文章编号：1006-3080(2025)06-0827-08 DOI：10.14135/j.cnki.1006-3080.20250225002

融合建模的图神经网络会话推荐模型

杜佳宇[1]， 郑 红[1]， 郭津延[1]， 罗俞建[1]， 李鹏威[1]， 单蓉胜[2]
（1. 华东理工大学信息科学与工程学院, 上海 200237； 2. 上海交通大学网络空间安全学院, 上海 200240）

摘要：针对传统会话推荐算法仅依赖显式信息而忽视会话间潜在交互关系的问题，本文提出了一种基于门控和图注意力机制的融合建模模型 IM-GGN(Integrated Modeling Gated Graph Network)，它对物品间的结构化关系和会话间的非结构化关系同时进行建模，从而提升推荐性能。该模型由结构化关系学习(Structured Pattern Learning，SPL)模块与非结构化关系学习(Unstructured Pattern Learning，UPL)模块组成：SPL 模块结合图神经网络和门控机制，捕捉会话内部的顺序依赖和长程关系；UPL 模块则利用图注意力机制对会话间非结构化的关联信息进行建模，以提取用户偏好上下文。实验结果表明，本文方法在多个公开数据集上均取得了一定程度的性能提升，验证了模型在会话推荐中的有效性。

关键词：会话推荐；门控图神经网络；图注意力机制；结构化关系；非结构化关系

中图分类号：TP183 **文献标志码：**A

基于会话的推荐旨在从给定匿名用户的短期物品访问序列中挖掘动态变化的兴趣偏好，以推荐用户下一个可能点击的物品[1]，由于它在电子商务、音乐推荐等领域具有广泛应用，因此受到越来越多的关注[2-3]。目前，该领域的大多数研究工作将会话视为有序序列，其中基于循环神经网络 (Recurrent Neural Network, RNN) 和基于图神经网络 (Graph Neural Network, GNN) 的方法在会话推荐技术中表现出了良好性能。

随着深度学习技术在各个领域的蓬勃发展，基于 RNN 的会话推荐技术也成为了该领域的热点。LSTM [4] 和多层门控循环单元 GRU[5] 等新兴技术被用来改善模型的序列建模与捕捉用户偏好的能力。多个深度学习模型，如 HARSAM[6]、NARM[7] 和 STAMP[8]，结合了自注意力和短期记忆网络，有效提升了推荐效果。然而，这些基于深度学习的方法只关注当前会话的相邻交互项，而不考虑交互项与其他位置的项目之间的依赖性，并且忽略了不同会话之间的依赖关系。

随着图神经网络 (Graph Neural Network, GNN) 的快速发展[9-11]，人们提出了许多图嵌入方案[12-14]，通过 GNN 有效地捕获项目之间的转换模式，获得了令人满意的推荐性能。然而，这些方法大多只关注相邻项之间的一阶转换关系，而忽略了无连接项的高阶信息，如郑楠等[15] 提出的 InterAtt-GNN 模型，虽然多层传播可以抑制这一问题，但过拟合问题变得更加突出。此外，由于会话图中缺乏序列位置信息，不同的会话可能被转换为相同的图。在这种情况下，序列模式无法从会话图中可靠地捕获序列位置信息从而无效。为了更好地识别用户偏好，部分研究者提出了共同利用会话的序列信息和项目的过渡关系来生成准确的图嵌入[16-18]。

由于会话推荐场景无法依赖用户历史行为数据，所以其主要依据当前会话的点击序列对用户偏好进行建模。然而，在短序列场景中，仅依赖会话的点击序列信息往往不足以生成准确的推荐。这是因为：(1)会话序列长度较短：会话序列通常仅包含有限的点击行为，仅依赖其序列信息可能无法全面反

收稿日期： 2025-02-25
基金项目： 上海市 2024 年度“科技创新行动计划”资助(24BC3200500, 24BC3200300)
作者简介： 杜佳宇(2000—)，浙江人，硕士生，主要研究方向为人工智能和推荐系统。E-mail: Y80220102@mail.ecust.edu.cn
通信联系人： 郑 红, E-mail: zhenghong@ecust.edu.cn
引用本文： 杜佳宇, 郑 红, 郭津延, 等. 融合建模的图神经网络会话推荐模型 [J]. 华东理工大学学报(自然科学版), 2025, 51(6): 827-834.
Citation: DU Jiayu, ZHENG Hong, GUO Jinyan, *et al*. Graph Neural Network Session Recommendation Model with Fusion Modeling[J]. Journal of East China University of Science and Technology, 2025, 51(6): 827-834.

