

分类号：TP391

密 级：公开

学校代码：10363

学 号：2220910106



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 基于信息增强的图神经网络会
话推荐方法研究

论 文 作 者	<u>孙先兰</u>
校 内 导 师	<u>孔超（副教授）</u>
校 外 导 师	<u>李公文（高级工程师）</u>
专业学位类别	<u>计算机技术</u>
研究方向（领域）	<u>数据分析与知识工程</u>

论文提交日期：2025 年 5 月 9 日

分类号：TP391

单位代码：10363

密 级：公开

学 号：2220910106

题 目 基于信息增强的图神经网络会话推荐方法研究

英文并列

题 目 Research on Information Enhanced Graph Neural
Network Session-based Recommendation Method

学生姓名： 孙先兰

校内导师： 孔超（副教授）

校外导师： 李公文（高级工程师）

专业学位类别： 计算机技术

研究方向（领域）： 数据分析与知识工程

论文答辩日期： 2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：孙先三
日期：2025年5月9日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☒。

学位论文作者签名：孙先兰

日期：2015年5月9日

指导教师签名：孙先兰

日期：2015年5月9日

基于信息增强的图神经网络会话推荐方法研究

摘 要

会话推荐系统旨在通过分析用户在短期匿名会话中的行为与偏好，为其提供个性化的项目推荐。会话推荐中常见的方法是利用用户自身会话数据来建模其项目转换模式，这类方法旨在通过捕捉会话中的项目转换规律，推测用户潜在的兴趣变化。然而，实际应用中的会话序列通常较短，蕴含的项目转换信息相对有限，导致模型难以从中提取足够的深层特征。因此，如何设计和构建更有效的图结构，以便图神经网络能够充分学习用户偏好和行为特征，成为会话推荐中的一个重要难题。此外，现实中的会话推荐场景还面临诸多挑战。第一，流行度偏差问题使得热门项目更容易出现在推荐列表中；第二，如果噪声数据未得到有效处理，将对推荐精度产生负面影响；第三，真实推荐场景中用户不一定只存在单一的目的意图。这些问题的存在，使得会话推荐在实际应用中难以提供精准的推荐结果。针对上述问题，本文进行了以下研究，具体工作如下：

（1）针对现有基于图神经网络的会话推荐模型在实际应用中常见的流行度偏差问题，提出一种全局增强的图神经网络会话推荐方法。首先，为了有效缓解流行度偏差问题，采用项目表示的规范化策略，使得在训练和推理过程中，项目与会话图的表示被约束在单位超球面

上；其次，结合自注意力机制和目标注意力机制，进一步增强图神经网络的表达能力；最后，综合考虑全局嵌入、局部嵌入以及目标嵌入，将它们融合成完整的会话嵌入，从而提升当前会话的推荐效果。

（2）针对现有的会话推荐方法中忽略项目之间的复杂高阶关系以及用户行为序列中包含嘈杂的偏好信息等问题，提出一种基于上下文增强的对比学习方法用于会话推荐。首先，通过采用基于度量的学习重构项目序列为紧密的项目兴趣图，在图中形成密集聚类来明确区分用户的核心兴趣；其次，通过为全局图中的每个会话添加上下文向量，聚合会话中每个结点的信息，提供一种自然的方式来传递相邻项之外的信息；最后，设计一个对比学习模块作为辅助任务，共同学习会话中项目的表示。

（3）针对现有的会话推荐方法中没有从高层次的角度捕捉会话语义从而忽略会话中可能存在的各种用户意图问题，提出一种基于多意图感知增强的会话推荐方法。首先，通过使用项目节点的重复频率作为注意力权重来更新项目嵌入，以增强重复项目的信息；其次，通过派生以每个项目为中心的多个会话表示并动态选择重要的会话表示来表明各种用户意图；最后，将多个会话表示解码为项目分布，并使用池函数将其聚合。

关键词：会话推荐，图神经网络，注意力机制，表示学习，对比学习

Research on Information Enhanced Graph Neural Network

Session-based Recommendation Method

ABSTRACT

The session-based recommendation system aims to provide personalized item recommendations by analyzing a user's behavior and preferences during a short-term anonymous session. A common approach in session-based recommendation is to model a user's item transition patterns using their session data. These methods aim to infer the user's potential interest changes by capturing item transition patterns within the session. However, in real-world applications, session sequences are often short, and the item transition information is relatively limited, making it difficult for models to extract enough deep features. Therefore, designing and constructing more effective graph structures so that graph neural networks can fully learn user preferences and behavior patterns becomes a critical challenge in session-based recommendation. In addition, real-world session-based recommendation scenarios face several challenges. First, the popularity bias problem makes popular items more likely to appear in the recommendation list. Second, if noise data is not effectively handled, it negatively impacts recommendation accuracy. Third, in real-world recommendation scenarios, users may not have a single intent or goal. The existence of these issues makes it difficult for session-based recommendation to provide accurate results in practical applications. To

address these problems, this paper conducts the following research, with the specific work outlined below:

(1) To address the common popularity bias issue in existing graph neural network session-based recommendation models in practical applications, a globally enhanced graph neural network session recommendation method is proposed. First, to effectively alleviate the popularity bias problem, a normalization strategy for item representations is employed, ensuring that both item and session graph representations are constrained to the unit hypersphere during training and inference. Next, by combining self-attention and target attention mechanisms, the expressive power of the graph neural network is further enhanced. Finally, in terms of session representation, the global embedding, local embedding, and target embedding are comprehensively considered and fused into a complete session embedding, thereby improving the recommendation performance for the current session.

(2) To address the issues in existing session-based recommendation methods, such as the neglect of complex higher-order relationships between items and the presence of noisy preference information in user behavior sequences, a context-enhanced contrastive learning method for session recommendation is proposed. First, by adopting a metric-based learning approach, the item sequence is reconstructed into a dense item interest graph, where dense clusters are formed to clearly differentiate the user's core interests. Next, context vectors are added to each session in the global graph, aggregating the information of each node in the session, providing a natural way to transmit information beyond adjacent items. Finally, a contrastive learning module is designed as an auxiliary task to jointly learn the representations of items within the session.

(3) To address the issue in existing session-based recommendation methods that fail to capture the session semantics from a higher-level perspective, thereby neglecting the various user intents that may exist within a session, a multi-intent perception-enhanced session recommendation method is proposed. First, the frequency of item node repetitions is used as attention weights to update item embeddings, enhancing the information of repeated items. Next, multiple session representations centered around each item are derived, and important session representations are dynamically selected to indicate various user intents. Finally, the multiple session representations are decoded into item distributions, and a pooling function is used to aggregate them.

KEY WORDS: session-based recommendation, graph neural network, attention mechanism, representation learning, contrastive learning

目录

第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	3
1.2.1 推荐系统.....	3
1.2.2 基于传统方法的会话推荐算法.....	3
1.2.3 基于深度学习的会话推荐算法.....	4
1.2.4 基于对比学习的会话推荐算法.....	6
1.3 研究基本框架及主要贡献.....	7
1.4 研究内容及组织结构.....	10
第 2 章 基础理论及相关技术.....	11
2.1 会话推荐系统概述.....	11
2.2 图神经网络.....	12
2.2.1 图卷积网络.....	13
2.2.2 图注意网络.....	14
2.2.3 门控图神经网络.....	15
2.3 注意力机制.....	17
2.3.1 自注意力机制.....	17
2.3.2 软注意力机制.....	18
2.3.3 多头注意力机制.....	19
2.4 对比学习.....	19
2.5 本章小结.....	20
第 3 章 全局增强的图神经网络会话推荐.....	22
3.1 设计背景.....	22
3.2 模型目标.....	23
3.3 模型结构.....	23
3.3.1 会话图构建.....	24
3.3.2 会话图表示.....	24
3.3.3 模型预测与训练.....	26
3.4 实验与分析.....	26
3.4.1 实验设置.....	27
3.4.2 性能比较.....	28

3.4.3 消融实验分析.....	30
3.4.4 超参实验分析.....	30
3.5 本章小结.....	31
第 4 章 基于上下文增强的对比学习会话推荐.....	32
4.1 设计背景.....	32
4.2 模型目标.....	33
4.3 模型结构.....	33
4.3.1 会话图构建.....	34
4.3.2 会话图嵌入.....	35
4.3.3 模型预测与训练.....	38
4.4 实验与分析.....	38
4.4.1 实验设置.....	38
4.4.2 性能比较.....	40
4.4.3 消融实验.....	41
4.4.4 超参实验分析.....	42
4.5 本章小结.....	43
第 5 章 基于多意图感知增强的会话推荐.....	45
5.1 设计背景.....	45
5.2 模型目标.....	46
5.3 模型结构.....	46
5.3.1 节点表示学习.....	47
5.3.2 多意图选择.....	48
5.3.3 多意图聚合.....	49
5.4 实验与分析.....	50
5.4.1 实验设置.....	50
5.4.2 性能比较.....	52
5.4.3 消融实验.....	52
5.4.4 超参实验分析.....	53
5.5 本章小结.....	54
第 6 章 总结与展望.....	55
6.1 工作总结.....	55
6.2 未来展望.....	56
参考文献.....	57
攻读学位期间发表的学术成果.....	62

第1章 绪论

1.1 研究背景及意义

在当今信息化社会，网络的飞速发展正以空前的速度改善着我们的生活。技术进步推动信息的多样化传播方式，尤其是文字、图像、声音、视频等各种媒介的快速传递，使得互联网的信息量呈现指数级增长^[1]，给社会带来巨大的便利。人们可以轻松获得各种类型的信息，从而促进了个人和社会的进步。然而，随着信息量的急剧增长，如何有效筛选出有价值的内容，成为一个亟待解决的难题。面对铺天盖地的资讯，人们常常难以迅速找到与自身需求相关的信息，进而陷入信息过载的困境。

信息过载问题不仅仅困扰着个人用户，还对整个社会和经济产生了深远影响。为了应对这一挑战，学者和技术专家们提出并研究了多种解决方案^[2]，其中推荐系统作为一种高效的技术手段，已经得到广泛的应用^[3]。推荐系统通过借助机器学习、深度学习等人工智能技术，能够精准捕捉用户的兴趣和需求，分析其行为习惯，预测并推送可能感兴趣的内容。这种系统的核心优势在于其个性化和智能化，使得用户能够在信息繁杂的环境中，快速找到最符合其需求的信息。例如，电商平台上的推荐系统通过分析用户的浏览历史、购买行为以及与其他消费者的相似度，为用户精准推荐商品。即便用户没有明确的搜索意图，系统也能通过推荐商品或服务，帮助用户发现潜在的购买需求，从而提高销售转化率和用户满意度。社交媒体平台则通过分析用户的兴趣、关注列表和互动行为，推送与其兴趣相符的内容，增强用户的互动性和粘性。短视频平台通过用户观看时长、点赞、评论等数据，智能推荐符合用户口味的视频内容，使用户在海量视频中迅速找到自己感兴趣的作品。

推荐系统的普及，已经成为现代互联网不可或缺的一部分，它不仅有效缓解信息过载带来的压力，还大大提升了用户体验。通过不断优化算法和增强数据分析能力，推荐系统能够更加精准地预测和满足个体用户的需求，推动个性化服务

的发展^[4]。同时，推荐系统的广泛应用也为企业带来巨大的商业价值，通过提高用户的活跃度和满意度，推动业务的增长和创新。然而，尽管传统推荐系统在很多场景中表现优异，但在一些特定情况下，它们的效果却大打折扣。特别是在匿名推荐场景中，例如用户以游客身份访问网站或未登录情况下浏览内容时，推荐系统难以获取用户的基本信息，如年龄、性别、历史偏好等。这使得系统无法像处理注册用户那样，追踪用户的长期行为，预测未来需求，并提供精准的推荐。为了解决这些问题，近年来科学家们提出了基于会话的推荐系统^[5,6]，这种系统将基于用户会话中已有的行为记录来预测即将发生的行为。相比传统的推荐系统，基于会话的推荐系统可以更好地为未注册、未登录或其他情况下无法提供年龄、性别等个人信息的匿名用户提供推荐服务。因此，基于会话的推荐系统成为推荐领域的一个重要研究方向，受到越来越多的关注^[7]。

随着人工智能的发展，深度学习方法^[8]已经在会话推荐领域得到广泛应用，其中基于图神经网络的方法表现突出。图神经网络可以有效地建模序列依赖关系和会话内物品之间的复杂依赖关系。因此，基于图神经网络的会话推荐系统成为推荐领域的一个重要研究方向^[9]，受到越来越多的关注。在推荐系统中应用图神经网络的首要步骤是构建合理且高效的图，在构建图数据时，需全面考量节点、边及路径等要素。部分研究致力于利用用户的会话数据来模拟项目间的转换模式。但由于会话序列通常较短，所包含的项目转换信息有限，这一局限性给图神经网络在深入捕捉用户行为特征以准确预测其偏好方面带来了困难。此外，在复杂数据环境中，有效提取用户潜在兴趣是提升推荐系统性能的核心挑战。注意力机制的引入，能够基于用户历史交互行为自适应聚焦会话上下文中的高价值信息，从而提升会话语义表征的准确性。具体而言，本研究通过构建细粒度的会话内项目转移模式并优化会话表示，着重探索增强会话表示层特征融合能力的方法，以支持模型更精准地预测用户后续行为意向。

1.2 国内外研究现状

1.2.1 推荐系统

推荐系统依据用户的个人信息及历史行为数据,学习并分析用户的偏好,进而为用户生成个性化的推荐序列。目前,推荐系统已广泛应用于多个行业领域,成为连接用户与信息、商品及服务的重要桥梁。传统推荐方法主要是基于内容的推荐方法^[10]和基于协同过滤的推荐方法^[11]。基于内容的推荐方法是根据用户的兴趣偏好与项目的特征描述实现推荐,这种推荐方法在一定程度上可以解决冷启动问题。但是难以挖掘用户的潜在兴趣偏好。基于协同过滤的推荐方法是基于用户或项目间的相关性进行推荐,能够挖掘出用户的潜在兴趣,但是难以解决冷启动和数据稀疏问题。传统推荐方法都只考虑用户长期静态偏好,而忽略了用户兴趣随时间的变化,因而出现了序列推荐^[12,13]。这类推荐模型通过解析用户行为的时间序列,致力于捕捉其兴趣演变轨迹,从而实现精准的个性化推荐。然而,随着研究的深入,研究者发现模型往往过度依赖长期行为数据,难以有效捕捉用户兴趣的短期波动。这种对长期兴趣的偏向性使得模型无法及时响应用户偏好的动态变化,从而降低了推荐结果的时效性和准确性,导致推荐内容与用户当前需求产生偏差。因此,将用户行为分解成更小的粒度,考虑用户行为的事务性结构,捕捉用户短期偏好的转移是十分有必要的,因此基于会话的推荐系统得到了快速的研究和发展^[14]。

1.2.2 基于传统方法的会话推荐算法

传统的会话推荐研究大致可以分为两类,分别是基于物品^[15,16]的邻域推荐方法和基于马尔可夫链^[17]的序列推荐方法。基于物品的邻域推荐方法通过计算物品之间的相似度,将与当前点击项相似度最高的物品推荐给用户。因此,它能够基于会话中物品的共现关系进行推荐,尤其适用于捕捉短期的物品兴趣。然而,单纯依赖物品的邻域信息可能缺乏足够的上下文信息,无法充分反映出会话序列的点击顺序和完整性。Item-KNN 模型^[18]通过计算用户历史交互项目与候选项目之间的余弦相似度,从而将与历史点击项相似度较高的物品推荐给用户,进而提

升推荐的相关性和准确性。Wang 等人^[19]提出了端到端神经网络模型 CSRM，该模型通过应用内部编码器来有效表示当前会话状态，并利用外部编码器来检索与当前会话相似的邻域会话，从而更好地学习潜在会话的隐含表示，进而提高了推荐预测的精度和泛化能力。基于马尔可夫链的序列推荐方法则主要通过将会话过程映射到马尔可夫链模型中，利用前一个点击来推测用户的下一个动作。Zhang 等人^[20]利用社交网络挖掘相似用户并联合个性化马尔可夫链模型以动态地模拟用户偏好。Rendle^[21]等人则提出一种结合矩阵分解和马尔可夫链的混合模型，旨在捕捉物品之间的顺序性行为及用户的动态兴趣。尽管基于马尔可夫链的模型在顺序推荐方面取得了较好的效果，但多数模型通常只关注最近的点击项，这使得它们无法有效捕捉复杂的高阶顺序模式和长期依赖关系。因此，如何增强模型的序列记忆能力和对长时间跨度的行为模式的建模仍然是基于马尔可夫链的推荐方法面临的重要挑战。

1.2.3 基于深度学习的会话推荐算法

在基于深度学习的会话推荐算法中，可以分为两大类：基于循环神经网络的会话推荐算法和基于图神经网络的会话推荐算法。

(1) 基于循环神经网络的会话推荐算法

随着深度学习技术的不断发展^[22,23]，基于会话的推荐系统取得了显著的性能提升，尤其是在捕捉用户的短期偏好和行为模式方面。越来越多的学者将神经网络技术应用于会话推荐^[24,25]，显著改善了推荐系统的准确性和推荐效果。通过深度神经网络模型，推荐系统不仅能够学习用户在当前会话中的即时兴趣，还能有效提取用户在多次交互中的潜在兴趣。其中，GRU4Rec 模型^[26]首次将循环神经网络（Recurrent Neural Network, RNN）应用于基于会话的推荐任务，采用改进的门控循环单元来对用户的交互行为进行建模。RNN 的优势在于能够通过递归的方式处理序列数据，捕捉到用户行为的顺序依赖性，从而有效地捕捉用户的短期偏好。然而，RNN 模型对于较长会话的表现并不理想，难以维持长期依赖关系，尤其是在用户会话较长、交互信息复杂的情况下，模型的性能可能会受到限制。为了解决这一问题，Li^[27]等人提出了 NARM 方法，结合了 RNN 与注意力机

制。该方法不仅能捕捉用户的短期偏好，还能通过引入注意力机制，对用户行为序列中不同位置的信息给予不同的权重，从而更好地处理具有顺序性的短期兴趣。同时，注意力机制能够突出重要特征，帮助模型更有效地捕捉全局性偏好和长期兴趣。进一步地，STAMP^[28]模型在 NARM 的基础上引入了多层感知机(Multilayer Perceptron, MLP)和注意力网络，旨在更全面地提取会话中的成对项转换信息。STAMP 通过结合 RNN、MLP 和注意力机制，不仅建模用户当前的偏好，还考虑了用户整体的历史偏好，进一步提高了推荐的个性化和准确性。

