非完备模态下的可靠多媒体推荐方法

檀彦超[1, 2, 3] 沈春旭[4] 陈佳敏[1, 2, 3] 马国芳[5, 6] 林政鸿[1, 2, 3] 王石平[1, 2, 3] 易玲玲[4]

摘 要 随着多模态内容的快速增长, 多媒体推荐系统在数据挖掘中发挥着重要作用. 然而, 现有方法通常假设项目具有完备的多模态信息, 难以适应真实场景中的模态缺失问题. 针对这一挑战, 提出一种融合稀疏超图与模态特定二分图的非完备多媒体推荐框架 (S2GRec). 该框架通过基于稀疏超图的自适应模态补全机制, 捕获模态内高阶相似性, 实现无监督的缺失模态补全, 并进一步利用模态特定二分图建模用户在不同模态视角下的偏好, 以提升推荐性能. 在多个公开数据集及大规模工业数据集上的实验结果表明, S2GRec 在召回率、准确率和 NDCG 等指标上较现有方法平均提升 4.42%, 验证了其在非完备多媒体推荐任务中的有效性.

关键词 推荐系统; 超图生成; 稀疏优化; 图卷积网络; 非完备多媒体推荐

引用格式 檀彦超, 沈春旭, 陈佳敏, 马国芳, 林政鸿, 王石平, 易玲玲. 非完备模态下的可靠多媒体推荐方法. 自动化学报, 2026, **52**(4): 805−820

DOI 10.16383/j.aas.c240659 **CSTR** 32138.14.j.aas.c240659

Reliable Multimedia Recommendation Method With Incomplete Modality Data

TAN Yan-Chao[1, 2, 3] SHEN Chun-Xu[4] CHEN Jia-Min[1, 2, 3] MA Guo-Fang[5, 6]
LIN Zheng-Hong[1, 2, 3] WANG Shi-Ping[1, 2, 3] YI Ling-Ling[4]

Abstract With the rapid growth of multi-modal content, multimedia recommendation systems play an important role in data mining. However, existing methods typically assume that items possess complete multi-modal information, making it difficult to adapt to the issue of missing modalities in real-world scenarios. To address this challenge, this paper proposes a novel framework named S2GRec (sparse hypergraph and modality-specific bipartite graphs for incomplete multimedia recommendation). The framework captures high-order intra-modal similarities via an adaptive modality completion mechanism based on sparse hypergraphs to achieve unsupervised missing modality completion. Furthermore, it utilizes modality-specific bipartite graphs to model user preferences from different modal perspectives, thereby enhancing recommendation performance. Experimental results on multiple public datasets and large-scale industrial datasets demonstrate that S2GRec achieves an average improvement of 4.42% over state-of-the-art methods in terms of Recall, Precision, and NDCG, validating its effectiveness in incomplete multimedia recommendation tasks.

Keywords recommendation systems; hypergraph generation; sparse optimization; graph convolutional network; incomplete multimedia recommendation

Citation Tan Yan-Chao, Shen Chun-Xu, Chen Jia-Min, Ma Guo-Fang, Lin Zheng-Hong, Wang Shi-Ping, Yi Ling-Ling. Reliable multimedia recommendation method with incomplete modality data. *Acta Automatica Sinica*, 2026, **52**(4): 805−820

近年来, 随着抖音、快手等多媒体共享平台用户数量和内容的爆发式增长, 如何帮助用户快速精确地找到他们感兴趣的内容变得比以往更具有挑战

收稿日期 2024-09-28 录用日期 2025-12-24
Manuscript received September 28, 2024; accepted December 24, 2025
国家自然科学基金 (62302098), 福建省人工智能产业发展技术项目 (2025H0042), 福建省自然科学基金 (2025J01540), 浙江省自然科学基金 (LQ23F020007), 浙江省“三农九方”科技协作项目 (2024SNJF044), 浙江省属高校基本科研业务费专项 (FR25008Q) 资助
Supported by National Natural Science Foundation of China (62302098), Fujian Provincial Artificial Intelligence Industry Development Technology Project (2025H0042), Fujian Provincial Natural Science Foundation (2025J01540), Zhejiang Provincial Natural Science Foundation (LQ23F020007), Zhejiang Provincial Department of Agriculture and Rural Affairs Project (2024SNJF044), and Fundamental Research Funds for the Provincial Universities of Zhejiang (FR25008Q)
本文责任编委 朱鹏飞
Recommended by Associate Editor ZHU Peng-Fei
1. 福州大学计算机与大数据学院 福州 350000 2. 大数据智能教育部工程研究中心 福州 350000 3. 福建省网络计算与智能信息处理重点实验室 (福州大学) 福州 350000 4. 腾讯科技有限公司 深圳 518057 5. 浙江工商大学计算机科学与技术学院 杭州 310000 6. 全省大数据与未来电子商务技术重点实验室 杭州 310000
1. College of Computer and Data Science, Fuzhou University, Fuzhou 350000 2. Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350000 3. Fujian Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), Fuzhou 350000 4. Tencent Technology Co., Ltd., Shenzhen 518057 5. School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310000 6. Provincial Key Laboratory of Big Data and Future E-Commerce Technology, Hangzhou 310000

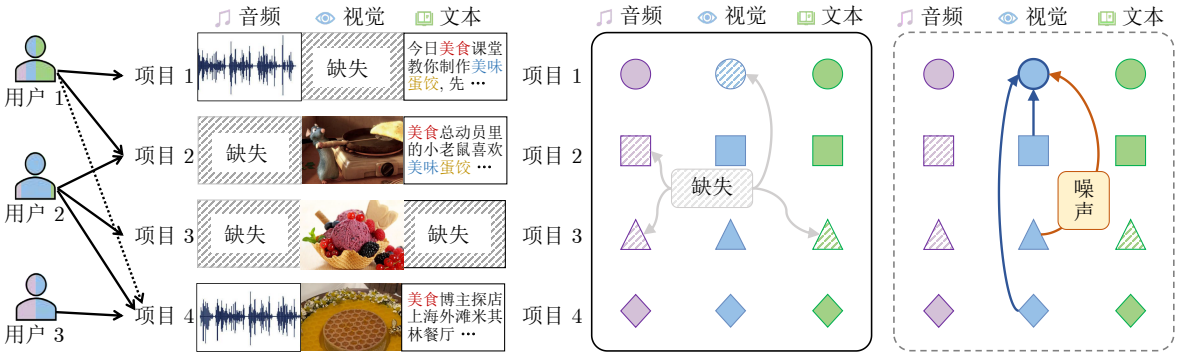
性[1]. 针对这一信息过载问题, 多媒体推荐系统[2−3]应运而生, 该系统通过分析用户在多媒体平台上的历史交互信息 (如点击、收藏等) 和项目的多模态特征 (如图片、文本、音频等), 向用户推荐符合个人偏好的多媒体项目, 提升用户在多媒体平台上的综合体验感[4]. 传统的多媒体推荐算法通过多模态数据来增强有关用户和项目的特征, 从而体现用户在多模态特征下的偏好. 例如, 一个用户可能会因为喜欢项目的视觉图案或者赞同其他用户的文本评价而购买某件衣服; 另一个用户可能因为非常喜欢最近爆火的某段背景音乐, 而给使用这段背景音乐的视频点赞. 随着图神经网络 (graph neural network, GNN)[5−6] 的崛起, 许多研究利用 GNN 将多模态特征集成到用户−项目二分图中[7−8], 通过消息传递模式的各种方法来细化用户和项目的多模态表征.

尽管现有的多媒体推荐系统研究已取得显著成效, 但它们通常基于一个假设: 每个项目的多模态数据都是完整的, 且模态始终可用[9−10]. 在实际情况中, 由于人为因素、设备故障等各种原因, 多媒体平台上的很多项目都存在模态缺失的现象[11]. 如图 1(a) 所示, 项目 1 是用户分享的美食课堂, 只包含音频描述和对应的文本信息; 项目 2 是用户录制的《美食总动员》电影, 但是由于视频录制时背景噪声过大, 声音信息无法有效提取. 这些项目的多模态特征缺失增加了多媒体推荐系统的复杂性. 面对这种非完备模态的现象, 传统多媒体推荐研究主要采取剔除模态缺失的项目以对齐数据训练的策略[12]. 然而, 这种数据预处理方法会导致部分项目间模态的关联性信息遗失[13]. 例如, 在图 1(a) 中, 某个用户 1 喜欢视觉模态缺失的美食语音课堂项目 1, 和音频模态缺失的美食寻味项目 2, 那么他可能也会喜欢项目 4 (一个美食博主探店的项目). 在本就面临数据稀疏性问题的多媒体推荐场景中, 移除缺失模态的项目会加剧数据稀疏性, 并且对多媒体推荐系统的性能产生负面影响.

为了在模态非完备的情况下充分挖掘多模态数据的潜力, 缓解模态缺失所引发的问题并提升多媒体推荐系统性能, 本文面临两大主要挑战:

1) 在无额外监督的条件下, 如何实现可靠的缺失模态补全. 在缺乏明确项目属性之间复杂关系图结构以及直接监督信号指导的情况下, 自适应地挖掘和利用项目的不同模态视角下的关联来补全高置信度的模态信息是具有挑战性的. 传统的模态补全算法采用基于插补[14]、表示[15] 和相似性[16] 的方法来学习不同模态下的关联矩阵, 利用完备项目的特定模态来补全非完备项目模态. 然而, 在多媒体推荐领域, 如果没有额外的监督, 学习到的模态关联矩阵是密集的. 在这样的情况下, 补全的项目模态信息可能来源于低置信度的节点连接, 导致学习的项目模态并不总是准确. 例如, 在图 1(b) 中, 项目 3 视觉模态对项目 1 视觉模态的补全是不可靠的, 因为项目 3 是一个包含多个缺失模态的低置信度节点. 也就是说, 低置信度的密集模态关联结构可能会引入噪声和误差, 在影响模态补全的同时, 也影响多媒体推荐的性能. 因此, 如何在没有额外监督的情况下获得高置信度的缺失模态补全, 是非完备多媒体推荐面临的首要挑战.

2) 如何根据补全的完整项目模态, 自适应地学习用户在不同模态视角下的偏好以提升推荐效果. 在多媒体推荐中, 除了交互信息之外, 推荐模型还可以利用项目的多模态特征来建模用户对项目的不同维度的偏好. 传统的算法主要将这些多模态项目特征融合到用户−项目知识图中[17−18], 平等地对待所有模态特征, 忽略了用户在不同模态视角下的偏好



(a) 多媒体推荐中的模态缺失

(a) Modality missing in multimedia recommendation

(b) 低置信度的密集模态关联补全

(b) Dense modality association completion with low confidence

图 1 非完备多媒体推荐示意图

Fig. 1 Illustration of incomplete multimedia recommendation

影响, 这可能会导致用户在特定模态视角下的表示不太准确. 如图 1(a) 中, 用户 1 可能更偏好于从项目的文本描述中查找兴趣点, 对视觉和音频关注度较低; 用户 2 并不在乎文本和音频, 只关心视觉体验. 同质化或统一多模态信息不足以识别不同模态的重要性, 可能会导致次优表示. 因此, 如何基于补全后的完备项目模态, 自适应地捕获细粒度模态级别的用户偏好, 优化多媒体推荐性能, 是非完备多媒体推荐的又一个挑战.

为了应对上述挑战, 实现非完备模态下的可靠多媒体推荐, 本文提出一个新的框架, 名为融合稀疏超图与模态特定二分图的非完备多媒体推荐框架 (sparse hypergraph and modality-specific bipartite graphs for incomplete multimedia recommendation, S2GRec). 针对第一个挑战, 本文设计一种基于稀疏超图的自适应模态补全模块. 该模块受到超图神经网络 (hypergraph neural network, HGNN)[19] 算法的启发, 将具有相同语义信息的项目聚集为一个组 (超边), 获取模态内高阶相似性, 寻找与缺失模态相似的最优邻居节点集合. 然而, 没有额外监督的情况下, 仅用 HGNN 构建的超图密集拓扑结构可能会包含噪声和误差. 因此, 在构建项目超图时融合稀疏最优传输机制, 通过稀疏约束实施特征选择, 重点关注那些更有信息量的特征. 这种减少输入变量的做法可以帮助模型在关键特征上形成更高的置信度, 模型训练集中于对结果影响最大的因素, 从而确保所学超图结构的可靠性. 最后, 基于稀疏最优传输的超图, 精确利用相似项目进行缺失模态补全, 获得高置信度的完备模态知识.

