软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(3):1049–1050 [doi: 10.13328/j.cnki.jos.007079]

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

面向多模态数据的新型数据库技术专题前言*

彭智勇[1], 高云君[2], 李国良[3], 许建秋[4]

[1](武汉大学 计算机学院, 湖北 武汉 430072)

[2](浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

[3](清华大学 计算机科学与技术系, 北京 100084)

[4](南京航空航天大学 计算机科学与技术学院, 江苏 南京 211106)

通信作者: 高云君, E-mail: gaoyj@zju.edu.cn; 许建秋, E-mail: jianqiu@nuaa.edu.cn

中文引用格式: 彭智勇, 高云君, 李国良, 许建秋. 面向多模态数据的新型数据库技术专题前言. 软件学报, 2024, 35(3): 1049–1050. http://www.jos.org.cn/1000-9825/7079.htm

以数字技术为标志的新一轮科技革命方兴未艾, 将人类带入数字经济时代. 全球各行各业数据量呈现爆炸式增长, 数据类型和数据格式也呈现多种形式, 例如结构化关系表、半结构化 JSON/XML、非结构化文本/图像/视频, 以及图数据、流数据和时序数据等. 这要求数据库系统能够同时高效地管理多种不同类型的数据. 多模态数据管理与分析成为亟需解决的问题. 目前的方法主要通过拓展现有的数据库或通过集成各种不同模态数据管理引擎来支持多模态数据管理与分析, 缺少新颖的理论、方法与技术的支撑. 本专题围绕多模态数据管理与分析的整个生命周期, 通过结合大数据技术和人工智能方法探讨新型数据库系统理论、方法和技术, 包括多模态数据统一建模、存储与索引、查询与挖掘、并发控制、多模态数据库系统构建及其典型应用等主题, 赋予数据库系统新的管理能力, 形成多模态数据管理与分析在各行各业的最新应用成果.

本专题公开征文, 共收到投稿 18 篇. 论文均通过了形式审查. 特约编辑先进行初审, 然后邀请了近 20 位专家参与审稿工作, 每篇投稿邀请 2 位同行专家进行评审. 稿件经初审、复审、NDBC 2023 会议宣读和终审共 4 个阶段, 历时 4 个月, 最终有 10 篇论文入选本专题. 根据主题, 这些论文可以分为 3 组.

(1) 多模态数据查询处理与优化方法

《基于邻域 *k*-核的社区模型与查询算法》针对多模态图数据开展研究, 主要解决了多模态图网络中稠密度阈值的多社区搜索问题, 提出了新的社区模型, 引入了边稠密度的概念, 提出了基于边稠密度的基线算法, 设计了索引树和改进索引树结构, 提升了搜索效率并证明了结果的完整性.

《基于细粒度特征融合的部分多模态哈希》针对多模态哈希开展研究, 提出了实现部分多模态哈希模型, 利用 Transformer 编码器实现细粒度的多模态特征融合, 解决样本模态不完整、学习能力有局限性和缺乏语义信息的问题. 实验结果表明所提模型能够有效地实现部分多模态哈希, 并可应用于大规模多模态数据检索.

《GPPR: 跨域分布式个性化 PageRank 算法》针对跨域分布环境下的大图分析个性化 PageRank 开展研究, 降低网络带宽异构对算法迭代速度的影响. 采用随机游走方式和相关算法减小了工作节点之间传输数据的带宽负载, 在 8 个开源大图数据进行性能测试, 相比于现有方法, 效率有显著提升.

(2) 多模态数据视图和融合技术

《融合多模态数据的小样本命名实体识别方法》提出了一种融合多模态数据的小样本命名实体识别模型, 通过将图像信息转化为文本信息作为辅助模态信息的方法, 有效解决了语义信息粒度不一致导致的模态对齐效果不佳的问题. 通过真实的多模态数据集进行了测试, 验证了所提方法的有效技术提升.

《面向多模态模型训练的高效样本检索技术》提出了一种面向多模态模型训练的高效样本检索技术, 通过感知模型训练类间边界点, 精确评估样本对模型的价值, 设计了半有序的高效样本索引. 采用多组多模态

* 收稿时间: 2023-11-09; jos 在线出版时间: 2023-11-09

数据集进行实验，验证了所提方法的有效性.

《面向视频的细粒度多模态实体链接》提出了面向视频的细粒度实体链接，构建了细粒度视频实体链接数据集，提出利用大模型抽取视频中的实体及其属性. 实验结果表明，所提方法能够有效地处理视频上细粒度实体链接任务.