(2) 基于图神经网络的会话推荐算法

近年来，图神经网络(Graph Neural Network, GNN)的发展极为迅猛，成为处理图结构数据的有力工具，尤其在推荐系统领域中展现出巨大的潜力。图神经网络的概念最早由 Gori 等人提出^[29]，由 Scarselli 等人^[30]和 Micheli 等人^[31]进一步发展与完善。与传统的基于循环神经网络的方法不同，基于 GNN 的会话推荐方法将会话序列映射为图结构数据^[32,33]，通过图的节点和边表示物品及其相互关系。在这一框架中，推荐系统通过图神经网络聚合图中相邻节点的信息，以捕捉物品之间的复杂转换关系和用户行为的深层次特征。SR-GNN 模型^[34]是图神经网络应用于会话推荐的前驱，首次提出将会话序列转化为会话图进行处理的方法。在这一模型中，会话中的物品作为图的节点，节点之间的边表示物品间的转化关系。SR-GNN 利用图神经网络来学习节点间的关系，并结合注意力机制，以便更精确地获取用户的兴趣偏好。然而，传统的 GNN 方法仅能从会话图中收集相邻节点的信息，这可能会限制其建模能力，尤其是在捕捉长距离依赖关系时。为了解决这一问题，Zhang 等人^[35]提出 DGTN 模型，该模型不仅构建当前会话的会话图，还融合其他会话图的信息，通过双通道学习，DGTN 显式地学习了两种类型的物品嵌入：一种代表当前会话的嵌入，另一种则表示其他会话中物品的关系，增强了当前会话图与其他会话图之间的交互，提升了模型的表现力。此外，Transformer^[36]作为一种基于自注意力机制的自然语言处理模型，在各类推荐系统中得到了广泛的应用。自注意力机制能够捕捉序列中物品之间的长期依赖关系，使其在会话推荐任务中表现出了良好的效果。例如，GC-SAN^[37]通过引入自注意力机制，能够更有效地学习会话序列中物品间的长期依赖，捕捉用户的全局偏好。

GCE-GNN^[38]将全局图和局部会话图的表示结合，通过图注意力网络从会话图和全局图中分别学习物品的表示，进一步提升了推荐的准确性和个性化。此外，TAGNN 模型^[39]通过自适应激活机制，动态地计算不同会话中用户对特定项目的兴趣，并使用注意力机制聚焦与目标项目相关的会话表示向量上，从而优化会话表示的质量。Ye 等^[40]提出的 CA-TCN 模型通过构建全局项目图和会话上下文图，模拟跨会话对项目 and 会话表示的影响，进一步扩展了会话推荐系统的表达能力，能够更好地捕捉用户在不同会话中的兴趣变化。Li 等^[9]提出的因子级项目嵌入方法，通过使用注意力机制精细捕捉用户对不同因素的偏好，从而深入挖掘用户的细粒度兴趣。这些细粒度兴趣的建模使得推荐系统能够更加精准地反映用户在不同情境下的行为。

1.2.4 基于对比学习的会话推荐算法

在过去的两年中，随着对比学习（Contrastive Learning, CL）^[41,42]在计算机视觉（Computer Vision, CV）和自然语言处理^[43]（Natural Language Processing, NLP）领域取得了显著的成功，它在推荐系统中的应用也引起了广泛关注^[44,45]。在推荐系统中，对比学习被用来有效挖掘用户与物品之间的潜在关系，尤其是在标注数据稀缺的情况下，能够显著改善模型效果。MHCN 模型^[46]通过从用户的社交关系中构建自监督信号，进一步增强了用户的表示能力，使得模型能够获取更全面、更细粒的用户特征。这种方法的创新之处在于它利用用户之间的交互和社交网络中的信息，来捕捉更加丰富的用户偏好，进而提升推荐系统的准确性。S²-DHCN^[47]模型则采用了超图来构建会话数据，并通过超图卷积网络来捕捉物品之间复杂的高阶关系。在传统图模型中，边仅连接两个节点，而在超图中，边可以连接多个节点，能够更有效地建模物品之间的多重关系，这对于捕捉复杂会话数据中的隐性联系至关重要。通过超图卷积，S²-DHCN 能够更好地反映物品之间的高阶关系，增强了推荐系统的表现力。此外，COTREC 模型^[48]提出了一种自监督图的共同训练机制，通过递归地利用两个视图（物品视图和会话视图）上的不同连通性信息生成真实样本。这一机制通过多视图的自监督学习，可以帮助模型挖掘出不同物品和会话之间的关系，并通过对比学习对这些生成的样本进行

监督，从而优化推荐系统的效果。COTREC 的创新之处在于它同时处理多种信息源，提升了模型对不同视角的理解和融合能力。

1.3 研究基本框架及主要贡献

现实中的会话推荐系统面临着诸多复杂的挑战，首先，推荐系统常伴有流行度偏差问题，某些热门项目由于具有广泛的受众基础，通常更容易出现在推荐列表中，导致模型过度关注这些高频物品，而忽视了那些更具个性化和潜在价值的项目。然后，用户行为序列中通常包含嘈杂的偏好信息，他们不能充分反映用户的实际偏好，比如用户可能会误点击或者出于好奇点击不相关的物品，能否从含有噪声的数据中有效提取用户的核心兴趣，是决定模型性能优劣的关键因素之一。此外，在用户会话序列中，不一定只存在单一的意图，例如，用户在为旅行做准备时，时尚和防晒同样符合用户意图，忽略会话中可能存在的各种用户意图会导致推荐性能显著下降。因此，当代复杂网络下的会话推荐任务存在以下挑战：

（1）如何有效缓解流行度偏差问题：现有的基于图神经网络的推荐模型在现实世界中经常存在流行偏差问题，它们主要关注并推荐流行项目，忽略相关的长尾项目（不太流行或不太频繁的项目）。因此，如何缓解上述问题并应用于会话推荐任务成为一个研究问题。

（2）如何有效缓解用户序列噪声问题：现有的基于图神经网络的推荐模型中，用户的行为序列通常包含嘈杂的偏好信息，他们不能充分反映用户的实际偏好。因此，如何有效的对他们的真实兴趣建模成为一大研究问题。

（3）如何有效缓解会话中可能存在的多种用户意图问题：在现实的推荐场景中，用户不一定只存在单一的目的意图，忽略会话中可能存在的各种用户意图会导致推荐性能显著下降。因此，如何在会话序列中提取重要的用户意图成为一个研究问题。

本文面向会话推荐任务，针对大数据背景下存在的噪声问题和现有的图神经网络模型存在的问题等设计了三种会话推荐算法并进行了应用研究。图 1-1 展示了本文的研究框架并表明主要研究贡献，具体来说，本文对基于信息增强的图神经网络会话的推荐方法展开三个方向的研究，它们分别是全局增强的图神经网络

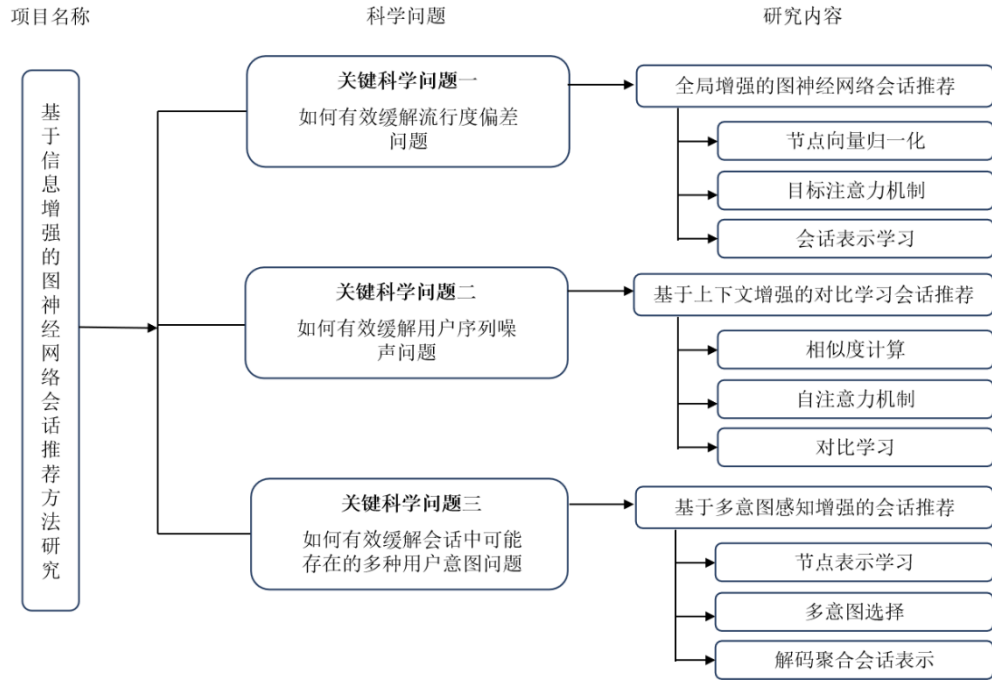


图 1-1 本文基本研究框架

Fig. 1-1 The basic research framework of this article

会话推荐、基于上下文增强的对比学习会话推荐，以及基于多意图感知增强的会话推荐，主要包含如下内容：

(1) 针对全局增强的图神经网络会话推荐。为了缓解流行偏见的问题，该模型首先对项目表示进行了规范化，并引入了自注意力机制来增强图神经网络。一方面，自注意力机制通过显式关注所有位置来捕获全局依赖关系。另一方面，GNN 可以捕获相邻项目之间丰富的局部依赖关系，以提供丰富的局部上下文信息。其次，考虑到过度依赖目标会话中最后一个项目的局限性，引入目标感知注意力来计算会话中所有项目相对于每个目标项目的注意力得分。通过注意力得分，了解目标会话中每个项目对用户偏好的贡献，并通过对目标会话中的每个项目进行加权来获得目标会话表示。对于会话表示学习，在使用注意力机制获得全局嵌入和目标嵌入后，同时考虑将全局嵌入、局部嵌入和目标嵌入集成到整个会话嵌入中，以提高当前会话的推荐性能。在训练和推理过程中，充分利用了将项目和会话图表示限制在单位超球面上的优势，提出了规范化的项目和会话表示模型。

(2) 针对上下文增强的对比学习会话推荐。首先，为了缓解用户的历史行为序列中的噪声导致了对用户真实偏好建模的不准确性，通过计算项目对之间的

相似性来构造兴趣图，将松散的项目序列转换为紧密的项目-项目图，更容易区分用户的主要兴趣和次级兴趣。其次，为了缓解 GNN 在非相邻项目之间信息传播和聚合的不足，通过为每个会话添加一个上下文向量来构建会话图，以学习会话图中不直接连接的项目对之间的关系，该上下文向量可以提供一种自然的方式来传递相邻项之外的信息。最后，在获得原全局会话图和对比图下的两种项目嵌入后，使用对比学习作为辅助任务使两种会话嵌入互信息最大化，最后根据会话特征计算候选物品的推荐概率，本研究使用对比学习方法有效弥补了会话数据稀疏的缺点，同时缓解了噪声信息所带来的影响。

(3) 基于多意图感知增强的会话推荐。目前大多数方法往往关注最后一项或者只独立考虑会话中的每个项目，并没有从高层次的角度捕捉会话语义，如果最后一项不太重要或者嘈杂，它们就无法准确地捕获用户意图。为了缓解没有从高层次的角度捕捉会话语义从而忽略会话中可能存在的各种用户意图问题，本课题首先通过使用项目节点的重复频率作为注意力权重来更新项目嵌入，降低对稀疏、噪声数据的敏感度，更好地理解 and 利用用户行为的序列特征，并且通过引入位置嵌入信息来区分序列的顺序，更好的描述会话中每个项目的重要性。其次，每个项目在序列中可能与多个其它项目相关联，并且这些关系可能在不同的方向上（前向和后向）具有不同的重要性，双向自注意力机制通过计算两个方向上的注意力权重来更全面地捕捉序列中项目之间的复杂关系和语义信息，高速网络是一种深度神经网络结构，能够有效地整合输入的多层特征信息。在推荐系统中，会话信息对于理解用户的整体行为模式至关重要，而每个项目的具体信息则有助于捕捉每个项目的个性化特征和重要性。因此采用双向自注意机制和高速网络来获取会话中不同的用户意图并自适应地提取会话中重要的用户意图。最后，将多个会话表示解码为项目分布，并使用池函数将其聚合从而进行推荐。

总的来说，上述三项内容是一个整体，且层层递进。第一项研究内容是本课题的基础部分，通过运用图神经网络和注意力机制算法来进行会话表示学习，但是并未考虑图神经网络忽略了非相邻项目之间的复杂高阶关系，并引入了噪声信息，因此本课题拟提出研究内容二，通过为每个会话添加一个上下文向量来传递相邻项之外的信息同时引入对比学习作为辅助任务使两种会话嵌入互信息最大

化。但在真实推荐场景中用户不一定只存在单一的目的意图，忽略会话中可能存在的各种用户意图会导致推荐性能显著下降，故本课题拟提出研究内容三，通过获取会话中不同的用户意图并自适应地提取会话中重要的用户意图，希望有效缓解上述研究出现的问题。

1.4 研究内容及组织结构

本文的研究内容和组织结构如下：

第 1 章，介绍了基于信息增强的图神经网络会话推荐方法的研究背景及意义，阐明了当前会话推荐系统所面临的挑战，并对目前国际上的研究状况做了简要的介绍，最后描述了本文研究的基本框架及主要贡献，并总结了本文的组织结构。

第 2 章，简要介绍了会话推荐的系统概述，同时介绍了会话推荐中用到的相关理论技术，包括图神经网络、注意力机制及对比学习等，并比较了它们的优势与局限性。

第 3 章，全局增强的图神经网络会话推荐。详细探讨了推荐系统中普遍存在的流行度偏差问题，具体描述了模型的设计细节，并深入分析了实验结果。

第 4 章，基于上下文增强的对比学习会话推荐。详细介绍了噪声数据问题和图神经网络在节点信息传播的局限性问题，具体描述了模型的设计细节，并深入分析了实验结果。

第 5 章，基于多意图感知增强的会话推荐。详细介绍了会话中可能存在的各种用户意图问题，具体描述了模型的设计细节，并深入分析了实验结果。

第 6 章，再次对本研究进行了全面的总结和未来展望。本章系统地回顾和整理了本文的研究工作，明确本文提出算法的优点和局限性，并对后续可能的研究方向提出了前瞻性的观点。

第2章 基础理论及相关技术

2.1 会话推荐系统概述

传统的推荐方法主要依靠用户的偏好信息或长期的历史交互信息来进行推荐。在匿名登录或受到隐私政策限制等场景中，例如用户以访客身份登录浏览网站时，这些数据变得无法访问，因此无法挖掘用户的历史偏好和长期偏好，从而导致推荐效果不佳。在此背景下，基于会话的推荐系统应运而生。

基于会话的推荐系统^[49]通过分析匿名用户的连续行为序列挖掘用户偏好，从而预测下一个可能与之交互的项目。在会话推荐中，一个会话是指用户在特定时间段内进行的一系列连续交互行为，可以是一次视频观看过程中的连续点击行为、一次连续播放音乐的列表记录等。会话数据^[50]通常以序列的形式存在，包括用户与物品的交互序列、交互行为的时间戳等信息。会话推荐的核心任务聚焦于对用户交互序列的建模分析，推断其可能产生交互的候选物品。这类系统能够有效捕捉到用户的动态变化和短期偏好，尤其在没有用户历史数据的情况下依然能够提供高质量的推荐服务。因此，基于会话的推荐系统在许多实际应用场景中表现出了良好的适应性，受到了国内外研究的广泛关注。随着算法的不断优化和计算能力的提升，基于会话的推荐系统逐渐成为研究和工业界的热门领域，特别是在电子商务、社交平台和新闻推荐等应用场景中展现了巨大的潜力。

会话推荐的关键技术融合了多个先进的机器学习和深度学习方法，主要包括用户行为序列建模、上下文信息整合和相似度计算。首先，通过对用户历史行为的深入分析，能够有效地捕捉用户的兴趣变化和潜在需求，尤其是利用图神经网络等方法建模用户与物品之间的复杂关系，从而增强推荐的准确性和个性化。其次，上下文信息的引入，诸如用户所在的时间、地点、设备等因素，能够帮助系统更好地理解用户的即时需求和偏好。注意力机制在这一过程中发挥着重要作用，通过动态调整权重，突出当前会话中的关键信息，提高推荐的相关性和即时性。此外，通过精准的相似度计算，系统能够识别与用户需求最相关的物品，并通过

对比学习等技术增强模型的泛化能力和鲁棒性,从而使得推荐结果更加贴合用户的短期兴趣。综合这些技术,会话推荐能够在复杂和动态的环境中提供高效、个性化的推荐服务。

2.2 图神经网络

图神经网络^[51]是一类专门设计用来处理图结构数据的深度学习模型。图^[52]是由节点 (Vertices) 和边 (Edges) 组成的非欧几里得结构数据。图神经网络通过节点之间的连接关系来捕捉图中复杂的结构信息和节点的高维表示。与传统的卷积神经网络^[53] (Convolutional Neural Network, CNN) 不同, CNN 主要用于处理图像数据, 而图神经网络则解决了图数据的特征学习问题。

GNN 的核心思想是通过节点间的相互作用进行信息传播, 让每个节点的代表能够充分融合其邻居节点的信息, 通过多次迭代 (即多层消息传递), 节点的代表能够逐渐融合整个图结构的信息, 从而能够更好地捕捉图中复杂的关系和特征。具体来说, 图神经网络通常分为三个主要步骤:

(1) 信息传递: 每个节点通过其与邻居节点的连接关系传递信息。这意味着一个节点不仅仅依赖于自己的特征, 还要聚合来自它的邻居节点的特征。在信息传递过程中, 节点通过边传递自己的特征, 并接收邻居节点的特征。

(2) 聚合: 每个节点会对其邻居节点的信息进行聚合操作。常见的聚合方式包括加权求和、平均池化或最大池化等。这一步骤的关键在于如何设计合适的聚合函数, 以便能够有效地合并不同邻居节点的信息。

(3) 更新: 聚合后的邻居信息会与节点本身的特征一起输入到神经网络中, 通常是一个全连接层, 进行更新生成新的节点表示。这一过程是通过激活函数完成的, 目的是让节点的代表能够适应下游任务的需求。

随着图神经网络研究不断深入, 提出了多种不同的 GNN 变种, 以应对不同的应用需求和数据规模。常见的 GNN 变种包括: 图卷积网络、图注意力网络和门控图神经网络。

