(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118535682 A
(43)申请公布日 2024.08.23

(21)申请号 202410745468.7
(22)申请日 2024.06.11
(71)申请人 小哆智能科技(北京)有限公司
地址 100089 北京市海淀区西北旺东路10号院东区23号楼三层346室
(72)发明人 刘晓玉
(74)专利代理机构 北京维创华成知识产权代理事务所(普通合伙) 16094
专利代理师 王玉琳
(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/332(2019.01)
G06F 16/338(2019.01)
G06F 40/30(2020.01)
G06F 40/284(2020.01)
G06F 40/216(2020.01)

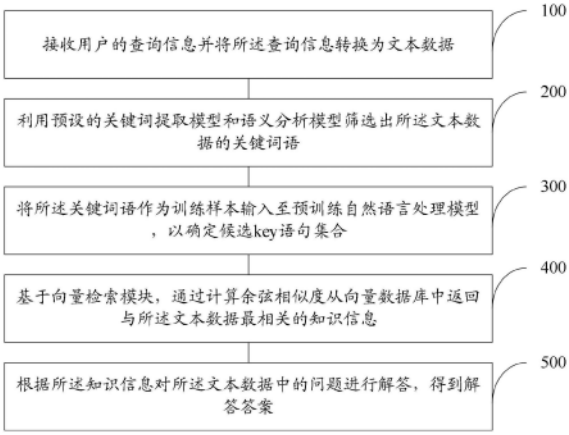
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种结合关键词提取与语义分析的检索增强方法

(57)摘要

本发明提供了一种结合关键词提取与语义分析的检索增强方法,包括:接收用户的查询信息并将所述查询信息转换为文本数据;利用预设的关键词提取模型和语义分析模型筛选出所述文本数据的关键词语;将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合;基于向量检索模块,通过计算余弦相似度从向量数据库中返回与所述文本数据最相关的知识信息;根据所述知识信息对所述文本数据中的问题进行解答,得到解答答案。本发明能够更准确地理解用户查询的信息,这大大提高了信息的检索效率和准确性,从而更好地满足用户的需要。



CN 118535682 A

1.一种结合关键词提取与语义分析的检索增强方法,其特征在于,包括:

接收用户的查询信息并将所述查询信息转换为文本数据;

利用预设的关键词提取模型和语义分析模型筛选出所述文本数据的关键词语;

将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合;

基于向量检索模块,通过计算余弦相似度从向量数据库中返回与所述文本数据最相关的知识信息;

根据所述知识信息对所述文本数据中的问题进行解答,得到解答答案。

2.根据权利要求1所述的结合关键词提取与语义分析的检索增强方法,其特征在于,所述关键词提取模型的构建方法包括:

构建第一微调数据集;所述第一微调数据集中包括用户提出的问题及问题对应的关键词;

根据所述第一微调数据集对所述预训练自然语言处理模型进行微调,得到所述关键词提取模型。

3.根据权利要求1所述的结合关键词提取与语义分析的检索增强方法,其特征在于,所述语义分析模型的构建方法包括:

构建第二微调数据集;所述第二微调数据集中包括用户提出的问题及问题对应的真实含义;

根据所述第二微调数据集对所述预训练自然语言处理模型进行微调,得到所述语义分析模型。

4.根据权利要求1所述的结合关键词提取与语义分析的检索增强方法,其特征在于,将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合,包括:

利用人工标注的方式将所述关键词语分为三类问题;所述三类问题包括:正确答案、错误答案和重复答案;

使用预训练自然语言处理模型对所述三类问题对应的关键词语分别进行微调,以使每个关键词语在对应的类别上表现出更高的准确率;

采用统计学的方法计算每条关键词语的相关性得分,并根据相关性得分排序后选择排名前3%的相关性最高的关键词语作为所述候选key语句集合。

一种结合关键词提取与语义分析的检索增强方法

技术领域

[0001] 本发明涉及信息检索技术领域,特别是涉及一种结合关键词提取与语义分析的检索增强方法。

背景技术

[0002] 随着互联网的快速发展,信息量爆炸式增长。然而传统的搜索引擎并不能满足用户多样化的需求和个性化体验的要求。因此需要一种能够提高检索结果多样性、准确性和个性化的技术来应对这一挑战。

