前后端分离架构下基于图神经网络的社交网络关系挖掘

谢鸿博 [1]
XIE Hongbo

摘 要 在社交网络环境中，因用户数量多，所生成数据规模极为庞大，传统挖掘方法在面对社交网络里复杂的用户关系时，难以实现高效挖掘。鉴于此，文章提出了一种在前后端分离架构下基于图神经网络的社交网络关系挖掘。在前端，部署基于 Selenium 的网页内容采集工具，采集社交网络网页数据，将数据传输至后端，进行缺失值填充和去噪等预处理后，构建一个图神经网络模型，输入预处理后的社交网络网页数据，预测输出社交网络关系挖掘结果。实验结果表明，在使用这一设计方法下，社交网络关系挖掘结果的 G 值为 0.93，这一结果充分验证了该方法在社交网络关系挖掘方面的可行性以及相较于传统方法的优越性。

关键词 前后端分离架构；图神经网络；社交网络；社交关系；关系挖掘

doi：10.3969/j.issn.1672-9528.2024.12.023

0 引言

随着互联网技术的飞速发展，社交网络已成为日常生活中不可或缺的一部分，它不仅改变了人们的沟通方式，还深刻影响着信息传播、社会舆论形成乃至个体行为模式。社交网络平台上积累的海量用户数据，包括用户基本信息、交互行为、内容发布等，为深入理解社会关系结构、个体行为规律以及信息传播机制提供了前所未有的机遇。因此，社交网络数据挖掘近年来受到学术界和工业界的广泛关注。文献 [1] 中通过构建决策树，进行社交网络用户行为数据的挖掘，可以解决传统数据挖掘方法中关联相似性计算困难的问题，但是决策数据过于复杂，实际应用中容易在未见过的数据上产生过拟合风险；文献 [2] 中构建了一个重叠节点挖掘模型，可以实现准确性更高的社交网络中社区数据挖掘，但是这种方法不可避免会涉及到用户个人信息和隐私数据，实际应用中面临隐私保护的局限。基于此，尽管传统社交网络数据挖掘方法取得了诸多成果，但仍存在一些局限，所以本文在前后端分离架构下基于图神经网络进行社交网络关系挖掘，以期为个性化推荐、舆情分析等应用提供新的思路和方法。

1 前端网页采集社交网络数据

本文创新性地在前后端分离架构下进行社交网络关系挖掘 [3]，前后端分离架构就是将前端用户界面与后端服务逻辑进行明确划分，通过 RESTful API 等接口进行数据交换，其优势在于降低系统耦合度并加快开发迭代速度，以此应对社交网络中海量数据的挖掘挑战。首先，为了精准捕获社交网络平台的复杂数据结构，本文构建了一种基于 Selenium 的网页内容采集工具，并将其以轻量化形式部署在社交网络前端页面上，进行数据采集。具体来说，通过 Selenium 提供的 WebDriver API 模拟用户操作，以触发 JavaScript 渲染的网页内容加载，确保数据的完整性与准确性，表 1 列举了基于 Selenium 的网页内容采集工具所支持的用户行为类别。

表 1 社交网络用户行为模拟类别

行为类别	Selenium 支持方式
浏览	Selenium 可以模拟点击链接或输入 URL，打开并浏览特定内容的详情页面
搜索	Selenium 可以模拟在搜索框中输入关键词，并执行搜索操作
点赞 / 取消点赞	Selenium 可以模拟点击点赞按钮，执行点赞或取消点赞操作
评论 / 回复评论	Selenium 可以模拟在评论框中输入文本，并点击提交按钮
分享	Selenium 可以模拟点击分享按钮，并选择分享的目标用户

在用户浏览行为模拟过程中，利用 Selenium 提供的 DOM 操作方法，定位并提取网页中的关键元素信息，如用户 ID、昵称、关注关系、互动数据等。同时，在爬取过程中，为了保留网页的原始状态以供后续分析，基于 Selenium 的网页内容采集工具进行了网页内容的离线存档 [4]，即通过捕获并保存网页的 HTML、CSS、JavaScript 等资源文件，构建一