2.2.1 图卷积网络

图卷积网络（Graph Convolutional Networks, GCN）是最经典和广泛使用的图神经网络之一，最早由 Thomas Kipf 和 Max Welling 在 2016 年提出，是一类专门用于处理图数据的深度学习模型，广泛应用于社交网络、推荐系统、知识图谱、化学分子结构等领域。与传统的卷积神经网络不同，GCN 通过将卷积操作推广到图结构数据，能够有效地学习节点之间的相互关系和节点的高维特征。在图数据中，节点并非以规则网格的形式存在，因此不能直接应用传统的卷积操作。GCN 的关键思想是通过多层卷积操作逐层传播节点的信息，并通过权重矩阵对节点特征进行线性变换，最后得到每个节点的嵌入表示，这些表示不仅包含节点自身的特征信息，还整合了其邻域节点的特征。此外，GCN 通过归一化的邻接矩阵对每一层的节点特征进行卷积计算，其中，邻接矩阵在计算时通常会进行归一化，以避免不同节点度数带来的偏差。通过这种方式，GCN 能够有效捕捉图中节点与其邻居之间的复杂关系，并生成能够反映图结构的节点表示。GCN 具有端到端的训练方式，即通过反向传播算法对模型参数进行优化，从而学习到图中的节点嵌入。

在 GCN 中，图卷积操作的公式通常表示为节点特征的更新过程。假设有一个图 $G=(V,E)$ ，其中， V 是节点集合， E 是边的集合，每个节点 i 都有一个特征向量 $\mathbf{h}_i$ ，GCN 的目标是通过多层信息传递来更新每个节点的特征，GCN 更新的基本公式如下：

$$\mathbf{H}^{(l+1)} = \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \quad (2-1)$$

其中， $\mathbf{H}^{(l)}$ 是第 l 层的节点特征矩阵， $\mathbf{W}^{(l)}$ 是该层的可学习权重矩阵， σ 是激活函数。 $\mathbf{A}$ 是归一化后的邻接矩阵，用于表示图的结构，归一化的目的是解决不同节点度数差异带来的影响。没有归一化的邻接矩阵会导致节点度数较高的节点会对信息的聚合产生更大的影响。通过归一化^[54]邻接矩阵，可以确保每个节点在聚合邻居信息时有更公平的影响。常见的归一化方法是 $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ，其中 $\mathbf{A}$ 是邻接矩阵， $\mathbf{D}$ 是对角矩阵。

尽管 GCN 在许多应用中取得了良好的效果，但它在处理大规模图数据时，

计算复杂度较高,且随着网络层数的增加,可能会出现节点特征过度平滑的问题,从而导致模型性能下降。为了解决这些问题,研究者提出了一些改进方法,例如通过引入邻居采样或注意力机制等手段,改善 GCN 在大规模图数据中的效率和表达能力。总的来说,GCN 在图数据处理领域展现出了强大的能力,尤其是在节点分类、图分类和链路预测等任务中,其应用前景广泛,且随着模型改进的不断深入,GCN 的适用范围和性能将持续提升。

2.2.2 图注意力网络

图注意力网络^[55] (Graph Attention Network, GAT) 作为图神经网络中的一种重要变体,凭借其在节点信息聚合中的灵活性和自适应性,已成为图学习领域的关键技术。GAT 的核心思想是通过引入自注意力机制,动态地为节点的邻居分配不同的权重,从而使得信息的传播更加精确且有选择性。与传统的图卷积网络依赖于固定邻接矩阵进行信息聚合不同,GAT 能够通过计算注意力系数,基于节点特征自适应地调整信息传播的权重。这种机制使得每个节点在聚合其邻居节点信息时,不仅考虑邻居的拓扑关系,还能够根据邻居节点的重要性赋予不同的权重。这种基于注意力的权重分配方式,使得 GAT 能够在图结构复杂、节点特征差异显著的场景下,展现出较强的学习能力。

首先,GAT 通过一组学习参数计算节点与邻居之间的注意力系数,这一系数反映邻居节点在信息传递中的重要性。然后,通过对邻居节点特征进行加权求和,得到新的节点表示。通过这种方式,GAT 不仅能够捕捉到节点的局部结构信息,还能自动识别哪些邻居节点对于该节点的表示更为重要。图注意力网络模型在多个图学习任务中表现出色,尤其是在节点分类、图分类以及链路预测等应用^[56]中,能够有效提升模型的表达能力和准确性。相比于传统的基于图卷积的方法,GAT 由于其灵活的注意力机制,在处理异构图^[57]和动态图时表现出了较高的适应性。

给定图中的一个节点 v , 它的邻接节点集合为 $N(v)$, GAT 更新过程如下: 对于每个邻居节点 $\mu \in N(v)$, 计算注意力系数 α_{ij} , 表示节点 i 和邻居节点 j 信息的关注程度, 通过两个节点的特征计算得到:

$$\alpha_{ij} = \frac{\exp(a(\mathbf{W}\mathbf{h}_i, \mathbf{W}\mathbf{h}_j))}{\sum_{k \in N(v)} \exp(a(\mathbf{W}\mathbf{h}_i, \mathbf{W}\mathbf{h}_k))} \quad (2-2)$$

其中, $\mathbf{h}_i$ 和 $\mathbf{h}_j$ 是节点 i 和节点 j 的特征, $\mathbf{W}$ 是节点特征的线性变换矩阵, a 是注意力系数的学习权重。计算出注意力系数后, 节点 i 对所有邻居节点的特征进行加权求和, 公式如下:

$$\mathbf{h}'_i = \sigma \left(\sum_{j \in N(v)} \alpha_{ij} \mathbf{W}\mathbf{h}_j \right) \quad (2-3)$$

其中, α_{ij} 是通过注意力机制计算的节点 i 和节点 j 之间的注意力权重, $\mathbf{W}$ 是权重矩阵, σ 是激活函数。

然而, 尽管 GAT 在图学习任务中取得了显著的成果, 仍然面临一些挑战, 特别是在大规模图上的计算效率问题。由于 GAT 需要计算所有邻居节点之间的注意力系数, 计算量随着图的规模增大而急剧增加, 这对计算资源和处理时间提出了较高的要求。此外, 虽然多头注意力机制能够提升模型的表达能力, 但它也引入了额外的计算开销。为了解决这些问题, 研究者们提出了多种改进方法, 如邻居采样、注意力系数的稀疏化等策略, 以期提高 GAT 在大规模图上的效率和性能。总体而言, GAT 通过其创新的注意力机制和灵活的特征聚合方式, 在图神经网络的发展中占据了重要地位, 且随着进一步研究的深入, 预计将会有更多高效的变种和优化方法出现, 推动其在更加复杂的图学习任务中的应用。

2.2.3 门控图神经网络

门控图神经网络^[58] (Gated Graph Neural Network, GGNN) 是一种基于图神经网络的先进模型, 它通过引入门控机制来优化信息在图结构中的传播过程, 从而提高了处理复杂图数据时的效果。传统的图神经网络如图卷积网络和图注意力网络依赖于固定的聚合函数来更新节点表示, 但在面对具有长距离依赖关系的图数据时, 这些方法可能会出现信息传播的过度平滑或梯度消失等问题, 限制了它们在某些任务中的表现。为了解决这些问题, GGNN 引入了类似于长短期记忆网络^[59]和门控循环单元中的门控机制, 使得节点间的信息传递过程能够灵活调整。具体来说, GGNN 通过在每一轮迭代中动态地调整信息的传播路径来更新节点

表示，避免了固定聚合方式的局限性，从而更好地捕捉图中的复杂依赖关系。模型的核心思想是通过门控单元控制节点状态的更新，确保信息传递的效率和有效性。节点在每一轮迭代中的状态不仅依赖于其自身的特征，还通过门控机制融合了来自邻居节点的特征，进而通过递归更新过程逐步优化节点表示。GGNN 的这种递归更新方式使其能够处理长距离的依赖关系，适应更加复杂的图结构。

每一轮迭代中，节点状态的更新不仅考虑其邻居节点的状态，还通过门控机制对状态的更新进行调节。迭代过程可以通过多个步骤进行，不同层之间的消息传递公式如下：

$$\mathbf{a}_{s,i}^t = \mathbf{A}_i^s [\mathbf{V}_{s,i}^{t-1} \mathbf{H}_1, \mathbf{V}_{s,i}^{t-1} \mathbf{H}_2]^T + \mathbf{b} \quad (2-4)$$

$$\mathbf{z}_{s,i}^t = \sigma(\mathbf{W}_z \mathbf{a}_{s,i}^t + \mathbf{U}_z \mathbf{v}_{s,i}^{t-1}) \quad (2-5)$$

$$\mathbf{r}_{s,i}^t = \sigma(\mathbf{W}_r \mathbf{a}_{s,i}^t + \mathbf{U}_r \mathbf{v}_{s,i}^{t-1}) \quad (2-6)$$

$$\hat{\mathbf{v}}_i^t = \tanh(\mathbf{W}_0 \mathbf{a}_{s,i}^t + \mathbf{U}_0 (\mathbf{r}_{s,i}^t \odot \mathbf{v}_{s,i}^{t-1})) \quad (2-7)$$

$$\mathbf{v}_i^t = (1 - \mathbf{z}_{s,i}^t) \odot \mathbf{v}_{s,i}^{t-1} + \mathbf{z}_{s,i}^t \odot \hat{\mathbf{v}}_i^t \quad (2-8)$$

其中， $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W}_0 \in \mathbb{R}^{2d \times d}$ 为权重矩阵， $\mathbf{z}_{s,i}^t$ 表示重置门，用于控制信息遗忘程度。 $\mathbf{r}_{s,i}^t$ 为更新门，用于控制当前时刻的状态更新量。 σ 为 sigmoid 函数， $\odot$ 为对应元素相乘算子。

然而，这种更新机制也带来了计算复杂度的增加，特别是在大规模图数据上，可能导致较高的计算成本和较长的训练时间。尽管如此，GGNN 在多个应用场景中展现出了强大的表现力，尤其是在图分类、节点分类、图生成以及链接预测等任务中，证明了在图数据学习中的优势。通过灵活的门控机制，GGNN 能够在多个领域中有效处理图数据的多样性和复杂性，成为图神经网络研究中的一个重要方向。

2.3 注意力机制

注意力机制^[60]（Attention Mechanism）最早起源于神经科学领域，用来描述人类在处理信息时对特定部分的重点关注。注意力机制的基本思想是：对于输入序列中的每一个元素，模型会根据其与当前状态或上下文的相关性，分配一个权重值（即注意力权重）。这个权重反映了模型在某一时刻对某个元素的关注度，并通过加权求和的方式对输入进行处理。通过这种方式，注意力机制能够帮助模型在处理信息时，着重关注与当前任务最相关的信息，从而克服了传统神经网络在处理长序列或复杂信息时容易丢失细节或难以捕捉长期依赖的问题。

在注意力机制的具体实现框架中，该模块通常由三个核心要素协同作用：查询向量（Query）、键向量（Key）和值向量（Value）。其中，查询向量表示当前时刻模型的状态，键向量表示输入不同部分的特征，而值向量则是与键相关的实际信息载体。该机制的核心运作逻辑体现在两个关键阶段：首先通过评估查询向量与每个键向量之间的相关性，计算相似度匹配得分；其次对得分进行归一化处理，生成反映不同位置重要程度的注意力权重系数，然后将该权重分配给对应的值，最终得到加权求和作为输出。具体公式表示如下：

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (2-9)$$

其中， $\mathbf{Q}$ 是查询， $\mathbf{K}$ 是键， $\mathbf{V}$ 是值， softmax 操作用于归一化权重，使得所有权重之和为 1。

注意力机制可以根据其计算方式和应用场景的不同，分为几种不同的类型，其中最常见的是有以下几种：自注意力机制，软注意力机制和多头注意力机制。

2.3.1 自注意力机制

自注意力机制（Self-Attention Mechanism）旨在通过动态调整序列中各个元素之间的相对重要性来捕捉远程依赖关系。与传统的循环神经网络或长短期记忆网络不同，自注意力机制不依赖于时间顺序来处理数据，而是通过对输入序列中的每个元素与其他元素进行逐对比较，从而计算出每个元素对其他元素的注意力

权重。这一过程通过查询、键和值之间的相似度来完成，通常通过点积运算计算出每一对元素之间的相似度，然后对这些相似度进行归一化处理，得到一个权重分布，进而加权求和所有值，从而得到每个元素的新的表示。这种机制能够有效捕捉长期依赖关系，并能够在计算上进行高度并行化，提高了处理效率。自注意力机制的优势在于其灵活性和高效性，通过为每个输入元素分配一个全局的权重，模型可以自主决定在特定任务中需要关注哪些部分的信息。这使得自注意力在多种自然语言处理任务中表现出色，尤其是在机器翻译、文本生成和问答系统等应用中，能够更好地捕捉句子中各个词之间的复杂关系和长距离依赖。通过这种方式，自注意力机制不仅提升了模型的表达能力，也为后续更复杂的模型如 Transformer 架构的提出奠定了基础，推动了深度学习在序列建模和语言理解等领域的突破。

2.3.2 软注意力机制

软注意力机制（Soft Attention Mechanism）主要用于通过加权求和的方式动态地聚焦输入数据的不同部分，从而提高模型对关键信息的捕捉能力。查询与键之间的相似度通常通过点积或其他相似度度量计算得到，然后对这些相似度值应用 Softmax 函数，生成一组归一化的权重分布。这些权重反映了模型对于不同输入元素的关注程度，权重越大，表示该元素对当前任务的贡献越大。随后，模型根据这些权重对对应的值进行加权求和，从而得到一个加权的表示，作为该输入元素的新表示。在这种机制下，模型能够根据任务需求灵活地关注输入数据中的不同部分，动态地调整其注意力分配。这种机制的核心优势在于它能够在处理输入序列时，对每个元素根据其重要性分配一个全局的注意力权重，从而使得模型能够自适应地捕捉到不同部分的信息。与硬注意力机制相比，软注意力机制具有较强的平滑性和可微性，这使得它能够通过反向传播进行端到端训练，并且具有较高的计算效率。软注意力机制在自然语言处理任务中得到了广泛应用，尤其是在机器翻译、图像描述和情感分析等任务中，能够有效地捕捉输入序列的局部与全局依赖关系，显著提高了模型的性能和表达能力。通过对输入数据进行加权求和，软注意力机制使得模型能够根据上下文动态地调整其关注点，灵活应对不同

的任务需求，极大地促进了深度学习模型的泛化能力和精确度。

2.3.3 多头注意力机制

多头注意力机制（Multi-Head Attention Mechanism）是一种在深度学习领域广泛应用的技术，旨在通过并行化计算多个注意力头（Attention Heads）来增强模型对输入数据不同子空间的表示能力。在多头注意力机制中，输入的查询、键和值首先被映射到多个不同的线性子空间，形成多个独立的查询、键和值的组合，每个组合对应一个独立的注意力头。每个注意力头在自己的子空间内进行标准的注意力计算，利用加权求和的方式从输入中提取出重要信息。在完成所有注意力头的计算后，这些结果会被连接在一起，并通过一个线性变换层进行映射，最终得到多头注意力机制的输出，公式如下所示：

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \mathbf{W}^O \quad (2-10)$$

每个头的计算公式为： $\text{head}_i = \text{Attention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V)$ ，其中， $\mathbf{W}_i^Q$ ， $\mathbf{W}_i^K$ ， $\mathbf{W}_i^V$ 是对应的权重矩阵， h 是头的数量。

通过引入多个注意力头，模型能够从不同的角度捕捉输入数据的多样化信息，从而更加全面地理解数据的内部结构与关系。与单头注意力机制相比，多头注意力机制具有更强的表达能力，因为它能够同时关注输入数据的多个子空间，而不会局限于单一的注意力模式。这种机制的优势在于能够捕捉到不同层次和不同粒度的依赖关系，尤其适用于复杂的任务，如自然语言处理中的机器翻译、文本生成等问题。通过将多个独立的注意力头的输出进行结合，多头注意力机制能够提升模型的表示能力和鲁棒性，增强模型对长距离依赖的建模能力，并有效解决信息丢失和梯度消失等问题。此外，多头注意力机制还具有较高的计算效率和并行性，因为每个注意力头的计算可以并行进行，这使得其在处理大规模数据时具有明显的优势。

2.4 对比学习

对比学习（Contrastive Learning）是一种无监督学习^[61]方法，旨在通过比较

不同样本之间的相似性和差异性，学习到有效的特征表示。在对比学习中，模型通过将输入数据映射到一个表示空间，优化这些表示，将相似的样本拉近，使得它们的表示更加相似，而将不相似的样本推远，使得它们的表示在特征空间中距离更远。具体地说，给定一对输入样本，其中一个样本被称为正样本，另一个则为负样本。正样本与当前样本在某些方面具有相似性或联系，而负样本则与当前样本在这些方面有较大差异。对比学习通过对比正负样本对之间的距离或相似度，来引导模型学习到更加区分性的特征。在实际应用中，对比学习一般通过设计一个合适的对比损失函数来优化模型。最常见的损失函数^[62]包括对比损失（Contrastive Loss）和信息量损失（InfoNCE Loss）。对比损失函数通过计算正样本对之间的距离小于负样本对之间的距离来优化模型参数。信息量损失函数则利用 softmax 计算正样本的概率，推向正确的分类，从而加强模型对区分样本的学习。

在推荐系统中，对比学习被用来提高个性化推荐的质量和效率，尤其是在没有丰富标签数据的情况下。推荐系统的任务通常是从大量的用户和项目中选择最相关的项目进行推荐。通过对比学习，推荐系统能够学习到用户和项目之间的潜在关系，而不仅仅依赖于显式的用户评分或交互数据。具体来说，通过对比学习，推荐系统可以为每个用户和每个项目学习一个低维度的表示向量。对比学习的一个常见策略是将用户与其历史交互的项目视为正样本，将与这些项目无关的其他项目视为负样本。然后，通过优化损失函数，使得用户和相关项目的表示靠近，而用户和不相关项目的表示远离。通过有效地训练用户和项目的嵌入向量，推荐系统能够更精准地匹配用户兴趣，提高个性化推荐的效果。

2.5 本章小结

本章全面介绍了会话推荐系统的核心技术，包括图神经网络、注意力机制和对比学习等方法。通过深入分析这些技术的工作原理与相互联系，阐述了它们在提升推荐精度和用户体验方面的巨大潜力。图神经网络作为处理复杂关系数据的有力工具，能够有效捕捉用户之间以及用户与物品之间的图结构信息，为会话推荐系统提供了强大的建模能力。注意力机制则通过自动关注信息中重

要的部分，进一步优化了模型的推荐效果，特别是在动态变化的会话上下文中。对比学习作为一种无监督学习方式，通过引导模型在不同样本之间进行相似度比较，强化了模型对用户偏好的敏感性，并提升了推荐的多样性和个性化水平。本章还详细讨论了这些方法的具体实现和评估标准，以确保模型在实际场景中的有效性与可操作性。通过对这些方法的比较与分析，本章为后续的算法优化和应用开发提供了理论指导和实践依据。

第3章 全局增强的图神经网络会话推荐

考虑到会话推荐系统中存在的流行度偏差从而导致推荐性能变差的问题,本章提出了全局增强的图神经网络会话推荐方法来提升系统的推荐效果。3.1 节剖析当前会话推荐方法的局限性,并针对性地提出改进策略;3.2 节阐述模型的基础定义,同时介绍会话推荐的目标;3.3 节基于模块化设计思想,系统构建了会话推荐模型的整体架构。该架构由规范化项目表示模块、注意力机制模块和会话表示学习模块这三大模块组成;3.4 节通过构建对比实验,在公开数据集上系统验证所提模型与主流基准方法的性能差异。

