体情况确定即可。开源大语言模型中代码开

源,数据集开源以及具有授权许可。

[0047] 本发明通过部署多源知识库,使用大语言模型重写用户的输入查询信息、抽取用户的输入查询信息的关键信息、识别用户的意图并进行分类,然后进行多路混合检索、检索结果重排,最后生成回复。本发明能够精准识别用户查询的需求和意图,在开放域中高精度检索相关信息,提高信息检索系统获取的信息的即时性、准确性和专业性,提高系统的回复质量。

[0048] 下面对本发明提供的基于大语言模型的检索增强生成系统进行描述,下文描述的基于大语言模型的检索增强生成系统与上文描述的基于大语言模型的检索增强生成方法可相互对应参照。

[0049] 请参考图2,图2为本发明提供的一种基于大语言模型的检索增强生成系统的结构示意图。

[0050] 本发明还提供一种基于大语言模型的检索增强生成系统,包括:获取模块201,用于获取用户的输入查询信息;查询指令模块202,用于基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;第一查询指令为包括重写后的查询信息的指令,第二查询指令为包括关键信息的指令;意图分类模块203,用于根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果;检索模块204,用于根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;增强提示模块205,用于根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词;结果生成模块206,用于将输入查询信息和增强提示词输入大语言模型,得到大语言模型输出的检索增强生成结果。

[0051] 本发明的基于大语言模型的检索增强生成系统包括获取模块、查询指令模块、意图分类模块、检索模块、增强提示模块和结果生成模块。各模块功能紧密衔接,通过多个专门设计的模块协同工作,旨在开放域中高精度检索相关信息,提高信息检索系统生成的回复的即时性、准确性和专业性。

[0052] 图3示例了一种电子设备的结构示意图,如图3所示,该电子设备可以包括:处理器(processor)301、通信接口(Communications Interface)302、存储器(memory)303和通信总线304,其中,处理器301,通信接口302,存储器303通过通信总线304完成相互间的通信。处理器301可以调用存储器303中的逻辑指令,以执行基于大语言模型的检索增强生成方法,该方法包括:获取用户的输入查询信息;基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;第一查询指令为包括重写后的查询信息的指令,第二查询指令为包括关键信息的指令;根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果;根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词;将输入查询信息和增强提示词输入大语言模型,得到大语言模型输出的检索增强生成结果。

[0053] 此外,上述的存储器303中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本

发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0054] 另一方面,本发明还提供一种计算机程序产品,所述计算机程序产品包括计算机程序,计算机程序可存储在非暂态计算机可读存储介质上,所述计算机程序被处理器执行时,计算机能够执行上述各方法所提供的基于大语言模型的检索增强生成方法,该方法包括:获取用户的输入查询信息;基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;第一查询指令为包括重写后的查询信息的指令,第二查询指令为包括关键信息的指令;根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果;根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词;将输入查询信息和增强提示词输入大语言模型,得到大语言模型输出的检索增强生成结果。

[0055] 又一方面,本发明还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各方法提供的基于大语言模型的检索增强生成方法,该方法包括:获取用户的输入查询信息;基于大语言模型分别对输入查询信息进行重写和关键信息提取,得到第一查询指令和第二查询指令;第一查询指令为包括重写后的查询信息的指令,第二查询指令为包括关键信息的指令;根据输入查询信息,基于大语言模型识别用户意图并进行意图分类,得到意图分类结果;根据第一查询指令和第二查询指令,通过预设多源知识库进行多路混合检索,得到初始检索结果;多源知识库包括互联网知识库、维基百科知识库和学术文献知识库;根据意图分类结果对初始检索结果进行结果重排和筛选,得到增强提示词;将输入查询信息和增强提示词输入大语言模型,得到大语言模型输出的检索增强生成结果。

[0056] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0057] 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各实施方式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。基于这样的理解,上述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,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,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)执行各个实施例或者实施例的某些部分所述的方法。

