

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来





开源向量数据库Vearch架构设计 及大模型AIGC时代新特征与展望

北京阿哇科技有限公司 李洁

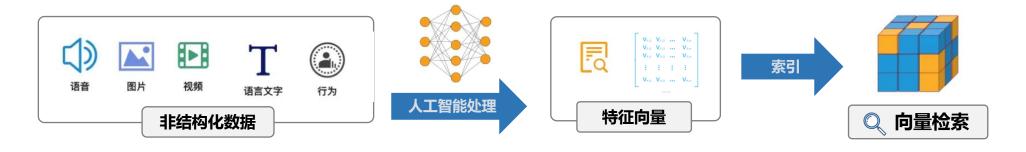


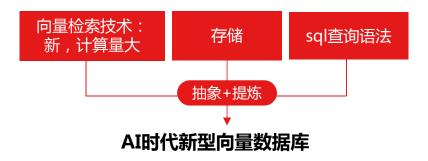




DTCC 2023 第十四届中国数据库技术大会 DATABASE TECHNOLOGY CONFERENCE CHINA 2023

向量数据库背景介绍





价值:应用场景广泛,解放算法工程师

以视觉为代表的第一代向量数据库主要场景:搜索/推荐/广告场景 向量召回

人脸识别

安防

以图搜图









DTCC 2023 第十四届中国数据库技术大会

向量数据库核心特征

与AI关系密切,向量一般需要通过神经网络来生成,表征数据内容

向量检索是其核心,不同数据量不同场景,有不同索引检索算法。数据量大需要采用近似最近邻算法来建索引

传统数据库查询具有0/1唯一性,向量检索是模糊相似性查找,用召回率评价查询结果好坏

向量检索为CPU密集性, CPU消耗远胜传统数据库

对事务没有要求





开源向量数据库Vearch

2018年初 2018年9月 2019年6月 2019年9月 2019年10月-

同款推荐/重复铺货等需求, 构建十亿规模的实时分布式图 片搜索系统

The Design and Implementation of a Real Time Visual Search System on JD E-commerce Platform.

国际中间件顶级会议

抽象出vearch系统,设计并实 现vearch支撑十亿量级同款推 荐和重复铺货等业务

业内第一个开源的支持百亿量 级向量存储检索的分布式向量 数据库

支持更多业务场景

- -Python SDK、gRPC、多索引
- -节省存储(内存+磁盘)空间
- -云原生、T-PAAS

重复铺货 | 同款推荐 | 图片侵权治理 | 敏感头像检测 | 新 图片场景 品识别 | ... 智能问答词向量 | 用户意图分析 | 风控团伙作弊治理 文本场景 视频场景 视频去重 | ...





