3.1 设计背景

尽管基于图神经网络的方法在推荐系统中取得了一定的成功,但现有的研究大多仅关注会话图中相邻项目之间的关系。传统方法往往仅收集紧密相连的相邻项目信息,并通过建模会话图中相邻项之间的依赖关系进行推荐。然而,这种方法在捕捉全局依赖关系时存在明显不足。尤其是在会话图中,项目节点之间可能并不直接相邻,传统的 GNN 难以有效捕捉这些远距离项目之间的潜在依赖关系,从而限制了模型的表达能力。此外,当前基于 GNN 的推荐系统还面临着流行偏差问题,推荐模型倾向于向用户推荐热门项目,而忽视了长尾项目^[68](即那些不太受欢迎或者频繁出现的项目)。这种偏向性会影响推荐系统的多样性与覆盖性,尤其是在处理用户兴趣差异较大、候选项目多样化的情况下,模型的性能往往受到影响。而且,待推荐项目通常是丰富的,用户的偏好通常是多元化的。以往的研究通过固定维度的向量代表用户的所有兴趣,难以捕捉用户兴趣的差异性,从而限制推荐系统的用户偏好表征能力。

针对上述问题提出全局增强的图神经网络会话推荐方法(Global Enhanced Graph Neural Network Embedding, GENE)。首先,为了增强处理流行偏差的问题,在训练和推理过程中,充分利用将项目和会话图表示限制在一个单位超球面上的优点,提出规范化项目和会话表示模型;此外,通过引入自注意力机制来加强图

神经网络。一方面，自注意力机制是通过明确关注所有位置来捕捉全局依赖。另一方面，GNN 可以捕捉相邻项之间丰富的局部依赖关系，提供丰富的局部上下文信息；最后，为克服仅依赖会话终端项目的固有局限，充分考虑到用户兴趣的动态演变特性，通过在建模过程中显式编码项目的时序位置信息。引入目标注意力机制来计算会话中所有项目相对于每个目标项目的软注意分数。经软注意力机制学习目标会话内每个项目对用户偏好的贡献度，加权目标会话中每个项目的会话级表示得到目标会话表示。在两个真实世界的数据集进行了广泛的实验。实证研究表明，GENE 在所有基线上都有显著的改进。

3.2 模型目标

本章分别用 $V = \{v_1, v_2, \dots, v_m\}$ 表示所有会话的物品集合，每个匿名会话 $S = \{v_{s,1}, v_{s,2}, \dots, v_{s,n}\}$ 由用户在特定时间范围内进行的一系列交互物品构成。具体来说， $v_{s,i} \in V$ 表示用户在会话 s 中的第 i 个交互项目，这些交互物品是按照时间顺序排列的，反映了用户在会话中的兴趣变化和行为轨迹。在会话推荐场景下，核心任务聚焦于建模用户当前交互序列的演进规律，以精准预测当前会话结束后的下一个目标交互对象。通过对所有潜在候选物品进行交互可能性排序，最终依据排序结果选取前 k 个项目生成推荐列表，以此实现用户行为的动态预测。

3.3 模型结构

为了增强基于图神经网络的会话推荐模型，本章综合考虑了流行度偏差和会话嵌入问题。提出全局增强的图神经网络会话推荐方法(GENE)，模型的整体框架如图 3-1 所示，首先将节点嵌入归一化为 $\tilde{V}$ ，然后将用户行为序列建模为会话图。在此基础上，应用图神经网络更新聚合会话图内的节点特征。之后，通过使用多层自注意力网络来捕捉会话中项目之间的全局协同偏好，同时目标感知注意力自适应地激活用户对不同待预测目标项的兴趣。在预测层，通过使用优化余弦相似度来计算每个候选项目的排序分数进行推荐。

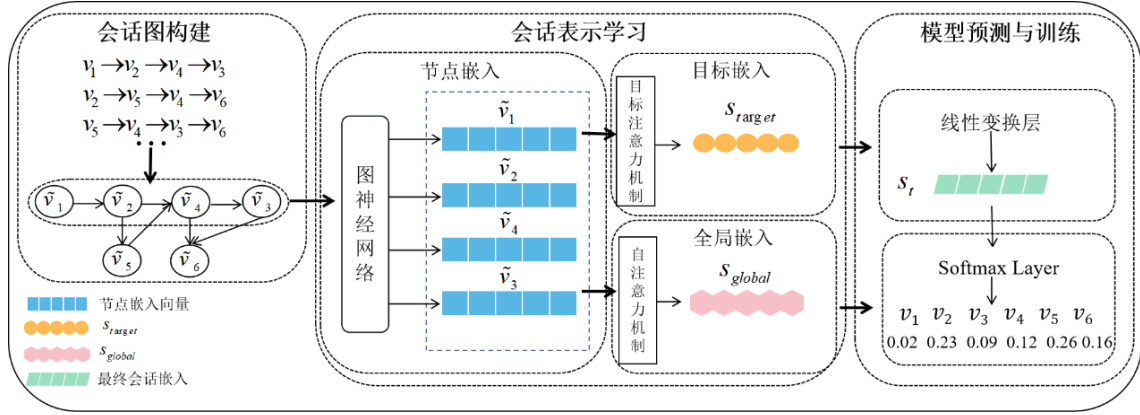


图 3-1 GENE 模型框架图

Fig. 3-1 The workflow of GENE model

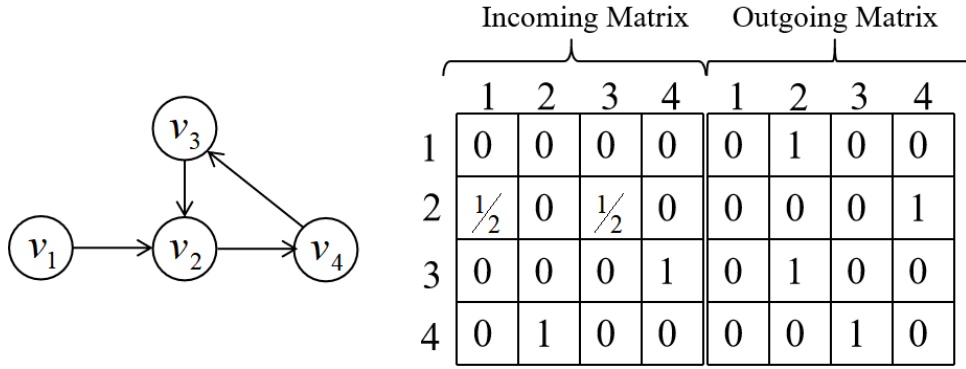


图 3-2 会话图构建示意图

Fig. 3-2 Schematic diagram of session diagram construction

3.3.1 会话图构建

每一个会话 S 可以被建模成图 $G_s = (V_s, E_s, A_s)$ ，其中 V_s 、 E_s 和 A_s 分别代表节点集，边集和邻接矩阵。在该会话图中，每个节点 $v_{s,i} \in V_s$ 表示一个物品，每一条边 $(v_{s,i}, v_{s,i+1}) \in E_s$ 代表一个用户在点击物品 $v_{s,i} \in V_s$ 后点击物品 $v_{s,i+1} \in V_s$ 。考虑到同一个物品可能多次出现在会话图中。因此在构建会话图中，为每条边引入归一化权重参数，该参数通过边出现频次与源节点出度的比值进行量化处理。这里定义邻接矩阵为两个出入邻接矩阵的串联如图 3-2 所示。

3.3.2 会话图表示

在流行偏差的背景下，受欢迎的项目可能会更频繁地被点击，这是由于他们在模型中更倾向于被推荐，因为模型是通过训练中调整权重参数等确保他们的向

量值更大来实现这一目的。这就导致热门商品在最后做排序计算的时候内积更大，排名更高。因此在热门商品较少时，基于图神经网络的会话推荐模型的性能也会下降。为了更好地解决流行偏见的问题，考虑将嵌入规范的影响降至最低，项嵌入被规范化为 $\tilde{\mathbf{v}}_k = \text{Dropout}(\text{Norm}(\mathbf{v}_k), p)$ ，其中，其中 p 是 dropout 概率， $\text{Norm}()$ 是 L_2 归一化函数。最终获得项集 $\tilde{\mathbf{V}}$ 以强制执行单位范数。图神经网络在处理会话推荐任务时具有优势，其核心机制在于能够显式建模会话交互图中的节点关联结构，通过消息传递范式自动聚合会话序列内的多阶邻域信息，从而捕捉用户行为轨迹中隐含的复杂交互模式。给定邻接矩阵 $\mathbf{A}_s$ ，以及规范化项嵌入 $\tilde{\mathbf{V}}$ 作为输入，对于图中的任何节点，使用节点的当前表示及其相邻节点的表示来迭代更新节点表示。具体而言，对于定向会话图中的节点，更新规则见公式 (2-4) - 公式 (2-8)。

在获得所有更新节点向量 $\tilde{\mathbf{V}}$ 之后，自注意力允许不同的注意力头关注不同位置的信息，以捕获会话中项目之间的全局偏好。将它们输入到自注意力层，第 i 个头注意力计算为：

$$\text{head}_i = \text{softmax}\left(\frac{(\tilde{\mathbf{V}}\mathbf{W}_i^Q)(\tilde{\mathbf{V}}\mathbf{W}_i^K)^T}{\sqrt{d}}\right)(\tilde{\mathbf{V}}\mathbf{W}_i^V) \quad (3-1)$$

其中， $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d(d/h)}$ 是参数矩阵， h 是注意力的头数。然后，将自注意力的 h 头与多头注意力相结合。多头注意力允许模型联合学习不同位置的依赖关系。

$$\mathbf{F} = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^o \quad (3-2)$$

其中， $\mathbf{W}^o$ 是投影矩阵。在此基础上， $\mathbf{F}$ 通过具有线性变换和激活函数的前馈网络如下所示：

$$\mathbf{Y} = \max(0, \mathbf{F}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 + \mathbf{F} \quad (3-3)$$

其中， $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n \in \mathbb{R}^{n \times d}]$ ， n 是会话长度， d 是项嵌入的向量维度。最后，对于每个会话 S ， $\mathbf{y}_n$ 是最后一个项目的潜在向量，自适应地聚合所有其他项目信息。因此，使用最后一项的潜在向量作为全局嵌入向量 $\mathbf{S}_{\text{global}} \in \mathbb{R}^d$ 。

在获取会话内各物品的节点表征后，模型通过目标注意力机制自适应地评估

目标项目与当前会话的关联强度，目标项目为所要预测的候选项目，具体而言，该机制通过计算会话中所有项目嵌入向量 $\tilde{\mathbf{v}}_i$ 和每个目标项目嵌入向量 $\tilde{\mathbf{v}}_t$ 之间的注意力分数，利用 **softmax** 函数进行归一化，其公式为：

$$\beta_{i,t} = \text{softmax}(e_{i,t}) = \frac{\exp(\tilde{\mathbf{v}}_t^T \mathbf{W} \tilde{\mathbf{v}}_i)}{\sum_{j=1}^m \exp(\tilde{\mathbf{v}}_t^T \mathbf{W} \tilde{\mathbf{v}}_j)} \quad (3-4)$$

最后，对于每个会话，目标嵌入 $\mathbf{S}_{target}^t \in \mathbb{R}^d$ ，如下所示：

$$\mathbf{S}_{target}^t = \sum_{i=1}^{s_n} \beta_{i,t} \tilde{\mathbf{v}}_i \quad (3-5)$$

3.3.3 模型预测与训练

由于用户的最终行为通常由用户的最后行为决定，因此简单地将用户的短期偏好表示为局部嵌入 $\mathbf{S}_{local} \in \mathbb{R}^d$ 。最后，通过对局部嵌入、全局嵌入和目标嵌入的级联执行线性变换来生成最终的会话嵌入 $\mathbf{S}_t = \mathbf{W}_3[\mathbf{S}_{global}; \mathbf{S}_{local}; \mathbf{S}_{target}^t]$ ，其中 $\mathbf{W}_3 \in \mathbb{R}^{d \times 3d}$ 为投影矩阵。在获得所有项目和会话嵌入后，计算每个目标项目的推荐分数：

$$\hat{\mathbf{y}} = \frac{\exp(\sigma \tilde{\mathbf{v}}_t^T \mathbf{S}_t)}{\sum_{i=1}^m \exp(\sigma \tilde{\mathbf{v}}_i^T \mathbf{S}_t)} \quad (3-6)$$

如文章^[63]所述，余弦相似性 $\tilde{\mathbf{v}}_t^T \mathbf{S}_t$ 的范围被限制为 $[-1,1]$ ，这表明训练中的 **softmax** 损失可能在高值下达到饱和，比例因子 $\sigma = 16$ 在实践中是有效的，允许更好的收敛。为了训练模型，使用交叉熵作为损失函数，公式^[34]为：

$$L(\hat{\mathbf{y}}) = -\sum_{i=1}^m \mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{y}}_i) \quad (3-7)$$

其中， $\mathbf{y}$ 表示地面的一个热编码向量。

3.4 实验与分析

为系统性验证本章所提的 GENE 模型的推荐性能，本研究在两个真实数据集上开展了多次实验。通过严谨的实证分析，重点探究 GENE 模型的性能表现和超参数的影响。

表 3-1 数据集详细描述

Table 3-1 The descriptive statistics of datasets

数据集	Diginetica	Nowplaying
点击数量	982961	1367963
训练集数量	719470	825304
测试集数量	68977	89824
项目总数	43097	60417
会话平均长度	5.12	7.42

3.4.1 实验设置

(1) 数据集。本章在两个真实世界的数据集上进行了实验：Nowplaying 和 Diginetica。Diginetica 是 2016 年 CIKM 挑战杯的数据集，实验中只使用了交易数据。Nowplaying 数据集包含用户的音乐收听行为。为了生成训练和测试集，最后几周的会话被用作 Diginetica 和 Nowplaying 的测试集。对于现有会话，生成一系列输入会话序列和相应的标签。筛选了测试集中没有出现在训练集中的项目。

(2) 评价指标。为了评估 GENE 模型的推荐性能，采用在会话推荐中广泛使用的度量，精确率(Precision@N)和平均倒数排名(MRR@N)，实验中 N 取 20，以评估所有模型的性能。Precision@20 表示推荐的准确性，即正确推荐的物品在前 N 个物品中的比例，计算公式如下：

$$P@N = n_{hit}/n \quad (3-8)$$

其中， n 为测试集样本总数， n_{hit} 为前 N 个物品中含有正确推荐的样本数。

MRR@20 表示正确推荐项目的平均倒数，其中值越高表示推荐质量越好，计算公式为：

$$MRR@N = \frac{1}{n} \sum_{i \in R} \frac{1}{rank(i)} \quad (3-9)$$

其中， n 为测试集样本总数， $rank(i)$ 表示第 i 个项目在推荐列表中的排名。

(3) 对比方法。通过构建对比实验，将所提模型与下列基准方法进行性能

对比分析，以客观验证性能优势。

FMPC^[21]：该模型结合矩阵分解和马尔可夫链进行推荐，可以捕捉用户的长期偏好和项目到项目的转换。

Item-KNN^[18]：该模型是通过计算会话项目之间的相似性来推荐的。

GRU4Rec^[26]：该模型首次将递归神经网络应用于基于会话的推荐领域，对用户行为序列进行建模。

NARM^[27]：将自注意力机制应用到序列建模中，根据用户过去与项目的交互序列识别出重要的行为模式，并基于这些模式预测用户的下一个交互项目。

SR-GNN^[34]：该模型通过将会话序列建模为会话图来捕捉项目之间的复杂转换，使用 GNN 生成项目的潜在向量，然后通过传统的注意力网络表示每个会话。

GC-SAN^[37]：该模型利用自注意机制来学习会话中非相邻项之间的全局依赖性信息。

TAGNN^[39]：该模型提出了一种新的目标注意力图神经网络，通过目标感知注意力机制获得目标嵌入，自适应地激活用户对不同目标项目的兴趣。

NISER+^[63]：该模型通过嵌入归一化来解决 GNN 模型中的普遍性偏差问题，并将位置信息引入到模型中。

Disen-GNN^[9]：该模型首先利用解纠缠学习技术将项目嵌入转化为多因素嵌入，并利用注意力机制来学习用户对会话中每个项目的不同因素的意图。

(4) 参数设置。GENE 模型及其所有对比方法均基于 Pytorch 实现，使用 python 进行开发。实验涉及的所有参数均通过网格实验选取。为了确保实验结果的公平性和公正性，按照前面提到的基线方法，使用平均值为 0、标准偏差为 0.1 的高斯分布对所有参数进行初始化。使用 ADAM 优化器对其进行优化。根据随机的 10% 验证集调整其他超参数。ADAM 的初始学习率设置为 0.001，每 3 个训练周期衰减 0.1。数据集的批量大小设置为 50，L₂ 惩罚系数设置为 10⁻⁵，迭代次数为 30。

3.4.2 性能比较

为了验证 GENE 的优越性，本章与较先进的模型进行了比较。从表 3-2 中可以看出：

(1) GRU4Rec 模型和 NARM 模型的推荐性能优于传统模型 FPMC 和 Item-KNN。前两个模型都使用 RNN 来捕捉用户偏好, 这表明忽略序列依赖性次优的, 验证了深度学习技术在基于会话的推荐中的有效性。NARM 的性能优于 GRU4Rec 表明注意力可以更准确地捕捉用户的主要兴趣和偏好, 将注意力机制引入推荐系统有助于提升推荐系统的性能。

(2) SR-GNN、GC-SAN、TAGNN 和 NISER+模型的推荐性能优于 GRU4Rec 和 NARM 模型, 表明了图神经网络在推荐系统中的有效性。基于图神经网络的会话推荐模型可以通过使用 GNN 来捕获会话序列中项目之间更复杂的转换信息。GC-SAN 在 SR-GNN 的基础上使用自注意机制来更准确地表示全局偏好。TAGNN 在 SR-GNN 的基础上, 引入目标感知注意力模块, 动态计算用户对候选目标的兴趣激活程度, 生成目标感知的嵌入向量, 提升了模型的个性化表达能力。NISER+通过规范化基于 SR-GNN 的项目和会话表示来缓解流行偏差问题, 同样提高了模型性能。Disen-GNN 模型通过解耦的方式, 将图数据中的信息分解为多个独立的子空间或潜在因素。方便更好地理解图数据中的不同因素, 并提高对噪声的鲁棒性。

