(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117909455 A
(43)申请公布日 2024.04.19

(21)申请号 202311644071.0
(22)申请日 2023.12.04
(71)申请人 深圳航天智慧城市系统技术研究院有限公司
地址 518000 广东省深圳市南山区粤海街道科技南十路6号深圳航天科技创新研究院大厦B座407室
(72)发明人 于文统 陈兴晖 邓亚运 潘晓雪 汪祖茂
(74)专利代理机构 深圳市添源创鑫知识产权代理有限公司 44855
专利代理师 周椿
(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 16/36(2019.01)
G06F 40/30(2020.01)

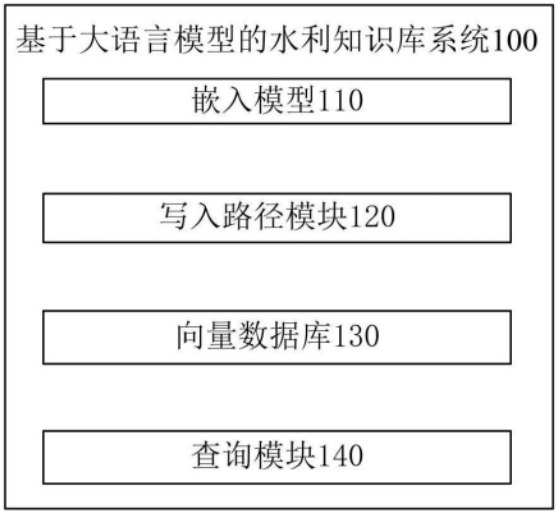
权利要求书1页 说明书7页 附图1页

(54)发明名称

一种基于大语言模型的水利知识库系统

(57)摘要

本发明公开了一种基于大语言模型的水利知识库系统,包括:嵌入模型,基于深度学习模型,将文本数据转换为向量,用于进行基于语义内容的检索;写入路径模块,将所述嵌入模型转换的向量添加到向量数据库索引中,用于迅速找到与特定查询语义上最接近的文档;向量数据库,用于存储和检索向量,并为快速语义查询提供索引;查询模块,所述查询模块将用户的查询输入转换为查询向量,在所述向量数据库中找到与查询向量语义最相近的文档,并返回关键字匹配和基于语义相似度的查询结果。本发明基于大语言模型的水利知识库系统,能自动分类、标注和更新相关内容,解决现有技术中信息更新不及时以及检索困难的问题。



CN 117909455 A

1.一种基于大语言模型的水利知识库系统,其特征在于,所述水利知识库系统包括:

嵌入模型,所述嵌入模型基于深度学习模型,将文本数据转换为向量,用于进行基于语义内容的检索;

写入路径模块,所述写入路径模块将所述嵌入模型转换的向量添加到向量数据库索引中,用于迅速找到与特定查询语义上最接近的文档;

向量数据库,所述向量数据库用于存储和检索向量,并为快速语义查询提供索引;

查询模块,所述查询模块将用户的查询输入转换为查询向量,在所述向量数据库中找到与查询向量语义最相近的文档,并返回关键字匹配和基于语义相似度的查询结果。

2.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述嵌入模型将文本数据中的词、短语或文本转化为固定长度的向量,所述向量捕获了文本的语义信息,使语义相似的文本数据拥有相近的向量。

3.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述深度学习模型为Transformer模型或BERT模型中的任意一种。

4.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述查询模块中利用的数据结构为KD-Tree或BallTree中的任意一种。

5.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述查询模块利用近似最近邻搜索法在所述向量数据库中找出最接近的向量。

6.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述查询模块还利用元数据/相似度量对查询结果进行筛选或排序,所述元数据包括文本数据的作者、日期以及相关项目。

7.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述嵌入模型将文本数据转换为向量前,先进行数据清洗,具体包括:去除无关字符并进行分词,提取关键术语。

8.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,当有大量数据输入所述向量数据库,对所述向量数据库进行索引更新。

9.根据权利要求1所述的基于大语言模型的水利知识库系统,其特征在于,所述查询模块利用余弦相似度、欧氏距离或曼哈顿距离中的任意一种方法计算查询向量与知识库文档向量之间的相似度。