针对第二个挑战, 本文设计基于模态特定二分图的多媒体推荐模块. 具体地, 该模块基于补全后的完备多模态特征, 融合项目和用户的 ID 表征, 构造模态特定二分图, 将特定模态感知协作学习注入自监督学习任务中, 自适应地捕获细粒度模态级别的用户偏好. 此外, 为了捕捉不同模态特定的用户偏好之间的隐式相互交织, 该模块还引入跨模态对比学习算法, 通过跨模态依赖建模来增强自监督学习范式. 最终, 实现完备模态知识优化的多媒体推荐, 在识别多模式用户偏好时, 共同捕获特定模式的协作效果和跨模式交互依赖性.

本文的主要贡献包括 3 个方面:

1) 针对多媒体推荐中的模态非完备性问题, 创新性地将稀疏约束引入超图结构中. 通过去除不相关的模态信息, 避免关系建模因过于密集而失去可靠性, 进而在没有额外监督的情况下获得高置信度的完备模态知识.

2) 提出融合稀疏超图与模态特定二分图的非完备多媒体推荐框架 (S2GRec). 该框架包含可靠缺失模态补全和自适应细粒度模态级别的用户偏好建模. 在没有额外监督的情况下, 利用对比学习捕获跨模态交互依赖性, 提升非完备模态缺失场景中的多媒体推荐性能.

3) 在四个真实公开数据集 (Amazon-Sports、Amazon-Baby、Allrecipes 和 TikTok 数据集) 上进行的大量实验结果表明, S2GRec 框架相较于其他主流多媒体推荐模型有更优的推荐表现.

1 相关工作

本节从协同过滤、图神经网络和非完备多媒体推荐模型三个方面回顾了相关研究领域的国内外研究现状.

1.1 协同过滤

协同过滤作为推荐系统领域的核心技术之一, 近年来得到了广泛的研究和应用. 其基本思想是通过分析用户与项目之间的交互数据 (如评分、点击或购买记录), 挖掘用户偏好或项目相似性, 从而实现个性化推荐. Goldberg 等[20] 提出 Tapestry 模型, 旨在通过协作过滤的方式帮助用户从大量的电子文档流中筛选出感兴趣的信息. 它不仅支持基于内容的过滤, 还允许用户通过记录和分享对文档的评价 (即注释) 来相互协助进行信息过滤. Resnick 等[21] 提出了基于用户协同过滤的 GroupLens 模型, 首次将协同过滤应用于电影推荐. 该方法通过计算用户之间的相似性, 为目标用户推荐与其兴趣相似的其他用户喜欢的物品.

近年来随着数据规模的增长, 传统协同过滤方法在稀疏数据和扩展性上面临挑战, 矩阵分解技术逐渐成为协同过滤的重要发展方向. Paterek 等[22] 提出的改进矩阵分解方法通过优化用户−项目评分矩阵的分解方式, 提升了推荐精度. 之后, Koren 等[23] 进一步提出奇异值分解模型, 结合隐式反馈 (如浏览记录) 和时间动态因素, 一定程度上优化了推荐效果. 随着深度学习技术的不断发展, 协同过滤开始融入神经网络模型. He 等[24] 提出的神经协同过滤 (neural collaborative filtering, NCF) 通过多层感知机建模用户−项目交互, 取代了传统的内积相似性计算, 显著提升了推荐性能. 此外, Wang 等[25] 提出了基于图神经网络的协同过滤方法, 通过建模用户−项目交互图, 进一步捕捉了复杂的高阶关系.

尽管协同过滤技术取得了显著进展, 但在处理冷启动问题、数据稀疏性以及可解释性不足等挑战时, 其性能仍有待提升.

1.2 图神经网络

图神经网络在图结构的信息传递和聚合方面得到了广泛应用[26]. Li 等[27] 将半监督图神经网络纳入元路径的自动选择中. Luo 等[28] 利用图神经网络捕捉谱域中节点之间的相关性和相异性, 从而建立项目之间互补关系并提高推荐性能. 然而传统的图神经网络无法建模高阶相互作用, 为此, 研究者们提出了超图神经网络[29] 来构建更深层次的图结构关系, 增强了图神经网络的协同过滤范式区分能力. 之后, Wei 等[30] 提出了动态超图协同过滤模型, 采用一个可微的轻量级多层超图学习器, 在统一的框架中学习动态超图结构以及用户和物品的表示, 以对用户不断变化的偏好进行建模. Yu 等[31] 提出了生成式超图框架来处理超图关联矩阵的稀疏性. Cai 等[32] 设计了超图结构学习算法, 基于原始数据集优化超图矩阵的两阶段消息传递方案. Guo 等[33] 提出了四元数超图网络, 基于片段级对象之间的关系构建边或超边.

但是, 现有的超图神经网络研究并没有探索模态非完备性限制下的多媒体推荐问题. 在没有额外监督的情况下, 多媒体推荐中基于超图神经网络构建学习到的补全模态可能并不能准确地表示节点之间的连接.

1.3 非完备多媒体推荐模型

多媒体推荐模型通过分析用户在多媒体平台上的历史交互信息和项目的多模态特征来向用户推荐符合个人偏好的多媒体项目[34]. 学习用户和项目的信息表示是多媒体推荐系统的中心主题. He 等[35] 和 Chen 等[36] 将多媒体内容 (例如视觉特征) 和项目的 ID 嵌入集成在传统的协同过滤框架中. Lin 等[37] 提出的邻域对比学习模型考虑了用户对物品的偏好, 将潜在的邻居纳入对比当中.

此外, 为了提升模型在信息不一致情况下的决策可靠性, 近期研究开始引入不确定性建模机制. Xu 等[38] 引入证据深度学习框架, 在融合多模态信息时显式建模模态间的冲突, 并输出预测结果的不确定性, 从而实现对决策可靠性的量化. Han 等[39] 提出可信多视图分类框架, 利用狄利克雷分布建模模态不确定性, 并通过证据理论进行动态融合, 以提升决策鲁棒性.

然而, 这些模型假设每个项目均涵盖完备的模态信息且所有模态随时可用, 并没有考虑模态非完备条件下建模的特殊性. 在实际情况中, 由于各种不可抗力, 多媒体推荐平台上的很多项目都存在模态缺失现象, 因此非完备多媒体推荐模型被提出. 例如, Zhang 等[40] 系统性地总结和分析了在多模态数据存在质量问题 (尤其是不完备模态) 的情况下, 如何进行有效的多模态学习与融合. Wang 等[41] 通过引入模态随机丢弃算法和多模态序列自动编码器来学习多模态表示以重构缺失模态数据, 填补了推荐系统中的数据空缺. Malitesta 等[42] 将用户−商品交互图转化为基于共同交互的商品−商品图, 然后利用商品之间的多模态相似性, 采用改进的特征传播方法来填补缺失的多模态特征. Liu 等[43] 将不完整多视图聚类中复杂的实例缺失问题简化为图结构的补全问题, 并通过自引导机制和统一的优化框架有效地挖掘多视图数据的一致性和互补信息.

然而, 上述方法并没有考虑补全模态的置信度问题, 过于密集的补全模态可能会包含不相关的噪声节点, 失去高阶关系的明确性, 降低多媒体推荐的有效性.

2 模型方法

在本节中, 将详细介绍适用于非完备多模态数据的可靠多媒体推荐框架 S2GRec. 具体而言, 第 2.1 节给出符号定义和待解决问题的数学定义, 并概述 S2GRec 框架的整体结构. 第 2.2 节介绍单模态无监督稀疏超图构建模块的设计与实现. 此模块旨在处理缺失信息的模态, 利用稀疏最优传输构建高置信度的稀疏超图, 挖掘模态内部的高阶相关性, 识别并链接最相近的模态节点, 从而补足缺失的模态信息. 第 2.3 节详细介绍基于模态特定二分图的多模态协同推荐模块. 此模块利用补全的模态特征及模态间的互补信息, 捕获特定模式的协作效果和跨模式交互依赖性, 实现协同过滤和精细化用户特征的挖掘. S2GRec 的总体框架如图 2 所示.

2.1 问题定义及模型概览

本文所提出的 S2GRec 框架, 其目标是为了解决: 如何基于有缺失模态的多模态数据, 执行可靠的多媒体推荐问题. 具体来说, 给定包含 N_U 个用户数量的用户集合 $\boldsymbol{U}$ 和包含 N_V 个项目数量的项目集合 $\boldsymbol{V}$, 可以使用交互矩阵 $\boldsymbol{R} \in \mathbf{R}^{N_U \times N_V}$ 来建模用户和项目之间的购买关系. 其中, 如果第 i 个用户与第 j 个项目存在历史交互, 则设置 $\boldsymbol{R}_{ij}$ 为 1, 否则设置 $\boldsymbol{R}_{ij}$ 为 0. 此外, 还定义了第 m 个模态下的项目多模态特征 $\overline{\boldsymbol{F}}^{V,\,m} \in \mathbf{R}^{N_V \times d}$, 其中 $\overline{\boldsymbol{f}}_i^{V,\,m} \in \mathbf{R}^{1 \times d}$ 代表 $\overline{\boldsymbol{F}}^{V,\,m}$ 中的第 i 行的原始特征, d 为低维隐空间中的维度数量. 本文将缺失模态处理为值等于全零的 $\boldsymbol{f}_i^{V,\,m}$ 向量, 并且使用 $\boldsymbol{F}^{U,\,m} \in \mathbf{R}^{N_U \times d}$ 来代表在第 m 个模态下的用户表征. 为了补全缺失模态,