(3) 可以看到, 在所有方法中, GENE 在 Precision 和 MRR 方面取得了最好的性能。结果证明 GENE 在会话推荐中的合理性。

表 3-2 模型性能比较实验结果

Table 3-2 Comparison of experimental results in model performance

Method	Diginetica		Nowplaying	
	Precision @20	MRR@20	Precision @20	MRR@20
FPMC	22.14%	6.66%	7.36%	2.82%
Item-KNN	35.75%	11.57%	15.94%	4.91%
GRU4Rec	29.45%	8.33%	7.92%	4.48%
NARM	49.70%	16.17%	18.59%	6.93%
SR-GNN	50.73%	17.59%	18.87%	7.47%
GC-SAN	51.70%	17.61%	18.85%	7.43%
TAGNN	51.31%	18.03%	19.02%	7.82%
NISER+	53.39%	18.72%	22.35%	8.52%
Disen-GNN	53.79%	18.99%	22.22%	8.22%
GENE	53.83%	19.21%	22.56%	8.94%

表 3-3 不同模块对算法性能的影响

Table 3-3 The impact of different modules on algorithm performance

Method	Diginetica		Nowpalying	
	Precision @20	MRR@20	Precision @20	MRR@20
GENE _{nN}	53.34%	18.29%	21.63%	8.12%
GENE _{nS}	53.45%	18.18%	21.52%	7.46%
GENE _{nT}	52.78%	18.25%	20.93%	8.16%
GENE	53.83%	19.21%	22.56%	8.94%

3.4.3 消融实验分析

在本节中，对话嵌入策略进行了消融研究。在保持所有其他条件相同的情况下，设计了四个模型变体来分析模型性能的影响：GENE_{nN} 模型，它去除了项目的归一化步骤；GENE_{nS} 模型，它去除了用于获得全局嵌入的自注意机制；GENE_{nT} 模型，它消除了用于获得目标嵌入的目标感知注意力。结果如表 3-3 所示。GENE 模型的性能优于 GENE_{nN} 模型，表明通过归一化嵌入处理流行度偏差对于提高基于 GNN 会话推荐的性能是有效的；为了进一步分析全局会话嵌入和目标会话嵌入对模型的影响，将 GENE_{nS} 和 GENE_{nT} 模型变体与 GENE 模型进行比较，结果表明多头注意力机制可以帮助从会话数据中提取重要的行为信息，并且目标注意力机制可以更好地完成目标项目的相关性推荐。总体而言，GENE 模型在两个数据集上都取得了最佳性能，这表明综合考虑归一化嵌入和自注意机制来建模全局依赖以及目标感知注意力的必要性和合理性。

3.4.4 超参实验分析

为了公平起见，除了测量的参数外，保持其他参数不变。使用接近测试时间 10% 的训练集作为超参数分析的保留验证集。发现 $\sigma = 16$ 在每个保留验证集上训练的大多数模型中效果最好，因此为了一致性，在数据集上使用相同的值。多头注意力使模型能够同时捕捉序列中不同位置的信息交互。然而，当注意力头的数量过度增加时，会导致单个注意力子空间的维度被压缩，进而削弱注意力机制对复杂依赖关系的建模能力。在表 3-4 中，Nowplaying 的最佳注意力头数为 4，Diginetica 的最佳注意力头数为 2。随着参数头的增加，推荐任务的性能先增加后降低。表明当注意力头数大于或小于最佳值时，性能略有下降。

表 3-4 自注意力头数的性能影响

Table 3-4 The impact of the number of self-attention heads on algorithm performance

Head	Diginetica		Nowpalying	
	Precision @20	MRR@20	Precision @20	MRR@20
1	53.62%	18.94%	21.69%	8.22%
2	53.83%	19.21%	21.75%	8.58%
4	52.63%	18.86%	22.56%	8.94%
6	53.17%	18.45%	22.42%	8.87%
8	53.23%	18.57%	21.99%	8.27%
10	53.01%	18.26%	21.78%	7.96%
12	53.16%	18.13%	21.49%	7.94%

3.5 本章小结

本章提出了一种在基于会话的推荐模型中学习更为丰富表示的方法，旨在克服图神经网络在消息传递过程中的局限性。传统的图神经网络在推荐系统中的应用往往受到消息传递机制的影响，导致在处理复杂用户行为时难以提取到足够的深层次信息。为了解决这一问题，提出一种结合注意力机制的图神经网络，能够更智能地进行特征提取，从而提升推荐系统的表现。具体来说，通过在模型中引入目标注意力模块，有效地捕捉用户与项目之间的复杂关系。与此同时，还通过多头注意力机制来学习全局项目依赖关系，并充分挖掘用户的整体兴趣偏好。这种方法能够更准确地表示用户在会话中的行为模式，从而为个性化推荐提供更加细致的依据。此外，通过标准化项目和会话图表示，缓解了传统 GNN 模型在处理流行度偏差时所面临的挑战。流行度偏差往往使得推荐系统过度偏向热门项目，忽略了冷门但可能更符合用户兴趣的项目。通过这种方法，能够更好地平衡热门与冷门项目之间的权重，进而显著提升推荐系统的准确性和多样性。在多个数据集上的实验结果表明，所提出的方法在现有最先进的推荐模型基础上取得了显著的性能提升。

第 4 章 基于上下文增强的对比学习会话推荐

虽然运用图神经网络和注意力机制算法来进行会话表示学习并通过项目规范化缓解流行度偏差问题,但是未考虑图神经网络忽略了非相邻项目之间复杂的高阶关系,并引入了噪声信息。因此,本章提出了基于上下文增强的对比学习会话推荐方法来提高推荐模型的预测性能。4.1 节剖析当前会话推荐方法的局限性,并针对性地提出改进策略;4.2 节阐述模型的基础定义,同时介绍会话推荐的目标;4.3 节基于模块化设计思想,系统构建了会话推荐模型的整体架构。该架构包含会话图构建、会话表示学习以及模型预测与学习三个核心模块组成;4.4 节通过构建对比实验,在公开数据集上系统验证所提模型与主流基准方法的性能差异。

4.1 设计背景

目前,会话推荐系统的最新进展主要集中在深度学习方法的应用上,尤其是图神经网络在该领域的广泛使用。许多研究采用图结构对用户的会话进行建模,并通过图神经网络在相邻项目之间传播信息,从而捕捉项目间的关联性。这种方法能够有效地提高推荐系统的性能,尤其是在建模用户短期兴趣和捕捉用户动态行为方面。然而,尽管基于 GNN 的方法在推荐系统中取得一定的成功,但它们依然面临着一些挑战,尤其是在处理用户行为数据中的噪声和复杂关系时。用户行为序列通常包含许多嘈杂的偏好信息,这些信息并不能准确反映用户的真实兴趣。例如,用户可能会因为好奇或误操作而点击一些与其实际兴趣无关的项目,这种交互行为通常不会在未来的会话中转化为真实的偏好选择。因此,这些无关的点击行为可能会被误认为是用户的兴趣,进而影响推荐系统对用户兴趣的准确建模。这些噪声信息在现有的推荐模型中,往往被直接传播和利用,进一步加剧了对用户真实兴趣的误解。此外,现有的基于图神经网络的推荐方法在构建全局图时,通常依赖于局部的邻接关系进行建模,这种方式在处理不相邻项之间的复杂关系时存在局限性,难以有效捕捉到项目间的高阶非相邻关系。事实上,许多情况下,用户的兴趣不仅仅体现在直接相邻项目之间的互动上,用户对某些项目的偏好可能通过间接的途径或跨越多个项目的关系进行体现。因此,这种局限性

使得推荐系统难以提供更加精确和个性化的推荐结果。

为了解决上述问题,本章提出一种基于上下文增强的对比学习方法(Context-augmented Contrastive Learning, CCL)用于会话推荐。首先,通过计算项目对之间的相似性来构造兴趣图,将松散的项目序列转换为紧密的项目-项目图,更容易区分用户的主要兴趣和次级兴趣。同时还为每个会话构建了一个当前会话兴趣图,通过对正在进行的会话中的成对项转换进行建模,可以获得会话级别的项嵌入;其次,通过为每个会话添加一个上下文向量来构建会话图,该上下文向量可以提供一种自然的方式来传递相邻项之外的信息;最后,为了获得更鲁棒的项目表示,生成 GSG 的增强图,并最大化原始图和增强图中同一会话的上下文向量的一致性来进行对比学习。该模型通过联合优化对比任务和推荐任务,显著提高了推荐性能。

4.2 模型目标

在会话推荐系统中,目标是通过用户历史行为和交互信息预测当前会话的下一点击物品。本章分别用 $V = \{v_1, v_2, \dots, v_m\}$ 和 $S = \{s_1, s_2, \dots, s_n\}$ 表示所有的会话物品集和所有的会话集,每个匿名会话 $S = \{v_{s,1}, v_{s,2}, \dots, v_{s,n}\}$ 由一系列按照时间顺序的交互物品组成。 $v_{s,i} \in V$ 表示用户在会话 s 中的第 i 个交互项目,在基于会话的推荐方法中,通过模型预测所有候选项目的相对重要性,为每个物品分配一个预测概率值。然后,根据这些预测概率,对所有候选物品进行排序,并生成一个排名列表,排名列表中前 k 个项目将会被选中进行推荐。

4.3 模型结构

模型的整体框架如图 4-1 所示:首先,通过采用基于度量的学习重构项目序列为紧密的项目兴趣图,在图中形成密集聚类来明确区分用户的核心兴趣。其次,通过为全局图中的每个会话添加上下文向量,该上下文向量聚合了会话中每个结点的信息,提供一种自然的方式来传递相邻项之外的信息。最后,设计了一个对比学习模块作为辅助任务,共同学习会话中项目的表示。它探索了兴趣图的构建、基于上下文向量的非相邻项目建模以及对比学习等内容。本章的体系结构分为四

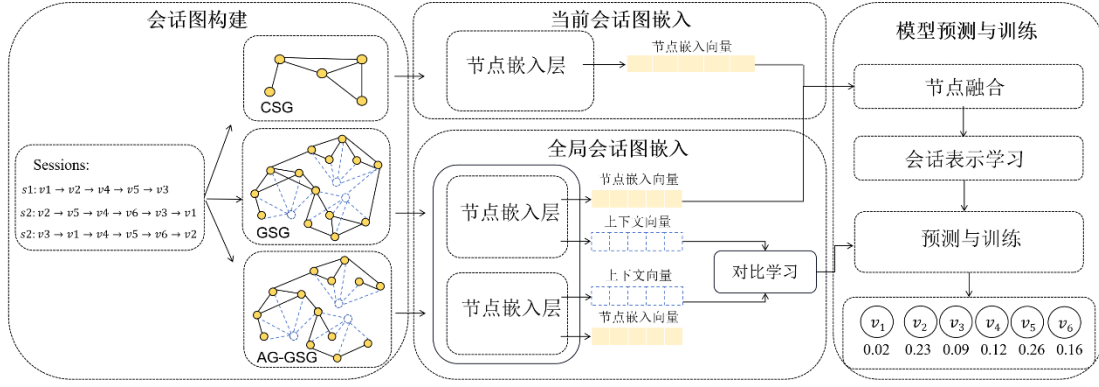


图 4-1 CCL 模型框架图

Fig.4-1 The workflow of CCL model

个部分：会话图构建、当前会话图嵌入、全局会话图嵌入以及模型预测和训练。

4.3.1 会话图构建

为了整合和区分用户丰富的历史行为中不同类型的偏好，可以通过将松散的项目序列转换为紧密的项目兴趣图。两个项目之间的同现关系是一个合理的构造标准，但挑战在于，同现关系的稀疏性不足以为每个用户生成连通图。因此，本章提出通过余弦相似度来计算项目对之间的相似度值 $\alpha_{ij} = \cos(\mathbf{W} \odot \mathbf{v}_i, \mathbf{W} \odot \mathbf{v}_j)$ ，为每个交互序列构建图结构，以探索其兴趣分布。考虑到噪声的引入，本章通过屏蔽那些小于非负阈值 α 的节点对来考虑最重要的连接节点对。核心兴趣节点由于连接了更多的相似兴趣而比外围兴趣节点具有更高的度，相似兴趣的频率越高，这种建模方法更容易区分用户的核心和次级兴趣。因此，本章通过计算项目对之间的相似度值来构建会话图。

考虑到仅依赖会话内的信息不足以全面捕捉用户偏好，本章提出从当前会话和跨会话两个方面构建兴趣图。对当前会话中的序列进行建模并构成的图定义为 $\text{CSG} = (\mathbf{V}_g, \mathbf{E}_g)$ 。和当前会话图建模方式类似，通过计算会话序列中项目对的相似度来判断两个项目之间是否有边的关系，同时，为增强会话结构的全局建模能力，为每个会话引入可学习的全局上下文嵌入，并将其称为上下文向量。该向量作为会话序列的整合表征，通过构建与会话内所有项目节点的双向全连接结构，形成信息交互通道。这种设计突破传统图结构的邻接约束，使得非直接相连的远端项目能够通过上下文向量实现隐式信息跃迁，从而为远程依赖关系的捕获提供结构支持。因此，对于用户的所有历史会话序列，全局会话图被定义为

$GSG = (V_g, E_g, C_g)$ ；其中 V_g 表示为所有会话的项目节点， E_g 表示通过计算相似度建立的边， C_g 表示为会话的上下文向量。同时，为了提高模型的鲁棒性，利用边缘丢弃策略来获得 GSG 的增强图 AG-GSG。

4.3.2 会话图嵌入

(1) 当前会话图嵌入

当前会话图 (CSG) 旨在通过对当前会话的建模来学习个性化项目嵌入。对于会话 s ，每个项目被嵌入到统一的嵌入空间中，并通过嵌入层为会话中的每个唯一项目生成 d 维嵌入，其中节点向量表示项目的潜在向量。图神经网络在考虑节点连接的情况下可以自动提取会话图的特征，因此，本章使用门控图神经网络来更新当前会话图 CSG 中嵌入的每个项目，公式如第二章 (2-4) 到 (2-8) 所示。

由于存在多层 GNN，容易发生拟合和过平滑。本章将模块最后一层的输出与初始输入聚合在一起，形成最终的项目表示。

$$g = \sigma(\mathbf{W}_s[\mathbf{v}_i^{s,0}; \mathbf{v}_i^{s,1}]) \quad (4-1)$$

$$\mathbf{v}_i^s = g\mathbf{v}_i^{s,0} + (1-g)\mathbf{v}_i^{s,1} \quad (4-2)$$

其中， $\mathbf{W}_s \in \mathbb{R}^{2d \times d}$ 表示可学习参数。

(2) 全局会话图嵌入

全局图 GSG 模块旨在通过会话的上下文向量对项目之间的复杂高阶关系进行建模，生成蕴含丰富协作信息的项目表示。本章将最后一个项目的表示指定为 $\mathbf{v}_m$ ，作为会话 s 的局部嵌入 $\mathbf{s}_l$ ，即 $\mathbf{s}_l = \mathbf{v}_m$ 。随后，使用软注意力机制来学习会话中的所有节点的优先级，并聚合节点向量，以生成全局偏好嵌入，表示为 $\mathbf{s}_g$ ，公式如下所示：

$$a_i = \mathbf{W}_0^T \sigma(\mathbf{W}_1 \mathbf{v}_m + \mathbf{W}_2 \mathbf{v}_i + \mathbf{b}) \quad (4-3)$$

$$\mathbf{s}_g = \sum_{i=1}^n a_i \mathbf{v}_i \quad (4-4)$$

其中, $\mathbf{W}_0 \in \mathbb{R}^d$, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times d}$ 是控制项目权重的可学习参数, $\mathbf{b} \in \mathbb{R}^d$ 是偏差向量。最后, 本章使用混合嵌入 $\mathbf{s}_h = \mathbf{W}_3[\mathbf{s}_i; \mathbf{s}_g]$ 作为相应的初始上下文向量 $\mathbf{c}^s$, 即 $\mathbf{c}^s = \mathbf{s}_h$ 。该上下文向量结合了会话的长期和短期兴趣偏好。

为了学习会话中非相邻项目之间的关系以及不同会话中项目之间的高阶关系, 需要迭代更新全局会话图上的项目嵌入和上下文向量。因此, 在信息传播过程中, 对于图中的每个节点, 信息都是从相邻项和上下文向量中获得的。首先, 忽略上下文向量, 按照当前会话图的处理方法处理全局会话图。对于图中的每个节点 v_i , 更新来自不同会话的相邻节点的表示。由于项目可能出现在多个会话中, 每个节点都可以连接到多个上下文向量。本章计算节点 v_i 和层 l 中的上下文向量 $\mathbf{c}_{i,j}$ 之间的相似度作为注意力权重, 并将它们线性组合。假设包含节点 v_i 的会话的上下文向量来自集合 $\mathbf{C}_i = [\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, \dots, \mathbf{c}_{i,n}]$, 其中 n 是包含节点 v_i 的会话数, 注意机制如下。

$$\alpha_{i,j}^l = \sigma\left(\frac{(\mathbf{W}_{q1}\mathbf{v}_i^{g,l})^T \mathbf{W}_{k1}\mathbf{c}_{i,j}^{l-1}}{\sqrt{d}}\right) \quad (4-5)$$

其中, $\mathbf{W}_{q1}, \mathbf{W}_{k1} \in \mathbb{R}^{d \times d}$, $\mathbf{v}_i^{g,l}$ 和 $\mathbf{c}_{i,j}^{l-1}$ 分别是节点 v_i 在层 l 处的表示和层 $l-1$ 处的上下文向量表示。并且通过对向量 $\mathbf{v}_i^{g,l}$ 和 $\mathbf{c}_{i,j}^l$ 执行非线性映射来计算级别优先级 β_i^l , 来平衡两个向量的重要性。

$$\mathbf{v}_i^{c,l} = \sum_{j=1}^m \alpha_{i,j}^l \mathbf{c}_j^l \quad (4-6)$$

$$\beta_i^l = \sigma(\mathbf{W}_4[\mathbf{v}_i^{g,l}; \mathbf{v}_i^{c,l}]) \quad (4-7)$$

其中, $\mathbf{v}_i^{g,l}$ 是从相邻项获得的表示, $\mathbf{v}_i^{c,l}$ 是从层 l 处的上下文向量获得的表示, $\mathbf{W}_4$ 是可学习的参数。然后, 整合来自相邻节点和相关上下文向量的信息。

$$\mathbf{v}_i^l = (1 - \beta_i^l) \mathbf{v}_i^{g,l} + \beta_i^l \mathbf{v}_i^{c,l} \quad (4-8)$$

其中, $\mathbf{v}_i^l$ 是层 l 处节点的表示。最后, 通过聚合最后一层的输出和初始输入, 以获得最终的项目表示 $\mathbf{v}_i^g$ 。

对于全局会话图中的每个上下文向量,其信息来自当前会话中的相应项目信息。由于相应项目的信息重要性有所不同,因此,本章首先为每个节点 v_i 分配不同的权重,如下所示。

$$\gamma_{j,i}^l = \text{softmax}\left(\frac{(\mathbf{W}_{k_2}\mathbf{v}_i^l)^T(\mathbf{W}_{q_2}\mathbf{c}_j^{l-1})}{\sqrt{d}}\right) \quad (4-9)$$