映用户偏好;(2)忽视会话间的潜在关联:现有方法通常仅关注各会话序列中项目的显式依赖关系,却忽视了跨会话间潜在的用户行为关联,导致上下文信息的丢失,其他方法虽然会附加简单会话级特征,但不系统建图,同样会导致上下文信息的缺失。

尽管以 Transformer 为基础的大规模预训练模型在推荐任务中展示了强大的建模能力,但它们在会话推荐的实际应用场景中,仍面临两个不可忽视的瓶颈。首先,结果稳定性与偏置校正方面,直接将大模型用于推荐时,由于大规模并行自注意力和复杂的预训练目标,会受到内置的"位置偏置"影响,导致推荐列表对输入顺序高度敏感、结果不稳定[19],从而导致推荐结果在不同会话或不同批次下波动较大,难以保证对用户意图的鲁棒捕捉;其次,在线实时性的计算开销方面,大模型通常包含数以亿计的参数,其推理过程需要多层全连接与多头注意力的密集矩阵运算,难以满足毫秒级响应要求[20]。

为解决上述问题,本文提出将会话推荐中的关系划分为两类:结构化关系和非结构化关系。结构化关系指不同物品之间通过点击顺序自然形成的显式依赖,用以捕捉会话中物品的转换模式;非结构化关系指不同会话之间隐含的用户行为相似性,这类关联无法通过简单规则直接构建,但可通过图注意力等学习机制自动挖掘并融入模型,从而补充短序列场景下的上下文信息。基于此,本文提出一种融合建模的门控图神经网络模型 IM-GGN(Integrated Modeling Gated Graph Network),同时建模结构化与非结构化会话信息。在结构化关系学习模块(SPL)中,引入图神经网络结合门控机制,有效捕捉会话内部的顺序与长距离依赖;在非结构化关系学习模块(UPL)中,采用跨会话图注意力机制动态挖掘邻近会话的语义关联,为目标会话提供外部上下文补充,从而生成更全面的会话表示向量。

1 相关模型

1.1 问题描述

基于会话的推荐任务是根据用户在当前会话中的物品交互记录,发掘用户偏好,预测用户接下来可能点击的物品,以下给出本文的符号定义:在会话推荐中,使用 $V=\{v_1,v_2,\cdots,v_N\}$ 代表 N 个物品的集合。每个会话被表示为 $s=[v_{s,1},v_{s,2},\cdots,v_{s,m}]$,其中每一项代表一个物品。为了学习物品和会话的向量表示,将每个物品嵌入到 d 维向量空间中并表示为 $v_i\in\mathbb{R}^d$。对于每个会话,将所有项目的概率设置为 $\hat{y}=\{y_i\}_{i=1}^{|V|}$,其中向量 $\boldsymbol{y}_i$ 是相应项目的推荐分数。会话推荐的目标是预测给定会话中用户点击的下一项物品 $v_{s,m+1}$。

1.2 模型整体架构

IM-GGN 的总体架构如图 1 所示,它由以下几个部分组成:

(1) 会话图构建:根据会话序列构建结构化关系图与非结构化关系图这两类关系图,前者由点击序列构建,后者基于会话间隐式相似性构建;

(2) 结构化关系学习模块:通过序列建模图神经网络(Sequential Modeling GNN, SM-GNN)捕捉物品间的显式依赖,并结合门控序列感知网络(Gated Sequence Perception Network, GSPN)提取顺序模式和长程依赖信息,生成会话内的物品嵌入表示;

(3) 非结构化关系学习模块:基于上下文驱动图注意力机制(Context-Driven Graph Attention, CD-GAT)挖掘跨会话的上下文信息,生成与用户偏好相似的物品表示;

(4) 融合与预测:将 SPL 和 UPL 模块生成的物品表示进行线性融合,并通过软注意力机制自适应地结合用户的短期兴趣和长期偏好,生成最终的会话表示,用于预测下一个可能点击的物品。

其中,结构化关系与非结构化关系如图 2 所示。结构化关系图中的每个节点 I 是一个项目,而非结构化关系图中的每一个节点 S 代表一个会话,边上的值代表两个会话之间的相似度,模型选择相似度 TOP-K 会话作为当前会话的邻域会话,用以补充上下文信息。实际上,每个会话之间都有一条双向边,但为了保持图的简洁,此处仅画出部分边。

1.3 结构化关系学习模块

在 IM-GGN 中,SPL 模块由 T 层门控序列感知网络及序列建模图神经网络构建,旨在学习会话序列中的长期依赖信息并提取结构化关系。

1.3.1 SM-GNN 为了动态更新隐藏状态,SM-GNN 采用门控机制,通过可学习的权重矩阵对隐藏状态 $X^{(0)}$ 进行线性变换,并通过 Chunk 函数均匀分块:

$$\boldsymbol{X}_w=\boldsymbol{X}^{(0)}\boldsymbol{W}_h \tag{1}$$

$$[\boldsymbol{X}_{w0},\boldsymbol{X}_{w1},\boldsymbol{X}_{w2}]=\mathrm{Chunk}(\boldsymbol{X}_w,3) \tag{2}$$

其中,$\boldsymbol{W}_h$ 为可学习的权重矩阵,$\boldsymbol{X}_w$ 为变换后的隐藏状态。

为了将隐藏状态的相关性信息引入全局邻接矩阵,首先根据第 t 次迭代的隐藏矩阵 $\boldsymbol{X}^{(t)}$ 计算隐藏状态之间的相关性矩阵 $\boldsymbol{C}$,然后对相关性矩阵进行 L2 归一化并与全局邻接矩阵融合:

$$\boldsymbol{C}=\boldsymbol{X}^{(t)}(\boldsymbol{X}^{(t)})^{\mathrm{T}} \tag{3}$$