[0003] 关键词提取是自然语言处理中的一项重要任务之一。它旨在从文本数据集中抽取具有代表性的词语或短语作为关键字进行查询匹配以获取相关知识。目前常用的方法包括基于词频统计的方法如BM25算法等以及深度学习模型如BERT-basedKeyword ExtractionModels等等。但是这些方法往往只考虑了单个单词或者短语的重要性而忽略了它们之间的关联关系。这导致了所提取的关键词可能不完整或不准确。此外,由于政务问答领域涉及的专业知识和术语较多,现有的关键词提取方法难以适应这种复杂的情况。因此有必要提出新的技术方案来解决这些问题。

发明内容

[0004] 为了克服现有技术的不足,本发明的目的是提供一种结合关键词提取与语义分析的检索增强方法。

[0005] 为实现上述目的,本发明提供了如下方案:

[0006] 一种结合关键词提取与语义分析的检索增强方法,包括:

[0007] 接收用户的查询信息并将所述查询信息转换为文本数据;

[0008] 利用预设的关键词提取模型和语义分析模型筛选出所述文本数据的关键词语;

[0009] 将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合;

[0010] 基于向量检索模块,通过计算余弦相似度从向量数据库中返回与所述文本数据最相关的知识信息;

[0011] 根据所述知识信息对所述文本数据中的问题进行解答,得到解答答案。

[0012] 优选地,所述关键词提取模型的构建方法包括:

[0013] 构建第一微调数据集;所述第一微调数据集中包括用户提出的问题及问题对应的关键词;

[0014] 根据所述第一微调数据集对所述预训练自然语言处理模型进行微调,得到所述关键词提取模型。

[0015] 优选地,所述语义分析模型的构建方法包括:

[0016] 构建第二微调数据集;所述第二微调数据集中包括用户提出的问题及问题对应的真实含义;

[0017] 根据所述第二微调数据集对所述预训练自然语言处理模型进行微调,得到所述语义分析模型。

[0018] 优选地,将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合,包括:

[0019] 利用人工标注的方式将所述关键词语分为三类问题;所述三类问题包括:正确答案、错误答案和重复答案;

[0020] 使用预训练自然语言处理模型对所述三类问题对应的关键词语分别进行微调,以使每个关键词语在对应的类别上表现出更高的准确率;

[0021] 采用统计学的方法计算每条关键词语的相关性得分,并根据相关性得分排序后选择排名前3%的相关性最高的关键词语作为所述候选key语句集合。

[0022] 根据本发明提供的具体实施例,本发明公开了以下技术效果:

[0023] 本发明提供了一种结合关键词提取与语义分析的检索增强方法,包括:接收用户的查询信息并将所述查询信息转换为文本数据;利用预设的关键词提取模型和语义分析模型筛选出所述文本数据的关键词语;将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合;基于向量检索模块,通过计算余弦相似度从向量数据库中返回与所述文本数据最相关的知识信息;根据所述知识信息对所述文本数据中的问题进行解答,得到解答答案。本发明能够更准确地理解用户查询的信息,这大大提高了信息的检索效率和准确性,从而更好地满足用户的需要。

附图说明

[0024] 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性的前提下,还可以根据这些附图获得其他的附图。

[0025] 图1为本发明实施例提供的方法流程图。

具体实施方式

[0026] 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。

[0027] 本发明的目的是提供一种结合关键词提取与语义分析的检索增强方法,能够更准确地理解用户查询的信息,这大大提高了信息的检索效率和准确性,从而更好地满足用户的需要。

[0028] 为使本发明的上述目的、特征和优点能够更加明显易懂,下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0029] 图1为本发明实施例提供的方法流程图,如图1所示,本发明提供了一种结合关键词提取与语义分析的检索增强方法,包括:

[0030] 步骤100:接收用户的查询信息并将所述查询信息转换为文本数据;

[0031] 步骤200:利用预设的关键词提取模型和语义分析模型筛选出所述文本数据的关键词语;

[0032] 步骤300:将所述关键词语作为训练样本输入至预训练自然语言处理模型,以确定候选key语句集合;

[0033] 步骤400:基于向量检索模块,通过计算余弦相似度从向量数据库中返回与所述文本数据最相关的知识信息;

[0034] 步骤500:根据所述知识信息对所述文本数据中的问题进行解答,得到解答答案。

[0035] 具体的,本实施例的关键词提取模型的构建过程如下:

[0036] 首先本实施例制作了微调数据集,其中包含用户提出的问题及其关键词。这些关键词的筛选基于词频、词性和语义等因素,以确保选出最具代表性的关键词。关键词提取大模型基于qwen大模型进行微调,通过评估词语的重要性来提取关键词。

[0037] 进一步地,本实施例的语义分析大模型的构建过程如下:

[0038] 本实施例制作了包含用户问题及其真实含义的微调数据集。这有助于模型理解和处理错别字、重复、啰嗦或难懂的语言。语义分析大模型同样基于qwen大模型(预训练自然语言处理模型)进行微调,以提高其语言理解能力和回答的准确性。

[0039] 更进一步地,本实施例的向量检索模块使用嵌入模型将非结构化数据编码为向量,并从向量库中检索知识。通过计算余弦相似度,返回与用户输入最相关的topk个知识。

[0040] 此外,本实施例还包括对话大模型,作为系统的输出模块,对话大模型接收用户输入和检索到的知识,并生成回答。该模型是在qwen大模型,即qwen1.5-14b模型和10W条专业问答知识的基础上进行微调的。

[0041] 可选地,所述方法还包括以下步骤:

[0042] 接收用户的查询信息并将其转换为文本形式;使用词频、词性和语义等因素筛选出该查询的关键词语;将这些关键词语作为训练样本提交给qwen1.5-14b大型预训练自然语言处理模型以获得相应的权重值,从而得到具有较高权重的候选关键语句集合;通过计算余弦相似度从向量数据库中返回与该查询最相关的topk个知识;根据返回的知识对query中的问题进行解答,并将答案发送至用户端。

[0043] 进一步地,本实施例首先获取包含用户提出的问题及其真实含义的微调数据集,然后利用人工标注的方式将问题分为三类,即正确答案、错误答案和啰嗦重复答案;接着使用qwen1.5-14b大型预训练的自然语言处理模型对这三类问题的句子分别进行微调,使得每个句子在对应的类别上表现出更高的准确率;最后,采用统计学的方法计算每条句子的相关性得分,并根据相关性得分排序后选择排名前3%的相关性最高的句子作为候选关键语句集合。

[0044] 更进一步地,本实施例通过遍历所有候选关键语句集合中的每一个句子来判断其是否满足条件;如果满足条件则对该句话赋予较高的分数值并在后续过程中优先考虑它所提供的相关信息;反之若不满足条件则对其给予较低的分值并且在后续过程中不予考虑它的贡献程度。

[0045] 更进一步地,本实施例通过对每一组候选关键语句集合内的各个句子依次进行评分操作来实现这一目标;其中评分的标准是根据它们能够提供有效帮助的程度而定;其次,当某一特定时间间隔内累计有超过一定数量的候选关键语句集合未能达到预期效果时,则

停止当前任务并对之前已完成的所有任务进行回顾检查以确定是否存在需要改进之处;最后,一旦发现存在需要改善之处则立即对其进行调整以确保整体性能得到提升。

[0046] 更进一步地,本实施例在对每一组候选关键语句集合内的各个句子依次进行评分操作的同时也记录下它们的原始分值以便于后续对比分析;其次,当某一特定时间间隔内累计有超过一定数量且平均分值低于某个阈值的候选关键语句集合仍未取得理想成绩时,则停止当前任务并进行回顾检查以确定是否有必要对现有策略进行优化;最后,一旦发现有必要对现有策略进行优化的情形出现即可及时采取相应措施予以解决。

[0047] 对应上述方法,本实施例还提供了一种基于qwen大模型的政务对话系统包括:一个微调数据集;一个用于构建关键字抽取模型的大模型和一个用于对用户输入进行语言理解的语义分析大模型;一个向量库和向量检索模块以及一个生成回答的对话大模型;所述方法还包括以下步骤:接收用户的查询信息并将其转换为文本形式;使用词频、词性和语义等因素筛选出该query的关键词语;将这些keyword作为训练样本提交给qwen1.5-14b大型预训练自然语言处理模型以获得相应的权重值从而得到具有较高权重的候选key语句集合;通过计算余弦相似度从vector数据库中返回与该query最相关的topk个knowledge;根据return的知识对Query中的问题进行解答并将answer发送至user端。

[0048] 本发明的有益效果如下:

[0049] (1)本发明通过使用qwen大模型进行微调,我们的技术能够更准确地理解用户查询的政务信息。这大大提高了政务信息的检索效率和准确性,从而更好地满足用户的需要;

[0050] (2)本发明结合关键词提取与语义分析的技术方法可以应用于其他领域的信息检索中,例如医疗保健或教育等领域,以提高这些领域的信息搜索能力并提供更好的服务体验;