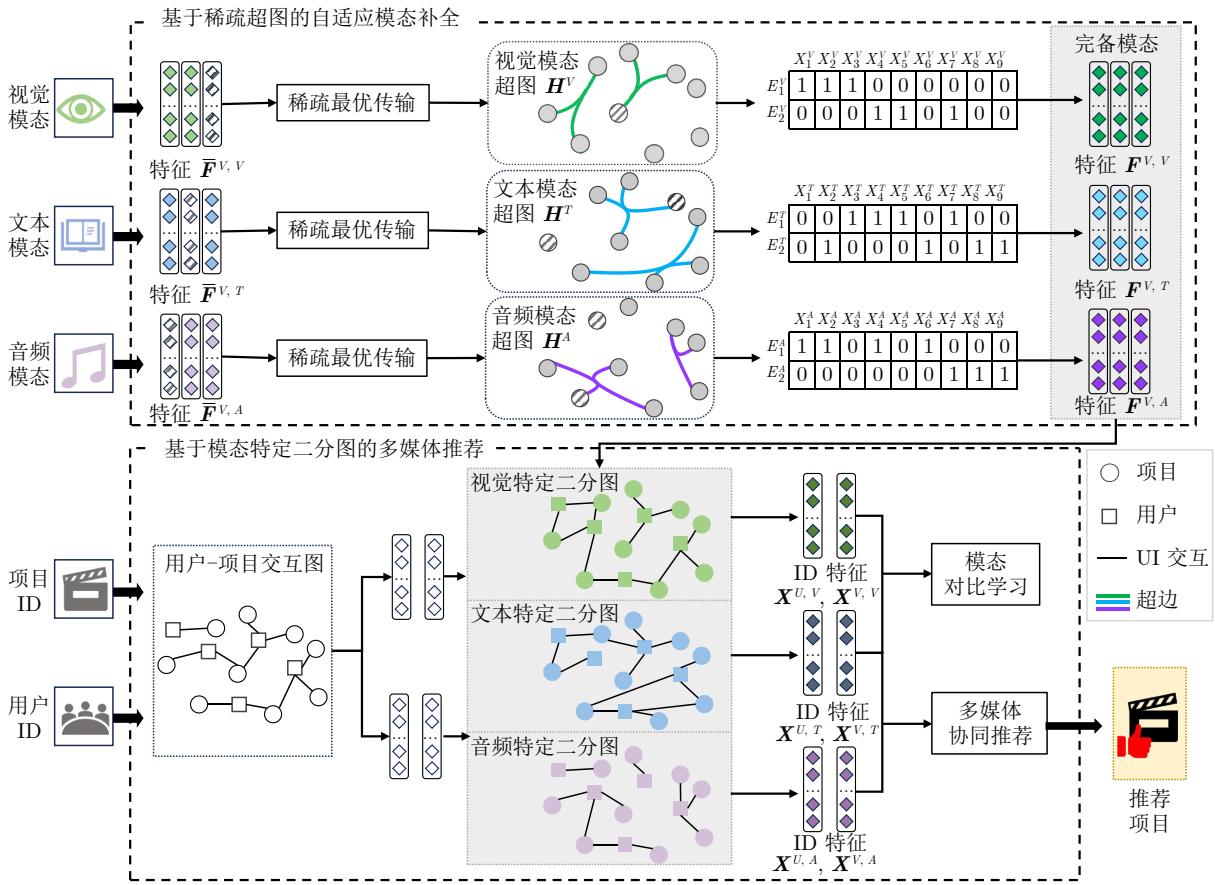


图 2　所提 S2GRec 的整体框架

Fig. 2　The overall framework of proposed S2GRec

使用稀疏超图 $\boldsymbol{H}^m \in \mathbf{R}^{N_V \times K}$ 来建模项目之间的高阶相关性, 其中 K 表示超边个数, 并且使用 $\boldsymbol{F}^{V,m} \in \mathbf{R}^{N_V \times d}$ 来表示经过稀疏超图补全后的完备的多模态项目特征. 为了基于完备的多模态表征执行个性化的多媒体推荐, 采用协同过滤的思想, 使用符号 $\hat{\boldsymbol{\Psi}}^U \in \mathbf{R}^{N_U \times d}$ 和 $\hat{\boldsymbol{\Psi}}^V \in \mathbf{R}^{N_V \times d}$ 来分别表示用户及项目的多模态协同表征, 并使用 $\hat{\boldsymbol{R}}_{ij}$ 评分矩阵来表示将第 j 个项目推荐给第 i 个用户的概率. 基本符号定义的详细信息如表 1 所示. 因此, 对于本文所提出的多媒体推荐框架 S2GRec, 整个模型的输入为用户的项目交互矩阵 $\boldsymbol{R}$, 以及非完备的多模态项目特征 $\overline{\boldsymbol{F}}^{V,m}$, 整个模型的输出为概率矩阵 $\hat{\boldsymbol{R}}_{ij}$.

表 1　基本符号定义表

Table 1　Basic notation definition table

符号	定义
$\boldsymbol{R}$	用户和项目间的购买关系描述矩阵
$\overline{\boldsymbol{F}}^{V,m}$	m 模态下项目的原始特征
$\boldsymbol{H}^m$	m 模态下项目的稀疏超图结构
$\boldsymbol{F}^{V,m}$	m 模态下补全后的项目的多模态特征
$\boldsymbol{E}^V$、$\boldsymbol{E}^T$、$\boldsymbol{E}^A$	视觉、文本、音频模态下的项目超边嵌入
$\hat{\boldsymbol{\Psi}}^U$、$\hat{\boldsymbol{\Psi}}^V$	用户或项目的多模态融合表征

本文 S2GRec 框架概览描述如下: 框架包括基于稀疏超图的自适应模态补全和基于模态特定二分图的多媒体推荐两个主要模块. 1) 在基于稀疏超图的自适应模态补全模块, 将不同模态的项目嵌入通过稀疏最优传输得到稀疏超图, 再将稀疏超图通过超图卷积得到可靠的完备模态表征, 最后将补全得到的完备模态特征输入多模态推荐模块; 2) 在基于模态特定二分图的多媒体推荐模块, 基于输入的完备多模态特征, 以及具有多模态协同信息的融合 ID 表征, 构造模态特定的用户–项目交互图, 最后通过两两对比学习模块以及协同推荐模块, 捕捉用户的个性化偏好进行高效的多媒体推荐.

2.2　基于稀疏超图的自适应模态补全

为了在没有额外监督的情况下实现可靠的缺失模态补全, 提出了基于稀疏超图的自适应模态补全模块, 通过寻找与缺失模态相似且可靠的最优邻居节点集合, 以补全缺失节点的模态信息. 具体而言,

该模块分为 3 个步骤: 首先, 基于 $\ell_{2,1}$-范数的最优超图构建, 通过稀疏约束实施特征选择, 构建稀疏最优超图, 在关键特征上形成更高的置信度, 从而确保所学超图结构的可靠性; 然后, 基于 Frank-Wolfe 算法[44] 的优化过程确保了所提出模型在全局上是可微分的; 最后, 自适应模态内的特征补全, 精确利用相似项目进行缺失模态补全, 获得高置信度的完备模态知识.

2.2.1 基于 $\ell_{2,1}$-范数的最优超图构建

在没有额外监督的情况下, 超图神经网络的密集拓扑结构可能会包含噪声和误差. 因此, 如何设计一种有效的学习机制来确保其结构的可靠性成了关键问题. 受到稀疏正则项在降噪方面的启发, 可通过限制超边内所含超节点数量 (即行稀疏) 来确保所学超图结构的可靠性, 避免超图过于密集而失去高阶关系的明确性. 传统的高阶稀疏约束方法 (例如 K-means), 虽能在一定程度上利用表征聚类构建超图结构, 但其在处理行稀疏性方面的能力有限, 且容易导致平凡解, 即聚类结果不均衡导致的大多项目都聚集到同一个类中的情况, 影响超图构建的有效性. 另外, 考虑到超图中的每个超节点 (即项目) 对超边 (即项目集合) 的贡献以及每个超边对超节点的影响是不一致的, 在相同模态下有缺失值的项目比没有缺失值的项目包含有更多的噪声信息. 因此, 含有不完备模态信息的项目应占有相对较低的权重.

本节基于 $\ell_{2,1}$-范数的最优超图构建通过稀疏约束实施特征选择, 模型只关注那些最有信息量的特征. 这种减少输入变量的做法可以帮助模型在这些关键特征上形成更高的置信度, 从而确保所学超图结构的可靠性. 具体地, 以音频模态下的超图结构为例, 其他模态下的超图构建和音频模态类似. 首先设计了一个超边感知的稀疏正则项 $\|\boldsymbol{H}^A\|_{2,1}$, 该技术已在计算机视觉、自然语言处理等多个领域得到应用[45]. 通过限制单模态下超图结构 $\boldsymbol{H}^A$ 的稀疏性, 实现去除超边中不相关的噪声节点、精确利用相似项目进行缺失模态补全的目标. $\ell_{2,1}$-范数的定义如下:

$$\|\boldsymbol{H}^A\|_{2,1}=\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{K}(\boldsymbol{H}_{ij}^A)^2} \tag{1}$$

其中, K 和 N 分别是 $\boldsymbol{H}^A$ 的列数和行数. 根据上述公式, $\ell_{2,1}$ 相当于找到列的 ℓ_2-范数和行的 ℓ_1-范数. 与已有的稀疏性范数研究相比, $\ell_{2,1}$ 可作为旋转不变的 ℓ_1-范数被引入, 并且比基于 ℓ_2-范数的损失函数更具鲁棒性[46].

考虑到最优传输可以计算由分布相似性引入的几何特性[47], 本文把限制超边稀疏的约束嵌入超图学习过程中. 最优传输可以直接在分布的经验估计上进行评估, 无需外部先验知识[48]. 超图生成的过程即两个超节点和超边分布之间的传输, 这里使用地球移动距离 (earth mover′s distances, EMD) 来测量每个超节点到任意超边的距离. 较小的移动距离意味着超节点与相应超边相似, 可以将相似的超节点聚集到一个超边中.

同样, 以音频模态为例, 初始化 $\boldsymbol{E}^A\in\mathbf{R}^{K\times d}$ 作为项目的可学习超边嵌入. 然后, 将节点的原始特征表示 $\overline{\boldsymbol{F}}^{V,A}$ 以及可训练的超边嵌入 $\boldsymbol{E}^A$ 输入最优传输模块, 以生成超图结构 $\boldsymbol{H}^A\in\mathbf{R}^{N_V\times K}$. 因此, 超图结构学习的目标是找到最优的传输方案 $\boldsymbol{H}^A$, 以最小化所有传输工作的总和.

利用最优传输的这些属性, 通过用 $\ell_{2,1}$-范数替换最优传输优化过程中最常使用的熵正则化, 以确保超边的结构稀疏性并减少噪声. 整个超图结构生成过程可以重写为:

$$\begin{aligned}&\min_{\boldsymbol{H}^A\in\Delta} J=<\boldsymbol{H}^A,\boldsymbol{C}^A>+\eta\|\boldsymbol{H}^A\|_{2,1}\\&\text{s.t. }\Delta=\left\{\boldsymbol{H}^A|\boldsymbol{H}^A\mathbf{1}_K=\frac{\mathbf{1}_N}{N_V},(\boldsymbol{H}^A)^{\mathrm{T}}\mathbf{1}_{N_V}=\frac{\mathbf{1}_K}{K}\right\}\end{aligned} \tag{2}$$

其中, η 根据经验常设置为 1, 是控制稀疏性的超参数; $\boldsymbol{H}^A$ 表示可学习的超图矩阵, $\boldsymbol{H}^A$ 的值表示项目嵌入和超边之间联合分布的概率; $\mathbf{1}_K$ 和 $\mathbf{1}_{N_V}$ 是维度为 K 和 N_V 的单位向量, 用于计算关联矩阵 $\boldsymbol{H}^A$ 中的行或列之和; $<\cdot,\cdot>$ 是 Frobenius 点积; 代价矩阵 $\boldsymbol{C}^A$ 代表运输成本. 本文可以通过测量超节点和超边之间的距离来计算 $\boldsymbol{C}^A$, 即 $\boldsymbol{C}^A$ 可以通过超节点嵌入 $\overline{\boldsymbol{F}}^{V,A}$ 和超边嵌入 $\boldsymbol{E}^A$ 得到, 计算方式为:

$$\boldsymbol{C}_{ij}^A=\|\overline{\boldsymbol{F}}_i^{V,A}-\boldsymbol{E}_j^A\|_2^2 \tag{3}$$

2.2.2 基于 Frank-Wolfe 算法的优化

为了满足式 (2) 中的约束, 需要使用具有非常高近似保证的投影梯度下降 (projected gradient descent, PGD) 算法. 但是基于 PGD 的方法迭代成本高且非常耗时[49]. 因此, 本文提出采用 Frank-Wolfe 算法[44] 的思想来实现同一域的最小化. 与基于 PGD 的方法的方向相比, 校准后的 Frank-Wolfe 梯度被视为与原始梯度的负值最一致的方向, 可以在可行区域内向最优解移动. 在整个优化过程中, 首先优化超图结构 $\boldsymbol{H}^A$, 取 $\boldsymbol{H}^A$ 对式 (2) 的导数并引入矩阵 $\boldsymbol{T}^A$ 作为对角项:

$$\nabla J(\boldsymbol{H}^A)=\boldsymbol{C}^A+\boldsymbol{H}^A\boldsymbol{T}^A=$$

$$\boldsymbol{C}^A+\boldsymbol{H}^A\begin{pmatrix}\dfrac{\eta}{\|\boldsymbol{H}_1^A\|_2} & & \\ & \ddots & \\ & & \dfrac{\eta}{\|\boldsymbol{H}_j^A\|_2}\end{pmatrix} \tag{4}$$

$\boldsymbol{T}^A$ 的值可以通过计算 $\boldsymbol{H}^A$ 中列的 ℓ_2 -范数, 即 $||\boldsymbol{H}_j^A||_2$ 得到; 然后根据以下目标优化当前的超图矩阵 $\boldsymbol{H}^A$.

$$\min_{\boldsymbol{H}^A\in\Delta} J_H=<\boldsymbol{H}^A,\ \nabla J(\boldsymbol{H}^A)> \tag{5}$$

其中 Frank-Wolfe 方法的思想是寻找与当前梯度方向夹角最大的迭代点 $\boldsymbol{s}$, $\boldsymbol{s}$ 可以通过计算得到:

$$\begin{cases}\boldsymbol{s}=\arg\min\limits_{\boldsymbol{s}\in\Delta}\boldsymbol{G}\boldsymbol{s}\\ \boldsymbol{G}=\mathrm{vec}(\nabla J(\boldsymbol{H}^A)),\ \boldsymbol{s}=\mathrm{vec}(\boldsymbol{H}^A)\end{cases} \tag{6}$$

其中, $\boldsymbol{s}\in\mathbf{R}^{NK}$ 是优化变量; $\mathrm{vec}(\cdot)$ 表示矩阵矢量化的过程, 使用 DeepEMD[50] 算法思想以可微分的方式学习 $\boldsymbol{s}$ 的最小化解, 根据 KKT 条件将式 (6) 变换为矩阵形式:

$$\min_{\boldsymbol{s}}\boldsymbol{G}\boldsymbol{s}\quad \text{s.t. } \boldsymbol{A}\boldsymbol{s}=\boldsymbol{b},\ \boldsymbol{Q}\boldsymbol{s}\le 0 \tag{7}$$

其中, $\boldsymbol{A}\boldsymbol{s}=\boldsymbol{b}$ 表示等式约束; $\boldsymbol{Q}\boldsymbol{s}\le 0$ 表示不等式约束. 通过拉格朗日问题的拉格朗日原理, 可以得出如下结论:

$$\mathcal{L}_{FW}(\theta,\ \boldsymbol{s},\ \boldsymbol{\mu},\ \boldsymbol{\lambda})=\boldsymbol{G}\boldsymbol{s}+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{s}+\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{s}-\boldsymbol{b}) \tag{8}$$

其中, $\boldsymbol{\mu}$ 是等式约束; $\boldsymbol{\lambda}\ge 0$ 表示不等式约束上的对偶变量; θ 表示模型前序模块的参数.

