苏云祥

手机：(+86) 18811382629
邮箱：suyx21@mails.tsinghua.edu.cn
个人主页：https://suyx1999.github.io

教育背景

清华大学，软件工程，*在读博士研究生* 2021/08 - 2026/06
清华大学，数理基础科学，*理学学士* 2017/08 - 2021/06

研究方向

主要研究方向为 **Data+AI**，探索**大模型增强型数据库**以及**数据库能力驱动的机器学习**。当前工作主要关注利用大模型能力增强传统数据库的数据处理能力。既往工作主要关注利用时间序列数据库的存储特性，提升数据库原生机器学习效率。相关成果以独立第一作者身份在数据库顶级会议 SIGMOD、KDD、VLDB 发表论文 4 篇，相关技术集成至开源时间序列数据库 Apache IoTDB，并完成系统部署与应用验证。

研究成果

(1) **Yunxiang Su**, Yikun Gong, Shaoxu Song. Time Series Data Validity. SIGMOD 2023. (**独立一作，CCF-A 类会议**)

- 贡献：首次定义时序数据的有效性，并提出度量方法，基于时序数据库存储特性设计高效算法。
- 效果：相比现有关系型数据基线取得更准确的度量结果，同时取得高达 4 个数量级的效率提升。

(2) **Yunxiang Su**, Wenxuan Ma, Shaoxu Song. Learning Autoregressive Model in LSM-Tree based Store. KDD 2023. (**独立一作，CCF-A 类会议**)

- 贡献：针对时序数据库内机器学习，改进传统自回归模型，通过时序数据库预计算信息加速学习。
- 效果：使数据库内自回归模型学习效率提升 10 倍，学习超过 100 万个数据点仅需约 10 毫秒。

(3) **Yunxiang Su**, Shaoxu Song, Xiangdong Huang, Chen Wang, Jianmin Wang: Distance-based Outlier Query Optimization in Apache IoTDB. VLDB 2024. (**独立一作，CCF-A 类会议**)

- 贡献：提出时序数据库内的基于距离的异常检测算法，实现数据库内的高效异常检测。
- 效果：在保证异常检测结果完全一致同时，使数据库内异常检测时间开销降低一个数量级。

(4) **Yunxiang Su**, Kenny Ye Liang, Shaoxu Song: In-Database Time Series Clustering. SIGMOD 2025. (**独立一作，CCF-A 类会议**)

- 贡献：提出基于形状的高效时序数据聚类方法，能够高效处理长序列，并针对数据库内优化。
- 效果：基于 20 余个数据集的实验，方法在效果和效率上显著优于所有基线，综合性能位居首位。

(5) Kenny Ye Liang, **Yunxiang Su**, Shaoxu Song, Chunping Li: Turn Waste Into Wealth: On Efficient Clustering and Cleaning Over Dirty Data. TKDE 2025. (**第二作者，CCF-A 类期刊**)

- 贡献：提出了基于网格的修复方法，将异常数据转化为增强聚类的有效信息。
- 效果：保持聚类效果同时，相较于基线方法降低一个数量级的时间开销。

获奖情况

(1) 国家奖学金 2023/11

(2) 清华大学综合奖学金（一等） 2024/11

实习经历

(1) 阿里巴巴-通义实验室-智能系统 2025/6-2025/10

课题：Benchmark LLM-Driven Enhancements to Native DBMS Operators

- 理论上，首次系统化、体系化提出了 LLM 增强型数据库的算子分类框架；
- 技术上，针对传统的传统关系代数算子，设计并实现了 6 大类、20 余个对应的 LLM 增强型数据库算子，能够完整覆盖、拓展整个传统关系代数的操作体系，从而实现增强型数据库；
- 实验上，设计基准测试集，覆盖 20 多个数据库、100 多张表、300 多个用户查询问题。同时对比了 10 个基线方法，评价了不同实现方式之间的优劣。

学术服务

(1) IEEE Transactions on Knowledge and Data Engineering (TKDE) 期刊审稿人 2024

(2) ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 审稿人 2022-2025

(3) International Conference on Very Large Databases (VLDB) 学生审稿人 2021-2025

(4) IEEE International Conference on Data Engineering (ICDE) 学生审稿人 2021-2025

(4) CIKM、AAAI、SIGIR、WWW 等会议学生审稿人

工程项目成果

(1) 国家重点研发计划：制造大数据驱动的预测运行与精准服务技术及系统 2021/03 - 2023/11

(2) 中冶赛迪信息技术（重庆）有限公司：时序数据库及数据质量研究及开发 2023/05 - 2024/11

(3) Apache IoTDB：*Apache* 顶级开源物联网数据库项目 2020/12 - 2024/11

(4) IoTDB-Quality：*Apache* 顶级开源项目 *IoTDB* 下二级项目 2020/12 - 至今