电 子 科 技 大 学

专业学位研究生学位论文开题报告表

攻读学位级别： □博士 ☑硕士

培养方式： ☑全日制 □非全日制

专业学位类别及领域： 计算机技术

学 院： 计算机科学与工程学院（网络空间安全学院）

学 号： 202122081122

姓 名： 段威威

论文题目： 复杂网络中社区发现

算法的研究与实现

校内指导教师： 吴立军

校外指导教师：

填表日期： 2023 年 2 月 18 日

电子科技大学研究生院

1. 学位论文研究内容

|  |  |  |
| --- | --- | --- |
| 课题类型 | | ☑应用基础研究 □应用研究 |
| 课题来源 | | □纵向 □横向 ☑自拟 |
| 学  位  论  文  研  究  内  容 | 学位论文的研究目标、研究内容及拟解决的关键性问题（可续页）  随着信息技术的快速发展，万事万物的关系都越来越紧密，无论人们在工作还是生活中，都存在着各式各样的网络，例如社交软件用户网络、论文引用网络、交通运输网络、蛋白质网络和病毒传播网络，这些网络通常是多样性的、大规模的和可变化的，其中包含了重要的属性信息和结构信息，并且可以抽象为图论中的复杂网络。在网络中运用社区发现有助于更好地理解并分析社会网络的拓扑结构及功能，了解网络结构形成原理，进行功能分析、行为预测等具有科研指导意义和实用价值的行为，能够更好的服务于人类社会，具有现实意义[1]。  在复杂网络中挖掘社区的这一过程被称为社区发现。社区发现旨在将一个网络划分为若干个社区内节点紧密连接的团体，它的特点是在同一社区中的节点密集连接，不同社区间的节点稀疏连接[2]。也正如人际交往中，人们总是和关系密切的人来往较多，形成了一个小团体，并且可能和朋友的朋友相结识。经过划分后的社区能够将底层的网络结构和重要信息显式地表现出来。可以应用于推荐系统、社交好友推荐、疫情的检测预防等各领域。  随着网络数据的不断增长和复杂度的不断提高，社区发现算法也面临着越来越多的挑战。首先，网络的规模和复杂度不断增加，计算复杂度和存储复杂度也随之增加；其次，网络的结构和特征不断变化，需要算法能够适应动态变化的网络环境；此外，不同类型的网络有着不同的特征和规律，需要算法能够针对不同类型网络进行优化。因此，本研究旨在提高社区发现算法的准确性、效率和鲁棒性，适应不同类型的网络环境，以更好地解决实际应用问题。本研究的目标是设计和实现一种高效、准确、鲁棒的复杂网络社区发现算法，不仅可以解决非重叠社区的社区发现，并且适用于重叠社区的社区发现，以揭示网络结构和功能的内在规律，解决现实应用中的问题。具体的研究内容包括以下几点。   1. 社区发现算法的设计，进一步提升社区发现的精确度   社区发现算法的设计是指如何利用数学方法和计算机技术识别出网络中具有社区特征的节点集合，这是复杂网络研究领域的重要问题之一。社区发现算法的设计需要考虑到多个方面的问题，包括算法的效率、可扩展性、准确性、鲁棒性等方面。首先，算法的效率和可扩展性是考虑复杂网络中大规模节点数的情况下，如何快速识别出社区结构。常用的算法包括基于聚类的算法、基于模块度的算法、基于最小割的算法等。对于大规模网络，可以采用分布式计算等方法来提高算法的效率。本课题拟打算使用图神经网络（GNN）的方法进行社区发现算法的设计，采用图卷积网络(GAT)进行图数据的特征提取，使用图注意力机制融合节点自身以及邻居节点的特征。  （2）社区发现算法的改进，解决有重叠社区的社区发现问题  在过去几年的研究中，很少有人关注到重叠社区的问题。重叠社区的社区发现是指网络中存在多个社区，而且同一个节点可以属于多个社区的情况。与传统的非重叠社区发现不同，重叠社区发现需要考虑节点的多重性，即一个节点可以同时属于多个社区。本课题考虑在实现非重叠社区检测的基础上对算法进行改进，以便于对重叠社区进行社区发现。  （3）复杂网络的社区发现系统的设计与实现  拟根据一定的研究成果或算法，实现复杂网络中的社区系统。该系统期望以桌面软件或网页的形式。具有一定良好的用户交互功能，期望核心功能为：用户输入复杂的网络数据，能够在一定时间内进行社区发现，输出不同的社区及社区中包含的节点信息。 | |