为了进一步说明本文所提出的基于 Frank-Wolfe 算法的优化过程, 提供了以上优化过程的算法流程, 详情请参考算法 1.

算法 1. Frank-Wolfe 优化算法流程

输入. 不完全项目模态特征 $\overline{\boldsymbol{F}}^{V,\,m}\in\mathbf{R}^{N_V\times d}$, 超边数量 K, 超参数 γ, 隐空间维度 d.

输出. 最终的可微损失函数 $\mathcal{L}_{FW}$.

1) 初始化超边嵌入 $\boldsymbol{E}$ 和统一异构超图结构 $\boldsymbol{H}$;
2) **While** 没有收敛 **do**
3) 通过式 (3) 中的 $\boldsymbol{C}_{ij}^m=||\overline{\boldsymbol{F}}_i^m-\boldsymbol{E}_j^m||_2^2$ 计算不同模态下的成本矩阵 $\boldsymbol{C}^m$;
4) 通过式 (2) 中的 $\min_{\boldsymbol{H}^m\in\Delta}J=<\boldsymbol{H}^m,\boldsymbol{C}^m>+\eta\|\boldsymbol{H}^m\|_{2,1}$ 更新不同模态下的超图结构 $\boldsymbol{H}^m$;
5) 根据式 (4) 求 J 的导数;
6) 利用 Frank-Wolfe 算法, 通过式 (5) 对当前不同模态下的超图矩阵 $\boldsymbol{H}^m$ 进行优化;
7) 根据式 (6), 通过校准的 Frank-Wolfe 方向找到迭代点 $\boldsymbol{s}$;
8) **End while**
9) 根据式 (8) 计算 $\mathcal{L}_{FW}(\theta,\ \boldsymbol{s},\ \boldsymbol{\mu},\ \boldsymbol{\lambda})=\boldsymbol{G}\boldsymbol{s}+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{s}+\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{s}-\boldsymbol{b})$.

根据 KKT 条件, 通过 $g(\theta,\ \boldsymbol{s},\ \boldsymbol{\mu},\ \boldsymbol{\lambda})=0$, 可以计算出损失函数的最优 $(\tilde{\boldsymbol{s}},\ \tilde{\boldsymbol{\mu}},\ \tilde{\boldsymbol{\lambda}})$.

$$g(\theta,\ \boldsymbol{s},\ \boldsymbol{\mu},\ \boldsymbol{\lambda})=\begin{bmatrix}\nabla_\theta\mathcal{L}_{\mathcal{FW}}(\theta,\ \int,\ \boldsymbol{\mu},\ \boldsymbol{\lambda})\\ \mathrm{diag}\{\boldsymbol{\lambda}\}\boldsymbol{Q}(\theta)\boldsymbol{s}\\ \boldsymbol{A}(\theta)\boldsymbol{s}-\boldsymbol{b}(\theta)\end{bmatrix} \tag{9}$$

根据凸优化的可微性[51] 可以计算关于 $\tilde{\boldsymbol{s}}$ 和 θ 的隐式函数:

$$J_\theta\tilde{\boldsymbol{s}}=-J_{\boldsymbol{s}}g^{-1}(\theta,\ \tilde{\boldsymbol{s}},\ \tilde{\boldsymbol{\mu}},\ \tilde{\boldsymbol{\lambda}})J_\theta g(\theta,\ \tilde{\boldsymbol{s}},\ \tilde{\boldsymbol{\mu}},\ \tilde{\boldsymbol{\lambda}}) \tag{10}$$

其中, $J_{\boldsymbol{s}}g$ 表示 g 关于 $\boldsymbol{s}$ 的雅可比矩阵; $J_\theta g$ 表示 g 关于 θ 的雅可比矩阵; $J_\theta\tilde{\boldsymbol{s}}$ 表示 $\tilde{\boldsymbol{s}}$ 关于 θ 的雅可比矩阵, 将隐函数定理应用于 KKT 条件得到雅可比行列式的公式, 并获得了关于参数 θ 的 $\tilde{\boldsymbol{s}}$ 梯度的封闭式, 即利用深度反向传播方法来迭代点 $\tilde{\boldsymbol{s}}$, 无需优化轨迹. 为了方便计算目标函数, 本文将超图矩阵 $\boldsymbol{H}=[\boldsymbol{h}_1;\boldsymbol{h}_2;\ \cdots;\boldsymbol{h}_{N_V}]\in\mathbf{R}^{N_V\times K}$, $\boldsymbol{h}_i=[\boldsymbol{H}_{i1}^A,\ \boldsymbol{H}_{i2}^A,\ \cdots,\ \boldsymbol{H}_{iK}^A]\in\mathbf{R}^{1\times K}$ 展开, 得到 $\boldsymbol{h}=[\boldsymbol{h}_1^{\mathrm{T}};\boldsymbol{h}_2^{\mathrm{T}};\cdots;\boldsymbol{h}_N^{\mathrm{T}}]\in\mathbf{R}^{KN_V\times 1}$ 作为列向量. $\boldsymbol{h}^{(k)}$ 是第 k 次迭代时的嵌入, 视为固定向量, 因此可以通过以下方式获得最终公式:

$$\boldsymbol{h}^{(k+1)}=(1-\gamma)\boldsymbol{h}^{(k)}+\gamma\boldsymbol{s} \tag{11}$$

其中, γ 控制点移动的强度. 其他模态下的超图结构也是根据式 (2) ~ (11) 得到.

2.2.3　自适应模态内特征补全

基于稀疏最优传输理论构建的稀疏最优超图, 需要利用相似项目进行缺失模态补全以获得高置信度的完备模态知识. 使用现有的模态集合来补全缺失的模态集合的前提是这两个集合属于同一集群. 为了建模这种集到集的相关性, 利用聚类质心 $\boldsymbol{E}^m$ 来表示超边嵌入, 并利用聚类概率矩阵 $\boldsymbol{H}^m$ 作为超图关联矩阵, 表示超节点可能属于哪个超边. 然后, 采用超图卷积网络进行集到集的补偿以更新缺失模态的表示:

$$\boldsymbol{F}^{m,\,(l+1)}=\sigma(\boldsymbol{D}^{-1}\boldsymbol{H}^m\boldsymbol{W}\boldsymbol{B}^{-1}\boldsymbol{H}^{m\mathrm{T}}\boldsymbol{F}^{m,\,(l)}) \tag{12}$$

其中, $\boldsymbol{F}^{m,\,(l)}$ 是在对 l 层进行超图卷积后的第 m 个模态的完全嵌入; $\boldsymbol{D}\in\mathbf{R}^{N_V\times N_V}$ 和 $\boldsymbol{B}\in\mathbf{R}^{K\times K}$ 是对角线矩阵, $\boldsymbol{D}$ 是顶点度矩阵, $\boldsymbol{B}$ 是边度矩阵; $\boldsymbol{W}$ 是超图神经网络的可训练权值; $\sigma(\cdot)$ 是引入非线性因素的激活函数; 零阶表示 $\boldsymbol{F}^{m,\,(0)}$, 由项目特征 $\overline{\boldsymbol{F}}^{V,\,m}$ 得到.

2.3 基于模态特定二分图的多模态协同推荐

为了充分利用多模态信息, 并基于补全后的完备多模态特征对用户进行个性化的推荐, 提出一种基于多模态特征的协同推荐机制, 该机制包含 2 个主要步骤: 模态特定二分图结构学习以及基于对比模态学习的多媒体推荐.

2.3.1 模态特定二分图结构学习

为了基于完备模态执行多媒体推荐任务, 需要构造特定模态的用户–项目交互图. 首先, 采用 GNN 来初始化输入的嵌入, 从稀疏的用户–项目交互中获取足够的信息, 并将协作模态效应整合到项目的表示学习中. 假设 $\boldsymbol{F}^{V,\,m}\in\mathbf{R}^{N_V\times d}$ 和 $\boldsymbol{F}^{U,\,m}\in\mathbf{R}^{N_U\times d}$ 分别是在 m 模态以及潜在维数 d 下不带有缺失值的项目嵌入和用户嵌入. $\boldsymbol{F}^{U,\,m}$ 可以被看作是用户对这些项目的偏好. 然后, 可以通过图形消息聚合来获得用户的输入模态表示:

$$\boldsymbol{F}^{U,\,m}=\boldsymbol{S}^{U^{-1}}\boldsymbol{R}\overline{\boldsymbol{F}}^{V,\,m} \tag{13}$$

其中, $\boldsymbol{S}^{U}\in\mathbf{R}^{N_U\times N_U}$ 是根据用户–项目交互矩阵 $\boldsymbol{R}\in\mathbf{R}^{N_U\times N_V}$ 所得到的度对角归一化矩阵; $\overline{\boldsymbol{F}}^{V,\,m}=[\overline{\boldsymbol{f}}_1^m;\cdots;\overline{\boldsymbol{f}}_{N_V}^m]$ 表示第 m 个模态下的原始项目特征.

为将学习到的项目特征融入多媒体推荐框架中, 将构建每个模态下特有的图结构, 其构造的具体过程可以形式化为计算在第 m 个模态下, 用户模态表征 $\boldsymbol{F}^{U,\,m}$ 与项目模态表征 $\boldsymbol{F}^{V,\,m}$ 的最大概率似然. 其中计算过程可以形式化为:

$$\begin{aligned}&pro(\boldsymbol{G}^m|\boldsymbol{F}^{V,\,m},\,\boldsymbol{F}^{U,\,m})=\\&\qquad pro(\boldsymbol{G}^m[i,\,k]=1|\boldsymbol{f}_i^{V,\,m},\,\boldsymbol{f}_k^{U,\,m})\end{aligned} \tag{14}$$

其中, $pro(\cdot)$ 表示基于用户和项目的模态特征, 推断它们之间是否存在超图连接关系的条件概率; $\boldsymbol{f}_i^{V,\,m}\in\boldsymbol{F}^{V,\,m}$ 是第 i 项的表示; $\boldsymbol{G}^m[i,\,k]$ 是将第 i 个项目分配给第 k 个用户的概率, $\boldsymbol{G}^m$ 可以被看作在第 m 个模态下用户对于项目的偏好矩阵. $\boldsymbol{G}^m[i,\,k]$ 可以通过余弦相似度得到, 计算如下:

$$\boldsymbol{G}^m[i,\,k]=\frac{\boldsymbol{f}_i^{V,\,m}\times\boldsymbol{f}_k^{U,\,m}}{||\boldsymbol{f}_i^{V,\,m}||_2\times||\boldsymbol{f}_k^{U,\,m}||_2} \tag{15}$$

为引入协同的多模态推荐信号, 首先初始化用户和项目的 ID 表征 $\boldsymbol{X}^U$ 和 $\boldsymbol{X}^V$, 然后对用户和项目的邻居执行 ID 对应的图信息聚合:

$$\begin{cases}\boldsymbol{X}^{U,\,(l+1)}=\sigma(\boldsymbol{S}^{U^{-1}}\boldsymbol{R}\boldsymbol{X}^{V,\,(l)}\boldsymbol{W}^{U,\,(l)})\\\boldsymbol{X}^{V,\,(l+1)}=\sigma(\boldsymbol{S}^{V^{-1}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{X}^{U,\,(l)}\boldsymbol{W}^{V,\,(l)})\end{cases} \tag{16}$$

其中, $\boldsymbol{X}^{U,\,(l)}\in\mathbf{R}^{N_U\times d}$、$\boldsymbol{X}^{V,\,(l)}\in\mathbf{R}^{N_V\times d}$ 是图神经网络第 l 层中用户和项目的 ID 对应嵌入, 零层嵌入 $\boldsymbol{X}^{U,\,(0)}$ 和 $\boldsymbol{X}^{V,\,(0)}$ 从可训练的查找表中初始化; $\boldsymbol{W}^{U,\,(l)}$ 和 $\boldsymbol{W}^{V,\,(l)}$ 是 GNN 第 l 层的可训练用户和项目权重.

