华东师范大学本科生毕业论文（设计）开题报告

论文题目	GitHub 仓库主题关系网络的构建与应用	数据科学与工程学院		数据科学与大数据技术专业	
		学生姓名	许瑞琪	学号	10205501450

一、选题的背景与意义

作为开源代码库以及版本控制系统，著名的 GitHub 社区对开源代码管理标准化产生了深远的影响，截止至 2021 年的统计数据显示 GitHub 已经拥有超过 7300 万的开发者用户。

随着越来越多的应用程序转移到了云上，Github 已经成为管理软件开发以及发现已有代码的首选方法。GitHub 提供给开发者们一套完整的生态工具，从代码托管、质量控制、自动化文档以及可持续集成与部署，都提供了很好的解决方案。另一个让 GitHub 成为黄金平台的原因是其强大的工具生态系统，GitHub 上有大量的开源项目，这些项目为开发者们供了丰富的工具和资源。

GitHub 一直不断更新优化以使得开发者们能够更高效地开发和发布自己的项目，但随着社区的发展，有效地组织仓库以使用户能够有效地检索和重用心仪的仓库变得越来越困难。在此之前，由于标签可以传达不同的概念，许多平台利用协作标记主题为用户在搜索或导航时提供更快、更准确的结果。基于同样的思考，2017 年 GitHub 推出了允许用户用主题标签注释仓库的功能。

在过去的几年中，关于 GitHub 仓库主题分类领域的开发研究层出不穷，已有大量研究与仓库多标签补齐推荐算法[3]、常见仓库主题排名[5]等方向相关，但关于分析仓库主题、探索这些主题之间关系的研究现如今还鲜少见。如果我能探索 GitHub 多标签仓库主题之间的关系，从中获取更多信息，从而对仓库自动标签分类方向的研究起到推进作用，用户将能够更高效准确地搜索和浏览目标代码仓库，还能够开发其上具有实际应用价值的分析任务。

二、研究的主要内容和预期目标

基于选题的背景与意义，我的研究打算通过 GitHub 上多标签仓库的视角，构建 GitHub 生态中的仓库主题关系网络。多标签仓库即一个仓库里存储的某项目有多个不同主题，同样，两个不同主题可以是同一个仓库的两个标签。我希望通过探索不同主题之间的亲疏关系，进而开发其上具有实际应用价值的分析任务，并将主题关系网络的信息添加到仓库主题标签生成的应用场景中用于对其功能进行辅助优化。

因此我的研究主要内容可以分为两大部分：

1. 构建 GitHub 仓库主题关系网络图

我预期构建 GitHub 仓库主题关系网络图，其中每个节点是一个主题，两个节点连接的边根据融合仓库信息数据设置权重，把标记为与边两端主题都有关的仓库们表示成边的权重，两个主题间共同仓库越多，则两主题对应节点的连边权重将会越大。我预计用多种方法设置边权重，后续通过探索其辅助预测多标签仓库主题的能力好坏比较不同边权重设置的优劣，找出当前应用场景下的最优方法。

2. 构建 GitHub 主题预测模型

接着我还打算构建预测模型，用于对上面提到的主题网络构建合理度评测，同时解释主题关系网络应用于优化仓库主题标签自动生成功能的可行性。模型的预期效果是能够自动补齐仓库标签，如遇到仓库所有者只给项目打了一个标签，则模型将学习项目仓库特征，加上已知标签和别的标签的关联度信息进行特征融合，来自动补齐项目还可能所属的主题。预测模型将对抹去部分标签的测试集进行多标签预测，预测结果与被抹去的正确标签进行对比，以此判断我构建的主题关系网络图提供的主题关联度信息好坏，即主题网络构建是否合理；表现优异的主题关系网络将能够用于给仓库主题标签自动生成功能带来辅助优化。