搜索/推荐/广告召回



Star 1.7k

https://github.com/vearch/vearch

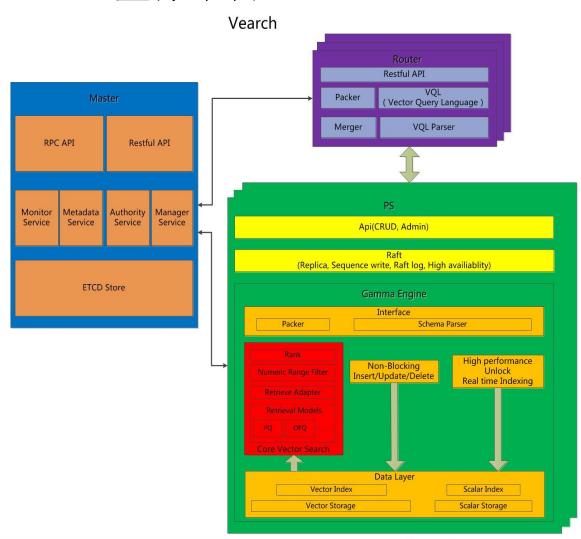






DTCC 2023 第十四届中国数据库技术大会

Vearch整体架构



- Master:元数据管理

-Router:提供restful API

- PartitionServer: 存储与检索引擎

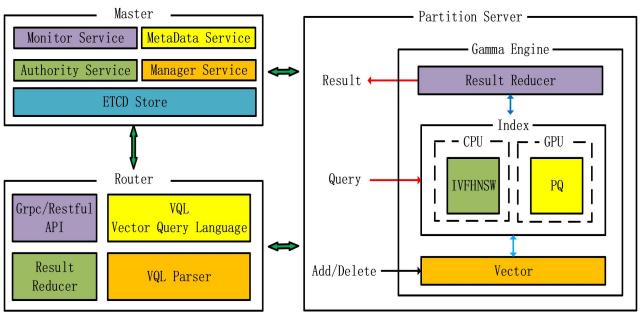




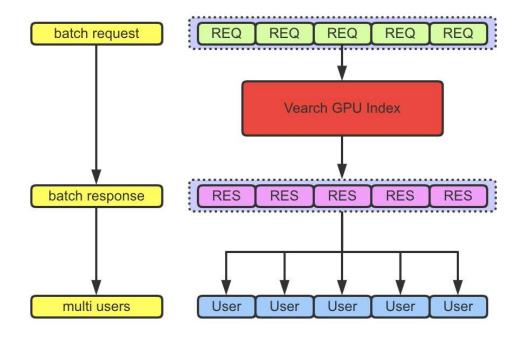


DTCC 2023 第十四届中国数据库技术大会

支持GPU/CPU异构混合架构







GPU Online 优化







核心模块及技术-核心功能

DTCC 2023 第十四届中国数据库技术大会 DATABASE TECHNOLOGY CONFERENCE CHINA 2023

实时向量索引



高性能无锁 内存可动态扩展,延迟释放等技术 实现添加数据同时支持索引

支持内存、磁盘两种数据存储方式



字段过滤



通过数值字段范围过滤、标签过滤大幅提升性能

自研gamma引擎,提供 高性能的向量检索



分布式



横向弹性扩展,支持几w到十亿,甚至百亿的数据规模 基于raft协议实现数据多副本存储

适应不同数据规模的最优召回匹配模型 IVFPQ针对亿级或者更多 HNSW耗内存但查询精度高,适用干万级数据 二进制索引用于处理unit8数据场景













核心模块及技术-向量检索模型

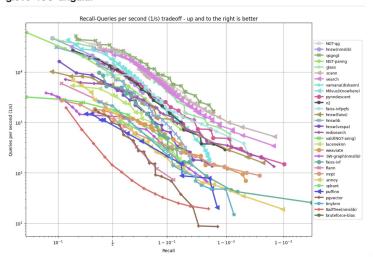
> IVFPQ

PQ量化索引,内存占用小,召回有一定精度损失,适 用亿级以上大规模数据

> HNSW & SSG

按多层次图构建索引, 召回精度高, 占用内存大, 适 用千万级规模数据

glove-100-angular



- 二进制索引

汉明距离计算

> IVFFLAT

索引按聚类质心分桶,对百万规模数据集暴力搜索(消耗较大计算资源),召回率能到100%针对大batch情况下,定向优化,性能提升比较明显

第三方公平公开评测

benchmark: https://github.com/erikbern/ann-benchmarks









DTCC 2023 第十四届中国数据库技术大会

核心模块及技术-存储&云原生&其他

- ▶ 基于Raft协议,支持多备份的数据强一致分布式系统
- > 支持单文档多向量的存储方式
- > 支持向量数据的压缩
- ▶ 支持原始数据(包括标量与向量)及索引的持久化
- ➤ 支持K8S集群以云原生的方式大规模部署
- ▶ 横向扩展副本方便
- ▶ 灵活可直接使用的Restful接口
- ▶ 通过算法插件服务,支持端到端、完整应用场景解决方案







大模型AIGC时代向量数据库的新特征



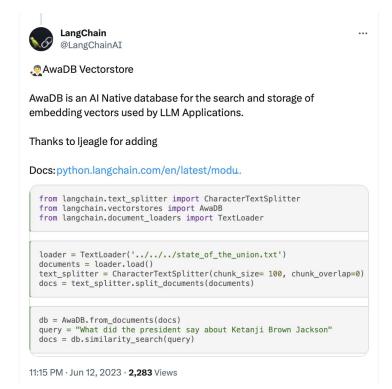
ChatGPT爆发对向量数据库场景需求由视觉转向自然语言

门槛大幅降低, 要求更灵活, 轻量

实时性要求更高

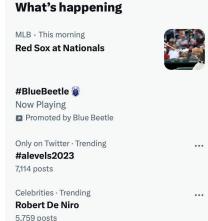
多表

高并发、低成本



Relevant people



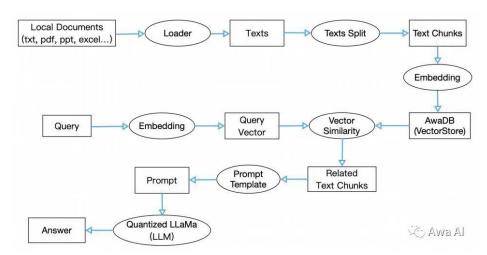








AI Native新型向量数据库-AwaDB



大模型+向量数据库构建本地知识库流程图

1: 加载本地关于《流浪地球2》的电影介绍文章

documents = TextLoader("./data/liulangdiqiu.txt", encoding="utf-8").load()