其次, 为了捕获每个模态特有的协同特征并充分利用完备的模态信息, 将 ID 特征输入模态特有的图结构, 并执行图卷积:

$$\begin{cases}\boldsymbol{X}^{U,\,m,\,(l+1)}=\sigma(\boldsymbol{D}^{U^{-1}}\boldsymbol{G}^m\boldsymbol{X}^{V,\,m,\,(l)}\boldsymbol{W}^{U,\,m,\,(l)})\\\boldsymbol{X}^{V,\,m,\,(l+1)}=\sigma(\boldsymbol{D}^{V^{-1}}(\boldsymbol{G}^m)^{\mathrm{T}}\boldsymbol{X}^{U,\,m,\,(l)}\boldsymbol{W}^{V,\,m,\,(l)})\end{cases} \tag{17}$$

其中, $\boldsymbol{X}^{U,\,m,\,(l+1)}$ 表示将初始化得到的 ID 表征送入第 m 个模态的图结构中进行卷积操作; $\boldsymbol{D}^{U^{-1}}$ 和 $\boldsymbol{D}^{V^{-1}}$ 为度对角的归一化矩阵.

2.3.2 基于对比模态学习的多媒体推荐

为了能够充分捕获模态间的互补信息, 使用对比学习来最大化不同模态间的互补信息:

$$\mathcal{L}_s=\sum_{m=1}^{M}\sum_{k=1}^{N_V}-\log\frac{\exp\left(\dfrac{s(\boldsymbol{x}_k^m,\,\hat{\boldsymbol{x}}_k^m)}{\tau}\right)}{\displaystyle\sum_{k'=1}^{N_V}\exp\left(\dfrac{s(\boldsymbol{x}_k^m,\,\hat{\boldsymbol{x}}_{k'}^m)}{\tau}\right)} \tag{18}$$

其中, $s(\cdot)$ 表示相似度度量函数, 本文使用余弦相似度; $\boldsymbol{x}_k^m$ 表示矩阵 $\boldsymbol{X}^m$ 的第 k 行; $\hat{\boldsymbol{x}}_k^m$ 和 $\hat{\boldsymbol{x}}_{k'}^m$ 分别为通过特定模态图卷积操作后的 ID 表征, 表示 $\boldsymbol{X}^m$ 第 k 项嵌入的正负样本对; τ 为温度参数. 这里采用 InfoNCE 对比损失从模态内和模态间视角监督生成项目表征.

为融合多模态表征, 将 ID 嵌入和项目内容结合起来作为完整的多模态嵌入:

$$\begin{cases}\hat{\boldsymbol{\Psi}}^U=Concat(\boldsymbol{X}^U,\,\boldsymbol{X}^{U,1},\,\boldsymbol{X}^{U,2},\,\cdots,\,\boldsymbol{X}^{U,M})\\\hat{\boldsymbol{\Psi}}^V=Concat(\boldsymbol{X}^V,\,\boldsymbol{X}^{V,1},\,\boldsymbol{X}^{V,2},\,\cdots,\,\boldsymbol{X}^{V,M})\end{cases} \tag{19}$$

其中, $\boldsymbol{X}^U\in\mathbf{R}^{N_U\times d}$ 和 $\boldsymbol{X}^{U,\,m}\in\mathbf{R}^{N_U\times d}$ 分别为用户的初始 ID 表征和经过多模态协同过滤的优化表征; 同理, $\boldsymbol{X}^V\in\mathbf{R}^{N_V\times d}$ 和 $\boldsymbol{X}^{V,\,m}\in\mathbf{R}^{N_V\times d}$ 分别为项目的初始 ID 表征和经过多模态协同过滤的优化表征; $Concat(\cdot)$ 是拼接表征操作.

然后, 将全局的 ID 表征送入推荐模块, 得到最终的全局评分矩阵 $\hat{\boldsymbol{R}}\in\mathbf{R}^{N_U\times N_V}$, $\hat{\boldsymbol{R}}$ 中的值 $\hat{\boldsymbol{r}}_{i,\,j}$ 表示向用户 i 推荐项目 j 的概率. 为了增强多媒体推荐, 采用推荐任务中常用的 BPR (Bayesian personalized ranking) 损失函数, 计算方式为:

$$\mathcal{L}_{Rec}=\sum_{(i,\,j_p,\,j_n)}^{D}-\log(\mathrm{Sigmoid}(\hat{\boldsymbol{r}}_{i,\,j_p}-\hat{\boldsymbol{r}}_{i,\,j_n})) \tag{20}$$

其中, i 表示采样锚点; j_p 和 j_n 分别表示采样锚点

所对应的正负样本点. 最后, 采用组合损失训练本文的推荐系统来共同优化 S2GRec 框架:

$$\mathcal{L} = \mathcal{L}_{Rec} + \lambda_1 \mathcal{L}_s + \lambda_2 ||\Theta||^2 \tag{21}$$

其中, $||\Theta||^2$ 表示针对过拟合的权重衰减正则化; λ_1、λ_2 都是超参数.

2.3.3　算法时间复杂度分析

在本节中, 将分析 S2GRec 的算法时间复杂度. 具体来说, S2GRec 的运行时间会被 2 个主要的核心模块所限制: 基于稀疏超图的自适应模态补全模块和基于模态特定二分图的多媒体推荐模块. 首先, 分析第一个基于稀疏超图的自适应模态补全模块. 该模块基于稀疏的范式约束来构建一个超图结构, 根据算法流程需要计算代价矩阵 $\boldsymbol{C}^A$ 和被稀疏正则项所约束的梯度方向 $\nabla J(\boldsymbol{H}^A)$, 其中计算代价矩阵 $\boldsymbol{C}^A$ 的时间复杂度是 $\mathrm{O}(N_V K d)$, 计算梯度导数 $\nabla J(\boldsymbol{H}^A)$ 的时间复杂度是 $\mathrm{O}(N_V K^2 + N_V K)$. 梯度导数 $\nabla J(\boldsymbol{H}^A)$ 由于需要计算稀疏约束的对角矩阵 $\boldsymbol{D}$ 及其导数, 时间复杂度分别为 $\mathrm{O}(N_V K)$ 和 $\mathrm{O}(N_V \times K^2)$. 然后, 再分析基于模态特定二分图的多媒体推荐模块. 该模块主要包含二分图构建和多媒体推荐 2 个步骤, 其中, 模态特定二分图构建的时间复杂度是 $\mathrm{O}(MN_V N_U)$, 多媒体推荐的时间复杂度为 $\mathrm{O}(N_U \times N_V d)$. S2GRec 中对比学习方法是通过 mini-batch 的采样方式实现的, 其算法时间复杂度相对于整体时间复杂度可以忽略不计, 本文所提出的 S2GRec 框架的最终时间复杂度为 $\mathrm{O}(N_V K d + N_V K^2 + N_V \times K + MN_V N_U + N_U N_V d)$. 本文算法的时间复杂度与用户数量 N_U 以及商品数量 N_V 呈线性关系. 与现有的多媒体推荐方法相比, 该时间复杂度达到了可接受的水平.

2.3.4　Frank-Wolfe 算法的收敛性分析

在本节中, 将阐述 Frank-Wolfe 优化算法收敛性保证的理论分析. 本文采用经过校准的 Frank-Wolfe 算法来处理目标函数. 具体来说, 它包括两个阶段. 首先计算 $\boldsymbol{s} = \arg\min_{\boldsymbol{s}\in\Delta} \boldsymbol{G}\boldsymbol{s}$, 然后更新 $\boldsymbol{h}^{(k+1)} = (1-\gamma)\boldsymbol{h}^{(k)} + \gamma \boldsymbol{s}$, 其中 $\boldsymbol{h}$ 是 $\boldsymbol{H}$ 的矩阵向量化形式, Δ 表示约束域, k 表示第 k 步, γ 表示点移动的强度. 定理 1 表明, 上述 Frank-Wolfe 算法的迭代解 $\boldsymbol{h}^{(k)}$ 在经过 $\mathrm{O}(\frac{1}{\mu})$ 次迭代后可达到 μ -近似解, 即满足 $f(\boldsymbol{h}^{(k)}) \le f(\boldsymbol{h}^*) + \mu$, 其中 $\boldsymbol{h}^*$ 是最优解, 该收敛定理已在文献 [52] 中被证明.

定理 1. 对于任意 $k \ge 1$, 上述 Frank-Wolfe 算法的迭代结果 $\boldsymbol{h}^{(k)}$ 满足:

$$f(\boldsymbol{h}^{(k)}) - f(\boldsymbol{h}^*) \le \frac{2C_f}{k+2} \tag{22}$$

其中, C_f 为曲率常数.

3　实验与分析

在四个公共多媒体数据集以及两个大规模工业数据集上评估本文提出的 S2GRec 的有效性, 重点关注以下研究问题:

问题 1. 与其他最先进的多媒体推荐模型相比, S2GRec 在四个多媒体公开数据集中表现如何.

问题 2. S2GRec 中设计的每个组件是否合理有效, 并且研究各组件如何对性能改进做出贡献.

问题 3. 超参数如何影响模型的预测性能以及如何选择最优值.

问题 4. S2GRec 在两个大规模工业数据集上性能表现如何.

3.1　实验设置

3.1.1　数据集

本次实验采用表 2 中四个公开的多模态推荐数据集. 其中, TikTok 是来自抖音短视频平台上的多模态数据, 包含视频的视觉、听觉和文本特征, 文本特征采用 Sentence-Bert 编码. Amazon-Baby 和 Amazon-Sports 数据集均来自亚马逊平台, 产品的图像和文本细节被用于生成 4 096 维的视觉特征嵌入和文本特征嵌入, 文本特征也用 Sentence-Bert 编码. Allrecipes 数据集来自最大的食谱社交网站, 该网站有 27 个不同类别的 52 821 种食谱. 每个食谱的图像被视为视觉特征, 20 种食材被采样为文本特征. 表 2 中的稀疏度是通过将用户和项目的实际交互数除以用户和项目的总交互数 (用户数乘以项目数) 获得.

表 2　包含视觉 (V)、音频 (A) 和文本 (T) 内容的多模态实验数据集统计

Table 2　Statistics of multimodal experimental dataset containing visual (V), audio (A), and textual (T) content

	数据集								
	Amazon-Baby		Amazon-Sports		TikTok			Allrecipes	
模态嵌入	V	T	V	T	V	A	T	V	T
嵌入维度	4 096	1 024	4 096	1 024	128	128	768	2 048	20
用户数量	19 445		35 598		9 319			19 805	
项目数量	7 050		18 357		6 710			10 067	
交互数量	160 792		296 337		59 541			58 922	
稀疏度	99.883%		99.955%		99.904%			99.970%	

3.1.2　基线模型

为了验证本文提出的模型在多媒体推荐任务中

的有效性, 并验证缺失模态信息对模型性能的影响, 选取了以下两类具有代表性的基线模型来进行实验分析.

