RUNHUI WANG

Website: runhuiwang.github.io Tel: +1-9089308736 E-mail: wangrunhui.pku@gmail.com

EDUCATION

Rutgers University, Computer Science Department 2019 - 2023 (expected)
Ph.D. Advisor: Prof. Yongfeng Zhang.

University of Queensland, School of ITEE, Australia 2017 - 2019
M.Phil. Advisors: Prof. Xiaofang Zhou (IEEE Fellow), Prof. Sibo Wang.

Peking University, School of EECS, China 2012 - 2016
B.S. in Computer Science.

RESEARCH INTERESTS

Similarity Search, Data Integration, Contrastive Learning, Recommender Systems, Parallel Computing

PUBLICATIONS

1. **Runhui Wang**, Yuliang Li, Jin Wang.
   Sudowoodo: Contrastive Self-supervised Learning for End-to-End Data Integration Under submission to Very Large Data Bases Conference ***(VLDB)*** 2022
2. **Runhui Wang**, Dong Deng.
   DeltaPQ: Lossless Product Quantization Code Compression for High Dimensional Similarity Search. Proceedings of the VLDB Endowment ***(PVLDB)*** 2020
3. **Runhui Wang**, Sibo Wang, Xiaofang Zhou.
   Parallelizing Approximate Single-Source Personalized PageRank Queries on Shared-Memory. Very Large Data Bases Journal (***The VLDB Journal***), 28(6), 923-940, 2019
4. Sibo Wang, Renchi Yang, **Runhui Wang**, Wenqing Lin, Xiaokui Xiao, Nan Tang.
   Efficient Algorithms for Approximate Single-Source Personalized PageRank Queries. Transactions on Database Systems (***TODS***), 44(4): 18:1-18:37, 2019

WORKING EXPERIENCES

Megagon Labs, Research Scientist Intern 2021
Contrastive Self-supervised Learning for End-to-End Data Integration. *Mountain View, CA, U.S.*

Worked on Sudowoodo, a multi-purpose Data Integration and Preparation (DI&P) framework based on contrastive representation learning (CL). Sudowoodo features a unified, matching-based problem definition capturing a wide range of DI&P tasks including Entity Matching (EM) in data integration, error correction in data cleaning, semantic type detection in data discovery, and more. CL enables Sudowoodo to learn similarity-aware data representations from a large corpus of data items (e.g., entity entries, table columns) without using any labels. The learned representations can be either directly used or facilitate fine-tuning with only a few labels to support different DI&P tasks. Our experiments show that Sudowoodo achieves state-of-the-art results on different levels of supervision and outperforms previous best specialized blocking or matching solutions for EM. Sudowoodo also achieves promising results in data cleaning and semantic type detection tasks showing its versatility in DI&P applications.

RESEARCH EXPERIENCES

1. **Rutgers University** 2019 - 2021
   Similarity Search on Billion-Scale High-Dimensional Data. *Piscataway, NJ, U.S.*

- Worked on generic framework for scaling up all-pair inference. Given two sets of vector representations X and Y, a score function (it can be trained models) for measuring pairs, the all-pair inference finds best k elements from Y for each elements in X. Our novel techniques aim to avoid enumerating every pair between two sets while achieving excellent results, and support various score functions. We achieve up to more than 1000x speedup with a nearly perfect recall on multiple real-world datasets and models.

- Developed techniques to compress and search high dimensional data. Specifically, we first apply vector quantization, which quantizes a high dimensional vector to a quantization code. Then, we propose an $O(n)$ algorithm, DeltaPQ, to compress n quantization codes in a lossless manner and perform read-intensive similarity search queries directly on compressed codes with reduced space overhead. Experimental results on five billion-scale real-world datasets show that DeltaPQ achieves a compression ratio of up to 5 (and often greater than 2) whereas state-of-the-art general-purpose lossless compression algorithms barely work. The on-disk search significantly outperforms state-of-the-art on-disk methods.

2. **University of Queensland** 2017 - 2019
Graph Data Management and Parallel Computing. *Brisbane, Australia*

- Proposed PAFO for parallel execution of single-source *Personalized PageRank* queries on large scale graph data. PAFO includes a forward push phase and a random walk phase, and we present optimization techniques to both phases. Our contributions include effective maintenance of active nodes, cache-aware scheduling for reducing contention, and counting-based random walk for lower memory overhead.
- Extensive experimental evaluation on graphs with billions of edges demonstrates that PAFO achieves up to 37× speedup on 40 cores and 3.3× faster than alternatives on 40 cores. Moreover, our parallel forward push algorithm alone is 4.8× faster than the state-of-the-art parallel graph processing framework.

PROFESSIONAL SERVICES

External Reviewer: *ICDE, VLDB, The VLDB Journal, WSDM*

AWARDS

UQ Research Training Tuition Scholarship (University of Queensland)	2018-2019
Research Higher Degree Scholarship (University of Queensland)	2018-2019
Merit Student (Outstanding undergraduates in Peking University)	2015
May 4th Scholarship (Outstanding undergraduates in Peking University, 9/160)	2015

SKILLS

C++, C, Cilk Plus, Java, Swift, Python; Developing Android, iOS applications; Linux