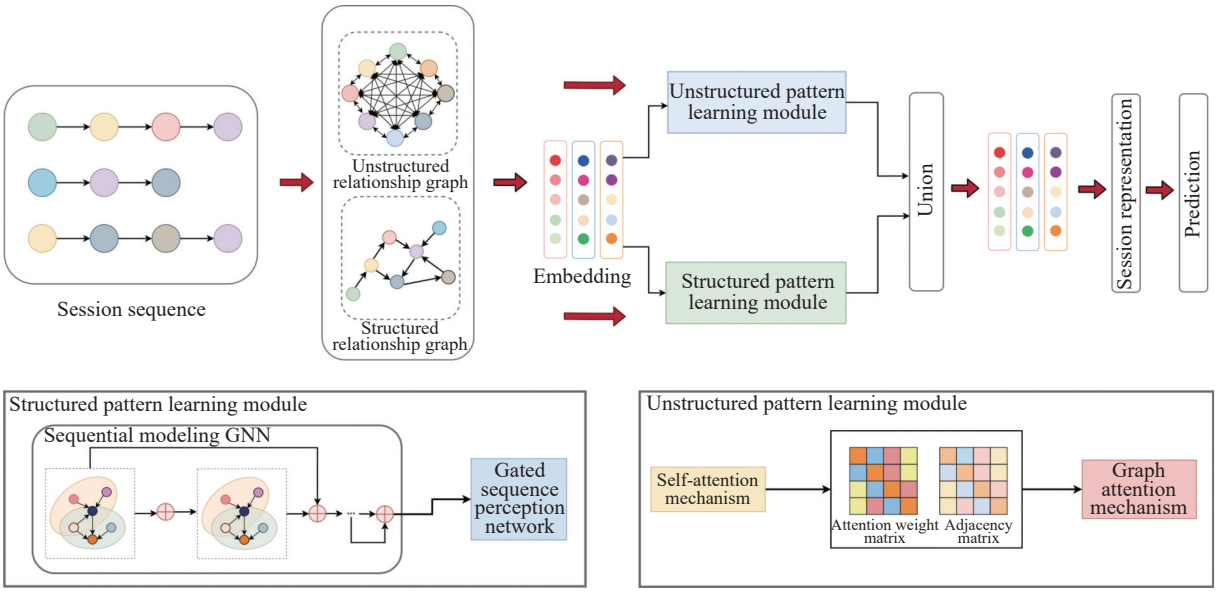

图 1　IM-GGN 模型架构图

Fig. 1　IM-GGN model architecture diagram

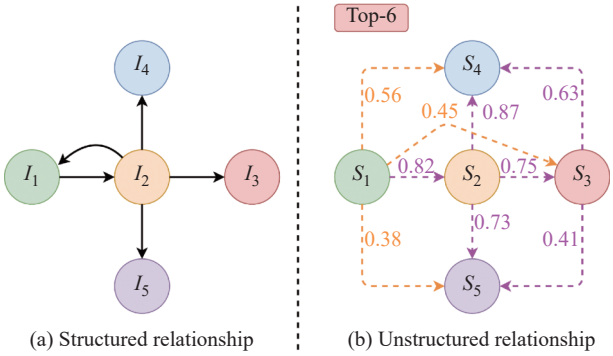

图 2　结构化关系图和非结构化关系图

Fig. 2　Structured relationship graph and unstructured relationship graph

$$A_{\text{new}} = A_{\text{Global}} + \frac{C}{\|C\|_2} \tag{4}$$

$\boldsymbol{A}_{\text{Global}} \in \mathbb{R}^{N\times N}$ 表示全局邻接矩阵, 表示会话中 N 个项目之间的关系。利用融合后的邻接矩阵 $\boldsymbol{A}_{\text{new}}$, 通过门控机制进一步得到更新隐藏状态 $\boldsymbol{X}_{\text{new}}$:

$$\boldsymbol{X}_{\text{new}} = \boldsymbol{X}_w + \text{ReLU}(\boldsymbol{A}_{\text{new}}\boldsymbol{X}_{w0}\boldsymbol{W}_{\text{hf}}) \odot \boldsymbol{X}_{w1} \tag{5}$$

其中, $\boldsymbol{W}_{\text{hf}} \in \mathbb{R}^{2d\times d}$ 为可学习的权重矩阵, $\odot$ 表示逐元素乘法操作, ReLU 表示激活函数。

该更新过程可迭代 T 步, 最终输出最后一次迭代得到的隐藏状态 $\boldsymbol{X}^{(t)}$:

$$\boldsymbol{X}^{(t)} = \text{SM-GNN}(\boldsymbol{A}_{\text{new}}, \boldsymbol{X}^{(t-1)}) \tag{6}$$

1.3.2　GSPN　传统的会话推荐系统[21-22](SBR)通常依赖于 Transformer 架构中的自注意力机制来捕捉会话序列中的顺序信号。然而, 自注意力机制的计算复杂度随着序列长度的增加呈二次增长, 这在处理长会话时会带来显著的计算开销。自注意力机制的有效性高度依赖位置编码(Positional encoding), 而现有的绝对位置编码和相对位置编码在 SBR 任务中仍存在一定的局限性。因此, 在 SPL模块中引入了 GSPN 模块, 以最大化利用序列位置信息, 并有效捕捉会话中项目之间的长期依赖关系。