一种基于大语言模型的水利知识库系统

技术领域

[0001] 本发明属于数据技术领域,具体涉及一种基于大语言模型的水利知识库系统。

背景技术

[0002] 传统水利知识库系统一般存在以下几个问题:数据分散:传统的水利知识库通常由不同部门或机构各自维护,导致重要水利数据和信息分散在多个库和平台上;标准不一:不同知识库采用各自的数据格式和标准,比如流量单位、雨量数据等,造成数据不一致,增加了后期数据整合的难度,导致信息整合异常困难;信息更新不及时:传统手动更新机制使得知识库的信息往往不是最新的,特别是在紧急或变化快速的水利事件中,信息时效性是关键。在洪水或干旱等紧急情况下,传统手动方式无法及时更新知识库,影响决策和应急响应;局限性和不完整性:传统知识库通常只能覆盖某一特定区域或主题,很难形成全面综合的知识体系,大多数知识库仅聚焦于某个特定区域(如某个流域或省份)或某个特定主题(如土壤侵蚀或水质监控),难以提供全面的水利知识;数据质量不稳定:多数数据录入和校验工作是手动进行的,易于产生错误或遗漏;检索困难:缺乏有效的数据结构和检索机制,用户常常难以快速准确地找到所需信息,传统知识库的检索功能通常仅限于基础的关键字搜索,不能满足复杂的水利业务查询,如“近五年某流域的平均降水量”等。

[0003] 现有技术中知识库系统构建,包括构建搜索引擎优化的知识库:重点关注搜索引擎的优化,确保其中存储的信息容易被搜索引擎抓取和索引,通常通过元标签、关键字密度、高质量的外部链接等SEO策略来实现;基于关键字的自动更新机制:依靠预定义的关键字或短语,自动从网络或其他数据源抓取和更新信息。一旦检测到与预定关键字相关的新信息,系统将自动进行更新。但是这些技术也存在以下缺点:缺乏行业特定复杂数据结构:虽然这种知识库对搜索引擎友好,但其内部数据结构通常较为简单,难以满足特定行业(如水利工程)的多维数据和复杂查询需求;更新不够智能:这类知识库通常依赖人工更新,因此在面对快速变化的行业信息(如水情报告、气象数据等)时,可能难以及时反应;信息质量不一:由于仅依赖于关键字匹配,该机制可能抓取到与主题不完全相关或质量不高的信息,如一些不经过专业评估的水利工程案例;缺乏内容深度:关键字匹配通常只能捕获表面信息,难以获取到深层次、结构化的知识,如水利工程中的计算模型、设计原则等。

发明内容

[0004] 本发明针对上述问题,提供了一种基于大语言模型的水利知识库系统,旨在提供一个智能化的水利知识库,能自动分类、标注和更新相关内容,解决现有技术中信息更新不及时以及检索困难的问题。

[0005] 本发明的技术方案如下:

[0006] 一种基于大语言模型的水利知识库系统,包括:

[0007] 嵌入模型,所述嵌入模型基于深度学习模型,将文本数据转换为向量,用于进行基于语义内容的检索;

[0008] 写入路径模块,所述写入路径模块将所述嵌入模型转换的向量添加到向量数据库索引中,用于迅速找到与特定查询语义上最接近的文档;

[0009] 向量数据库,所述向量数据库用于存储和检索向量,并为快速语义查询提供索引;

[0010] 查询模块,所述查询模块将用户的查询输入转换为查询向量,在所述向量数据库中找到与查询向量语义最相近的文档,并返回关键字匹配和基于语义相似度的查询结果。