1. 北京化工大学 北京 100029

个完整的网页快照。总之，本文在前后端分离架构下，基于 Selenium 的网页内容采集工具，实现了对社交网络网页数据的高效且准确爬取，为后续挖掘提供可靠数据基础。

2 后端预处理社交网络数据

在前端成功采集到社交网络网页数据并将其安全传输至后端之后，为进一步构建出适合图神经网络学习和训练的高质量数据集，需要针对原始社交网络数据中普遍存在的缺失值问题与噪声干扰，进行一系列预处理 [5]。首先，社交网络数据由于用户行为的多样性与随机性，伴随着大量的缺失值，这些缺失值需及时处理，否则将直接影响图神经网络训练的效果。为有效解决这一问题，本文采用了基于统计学的填充方法，以社交网络用户属性缺失为例，本文采用估算同类用户均值进行填充的方法，即假设同类用户在某一属性上呈现出相似的分布特征，如地理位置或者社交圈层等特征相似，那么可以计算出该类用户在该属性上的非缺失值的均值填充在缺失值位置，计算公式为：

$$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{N_1-N_2} \tag{1}$$

式中：$\overline{X}$表示社交网络用户数据非缺失值的均值；X_i 表示第 i 个社交网络用户数据非缺失值；n 表示社交网络用户数据非缺失值的数量；N_1、N_2 分别表示社交网络用户数据总数量和缺失值数量。通过上述方法可以在保留社交网络用户数据的统计特性与分布规律的基础上，填补缺失值，为后续图神经网络的学习与训练提供更为完整的数据集 [6]。除了缺失值外，社交网络数据中还可能包含噪声，噪声通常表现为孤立点或异常值，可能由错误的数据录入、数据损坏或极端用户行为等原因造成的，本文采用 K-means 聚类算法来识别并移除这些孤立点。即已知社交网络数据中孤立的噪声点往往远离大多数的数据点，形成单独的簇，所以本文通过迭代地将原始数据点分配到 K 个簇中，使得每个簇内的点尽可能相似，而不同簇间的点尽可能不同，在簇划分时，本文以最小化簇内平方和为目标，具体函数表达式为：

$$f=\sum_{i=1}^{K}\sum_{X\in H}\left\|X-\gamma_i\right\| \tag{2}$$

式中：f 表示 K-means 聚类目标函数；X 表示社交网络数据点；γ_i 表示第 i 个簇的质心；H 表示簇的集合；$\|\cdot\|$ 表示欧几里得距离。通过迭代调整簇的划分与质心的位置，直至目标函数收敛，即可得到最终的聚类结果。通过 K-means 聚类后，设定一个合理阈值，将低于该阈值的簇视为噪声簇，进而将其中的数据点视为孤立点并移除。综上所述，本文在后端对社交网络数据进行了缺失值填充与去噪处理 [7]，为后续社交网络关系挖掘提供更可靠数据基础。

3 后端构建图神经网络模型挖掘社交网络关系

在得到高质量的社交网络数据后，本文在后端构建了图神经网络模型，进行社交网络关系的挖掘 [8]。通过构建的图神经网络模型采用多层结构，集成了输入层、图卷积层、会话表示层及预测层，旨在从预处理后的社交网络网页数据中提取关键特征，进而预测用户间的关联模式。输入层作为整个模型的数据入口，负责接收经过预处理的社交网络数据，当输入层接收到社交网络用户及其对应关系的数据后，会将其映射转化为 One-Hot 向量，以供图卷积层的使用。在图卷积层中，主要负责将输入层提供的 One-Hot 向量映射为用户 - 关系类别图，节点代表用户或实体，边则根据用户间的关系类别建立，形成多类型边的图结构 [9]。为有效表示节点的特征，图卷积层采用消息传播机制与图神经网络算法，通过聚合其邻居节点的信息来更新节点自身的嵌入向量，其公式为：

$$\boldsymbol{x}_j^{k+1}=\varphi\left(\sum_{g\in M(j)}\frac{1}{\lambda_{jg}}\omega^k\boldsymbol{x}_j^k+p^k\right) \tag{3}$$

