Yue LIU (刘悦)

yliumh@connect.ust.hk

Biography

I am a first-year Ph.D. student in Computer Science in DSF Lab, HKUST, supervised by Prof. Xiaofang Zhou. Before that, I did four months' Research Assistant in DSF Lab, where I focused on the large-scale data cleaning frameworks for Large Language Models. Before that, I obtained my master's degree in Computer Science from Zhejiang University, supervised by Prof. Xiaoye Miao, and obtained my bachelor's degree in Applied Mathematics from Zhejiang University. I have worked with Prof. Yunjun Gao, Prof. Lu Chen, and Prof. Qinghai Zhang from Zhejiang University.

I maintain an online personal website (https://yliuhz.github.io/) and a blog site (https://yliuhz.github.io/blogs/) to share my ideas about research papers.

Recent Research Projects

[1] **The Hong Kong GPT Project**

We implement our corpus data cleaning framework based on CC-Net, which is targeted for large-scale dataset, like CommonCrawl (~120TB raw data for 15 months), and based on Clean-dialog, which implements other fine-grained rules for smaller datasets, like keywords filtering.

Regarding large-scale data cleaning and data quality screening, we are working on three research topics:

- How to optimize the data mixtures for pretraining?
- How to filter out low quality corpus data?
- How to deduplicate the corpus data more effectively and efficiently?

[2] **Trajectory-User Linking (TUL) Problem**

The TUL problem aims to link anonymous trajectories with users, which has broad applications in criminal investigation to personalized POI recommendation. To effectively embed the locations and trajectories is essential to improve the linking accuracy.

[3] **Statistical Methods for (Self-supervised) Network Analysis**

The community (or block, cluster) is a powerful concept to analyze the graph structure. However, real-word graphs usually exhibit very complex community structures, namely assortativity to disassortativity, or even the combination of them. Modern techniques for graph mining, such as graph neural networks (GNNs), usually only deal with assortative structures. Stochastic block models, on the other hand, is a general mechanism, which can theoretically mine all kinds of structures with good interpretability. Currently, I am working on designing more regularized and more efficient stochastic block models, and may also combine them with GNNs.

Publications

[1] Miao, Xiaoye (past supervisor), **Yue Liu**, Lu Chen, Yunjun Gao, and Jianwei Yin. "Reliable community search on uncertain graphs." In 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 1166-1179. IEEE, 2022.

In this work, we study the community search problem on uncertain graphs.

[2] Shao, Jinning, **Yue Liu** (equal contribution), Jiaqi Yan, Ze-Yi Yan, Yangyang Wu, Zhongying Ru, Jia-Yu Liao, Xiaoye Miao, and Linghui Qian. "Prediction of maximum absorption wavelength using deep neural networks." Journal of Chemical Information and Modeling 62, no. 6 (2022): 1368-1375.

In this paper, we develop AI models to predict the wavelengths of molecules.

Invited Talks

2022.05 Reliable Community Search on Uncertain Graphs, ICDE 22

Work Experience / Internships

2023.03 - 2023.08 Research Assistant, DSF Lab, HKUST
2021.10 - 2022.05 AI Engineer, Hikvision Research Institute, Hangzhou, China.

Educations

2023.09 - now, Ph.D. student, Computer Science, HKUST
2020.09 - 2023.03, Master, Computer Science, Zhejiang University
2016.09 - 2020.06, Undergraduate, Applied Mathematics, Zhejiang University

Honors and Awards

2023.03 Outstanding Graduate at Zhejiang University
2022.10 China Optics Valley Scholarship (Top 1%)
2022.10 Outstanding Graduate Student
2022.10 Merit Graduate Student

Teaching Assistant

2022 Spring, Database System (credit 4.0), Zhejiang University