1. 学位论文研究依据

|  |
| --- |
| 学位论文的选题依据和研究意义，国内外研究现状和发展态势；选题在理论研究或实际应用方面的意义和价值；主要参考文献，以及已有的工作积累和研究成果。（2000字）   1. **选题依据**   本课题主要针对复杂网络中的社区发现算法进行研究或改良，以及计算机软件形式的复杂网络中社区发现系统的软件实现。主要依据在于人工智能、图神经网络等计算机科学技术的发展和应用，以图神经网络GNN[3]，图注意力网络GAT[4]等在社区发现方面已经有了许多成果和应用，深度学习在社区发现模型的应用被证明有一定的可行性。  复杂网络的数据集很丰富，分为人工数据集和现实世界数据集。Karate数据集是 20世纪70年代美国一所大学空手道俱乐部34名成员之间的社交网络，包括34个节点和78条边[5]。Dolphins数据集是新西兰某海峡由 62 只海豚群体组成的海豚交流情况网络，经过观察人员 7 年时间的记录，包括了 62个节点和159条边[6]。Cora数据集是由2708份科学出版物组成的论文引用网络，论文分为七个类别，引文网络包含2708个节点和5429条边[7]。Citeseer数据集和Cora相同也是论文引用网络，论文分为六个类别，网络包含3327个节点和4732条边[8]。利用丰富的数据集可以进行深度学习模型的训练和测试，非常方便，具有较高的可行性。   1. **研究意义**   在复杂网络中挖掘社区的这一过程被称为社区发现。社区发现旨在将一个网络划分为若干个社区内节点紧密连接的团体，它的特点是在同一社区中的节点密集连接，不同社区间的节点稀疏连接。也正如人际交往中，人们总是和关系密切的人来往较多，形成了一个小团体，并且可能和朋友的朋友相结识。经过划分后的社区能够将底层的网络结构和重要信息显式地表现出来[9]。  社区发现是用来揭示网络聚集行为的一种技术，它主要从社区内边的密度和社区间边的稀疏程度这一角度，对网络的拓扑结构以及相关性质进行分析和研究，在过去的几十年中取得了众多成果，其成果已成功地应用于社交好友推荐、商品推荐系统、公众舆情分析和病毒传播预测等各个领域。通过社区发现，微博、Facebook等社交软件可以将有共同兴趣爱好的朋友推荐给用户。用户在使用淘宝、京东等购物软件时也能更好定位想要的商品。当科研人员想了解一个领域时，也能够通过论文的引用关系快速找到该领域高质量的论文[10]。另外，自2019年12月新冠疫情爆发以来，给世界各个国家和人民造成了巨大冲击，如果能利用社区发现模型对新型冠状病毒的传播规律进模拟和预测，提前做出预防措施，则能更好的阻断疫情的传播、保障人民的生命财产安全。因此，社区发现能够在各领域取得出色的效果，对社区发现的研究也已成为复杂网络研究的热点。  本课题旨在研究如何从复杂网络中识别出具有社区特征的节点集合，从而更好的理解网络的结构和功能，这种研究是很有价值、也很有必要的。   1. **国内外研究现状和发展态势**   目前主流的传统社区发现方法大致可分为基于层次聚类、基于标签传播、基于模块度优化、基于派系过滤和基于随机游走等方法[11]。传统的社区发现方法主要基于邻接矩阵，然而深度学习可以创建更强大的节点属性和社区结构的表示。深度学习在社区发现领域的明显优势是它能够对高维数据的特征表示进行编码，可以将拓扑网络的高维输入转换为易于处理且能表示网络重要特征的低维表示。深度学习对与大规模图数据相关的稀疏特征更有弹性，同时在许多现实场景中，大多数节点信息都是未标记的，而且数据中的社区几乎没有先验知识，因此深度学习是无监督学习任务的首选。此外，在深度学习过程中，可以共同识别来自节点、边、邻居或多图的信息组，从而得到有效的社区检测结果[12]。基于深度学习的社区发现方法主要分为深度神经网络、深度图嵌入和图神经网络三类。   1. 深度神经网络   深度神经网络（Deep Neural Network, DNN）在构建和获取全局关系方面具有得天独厚的优势。在社区检测领域中，最流行的三种深度神经网络模型分别是卷积神经网络（Convolutional Neural Network, CNN）[13]、自动编码器（AutoEncoder,AE）[14]和生成对抗网络（Generative Adversarial Network, GAN）[15]。卷积神经网络（CNN）的两个关键组成部分是卷积操作和在卷积层上的池化操作。卷积层主要用来提取数据的关键特征，随后池化层降低数据的维度以确保CNN的鲁棒性。自动编码器（AE）作为一种无监督学习的人工神经网络，在特征提取方面应用十分广泛，其在图像处理领域的成功使得研究者们尝试着将自动编码器应用于社区发现领域。生成对抗网络（GAN）涉及到两个深度神经网络的相互竞争，从而实现训练精度的快速调整，通常使用无监督生成与训练集相同规模的新数据。   1. 深度图嵌入   深度图嵌入是一种将网络中的节点映射到低维向量空间的技术，同时在表示中保存尽可能多的结构信息。图嵌入方法适用于基于网络分析的深度学习任务，如链路预测、节点分类和节点聚类，因为它们可以引用表示中的潜在特征。非负矩阵分解（Nonnegative Matrix Factorization , NMF）[16]是一组计算算法，它将一个矩阵分解成两个矩阵，其特性是三个矩阵都没有负元素。对于社区发现来说，NMF 方法通过最小化误差函数将网络的邻接矩阵近似为两个因式分解矩阵的乘积，以进一步完成聚类任务。   1. 图神经网络   图卷积神经网络（Graph Neural Network, GCN）[17]作为图神经网络中的一项重要技术，它继承了卷积神经网络（CNN）快速学习的优点，在拓扑图中能高效的提取空间特征。GCN 采用了一些传统的社区发现方法作为深度图算子，如用于统计推断的SBMs[18,19]、用于频谱分析的拉普拉斯矩阵和用于信念传播的概率图模型。基于图注意力网络的方法可以在更复杂的场景检测社区。图注意力网络（Graph Attention Network，GAT）[20]通过可训练的权重聚合邻域中节点的特征并通过最大化全局共享特征来计算网络嵌入以检测社区，对注意力系数进行正则化降低了在嵌入时学到不重要关系的可能性。   1. **选题在理论研究或实际应用方面的意义和价值**   随着信息时代的到来，各式各样的网络已经融入到人们生活的方方面面，例如人际关系网络、论文引用网络，这些网络具有复杂性、多样性的特点[21]。社区发现则是指在网络图中寻找具有相似特征的社区结构，以了解它们的拓扑结构和属性信息，从而能够运用于分类、预测等任务，服务于现实社会，因此社区发现具有重要的现实意义。  虽然现今在简单网络数据中社区发现已经有了较高的准确率和广泛的应用，但是当面对着复杂的网络数据。网络数据信息包含着很多无效信息、干扰信息，如网络结构较为复杂以及存在重叠社区等。这种干扰信息导致网络中的社区发现算法相对于在控制条件下的准确率急剧下降。本课题旨在研究如何更加高效的利用网络中的结构信息以及结点特征，这种研究是很有价值、也很有必要的。   