[0011] 在一些实施例中,所述嵌入模型将文本数据中的词、短语或文本转化为固定长度的向量,所述向量捕获了文本的语义信息,使语义相似的文本数据拥有相近的向量。

[0012] 在一些实施例中,所述深度学习模型为Transformer模型或BERT模型中的任意一种。

[0013] 在一些实施例中,所述查询模块中利用的数据结构为KD-Tree或BallTree中的任意一种。

[0014] 在一些实施例中,所述查询模块利用近似最近邻搜索法在所述向量数据库中找出最接近的向量。

[0015] 在一些实施例中,所述查询模块还利用元数据/相似度量对查询结果进行筛选或排序,所述元数据包括文本数据的作者、日期以及相关项目。

[0016] 在一些实施例中,所述嵌入模型将文本数据转换为向量前,先进行数据清洗,具体包括:去除无关字符并进行分词,提取关键术语。

[0017] 在一些实施例中,当有大量数据输入所述向量数据库,对所述向量数据库进行索引更新。

[0018] 在一些实施例中,所述查询模块利用余弦相似度、欧氏距离或曼哈顿距离中的任意一种方法计算查询向量与知识库文档向量之间的相似度。

[0019] 本公开实施例提供的技术方案:基于大语言模型的水利知识库系统,水利信息整合:将水资源管理、灌溉系统、洪水控制等多个水利部门和子领域的信息整合到一个统一的平台;数据和标准统一:实现流量、雨量、水文地质等多方面水利数据的统一标准化处理;紧急水情和灾害信息的实时更新:利用大语言模型进行快速数据分析和更新,提高应急响应能力;全面覆盖水利业务:通过自然语言处理和大数据分析,从各个角度和维度提供全面的水利知识和信息;提高水文数据和工程信息的质量:自动校验水文数据、渠道尺寸、水坝数据等,减少人为错误;多维度的水利业务检索:构建高级检索功能,支持复杂的水利业务查询需求。综上所述,本发明提供了一个全面、高效、可靠和智能化的水利知识库系统,其产生的有益效果如下:

[0020] 1.智能化:利用大语言模型的自然语言处理能力,提供了一个智能化的水利知识库,能自动分类、标注和更新相关内容;

[0021] 2.高维数据处理:通过集成向量数据库,能高效地处理和检索高维数据,如水质参数、气象数据等,从而满足水利工程专业需求;

[0022] 3.质量与深度:采用先进的自然语言理解和数据分析技术,以确保知识库中的信息不仅广泛而且深入,涵盖从基础数据到高级分析的所有方面。

[0023] 4.成本与效率:能自动进行数据抓取、分类和更新,可以大大降低人工成本和时间成本。

[0024] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不

能限制本公开。

附图说明

[0025] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本发明的实施例,并与说明书一起用于解释本发明的原理

[0026] 图1是本发明实施例中基于大语言模型的水利知识库系统结构图。

具体实施方式

[0027] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是,此处所描述的具体实施例仅用于解释本发明,而非对本发明的限定。另外还需要说明的是,为了便于描述,附图中仅出示了与本发明相关的部分而非全部结构。

[0028] 在更加详细地讨论示例性实施例之前应当提到的是,除了实施例中涉及的模块以外,实施例系统还可以包括其他部件,由于这些部件与本公开实施例的内容无关,因此在实施例中省略其中一些图示和描述。

[0029] 大语言模型(LLM)和向量数据库,大语言模型是基于神经网络的模型,用于处理和理解自然语言,特别是在处理大规模文本数据集方面表现出色。模型通常采用深度学习方法,如Transformer架构,具有数百万到数十亿的参数;向量数据库是用于高效存储和检索高维向量数据的数据库系统,通常用于处理机器学习模型生成的复杂数据结构,其可使用一种或多种高效的索引结构,如K-D树、球树(Ball Tree)或近似最近邻搜索(ANN)等,以实现快速查询。与传统的关系数据库相比,向量数据库在处理高维、非结构化数据方面有明显的优势。

[0030] 本发明实施例基于大语言模型的水利知识库系统,如图1所示,基于LLM构建的知识库系统100主要由嵌入模型110(Embedding Model)、写入路径模块120(Write Path)、向量数据库130(Vector Database)、查询模块140(Query Path)4个模块组成,各模块说明如下:

[0031] 嵌入模型110,嵌入模型110基于深度学习模型,将文本数据转换为向量,用于进行基于语义内容的检索,其中文本数据包括与水利相关的数据。

[0032] 嵌入模型110将文本数据转换为向量前,先进行数据清洗,具体包括:去除无关字符并进行分词,提取关键术语。

[0033] 嵌入模型110将文本数据中的词、短语或文本转化为固定长度的向量,向量捕获了文本的语义信息,使语义相似的文本数据拥有相近的向量。

[0034] 嵌入模型110实现的基础具体包括:

[0035] 语义编码:嵌入模型110是将词、短语或文本转化为固定长度的向量。这些向量能够捕获文本的语义信息,使得语义上相似的内容拥有相近的向量。

[0036] 深度学习:利用神经网络模型/深度学习模型,优选Transformer或BERT,从大量的水利文本中学习语义。优选实施例中,同义词“水库”和“蓄水池”在嵌入空间中会有相近的向量表示。