式中：$\boldsymbol{x}_j^k$、$\boldsymbol{x}_j^{k+1}$分别表示节点 j 在第 k 层和第 k+1 层的嵌入向量；φ 表示非线性激活函数；λ_{jg} 表示节点 j 与节点 g 之间边的归一化常数；$M(j)$ 表示节点 j 的邻居节点集合；ω^k、p^k 分别表示第 k 层的权值与偏置。然后，会话表示层接收图卷积层输出的用户和关系类别嵌入向量 [10]，并利用注意力机制计算表示会话的嵌入向量，表达式为：

$$\boldsymbol{x}_u=\sum_{v\in V_u}\sigma_v x_v^0 \tag{4}$$

式中：$\boldsymbol{x}_u$ 表示会话 u 的嵌入向量；$\boldsymbol{x}_v^0$表示图卷积层输出的最终嵌入向量 [11]；σ_v 表示节点 v 的注意力权重；V_u 表示会话 u 中涉及的节点集合。最后，预测层作为模型的输出端，负责根据会话表示层输出的会话嵌入向量 $\boldsymbol{x}_u$ 以及其他用户的最终嵌入向量 $\boldsymbol{x}_e$，预测出下一个会话中用户关系的概率，表达式为：

$$P\left(e\middle|u\right)=\frac{\exp\left(\mathrm{sim}\left(x_u,x_e\right)\right)}{\sum_{v\in U}\exp\left(\mathrm{sim}\left(x_u,x_v\right)\right)} \tag{5}$$

式中：$P(e|u)$ 表示用户关系预测概率，取概率最大的用户关系类别为用户特征最终预测结果；sim 表示相似度计算函数；exp 表示 softmax 函数，将用户嵌入向量之间相似度分数转换为概率分布；U 表示社交网络用户集合。因此，本文向上述图神经网络模型中输入预处理后的社交网络网页数据，即可预测输出社交网络关系挖掘结果。

4 实验分析

4.1 实验准备

为验证前后端分离架构下基于图神经网络的社交网络关系挖掘方法的有效性和正确性，本文在 Intel Core i7 处理器、

16 GB RAM、NVIDIA GeForce GTX 1080 Ti GPU 的高性能计算机上，以设计方法为实验组，并以基于决策树的社交网络关系挖掘方法和基于重叠节点挖掘模型的社交网络关系挖掘方法为对照组，展开社交网络关系挖掘的对比实验。实验数据来源于全国常用社交网络新浪微博，随机抓取 1000 名用户的相关数据，包括用户 ID、所在地、关注数、粉丝数、微博发布数、转发数、评论数及点赞数等。部分数据示例如表 2 所示。

表 2 实验社交网络用户基本信息数据（部分）

用户 ID	所在地	关注数	粉丝数	微博发布数	转发数	评论数	点赞数
U0001	重庆	12	986	5	93	9397	8130
U0002	浙江	301	31	94	64	33	8456
U0003	北京	80	252	245	104	68	6007
U0004	上海	46	192	227	3	869	429
U0005	江苏	24	687	423	86	544	257
U0006	南京	235	99	305	33	502	1177
U0007	河南	300	102	283	92	1080	1683
…	…	…	…	…	…	…	…
U1000	辽宁	19	6058	44	43	6692	397

在此基础上，为避免实验结果的偶然性，将上述 1000 个新浪微博用户相关数据划分为 10 组，分别用实验组方法和对照组中两种方法进行各组社交网络关系挖掘，对比不同用户规模下的挖掘结果。

4.2 结果分析

完成实验组方法和对照组中两种方法下的社交网络关系挖掘后，本文采用了结合召回率与精确度的 G 值作为评价指标，具体计算公式为：

$$G=\sqrt{P\times R}=\sqrt{\left(\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}}\right)\times\left(\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}\right)} \tag{6}$$

