Research Group:	The institute of applied informtics and formal description methods (AIFB) The institute of theoretical informarics, algorithm engineering(ITI)
Supervisors:	Chen Shao
Project Type:	Master Thesis/Hiwi in Computer Science Engineering/Economic Informatics

Research Project Proposal

Large Language Model-enhanced Graph Message Passing Network for Link Prediction

Improving text features is a fundamental aspect of academic recommendation networks. Within academic citation networks, it's essential to grasp the essence of texts, assess the innovativeness of new documents, and uncover connections between texts. In the fast-paced world of large technology companies, keeping pace with emerging algorithms and publications is vital. Identifying relevant work, evaluating it accurately, and systematically characterizing existing methods are critical to maintaining competitiveness.

The graph enriched with textual features is termed as a Text-Attributed Graph (TAG). In citation graph, each paper is represented as a node, encompassing details like the title and abstract. The adjacency matrix of the graph describes the relationships between the papers. Other TAG encompasses knowledge graphs [1][2] and product graphs/product co-purchasing network **??** as well. In these graphs, nodes represent entities and users/items, and the adjacency matrix reflects their knowledge structure and commercial behavior, respectively. **In this project, we mainly focus on citation network, and simplify it as an undirected and unweighted graph. Our objective is to predict the latent citation relationships between papers. We randomly select part of existing citation relationships among papers, use their link relationships and text features to train the network**, and output results either directly predicting new citation relationships, or performing other graph-supervised learning tasks through transfer learning, such as community detection and node classification of citation networks, subgraph classification.

Existing relationship predictions mostly rely on using Bag of Words (BoW)[3] to extract text features, which cannot comprehend and extract concept-level text features based on understanding. However, for text citation networks, comprehension is a necessary algorithmic capability, especially when the citation data itself contains some noisy and missing links. Besides, it also enables the generation of human-readable interpretations for error analysis and optimization of the datasets. However, recent works have indicated that there is a certain trade-off between the effective utilization of features and the effective utilization of structural features [4]. In other words, relationship prediction algorithms based on message passing algorithms can either fully extract structural features or efficiently utilize node features, but not both simultaneously.

Work Package 1: Benchmark for relationship prediction on TAG

Therefore, in this project, our aim is to delve deeply into the maximum accuracy achievable by these two strategies: namely, utilizing language models exclusively and relying solely on the structural characteristics of graphs for relationship prediction. We also intend to thoroughly investigate several key factors influencing accuracy. 1) Regarding the former approach, we will explore aspects such as text length, pre-training tasks, label function design, and analysis of error patterns. 2) As for the latter approach, our focus includes assesssing its capacity to fully reconstruct the existing structure from a statistical standpoint of the graph. Building upon this analysis, we endeavor to devise a graph algorithm capable of harnessing both sets of features simutaneously. 3) In addition, all above mentioned work are based on the fair benchmark of the currently popular large language models with comparable metrics and unique data split and conduction of a cost-effective analysis from the perspective of performance vs. complexity.

Many exisiting pretrain methods in TAG primarily emphasize node classification, where node features are derived from pretraining task such as supervised text classification. However, the identification of an representative pretrain task specifically tailored for relationship prediction has not been thoroughly examined and investigated. Essentially, any pretraining task that encompasses the relationship between two textual entities could be employed, effectively capturing first-order structural features based on textual relevance. Nevertheless, the exploration of methods to leverage higher-order graph structural features derived from these textual features remains an unexplored area of research.

A foundational step in this research involves benchmarking existing algorithms, encompassing heuristic methods [5], embedding techniques [6, 7, 8], vanilla graph message passing networks [9][**?**], and subgraph convolutional neural networks[10][11][12] tailored for relationship prediction, all of which have publicly available code. In this phase, we aim to scrutinize the characteristics of diverse algorithm types and analyze their error patterns. This analysis will extend to a broader array of datasets, including fb238k[12], wn12rr[12], and product networks[13]. Following the benchmarking process, the dataset will undergo evaluation using a recent paperwithcode dataset compiled by AIFB, containing abstracts and titles until 2023. Although this dataset may be limited in size and may not be sufficient for fully training a state-of-the-art relationship predictor, it presents numerous compelling test cases for evaluating such a method and potential fine-tuning. In a subsequent phase of the project, should time permit, we will consider assessing more advanced deep learning techniques, such as employing Transformers or other attention-based methods on subgraphs. Ultimately, the resulting method will be integrated and compared with currently benchmarked algorithms under comparable conditions.

Next we detail the milestones:

- Preliminary steps: Get familiar with the topic, understanding the task and problem formulation as well as existing datasets and baseline methods. Get familiar with the tools, e.g. Python, PyTorch or TensorFlow.

- Milestone 1: Conduct a literature review and identify the most promising method(s) and dataset(s) to address the task at hand, including e.g. also approaches that work with depth from Cora, Citeseer and related datasets.

- Milestone 2: Study and evaluate existing publicly available baseline methods on data provided by AIFB. Identify sucecess and failure cases both qualitively and quantitatively, by implementing relevant algorithms and metrics.

- Milestone 3: Extend or fine-tune existing baseline methods for better performance on Cora, Citeseer, Pubmed, Arxiv2023, FBK235k and WN18RR.

- Milestone 4: Propose, implement and evaluate a novel approach for link prediction, by benchmarking listed methods or investigating more complex network architectures.
- Milestone 5: Discussion and possible re-iteration of Milestones 2-4.
- Milestone 6: Thesis writing and final presentation. There will be regular meetings with the project supervisors at times/dates to be defined.

The candidate's judgment will be based on achieving goals, along with criteria such as motivation, autonomy, understanding of the topic, academic communication skills, creativity, theoretical soundness, implementation quality, and report/presentation quality.

References

[1] Mohamad Zamini, Hassan Reza, and Minou Rabiei. A review of knowledge graph completion. *Information*, 13(8):396, 2022.

[2] Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4710–4723, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[4] Haitao Mao, Juanhui Li, Harry Shomer, Bingheng Li, Wenqi Fan, Yao Ma, Tong Zhao, Neil Shah, and Jiliang Tang. Revisiting link prediction: A data perspective. *arXiv preprint arXiv:2310.00793*, 2023.

[5] David Liben-Nowell and Jon Kleinberg. The Link Prediction Problem for Social Networks. *Proceedings of the conference on information and knowledge management*, 2003.

[6] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, August 2014. arXiv:1403.6652 [cs].

[7] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, San Francisco California USA, August 2016. ACM.

[8] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, May 2015. arXiv:1503.03578 [cs].

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[10] Seongjun Yun, Seoyoon Kim, Junhyun Lee, Jaewoo Kang, and Hyunwoo J Kim. Neo-gnns: Neighborhood overlap-aware graph neural networks for link prediction. *Advances in Neural Information Processing Systems*, 34:13683–13694, 2021.

[11] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.

[12] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, 34:29476–29490, 2021.

[13] K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. The extreme classification repository: Multi-label datasets and code, 2016.