软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(3):969−970 [doi: 10.13328/j.cnki.jos.007520] [CSTR: 32375.14.jos.007520]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

向量数据库及 DB4LLM 技术专题前言*

高 宏[1], 李国良[2], 张 蓉[3]

[1](浙江师范大学 计算机科学与技术学院, 浙江 杭州 321004)

[2](清华大学 计算机科学与技术系, 北京 100084)

[3](华东师范大学 数据科学与工程学院, 上海 200062)

通信作者: 李国良, E-mail: liguoliang@tsinghua.edu.cn

中文引用格式: 高宏, 李国良, 张蓉. 向量数据库及DB4LLM技术专题前言. 软件学报, 2026, 37(3): 969–970. http://www.jos.org.cn/1000-9825/7520.htm

向量数据库是一种专门用于管理高维数据的数据库系统. 与传统数据库系统主要处理结构化数据不同, 向量数据库专注于存储和检索以向量形式表示的数据, 这些数据可以是文本、图像、音频等复杂数据类型. 在大数据和人工智能技术的驱动下, 向量数据库已经广泛应用于信息检索、推荐系统、语音识别、图像分析等领域, 成为数字化转型和智能化发展的重要推动力. 然而, 面对海量高维向量数据的实时更新、存储、检索与分析需求, 特别是来自大语言模型 (large language model, LLM) 等人工智能技术的应用需求, 传统数据库技术还面临着许多挑战, 亟须发展新的数据库理论、方法与技术, 提高向量数据库对实际应用场景的支持能力. 本专题聚焦于“向量数据库及 DB4LLM 技术”主题, 重点关注向量数据库及 DB4LLM 技术研究中具有创新性和突破性的高水平研究成果, 探讨相关基础理论、关键技术, 以及在系统研发过程中关于系统设计原理、范式、架构、经验等方面的实质性进展, 探讨其在相关产业和领域的应用前景以及关键挑战.

本专题公开征文, 共收到投稿 16 篇. 论文均通过了形式审查, 内容涉及向量数据库的数据存储、索引维护、查询处理与优化、向量检索. 特约编辑先后邀请了 20 多位专家参与审稿工作, 每篇投稿至少邀请 2 位专家进行评审. 稿件经初审、复审、NDBC 2025 会议宣读和终审 4 个阶段, 历时 4 个月, 最终有 8 篇论文入选本专题. 根据主题, 这些论文可以分为 4 组.

(1) 向量检索

《向量数据库中近似最近邻搜索关键技术综述》针对面向高维向量近似最近邻搜索 (approximate nearest neighbor search, ANNS) 的新方法和研究成果系统性梳理缺失引起的技术理解难和应用推进慢的问题, 全面地总结和分类现有方法, 概述了它们所针对的问题和解决思路; 并基于当前的研究成果和研究趋势, 完成了对未来研究方向的展望, 从而为向量搜索和向量数据库的研究和设计提供可靠的参考和借鉴.

《向量数据库的 K 近邻图高效更新方法》针对嵌入模型微调通常会导致全部数据的向量表示发生系统性变化, 从而使原有 K 近邻图的邻接关系失效的问题, 提出一种面向嵌入模型微调场景的高效 K 近邻图更新方法, 解决了面向嵌入模型微调后的 K 近邻图质量问题. 该方法基于嵌入模型微调为每条数据嵌入带来的影响较小的观察, 通过局部更新策略对原始 K 近邻图进行增量调整, 最终实现了在确保最终 K 近邻图质量的同时, 显著提升更新效率.

《GoVector: I/O-高效的高维向量近邻查询缓存策略》针对存储大量邻接关系的图结构高维向量索引 (索引图) 存在由于索引访问频繁发生 I/O 操作导致的性能瓶颈问题, 提出了一个静态-动态混合缓存策略, 实现了缓存命中率的显著提升, 并有效降低了整体 I/O 开销. 论文利用静态缓存区预加载入口点及其高频近邻, 利用动态缓存区自适应地缓存空间局部性高的顶点. 这种双重优化机制提升了吞吐降低了延时.

(2) 向量数据管理

《GPU 加速的高维向量聚类算法》针对现有的基于密度的聚类算法在处理高维向量数据时将产生极高的时

* 收稿时间: 2025-09-08; jos 在线出版时间: 2025-09-09

间代价, 且面临维度灾难问题, 以及面向大规模高维向量基于 CPU 的聚类算法产生的高时间代价和低可扩展性问题, 提出一种 GPU 加速的高维向量聚类算法, 通过引入 K 近邻图索引实现了加速 DBSCAN 计算的目的, 推动了该类方法在实际场景中的部署和应用.