本项目的预期目标是通过构建 GitHub 仓库主题关系网络图，提供拥有实用价值的主题之间关联度的信息，辅助优化仓库主题标签自动生成的应用场景，从而帮助开发人员更好地找到和发现目标主题高关联度的项目仓库。此外，期望主题关系网络能够被用于优化根据主题给开发者推荐仓库项目的应用场景，对研究 GitHub 上的推荐算法带来帮助。

三、拟采用的研究方法、步骤

事先声明由于我现在处于阅读整理选题相关文献阶段，还没有实际开始代码编写尝试，该部分提到的所有研究方法都还处于初步构思阶段，后续具体将采用何种方法会根据可行性做适当调整。本项目拟采用研究方法及步骤如下：

1. 首先进行数据收集与预处理

项目目标数据将基于 GitHub 生态中的所有仓库获取。已知 GitHub 提供了很多公开的 API 让开发者进行几乎一切的 GitHub 操作，开发者可以通过访问 API 接口进行查询并获取数据。因此数据收集方面，我预计通过 GitHub API 实现对符合搜索条件的各个仓库中感兴趣信息项的获取，然后再用 python 代码对返回的 json 结果作处理，提取出各仓库的 id、仓库名、仓库描述、README、topics、language 等字段对应内容作为初始数据内容。数据预处理方面，由于目标是多标签仓库数据，若某仓库只与某主题相关，则不对两个主题的关联度有贡献，将面临去掉没有或只有一个主题标签的项目数据。此外，有的主题可能出现频数极少，我也将考虑适当删去这部分主题。最后把数据存为 csv 文件方便使用。

2. 构建 GitHub 主题关系网络图

仓库主题关联度的计算基于异质信息网络的构建和降维，构造网络的方式是构建一个主题和仓库的二部图。直接对于如图 1(a) 所示的异构图进行分析是较为复杂的，传统的图算法分析大多基于同质信息网络，需对该网络图进行关系与节点降维，使其变为一个同质信息网络。因此，构建 GitHub 主题关系网络图的思路为：每个节点代表一个主题，任意仓库项目有主题 1、主题 2 两个主题，那么这两个主题之间就必然有一条边，一般情况下主题间项目越多边权重越大；多标签仓库以此类推。不同的边权重设置可能会影响主题关联度信息质量，因此我会设计多种图权重分配法，如将边权赋值为两端主题共同仓库之和、仓库和取对数以及统计仓库数据并加权求和等，如图 1(b) 所示。预期使用可视化工具 CodePen、echarts 或 d3 绘制这些主题关系网络图，该过程还要适当对数据进行处理，将它们转换成适用于作图工具的形式，以呈现复杂的网络关系。

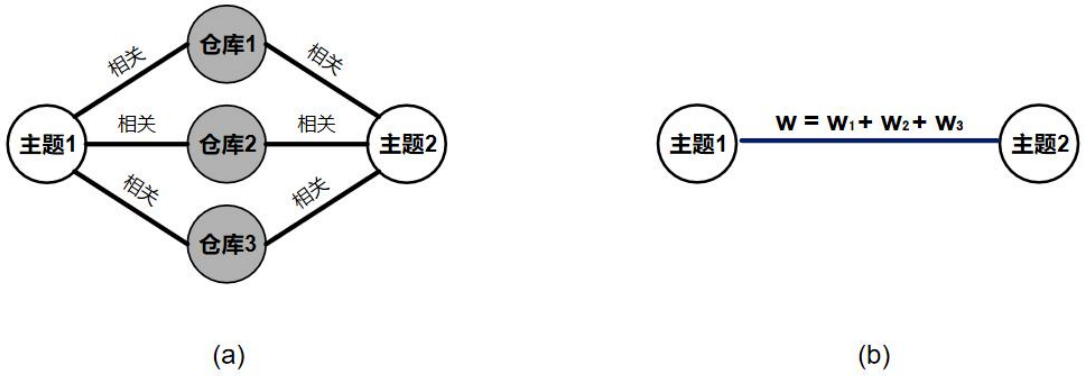

图 1　(a)异质仓库主题网络 (b)经关系与节点类型降维后的同质主题网络
其中，w_1 ，w_2 ，w_3 分别为仓库 1、2、3 对主题 1、2 之间的关联度贡献