GSPN 使用 SM-GNN 的输出作为输入, 其模块由两个主要部分组成: 空间特征门控模块(Spatial Feature Gate, SFG)和多层感知机(Multi-Layer Perception, MLP)。

空间特征门控模块是 GSPN 的核心组件, 包含对空间维度的收缩操作, 用来捕捉序列中的位置信息和长期依赖关系。GSPN 可通过 SFG 来捕捉会话的复杂结构化关系, 而不需要使用位置编码, 避免了显式位置偏码带来的额外计算。

首先, 使用层归一化(Layer Normalization)稳定特征分布, 提升训练稳定性, 随后通过一维卷积(Conv1d)对归一化后的特征 $\hat{v}$ 进行线性变换, 捕获序列中的局部关系, 并通过门控机制实现动态特征调节, 得到输出特征张量 o。其中, 输入 $\boldsymbol{X}$ 被均匀拆分为 $\boldsymbol{u}$ 和 $\boldsymbol{v}$。μ 和 σ^2 分别表示 $\boldsymbol{v}$ 在特征维度上的均值和方差, ε 是一个小常数, 用于防止除零错误。

$$\boldsymbol{u}, \boldsymbol{v} = \text{Chunk}(\boldsymbol{X}, 2, \text{dim} = -1) \tag{7}$$

$$\hat{v} = \frac{\boldsymbol{v} - \mu}{\sqrt{\sigma^2 + \varepsilon}} \tag{8}$$

$$o = \boldsymbol{u} \odot \text{Conv1d}(\hat{v}) \tag{9}$$

在上述操作的基础上, 将 SFG 与 MLP 结合, 通过线性变换与非线性激活, 提升了特征表示的复杂

性和表达能力。同时,通过残差连接将输出与输入特征 $\boldsymbol{X}$ 相加,确保信息在多层网络中的有效传递。

$$\boldsymbol{H}' = \mathrm{GELU}(\mathrm{Dropout}(\mathrm{Linear}_1(o))) \tag{10}$$

$$\boldsymbol{H}_s = \mathrm{Dropout}(\mathrm{Linear}_2(\boldsymbol{H}')) + \boldsymbol{X} \tag{11}$$

其中,$\boldsymbol{H}'$ 是通过门控机制聚合项目嵌入和结构化关系的会话表示。$\boldsymbol{H}_s$ 是 GSPN 的最终输出,GELU 是激活函数,Dropout 函数会随机丢弃部分神经元节点,以缓解过拟合问题,Linear 是线性变换函数。通过这种方式,会话项目间的结构化关系可以通过 SPL 模块中使用的空间映射矩阵来捕获。

1.4 非结构化关系学习模块

如前文所述,不同会话之间的非结构化关系也有利于预测用户的下一个点击项,为了利用这种关系,UPL 模块建立在一个完善的上下文驱动图注意力机制基础上,进一步探索来自 K 近邻会话的上下文信息,从而得到具有用户兴趣和行为相似性的上下文表示。

1.4.1 非结构化关系图构建 在应用 CD-GAT 进行探索之前,需要构造非结构化关系图并确定当前会话的 K 近邻。首先将每个会话视为一个节点,并计算当前会话与其他会话之间的相似性。也就是说,会话的全局表示之间的相互依赖关系是从当前会话和其他会话之间的相似性建立。

将会话的全局表示作为输入,利用一种简化的自注意力机制来计算会话之间的相似性得分矩阵:

$$\boldsymbol{B}_c = \mathrm{softmax}\left(\frac{\boldsymbol{H}^i(\boldsymbol{H}^j)^{\mathrm{T}}}{\sqrt{D}}\right) \tag{12}$$

其中,$\boldsymbol{B}_c$ 的元素 $\boldsymbol{B}_c^{i,j}$ 表示节点 i 与节点 j 之间的相似度,$\boldsymbol{H}^i$、$\boldsymbol{H}^j$ 分别表示会话 i、j 的全局表示向量。$\sqrt{D}$ 为缩放因子,用于稳定梯度并对相似度得分进行缩放。

为了缓解过度平滑问题,构造非结构化关系图的关键问题是搜索 K 近邻会话。因此,UPL 模块还引入近似 KNN 算法,选择 K 近邻会话作为当前会话的邻域节点。

$$\boldsymbol{F}_s(i,j) = \begin{cases} 1 & \text{if } j \in \text{Top-}k(\boldsymbol{B}_c^i) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

其中,$\boldsymbol{F}_s(i,j)$ 表示会话 i 与 j 是否为近邻关系,Top-$k(\boldsymbol{B}_c^i)$ 表示节点 i 与其他节点的相似度中排名前 k 的节点。