式中：G 表示社交网络关系挖掘结果的 G 值，反映了精确度 P 和召回率 R 的平衡情况，其值越大挖掘结果越佳；T_{P} 表示正例样本被预测为正的数量；F_{P} 表示负例样本被预测为正的数量；F_{N} 表示正例样本被预测为负的数量。

因此，根据式（6）分别计算出实验组方法和对照组中两种方法下的社交网络关系挖掘结果的 G 值，如图 1 所示。从图 1 可以看出，在不同用户规模下，实验组方法均表现出优于对照组中两种方法的社交网络关系挖掘性能。具体来说，在本文设计方法下，社交网络关系挖掘结果的 G 值平均为 0.93，较对照组中两种方法分别提升了 0.08、0.10。因此，前后端分离架构下基于图神经网络挖掘社交网络关系是可行且优越的，设计方法能够有效、精准地预测出社交网络中用户间的关系。

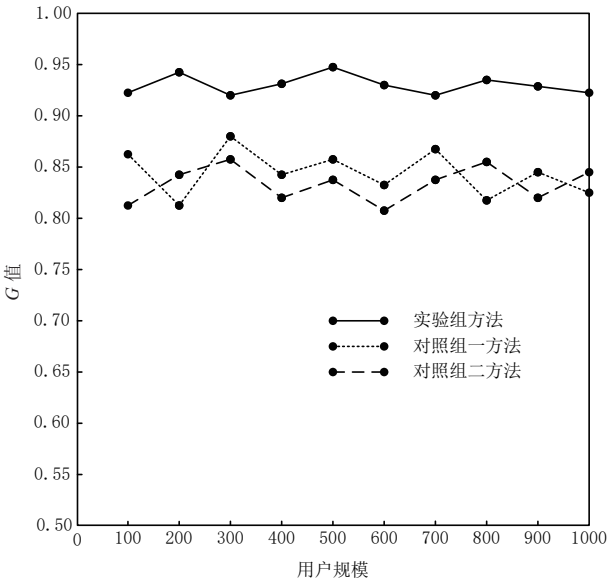

图 1 社交网络关系挖掘结果对比

5 结语

本文成功在前后端分离架构下实现了社交网络关系的挖掘，文中通过前端网页内容采集技术获取社交网络数据，并在后端进行预处理与图神经网络建模，有效挖掘了用户间的潜在关系。实验结果表明，该方法在提升挖掘精度方面展现出显著优势。然而，社交网络关系的复杂性及动态性仍对研究提出挑战。未来，将进一步探索图神经网络在动态社交网络分析中的应用，优化模型架构以提高其泛化能力，并深入研究社交网络中的多模态数据融合技术，以期在更广阔的领域内推动社交网络关系挖掘的发展。

参考文献：

[1] 韩永印，王侠，王志晓．基于决策树的社交网络隐式用户行为数据挖掘方法 [J]. 沈阳工业大学学报，2024, 46(3): 312-317.

[2] 魏会廷，陈永光．面向社交网络重要信息传播的重叠节点挖掘模型研究 [J]. 西南大学学报（自然科学版），2024, 46(2): 150-158.

[3] 钱忠胜，杨家秀，李端明，等．结合用户长短期兴趣与事件影响力的事件推荐策略 [J]. 计算机研究与发展，2022, 59(12): 2803-2815.

[4] 范伟，刘勇．基于时空 Transformer 的社交网络信息传播预测 [J]. 计算机研究与发展，2022,59(8):1757-1769.

[5] 汪子航，言鹏韦，蒋卓人．基于可解释图神经网络模型的社交媒体谣言识别研究 [J]. 情报学报，2023,42(11):1369-1381.