3. GitHub 主题关系网络的应用与评价

本文预期主题关系网络的主要应用场景为仓库主题标签自动生成，要实现该功能首先应构建主题预测模型。本文将研究“主题关系网络信息辅助优化仓库主题标签自动生成功能的程度”的方法为控制变量法，即保持其他条件不变，将主题关系网络图得到的多种主题关系网络信息作为多个对照组，分别与数据预处理步骤中得到的仓库基本数据做数据融合，其中文本数据预计使用 bert 处理得到特征向量；为了揭示主题关系网络信息确实对仓库主题标签自动生成产生了优化作用，将进行数据消融，即设置只以仓库基本数据为特征输入的空白对照组，缺少的网络信息部分用零向量补齐，探究模型在其上的表现并与加入主题关系网络信息后的表现作比较。数据特征处理方法如图 2 所示。接着划分训练集、验证集和测试集，用 PyTorch 设置一个简单的神经网络，将已经处理好的训练集特征作为输入进行多标签分类模型训练，每轮训练得到的模型经验证集预测准确率并输出，方便对模型训练过程的观测，最后将训练好的模型对测试集进行预测，取得分高的前几个主题作为预测结果输出。

比较预测结果与真实标签，预计将预测结果的准确率、召回率以及两者的加权 F_β 值作为评价指标对主题网络构建合理度及有效性进行评估，其中 β 表示召回率的重要程度是正确率的 β (>=0)倍， β 将根据数据预处理结果中各仓库标签数视情况而定。F_β 值越高说明主题关系网络提供的信息越好，据此选出使结果最优的边权重分配法得到的主题关系网络为信息源，使其对仓库主题标签自动生成算法起到最佳优化效果。

