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, 特别是来自于大语言模型 (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》: 高维向量近似最近邻搜索 (approximate nearest neighbor search, ANNS) 是向量数据库的基础和核心查询之一. 近些年随着该领域的快速发展, 涌现出来的新方法和研究成果亟须系统性梳理. 综述中总结了现有方法的问题和解决思路, 并对未来研究方向进行展望.

《向量数据库的 K 近邻图高效更新方法》: 随着预训练嵌入模型在非结构化数据建模与检索中的广泛使用, 嵌入模型的微调逐渐成为提升嵌入向量的语义表示能力的核心步骤. 然而, 嵌入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. 针对现有研究缺乏对微调后的嵌入向量进行快速适应的 K 近邻图研究, 论文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.

《GoVector: I/O-高效的高维向量近邻查询缓存策略》: 基于图结构的高维向量索引 (索引图) 因其高效的近似最近邻搜索能力, 已成为大规模向量检索的主流方法. 然而, 由于索引图需存储大量邻接关系, 导致内存开销大, 因此实际部署时通常需将其存储于外存. 当执行近似最近邻搜索时, 按需加载索引图和向量数据会导致频繁发生 I/O 操作, 并成为检索性能的主要瓶颈. 针对该问题, 论文设计了一个静态-动态混合缓存策略, 使得缓存命中率得到显著提升, 有效降低了整体 I/O 开销.

(2) 向量数据管理

《GPU 加速的高维向量聚类算法》: 基于密度的聚类算法 (DBSCAN) 因其无需预先指定聚类数量、能够发

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

现复杂聚类结构并有效识别噪声点的特性, 在数据分析领域得到了广泛应用. 然而, 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间代价, 且面临维度灾难等问题, 阻碍其在实际场景中的部署和应用. 本文提出一种 GPU 加速的高维向量聚类算法, 通过引入 $k$ 近邻图索引加速 DBSCAN 的计算.

(3) 索引构建与优化

《LSMDiskANN: 更新友好型磁盘向量索引框架》: 在大模型时代, 向量数据库的广泛应用推动了向量索引规模的急剧膨胀. 在磁盘级向量索引中高效支持大规模向量的更新操作并同时提供高性能的查询服务已成为重要的研究课题. 论文针对当前先进的 FreshDiskANN 算法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, 受到日志合并思想在次级索引中的成功应用的启发, 提出了一种基于 LSM 思想的更新友好型磁盘向量索引框架, 提升了系统在混合负载场景下的整体性能与稳定性.

《面向批量更新的向量索引召回率优化》: 近似最近邻 (approximate nearest neighbor, ANN) 搜索是支撑向量数据库、推荐系统及大语言模型等上层应用的关键技术. 其中, 分层可导航小世界 (hierarchical navigable small world, HNSW) 图索引通过构建层级化结构, 迅速定位结果至目标区域, 从而以较低的计算成本实现较高的检索召回率. 然而, 现有 HNSW 算法主要面向静态数据检索场景而设计, 数据库中数据以批量方式进行更新时, 会造成 HNSW 算法中启发式剪枝算法的有效性降低以及相似向量连接的稀疏化等问题, 导致查询召回率的下降. 针对上述问题, 论文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, 实现保证检索精度的同时提升邻居连接的多样性.

《权重残差向量量化: 向量压缩与分层索引结构》: 随着多源异构数据、多模态数据等在大模型和数据湖等场景的广泛应用, 基于向量的数据检索和存储管理需求增长迅速. 通过将异构数据映射为高维向量, 向量数据库将多种数据统一管理并支持相似性检索, 成为生成式检索和 AI 数据库的重要基础. 然而, 现有向量数据库在索引效率、索引构建复杂度及检索准确性方面还存在显著瓶颈. 针对该问题, 论文提出一种基于权重残差向量量化 (weight residual vector quantization, WRVQ) 的新型框架, 实现了低失真率下的高效索引压缩与存储, 以及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》: Text2SQL 技术通过减少非专家用户与关系数据库交互的技术障碍, 已逐步发展为数据分析和数据库管理的重要工具. 尽管以 GPT 为代表的大语言模型的引入可以进一步提升 Text2SQL 系统的性能. 然而, 由于空间数据具有复杂的几何关系、多样化的查询模式, 并需要满足对高精度语义理解的严苛需求, 现有的 Text2SQL 技术难以直接适用于空间数据库领域. 为解决该问题, 降低用户与空间数据库的交互门槛, 论文提出了面向空间数据库的自然语言查询转换方法, 实现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.

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. 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.