Ziyan Han

Website: philo-vanguard.github.io
hanzy@buaa.edu.cn | +86 16601163504 | Shenzhen, China

EDUCATION

Beihang University (BUAA), Beijing, China Sept. 2018 – Jun. 2025
School of Computer Science and Engineering
Ph.D. in Computer Software and Theory
GPA 3.76/4.0, ranked 11/91 (12%)
Advisor: Prof. Wenfei Fan

Xidian University (XDU), Xi'an, China Aug. 2014 – Jun. 2018
School of Computer Science and Technology
B.E. in Computer Science and Technology
GPA 3.74/4.0, ranked 3/175 (1.7%)

RESEARCH INTERESTS

Data Mining/Rule Discovery: Sampling, Top-k, Diversified, etc.

Logic Deduction combined with ML Models: AI for DB, DB for AI, ML interpretability, etc.

Data Management: Error Detection, Data cleaning, etc.

Data Quality: Conflict resolution, Entity resolution, Tuple splitting, etc.

I have a strong interest in the intersection of DB and AI. I am particularly intrigued by how database techniques can enhance AI model performance, and vice versa. I expect to explore applying logic rules to make black-box ML models more interpretable and robust, collaborative optimization between logic rule discovery and specific downstream ML tasks, utilizing ML techniques to optimize various database management processes, etc.

I am also open to exploring other research directions and enthusiastic about investigating new areas of interest.

PUBLICATIONS

(Note: In papers 1–4, all authors are co-authors, and thus, sorted by alphabetic order.)

1. Wenfei Fan, **Ziyan Han**, Min Xie, and Guangyi Zhang. *Discovering Top-k Relevant and Diversified Rules*. In SIGMOD (2025). ACM. **[CCF-A; Top-tier; Co-first author]**

2. Wenfei Fan, **Ziyan Han**, Weilong Ren, Ding Wang, Yaoshu Wang, Min Xie, and Mengyi Yan. *Splitting Tuples of Mismatched Entities*. In SIGMOD (2024). ACM. **[CCF-A; Top-tier; Co-first author]**

3. Wenfei Fan, **Ziyan Han**, Yaoshu Wang, and Min Xie. *Discovering Top-k Rules using Subjective and Objective Criteria*. In SIGMOD (2023). ACM. **[CCF-A; Top-tier; Co-first author]**

4. Wenfei Fan, **Ziyan Han**, Yaoshu Wang, and Min Xie. *Parallel Rule Discovery from Large Datasets by Sampling*. In SIGMOD (2022). ACM. **[CCF-A; Top-tier; Co-first author]**

5. Ting Deng, Lei Hou, and **Ziyan Han**. *Keys as features for graph entity matching*. In ICDE (2020). IEEE. **[CCF-A; Top-tier]**

6. **Ziyan Han**, Weilong Ren, and Wanjia Chen. Fast Diversified Top-k Rule Discovery via Embedding. In VLDB (2026). IEEE. (Under review) **[CCF-A; Top-tier]**

RESEARCH EXPERIENCES

My research primarily focuses on data mining, rule discovery, and logic deduction combined with machine learning models, specifically on the discovery and application of data quality rules. My work has been published in top-tier database conferences, i.e., SIGMOD and ICDE. Below is a concise overview of my contributions across various domains.

- **Data Mining and Data Analysis**

  I have tackled several challenges in rule discovery, including high computational costs and extensive resource consumption [SIGMOD22], the limitations of non-comprehensive rule evaluation metrics that lack subjective criteria [SIGMOD23], and redundancy within mined rule sets [SIGMOD25].

- **Data Management and Data Quality**

  I have developed methods for resolving conflicts within tuples of mismatched entities [SIGMOD24], and for graph entity resolution using graph keys [ICDE20].