[0058] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管

参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

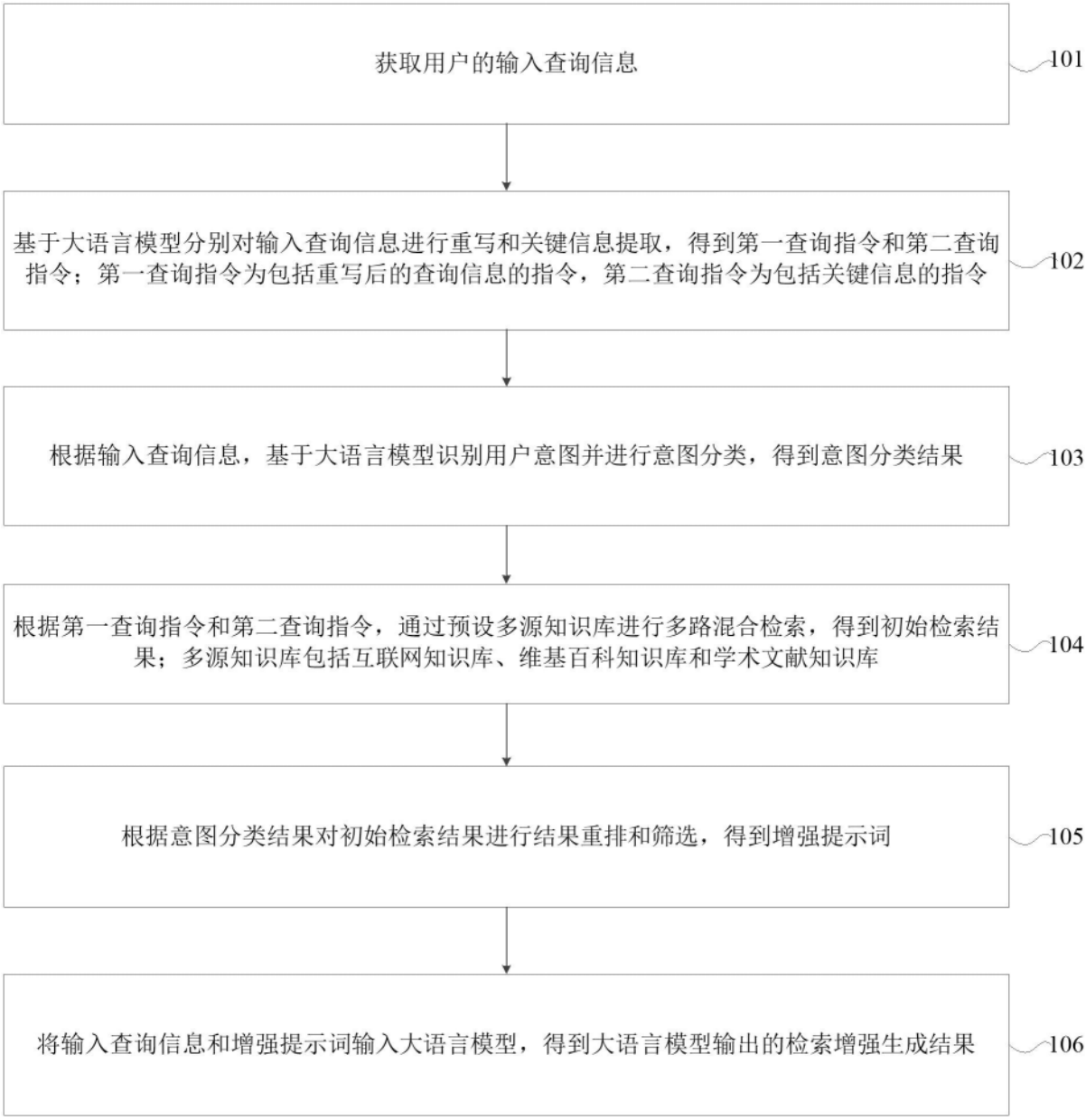

获取用户的输入查询信息
101
基于大语言模型分别对输入查询信息进行重写和关键信息提取，得到第一查询指令和第二查询指令；第一查询指令为包括重写后的查询信息的指令，第二查询指令为包括关键信息的指令
102
根据输入查询信息，基于大语言模型识别用户意图并进行意图分类，得到意图分类结果
103
根据第一查询指令和第二查询指令，通过预设多源知识库进行多路混合检索，得到初始检索结果；多源知识库包括互联网知识库、维基百科知识库和学术文献知识库
104
根据意图分类结果对初始检索结果进行结果重排和筛选，得到增强提示词
105
将输入查询信息和增强提示词输入大语言模型，得到大语言模型输出的检索增强生成结果
106