1.4.2 上下文驱动图注意力机制 引入了动态图注意力网络 GATv2[23] 从非结构化关系图中提取非结构化关系,进一步提高图结构关系的特征表达能力。GATv2 解决了标准 GAT 中静态注意力的问题,它在计算注意力权重时引入非线性变换,使得每个节点可以关注任何其他节点。利用该网络,UPL 模块通过动态图注意力机制自适应地为不同的邻域节点分配适当的边权。根据动态图注意力网络的论文中的相关设置,在 GATv2内部,动态图注意力系数的计算如式(14)所示:

$$\alpha_{i,j} = \frac{\exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{LeakyReLU}(\boldsymbol{W}_s\boldsymbol{H}^i + \boldsymbol{W}_t\boldsymbol{H}^j))}{\sum_{k\in\mathcal{N}(i)}\exp\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{LeakyReLU}\left(\boldsymbol{W}_s\boldsymbol{H}^i + \boldsymbol{W}_t\boldsymbol{H}^k\right)\right)} \tag{14}$$

其中,$\boldsymbol{W}_s$ 和 $\boldsymbol{W}_t$ 是可学习的权重矩阵,$\boldsymbol{\alpha}$ 是可训练的变换向量,$\mathcal{N}(i)$ 是节点 i 的邻居集合。

因此,非结构化关系表示通过相邻会话表示与注意力系数聚合来计算:

$$\boldsymbol{H}_c^i = \sum_{j\in\mathcal{N}_i}\alpha_{i,j}\boldsymbol{H}^j \tag{15}$$

通过这种方式,可以从相邻会话中提取的非结构化线索获得非结构化关系。

1.5 融合模块

给定 SPL 模块与 UPL 模块中项目的结构化与非结构化表示,模型使用线性变换将其融合,从而得到最终的项目表示 $\widehat{X}$:

$$\widehat{X} = [X_{\mathrm{SPL}}^{(M)} \| X_{\mathrm{UPL}}^{(L)}]\boldsymbol{W}_F \tag{16}$$

其中 $\boldsymbol{W}_F$ 是可学习的参数矩阵。

一般来说,用户的短期兴趣是随时间变化的当前偏好,而用户的长期兴趣则代表会话中稳定的偏好。从用户点击序列中提取的短期和长期兴趣对推荐都至关重要。此处假设用户当前的偏好可以通过最后一个条目来反映,因此模型使用最后单击项的表示形式 $\hat{x}_m$ 作为会话的本地表示形式,即 $s_l = \hat{x}_m$。对于会话 s,基于会话中所有项的表示生成其全局表示 s_g。软注意力机制是指在模型中通过可学习的注意力权重,对一组表示进行加权求和,从而自适应地聚焦于最相关信息,并保持端到端可微。模型采用软注意力机制来融合所有项目的信息,同时考虑到它们的不同重要性:

$$\alpha_i = \boldsymbol{q}^{\mathrm{T}}\sigma(\boldsymbol{W}_1\hat{x}_m + \boldsymbol{W}_2\hat{x}_i + c) \tag{17}$$

$$s_g = \sum_{i=1}^{m}\alpha_i\hat{x}_i \tag{18}$$

其中,$\boldsymbol{q}^{\mathrm{T}}$、$\hat{x}_i$、$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 都是可学习的参数。α_i 控制物品表示的权重。基于匿名数据源的用户短期兴趣和长期兴趣的重要性很难界定,这可能取决于用户的个人偏好。因此,有必要平衡用户短期利益和

长期利益的重要性。

为了克服这一问题, 模型利用门控融合函数, 通过综合挖掘用户的不同行为偏好, 自适应地聚合用户的短期利益和长期利益。混合利益表示可以定义为:

$$\boldsymbol{S}_h = \boldsymbol{G}_s \odot \boldsymbol{S}_l + (1-\boldsymbol{G}_s) \odot \boldsymbol{S}_g \tag{19}$$

式中, $\boldsymbol{S}_h$ 表示用户偏好的混合表示, $\boldsymbol{G}_s$ 是由 $\boldsymbol{S}_l$ 和 $\boldsymbol{S}_g$ 决定的门控融合向量:

$$\boldsymbol{G}_s = \sigma(\boldsymbol{S}_l \boldsymbol{W}_l + \boldsymbol{S}_g \boldsymbol{W}_g) \tag{20}$$

通过公式(21)计算分数 $\hat{z}$:

$$\hat{z} = \boldsymbol{S}_h^{\mathrm{T}} \boldsymbol{X} \tag{21}$$

然后对分数使用 Softmax 函数进行下一项预测, 生成的概率表示每个项目是用户下一个点击的可能性:

$$\hat{y} = \mathrm{Softmax}(\hat{z}) \tag{22}$$

对于每一个会话, 损失函数被定义为预测结果与真实结果的交叉熵:

$$\mathcal{L}(\hat{y}) = -\sum_{i=1}^{n} [y_i \lg(\hat{y}_i) + (1-y_i)\lg(1-\hat{y}_i)] \tag{23}$$