其中, $\mathbf{W}_{q_2}, \mathbf{W}_{k_2} \in \mathbb{R}^{d \times d}$ 是可训练的参数, 并且 $\gamma_{j,i}^l$ 表示第 i 项对第 l 层中的会话 j 的重要性。最后,线性地组合项目表示并聚合更新上下文向量 $\mathbf{c}_j^l$ 和 $l-1$ 层的表示。

$$\mathbf{c}_j^{l'} = \sum_{i=1}^n \gamma_{j,i}^l \mathbf{v}_i^l \quad (4-10)$$

$$\phi^l = \sigma(\mathbf{W}_5[\mathbf{c}_j^{l'}; \mathbf{c}_j^{l-1}]) \quad (4-11)$$

$$\mathbf{c}_j^l = \phi^l \mathbf{c}_j^{l-1} + (1 - \phi^l) \mathbf{c}_j^{l'} \quad (4-12)$$

其中, $\mathbf{W}_5 \in \mathbb{R}^{d \times 2d}$ 是一个可学习的参数。同样,使用高速公路网络将初始上下文向量与最后一层的输出相结合,以获得最终表示 $\mathbf{c}_j$ 。

(3) 对比学习

由于跨会话图生成阶段易引入冗余噪声信息,造成模型对异常数据的抵抗能力降低。因此,本章引入了一个自监督对比学习模块来辅助学习,通过边缘丢弃方法生成了 GSG 的数据增强图 AG-GSG,并将原始图 GSG 和增强图 AG-GSG 中的同一会话对视为正样本对,将两个图中的其他 $2(N-1)$ 个会话视为负样本对。对于每个会话对,由于每个会话的上下文向量可以视为当前会话的整体表示,因此,可以将从会话对获得的上下文向量视为一对正样本。本章使用两个图中上下文向量的 InfoNCE 损失作为下面定义的优化对象。

$$L_{ssl}(\mathbf{c}_n, \mathbf{c}_n^{aug}) = -\log \frac{\exp(\text{sim}(\mathbf{c}_n, \mathbf{c}_n^{aug})\tau)}{\sum_{m=1}^{2N} \exp(\text{sim}(\mathbf{c}_n, \mathbf{c}_m^{aug})/\tau)} \quad (4-13)$$

其中, $\text{sim}()$ 是相似函数, τ 是控制缩放的超参数。

4.3.3 模型预测与训练

(1) 模型预测

对于每个项目 v_j ，它们可以分别从 CSG 模块和 GSG 模块获得项目表示。因此，项目的最终表示是通过合并来自不同来源的信息而获得的。

$$\mathbf{v}'_j = \mathbf{v}_j^s + \mu \mathbf{v}_j^g \quad (4-14)$$

其中， μ 是一个超参数，用于控制从 CSG 模块学习项目表示的比率。同时，通过与获得初始会话表示相同的处理方法，使用软注意力聚合节点来获得最终会话表示。在获得所有项目和会话嵌入后，计算每个目标项目的推荐分数：

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{s}_i^T \mathbf{v}'_j) \quad (4-15)$$

(2) 模型训练

本章使用预测结果的交叉熵作为主要损失函数，定义如下：

$$L_{rec}(\hat{\mathbf{y}}) = -\sum_{i=1}^m \mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{y}}_i) \quad (4-16)$$

其中， $\mathbf{y}$ 表示目标会话中真实点击项目的独热编码向量。

然后，通过将推荐损失与对比学习损失相结合进行性能优化，如下所示：

$$L = L_{rec} + \lambda L_{ssl} \quad (4-17)$$

其中， λ 是控制对比学习损失率的超参数。

4.4 实验与分析

为系统性验证本章所提的 CCL 模型的推荐性能，本研究在两个真实数据集上开展了多次实验。通过严谨的实证分析，重点探究 CCL 模型的性能表现和核心超参数的影响。

4.4.1 实验设置

(1) 数据集。在两个基准数据集 Diginetica 和 Tmall 上进行了实验。Diginetica

表 4-1 数据集详细描述

Table 4-1 The descriptive statistics of datasets

数据集	Diginetica	Tmall
点击数量	982961	818479
训练集数量	719470	351268
测试集数量	68977	25898
项目总数	43097	40728
会话平均长度	5.12	6.69

来自 2016 年 CIKM 杯，包含从电子商务搜索引擎日志中提取的会话信息。Tmall 来自 IJCAI-15 竞赛官方数据集，包含天猫商城平台匿名用户在真实购物场景下的交易行为序列日志，数据集详细描述如表 4-1 所示。

(2) 对比方法。通过构建对比实验，将所提模型与下列基准方法进行性能对比分析，以客观验证性能优势。

FMPC^[21]：该模型通过矩阵分解提取用户和物品之间的潜在关系，同时利用马尔可夫链捕捉用户行为的转变模式，提升推荐系统对用户需求的理解和预测准确性。

GRU4Rec^[26]：该模型通过学习用户行为之间的时间顺序和依赖关系，从而预测用户未来可能感兴趣的物品。

NARM^[27]：该模型将 RNN 和注意力机制相结合运用到基于会话的推荐模型中，以捕获用户偏好。

SR-GNN^[34]：该模型首次将会话序列建模为会话图，并使用图神经网络来捕捉项目之间的复杂转换。

GCE-GNN^[38]：该模型通过构建当前会话图和全局会话图来捕获局部和全局信息。

S²-DHCN^[47]：该模型通过构造超图来学习会话间和会话内信息，并使用自监督学习来提供补充信息。

COTREC^[48]：该模型通过构建两个视图来捕获会话间和会话内信息，并使用协作训练策略迭代选择和演化伪标签作为信息自监督示例。

CORE^[64]: 该模型设计了一种编码器, 可以将输入项的线性组合嵌入为会话嵌入, 确保会话和项在相同的表示空间中。

MGIR^[65]: 该模型使用不同的聚合层来编码不同的关系, 并通过融合正关系和负关系来生成增强的会话表示。

MGS^[66]: 该模型利用两种类型的图来学习项目表示, 并应用迭代双重细化机制在会话图和镜像图之间传播信息。

(3) 参数设置。本章使用平均值为 0, 标准偏差为 0.1 的高斯分布来初始化参数。对于常规设置, 将嵌入大小和批处理大小分别设置为 256 和 1024。每个会话被截断, 最大长度为 20。使用 Adam 来优化模型。初始学习率为 $1e^{-3}$, 每三个 epoch 的衰减因子为 0.1, L_2 正则项为 10^{-5} 。

(4) 评价方法。实验采用分层抽样策略, 将各数据集按 8:1:1 的比例划分为训练集、验证集和测试集。模型性能通过会话推荐领域两个核心评估指标进行量化: Precision@k 和 MRR@k, 以评估推荐性能。Precision@K 用于衡量推荐结果的精准度, MRR@k 表示目标项目在推荐项目列表中的位置。更高的指标表示更好的排名准确性。

4.4.2 性能比较

为了验证 CCL 的优越性, 本章与较先进的模型进行了比较。从表 4-2 中可以看出:

(1) FPMC 模型的性能不如 GRU4Reg 和 NARM。这两个模型都使用 RNN 对用户序列进行建模并捕获用户偏好, 这表明序列相关性至关重要。NARM 的性能优于 GRU4Rec, 表明注意力机制可以更准确地建模用户的主要兴趣。

(2) 基于图神经网络模型的性能优于 NARM 和 GRU4Reg 等传统模型, 证明了图神经网络在数据建模中的有效性, 因为它们可以捕获更复杂项目之间的转换信息。此外, GCE-GNN、S²-DHCN、COTREC、MGIR、MGS 等其他模型通过构建两种类型的会话图来捕获本地和全局信息, 以获得更准确的兴趣偏好。表明仅考虑会话内信息来捕捉兴趣偏好是次优的。

表 4-2 模型性能比较实验结果

Table 4-2 Comparison of model performance experimental results

Method	Diginetica		Tmall	
	Precision @20	MRR@20	Precision @20	MRR@20
FMPC	22.14%	6.66%	16.06%	7.32%
GRU4Rec	30.79%	8.22%	10.93%	5.89%
NARM	48.32%	16.00%	23.30%	10.70%
SR-GNN	51.26%	17.78%	27.57%	13.72%
GCE-GNN	54.22%	19.04%	33.42%	15.42%
S ² -DHCN	53.66%	18.51%	31.72%	15.05%
COTREC	54.18%	19.07%	36.35%	18.04%
CORE	53.73%	18.77%	36.31%	17.42%
MGIR	52.84%	18.47%	39.16%	19.52%
MGS	55.02%	19.14%	41.55%	18.39%
CCL	55.42%	19.32%	41.94%	21.37%

(3) S²-DHCN 和 COTREC 等模型使用自监督模块通过构建对比图来捕获更复杂项目之间的转换信息，证明了对比学习在推荐系统中的有效性。

(4) 可以看到，在所有方法中，CCL 在 Precision 和 MRR 方面取得了最佳性能。这些结果证明 CCL 在基于对话的推荐中的合理性和有效性。

4.4.3 消融实验

为了证明本章解决方案的合理性，通过对 CCL 进行消融研究以验证核心组件的有效性，结果如表 4-3 所示。

在本节中，在保持所有其他条件不变的情况下，设计了四个模型变体来分析对模型性能（RQ2）的影响：模型 CCL_{ns}，它删除了通过相似性构建的兴趣图模块；模型 CCL_{nsD}，它删除了上下文向量和对比学习模块；模型 CCL_{nd}，它删除了对比学习模块。结果显示，CCL 模型的性能优于 CCL_{ns} 模型，表明通过相似性构建兴趣图可以更好地消除噪声的影响，获得用户兴趣偏好，有效地提高了基于 GNN 的会话推荐的性能；将 CCL_{nsD} 和 CCL_{nd} 模型变体与 CCL 模型进行比较，表明上下文向量有助于从会话数据中提取非相邻项之间的重要行为信息，并且通过对比学习可以更好地增强会话数据，从而获得更好的推荐结果。总体而言，CCL 模型在两个数据集上都取得了最佳性能，这表明通过相似性构图以及引入

上下文向量的必要性和合理性。

4.4.4 超参实验分析

在本节中，研究了三个重要超参数的影响。为了公平起见，除了测量的参数外，保持所有其他参数不变。通过控制自监督学习比率的数值变化来研究其超参数 λ ，该值的性能如图 4-2 所示。实验表明当 λ 分别为 0.3 和 0.2 时，推荐任务与对比学习任务联合优化可以对会话图进行更好的数据增强，从而使推荐任务具有良好的性能和鲁棒性。为了研究两个模块的比率的影响，测试了一组具有代表性的 μ 值。从图 4-3 的结果可以看出，当 Diginetica 和 Tmall 数据集中的比率 μ 分别设置为 0.1 和 0.6 时，可以更好地聚合从全局和会话级别学习到的项目嵌入，以提高当前会话的推荐性能。同时，为了研究在构建兴趣图时相似性阈值大小对性能的影响，分别在两个数据集上进行了测试。该值的性能如图 4-4 所示。发现 Diginetica 数据集和 Tmall 数据集分别在 $\alpha = 0.2$ 和 $\alpha = 0.3$ 时达到了最佳性能。

表 4-3 不同模块对算法性能的影响

Table 4-3 The impact of different modules on algorithm performance

Method	Diginetica		Tmall	
	Precision @20	MRR@20	Precision @20	MRR@20
CCL _{ns}	55.24%	19.02%	41.42%	20.89%
CCL _{nsD}	54.48%	18.77%	39.75%	19.91%
CCL _{nsD}	55.26%	18.95%	41.63%	21.04%
CCL	55.42%	19.32%	41.94%	21.37%

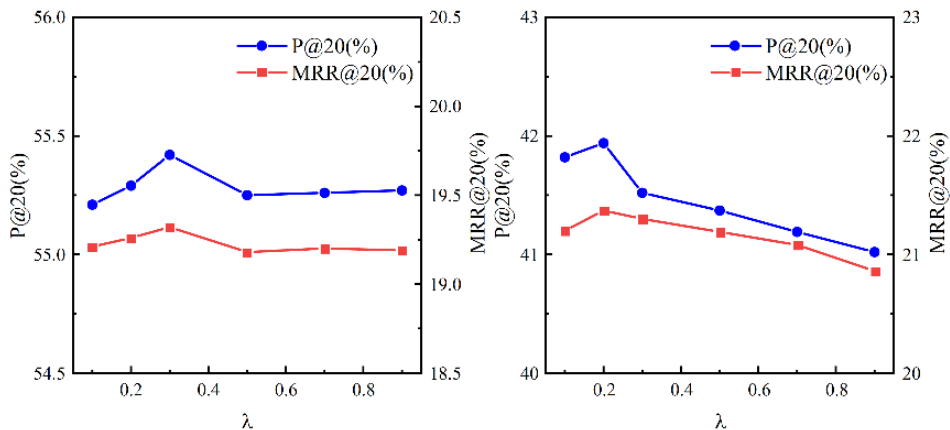


图 4-2 对比学习对算法性能的影响。

Fig.4-2 The impact of contrastive learning on algorithm performance

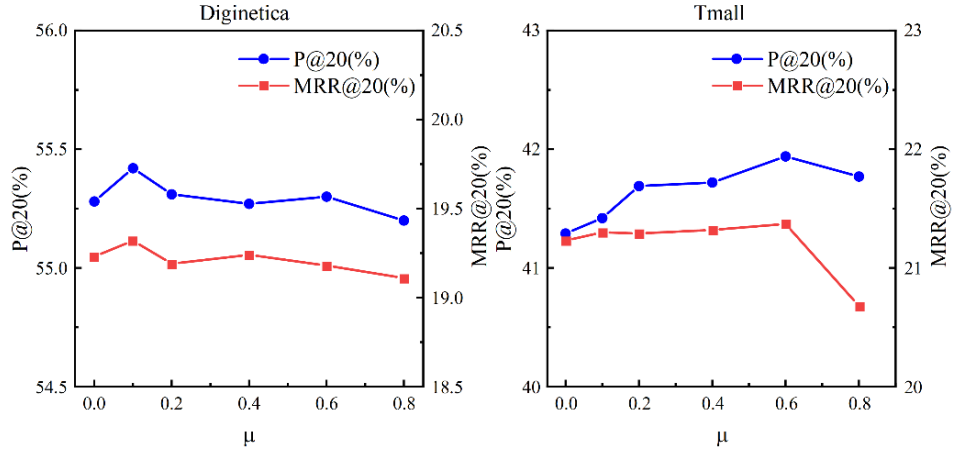


图 4-3 当前会话图模块比率对算法性能的影响

Fig.4-3 The impact of similarity threshold on algorithm performance

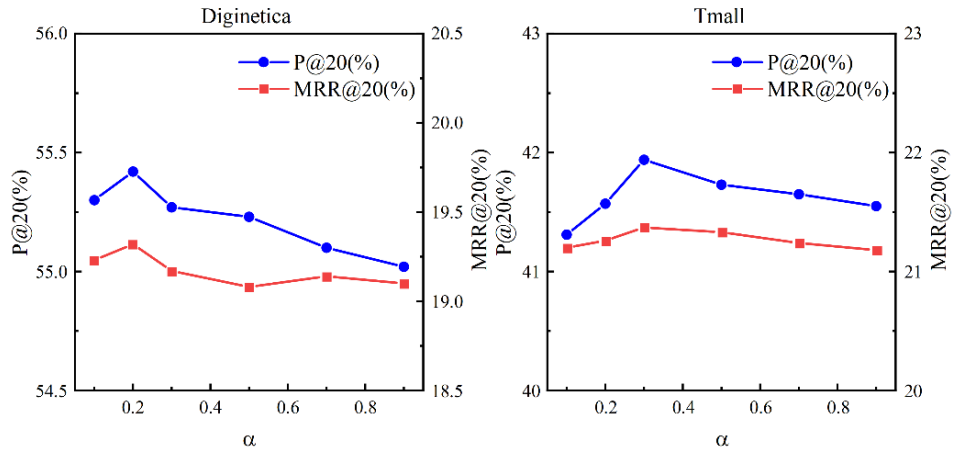


图 4-4 相似度阈值对算法性能的影响

Fig.4-4 The impact of CSG module ratio on algorithm performance

4.5 本章小结

本章提出了一种在基于会话的推荐模型中学习更丰富表示的方法，旨在克服传统图神经网络在消息传递过程中的局限性。在该模型中，通过计算项目对之间的相似性来构建会话兴趣图，同时为全局会话图中的每个会话添加上下文向量，作为不同会话中的非相邻项之间传递信息的桥梁。这一创新设计能够有效捕捉会话历史信息，并为推荐任务提供更强的语境理解，从而增强模型对复杂用户行为模式的适应能力。为了进一步提升模型的鲁棒性与准确性，引入了对比学习模块。该模块通过与推荐任务联合优化对比任务，促进了模型在学习过程中更好地对比不同项目间的特征差异，从而帮助模型生成更加精确且具有判别力的项目表示。

通过这种联合优化策略，模型能够从多个维度提升对用户需求的理解，显著提高了推荐的性能。本章已经在多个广泛应用的推荐场景中进行了大量实验，实证结果表明，所提出的 CCL 模型在多项基准数据集上均显著优于几种当前最先进的推荐方法，证明了其有效性和优越性。

第5章 基于多意图感知增强的会话推荐

虽然通过添加上下文向量和引入对比学习来缓解噪声数据问题。但在真实推荐场景中用户不一定只存在单一的目的意图，忽略会话中可能存在的各种用户意图会导致推荐性能显著下降，故本章提出了基于多意图感知增强的会话推荐方法来提高推荐模型的预测性能。5.1 节剖析当前会话推荐方法的局限性，并针对性地提出改进策略；5.2 节阐述模型的基础定义，同时介绍会话推荐的目标；5.3 节基于模块化设计思想，系统构建了会话推荐模型的整体架构。该架构由会话图嵌入模块、多意图提取模块和模型预测与学习模块这三大模块组成；5.4 节通过构建对比实验，在公开数据集上系统验证所提模型与主流基准方法的性能差异。

5.1 设计背景

目前，基于会话的推荐系统已经取得了显著的进展，尤其是在深度学习的驱动下。许多先进的方法使用图结构来对会话进行建模，并利用图神经网络在项目之间传播信息，这为推荐系统提供了更好的上下文理解能力和信息传递效果。然而，这些方法大多聚焦于会话的最后一项，或独立地考虑会话中的每个项目，往往忽视会话的整体语境和高层次的语义理解。例如，在许多情况下，最后一项的选择可能并不代表用户真正的意图，或者会话中的某些项目可能是无关的噪声，这使得这些方法在捕捉用户意图方面存在局限。此外，用户的意图在一个会话中通常是多样化和动态变化的。在许多实际场景中，用户的需求并非简单线性的。例如，当用户在为旅行做准备时，他们的需求可能同时涉及时尚、美妆、防晒等多个方面，而这些需求之间并非孤立存在，而是有一定的内在联系。因此，单一地聚焦于最后一项或将会话中的每个项目视为独立的单元，无法全面而精确地捕捉到多样化的用户意图。为了更好地应对这些挑战，未来的会话推荐系统亟需进行更复杂的建模，能够同时识别和理解多个潜在的用户意图，并在不同的意图之间进行有效的协调和融合。这要求推荐系统不仅仅依赖于图神经网络在局部项目之间的传播，还需要能够捕捉会话全局的信息，识别出会话中的多重意图，并通过更精细的语义分析来推断用户的真实需求。通过这种方式，推荐系统能够更准

