A heterogeneous network-enhanced approach for knowledge recombination prediction

Mengjia Wu (Mengjia.Wu@uts.edu.au), University of Technolog Sydney
Yi Zhang (Yi.Zhang@uts.edu.au), University of Technology Sydney

1. Introduction

Knowledge recombination is one of the major mechanisms of producing scientific innovations (Fleming, 2001; Rosenkopf & Nerkar, 2001). Significant scientific breakthroughs always originate from the recombination of various types of knowledge that bridge diverse ideas and break the established scientific bonds (Audia & Goncalo, 2007). However, the continuous expansion of the depth and breadth of various disciplines is challenging researchers to think beyond their knowledge boundaries and conduct a long-jump search to realize knowledge recombination (Gavetti & Levinthal, 2000; Rosenkopf & Nerkar, 2001). Under such circumstances, leveraging the power of big data analysis and artificial intelligence has become a significant approach to assist researchers in identifying possible knowledge recombination outside their fields.

While most existing studies rely on text data (Chen et al., 2018; Liang et al., 2021), incorporating heterogeneous network features, including author, venue, and citation information, can add significant value in predicting reliable and accurate knowledge recombination. In this study, we characterize the knowledge recombination by diverse topic co-occurrence and mine the Microsoft Academia Graph (MAG) to construct a methodological framework to predict knowledge recombination. The meta-data of papers, authors, venues, and topics from MAG compose a multi-relational heterogeneous entity graph, based on which we apply heterogeneous graph neural networks (HetGNN) to transform nodes into vector representation by preserving the heterogeneous information of the graph. Following this, we define the quantitative judging criteria of knowledge recombination and build the train and test datasets to train the prediction model.

2. Data and Methods

Our research framework is illustrated in Figure 1. In the following subsections, we will detail the input and our experiment steps.

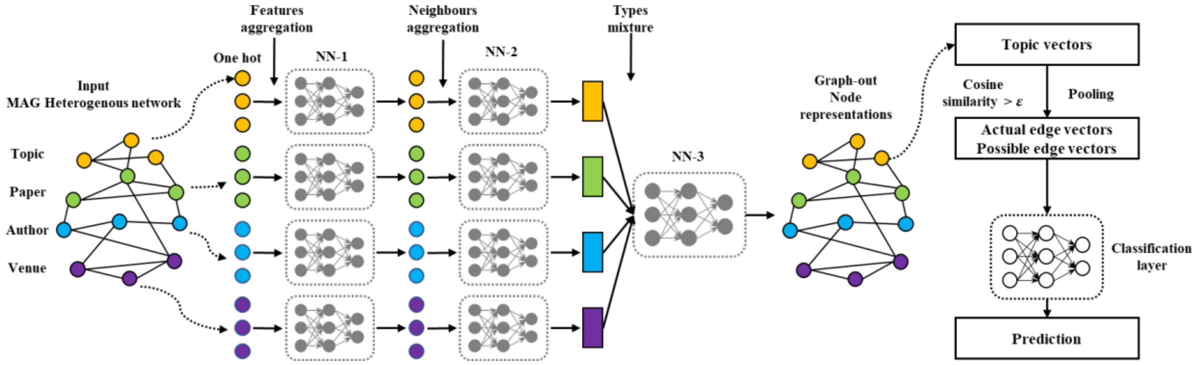

Figure 1. Research framework **2.1.**

Data collection

With the aid of the AMiner platform (Tang et al., 2008), we accessed Microsoft Academic Graph (MAG) and collected bibliographic metadata of 240 million articles, 243 million authors, and 53,422 venues. The collected data composes a large heterogeneous network consisting of four categories of heterogeneous bibliometric entities and nine types of relationships between pairwise entities: Co-authorship, authorship, paper citing, and co-occurrences of author-venue, author-topic, paper-topic, paper-venue, topic-topic, and topic-venue.

2.2. Heterogeneous graph neural networks (HetGNN)

Graph neural networks (GNNs) are designed to learn the connectivity of graphs and transform nodes, edges, and entire graphs into vector representations. A GNN adopts graph-in, graph-out architecture that preserves its symmetries and does not change the connectivity of the graph. Since we have a heterogeneous network, we decide to adopt the heterogeneous graph neural networks (HetGNN) to realize the transformation. HetGNN is a refined GNN technique that utilizes the heterogeneous information of node and edge types (Zhang et al., 2019). It adopts the heterogeneous network as input and outputs the nodes in vector representations.

2.3. Knowledge recombination definition

We use research topics provided by MAG to characterize knowledge in the scientific literature. Given that, we will use the co-occurrence of two topics to proxy knowledge recombination in the following explanation. Specifically, we design the following three criteria for identifying knowledge recombination of two topics referring to (Kaplan & Vakili, 2015):

1. The two topics appear together (co-occur) in a paper for the first time.
2. The two topics both exist already; None of them is newly generated.
3. The two topics are diverse and distant enough based on the established knowledge system; Otherwise, their co-occurrence will become an in-domain local search (Ethiraj & Levinthal, 2004; Gavetti & Levinthal, 2000).

The previous two criteria guarantee the novelty of topic recombination in a retrospective view; The last criterion guarantees that we focus on long-jump recombination instead of local search according to knowledge recombination theories (Ethiraj & Levinthal, 2004; Gavetti & Levinthal, 2000). For example, we deem the co-occurrence of *genome sequencing* and *machine learning* as knowledge recombination but ignore the case of *feature engineering* and *machine learning*.

2.4. Topic combination prediction

With the HetGNN aligning each topic to a unified feature space, the remaining task becomes a typical link prediction problem. The concatenation of two node vectors will represent any possible edges. According to our third topic recombination criterion, the topics are supposed to be diverse from each other. Hence, we will only sample the node pairs with a cosine similarity lower than ε when building train and test datasets, in which we empirically set ε as 0.5. Following the pilot studies (Zhang et al., 2021), we will design random removal and roll-back experiments to test and validate the performance of the built model.

References

Audia, P. G., & Goncalo, J. A. (2007). Past success and creativity over time: A study of inventors in the hard disk drive industry. *Management Science, 53*(1), 1-15.

Chen, C., Wang, Z., Li, W., & Sun, X. (2018). *Modeling scientific influence for research trending topic prediction.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Ethiraj, S. K., & Levinthal, D. (2004). Modularity and innovation in complex systems. *Management Science, 50*(2), 159-173.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117-132.

Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly, 45*(1), 113-137.

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435-1457.

Liang, Z., Mao, J., Lu, K., Ba, Z., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management, 58*(5), 102611.

Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal, 22*(4), 287-306.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *Arnetminer: extraction and mining of academic social networks.* Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). *Heterogeneous graph neural network.* Paper presented at the Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.

Zhang, Y., Wu, M., Miao, W., Huang, L., & Lu, J. (2021). Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. *Journal of Informetrics, 15*(4), 101202. doi:https://doi.org/10.1016/j.joi.2021.101202