1) 单模态推荐模型

MF-BPR[4]: 采用一种扩展抽样方法来充分利用多样化的反馈信息, 反映不同类型用户的评论和偏好.

NGCF[25]: 通过在用户−项目图结构上传播嵌入, 能够对高阶连接进行有效的建模, 有效地将协同信号注入嵌入过程.

LightGCN[53]: 通过线性传播用户和项目嵌入在用户−项目交互图上进行学习, 并使用在所有层学习的嵌入的加权和作为最终嵌入.

SGL[54]: 生成一个节点的多个模态图, 最大限度地提高同一节点不同模态间的一致性.

NCL[55]: 包括针对每个组件进行监督学习的模块和针对多个组件之间知识传递的模块.

HCCF[56]: 使用超图增强的跨模态对比学习结构来联合捕获局部和全局协作关系.

2) 多媒体推荐模型

LightGCN-M[53]: 在 LightGCN 模型的基础上融合多模态特征构建 LightGCN-M.

MMGCL[57]: 旨在以自监督学习的方式显著增强多模态表征学习.

SLMRec[58]: 为了捕捉数据本身的多模态特征, 超越了监督学习的范式, 并将自监督学习的思想融入多媒体推荐的表示增强中.

MMSSL[59]: 是一种模态感知结构学习范式, 通过对抗性扰动来增强数据, 以表征用户−项目协同图和项目多模态语义图之间的相互依赖性.

CI2MG[2]: 提出模态内和模态间生成对比框架, 以增强非完备模态下的多媒体推荐.

3.1.3 超参数设置

本文基于 Pytorch 实现了所提出的框架, 并采用 AdamW 和 Adam 作为生成器的优化器. 具体地, 设置学习率为 {4.5e−4, 5.0e−4, 5.4e−3, 5.6e−3}, L_2 正则化项的衰减率在 {1.2e−2, 1.4e−2, 1.6e−2} 范围内进行优化, 图层数为 {1, 2, 3, 4}. 本文提出的模型和其他对比模型的嵌入维数均设置为 64. 超参数 λ_1、λ_2 和 λ_3 在 {1e−3, 1e−2, 1e−1, 1} 中选择, 在对比学习的参数 τ 的取值范围为 [0, 1]. 请注意, 本文中所有给定的超参数都使用了交叉验证的方式, 并且在所有数据集上对超参数进行微调, 选取最优性能值对应的超参数值作为最终结果.

3.1.4 实验环境设置

为了更公平地比较 S2GRec 的性能, 本文按照 7 : 1 : 2 的比例将数据集分成训练集、验证集和测试集. 采用召回率 (Recall@K)、准确率 (Precision@K) 以及 NDCG (NDCG@K) 这三个推荐系统中被广泛采用的指标来衡量模型性能. 此外, 所有的对比算法均采用原始的参数设置并在对应的数据集上进行调优选择最优参数, 并且针对模态缺失的情况, 对所有性能进行 10 次实验, 选取总体实验的均值作为最终性能. 本文所有的实验均在具有 2 张 NVIDIA GeForce RTX 3090 GPU 显卡的服务器环境中完成.

表 3 在 Amazon-Baby、Amazon-Sports、Allrecipes 和 TikTok 多媒体数据集上的实验对比结果 (%)

Table 3 Experimental comparison results on the Amazon-Baby, Amazon-Sports, Allrecipes, and TikTok multimedia datasets (%)

数据集	指标	MF-BPR	NGCF	Light-GCN	SGL	NCL	HCCF	Light-GCN-M	MM-GCL	SLM-Rec	MM-SSL	CI2-MG	S2GRec	Imp.
Amazon-Baby	R@20	4.40	5.90	6.98	6.78	7.00	7.05	5.29	5.70	7.01	7.78	8.51	**8.89**	**4.47**
	P@20	0.24	0.32	0.37	0.36	0.38	0.37	0.28	0.31	0.39	0.41	0.45	**0.47**	**4.44**
	N@20	2.00	2.61	3.19	2.96	3.11	3.08	2.24	2.57	3.21	3.41	3.69	**3.88**	**5.15**
Amazon-Sports	R@20	4.30	6.95	7.82	7.79	7.65	7.79	4.27	6.90	7.87	8.16	8.65	**8.94**	**3.35**
	P@20	0.23	0.37	0.42	0.41	0.40	0.41	0.23	0.37	0.40	0.43	0.46	**0.48**	**4.35**
	N@20	2.02	3.18	3.69	3.61	3.49	3.61	2.48	3.28	3.77	3.81	3.98	**4.23**	**6.28**
Allrecipes	R@20	1.37	1.65	2.12	1.91	2.24	2.25	3.38	3.82	3.28	3.35	4.12	**4.33**	**5.10**
	P@20	0.07	0.08	0.10	0.10	0.10	0.11	0.17	0.19	0.15	0.17	0.21	**0.22**	**4.76**
	N@20	0.50	0.59	0.76	0.69	0.77	0.82	1.34	1.70	1.41	1.51	1.85	**1.91**	**3.24**
TikTok	R@20	3.40	6.04	6.53	6.03	6.58	6.62	6.82	5.84	7.11	7.64	8.03	**8.28**	**3.11**
	P@20	0.17	0.30	0.33	0.30	0.34	0.29	0.34	0.29	0.32	0.38	0.41	**0.43**	**4.88**
	N@20	1.30	2.30	2.82	2.38	2.69	2.67	2.83	2.59	4.25	4.42	4.58	**4.76**	**3.93**

注: 其中 R@20、P@20 和 N@20 分别是评价指标 Recall@20、Precision@20 和 NDCG@20 的缩写; 下划线表示次佳实验结果; 效果提升 (Imp.) 为 S2GRec 相比次佳模型的性能提升.

3.2　总体性能比较 (问题 1)

表 3 显示了 S2GRec 与其他对比模型在四个多媒体数据集上的性能对比. 通过分析实验结果, 得出以下结论:

首先, S2GRec 在所有评估指标和所有数据集上均超过了所有基线算法, 这回答了问题 1, 表明 S2GRec 在多媒体推荐能力上具有显著优势. 与第二好的基线 CI2MG 相比, S2GRec 在 Amazon-Baby、Amazon-Sports、Allrecipes 和 TikTok 四个数据集上的性能提升显著, 从在 TikTok 数据集 Recall@20 上提升 3.11% 到在 Amazon-Sports 数据集 NDCG@20 上提升 6.28%.

其次, 相比无多媒体信息的单模态推荐算法, S2GRec 取得了显著性能提升. 例如, 相比于仅使用历史交互的 LightGCN, S2GRec 在四个数据集以及三个指标上平均提升达到 51.73%. 尽管 HCCF 使用基于超图的协同过滤方法, 能够捕获更高阶的语义信息, 但是其没有考虑多模态信息以及其在高阶语义上的关联, 因而与 HCCF 相比, S2GRec 在四个数据集以及三个指标上平均提升达到约 50.42%. 上述两个观察均说明多模态信息的使用具有重要意义.

此外, 相比成功处理完全多模态数据的模型 (例如 MMSSL), S2GRec 的提升是显著的 (从在 TikTok 数据集 NDCG@20 上提升 7.69% 到在 Allrecipes 数据集 Precision@20 上提升 29.41%). 这表明, 在处理具有模态缺失的实际应用场景时, 模态补全的可靠性十分重要. 尽管 CI2MG 利用同一模态和同一项目中不同模态特征之间的相似性, 对齐异构特征的分布以补全模态, 然而 CI2MG 并没有考虑补全模态的置信度问题, 过于密集的补全模态可能会包含不相关的噪声节点, 失去高阶关系的明确性. 由于 S2GRec 设计了可靠补全模块, 因而能够在四个数据集以及三个指标上相较于 CI2MG 平均提升达到约 4.42%.

最后, 为了全面展示本文的模型在处理模态非完备推荐任务方面的效果, 提供了不同缺失率下的 S2GRec 与完备的多模态推荐算法 (例如, MMGCL 和 MMSSL), 以及非完备的多模态推荐算法 (例如, LRMM, CI2MG, FeatProp) 的对比 (如图 3 所示). 总体而言, S2GRec 在所有数据集和缺失率设置 (10% ~ 90%) 下均取得了最佳 Recall@20. 同时观察到, 在 Amazon-Sports 数据集下, 缺失率为 60% 时, S2GRec 的性能稍高于缺失率为 50% 的情况. 这可能是因为原数据中的模态信息存在噪声, 如图片背景模糊、文本错误等. 此外, 注意到部分方法在不同缺失率下会发生性能突变, 这可能是模拟缺失设置影响和核心节点数据缺失所造成的.

3.3　消融实验 (问题 2)

为了更好地理解 S2GRec 框架的主要设计以及各个组件的作用, 通过逐个添加组件来仔细研究本文的框架. 具体含义如下:

1) 基础模型: 使用推荐中常用的 LightGCN-M[53] 作为基本的基础模型组件;

2) 基础 + 补全: 表示在基础模型的基础上添加了基于稀疏超图的自适应模态补全组件;

3) 基础 + 补全 + 多模态: 表示基于补全后的超图, 添加基于模态特定二分图的多媒体推荐组件.

表 4 展示了消融实验结果, 通过分析可得:

首先, 当本文的模型仅具有基于稀疏超图的自适应模态补全组件时, 即基础 + 补全模型, 该模型在具有 2 个模态的数据集 (例如 Amazon-Baby) 和具有更多模态的数据集 (例如 TikTok) 上的指标都

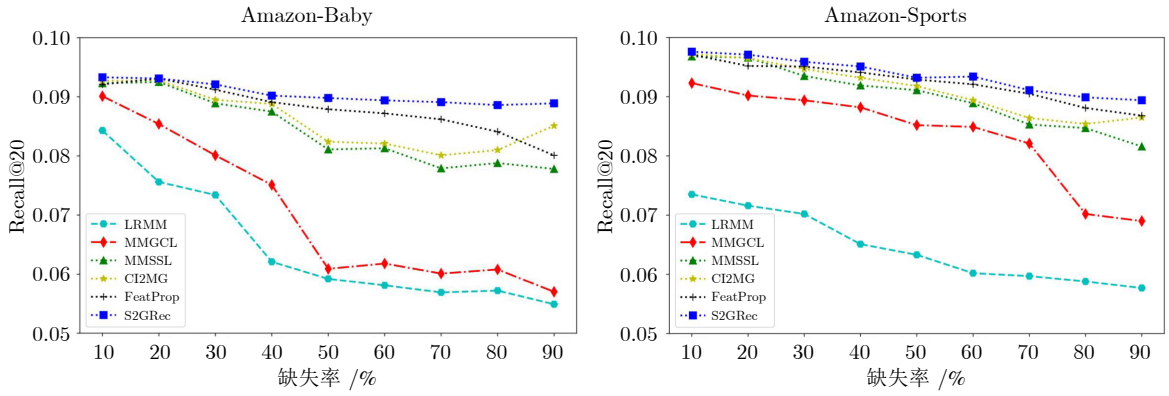


图 3　不同缺失率下的 S2GRec 与其他方法在 Amazon-Baby 和 Amazon-Sports 数据集上的 Recall@20 性能比较

Fig. 3　Recall@20 performance comparison of S2GRec and other methods under different missing rates on Amazon-Baby and Amazon-Sports datasets

表 4 消融实验结果 (%)
Table 4 Ablation experimental results (%)

数据集	指标	基础模型	基础 + 补全	基础 + 补全 + 多模态
Amazon-Baby	R@20	5.29	7.31	**8.89**
	P@20	0.28	0.35	**0.47**
	N@20	2.24	3.09	**3.88**
Amazon-Sports	R@20	4.27	7.26	**8.94**
	P@20	0.23	0.33	**0.48**
	N@20	2.48	3.87	**4.23**
Allrecipes	R@20	3.38	4.11	**4.33**
	P@20	0.17	0.19	**0.22**
	N@20	1.34	1.45	**1.91**
TikTok	R@20	6.82	7.68	**8.28**
	P@20	0.34	0.40	**0.43**
	N@20	2.83	3.96	**4.76**

优于基础模型. 具体来说, 相比于基础模型, 基础 + 补全模型在 Amazon-Baby、Amazon-Sports、Allrecipes 和 TikTok 数据集上的性能提升显著 (从在 Allrecipes 数据集 NDCG@20 上提升 8.21% 到在 Amazon-Sports 数据集 Recall@20 上提升 70.02%), 这验证了基于稀疏超图模态补全组件的有效性.