- **Logic Deduction combined with ML models**

  I have integrated machine learning techniques with logic rules to enhance data quality. Specifically, I utilize machine learning techniques to accelerate the rule discovery process [SIGMOD22, SIGMOD23, SIGMOD25]. Additionally, rules discovered can be further applied to improve data quality, such as entity resolution, conflict resolution, and tuple splitting [SIGMOD24, ICDE20].

WORKING EXPERIENCE

***Intern Researcher*, Shenzhen Institute of Computing Sciences, Shenzhen, China** 2021 – 2024

- ***Project 1: Parallel Rule Discovery from Large Datasets by Sampling*** (link)
  - *Challenge*: Extensive computational costs and substantial resource consumption encountered in rule discovery when processing large datasets or mining complex rules.
  - *Overview*: We propose a multi-round sampling strategy to independently discover rules in each sample and ultimately obtains the union of all rule sets. It focuses on Entity Enhancing Rules (REEs) for collective entity resolution and conflict resolution. By proving accuracy and validity bounds, the method aims to guarantee that the rules discovered from samples are effective across the entire dataset.
  - *Technical Contributions*: i. Propose a multi-round sampling strategy based on random-walk or breath-first search, and prove an accuracy bound and deduce a sampling size for the sampling method with BFS; ii. Design a generic representation of REEs by means of a tableau that specifies constant patterns; iii. Formulate the REE discovery problem with sampling and develop a parallelized discovery algorithm; iv. Train an DQN model to select semantically correlated predicates during expansion; v. Populate tableaux to efficiently retrieve constant patterns without enumerating numerous constant values.
  - *Academic Output*: Published paper in SIGMOD'22 [Top-tier / CCF-A Database Conference]
- ***Project 2: Discovering Top-k Rules using Subjective and Objective Criteria*** (link)
  - *Challenge*: Existing rule discovery algorithms typically consider only objective criteria such as support and confidence, lacking subjective criteria and neglecting users' preferences. This yields rules that may be already known or common sense, failing to meet user preferences and practical application needs.
  - *Overview*: We study the top-k rule discovery algorithm that integrates both subjective and objective metrics. It aims to learn users' preference and discover top-ranked rules that most fit users' needs, suitable for highly customized scenarios that require detailed evaluation and interpretation.

- *Technical Contributions*: i. A bi-criteria model to characterize rules, in terms of conventional objective measures and new subjective measures to fit users' needs; ii. Propose interacting with users through active learning to learn the subjective measures and weight vectors; iii. A top-k rule discovery algorithm based on the bi-criteria model; iv. An anytime top-k discovery algorithm for successive discovery via lazy evaluation; v. Effective pruning strategies, e.g., learned score upper bounds for early termination.
- *Academic Output*: Published paper in SIGMOD'23 [Top-tier / CCF-A Database Conference]

- ***Project 3: Discovering Top-k Relevant and Diversified Rules*** (link)
  - *Challenge*: Current rule discovery methods usually lack metrics for measuring the diversity of rules, resulting in rules that are often too "homogeneous" to each other and consequently, functionally similar and redundant. This reduces the quality of rule based decision-making.
  - *Overview*: We study the diversity-driven top-k rule discovery problem. It aims at achieving an optimal balance between relevance and diversity in rule discovery, effectively reducing redundancy in rule set, suitable for scenarios that require extensive decision support.
  - *Technical Contributions*: i. Develop a more user-friendly model to learn the relevance of rules; ii. Propose four diversity measures to assess the diversity among rules; iii. Formulate a new discovery problem that integrates both relevance and diversity, proving its NP-completeness; iv. Propose a practical algorithm and establish approximation bounds under specific conditions.
  - *Academic Output*: Published paper in SIGMOD'25 [Top-tier / CCF-A Database Conference]