其中 y_i 是 one-hot 编码。

2　实验部分

2.1　数据集

本文基于会话的推荐中常用的 Diginetica、Yoochoose 和 Gowalla 这 3 个数据集进行实验, 其统计数据如表 1 所示。

表 1　数据集详细信息

Table 1　Details of the dataset

Dataset	Number						
	#clicks	#train session	#test session	#item	#length≤5	#length > 5	Average length
Diginetica	981 620	716 835	60 194	42 596	537 546	239 483	4.80
Yoochoose1/64	557 248	369 859	55 898	16 766	289 490	136 627	6.16
Gowalla	1 122 78	675 561	155 332	29 510	627 100	203 793	4.32

Diginetica 是 2016 年 CIKM 杯的个性化电子商务研究挑战数据集, 数据集包含转换历史, 适合基于会话的推荐。Yoochoose 来自 2015 年的 RecSys 挑战赛, 由于 Yoochoose 的训练集非常大, 所以本文使用所有训练会话的最近部分 1/64 子样本作为训练集, 分别记为“Yoochoose1/64”。Gowalla 是一个广泛用于兴趣点推荐的数据集。基于之前研究者的工作, 本文保留了前 30 000 个最受欢迎的地点, 并通过分割相邻记录之间的间隔(超过一天)将用户的签到记录分组为不相交的会话。在本章的实验中使用最后 20% 的会话作为测试集。根据此前研究, 对数据集进行相应的预处理, 过滤掉长度为 1 的会话和出现次数少于 5 次的项目。

2.2　实验评价指标

为了便于与基线模型进行比较, 文章选择了常用的命中率 (HR) 和平均倒数排名 (MRR) 作为评估指标。在实际推荐中系统通常同时推荐多个物品, 为了评估不同物品数量的推荐效果, 使用 HR@K 和 MRR@K 以测量模型的性能, 其中 K 表示推荐物品的数量。HR@K 表示前 K 个项目中正确推荐的项目的比例, 并定义为:

$$\mathrm{HR@K} = \frac{n_{\mathrm{hit}}}{N} \tag{24}$$

其中, N 是测试集中的序列数, n_{hit} 是在排名列表中前 K 个物品中正确推荐的物品数。

MRR@K 用于评价正确推荐的物品在长度为 K 的推荐列表中所处的位置, 其具体值等于该物品 v_t 在列表 I 中的排名的倒数。如果 v_t 位于列表 I 的第 1 位时 MRR@K 为 1, 当 v_t 未出现在列表 I 中时 MRR@K 为 0。假设测试集的大小为 N, 取平均值 MRR@K 作为评估的度量:

$$\mathrm{MRR@K} = \frac{1}{N} \sum_{v_t \in S_{\mathrm{test}}} \frac{1}{\mathrm{rank}(v_t)} \tag{25}$$

其中 $\mathrm{rank}(v_t)$ 是 v_t 在推荐列表中的排名。

2.3　实验设置

为了实验的公平性, 本实验选择初始学习率为 0.001 的 Adam 优化器, 每 5 个 epoch 后学习率衰减 0.5。将 L2 正则化设置为 10^{-5}, 并使用早停策略(连续 5 个 epoch 的评估指标没有改进)来缓解过拟合问题。使用均值为 0、标准差为 0.1 的高斯分布初始化所有参数。将嵌入维度和批大小都固定为 100。对于结构化关系学习模块, 层数在{1,2,3,4}内变化; 对于非结构化关系学习模块, 块数在{4,5,6}范围内。

2.4 实验分析

2.4.1 对比模型 本文选择了11个基准模型进行对比实验:

(1) POP: 是一个简单的基准模型, 为用户推荐最受欢迎(排名最高)的项目。

(2) Item-KNN: 通过当前会话的每个项目与其他项目之间的相似性来推荐项目。

(3) FPMC: 将一阶马尔可夫链与矩阵分解相结合, 以捕获顺序效应和用户偏好。

(4) GRU4Rec: 采用门控循环单元来模拟会话中项目的顺序行为。

(5) NARM: 通过引入RNN, 关注基于会话的推荐, 改进了GRU4Rec。

(6) SR-GNN: 通过图神经网络对会话内的显式依赖关系进行建模, 然后应用软注意力机制生成会话级嵌入。

(7) SGNN-HN: 应用图神经网络来模拟正在进行的会话中没有直接连接的项目之间的复杂转换关系。

(8) LESSR: 引入了两种带有自跳闸和捷径的会话图来捕获隐式连接, 解决信息丢失和长期依赖问题。

(9) MSGIFSR: 提出了一个连续意图单元, 基于当前会话中不同的项目组, 从不同粒度提取用户意图。

(10) GC-SAN: 通过使用GGNN获取本地上下文信息, 然后利用自关注机制捕获显式依赖。

(11) GCE-GNN: 考虑 ε-neighbor(ε = 2) 连接, 构建基于会话的推荐的会话间图。

2.4.2 对比实验 表2记录了11个基准模型在Diginetica、Yoochoose1/64和Gowalla数据集上的HR@20和MRR@20指标结果。比较结果表明, IM-GGN明显优于所有基线模型结果。从表2可以看出, 深度学习方法的表现明显优于传统方法(如POP、Item KNN和FPMC), 证明了它们优越的复杂特征提取和表示能力。NARM的表现优于GRU4Rec, 因为NARM不仅可以捕获会话中潜在的序列信息, 还可以通过注意力机制学习项目相关性。