(3) 索引构建与优化

《LSMDiskANN: 更新友好型磁盘向量索引框架》针对当前领先算法 FreshDiskANN 在查询与更新混合负载场景中面临的查询吞吐瓶颈和极端查询延迟过高等问题, 提出了一种基于 LSM 思想的更新友好型磁盘向量索引框架; 在继承 FreshDiskANN 架构的基础上, 设计并实现了包含磁盘中间层的三层架构, 同时引入了磁盘组件搜索参数的动态确定机制以及面向合并操作删除阶段的重布局算法, 解决了在磁盘级向量索引中高效支持大规模向量的更新操作的同时提供高性能查询服务的问题, 实现了系统在混合负载场景下的服务整体性能与稳定性的有效提升.

《面向批量更新的向量索引召回率优化》针对向量数据库中数据以批量方式进行更新时, 造成 HNSW (hierarchical navigable small world) 算法中启发式剪枝算法的有效性降低以及相似向量连接稀疏化的问题, 提出一种基于图结构局部调整的自适应细粒度剪枝策略, 构建了融合“识别与修复”机制的优化方案, 实现了在保证检索精度的同时提升邻居连接的多样性目标, 有效缓解了过度剪枝与连接稀疏化问题.

《权重残差向量量化: 向量压缩与分层索引结构》针对现有向量数据库在存储索引效率、索引构建复杂度及检索准确性方面存在的显著瓶颈问题, 提出一种基于权重残差向量量化 (weight residual vector quantization, WRVQ) 的新型框架, 实现了对近似最近邻检索的有效支持. 论文通过将量化方向与残差长度分离处理, 以单位向量形式存储残差方向并附加权重标记, 完成了低失真率下的高效压缩与存储; 设计了适配 WRVQ 量化特性的三层倒排索引结构, 有机结合非对称距离计算与近邻搜索技术, 达成了高准确度与高效率兼具的近似最近邻检索.

(4) 自然语言查询

《基于大语言模型的空间数据库自然语言查询转换方法》针对空间数据固有特征引起的现有 Text2SQL 技术难以直接适用于空间数据库领域的问题, 提出了面向空间数据库的自然语言查询转换方法, 实现了从用户自然语言查询到空间数据库可执行查询语言的高效转换. 论文在自然语言理解阶段使用实体信息提取算法提取关键查询实体, 并基于大语言模型构建空间数据查询语料库, 进而确定查询类型; 在可执行语言生成阶段根据查询类型选择结构化语言模型, 然后将实体映射到结构化语言模型中生成空间数据库可执行语言. 最终达到了降低用户与空间数据库的交互门槛的目标.

本专题主要面向数据库、数据挖掘、大数据、人工智能、推荐系统等多领域的研究人员和工程人员, 展示了我国学者在向量数据库及 DB4LLM 技术领域中具有创新性和突破性的高水平研究进展和研究成果. 感谢《软件学报》编委会和数据库专委会对专题工作的指导和帮助, 感谢专题全体评审专家及时、耐心、细致的评审工作, 感谢踊跃投稿的所有作者. 希望本专题能够对向量数据库及 DB4LLM 相关领域的研究工作有所促进.

作者简介

高宏, 博士, 浙江师范大学教授, 校“杰出教授”, 博士生导师, CCF 杰出会员. 主要研究领域为大数据管理与计算, 数据库系统, 多模态大数据智能分析, 物联网数据获取与分析. 先后主持国家自然基金重大项目、国家自然基金重点项目、科技部重点研发计划等 20 余项. 曾获得国家科技进步二等奖 1 项、省部级自然科学一等奖 1 项、省部级自然科学二等奖 1 项.

李国良, 博士, 清华大学教授, 博士生导师, 计算机科学与技术系副主任, CCF 杰出会员, 国家杰出青年科学基金获得者, IEEE Fellow、openGauss 社区技术委员会主席. 主要研究领域为大数据, 数据库, 数据科学. 曾获得国家科技进步二等奖 1 项、江苏省科技进步一等奖 1 项、电子学会科技进步一等奖 1 项、CCF 科技进步特等奖 1 项.

张蓉, 博士, 华东师范大学教授, 博士生导师, CCF 专业会员, CCF 数据库专委执委. 主要研究领域为分布式数据管理, 数据库基准评测, 数据库测试, 数据流管理. 主持多项国家自然科学基金项目, 参与多项 863、973 项目. 曾获得国家科学技术进步奖二等奖 1 项、上海市科学技术奖一等奖 1 项.