1. **主要参考文献**   [1] RANI S, MEHROTRA M. Community Detection in Social Networks: Literature Review[J]. Journal of Information & Knowledge Management, 2019, 18(02):1950019.  [2] TENG X, LIU J, LI M. Overlapping Community Detection in Directed and Undirected Attributed Networks Using a Multiobjective Evolutionary Algorithm[J]. IEEE Transactions on Cybernetics, 2021, 51(1): 138-150.  [3] DING Z, CHEN X, DONG Y, et al. Consensus reaching in social networkDe Groot Model: The roles of the Self-confidence and node degree[J].Information Sciences, 2019, 486: 62-72.  [4] LIU F, XUE S, WU J, et al. Deep Learning for Community Detection: Progress,Challenges and Opportunities[J]. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020: 4981-4987.  [5] XU R, CHE Y, WANG X, et al. Stacked autoencoder-based community detection method via an ensemble clustering framework[J]. Information Sciences,2020, 526: 151-165.  [6] ATAY Y, KOC I, BABAOGLU I, et al. Community detection from biological and social networks: A comparative analysis of metaheuristic algorithms[J].Applied Soft Computing, 2017, 50: 194-211.  [7] TANG J, HU X, LIU H. Social recommendation: a review[J]. Social Network Analysis and Mining, 2013, 3(4): 1113-1133.  [8] SHCHUR O, GÜNNEMANN S. Overlapping Community Detection with Graph Neural Networks[J]. ar Xiv e-prints, 2019: ar Xiv:1909.12201.  [9]尚敬文，王朝坤，辛欣，等. 基于深度稀疏自动编码器的社区发现算法[J]. 软件学报，2017，28：648-662.  [10] WANG F, ZHANG B, CHAI S. Deep Auto‐encoded Clustering Algorithm for Community Detection in Complex Networks[J]. Chinese Journal of Electronics,2019, 28(3): 489-496.  [11] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.  [12] NEWMAN M E J, MOORE C, CLAUSET A. Finding community structure in very large networks[J]. Physical Review E, 2004, 70(6): 066111.  [13] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007,76(3): 036106.  [14] GUO K, GUO W, CHEN Y, et al. Community discovery by propagating local and global information based on the MapReduce model[J]. Information Sciences,2015, 323: 73-93.  [15] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2010, 12(10): 103018.  [16] BRANDES U, DELLING D, GAERTLER M, et al. On Finding Graph Clusterings with Maximum Modularity[C]. BRANDSTÄDT A, KRATSCH D,MÜLLER H. Graph-Theoretic Concepts in Computer Science. Berlin, Heidelberg:Springer, 2007: 121-132.  [17] LI S, CHEN Y, DU H, et al. A genetic algorithm with local search strategy for improved detection of community structure[J]. Complexity, 2010, 15(4): 53-60.  [18] XIE Y, WANG X, JIANG D, et al. High-performance community detection insocial networks using a deep transitive autoencoder[J]. Information Sciences,2019, 493: 75-90.  [19]刘苗苗，郭景峰，潘晓. 加权网络社区发现与链接预测方法研究[D]. 燕山大学, 2017.  [21]CHEN Z, LI X, BRUNA J. Supervised Community Detection with Line Graph Neural Networks[J]. 2020, ar Xiv:1705.08415.   1. **已有的工作积累和研究成果**   曾在导师的指导下与师兄研究过深度学习相关方法，并，熟悉tensorflow等机器学习相关框架、熟悉python语言,熟悉机器学习相关知识。对本文所述相关算法的研究有一定的基础。参与过一些小型网站的开发，对本文所述复杂网络社区发现系统的设计与实现在软件编程方面有一定的把握。 |

