Mengyi Yan

G507 New main building of Beihang University, XueYuan Road No.37, HaiDian District, BeiJing, China

(+86) 13240300017 | yanmy1008@buaa.edu.cn | authurlord

Education

Beihang University Beijing, China

BDBC (Beijing Advanced Innovation Center for Big Data and Brain Computing), School of Computer Science and Engineering *Sep. 2017 - Present*

- Ph.D. Candidate in Computer Software and Theory
- Advisor: Prof. Jianxin Li
- Expected date of the PhD graduation: 06/2025
- GPA: 3.23/4

Beihang University Beijing, China

School of Mathematics *Sep. 2013 - Jun. 2017*

- B.E. in Applied Mathematics
- GPA: 3.59/4
- RANK: 4/42

Publications

* means corresponding author; $^{+}$ indicates that author names are listed in alphabetical order.

- **Mengyi Yan**, Wenfei Fan, Yaoshu Wang, Min Xie*. Enriching Relations with Additional Attributes for ER. VLDB'24, 2024(CCF-A Conference), Link
- **Mengyi Yan**, Yaoshu Wang*, Yue Wang, Xiaoye Miao, Jianxin Li. GIDCL: A Graph-Enhanced Interpretable Data Cleaning Framework with Large Language Models. SIGMOD'25, 2024(CCF-A Conference). Link, Camera-Ready
- **Mengyi Yan**, Yaoshu Wang*, Kehan Pang, Min Xie, Jianxin Li*. Efficient Mixture of Experts based on Large Language Models for Low-Resource Data Preprocessing. SIGKDD'24, 2024(CCF-A Conference),Link
- **Mengyi Yan**, Weilong Ren*, Yaoshu Wang, Jianxin Li*. A Retrieval-Augmented Framework for Tabular Interpretation with Large Language Model. DASFAA'24, 2024(CCF-B Conference). Link
- Wenfei Fan, Ziyan Han, Weilong Ren*, Ding Wang, Yaoshu Wang, Min Xie, **Mengyi Yan**. Splitting Tuples of Mismatched Entities^{+}. SIGMOD'24, 2024(CCF-A Conference). Link
- Haoyi Zhou, Jianxin Li*, Shanghang Ji, Shuai Zhang, **Mengyi Yan**, Hui Xiong. Expanding the Prediction Capacity in Long Sequence Time-Series Forecasting. Artifical Intelligence, 2023.(CCF-A Journal). Link
- Shuai Zhang, Jianxin Li, Pengtao Xie, Yingchun Zhang, Minglai Shao, Haoyi Zhou, **Mengyi Yan**. Stacked kernel network. arXiv preprint arXiv:1711.09219, 2017. Link
- Yaoshu Wang, **Mengyi Yan**. Unsupervised Domain Adaptation for Entity Blocking Leveraging Large Language Models. IEEE BigData'24(CCF-C Conference) Link

Manuscripts

- **Mengyi Yan**, Yaoshu Wang, Xiaohan Jiang, Haoyi Zhou, Jianxin Li*. Towards Uncertainty-Calibrated Structural Data Enrichment with Large Language Model for Few-Shot Entity Resolution. Frontiers of Computer Science, 2024 (CCF-B Journal, Major Revision). Link

Research Experience

My research are broadly in the field of databases and data quality, with a strong emphasis on data cleaning, entity resolution, and the application of large language models in data preprocessing. I develop solutions to enhance data consistency and accuracy, particularly in low-resource environments, and explore methods for integrating heterogeneous data sources effectively. These works have been published in top-tier venues, including SIGMOD, VLDB, and KDD. A brief summary of my past work can be found below.

Knowledge-Enhanced Entity Resolution

- I developed knowledge-enhanced algorithms, notably ENRICH, which integrates external data sources to address limitations in traditional entity resolution (ER). Using reinforcement learning, ENRICH optimizes attribute selection to enhance data quality and efficiency in relational databases, significantly improving accuracy. This work was published in [VLDB24].

Graph-Enhanced Interpretable Data Cleaning

- I proposed GIDCL, a framework that leverages large language models and graph neural networks (GNNs) for interpretable data cleaning. GIDCL converts relational tables into graph structures, enabling structural correlations among data to be captured effectively. This framework also generates interpretable data cleaning rules through LLMs, improving accuracy and accessibility in data cleaning. This research was published in [SIGMOD25].

Low-Resource Multi-Task Data Preprocessing

- I introduced MELD, a mixture-of-experts (MoE) framework tailored for data preprocessing in low-resource environments. MELD combines retrieval-augmented generation and expert refinement to optimize multi-task tabular data cleaning, advancing the performance of data preprocessing with minimal labeled data. Relevant work has been presented in [KDD24].

Work Experience

Research Intern, Shenzhen Institute of Computing Sciences *Sep. 2021 - Jan. 2025*

- Engaged in research related to databases, data quality, and large language model for data quality tasks. Contributed to several academic papers ([SIGMOD24, VLDB24]), and collaborated on algorithm design for database products such as RockDQ.

Teaching Assistant, Beihang University *Mar. 2021 - Jun. 2021*

- Graduate Course: Discrete Mathematics

Data Mining Intern, United Nations Environment Programme (Bangkok) *Jul. 2015 - Oct. 2015*

- Organized compliance documents by compiling and analyzing environmental protection and compliance policies from multiple Asian and European countries, and prepared a comparative policy report.
- Collected and processed data by scraping environmental reports and documents published by government agencies in Asia and Europe over the past 3-5 years. Extracted relevant statistical tables and data, developed MATLAB scripts to consolidate, organize, and visualize key statistical results, and performed trend analysis on critical indicators to create a comprehensive statistical report.

Skills & Others

Programming Python, Bash & Linux, SQL

Tools Git, LaTeX, Torch, vLLM

Services

Subreviewer ICDE 2024, KDD 2024