确地预测用户的兴趣，提高推荐的相关性和精准度，从而显著提升用户体验。

为了解决上述问题，提出了一种基于多意图感知增强的会话推荐模型(Multi-intent-Aware Enhancement, MIAE)。首先，通过使用项目节点的重复频率作为注意力权重来更新项目嵌入，以增强重复项目的信息，降低对这些稀疏、噪声数据的敏感度，更好地理解 and 利用用户行为的序列特征。其次，采用自注意机制和高速网络来获取会话中不同的用户意图并自适应地提取会话中重要的用户意图。最后，将多个会话表示解码为项目分布，并使用池函数将其聚合从而进行推荐。模型显著提高了推荐性能，增强了模型对交互噪声的鲁棒性。

5.2 模型目标

在本章中，用 $V = \{v_1, v_2, \dots, v_m\}$ 来表示所有的会话物品集，其中，每个物品 v_i 代表一个候选物品，每个匿名会话 $S = \{v_{s,1}, v_{s,2}, \dots, v_{s,n}\}$ 由一系列按照时间顺序的交互物品组成。 $v_{s,i} \in V$ 表示用户在会话 s 中的第 i 个交互项目，在基于会话的推荐模型中，目标是根据用户在当前会话中的历史行为，推测其可能的下一个兴趣物品。为此，为所有候选项目生成一个概率排名列表，并根据模型对每个候选物品的兴趣度预测，排序所有项目。在此排名列表中，排名前 k 的项目将被选中作为推荐物品，展示给用户。通过这种方式，不仅能够捕捉到用户当前的即时需求，还能依据其历史行为模式，动态调整推荐策略，从而提高推荐的准确性和相关性。

5.3 模型结构

模型的整体框架如图 5-1 所示：首先，通过使用项目节点的重复频率作为注意力权重来更新项目嵌入，以增强重复项目的信息。其次，通过派生以每个项目为中心的多个会话表示并动态选择重要的会话表示来表明各种用户意图。最后，将多个会话表示解码为项目分布，并使用池函数将其聚合。通过利用增强的信息和多样化的意图表示，实现更为精准和全面的推荐结果。本章的体系结构分为三个部分：节点表示学习、多意图选择和模型预测与优化模块。

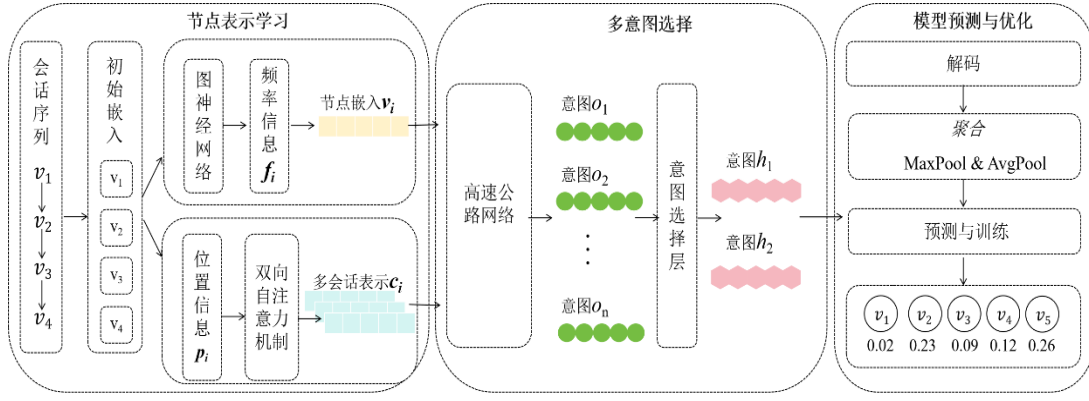


图 5-1 MIAE 模型框架图

Fig. 5-1 The workflow of MIAE model

5.3.1 节点表示学习

(1) 节点嵌入

首先将每个会话项 v_i 嵌入到统一的嵌入空间中，并通过嵌入层为会话中的每个唯一项生成 d 维嵌入，其中节点向量表示项目的潜在向量。图神经网络通过聚合邻域信息，能够有效建模会话图中节点间的高阶交互模式。为增强对复杂会话结构的解析能力，本研究利用 GGNN 特有的门控机制动态调控消息传递过程，通过多轮迭代优化各节点表示。在给定邻接矩阵和项目嵌入向量的情况下迭代更新项目节点向量，节点嵌入公式如第二章（2-4）到（2-8）所示。

为了更好地描述会话中每个项目的重要性，引入了位置嵌入向量 p_i 来区分序列的顺序，同时，引入了一种基于项目序列中重复频率的注意力机制，用于动态调整项目嵌入的表示。首先，每个项目在序列中的重复频率向量表示为 $F = [f_1, f_2, \dots, f_L]$ 为，其中 f_i 表示序列中第 i 个项目的重复频率。对于每个项目 v_i ，重复频率表示在序列中与之相同的项目的数量，通过以下公式计算注意力权重 a_i 来调整项目嵌入。

$$a_i = \frac{f_i}{\sum_{i=1}^L f_i} \quad (5-1)$$

其中， a_i 表示项目 v_i 的注意力权重， L 是序列的长度。将注意力权重应用于项目的嵌入表示，以动态调整嵌入向量 $\mathbf{v} = \mathbf{v}_i \odot \mathbf{a}_i$ 。这样，更频繁出现的项目在嵌入表示中将得到更大的权重，从而更显著地影响模型的学习和推理过程。

(2) 多会话表示

每个项目在序列中可能与多个其它项目相关联,并且这些关系可能在不同的方向上(前向和后向)具有不同的重要性,双向自注意力机制通过计算两个方向上的注意力权重来更全面地捕捉序列中项目之间的复杂关系和语义信息。因此本章采用双向自注意机制来捕捉会话项目之间的复杂关系,自适应生成与每个项目相关联的会话信息的多个上下文表示,公式如下所示:

$$\mathbf{c}_i = \sum_{j=1}^n \text{soft max}_j \left(\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}} \right) \mathbf{v}_j \quad (5-2)$$

$$\mathbf{c} = \mathbf{W}_q \mathbf{c}_i^q + \mathbf{W}_h \mathbf{c}_i^h \quad (5-3)$$

其中, $\mathbf{W}_q$ 和 $\mathbf{W}_h$ 为权重矩阵,用于对前向和后向上下文表示进行加权组合。

5.3.2 多意图选择

高速公路网络是一种深度神经网络结构,能够有效地整合输入的多层特征信息。在推荐系统中,会话信息对于理解用户的整体行为模式至关重要,而每个项目的具体信息则有助于捕捉每个项目的个性化特征和重要性。因此为了确保多个上下文表示不会变得相似,并更好地反映不同的用户意图,本章通过高速公路网络来强调项目嵌入。具体地说,将上下文表示 $\mathbf{c}$ 和项目嵌入 $\mathbf{v}$ 相结合,得到既考虑会话整体信息又考虑每个项目信息的用户意图表示 $\mathbf{o}$ 。

$$\mathbf{o}_i = g \odot \mathbf{v}_i + (1-g) \odot \mathbf{c}_i \quad (5-4)$$

$$g = \sigma(\mathbf{W}_g [\mathbf{v}_i; \mathbf{c}_i]^T) \quad (5-5)$$

其中, $\mathbf{W}_g$ 为权重矩阵, g 是门控向量, $\sigma()$ 是 sigmoid 函数。

本章采用多个会话表示,以充分利用每个会话项的潜力。然而,并非所有会话项都是必要的,并且一些会话项可能是嘈杂的。为了在会话中提取重要的用户意图,通过计算多个表示的重要性并删除不重要的表示。同时利用稀疏变换矩阵 $\alpha - \text{entmax}$, 将零概率分配给不重要的表示。

$$\alpha - \text{entmax}(\mathbf{z}) = \arg\max \mathbf{p}^T \mathbf{z} + H_\alpha^T(\mathbf{p}) \quad (5-6)$$

如果 $\alpha \neq 1$ ，则 $H_\alpha^T(\mathbf{p}) = \frac{1}{\alpha(\alpha-1)} \sum_j (\mathbf{p}_j - \mathbf{p}_j^\alpha)$ ，否则 $H_\alpha^T(\mathbf{p}) = -\sum_j \mathbf{p}_j \log \mathbf{p}_j$ 。根据不同的情况，当 $\alpha = 1$ 时 entmax 等价于 softmax 函数，生成接近均匀分布的概率。当 $\alpha = 2$ 时 entmax 等价于 sparsemax 函数，生成更加稀疏的概率分布，更多地倾向于产生零概率。根据经验将 α 设置为 1.5。通过屏蔽不必要的用户意图来提取会话表示 $\mathbf{h}_i$ 。

$$\gamma = \alpha - \text{entmax}(\mathbf{W}_h[\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_m]^T) \quad (5-7)$$

$$\mathbf{h}_i = \{\gamma_i \mathbf{o}_i \mid \gamma > 0, i \in \{1, \dots, |s|, m\}\} \quad (5-8)$$

其中， $\mathbf{W}_h$ 是一个可学习的参数， γ 表示每个项目的重要性权重。本章通过稀疏变换矩阵和阈值设置，可以有效地减少嘈杂或不相关的信息，从而提升模型的精确度和效率。

5.3.3 多意图聚合

对于多意图聚合过程主要分为两个部分：从多个会话表示中解码项目分布和聚合分布以获得最终推荐。

(1) 解码

在解码每个会话表示为项目分布时，利用余弦相似度结合项目嵌入矩阵来实现，即，带有 L_2 归一化的点积。为了简化模型，重新利用了预先训练好的项目嵌入矩阵 $\mathbf{V}$ ，在无需增加额外的模型参数下，同时能够利用这些嵌入向量来表示项目的特征。

$$\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_k = [\mathbf{h}_1^g \mathbf{V}_g^T, \dots, \mathbf{h}_k^g \mathbf{V}_g^T] \quad (5-9)$$

其中， $\mathbf{h}_i^g$ 和 $\mathbf{V}_g$ 分别是归一化会话向量和归一化项嵌入矩阵。 $\hat{\mathbf{y}}_i$ 表示由会话向量 $\mathbf{h}$ 解码的项分布。

(2) 聚合

为了聚合多个项目分布，采用最大池化和平均池化函数。Max-pooling 保留了多个意图的主要特征，而 Average-pooling 则捕捉了会话中一致的意图。结合最大池化和平均池化函数，能够充分利用它们各自的优势，既能突出显著信息又能捕捉一致的用户意图，从而提高推荐系统的准确性。

$$\hat{\mathbf{y}} = \beta \text{MaxPool}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_k) + (1 - \beta) \text{AvgPool}(\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_k) \quad (5-10)$$

其中， $\hat{\mathbf{y}}$ 是最终的聚合项分布，而 β 是组合超参数。MaxPool() 和 AvgPool() 分别代表从多个向量中提取每个维度的最大（或平均）值的最大池化（平均池化）函数。

本章使用预测结果的交叉熵作为损失函数，定义如下：

$$L(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{i=1}^m \mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1 - \mathbf{y}_i) \log(1 - \hat{\mathbf{y}}_i) \quad (5-11)$$

其中， $\mathbf{y}_i$ 表示地面的一个热编码向量。

5.4 实验与分析

为系统性验证本章所提的 MIAE 模型的推荐性能，本研究在两个真实数据集上开展了多次实验。通过严谨的实证分析，重点探究 MIAE 模型的性能表现和超参数的影响。

5.4.1 实验设置

表 5-1 数据集详细描述

Table 5-1 The descriptive statistics of datasets

数据集	RetailRocket	Tmall
点击数量	710586	818479
训练集数量	433648	351268
测试集数量	26968	25898
项目数量	36968	40728
会话平均长度	5.43	6.69

(1) 数据集。本章在 RetailRocket 和 Tmall 两个真实世界的数据集上进行了实验。前者是一家电子商务公司发布的 Kaggle 竞赛数据集。它包括 5 个月内匿名用户的浏览统计数据，后者取自 IJCAI-15 竞赛的官方数据集，包含天猫商城平台匿名用户在真实购物场景下的交易行为序列日志。基于之前的工作，筛选出长度为 1 的会话和出现次数少于 5 次的项目，按照 8: 1: 1 的比例按时间顺序划分训练集、验证集和测试集。表 5-1 总结了所有基准数据集的详细统计数据。

(2) 对比方法。通过构建对比实验，将所提模型与下列基准方法进行性能对比分析，以客观验证性能优势。

FMPC^[21]: 该模型同时捕捉用户的稳定兴趣和短期行为变化，精确地预测用户的需求和偏好。

GRU4Rec^[26]: 该模型是一种基于 GRU 的深度学习模型，通过建模用户行为序列来预测用户的下一步行为。

NARM^[27]: 该模型通过结合注意力机制和序列建模，从用户的历史行为中提取关键信息，预测用户的下一个行为。

SR-GNN^[34]: 该模型将用户的会话建模为图结构，并利用图神经网络来学习会话中项目之间的关系。通过这种方式，模型能够捕捉用户行为序列中的复杂模式，并预测用户的下一个行为。

CORE^[64]: 该模型将上下文信息嵌入到推荐系统中，通过联合建模用户行为数据和上下文信息，生成更精准的个性化推荐。

Atten-Mixer^[67]: 该模型提出了多层次注意力混合网络，利用概念视图和实例视图读数来实现项目转换的多层次推理，提高了推荐性能。

MSGIFSR^[68]: 该模型提出一个意图融合排序模块来合成不同粒度的用户意图，利用不同粒度的用户意图生成更准确、更全面的会话表示。

MiaSRec^[69]: 该模型采用动态选择重要的会话表示来表明各种用户意图，提高了推荐性能。

(3) 参数设置。本章使用平均值为 0，标准偏差为 0.1 的高斯分布来初始化参数。对于常规设置，将嵌入大小和批处理大小分别设置为 120 和 512。每个会话被截断，最大长度为 20。使用 Adam 来优化本章的模型。初始学习率为 $1e^{-3}$ ，每三个 epoch 的衰减因子为 0.1， L_2 正则项为 10^{-5} 。

(4) 评价方法。本章选用了基于会话推荐领域中常用的两项评价指标： $Precision@k$ 和 $MRR@k$ ，用于衡量推荐系统的性能表现。 $Precision@K$ 反映了模型预测结果的准确率， $MRR@k$ 用于评估目标物品在推荐列表中的排序质量。两项指标的数值越高，表明模型的推荐效果越优。

5.4.2 性能比较

为了验证 MIAE 的优越性，本章与较先进的模型进行了比较。从表 5-2 中可以看出：

(1) 与其他模型相比，FPMC 的表现相对较弱，而 GRU4Rec 和 NARM 则展现出较强的性能。这两种模型均采用 RNN 对用户行为序列进行建模，能够有效捕捉用户的动态偏好，进一步验证了序列相关性在推荐系统中的重要性。此外，NARM 的表现优于 GRU4Rec，表明注意力机制的引入能够更精准地识别并建模用户的主要兴趣。

(2) 基于图神经网络的模型性能优于 NARM 和 GRU4Reg 等传统模型，证明了图神经网络在数据建模中的有效性，因为它们可以捕捉更复杂项目之间的转换信息。此外，Atten-Mixer、MSGIFSR 和 MiaSRec 等模型往往比单一表示模型显示出更高的准确性。这意味着会话中可以存在各种意图，这证明了多意图在推荐系统中的有效性。

(3) 相较于基准方法，本章的方法在 $Precision$ 和 MRR 两个关键指标上均显著优于现有方法，充分证明了所提方法在会话推荐领域的适用性和有效性。

5.4.3 消融实验

为了证明本章的解决方案的合理性，通过对 MIAE 进行消融研究以验证核心组件的有效性，结果如表 5-3 所示。

在本节中，在保持所有其他条件不变的情况下，设计了三个模型变体来分析对模型性能 (RQ2) 的影响：模型 $MIAE_{nF}$ ，它删除了基于重复频率的注意力机制模块；模型 $MIAE_{nP}$ ，它删除了位置嵌入向量模块；模型 $MIAE_{nD}$ ，它删除了多意图模块，使用均值向量 ($\mathbf{o}_m$) 保留单一意图表示。结果显示，MIAE 模型的性能优于 $MIAE_{nF}$ 模型，表明增强重复项目的信息，可以降低对这些稀疏、噪声数

表 5-2 不同模型在两个数据集上的性能对比

Table 5-2 Comparison of the performance of different models on two datasets

Method	RetailRocket		Tmall	
	Precision @20	MRR@20	Precision @20	MRR@20
FMPC	46.04%	21.95%	20.30%	13.07%
GRU4Rec	55.32%	33.18%	23.25%	15.78%
NARM	58.65%	34.69%	31.67%	21.83%
SR-GNN	58.71%	36.42%	33.65%	25.27%
CORE	61.77%	38.49%	44.91%	31.59%
Atten-Mixer	59.49%	36.25%	38.76%	28.52%
MSGIFSR	63.04%	38.42%	35.39%	22.19%
MiaSRec	63.22%	39.15%	55.89%	33.41%
MIAE	63.79%	39.67%	56.36%	33.74%

表 5-3 不同模块对于模型性能的影响

Table 5-3 The impact of different modules on algorithm performance

Method	RetailRocket		Tmall	
	Precision @20	MRR@20	Precision @20	MRR@20
MIAE _{nF}	63.23%	39.14%	55.62%	33.14%
MIAE _{nP}	65.45%	38.31%	54.47%	32.69%
MIAE _{nD}	63.32%	38.47%	54.62%	32.49%
MIAE	63.79%	39.67%	56.36%	33.74%

据的敏感度，更好地理解并利用用户的序列特征；将 MIAE_{nP} 模型变体与 MIAE 模型进行比较，表明通过顺序信息反映每个项目的重要性对于提高性能是有效的。将 MIAE_{nD} 模型变体与 MIAE 模型进行比较，表明多个意图表示比单一意图表示效果更好。总体而言，MIAE 模型在两个数据集上都取得了最佳性能，这表明通过重复频率注意力机制以及位置信息和多意图表示的必要性和合理性。

5.4.4 超参实验分析

选取训练集中接近测试时间 10% 的数据作为验证集，用于超参数调优分析。双向自注意力机制使模型能够捕捉序列中不同位置的信息，但过多的注意力头可能导致单个注意力维度过小，从而削弱其表达能力。如图 5-2 所示，RetailRocket