[0037] 特定领域训练:在水利领域内,使用相关文献、研究报告等对模型进行微调,使其更好地理解水利专业术语与知识。

[0038] 具体实施例中,嵌入模型110模型基于深度学习技术,优选为Transformer或BERT结构,将文本数据转换为高维空间向量。在此空间里,语义上相似的文档将被映射到相近的位置。水利专业的文档和资料,如河流数据、土壤类型、工程报告等,通过嵌入模型110被转换为向量,这使得在后续的查询中可以基于语义内容而非仅基于关键字来检索信息。

[0039] 写入路径模块120,写入路径模块120将嵌入模型110转换的向量添加到向量数据库索引中,用于迅速找到与特定查询语义上最接近的文档。

[0040] 写入路径模块120(Write Path)实现基础包括:

[0041] 文档/数据输入:当新的水利文档、研究报告或实时数据被添加到知识库时,它们首先被传入嵌入模型110。

[0042] 元数据(Metadata):每份文档或数据都可能带有元数据,实施例中有作者、出版日期、所属项目等。这些元数据对于后续的分类、过滤或排序检索结果非常重要。

[0043] 索引(Indexing):一旦文档被转换为向量形式,它们将被添加到向量数据库的索引中。这使得在后续的查询中,系统可以迅速找到与特定查询语义上最接近的文档。

[0044] 向量数据库130,向量数据库130用于存储和检索向量,并为快速语义查询提供索引;

[0045] 向量数据库130(Vector Database)使用高效的数据结构,优选为KD-Tree或BallTree,来存储和检索高维向量。在水利知识库中存储与水利相关的所有文档的向量表示,并为快速查询提供索引。当有大量数据输入向量数据库130,对向量数据库130进行索引更新。

[0046] 向量数据库130实现方法具体包括:

[0047] 高维空间存储:向量数据库130是设计来存储和管理高维向量的。在水利知识库中,每一篇文献、报告或文档的向量表示都被存储在此。

[0048] 分布式存储:为了应对大量数据的存储和检索需求,向量数据库130可以分布在多个服务器或集群上。

[0049] 实时更新:随着新的水利数据的产生,可以实时将其转化为向量并加入到向量数据库130中,确保知识库的时效性。

[0050] 查询模块140,查询模块将用户的查询输入转换为查询向量,在向量数据库中找到与查询向量语义最相近的文档,并返回关键字匹配和基于语义相似度的查询结果。查询模块140利用余弦相似度、欧氏距离或曼哈顿距离中的任意一种方法计算查询向量与知识库文档向量之间的相似度。

[0051] 查询模块140(Query Path)包括:

[0052] 查询输入:用户可能基于特定的问题、关键字或主题进行查询,一些实施例中的“河流侵蚀的主要原因是什么”;

[0053] 查询模块140利用近似最近邻搜索法在向量数据库中找出最接近的向量。

[0054] 近似最近邻搜索(ANN):向量空间中的高效搜索算法,当用户发起查询时,查询首先被转换为向量,然后使用ANN在向量数据库中找到与查询语义上最相近的文档;

[0055] 结果呈现:返回的文档或数据不仅基于关键字匹配,而且基于其与查询的语义相似度,这确保了用户获得的答案是与其查询最相关的。

[0056] 查询模块140还利用元数据对查询结果进行筛选或排序,元数据包括文本数据的

作者、日期以及相关项目。

[0057] 实施例中向量索引与查询的具体实施方式包括:

[0058] 高效的数据结构:利用如KD-Tree、BallTree等高效数据结构,对高维向量进行索引,从而实现高速查询。

[0059] 近似最近邻搜索(ANN):当用户提出查询时,首先将查询转化为向量,然后在向量数据库中快速找出最接近的向量,也即最相关的文档。

[0060] 元数据匹配:结合向量查询的结果,再利用元数据(如作者、日期、相关项目等)或相似度量进行进一步的筛选或排序,以提供更为精准的答案。

[0061] 实施例中水利知识库系统利用了大语言模型的嵌入技术和向量数据库的高效检索技术,为水利领域的工程师提供了一个强大、灵活且高效的查询工具。

[0062] 实施例中提供的水利知识库系统具有以下特定:

[0063] 大语言模型集成:用于自然语言处理和理解,实现知识库中的信息自动分类、标注和更新。

[0064] 向量数据库应用:高效地处理和检索高维数据,支持复杂查询。

[0065] 智能匹配与过滤机制:采用先进的机器学习算法,以确保信息质量和内容深度。

[0066] 动态更新与校验:结合大语言模型和向量数据库,实现知识库的实时或定期自动更新和数据质量校验。

[0067] 行业特定模板:针对水利工程的多维需求,设计特定的数据模板和查询界面。

[0068] 实施水利知识库系统中大语言模型和向量数据库相结合的具体算法和工作流程。

[0069] 具体实施例还包括如下:

[0070] 数据预处理和嵌入:文本清洗、分词、去噪。

[0071] 工作流程:

[0072] (1)收集数据:汇集水利领域的文本数据,如报告、文章、研究论文等。

[0073] (2)预处理:清洗数据,去除无关字符,分词,提取关键术语。

[0074] (3)嵌入:使用训练好的大语言模型(如BERT、GPT等)将预处理后的文本转换为向量。

[0075] 向量索引

[0076] 算法:KD-Tree、BallTree、或其他近似最近邻(ANN)算法。

[0077] 工作流程:

[0078] (1)构建索引:将文本向量输入向量数据库,创建索引。

[0079] (2)优化索引:对索引结构进行优化,以加快查询速度。

[0080] 查询执行

[0081] 算法:向量相似度计算(如余弦相似度)。

[0082] 工作流程:

[0083] (1)接收查询:用户输入查询请求,可能是自然语言问题或特定的查询指令。

[0084] (2)查询嵌入:将查询文本通过嵌入模型转换为向量。

[0085] (3)执行搜索:使用ANN算法在向量数据库中找到最接近的向量。

[0086] (4)返回结果:根据向量的匹配程度,返回最相关的文档列表。

[0087] 用户交互和反馈

[0088] 算法:排序算法、反馈学习机制。

[0089] 工作流程:

[0090] (1)呈现结果:将搜索结果按相关性排序并展示给用户。

[0091] (2)收集反馈:用户对搜索结果的点击、阅读时间等行为作为反馈。

[0092] (3)优化模型:利用反馈信息对嵌入模型和搜索算法进行微调。

[0093] 系统更新与维护

[0094] 算法:在线学习、增量学习。

[0095] 工作流程:

[0096] (1)持续学习:随着新数据的累积,定期更新嵌入模型。

[0097] (2)索引维护:定期重新索引以优化性能,特别是在大量新数据加入时。

[0098] (3)性能监控:监控查询响应时间和准确率,确保系统质量。

[0099] 系统数据结构是专为水利工程设计的高维数据结构和模板。

[0100] 系统用户界面是专门针对水利工程需求设计的用户查询和管理界面。

[0101] 系统安全机制包括数据加密、用户验证和权限管理等安全措施。

[0102] 系统智能匹配与过滤机制的具体算法和实现细节,以保证信息质量:

[0103] 自然语言理解(NLU)算法:

[0104] 目的:理解用户查询的语义意图。

[0105] 实现:利用预训练的大语言模型(如BERT、GPT)来提取语义特征。

[0106] 文本分类算法:

[0107] 目的:将查询分类到水利知识库的相关子领域。

[0108] 实现:使用支持向量机(SVM)、神经网络或决策树等分类器。

[0109] 相似度计算算法:

[0110] 目的:计算查询向量与知识库文档向量之间的相似度。

[0111] 实现:常用的相似度量包括余弦相似度、欧氏距离或曼哈顿距离。

[0112] 排序和推荐算法:

[0113] 目的:根据匹配程度对搜索结果进行排序。

[0114] 实现:使用学习到排序(Learning to Rank)算法,如RankNet、LambdaMART。反馈学习算法:

[0115] 目的:根据用户行为优化匹配和过滤效果。

[0116] 实现:采用在线学习策略,如在线梯度下降法,不断微调模型参数。

[0117] 具体实施过程中:

[0118] 特征工程:

[0119] 文本预处理:清洗、标准化、去噪和分词等。

[0120] 特征提取:从文本中提取关键词、短语和其他语言特征。

[0121] 模型训练与验证:

[0122] 数据标注:对水利领域的文档进行标注,为模型训练提供监督信号。

[0123] 交叉验证:对模型进行验证,确保其泛化能力。

[0124] 查询处理:

[0125] 实时嵌入:将用户查询转换为向量。

[0126] 查询优化:根据历史数据调整查询参数,优化检索效果。

[0127] 结果过滤与排序:

[0128] 过滤逻辑:根据用户设定的条件(如时间范围、地理位置等)过滤结果。

[0129] 动态排序:结合用户反馈和行为数据,动态调整结果的排序。

[0130] 用户反馈循环:

[0131] 监测与分析:跟踪用户与系统的互动,分析点击率、阅读时间等。

[0132] 模型更新:使用反馈数据来微调和优化匹配与过滤模型。

[0133] 综合上述各实施例提供的技术方案,一种基于大语言模型的水利知识库系统,其具体方案及有益效果如下:

[0134] 1.智能化与高度自动化:与基于搜索引擎优化和关键字的现有技术相比,本发明通过集成大语言模型,实现了高度智能化和自动化的知识库构建与维护,减少了人工介入的需要,提高了效率,并且能更准确地获取和分类信息。

[0135] 2.支持高维和复杂查询:现有技术主要侧重于文本数据,并不支持高维数据或复杂查询。通过集成向量数据库,本发明能够方便地处理和检索水利工程中的高维数据,如气象参数、水质数据等。

[0136] 3.内容质量与深度:利用大语言模型的先进自然语言理解能力,本发明不仅能捕捉更广泛的信息,还能提供更深入的分析和理解,如涉及水利工程的计算模型、法规和设计原则等。

[0137] 4.行业特定优化:本发明针对水利工程领域的特点和需求,设计了行业特定的数据结构和用户界面,从而更加贴近实际应用需求。

[0138] 5.成本效益:由于减少了人工干预和提高了自动化程度,本发明在长期运营中可能具有更低的维护成本和更高的数据准确性,从而具有更好的成本效益。

[0139] 6.数据安全与完整性:本发明通过先进的安全机制如数据加密、用户验证和权限管理等,确保了数据的安全性和完整性。

[0140] 7.可扩展性:本发明采用模块化设计,具有很强的可扩展性,意味着可以轻松地添加更多的数据源、分析工具或其他功能,以适应水利工程领域不断发展和变化的需求。

[0141] 8.多语言与跨平台支持:利用大语言模型的多语言能力,本发明可以支持多种语言,从而方便全球范围内的水利工程专家和决策者使用。同时,由于采用了云基础架构,本发明可以跨多种设备和操作系统运行。

[0142] 9.环境适应性:针对水利工程中可能出现的各种复杂环境因素(如极端天气、地理差异等),本发明可以通过实时更新和智能算法,提供更为精准和实用的信息和解决方案。

[0143] 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步骤、方法不仅包括那些要素,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步骤、方法所固有的要素。

[0144] 以上内容是结合具体的优选实施方式对本发明所作的进一步详细说明,不能认定本发明的具体实施只局限于这些说明。对于本发明所属技术领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干简单推演或替换,都应当视为属于本发明的保护范围。

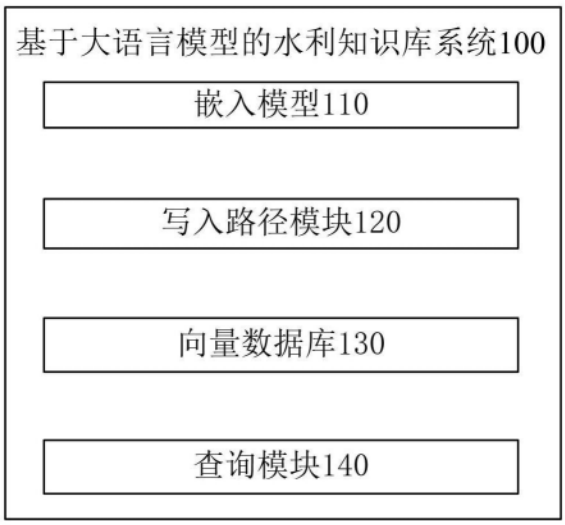


图1