其次, 与仅具有补全后特征的基础 + 补全模型相比, 添加了模态特定二分图的多媒体推荐组件的模型, 即基础 + 补全 + 多模态模型, 在有效性指标上实现了提升. 具体来说, 基础 + 补全 + 多模态模型在四个数据集上平均提升 20.65% (从在 Allrecipes 数据集 Recall@20 上提升 5.35% 到在 Amazon-Sports 数据集 Precision@20 上提升 45.45%). 结果表明了本文设计的基于模态特定二分图的协同过滤机制的有效性. 基础 + 补全 + 多模态模型在所有消融模型中表现最好, 这也验证了本文提出的 S2GRec 框架中每个组件的合理性和有效性.

3.4 超参数研究 (问题 3)

本文提出的 S2GRec 框架主要引入了四个超参数, 即稀疏超图层数 L_h、模态特定二分图层数 L_b、温度参数 τ 和超边个数 K. 将展示这四个超参数如何影响性能. 从图 4 中可以观察到以下结果:

1) L_h 表示式 (12) 中稀疏超图的层数, 其中最优值为 3. 当层数增多时, 超图卷积神经网络一样会产生过度平滑问题. 因此, 设置 $L_h = 3$ 来缓解过度平滑问题.

2) L_b 表示式 (17) 中模态特定二分图的层数, 发现在不同数据上, L_b 的最优值为 2. 这是因为随着二分图层数的增加, 更多消息的传递和聚合可能加剧数据过平滑问题, 导致模态表征无法具有区分度. 因此, 在挑选最优图层数时, 设置 $L_b = 2$.

3) τ 为式 (18) 中的温度参数, 其最优值为 0.1. 事实上, τ 设为 0.1 同样是符合经验的现象, 并遵从大多数实验设置.

4) K 表示超边个数, 其最优值为 128. 因此, 通过轻微调整可以得到最优的参数.

为了证明在模态非完备情况下, S2GRec 框架基于稀疏超图自适应模态补全的可靠性, 选取了一个研究案例, 对模型构建的超图结构以及补全模态

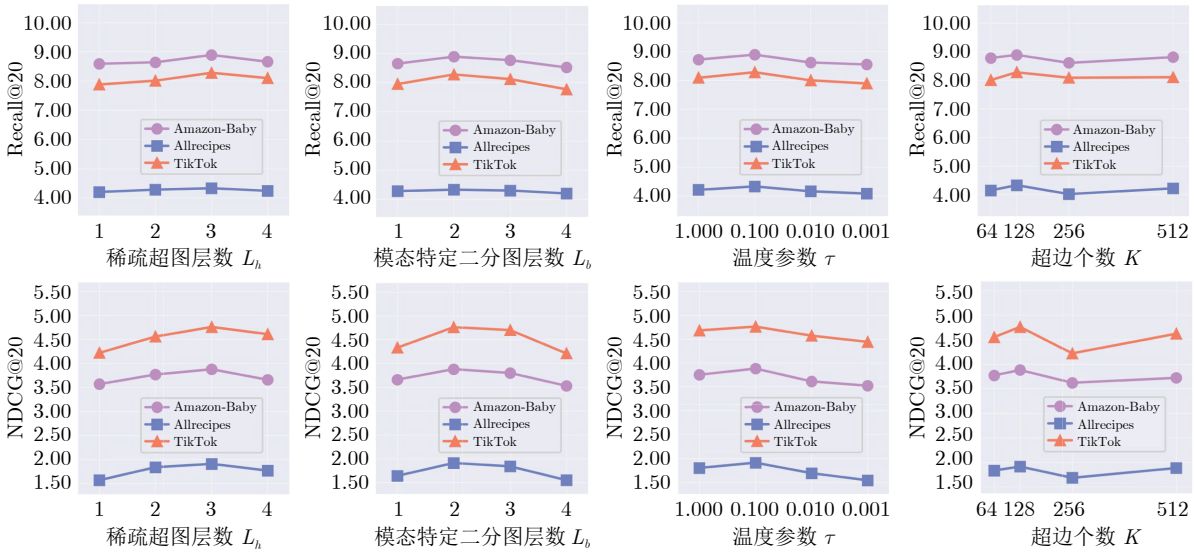
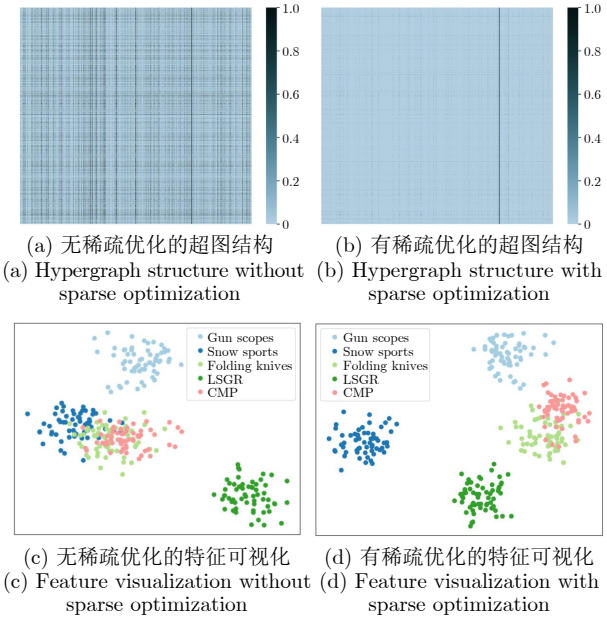


图 4 不同超参数下的 S2GRec 框架在 Amazon-Baby、Allrecipes 和 TikTok 数据集上关于 Recall@20、NDCG@20 的性能
Fig. 4 Performance of the S2GRec framework with different hyperparameters on Amazon-Baby, Allrecipes, and TikTok datasets for Recall@20 and NDCG@20

的特征进行可视化. 图 5(a) 和图 5(b) 分别是基于聚类 (无稀疏正则项) 生成的超图结构与 S2GRec 学习得到的超图结构, 两者对比可以看出, S2GRec 得到的超图结构更稀疏. 图 5(c) 和图 5(d) 分别是基于两种约束状态中学习到项目表征绘制的 T-SNE 图, 使用相同的色谱来表示同一个项目类别, 这些类别由数据集中的原始文本筛选关键字给出. 结果表明, 与无稀疏优化的方法相比, 有稀疏优化的方法可以分离得到更加清晰的节点边界. 因此, 在模态非完备情况下, S2GRec 框架基于稀疏超图得到的补全模态具有更高的可靠性. 同时, S2GRec 框架构造的超图更稀疏, 但最终多媒体推荐的性能更优, 也就意味着本文的模型可以利用更少的连接来捕获更有效的信息.



(a) 无稀疏优化的超图结构 (a) Hypergraph structure without sparse optimization　(b) 有稀疏优化的超图结构 (b) Hypergraph structure with sparse optimization

(c) 无稀疏优化的特征可视化 (c) Feature visualization without sparse optimization　(d) 有稀疏优化的特征可视化 (d) Feature visualization with sparse optimization

图 5　超图的结构及特征可视化

Fig. 5　Visualization of the structure and features of the hypergraph

3.5　工业场景下的应用 (问题 4)

为了验证所提出的非完备多媒体推荐系统在真实复杂场景下的有效性和实用性, 本研究选取了腾讯公司微信直播间平台的两个大规模工业数据集 WeStream-Small 和 WeStream-Large 进行实验评估. 该平台是中国最大的直播平台之一, 数据涵盖了视频流媒体和在线直播推荐等典型多模态场景.

3.5.1　工业数据集描述

大规模工业数据集相较于通用的非完备多媒体数据集具有数据集规模更大、模态信息更丰富且存在可控缺失的特点, 更能反映实际应用挑战. 本研究采用留一法 (leave-one-out) 评估策略. 两个数据集均经过严格的隐私保护处理, 其关键统计信息如表 5 所示.

表 5　工业数据集统计

Table 5　Industrial dataset statistics

数据集	用户数量	项目数量	交互数量
WeStream-Small	649 万	58 万	3.2959 亿
WeStream-Large	2225 万	195 万	12.3531 亿

3.5.2　工业数据集对比实验

为验证本文所提出的 S2GRec 在工业场景下的有效性, 将其与当前性能领先的基线方法 CI2MG 在微信工业数据集上进行了对比实验. 实验遵循工业环境通用的评价标准, 采用 Recall@K 和 NDCG@K (K = 50, 100, 150) 作为评价指标. 其中, 直播场景多模态缺失呈技术级联效应: 图像缺失源于硬件损伤、光照异常或视频流传输层故障; 音频中断可归因于拾音装置失效、环境噪声掩蔽或编码协议故障; 文本缺失源于自动语言识别技术产生的误差、元数据同步中断或跨语言解析层失效. 观察表 6 可以看出 S2GRec 取得了显著的提升. 具体来说, S2GRec 框架在 WeStream-Small 数据集上, 分别在指标 Recall@50 和 NDCG@50 上取得了 4.12% 和 8.33% 的提升. 在规模更大的 WeStream-Large 数据集上, 分别在 Recall@50 和 NDCG@50 上取得了 3.41% 和 10.38% 的提升. 以上实验充分说明了, 本文方法在真实的大规模工业数据集上, 具有更强的实用性和有效性.

表 6　在 WeStream-Small 和 WeStream-Large 数据集上的实验对比结果 (%)

Table 6　Experimental comparison results on WeStream-Small and WeStream-Large datasets (%)

数据集	指标	CI2MG	S2GRec	Imp.
WeStream-Small	R@50	3.40	3.54	4.12
	R@100	5.02	5.19	3.39
	R@150	6.29	6.50	3.34
	N@50	1.80	1.95	8.33
	N@100	2.78	3.01	8.27
	N@150	3.55	3.80	7.04
WeStream-Large	R@50	3.81	3.94	3.41
	R@100	5.63	5.82	3.37
	R@150	7.13	7.33	2.80
	N@50	2.12	2.34	10.38
	N@100	3.26	3.53	8.28
	N@150	4.17	4.45	6.71

4 结论

本文针对现实多媒体平台中的模态缺失场景, 研究非完备模态下的可靠多媒体推荐方法, 提出了一种稀疏超图与模态特定二分图融合的非完备多媒体推荐框架, 即 S2GRec. 该框架由基于稀疏超图的自适应模态补全机制和基于模态特定二分图的推荐模块组成, 能够在缺失场景中泛化推荐性能. 在基于稀疏超图的自适应模态补全机制中, 通过稀疏最优传输构建的超图捕获模态之间的高阶相似性, 在无监督情况下获得可靠的补全模态. 在基于模态特定二分图的推荐方法中, 本文的模型可以充分融合多模态表征下的协同信号, 以执行多媒体推荐任务. 在真实世界数据集上的大量实验表明, 本文提出的 S2GRec 框架能够取得优异的泛化性能.