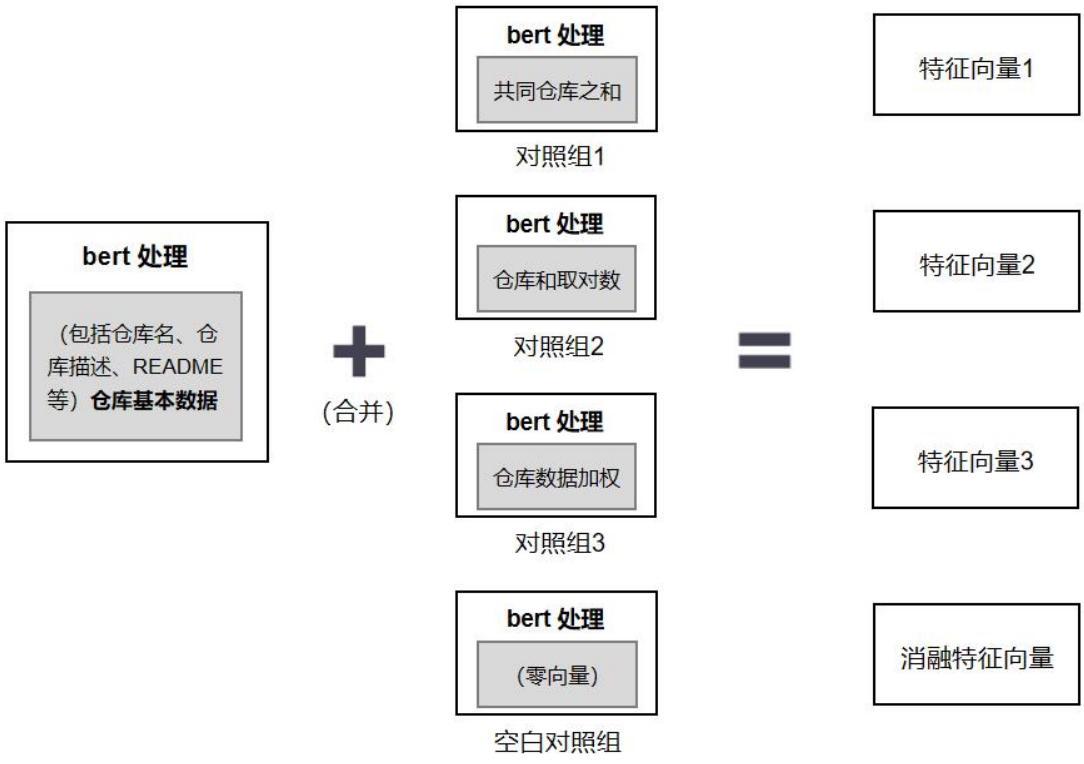

图 2　控制变量下的数据特征处理方法

四、研究的总体安排与进度

我将本人毕业设计分为几个主要步骤，计划分任务按时段完成，预期进度及大致时间安排如下：

1. 毕设选题与材料准备阶段 2023.11-2023.12
2. 文献收集整理与开题报告撰写 2023.12
3. 实现相关数据收集及数据预处理 2023.12-2024.01
4. 实现 GitHub 主题关系网络构建 2024.01-2024.02
5. 利用主题预测模型评测主题网络构建合理度 2024.02-2024.03
6. 完成论文初稿 2024.03

其中，毕设选题与材料准备已完成，现阶段的安排是阅读相关文献并开始尝试收集、处理数据，从而为后续的毕设细节寻找更多灵感与尝试。

五、参考文献

[1] Y. Li, J. Fan, Y. Wang and K. -L. Tan, "Influence Maximization on Social Graphs: A Survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 10, pp. 1852-1872, 1 Oct. 2018, doi: 10.1109/TKDE.2018.2807843.

[2] Braunstein, A., Dall'Asta, L., Semerjian, G., Zdeborová, L.: Network dismantling. Proc. Natl. Acad. Sci. U.S.A. 113(44), 12368–12373 (2016)

[3] Izadi M, Heydarnoori A, Gousios G. Topic recommendation for software repositories using multi-label classification algorithms[J]. Empirical Software Engineering, 2021, 26: 1-33.

[4] Xia X, Zhao S, Han F, et al. OpenDigger: Data Mining and Information Service System for Open Collaboration Digital Ecosystem[J]. arXiv preprint arXiv:2311.15204, 2023.

[5] Sas C, Capiluppi A, Di Sipio C, et al. Gitranking: A ranking of github topics for software classification using active sampling[J]. Software:

Practice and Experience, 2023, 53(10): 1982-2006.

[6] https://github.com/MalihehIzadi/SoftwareTagRecommender

[7] https://github.com/X-lab2017/open-digger?tab=readme-ov-file

[8] https://github.com/X-lab2017/open-research/issues/231

[9] https://github.com/anvaka/map-of-github

论文题目	GitHub 仓库主题关系网络的构建与应用	数据科学与工程学院		数据科学与大数据技术专业	
		学生姓名	许瑞琪	学号	10205501450

六、指导教师意见

该论文选题针对 GitHub 主题关系问题，尝试提出通过主题关系网络优化仓库自动标签分类的解决方法，立意明确，拥有较强的现实需求和应用价值。

该课题是学生所学数据可视化及机器学习知识的延申，符合学生专业发展方向，有益于提高学生研究能力，实验设计切实可行。

该生对相关知识与理论研究透彻，阅读参考了大量文献资料，有一定的前期经验累积，有能力完成论文设计。

建议进一步细化方案，充分调研，特别是应用及网络评价部分。

综上所述，同意开题。

	签字：		年	月	日	

七、开题答辩小组意见

	小组成员签字：		年	月	日	

教务处编制