Proposal: Suggested Academic Reading Sequence Recommendation Using Graph Neural Networks

Bowen Li

February 16, 2024

Abstract

Increasing output in the volume of academic publications over the years will inevitably lead to problems when it comes to allowing a user to be able to pick out papers that are relevant and/or useful to them, with citations being the only sense of structure one can rely on. Incidentally, graph neural networks has seen a recent surge in interest and developments, presenting new ways to apply deep learning techniques on data that can be represented in the structure of a graph. This project seeks to apply graph neural networks to create a recommendation system that, given a paper/topic, will provide users with a suggested sequence of papers to read in order to further understand the paper/topic, ultimately assisting users in navigating the ever increasingly complex landscape of academic publications.

Introduction

As interest in the fields of science and technology has grown over the past few years, and inevitably will continue to grow in the coming years, the number of academic publications will undoubtedly skyrocket. Although this certainly serves as an indicator of the accelerating advances made across these fields, it could also pose a problem for future generations, hobbyists and academics alike, as it increasingly makes the overhead of searching through and selecting relevant reading material more costly.

Typically, when an interested party happens upon a relevant paper, the references of that paper could serve as a suitable starting point for further reading. However, not only could the references of the paper be dubious due to some of the meta motivations concerning the publication and review process, it is still lacking in terms of reading about future developments. After all, a paper is not physically able to cite something that, as of its writing, has yet to exist.

This leaves the interested party to rely on search engines, scouring relevant conferences (the volume of which is only increasing), and gaming their own social media recommendations.

The last option could involve one following relevant figures on social media, whom in turn follow other relevant figures that may eventually all be taken into account when recommending new papers to users. This is reminiscent of a graph structure, and is suitable for users who may wish to keep up with the latest in terms of the field. In seeking to build a recommendation system that provides more structure to the users however, we can still leverage this idea of graphs.

Graph neural networks (GNNs) have seen a recent surge in popularity as a paradigm that applies deep learning techniques to data with a graph structure. One of the main problems that it addresses is link prediction, wherein a graph is assumed to have missing edges, and the task is, for a given node, to predict which of the other nodes in the graph it would most feasibly share an edge with. With a good choice of nodes and edges, this task steadily lends itself to being applied as a recommendation system.

This project will explore the potential of link prediction to be applied to a recommendation system that will provide users with a suggested reading sequence of academic publications in order to familiarize themselves with a given topic. This way, if a user is considering a paper, they would be recommended papers to read as prerequisites and suggestions for further reading. In the same vein, if a user has a topic in mind (e.g. attention mechanism), they would be recommended readings to familiarize themselves with the fundamentals of the topic, and readings that explore the applications and further developments of that topic.

Related Work

The task of link prediction is a well-studied problem across various fields. [1] provides a survey on various classical link prediction approaches and comparisons. [2, 3] both provide classical link prediction solutions in the context of recommender systems, one based on vector similarity and the other using bipartite graphs.

In terms of GNNs, [4, 5] provide surveys and taxonomies of GNN techniques applied to recommender systems. [6] presents a particular GNN approach to the recommendation problem based on a new version of the dual-tower model for heterogeneous graphs. [7] presents another GNN approach to creating a recommender system based on a heterogenous graph with users, articles, and article topics as nodes. Most notably, the topic nodes are latent, and are thus learned from the dataset. [8] presents a GNN approach that reconstructs graph components into dense item-item graphs based on a user's history and perceived interests. [9] presents a GNN approach to creating a recommender system for Massively Open Online Courses (MOOCs), though doesn't provide a recommended sequence to take courses in.

Proposed Work

The target recommendation system is one wherein a user is able to provide either a topic or a published paper, and the system provides a sequence of recommended prior readings and recommended further readings. The length of the sequence and number of successive prerequisites can be adjusted via the user's preferences.

This project will primarily involve creating a proof-of-concept for the target recommendation system using GNNs.

The graph representation of the dataset of papers will be instrumental in deciding the approaches and limitations. The current plan is to have a heterogeneous graph with two node types: one for publication embeddings and one for topic embeddings. The publication embeddings can be learned from the paper's details such as the abstract and conference. The topic embeddings can be learned through techniques such as LDA as in [Hu et al.]. The edges, then, will be directed, pointing from citing papers to cited papers, from topics to papers that apply them, from papers that provide foundations for a certain topic to that topic, and from prerequisite topics to more advanced topics.

This way, when a user provides a new paper's information, that information can be embedded and the system will predict incoming edges (prerequisite readings) and outgoing edges (further readings). Alternatively, if a user were to provide a description of a topic, the topic can be embedded (e.g. using a language model), then either its edges can be predicted directly or the edges of its nearest topic node neighbor.

The GNN architecture used will likely be a foundational architecture such as graph convolutional networks through the PyTorch Geometric (PyG) library. Time-permitting, other architectures such as graph attention networks could be explored. Another stretch goal could include accounting for more information in the paper embeddings such as citation count and date published in order to give preference to more popular or timely papers.