- ***Project 4: Splitting Tuples of Mismatched Entities*** (link)
  - *Challenge*: Fixing tuples of mismatched entities is challenging. When data is erroneously merged, it often involves structural or logical errors, where different entities is mistakenly combined or concatenated. This goes beyond traditional data cleaning.
  - *Overview*: We study the inverse of entity resolution, to identify tuples to which distinct real-world entities are matched by mistake, and split such tuples into a set of tuples, one for each entity. We formulate the tuple splitting problem and propose a scheme for this new problem. By chasing data with REEs, it enables detect and repair errors, inconsistencies, and missing values within a unified framework. This not only improves data quality but also provides support for the practical application of data quality rules.
  - *Technical Contribution*: i. Formulate a new problem named Tuple Splitting (TS), the inverse of Entity Resolution (ER); ii. Propose a scheme to decide what tuples to split and what tuples to correct without splitting, fix errors/assign attribute values to the split tuples, and impute missing values; iii. Extend REEs to REE+s by supporting predicates for assessing correlation between attributes, aligning entities across relations and knowledge graphs, and extracting data from knowledge graph; iv. It unifies logic deduction, correlation models, and data extraction by chasing the data with the rules.
  - *Academic Output*: Published paper in SIGMOD'24 [Top-tier / CCF-A Database Conference]

***Research Assistant*, Beihang University, Beijing, China** 2019 – 2020

- ***Project: Keys as features for graph entity matching*** (link)
  - *Challenge*: Entity matching encounters the challenge of balancing interpretability and accuracy. Rule-based methods provide clear explanations but may lack accuracy, while ML-based approaches achieve higher accuracy at cost of interpretability.
  - *Overview*: We introduce Graph Matching Keys (GMKs), an extension of graph keys that

incorporates similarity predicates to support approximation entity matching. Treating entity matching as a classification problem, we propose a supervised learning method that combines ML techniques with graph dependencies, achieving high accuracy while providing interpretability.

- *Technical Contributions*: i. Extend graph keys (GKeys) to graph matching keys (GMKs) with similarity predicates on attributes; ii. Introduce a supervised learning method for graph entity matching by using GMKs as features in vector representation for node pairs to train a classifier; iii. Present GMK feature extraction algorithm and feature selection method; iv. GMKs help to explain the classification results and classification models. We analyze the interpretability provided by GMKs from local and global explanations.
- *Academic Output*: Published paper in ICDE'20 [Top-tier / CCF-A Database Conference]

Teaching Assistant, Beihang University, Beijing, China 2020 – 2021

- ***Class: Formal Languages and Automata***
  - *Responsible for:* Assigning Homework, Grading Assignments, Proctoring Exams, and Scoring Exams, etc.

AWARDS

2018 – 2024 ***Beihang University***

Distinguished Graduate of Beihang University, 2025

SIGMOD 2025 Student Travel Grants, 2025

SIGMOD 2024 Student Support Scholarship, 2024

SIGMOD 2023 Student Travel Award, 2023

Outstanding Freshman Scholarship, BUAA, 2018

Outstanding Graduate Students Award, BUAA, 2020

Merit Student Award, BUAA, 2019/2020/2021

The Second Prize Scholarship, BUAA, 2019/2021

The First Prize Scholarship, BUAA, 2020

CASC Scholarship, BUAA, 2022

2014 – 2018 ***Xidian University***

National Scholarship for Encouragement, 2016

The Special Scholarship, XDU, 2017

National/Provincial College Student Innovation and Entrepreneurship Training Program Completion Certificate, 2017

Outstanding Student Model Award, XDU, 2017

The Second Prize Scholarship, XDU, 2015

The Second Prize Scholarship, XDU, 2014

SKILLS & Hobbies

Programming: Python, Java, Bash, C/C++, Markdown, etc.

Tools: LaTeX, Git, Spark, etc.

Languages: English, Chinese (native).

Hobbies: Tennis, Fitness (Strength Training), Cooking, etc.

SERVICES

Sub-reviewer: AAAI 2023, APWEB 2023, TKDE 2023, ICDE 2024, ICDE 2025, APWEB 2025

Volunteer: SIGMOD 2023