基于GNN的模型通常优于基于序列的方法, 这表明会话图在表示不同项目之间的转换关系方面的重要性。MSGIFSR设计了各种粒度意图单元对项目之间的隐式和多粒度关系进行建模, 在基线模型中表现最优, 这表明设计复杂的模块来捕捉基于会话的推荐项目之间的隐含相关性十分必要。

不同于基线模型, 本文将历史会话序列构建为结构化与非结构化关系图, 并采用IM-GGN来学习物品之间的复杂高阶关系, 同时整合了结构化和非结构化的信息, 因此能够有效地表征当前会话序列中物品的特征。

2.5 消融实验

为了验证SM-GNN、GSPN和CD-GAT在IM-GGN中的有效性, 实验从IM-GGN中移除或替换其中一个模块来分析性能变化, 具体如下:

(1) MLP-SR: 用一个单层MLP替代SM-GNN、GSPN和UPL模块, 保留预测层和损失函数, 验证整体图神经网络结构的有效性;

(2) w/o GSPN: 去掉GSPN模块, 仅使用SPL的SM-GNN模块和UPL模块生成物品表示;

(3) w/o SM-GNN: 去掉SM-GNN模块, 仅使用

表2 不同模型在3个数据集上实验性能对比

Table 2 Performance comparison of different models on three datasets

Model	Diginetica		Yoochoose1/64		Gowalla	
	HR@20	MRR@20	HR@20	MRR@20	HR@20	MRR@20
POP	0.89	0.28	6.71	1.65	1.46	0.38
Item-KNN	37.75	11.57	51.60	21.81	38.60	16.66
FPMC	26.53	6.66	45.62	15.01	29.91	11.45
GRU4Rec	29.45	8.22	60.64	22.89	41.98	18.37
NARM	49.70	16.00	68.32	28.63	50.07	23.92
SR-GNN	50.73	17.78	70.57	30.94	50.32	24.25
SGNN-HN	55.67	19.45	72.13	32.60	55.28	27.58
LESSR	51.71	18.15	70.59	31.46	51.34	25.49
MSGIFSR	57.11	20.05	73.13	33.50	56.64	29.02
GC-SAN	51.70	17.61	70.66	30.04	50.68	24.67
GCE-GNN	54.02	19.04	70.91	30.63	53.96	24.53
IM-GGN	73.92	30.28	82.33	52.34	68.49	48.49

SPL 的 GSPN 模块和 UPL 模块生成物品表示;

(4) w/o SPL: 去掉整个 SPL 模块, 仅使用 UPL 模块验证结构化关系学习模块的作用;

(5) w/o UPL: 去掉 UPL 模块, 仅使用 SPL 模块验证非结构化关系学习模块的作用。

结果(表 3)表明, MLP-SR 在 3 个数据集上几乎都优于序列模型, 原因可能在于, 简单的 MLP 就可以从较短的会话中捕获到全局的信息, 而序列模型在面对较长的会话序列时可能会更有效, 并且所有模块都是有效的, 因为去除任何一个模块, 都会对模型结果造成影响。

表 3　消融实验结果

Table 3　Result of ablation experiment

Dataset	MLP-SR		w/o GSPN		w/o SM-GNN		w/o SPL		w/o UPL		IM-GGN	
	HR@20	MRR@20	HR@20	MRR@20	HR@20	MRR@20	HR@20	MRR@20	HR@20	MRR@20	HR@20	MRR@20
Diginetica	49.55	15.89	48.81	15.75	56.50	20.94	56.69	21.42	72.44	28.22	73.92	30.28
Yoochoose1/64	69.21	29.76	68.99	29.91	70.63	30.79	69.96	30.35	80.50	51.42	82.33	52.34
Gowalla	47.52	22.22	49.85	23.45	52.95	26.04	53.36	26.25	67.80	47.22	68.49	48.49

从实验结果中发现, UPL 模块的影响性相对较小, 这可能是由于模块之间功能重叠导致。UPL 模块的目标是通过捕捉非结构化的全局上下文信息来提升推荐性能, 然而 SPL 和 SM-GNN 模块已经能够较好地建模局部和结构化的关系。这种功能上的部分重叠可能导致 UPL 模块在某些简单场景中的边际收益减小。同时, UPL 模块的全局上下文信息是基于非结构化关系进行建模的, 而推荐任务中结构化的局部信息(如项目之间的顺序依赖和邻接关系)通常更具有决定性。在数据分布简单的场景下, 模型对非结构化信息的依赖性较低, 但其在复杂场景(如 Diginetica 数据集)中的表现证明了其必要性。

3　结　论

文章提出了一种基于图神经网络的推荐模型 IM-GGN, 融合了结构化和非结构化关系建模, 同时结合长短期兴趣建模、会话图神经网络和门控融合机制, 在一定程度上提升了推荐性能。实验表明, IM-GGN 在多项评估指标上优于现有方法, 特别是在 MRR@20 上表现突出, 证明了其在捕捉用户兴趣和会话信息方面的有效性。尽管取得了良好效果, 但模型在数据稀疏性和冷启动场景下仍需优化, 未来可考虑引入多模态信息以增强特征表达能力。同时, 如何提高模型效率、降低计算复杂度也是未来的挑战。总体而言, IM-GGN 为会话推荐提供了新的思路, 未来将优化模型结构, 并在真实场景中验证其应用价值。

参考文献:

[1] WANG S, CAO L, WANG Y, *et al*. A survey on session-based recommender systems[J]. ACM Computing Surveys, 2021, 54(7): 1-38.

