

Data Science Seminar Series

In Collaboration with
The Department of Data Science

Structure-Enhanced Text Mining for Understanding and Augmenting Scientific Discovery

Hosted by Shuai Zhang

Yu Zhang

University of Illinois Urbana-Champaign

Date: Tuesday, February 6, 2024
Time: 2:30 PM - 3:30 PM (Coffee served at 2:15 PM)
Location: GITC Building Room 4402 (4th floor lecture hall)
Web Link: [Zoom Meeting Link](#)

Language models pre-trained on large-scale text corpora have achieved remarkable success in building text mining systems. Meanwhile, text is usually accompanied by various types of structural signals, such as document metadata, concept ontologies, and citation networks, that can potentially benefit the understanding of text. To enhance the effectiveness of text mining methods, my research focuses on teaching language models to exploit structural information for both fundamental tasks and advanced domain-specific applications, with an emphasis on understanding and augmenting scientific discovery.

In the first part of the talk, I will present structure-aware classification algorithms that can predict relevant categories of a scientific paper from hundreds of thousands of candidate classes. These methods have been adapted into the Microsoft Academic Graph production pipeline. The second part of the talk will introduce seed-guided topic mining approaches that find category-indicative entities and structural signals. In the third part, I will discuss how to leverage multi-task language model pre-training techniques to facilitate advanced applications in the scientific domain, such as patient-to-article retrieval and paper-reviewer matching. Finally, I will outline future research directions, including structure-aware usage of large language models, flexible translation between different types of scientific data, and data mining for accelerating science and innovation.

Yu Zhang is a Ph.D. candidate in the Department of Computer Science at the University of Illinois Urbana-Champaign, advised by Prof. Jiawei Han. Prior to UIUC, he received his B.Sc. degree in Computer Science from Peking University. Yu's research focuses on structure-enhanced text mining and its applications in scientific literature understanding. His first-authored papers have been published in top-tier venues in the fields of data mining, natural language processing, and information retrieval. Yu has been awarded the UIUC Dissertation Completion Fellowship and the Yunni & Maxine Pao Memorial Fellowship.