Yiming Lin

6330, Adobe Cir South, Irvine, CA, USA
92617
☎ *(+1) 9495222578*
✉ *yiminl18@uci.edu*

Education

2017–Present **Ph.D Candidate**, *Dept. of Computer Science, University of Calfornia Irvine.*

2015–2017 **Master of Science**, *Computer Science and Technology, Harbin Institute of Technology.*

2011–2015 **Bachelor of Science**, *Computer Science and Technology, Harbin Institute of Technology.*

Areas & Technical Skills

Areas **Data cleaning, Data Integration, Query Optimization, Query Processing**.

Skills C/C++, Java, Python programming.

Advisor Prof. Sharad Mehrotra in UCI.

Experiences

Summer, 2022 **Research Internship** in **Microsoft Research**.
- Working with Yeye He, on AutoBI: automatical BI model building. Our work is under submission to PVLDB 2023.

2020-2023 Recipient of **Hasso-Plattner-Institute**(**HPI**) Fellowship.

Summer, 2021 **Applied Scientist Internship** in **Amazon**.
- Working with Dmitri, Kalashnikov and Vidit, Bansal on holistic approach to resolve suspicious duplicated entity problem.

2017-present **Research Assistant** in ISG group, UCI.

2013-2016 **ACM/ICPC** Asia Programming Contest, Silver Medal, 1 time, Bronze Medal, 3 times.

Publications

[1] **Yiming Lin**, Daokun Jiang, Roberto Yus, Andrew Chio, Georgios Bouloukakis, Sharad Mehrotra, Nalini Venkatasubramanian: LOCATER: Cleaning WiFi Connectivity Datasets for Semantic Localization. **PVLDB** 14(3): 329 - 341, 2021.

[2] **Yiming Lin**, Pramod Khargonekar, Sharad Mehrotra, Nalini Venkatasubramanian: T-Cove: An exposure tracing System based on Cleaning Wi-Fi Events on Organizational Premises. **PVLDB**, 14(12): 2783 - 2786, 2021.

[3] **Yiming Lin**, Hongzhi Wang, Jianzhong Li, Hong Gao: Efficient entity resolution on heterogeneous records. (Extended Abstract) **ICDE** 2020.

[4] **Yiming Lin**, Hongzhi Wang, Jianzhong Li, Hong Gao: Efficient entity resolution on heterogeneous records. (**TKDE**) VOL. 32, NO. 5, MAY 2020.

[5] **Yiming Lin**, Hongzhi Wang, Jianzhong Li, Hong Gao: Data source selection for information integration in big data era. **Information Sciences** 479 (2019): 197-213.

[6] **Yiming Lin**, Hongzhi Wang, Shuo Zhang, Jianzhong Li, Hong Gao: Efficient quality-driven source selection from massive data sources. **Journal of Systems and Software** 118 (2016): 221-233.

[7] **Yiming Lin**, Yeye He. Auto-BI: Automatically Build BI-Models Leveraging Local Join Prediction and Global Schema Graph. (Under submission in **PVLDB 2023**)

[8] **Yiming Lin**, Sharad Mehrotra. ZIP: Lazy Imputation during Query Processing. (Under review in **PVLDB 2023**)

[9] **Yiming Lin**, Sharad Mehrotra. Filter Optimization: Learning from Rejection. (Under preparation)

Research Projects

2022. (Recent) **Auto-BI**.

- We developed an Auto Business Intelligence (BI) system that helps end-users by accurately predicting BI models given a set of input tables, i.e., to discover join columns accurately. We propose a principled graph-based optimization problem that considers both local join prediction and global schema-graph structures, which achieves over 90% F1-score on real-world and TPC benchmarks. [7] (Work is done during internship in Microsoft Research, with Yeye He.)

2021–present. (Ongoing) **Analysis-aware Data Cleaning**.

- **ZIP: lazy imputation during query processing**. Given SQL query on relational data set containing missing values, we develop ZIP which only imputes minimal number of missing values to answer query exactly. Quip co-optimizes query processing and missing value imputation by modifying the physical implementations of given query plan tree to minimize the query execution and imputation overhead. [8]
- **Filter optimization**. This work discovers new filters in query processing at run-time, and considers proper query re-optimization to speed up the overall query execution. [9]

2017–2019. (Recent) **Sensor Data Cleaning**.

- **LOCATER: Semantic Localization**. LOCATER uses data cleaning technologies over WiFi events to locate people inside buildings, which is passive, server-side and free of cost. [1]
- **T-COVE: Occupancy Estimation**. T-COVE targets to compute real-time occupancy (the number of occupants in a given area) estimation by leveraging data cleaning methods using WiFi connectivity events. [2]
- **Zero Cost Contact Tracing**. This work provides a practical solution to expose people's trajectories as required without new hardware and software, by using data cleaning and optimized query processing in WiFi sensor real data sets. [2]

2016,2021. (Recent & Past) **Efficient and Accurate Entity Resolution**.

- **Post-clustering for Suspicious Clusters** (Recent). This work tries to resolve *super dirty* clusters produced by ER algorithms, which contain multiple errors, incorrect/missing/incomplete/copied values. Our proposed algorithm SCC improves the old method used in Amazon by around 61% precision (from 34.1% to 95.5%) and by around 52% F-1 score (from 42.4% to 94.7%). This work was done during internship in Amazon.
- **Entity Resolution on Heterogeneous Records** (Past). We presented a new framework of entity resolution (ER) based on heterogeneous records and proposed a heterogeneous entity resolution algorithm (HERA). [4]

2014-2016. (Past) **Research of Source Selection**.

- **Incremental Integration over Massive Data Sources**. We studied online integration on massive data sources, and proposed an incremental integration algorithm, which can reduce the response time and return results with quality guarantee efficiently.
- **Data Source Selection for Information Integration in Big Data Era**. We first proposed a probabilistic coverage by considering the coverage, accuracy and overlaps of sources. To improve scalability, we designed a novel index, and proposed a scalable algorithm based on it, with two pruning strategies without sacrificing precision. [5]
- **Quality-driven Source Selection**. I developed algorithms of source selection focusing on the uneven quality of data source, considering the data quality, the limitation of resources and the completeness of data source. [6]

Systems

2017-2019 **Sensor Data Cleaning Systems (LOCATER and T-COVE) (Demo , Website)**.

- LOCATER and T-COVE systems have been deployed and being operational in more than **30 buildings** in **3 universities** (UCI, BSU, Plaksha University in India) and an **elderly living facility**, Walnut Village, located in Orange County.
- T-COVE has been adopted as **Covid-19 protection strategy** in UCI to display the real-time occupancy in campus building for **over 3 years**.
- LOCATER achieves similar accuracy to that achieved by expensive dedicated hardware based solutions available commercially in UCI testbed.