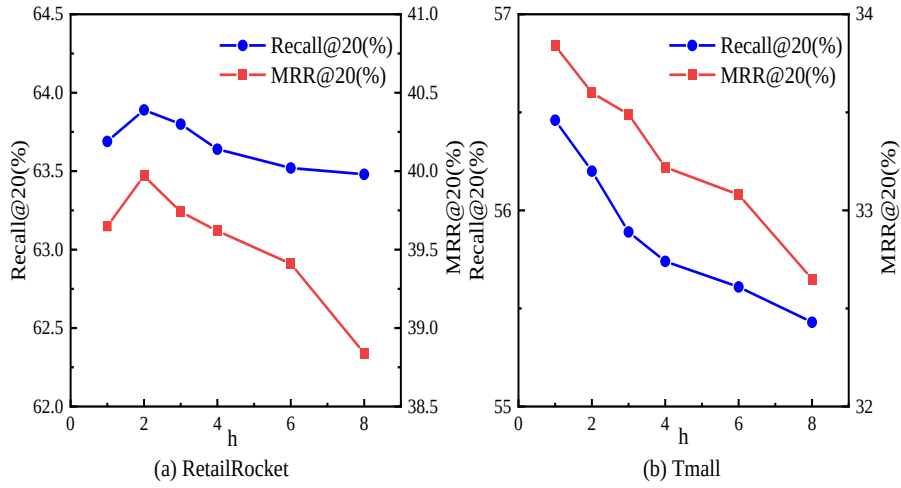


图 5-2 自注意力头数的性能影响

Fig. 5-2 The impact of the number of self-attention heads on algorithm performance

数据集的最佳注意力头数为 2，而 Tmall 数据集的最佳头数为 1。实验结果表明，随着注意力头数的增加，推荐性能会发生变化，当注意力头数偏离最佳值时，性能会出现下降。

5.5 本章小结

本章提出了一种在基于会话的推荐模型中学习更丰富表示的方法，来捕捉各种用户意图。在该模型中，通过位置和频率嵌入对会话项进行编码，以反映序列信息和重复信息。其次，通过使用自注意机制和高速网络来获取会话中不同的用户意图并自适应地提取会话中重要的用户意图，最后，有效解码和聚合多个表示从而进行推荐，模型显著提高了推荐性能。已经在广泛的推荐场景下进行了大量实验。实证结果表明，MIAE 模型显著优于几种最先进的方法。

第 6 章 总结与展望

6.1 工作总结

本课题主要将建模会话项目转换和会话表示层作为研究的重点，主要从以下几个方面来完成研究工作：它们分别是全局增强的图神经网络会话推荐，基于上下文增强的对比学习会话推荐，以及基于多意图感知增强的会话推荐的研究。

(1) 全局增强的图神经网络会话推荐。为了增强处理流行偏差的问题，在训练和推理过程中，本课题充分利用了将项目和会话图表示限制在一个单位超球面上的优点，提出了 SR 的规范化项目和会话表示模型。并且，通过引入自注意力机制来加强图神经网络。一方面，自注意机制是通过明确关注所有位置来捕捉全局依赖。另一方面，GNN 可以捕捉相邻项之间丰富的局部依赖关系，提供丰富的局部上下文信息。综合考虑了流行度偏差以及会话嵌入相关问题，同时将局部嵌入以及通过注意力机制获取的全局嵌入和目标嵌入相结合。

(2) 基于上下文增强的对比学习会话推荐。为了缓解用户的历史行为序列中的噪声导致了对用户真实偏好建模的不准确性，本课题首先通过计算项目对之间的相似性来构造兴趣图，通过将松散的项目序列转换为紧密的项目-项目兴趣图，更容易区分用户的主要兴趣和次级兴趣。同时，为每个会话添加一个上下文向量来构建会话图，该上下文向量可以提供一种自然的方式来传递相邻项之外的信息。最后使用对比学习作为辅助任务使两种会话嵌入互信息最大化，以缓解数据稀疏问题。

(3) 基于多意图感知增强的会话推荐。为了缓解没有从高层次的角度捕捉会话语义从而忽略会话中可能存在的各种用户意图问题，本课题首先通过使用项目节点的重复频率作为注意力权重来更新项目嵌入，降低对稀疏、噪声数据的敏感度，更好地理解 and 利用用户行为的序列特征。其次，采用双向自注意力机制和高速网络来获取会话中不同的用户意图并自适应地提取会话中重要的用户意图。最后，将多个会话表示解码为项目分布，并使用池函数将其聚合从而进行推荐。

6.2 未来展望

会话推荐作为个性化推荐系统中的一个重要方向,随着自然语言处理技术和深度学习方法的不断进步,得到了广泛的关注和应用。尤其是在处理多用户行为序列时,基于会话上下文的推荐算法能够显著提升推荐的准确性和用户体验。尽管本研究对会话推荐算法进行了部分改进,通过引入图神经网络和信息增强机制有效提升了会话推荐的性能,但由于研究时间、数据集的局限性以及算法的复杂性,仍有许多可以优化和改进的空间。随着对会话数据建模方法的进一步深入以及新型算法的提出,未来的研究将能够在处理更复杂的用户意图、多模态数据融合、实时推荐等方面取得更大突破,从而进一步提升会话推荐系统的精度、灵活性和用户满意度。对未来的研究有着以下的展望:

(1) 当前的研究主要集中在单一领域的会话推荐中,未来可以扩展到跨领域推荐系统,探讨如何结合不同领域的信息,尤其是如何通过多模态数据(如文本、图像、音频等)来增强会话推荐系统的泛化能力。例如,通过融合来自不同数据源的特征,提升系统对用户偏好的理解和推荐质量。

(2) 现有方法主要在离线环境中进行训练,未来可以进一步研究实时推荐系统,探索如何在用户与系统交互的过程中动态调整推荐策略。在线学习和增量学习的结合,能够帮助系统及时适应用户行为的变化,并根据最新的会话信息快速更新推荐结果。

(3) 未来的研究可以考虑将多任务学习与协同推荐相结合,通过联合优化不同的任务(如会话推荐、个性化推荐、点击预测等),提升推荐系统的整体性能。此外,探索基于协同过滤的图神经网络方法,进一步提升推荐精度。

参考文献

- [1] 王唯先. 数字霸权语境下个人信息传播向度的研究[J]. 新闻传播, 2023(20): 30-32.
- [2] 王惠敏, 田新瑞. 生成式人工智能引发的信息过载风险及其对策[J]. 科技管理研究, 2024, 44(20): 149-156.
- [3] 李邵莹, 孟丹, 孔超, 等. 面向社交推荐的自适应高阶隐式关系建模[J]. 软件学报, 2023, 34(10): 4851-4869.
- [4] 杨新新, 刘真, 卢思博, 等. 基于因果推断的推荐系统去偏研究综述[J]. 计算机学报, 2024, 47(10): 2307-2332.
- [5] Lin R, Tang F, Yuan C, et al. DeHier: decoupled and hierarchical graph neural networks for multi-interest session-based recommendation[J]. World Wide Web, 2024, 28(1): 1.
- [6] Guo S, Liao X, Meng F, et al. FSASA: Sequential recommendation based on fusing session-aware models and self-attention networks[J]. Computer Science and Information Systems, 2024, 21(1): 1-20.
- [7] Sang S, Yuan W, Li W, et al. Position-aware graph neural network for session-based recommendation[J]. Knowledge-Based Systems, 2023, 262: 110201.
- [8] Shi L, Zhang Z, Wang S, et al. Session-based Interactive Recommendation via Deep Reinforcement Learning[C]//2023 IEEE International Conference on Data Mining (ICDM), Shanghai, China, 2023: 1319-1324.
- [9] Li A, Cheng Z, Liu F, et al. Disentangled Graph Neural Networks for Session-Based Recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(08): 7870-7882.
- [10] 娄铮铮, 朱军娇, 张万闯, 等. 用户生成内容场景下角色导向图神经推荐方法[J]. 计算机学报, 2024, 47(6): 1288-1303.
- [11] Yin Y, Ling R, Xu Y, et al. A collaborative filtering recommendation method based on emotional evaluation relations[J]. Soft Computing, 2024, 28(13): 8167-8181.
- [12] Gu P, Hu H, Xu G. Modeling multi-behavior sequence via HyperGRU contrastive network for micro-video recommendation[J]. Knowledge-Based Systems, 2024, 295: 111841.
- [13] 赵容梅, 孙思雨, 鄢凡力, 等. 基于对比学习的多兴趣感知序列推荐系统[J]. 计算机研究与发展, 2024, 61(7): 1730-1740.
- [14] Li Z, Yang C, Chen Y, et al. Graph and Sequential Neural Networks in Session-based Recommendation: A Survey[J]. ACM Computing Surveys, 2024, 57(2): 40:1-40:37.
- [15] Zhang W, Yan J, Wang Z, et al. Neuro-Symbolic Interpretable Collaborative Filtering for Attribute-based Recommendation[C]//Proceedings of the ACM Web Conference 2022, Online, France, 2022: 3229-3238.
- [16] Musa J M, Zhihong X. Item Based Collaborative Filtering Approach in Movie Recommendation System Using Different Similarity Measures[C]//Proceedings of

- the 2020 6th International Conference on Computer and Technology Applications, Antalya, Turkey, 2020: 31-34.
- [17]Cui J, Yan C, Wang C. ReMEMBeR: Ranking Metric Embedding-Based Multicontextual Behavior Profiling for Online Banking Fraud Detection[J]. IEEE Transactions on Computational Social Systems, 2021, 8(3): 643-654.
- [18]Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web, Hong Kong, China, 2001: 285-295.
- [19]Wang M, Ren P, Mei L, et al. A Collaborative Session-based Recommendation Approach with Parallel Memory Modules[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 2019: 345-354.
- [20]Zhang Y, Shi Z, Zuo W, et al. Joint Personalized Markov Chains with social network embedding for cold-start recommendation[J]. Neurocomputing, 2020, 386: 208-220.
- [21]Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation[C]//Proceedings of the 19th international conference on World wide web, Carolina, USA, 2010: 811-820.
- [22]朱志国, 李伟玥, 姜盼, 等. 图神经网络会话推荐系统综述[J]. 计算机工程与应用, 2023, 59(5): 55-69.
- [23]王利娥, 李东城, 李先贤. 基于跨域关联与隐私保护的深度推荐模型[J]. 软件学报, 2023, 34(7): 3365-3384.
- [24]Wang S, Cao L, Wang Y, et al. A Survey on Session-based Recommender Systems[J]. ACM Comput. Surv., 2021, 54(7): 154:1-154:38.
- [25]Wang B, Cai W. Attention-Enhanced Graph Neural Networks for Session-Based Recommendation[J]. Mathematics, 2020, 8(9): 1607.
- [26]Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based Recommendations with Recurrent Neural Networks[J]. arXiv:1511.06939, 2016.
- [27]Li J, Ren P, Chen Z, et al. Neural Attentive Session-based Recommendation[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 2017: 1419-1428.
- [28]Liu Q, Zeng Y, Mokhosi R, et al. STAMP: Short-Term Attention/Memory Priority Model for Session-based Recommendation[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London , United Kingdom, 2018: 1831-1839.
- [29]Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains[C]//Proceedings. 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 2005: 729-734.
- [30]Scarselli F, Gori M, Tsoi A C, et al. The Graph Neural Network Model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.
- [31]Micheli A. Neural Network for Graphs: A Contextual Constructive Approach[J].

- IEEE Transactions on Neural Networks, 2009, 20(3): 498-511.
- [32] Wang J, Ding K, Zhu Z, et al. Session-based Recommendation with Hypergraph Attention Networks[C]//Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), Online, 2021: 82-90.
- [33] Zhu T, Sun L, Chen G. Graph-Based Embedding Smoothing for Sequential Recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 496-508.
- [34] Wu S, Tang Y, Zhu Y, et al. Session-Based Recommendation with Graph Neural Networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 346-353.
- [35] Zheng Y, Liu S, Li Z, et al. DGTN: Dual-channel Graph Transition Network for Session-based Recommendation[C]//2020 International Conference on Data Mining Workshops (ICDMW), Sorrento, Italy, 2020: 236-242.
- [36] Wang X, Liu X, Wang L, et al. A simple yet effective self-debiasing framework for transformer models[J]. Artificial Intelligence, 2025, 339: 104258.
- [37] Xu C, Zhao P, Liu Y, et al. Graph Contextualized Self-Attention Network for Session-based Recommendation[J]. 2019: 3940-3946.
- [38] Wang Z, Wei W, Cong G, et al. Global Context Enhanced Graph Neural Networks for Session-based Recommendation[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, China, 2020: 169-178.
- [39] Yu F, Zhu Y, Liu Q, et al. TAGNN: Target Attentive Graph Neural Networks for Session-based Recommendation[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, China, 2020: 1921-1924.
- [40] Ye R, Zhang Q, Luo H. Cross-Session Aware Temporal Convolutional Network for Session-based Recommendation[C]//2020 International Conference on Data Mining Workshops (ICDMW), Sorrento, Italy, 2020: 220-226.
- [41] Lin Z, Tian C, Hou Y, et al. Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning[C]//Proceedings of the ACM Web Conference 2022, Online, France, 2022: 2320-2329.
- [42] Yao T, Yi X, Cheng D Z, et al. Self-supervised Learning for Large-scale Item Recommendations[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Online, Australia, 2021: 4321-4330.
- [43] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, Dominican Republic, 2021: 6894-6910.
- [44] Wang L, Xu X, Ouyang K, et al. Self-Supervised Dual-Channel Attentive Network for Session-based Social Recommendation[C]//2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 2022: 2034-2045.

- [45]Wu J, Wang X, Feng F, et al. Self-supervised Graph Learning for Recommendation[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, Canada, 2021: 726-735.
- [46]Yu J, Yin H, Li J, et al. Self-Supervised Multi-Channel Hypergraph Convolutional Network for Social Recommendation[C]//Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 2021: 413-424.
- [47]Xia X, Yin H, Yu J, et al. Self-Supervised Hypergraph Convolutional Networks for Session-based Recommendation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, Pennsylvania, 2021: 4503-4511.
- [48]Xia X, Yin H, Yu J, et al. Self-Supervised Graph Co-Training for Session-based Recommendation[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Online, Australia, 2021: 2180-2190.
- [49]Wang Y, Zhu J, Duan L, et al. Disentangled Sparse Graph Attention Networks with Multi-Intent Fusion for Session-based Recommendation[J]. Knowledge-Based Systems, 2025, 311: 113082.
- [50]Zhu X, Zhang Y, Wang J, et al. Graph-enhanced and collaborative attention networks for session-based recommendation[J]. Knowledge-Based Systems, 2024, 289: 111509.
- [51]肖国庆, 李雪琪, 陈玥丹, 等. 大规模图神经网络研究综述[J]. 计算机学报, 2024, 47(1): 148-171.
- [52]刘壮, 董子宸, 董宜琳, 等. 图终身学习: 综述[J]. 计算机研究与发展, 2024, 61(8): 2067-2096.
- [53]Diao H, Li G, Xu S, et al. Self-distillation enhanced adaptive pruning of convolutional neural networks[J]. Pattern Recognition, 2025, 157: 110942.
- [54]Bourke T, Jeanmaire P, Pesin B, et al. Verified Lustre Normalization with Node Subsampling[J]. ACM Trans. Embed. Comput. Syst., 2021, 20(5s): 98:1-98:25.
- [55]Ben J, Sun Q, Liu K, et al. Multi-head multi-order graph attention networks[J]. Applied Intelligence, 2024, 54(17): 8092-8107.
- [56]Tofighy S, Charkari N M, Ghaderi F. A meta-learning based approach for temporal link prediction in multiplex networks[J]. Knowledge-Based Systems, 2025, 309: 112803.
- [57]Bing R, Yuan G, Zhang Y, et al. Automated message selection for robust Heterogeneous Graph Contrastive Learning[J]. Knowledge-Based Systems, 2025, 307: 112739.
- [58]Bai X, Zhang L, Jiang M, et al. Gated graph spiking neural P network for session-based recommendation[J]. Knowledge-Based Systems, 2024, 300: 112162.
- [59]De Smedt J, De Weerd J. Predictive process model monitoring using long short-term memory networks[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108295.
- [60]Astudillo A, Barrera A, Guindel C, et al. DAttNet: monocular depth estimation

- network based on attention mechanisms[J]. *Neural Computing and Applications*, 2023, 36(7): 3347-3356.
- [61]Moslemi A, Jamshidi M. Unsupervised feature selection using sparse manifold learning: Auto-encoder approach[J]. *Information Processing & Management*, 2025, 62(1): 103923.
- [62]Yessou H, Sumbul G, Demir B. A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification[C]//IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, Online, 2020: 1349-1352.
- [63]Gupta P, Garg D, Malhotra P, et al. NISER: Normalized Item and Session Representations with Graph Neural Networks[J]. *ArXiv*, 2019.
- [64]Hou Y, Hu B, Zhang Z, et al. CORE: Simple and Effective Session-based Recommendation within Consistent Representation Space[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022: 1796-1801.
- [65]Han Q, Zhang C, Chen R, et al. Multi-Faceted Global Item Relation Learning for Session-Based Recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022: 1705-1715.
- [66]Lai S, Meng E, Zhang F, et al. An Attribute-Driven Mirror Graph Network for Session-based Recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022: 1674-1683.
- [67]Zhang P, Guo J, Li C, et al. Efficiently Leveraging Multi-level User Intent for Session-based Recommendation via Atten-Mixer Network[C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, Singapore, 2023: 168-176.
- [68]Guo J, Yang Y, Song X, et al. Learning Multi-granularity Consecutive User Intent Unit for Session-based Recommendation[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Online, USA, 2022: 343-352.
- [69]Choi M, Kim H young, Cho H, et al. Multi-intent-aware Session-based Recommendation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, USA, 2024: 2532-2536.

攻读学位期间发表的学术成果

已录用的论文:

- [1] Xianlan Sun, Dan Meng, Xiangyun Gao, et al. GENE: Global Enhanced Graph Neural Network Embedding for Session-Based Recommendation[C], International Conference on Web Information Systems and Applications. Singapore: Springer Nature Singapore, 2023: 173-184. (EI, 已发表)
- [2] Xianlan Sun, Xiangyun Gao, Subin Huang, et al. Context-Augmented Contrastive Learning Method for Session-based Recommendation[C], International Conference on Advanced Data Mining and Applications. Singapore: Springer Nature Singapore, 2024: 19-33. (CCF-C, 已发表)

攻读学位期间参与的科研项目

- [1] 国家自然科学基金青年科学基金项目, 面向互联网开放数据的实体匹配及其应用研究 (项目号 61902001)
- [2] 安徽省高等学校科学研究项目 (自然科学类), 面向数据安全性与隐私保护的个性化推荐方法研究 (项目号 2023AH050914)