(3) 多模态数据管理系统应用技术

《面向云边端协同的多模态数据建模技术及其应用》从云边端三层数据的数据类型出发，提出了面向云边端协同的多模态数据建模技术，给出了基于元组的多模态数据模型定义，以解决多模态数据统一表征困难的问题. 同时，给出了多模态数据模型的完整性约束以及面向云边端协同多模态数据模型的示范应用.

《Apache IoTDB 中的多模态数据编码压缩》基于 Apache IoTDB 系统中时间戳数据、数值数据、布尔值数据、频域数据、文本数据等多个不同的模态，提出了利用不同模态数据特点的数据编码压缩方法，将数据质量因素纳入到编码算法的设计中，在多个数据集上进行实验评估，验证了多模态数据编码压缩的效果.

《Navi: 基于自然语言交互的数据分析系统》提出了基于自然语言交互的数据分析系统 Navi. 该系统采用模块化的设计原则，抽象出主流数据分析流程的 3 个核心功能模块: 数据查询、可视化生成和可视化探索模块，从而降低系统设计的耦合度.

《支持深度学习的视觉数据库管理系统研究进展》从文本、图像和视频等多模态数据的相互融合处理出发，总结了视觉数据库管理系统在不同层面上面临的挑战，包括数据存储、查询优化、执行调度以及编程接口，探讨了上述 4 个层面上的相关技术，并对视觉数据库管理系统未来的研究方向进行了展望.

本专题主要面向数据库、大数据、人工智能等多领域的研究人员和工程人员，反映了我国学者在多模态数据的新型数据库技术上的最新研究进展. 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. 希望本专题能够对多模态数据的新型数据库技术相关领域的研究工作有所促进.

彭智勇(1963－)，男，博士，武汉大学计算机学院教授，武汉大学大数据研究院副院长，获国务院政府特殊津贴，国务院软件工程学科评议组成员，CCF 会士，CCF 常务理事. 主要研究领域为对象代理数据库，大数据管理系统，制造业大数据，科技大数据，教育大数据，可信云数据，地理数据水印. 主持了国家自然科学基金重点项目和国家 863 数据库重大专项课题等，提出了一个新的数据库模型: 对象代理模型. 曾获得中创软件人才奖、国防科工委科技进步一等奖、教育部科技进步二等奖.

高云君(1977－)，男，博士，浙江大学求是特聘教授，博士生导师，国家杰出青年科学基金获得者，现为 ACM 中国 SIGSPATIAL 分会副主席，浙江省大数据智能计算重点实验室副主任，浙江大学软件学院副院长，浙江大学计算机软件研究所副所长，CCF 高级会员. 主要研究领域为数据库，大数据管理与分析，DB 与 AI 融合. 主持了国家杰出青年科学基金、国家重点研发计划、973 计划等，曾获得 2019 年度中国电子学会科技进步特等奖、2016 年度教育部科技进步一等奖、2011 年度浙江省科学技术一等奖.

李国良(1981－)，男，博士，清华大学长聘教授，博士生导师，计算机系副主任，国家自然科学基金杰出青年基金获得者，数据库专委会副主任，CCF 杰出会员. 主要研究领域为大数据，数据库，数据科学. 2014–2022 年入选爱思唯尔高被引学者榜单，获得 VLDB 2017 Early Research Contribution Award. 曾任 SIGMOD 2021 大会主席、ICDE 2022 Industry 主席. 曾获得 VLDB 青年杰出贡献奖、IEEE TCDE 杰出新人奖、国家科技进步二等奖、江苏省科技进步一等奖、电子学会科技进步一等奖、计算机学会科技进步特等奖.

许建秋(1982－)，男，博士，南京航空航天大学教授，博士生导师，计算机系主任，CCF 高级会员. 主要研究领域为时空数据管理. 主持国家自然科学基金项目、国防 173 领域基金、CCF 华为胡杨林数据库专项基金等. 发表学术论文 40 余篇，包括 CCF 推荐 A 类论文，如 IEEE TKDE、ICDE、PVLDB 等. 授权国家发明专利 3 项. 曾获得 APWeb/WAIM 2017 最佳系统演示论文奖、SSTD 2019 最佳展望论文奖、SSTD 2021 最佳研究论文提名奖、NDBC 最佳系统演示论文奖(2023).