1. 学位论文研究计划及预期目标

|  |
| --- |
| 1.拟采取的主要理论、研究方法、技术路线和实施方案（可续页）  （1）拟采取的主要理论和研究方法  目前，复杂图数据的结构特征和节点特征还不能够充分的利用，同时，在获得图潜在特征时可能会包含较多的噪声，影响社区发现的精确度和稳定性。此外，部分深度学习模型只考虑单独的节点自身，而忽略了节点邻居信息的影响，按照以往的研究结果来看，相似的节点往往在图中的距离更加接近。本课题基于以上问题，拟设计一种拉普拉斯滤波器，首先对图中的特征进行过滤，过滤掉图中的高频噪声，使获得的图数据特征更加平滑。获取的平滑特征将作为图注意力网络编码器GAT的输入，使图中的节点在获得嵌入向量时不仅考虑自身特征还会考虑邻居节点的特征表示，使图的结构信息和节点特征信息更好地融合。最后在通过图注意力网络编码器编码后的向量表示上使用谱聚类算法，进行社区分类。  研究方法主要为阅读相关论文，研究相关算法。通过tensorflow应用、并搜集相关数据集进行实验，调整、优化、修改算法或代码、研究并设法实现一定的准确率或效率。  （2）技术路线  基于以上，本课题拟采用图卷积网络GCN或图注意力网络GAT对，以提出或改良并实现一定的优势。相关的模型或算法的研究、实现将主要以tensorflow进行实验，编程语言拟采用python。其他可能的优化方法的研究如：滤波器的设计、自监督训练模块的设计等。复杂网络社区发现系统拟以PC桌面软件的形式，或者网站Web APP的形式实现。对于PC桌面软件的形式完成复杂网络社区发现系统的设计与实现拟采用微软.Net Framework或者QT，编程语言拟采用c#或c++。对于网站Web APP的实现，拟采用前后端的形式实现。前端网页拟采用js、html等完成，后端的实现拟采用Spring boot框架,编程语言拟采用java。   1. 实施方案   在本课题研究前期，主要针对上述的相关算法的研究：拟首先将主要广泛阅读国内外相关文献与方法，尽可能实现一些优秀的算法或方法。再次通过大量的实验提出或改进相关的算法或方法以达到一定的准确率或者效率。最后完成复杂网络社区发现系统的设计与实现。系统的设计方面拟采用前后端分离，前端用于用户操作交互，作为系统的输入输出部分即展示层，后端处理业务层逻辑。 |
| 2.研究计划可行性，研究条件落实情况，可能存在的问题及解决办法（可续页）  （1）可行性  以图卷积网络GCN，图注意力网络GAT等在社区发现、链接预测等方面已经有了许多成果和应用，深度学习在社区发现的应用被证明有一定的可行性。  目前，Karate、Cora、Dolphins、Citeseer数据集被广泛用于社区发现算法的研究，许多方法包括使用深度学习方法已经较高的识别精度。这意味着采用深度学习等方法的社区发现算法已经有了很高的可行性。  故该课题基于现有的算法、方法提出或改进有一定的可行性。   1. 研究条件   已有实验条件：windows或linux系统的pc、英特尔I5-7500CPU、Nvidia 3050-6G显卡。   1. 可能存在的问题及解决办法   可能存在的问题：  （1）大量的图结构数据进行实验，需要很强的CPU、GPU性能。现有条件可能不足。  解决办法：先用较小图数据网络在PC上进行测试，达到准确率后，再购买GPU进行测试。   1. 知识储备的不足：虽然有一定的深度学习、机器学习的知识储备，但过于基础，对该课题的研究有一定明显的不足。   解决办法：拟先安排课外时间进行听课、自学、实验以能够进行相关模型、算法的实现与进行实验。 |