图1

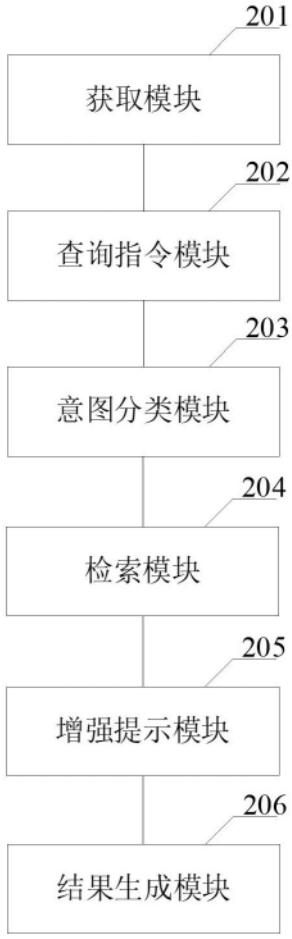

201
获取模块
202
查询指令模块
203
意图分类模块
204
检索模块
205
增强提示模块
206
结果生成模块

图2

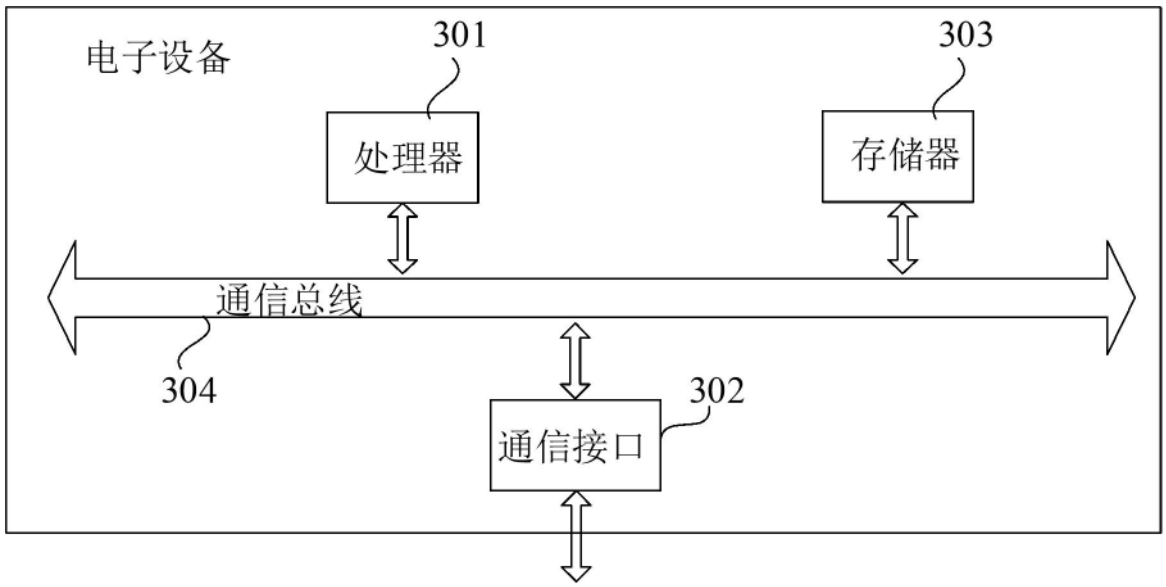

电子设备
301
处理器
303
存储器
通信总线
304
302
通信接口

图3