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

Yiming Lin

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, Query Optimization, Scalable Data Analysis, Query Processing**.
Skills C/C++, Java, Python programming.
Advisor Prof. **Sharad Mehrotra** in UCI.

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.

Experiences

2020-present Recipient of **Hasso-Plattner-Institute**(**HPI**) Fellowship.
summer, 2021 **Applied Scientist Internship** in Amazon.
2017-present **Research Assistant** in ISG group, UCI.
2013-2016 **ACM/ICPC** Asia Programming Contest, Silver Medal, 1 time, Bronze Medal, 3 times.

Research Projects

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

- **Quip: Query-driven Missing Value Imputation**. Given a SQL query on relational data set containing missing values, we develop Quip 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.
- **Entity Resolution on Streaming Data**. We are working on the approach to identify *only* the record pairs which can affect the results of query over streaming data to support a (near) real-time analysis.

2017–2019. **Sensor Data Cleaning**.

(Recent)

- **Semantic Localization**. We formulated localization as a set of data cleaning problems, missing value imputation and entity disambiguation, and proposed techniques only based on WIFI sensor data to localize people precisely. [2]
- **Occupancy Estimation**. Based on WiFi connectivity data, we built systems to compute real-time occupancy (the number of occupants in a given area) estimation by leveraging data cleaning methods. [1]
- **Free Cost Contact Tracing**. Due to the low adoption of most current contact tracing techniques, we are working on building a passive contact tracing system based on WiFi connectivity logs without requiring any new hardware and new software. We explore the space as well as time constraints to improve accuracy and efficiency. [1]
- **Applications**. Sensor data cleaning researches are implemented in LOCATER and T-COVE systems, which have been deployed at two campus, UCI and Ball State Univ. It is operational on over 20 buildings in UCI campus. Demo links.

2016,2021. **Efficient and Accurate Entity Resolution**.

(Recent & Past)

- **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 when I was doing internship in Amazon and SCC was adopted in Amazon product pipelines.
- **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. **Research of Source Selection**.

(Past)

- **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]