Datasets

Currently, the planned datasets to use are publicly available citation graphs. [10] provides two citation graph datasets from the Stanford Network Analysis Project (SNAP). One is from Arxiv HEP-PH (high energy physics phenomenology) and the other from Arxiv HEP-TH (high energy physics theory). HEP-PH includes 34,546 papers with 421,578 edges published in the period from January 1993 to April 2003. HEP-TH includes 27,770 papers with 352,807 edges among them in the same period. In both of these datasets, edges are directed from the citing paper to the cited paper.

[11] provides much larger datasets that will likely only be explored time-permitting. The citation data is represented in a similar way, except the first version contains 629,814

papers and 632,752 citations while the most recent version contains 4,894,081 papers and 45,564,149 citations.

Evaluation

Based on the similarity of the problems, the planned evaluation metric will be similar to [9] wherein Rec@k scores are measured, which represent the percentage of items in the the top-k recommended list. As such, the initial baselines will also be based on the baselines from [9], one of which being POP, wherein the most popular items are recommended, which the paper claims is a common baseline in recommender systems. Another possible baseline that may perform better would be selecting recommendations using purely similarity metrics between embeddings, disregarding the graph structure. These will account for the relevance of the recommendations themselves, however further reading and another metric may be needed to evaluate the quality of the recommended ordering of content. One possibility could be to favor a chronological ordering of papers, as in most cases, earlier papers tend to build the foundations that later papers work off of.

Timeline

Week 1 (Starts 2/19): Complete bulk of research and reading of relevant literature.

Weeks 2-3: Initial EDA and data visualization.

Deliverables: Some visualization of the data.

Weeks 4-5: Data augmentation and feature extraction.

Deliverables: Mid-project check-in. Finalized feature graph to run the GNN on, including content and topic embeddings.

Weeks 6-7: Creating and tuning model.

Deliverables: Initial runs and results. Baseline measurements.

Weeks 8-9: Experimentation and evaluation.

Deliverables: More detailed results and visualizations.

Week 10: Wrap-up and report.

Conclusion

The project aims to apply GNNs to build a proof-of-concept for a recommendation system to assist an interested party in navigating the ever-growing expanse of available academic works.

The real-world dataset of academic publications would be represented in an embedding space as nodes along with relevant topics, and edges would be directed from citing papers to cited papers. In solving a link prediction task, one could retrieve a set of nodes that are likely to point towards a given node, and a set of nodes that the given nodes are likely to point towards.

After some refinement based on the user's preferences, the user would be presented with a recommended sequence of papers to read in order to both be onboarded to and to gain deeper understanding of a desired topic.

References

[1] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, Bhaskar Biswas. *Link prediction techniques, applications, and performance: A survey.* Physica A: Statistical Mechanics and its Applications, 553, 2020. `https://doi.org/10.1016/j.physa.2020.124289`

[2] Zhan Su, Xiliang Zheng, Jun Ai, Yuming Shen, Xuanxiong Zhang. *Link prediction in recommender systems based on vector similarity.* Physica A: Statistical Mechanics and its Applications, 560, 2020. `https://doi.org/10.1016/j.physa.2020.125154`

[3] T. Jaya Lakshmi, S. Durga Bhavani. *Link prediction approach to recommender systems.* Computing, 2023. `https://doi.org/10.1007/s00607-023-01227-0`

[4] Shiwen Wu, Fei Sun, Wentao Zhang, X Xie, Bin Cui. *Graph Neural Networks in Recommender Systems: A Survey.* ACM Computing Surveys, 55(5): 1-37, 2022. `https://doi.org/10.48550/arXiv.2011.02260`

[5] Zijian Liang, Hui Ding, Wenlong Fu. *A Survey on Graph Neural Networks for Recommendation.* International Conference on Culture-oriented Science & Technology (ICCST): 383-386, 2021. `https://doi.org/10.1109/ICCST53801.2021.00086`

[6] Qiang He, Xinkai Li, Biao Cai. *Graph neural network recommendation algorithm based on improved dual tower model.* Scientific Reports, 14, 3853, 2024. `https://doi.org/10.1038/s41598-024-54376-3`

[7] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, Chao Shao. *Graph neural news recommendation with long-term and short-term interest modeling.* Information Processing & Management, 57(2), 2020. `https://doi.org/10.1016/j.ipm.2019.102142`

[8] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, Yong Li. *Sequential Recommendation with Graph Neural Networks.* Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval: 378-387, 2021. `https://doi.org/10.48550/arXiv.2106.14226`

[9] Jingjing Wang, Haoran Xie, Fu Lee Wang, Lap-Kei Lee, Oliver Tat Sheung Au. *Top-N personalized recommendation with graph neural networks in MOOCs*. Computers and Education: Artificial Intelligence, 2, 2021. `https://doi.org/10.1016/j.caeai.2021.100010`

[10] Johannes Gehrke, Paul Ginsparg, Jon Kleinberg. *Overview of the 2003 KDD Cup*. SIGKDD Explorations 5(2): 149-151, 2003. `http://www.cs.cornell.edu/home/kleinber/kddcup2003.pdf`. Retrieved from `https://www.kaggle.com/datasets/wolfram77/graphs-citation/data?select=cit-HepTh-abstracts`.

[11] Mathurin Aché. *Citation Network Dataset.* March 2021. Retrieved from `https://www.kaggle.com/datasets/mathurinache/citation-network-dataset/data`