|  |  |  |
| --- | --- | --- |
| **3.研究计划及预期成果** | | |
| 研  究  计  划 | 起止年月 | 完成内容 |
| 2023/1-2023/3 | 深入进行文献调研，进一步明确课题的研究内容和方案、完成开题。 |
| 2023/3-2023/6 | 对现有的一些算法、模型进行实验、对比。提出优化改进的方法。 |
| 2023/6-2023/9 | 优化、改进相关的算法、并初步实现达到一定的效果。 |
| 2023/9-2023/12 | 完成整体系统、对前期工作进行优化、总结并撰写学位论文。 |
|  |  |
|  |  |
| 预  期  创  新  点  及  成  果  形  式 | 1. 预期创新点   （1）精心设计一种自适应的拉普拉斯滤波器，可以自适应得调节卷积的层数，获得的图数据特征将更加平滑，具有更好的过滤效果。  （2）精心设计图嵌入学习的编码器，采用图注意力网络，不仅自考虑当前结点自身的特征，还会考虑邻居结点的特征，将网络数据的结点特征和结构特征更好地融合，相对于一些方法、算法有一定的优势。  （3）拟采用自监督学习的方式，监督社区发现的结果，对图嵌入向量学习模块和社区发现模块同时进行调整，提升算法的准确性和稳定性。  （4）拟对非重叠社区发现算法进行改进，设计一种既能发现非重叠社区，又能发现重叠社区的高鲁棒性算法。  （5）形成较完整的系统，以软件的形式呈现。  2. 成果形式  期望最终成果为软件的形式：复杂网络的社区发现功能。期望核心功能为：输入复杂网络数据能够在一定时间内输出检测出来的社区以及对应的社区节点信息，可视化展示社区结构。 | |

1. 开题报告审查意见

|  |  |  |  |
| --- | --- | --- | --- |
| 1.导师对学位论文选题和论文计划可行性意见，是否同意开题：  校内导师（组）签字： 年 月 日  校外导师签字： 年 月 日 | | | |
| **2.开题报告考评组意见** | | | |
| 开题日期 |  | 开题地点 |  |
| 考评专家 |  | | |
| 考评成绩 | 合格 票 基本合格 票 不合格 票 | | |
| 结 论 | □通过 □原则通过 □不通过  **通过：**表决票均为合格  **原则通过：**表决票中有1票为基本合格或不合格，其余为合格和基本合格  **不通过：**表决票中有2票及以上为不合格 | | |
| 考评组对学位论文的选题、研究计划及方案实施的可行性的意见和建议： | | | |
| 考评组签名：  年 月 日 | | | |
| **3.学院意见：** | | | |
| 负责人签名： 年 月 日 | | | |