[0051] (3)本发明整合了多个模块和技术手段来处理不同类型的数据输入,包括非结构化文本、图像等,使得整个系统的功能更加全面且灵活性更高。这将有助于解决当前互联网快速发展带来的海量信息和个性化需求之间的矛盾问题,为广大网民提供一个高效便捷的数据获取渠道和服务平台。

[0052] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似部分互相参见即可。

[0053] 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。

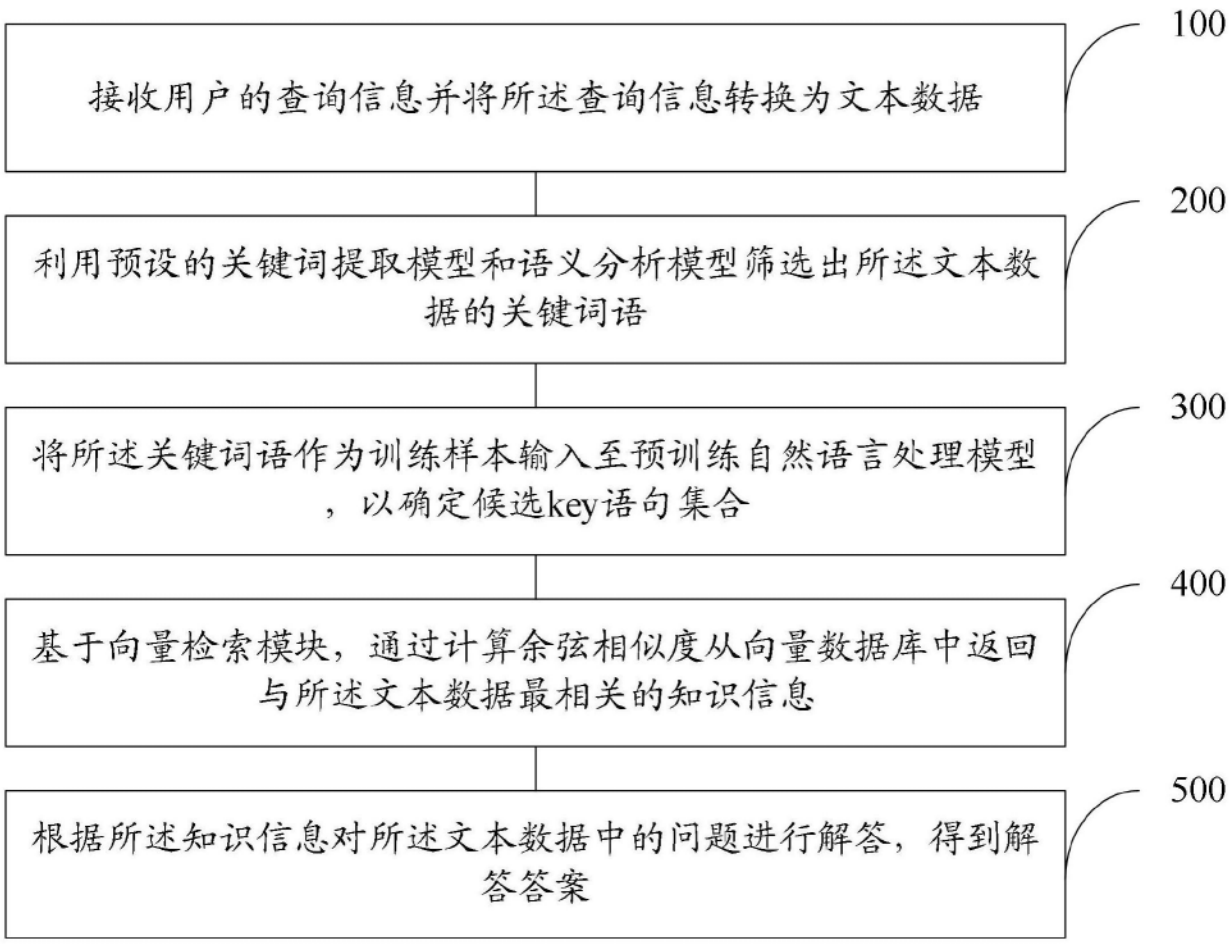

100
接收用户的查询信息并将所述查询信息转换为文本数据
200
利用预设的关键词提取模型和语义分析模型筛选出所述文本数据的关键词语
300
将所述关键词语作为训练样本输入至预训练自然语言处理模型，以确定候选key语句集合
400
基于向量检索模块，通过计算余弦相似度从向量数据库中返回与所述文本数据最相关的知识信息
500
根据所述知识信息对所述文本数据中的问题进行解答，得到解答答案

图1