参考文献

1 Yuan Yong, Zhou Tao, Zhou Ao-Ying, Duan Yong-Chao, Wang Fei-Yue. Blockchain technology: From data intelligence to knowledge automation. *Acta Automatica Sinica*, 2017, **43**(9): 1485−1490
(袁勇, 周涛, 周傲英, 段永朝, 王飞跃. 区块链技术: 从数据智能到知识自动化. 自动化学报, 2017, **43**(9): 1485−1490)

2 Lin Z H, Tan Y C, Zhan Y F, Liu W M, Wang F, Chen C C, et al. Contrastive intra- and inter-modality generation for enhancing incomplete multimedia recommendation. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 6234−6242

3 Bai H Y, Hou M, Wu L, Yang Y H, Zhang K, Hong R C, et al. GoRec: A generative cold-start recommendation framework. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 1004−1012

4 Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal, Canada: AUAI Press, 2009. 452−461

5 Bronstein M M, Bruna J, Cohen T, Veličković P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv: 2104.13478, 2021.

6 Scarselli F, Gori M, Tsoi A C, Hagenbuchner M, Monfardini G. The graph neural network model. *IEEE Transactions on Neural Networks*, 2009, **20**(1): 61−80

7 Tao Z L, Wei Y W, Wang X, He X N, Huang X L, Chua T S. MGAT: Multimodal graph attention network for recommendation. *Information Processing & Management*, 2020, **57**(5): Article No. 102277

8 Wang Q F, Wei Y W, Yin J H, Wu J L, Song X M, Nie L Q. DualGNN: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 2023, **25**: 1074−1084

9 Zhang Ying, Zhang Bing-Bing, Dong Wei, An Feng-Min, Zhang Jian-Xin, Zhang Qiang. Multi-modal video action recognition method based on language-visual contrastive learning. *Acta Automatica Sinica*, 2024, **50**(2): 417−430
(张颖, 张冰冰, 董微, 安峰民, 张建新, 张强. 基于语言−视觉对比学习的多模态视频行为识别方法. 自动化学报, 2024, **50**(2): 417−430)

10 Du X Y, Yuan H H, Zhao P P, Fang J H, Liu G F, Liu Y C, et al. Contrastive enhanced slide filter mixer for sequential recommendation. In: Proceedings of the 39th International Conference on Data Engineering (ICDE). Anaheim, USA: IEEE, 2023. 2673−2685

11 Wang S W, Liu X W, Liu L, Tu W X, Zhu X Z, Liu J Y, et al. Highly-efficient incomplete largescale multiview clustering with consensus bipartite graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 9766−9775

12 Li H X, Zheng C Y, Wu P. StableDR: Stabilized doubly robust learning for recommendation on data missing not at random. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023.

13 Cho J W, Kim D J, Choi J, Jung Y, Kweon I S. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2021. 1592−1601

14 Xia D X, Yang Y, Yang S H, Li T R. Incomplete multi-view clustering via kernelized graph learning. *Information Sciences*, 2023, **625**: 1−19

15 Deng S J, Wen J, Liu C L, Yan K, Xu G H, Xu Y. Projective incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(8): 10539−10551

16 Liu S Y, Zhang J P, Wen Y, Yang X H, Wang S W, Zhang Y, et al. Sample-level cross-view similarity learning for incomplete multi-view clustering. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 14017−14025

17 Hu J W, Liu Y C, Zhao J M, Jin Q. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual Event: Association for Computational Linguistics, 2021. 5666−5675

18 Du W Z, Su H Y, Cam-Tu N, Sun J. Enhancing product representation with multi-form interactions for multimodal conversational recommendation. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 6491−6500

19 Huang J, Lu T, Zhou X B, Cheng B, Hu Z B, Yu W H, et al. HyperDNE: Enhanced hypergraph neural network for dynamic network embedding. *Neurocomputing*, 2023, **527**: 155−166

20 Goldberg D, Nichols D, Oki B M, Terry D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 1992, **35**(12): 61−70

21 Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work. Chapel Hill, USA: ACM, 1994. 175−186

22 Paterek A. Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of the KDD Cup and Workshop. San Jose, USA: ACM, 2007. 39−42

23 Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer*, 2009, **42**(8): 30−37

24 He X N, Liao L Z, Zhang H W, Nie L Q, Hu X, Chua T S. Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: ACM, 2017. 173−182

25 Wang X, He X N, Wang M, Feng F L, Chua T S. Neural graph collaborative filtering. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris, France: ACM, 2019. 165−174

26 Lin Jing-Dong, Wu Xin-Yi, Chai Yi, Yin Hong-Peng. Structure optimization of convolutional neural networks: A survey. *Acta*

Automatica Sinica, 2020, **46**(1): 24−37
(林景栋, 吴欣怡, 柴毅, 尹宏鹏. 卷积神经网络结构优化综述. 自动化学报, 2020, **46**(1): 24−37)

27 Li Y, Jin Y L, Song G J, Zhu Z H, Shi C, Wang Y M. GraphMSE: Efficient meta-path selection in semantically aligned feature space for graph neural networks. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2021. 4206−4214

28 Luo H T, Meng X Y, Wang S H, Cao H Y, Zhang W Y, Wang Y Q, et al. Spectral-based graph neural networks for complementary item recommendation. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 8868−8876

29 Feng Y F, You H X, Zhang Z Z, Ji R R, Gao Y. Hypergraph neural networks. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press, 2019. 3558−3565

30 Wei C Y, Liang J, Bai B, Liu D. Dynamic hypergraph learning for collaborative filtering. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta, USA: ACM, 2022. 2108−2117

31 Yu J L, Yin H Z, Li J D, Wang Q Y, Hung N Q V, Zhang X L. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In: Proceedings of the Web Conference. Ljubljana, Slovenia: ACM, 2021. 413−424

32 Cai D R, Song M X, Sun C X, Zhang B F, Hong S D, Li H Y. Hypergraph structure learning for hypergraph neural networks. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI). Vienna, Austria: IJCAI, 2022. 1923−1929

33 Guo Z C, Zhao J X, Jiao L C, Liu X, Liu F. A universal quaternion hypergraph network for multimodal video question answering. *IEEE Transactions on Multimedia*, 2023, **25**: 38−49

34 Rao Zi-Yun, Zhang Yi, Liu Jun-Tao, Cao Wan-Hua. Recommendation methods and systems using knowledge graph. *Acta Automatica Sinica*, 2021, **47**(9): 2061−2077
(饶子昀, 张毅, 刘俊涛, 曹万华. 应用知识图谱的推荐方法与系统. 自动化学报, 2021, **47**(9): 2061−2077)

35 He R N, McAuley J. VBPR: Visual Bayesian personalized ranking from implicit feedback. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI Press, 2016. 144−150

36 Chen J Y, Zhang H W, He X N, Nie L Q, Liu W, Chua T S. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo, Japan: ACM, 2017. 335−344

37 Lin Z H, Tian C X, Hou Y P, Zhao W X. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In: Proceedings of the ACM Web Conference. Lyon, France: ACM, 2022. 2320−2329

38 Xu C, Si J J, Guan Z Y, Zhao W, Wu Y, Gao X Y. Reliable conflictive multi-view learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 16129−16137

39 Han Z B, Zhang C Q, Fu H Z, Zhou J T. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(2): 2551−2566

40 Zhang Q Y, Wei Y K, Han Z B, Fu H Z, Peng X, Deng C, et al. Multimodal fusion on low-quality data: A comprehensive survey. arXiv preprint arXiv: 2404.18947, 2024.

41 Wang C, Niepert M, Li H. LRMM: Learning to recommend with missing modalities. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 3360−3370

42 Malitesta D, Rossi E, Pomo C, Malliaros F D, di Noia T. Dealing with missing modalities in multimodal recommendation: A feature propagation-based approach. arXiv preprint arXiv: 2403.19841, 2024.

43 Liu C, Li R, Wu S, Che H J, Jiang D Z, Yu Z W, et al. Self-guided partial graph propagation for incomplete multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(8): 10803−10816

44 Frank M, Wolfe P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 1956, **3**(1−2): 95−110

45 Yang Y, Shen H T, Ma Z G, Huang Z, Zhou X F. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI). Barcelona, Spain: IJCAI, 2011. 1589−1594

46 Nie F P, Huang H, Cai X, Ding C. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2010. 1813−1821

47 Li X C, Chin J Y, Chen Y L, Cong G. Sinkhorn collaborative filtering. In: Proceedings of the Web Conference. Ljubljana, Slovenia: ACM, 2021. 582−592

48 Courty N, Flamary R, Tuia D, Rakotomamonjy A. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(9): 1853−1865

49 Bahmani S, Raj B, Boufounos P T. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 2013, **14**(1): 807−841

50 Zhang C, Cai Y J, Lin G S, Shen C H. DeepEMD: Few-shot image classification with differentiable earth mover′s distance and structured classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 12200−12210

51 Barratt S. On the differentiability of the solution to convex optimization problems. arXiv preprint arXiv: 1804.05098, 2018.

52 Jaggi M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In: Proceedings of the 30th International Conference on Machine Learning (ICML). Atlanta, USA: JMLR.org, 2013. 427−435

53 He X N, Deng K, Wang X, Li Y, Zhang Y D, Wang M. LightGCN: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Xi′an, China: ACM, 2020. 639−648

54 Wu J C, Wang X, Feng F L, He X N, Chen L, Lian J X, et al. Self-supervised graph learning for recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event: ACM, 2021. 726−735

55 Li J, Tan Z C, Wan J, Lei Z, Guo G D. Nested collaborative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 6939−6948

56 Xia L H, Huang C, Xu Y, Zhao J S, Yin D W, Huang J. Hypergraph contrastive collaborative filtering. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain: ACM, 2022. 70−79

57 Yi Z X, Wang X, Ounis I, Macdonald C. Multi-modal graph contrastive learning for micro-video recommendation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain: ACM, 2022. 1807−1811

58 Tao Z L, Liu X H, Xia Y W, Wang X, Yang L F, Huang X L, et al. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 2023, **25**: 5107−5116

59 Wei W, Huang C, Xia L H, Zhang C X. Multi-modal self-super-

vised learning for recommendation. In: Proceedings of the ACM Web Conference. Austin, USA: ACM, 2023. 790−800

檀彦超 福州大学计算机与大数据学院副教授. 2022 年获得浙江大学博士学位. 主要研究方向为数据挖掘, 推荐系统和医疗保健.
E-mail: yctan@zju.edu.cn
(**TAN Yan-Chao** Associate professor at the College of Computer and Data Science, Fuzhou University. She received her Ph.D. degree from Zhejiang University in 2022. Her research interests include data mining, recommendation system, and healthcare.)

沈春旭 腾讯科技有限公司资深算法研究员. 主要研究方向为多模态语言模型, 深度强化学习与信息检索.
E-mail: lineshen@tencent.com
(**SHEN Chun-Xu** Senior algorithm researcher at Tencent Technology Co., Ltd. His research interests include multimodal language models, deep reinforcement learning, and information retrieval.)

陈佳敏 福州大学计算机与大数据学院硕士研究生. 2020 年获得福州大学学士学位. 主要研究方向为推荐系统和多模态学习.
E-mail: Jiamin020316@163.com
(**CHEN Jia-Min** Master student at the College of Computer and Data Science, Fuzhou University. She received her bachelor degree from Fuzhou University in 2020. Her research interests include recommendation system and multimodal learning.)

马国芳 浙江工商大学计算机科学与技术学院讲师. 2021 年获得浙江大学博士学位. 主要研究方向为推荐系统和金融科技. 本文通信作者.
E-mail: maguofang@zjgsu.edu.cn
(**MA Guo-Fang** Lecturer at the School of Computer Science and Technology, Zhejiang Gongshang University. She received her Ph.D. degree from Zhejiang University in 2021. Her research interests include recommendation system and Fintech. Corresponding author of this paper.)

林政鸿 福州大学计算机与大数据学院博士研究生. 2019 年获得福建农林大学学士学位. 主要研究方向为推荐系统, 多模态学习, 大语言模型. E-mail: hongzhenglin970323@gmail.com
(**LIN Zheng-Hong** Ph.D. candidate at the College of Computer and Data Science, Fuzhou University. He received his bachelor degree from Fujian Agriculture and Forestry University in 2019. His research interests include recommendation systems, multimodal learning, and large language models.)

王石平 福州大学计算机与大数据学院教授. 2014 年获得电子科技大学博士学位. 主要研究方向为机器学习, 计算机视觉与粒计算.
E-mail: shipingwangphd@163.com
(**WANG Shi-Ping** Professor at the College of Computer and Data Science, Fuzhou University. He received his Ph.D. degree from University of Electronic Science and Technology of China in 2014. His research interests include machine learning, computer vision, and granular computing.)

易玲玲 腾讯科技有限公司资深技术专家. 主要研究方向为多模态语言模型, 大规模异构图与个性化推荐算法.
E-mail: chrisyi@tencent.com
(**YI Ling-Ling** Senior technical expert at Tencent Technology Co., Ltd. Her research interests include multimodal language models, large-scale heterogeneous graphs, and personalized recommendation algorithms.)