2: 对文章按每70个字符做切分,同时AwaDB默认会对每70个字符的文本片段执行embedding成语义向量的操作,写入指定表langchain_awadb_qa
docsearch = AwaDB.from_documents(RecursiveCharacterTextSplitter(chunk_size=70, chunk_overlap=0, keep_separator=False).split_documents(documents), table_name='langchain_awadb_qa')

3-6: 自定义上下文问答提示词模版

prompt_template = """使用以下上下文来回答最后的问题。如果你不知道答案,就说你不知道,不要试图编造答案。{context}

问题: {question}

答案:"""

PROMPT = PromptTemplate(template=prompt_template, input_variables=["context", "question"])

- #7: langchain内置的问答检索类链式调用。大模型基于Llama系列中斯坦福微调的Alpaca-7b,并通过中文数据微调,最后经llama.cpp量化压缩到原始大小的1/8。
- # LlamaCpp类中的model path请设置为本地经过量化后的实际大模型地址
- qa = RetrievalQA.from_chain_type(llm=LlamaCpp(model_path="../models/alpaca_chinese_lora_7b/ggml-model-q4_0.bin", n_ctx=2048, temperature=0), chain_type="stuff", retriever=docsearch.as_retriever(), chain_type_kwargs={"prompt": PROMPT})

8: 问答检索类链式调用执行对应的问题推理,并返回答案qa.run("流浪地球讲的什么故事?")

8行代码LLaMa+AwaDB本地电脑CPU推理构建知识库示例https://github.com/awa-ai/awadb/tree/main/examples/llama.cpp

DTCC 2023 第十四届中国数据库技术大会 DATABASE TECHNOLOGY CONFERENCE CHINA 2023

Demo

以有关三体问题 你知道面壁计划吗,你知道面壁者都有谁吗 为例:

未使用接入本地文档时:

面壁计划是指爱因斯坦在1921年提出的一个计划,旨在解决数学和物理学中的一些难题,包括狭义相对论和广义相对论等。这个计划被称为"面壁者计划",因为它是指爱因斯坦在实验室里进行的研究,而不是指他在实际环境中进行的科学研究。由于爱因斯坦在二战期间遭受了严重的头郎创伤,他身体健康状况逐渐恶化,他最终在1955年逝世,享年76岁。因此,"面壁者计划"通常指的是他的相对论研究计划。在"面壁者计划"中,爱因斯坦提出了一些重要的数学和物理学理论,包括狭义相对论、广义相对论、质能方程、光电效应、波动方程等。他的理论和思想对现代物理学和数学的发展产生了深远的影响。

使用 awadb 构建了本地三体数据库后 ChatGLM 的回答

面壁计划是指由联合国行星防御理事会选定的四位面壁者,他们的任务是领导战略计划的制订和执行,并建立一个扑朔迷离的假象迷宫,以推迟敌人判明我方真实战略意图的时间。面壁者将被授予很高的权力,调集和使用地球已有的战争资源中的一部分,但不需要对自己的行为和命令做出解释。面壁计划是一项人类历史上最艰难的使命,所有面壁者将孤独地走过漫长的岁月。

以关于本数据库 awadb 的问题 你知道 awadb 是什么吗? 为例:

未使用接入本地文档时

抱歉,我不知道awadb是什么。我是一个基于自然语言处理技术的人工智能助手,我的知识来自于已经存储在我的数据库中的大量文本 数据。我无法浏览互联网或访问外部信息,因此我可能无法提供有关最新事件或信息的答案。如果您有任何其他问题,我将尽力回 签

增加了 awadb 的 README.md 文件后 ChatGLM 的回答

是的,我知道awadb是什么。awadb是一款为大语言模型LLM打造的向量数据库,用于存储和快速搜索模型编码后的向量,帮助打造基于个人知识库的大模型应用。它是基于它语言开发的,并提供Python接口,可以直接通过pip安装。awadb支持OpenAl, Llama, Vicuna, Alpaca, ChatGLM, Dolly等模型,以及LangChain库

ChatGLM+AwaDB构建本地知识库示例 https://github.com/awa-ai/awadb/tree/main/examples/chatglm









TemporalDat

CloudnativeDat

Alalgorithn

Distribute