[2] QIAO S, ZHOU W, WEN J, *et al*. Multi-perspective enhanced representation for effective session-based recommendation[J]. Knowledge-Based Systems, 2023, 263: 110284.

[3] 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.

[4] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[5] HIDASI B, KARATZOGLOU A, BALTRUNAS L, *et al*. Session-based recommendations with recurrent neural networks[EB/OL]. (2015-11-21) [2024-10-23]. https://arxiv.org/abs/1511.06939.

[6] PENG D, YUAN W, LIU C. HARSAM: A hybrid model for recommendation supported by self-attention mechanism[J]. IEEE Access, 2019, 7: 12620-12629.

[7] 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. New York: ACM, 2017: 1419-1428.

[8] LIU Q, ZENG Y, MOKHOSI R. 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, UK: ACM, 2018: 1831-1839.

[9] LI Y, ZEMEL R, BROCKSCHMIDT M, *et al*. Gated graph sequence neural networks[EB/OL].(2015-11-17) [2023-05-16]. https://arxiv.org/abs/1511.05493.

[10] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, *et al*. Graph attention networks[EB/OL]. (2017-10-30) [2023-06-

20]. https://arxiv.org/abs/1710.10903.
[11] XU K, HU W, LESKOVEC J, *et al*. How powerful are graph neural networks?[EB/OL]. (2018-10-01) [2024-03-21]. https://arxiv.org/abs/1810.00826.
[12] WU S, TANG Y, ZHU Y, *et al*. Session-based recommendation with graph neural networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. USA: AAAI, 2019: 346-353.
[13] WANG W, ZHANG W, LIU S, *et al*. Incorporating link prediction into multi-relational item graph modeling for session-based recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(3): 2683-2696.
[14] YIN Z, HAN K, WANG P, *et al*. Multi global information assisted streaming session-based recommendation system[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(8): 8245-8256.
[15] 郑楠, 过弋, 李智强, 等. 融合交互注意力和参数自适应的商品会话推荐 [J]. 中文信息学报, 2022, 36(11): 131-139.
[16] PAN Z, CAI F, CHEN W, *et al*. Star graph neural networks for session-based recommendation[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 1195-1204.
[17] XU C, ZHAO P, LIU Y, *et al*. Graph contextualized self-attention network for session-based recommendation[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI, 2019: 3940-3946.
[18] QIU R, LI J, HUANG Z, *et al*. Rethinking the item order in session-based recommendation with graph neural networks[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM, 2019: 579-588.
[19] MA T, CHENG Y, ZHU H, *et al*. Large language models are not stable recommender systems[EB/OL]. (2023-12-25) [2025-04-15]. https://arxiv.org/abs/2312.15746.
[20] ZHOU Z, NING X, HONG K, *et al*. A survey on efficient inference for large language models[EB/OL]. (2024-04-22) [2025-04-15]. https://arxiv.org/abs/2404.14294.
[21] YUAN J, SONG Z, SUN M, *et al*. Dual sparse attention network for session-based recommendation[C]// Proceedings of the AAAI conference on artificial intelligence. USA: AAAI, 2021: 4635-4643.
[22] LUO A, ZHAO P, LIU Y, *et al*. Collaborative self-attention network for session-based recommendation[C]// Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: AAAI, 2020: 2591-2597.
[23] BRODY S, ALON U, YAHAV E. How attentive are graph attention networks?[EB/OL]. (2021-05-30) [2023-05-23]. https://arxiv.org/abs/2105.14491.

Graph Neural Network Session Recommendation Model with Fusion Modeling

DU Jiayu[1], ZHENG Hong[1], GUO Jinyan[1], LUO Yujian[1], LI Pengwei[1], SHAN Rongsheng[2]
(1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China; 2. School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: To address the limitations of traditional session recommendation algorithms that rely solely on explicit information while overlooking potential interactions between sessions, this paper proposes a novel integrated modeling approach based on gating mechanisms and graph attention networks—IM-GGN (Integrated Modeling Gated Graph Network). This model simultaneously captures structured relationships between items and unstructured associations across sessions to enhance recommendation performance. Specifically, the model comprises two main components: the Structured Pattern Learning (SPL) module and the Unstructured Pattern Learning (UPL) module. The SPL module integrates graph neural networks with gating mechanisms to model sequential dependencies and long-range relationships within sessions. Meanwhile, the UPL module leverages graph attention mechanisms to capture unstructured inter-session correlations, thereby extracting contextual user preferences. Experimental results on multiple public datasets demonstrate that the proposed method achieves notable performance improvements, confirming its effectiveness in session-based recommendation tasks.

Key words: session recommendation; gated graph neural network; graph attention mechanism; structured relationship; unstructured relationship

（责任编辑：王晓丽）