[6] 马锐垚，王鑫，李树，等．基于神经网络的粒子输运问题高

基于上下文信息的知识追踪方法

何萌红 [1]
HE Menghong

摘 要 知识追踪（KT）旨在通过分析学生的学习过程及其知识掌握情况来预测学生未来的学习表现。随着科技的进步，知识追踪技术不断成熟。然而，传统的知识追踪模型通常忽视了问题之间潜在的关联，并且在处理长序列数据时，模型难以有效捕捉长期依赖关系。为解决这些问题，文章提出了一种基于上下文信息的知识追踪模型。该模型通过引入注意力机制计算问题之间的相关性，确保当前问题获得更多相关信息，并有效捕捉问题的长序列信息。这种方法减少了信息在传递过程中的丢失，通过实验表明，与现有的模型相比本模型具有更好的预测能力。

关键词 深度学习；知识追踪；教育系统；上下文信息；注意力机制

doi：10.3969/j.issn.1672-9528.2024.12.024

0 引言

随着社会的快速发展，人工智能在生活中的应用越来越广泛，教育领域也因人工智能技术的快速演进而向智能化教育转型 [1]，在智能辅导系统中，通过计算机辅助技术帮助学生进行个性化学习，使用知识追踪来跟踪和掌握学生的知识状态。一般来说，知识追踪通过学生的历史学习行为进行定量分析，评估学生对知识点掌握的动态演进并以此预测学生能否回答新问题。根据此来跟踪学生在练习过程中的知识获取变化，使得学生可以及时了解自己的知识点掌握情况，并对知识薄弱部分进行针对性的练习。在此背景下，知识追踪（knowledge tracing，KT）技术应运而生。其核心在于通过分析学生的历史答题记录，对其知识状态进行建模，并预测其答对下一题的概率，从而揭示学习过程中学生知识掌握的内在规律。这一技术极大地推动了智能辅助教育系统的发展 [2]。

现有的知识追踪模型大致可分为三类：基于概率图模型的知识追踪、基于矩阵分解的知识追踪，以及基于深度学习的知识追踪。Corbett 等人 [3] 于 1995 年提出的贝叶斯知识追踪（Bayesian Knowledge Tracing）是一种典型的基于概率图模型的知识追踪方法，它假设学生的知识状态是一个隐含的二值变量，通过观察学生的学习行为（例如做题记录）来更新该知识状态的后验概率分布。而基于矩阵分解的知识追踪是通过将学生的学习记录转化为一个矩阵，利用矩阵分解等方法来挖掘学生的知识状态和学习行为之间的潜在关联，以实现知识追踪。然而，这些模型仍存在局限性，无法有效建模知识点之间的相关性，难以捕捉学生学习过程中的复杂行为。随着技术的进步，Piech

1. 贵州民族大学 贵州贵阳 550025

[基金项目] 贵州省教育厅（黔教技〔2022〕047 号），贵州省高等学校智慧教育工程研究中心；贵州民族大学自然科学研究基金（GZMUZK〔2021〕YB22）

效计算方法 [J]. 物理学报 ,2024,73(7):114-123.
[7] 张佳宁 , 沈慧 , 周广东 . 结构监测无线传感网络数据传输优化方法 [J]. 哈尔滨工程大学学报 ,2024,45(8):1543-1551.
[8] 邵云飞 , 宋友 , 王宝会 . 基于社交网络图节点度的神经网络个性化传播算法研究 [J]. 计算机科学 ,2023, 50(4):16-21.
[9] 杨慎 , 陈磊 , 周绮凤 . 基于图增强和图神经网络的层次社区发现方法 [J]. 厦门大学学报 (自然科学版), 2024, 63(2): 209-220.
[10] 仝卫国 , 曾世超 , 张立峰 . 多尺度卷积神经网络的电阻层析成像算法 [J]. 计算机应用与软件 , 2024,41(5):177-182.
[11] 孙乾 , 蒋楠 . 基于卷积神经网络的高效量子态层析方法 [J]. 北京师范大学学报 (自然科学版), 2024,60(3):325-330.

【作者简介】

谢鸿博（2004—），男，河北秦皇岛人，本科，研究方向：计算机科学与技术。

（收稿日期：2024-08-09）