CS3319-01 Project (2024 Spring)

将图算法和图神经网络应用于现实世界问题的大作业项目。

在这个项目中，你需要从 2 项作业中选择 1 项作业（同时进行这两项作业也可以）。

Assignment 1. GNN over Recommendation Senario

Introduction

GNN 在异质图和同构图的节点分类和链路预测任务中表现非常好。在很多推荐场景中，利用用户之间的信任关系和产品的本质属性，结合用户的购买记录，使用 GNN 生成的表征也能达到很好的效果。同时，在学术平台中，合作者推荐、论文推荐、期刊和会议的审稿人推荐是主要任务，GNN 在这些任务上的表现依然良好。

对于上述问题，都有相对较好的基准指标和模型。但是，在论文推荐的场景中，如果是通过合作者和论文所在的领域和合作社群进行科学研究的学者，引文推荐仍然缺乏更好的数据集和模型。本选题提供的数据集用在一个学术推荐系统中，其中"用户"为学术论文的作者，"产品"为学者的论著。当用户和产品都有关联网络时，我们可以提取用户群体和产品类别的群体特征，基于产品之间的关系和用户之间的联系，为用户提供更好、更多样的推荐，这可能有助于解决推荐系统冷启动造成的用户行为较少的问题。

由此，你需要解决一个学术网络中的推荐问题。我们从地理科学领域的顶级期刊中收集了 6,611 位作者和相应的 79,937 篇论文，以及他们出版物的引文信息。你需要利用收集到的信息形成一个学术网络，这里有一个可行的方法：

建立一个异质网络，其中包含两类节点，一类节点代表作者，另一类代表论文。在这个网络中，作者节点和论文节点之间的每条边都表示作者阅读过该论文（连接作者和被作者所写的论文所引用的论文），两个作者节点之间的每条边表示合著关系，两个论文节点之间的每条有向边表示引用关系。

你可以使用其他方式来构建学术网络。请注意，我们所提供的是迄今为止的作者和论文信息，它揭示了不同作者的工作之间的关联性。因此，假设你正在设计一个学术阅读推荐系统，你需要挑选出与作者以前研究相关的论文。这个问题可以被模拟成一个链路预测问题，你的任务是根据所提供的信息预测测试集中的每个作者-论文对。如果该论文被推荐给作者，则标记为 1，否则标记为 0。

Data

节点信息：6611 个作者，79939 篇论文

边信息：引用信息

bipartite_train(test)_ann.txt：(author, paper)代表某个作者引用了某篇论文

author_file_ann.txt：(author, author) 代表两个作者合作过

paper_file_ann.txt ：(paper1, paper2)代表论文 1 引用了论文 2

提交格式：根据 bipartite_train_ann.txt 给出的作者论文对，判断是否把论文推荐给作者

Required Files

1. 算法的设计报告，以会议论文的形式呈现。
2. 最终用于性能评估的代码（与 Kaggle 最终决定用于测试的代码一致）

References

1. 【arXiv 2020】 Graph neural networks in recommender systems: a survey.
2. 【arXiv 2021】 Graph learning based recommender systems: A review.
3. 【KDD 2018】 Graph Convolutional Matrix Completion (GC-MC).
4. 【KDD 2018】Graph convolutional neural networks for web-scale recommendersystems (PinSage).
5. 【RecSys 2018】 Spectral collaborative fltering (SpectralCF).
6. 【SIGIR 2019】 Neural graph collaborative filtering (NGCF).

7. 【SIGIR 2020】 Lightgcn: Simplifying and powering graph convolution network for recommendation (LightGCN).

8. 【SIGIR 2019】 A neural influence diffusion model for social recommendation (DiffNet).

9. 【WWW 2019】 Graph neural networks for social recommendation (GraphRec).

10. 【WWW 2019】 Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems (DANSER).

11. 【RecSys 2019】 Deep social collaborative filtering (DSCF).

Attention

该项目将在 Kaggle 平台上进行。你需要在 Kaggle 上提交结果以参与性能评估和排名，同时你需要在 Canvas 上提交其他材料。关于项目细节、数据格式、评估方法等，请查看 Kaggle 竞赛页面。

注意：大作业为小组形式，在组队完成后，不要忘记在 Canvas 上登记并加入小组。

为了公平起见，我们制定了一些规则------违反规则将导致额外的扣分：

1. 请不要复制别人的代码。我们会在提交后进行抄袭检查。

2. 请不要在其他地方下载数据集来训练你的模型。我们已经对数据集进行了随机打乱，并将使用该数据集重现你的实验结果。如果你报告的结果和我们重现的结果之间有很大的差距，将被视为违规。

3. 请不要使用预先训练好的模型。

Other references

图机器学习的最新成果和性能最先进的模型在这可以找到：

1. OGB leaderboard

2. 机器学习/数据挖掘领域顶级国际会议： 要找到与图机器学习相关的论文，你可以在对应的会议网站中搜索 "graph" 这个词。

ICML 2021: https://proceedings.mlr.press/v139/

NeurIPS 2022: https://papers.nips.cc/paper/2022
ICLR 2022: https://openreview.net/group?id=ICLR.cc/2022/Conference
KDD 2022: https://kdd.org/kdd2022/toc.html

Assignment 2. Graph Augmentation

Introduction

在常用的 GNN 模型中，图拓扑以低通滤波器的形式发挥作用，其原理是图的同配性。然而，在常用的数据集中，拓扑并不会严格按照同配性原理连接。例如，在引文网络中，计算机领域的论文就有可能被医疗、材料、经融等领域的论文引用。这样的异配连接可能给 GNN 的性能带来不利影响。然而，一些异配连接可能是沟通图的不同子图的桥梁，在信息传递中发挥着举足轻重的作用。因此，在这个任务中，你需要为图中存在/不存在的拓扑连接给出评价指标，并给出算法增加/删除一定数量的边，使图在下游任务上有更好的性能。下面是两个可行的思路：

1. 使用节点标签作为标准，尝试增加同标签的边，并删除不同标签的边。如果性能提升，则保留修改，否则回退。如此迭代进行。

2. 使用链路预测给出的链路存在概率作为标准，增加那些大概率存在的链路，删除那些已有但预测置信度低的链路。

Data

数据来源：引文网络，一张连通的同配图

数据格式：PYG 格式

数据简介：数据共有 2485 个节点， 5069 条边，每个节点有 1433 维特征，共 7 个类别的节点

数据划分：训练集每类 20 个节点，验证集共 500 个节点

References

1. 【ScienceDirect 2023】 Towards data augmentation in graph neural network: An overview and evaluation

2. 【arXiv 2022】 Data Augmentation on Graphs: A Technical Survey.

3. 【AAAI 2020】 Measuring and Relieving the Over-smoothing Problem for Graph Neural Networks from the Topological View.

4. 【IJCNN 2021】 Automated Graph Representation Learning for Node Classification.

5. 【AAAI 2021】 Data Augmentation for Graph Neural Networks.

6. 【CIKM 2020】 Data Augmentation for Graph Classification.

7. 【Graph Data Mining 2021】Subgraph Augmentation with Application to Graph Mining.

8. 【NeurIPS 2021】 An Empirical Study of Graph Contrastive Learning..

9. 【ICML 2020】 Deep Graph Contrastive Representation Learning.

10. 【WWW 2021】 Graph Contrastive Learning with Adaptive Augmentation.