中国科学: 信息科学 2024 年 第 54 卷 第 9 期: 2116–2135
SCIENTIA SINICA Informationis
论文

基于上下文增强的多级注意力会话推荐模型

曾碧卿*, 池俊龙, 陈嘉涛, 谢梁琦

华南师范大学软件学院, 佛山 528225
* 通信作者. E-mail: zengbiqing@scnu.edu.cn

收稿日期: 2023–04–18; 修回日期: 2023–09–12; 接受日期: 2023–10–04; 网络出版日期: 2024–09–10

国家自然科学基金面上项目 (批准号: 62271217)、国家自然科学基金项目 (批准号: 12001208)、广东省基础与应用基础研究基金项目 (批准号: 2021A1515011171)、广东省普通高校人工智能重点领域专项 (批准号: 2019KZDZX1033) 和广州市基础研究计划基础与应用基础研究项目 (批准号: 202102080282) 资助

摘要 会话推荐的目标是仅根据用户在匿名会话中有限的交互行为, 来预测用户的下一次点击行为. 最近几年, 许多基于图神经网络的会话推荐方法取得了可喜的结果. 然而, 这些方法仍然存在不足之处. 一方面, 基于图神经网络的方法只考虑物品之间的转换模式, 忽略了会话中的序列模式. 另一方面, 现有的大多数方法都只关注当前会话内部的信息, 忽略了来自邻居会话的外部协作信息, 即上下文模式. 为了解决上述问题, 本文提出了一种新颖的基于上下文增强的多级注意力会话推荐模型 (CEMA), 通过多级注意力机制分别在物品级和会话级这两个粒度上学习物品特征和建模用户偏好, 以增强模型的个性化推荐能力. CEMA 模型利用多层 GraphSAGE 来学习物品之间复杂的转换模式, 以捕获用户的局部偏好. 特别地, 在 CEMA 模型中设计了一种物品级注意力机制, 通过门控注意力单元计算会话中不同物品的重要性, 以识别用户真正感兴趣的物品, 避免噪声物品的干扰. 这有助于准确地捕获会话的序列模式, 以建模用户的全局偏好. 此外, 所提出的方法还设计了一种会话级注意力机制, 通过简单的软注意力高效地计算不同会话之间的相似度, 以聚焦于那些与当前会话最相似的邻居会话, 并从中提取上下文模式, 以帮助预测用户的下一次点击. 本文在 3 个公开的基准数据集上进行了一系列实验, 实验结果表明 CEMA 的推荐性能超过了现有最好的方法, 充分验证了 CEMA 的有效性和优越性.

关键词 会话推荐, 多级注意力机制, 图神经网络, 序列模式, 上下文模式

1 引言

随着大数据时代的发展, 互联网上的多媒体信息不断增长, 给信息筛选和处理带来了挑战. 为解决信息过载问题, 推荐系统成为各个领域的重要工具, 主要用于高效地识别用户感兴趣的信息 [1]. 然而, 传统推荐系统过度依赖于用户身份标识和长期历史记录 [2]. 在实际场景中, 出于隐私和安全的考

引用格式: 曾碧卿, 池俊龙, 陈嘉涛, 等. 基于上下文增强的多级注意力会话推荐模型. 中国科学: 信息科学, 2024, 54: 2116–2135, doi: 10.1360/SSI-2023-0104
Zeng B Q, Chi J L, Chen J T, et al. Context enhanced multi-level attention model for session-based recommendation (in Chinese). Sci Sin Inform, 2024, 54: 2116–2135, doi: 10.1360/SSI-2023-0104

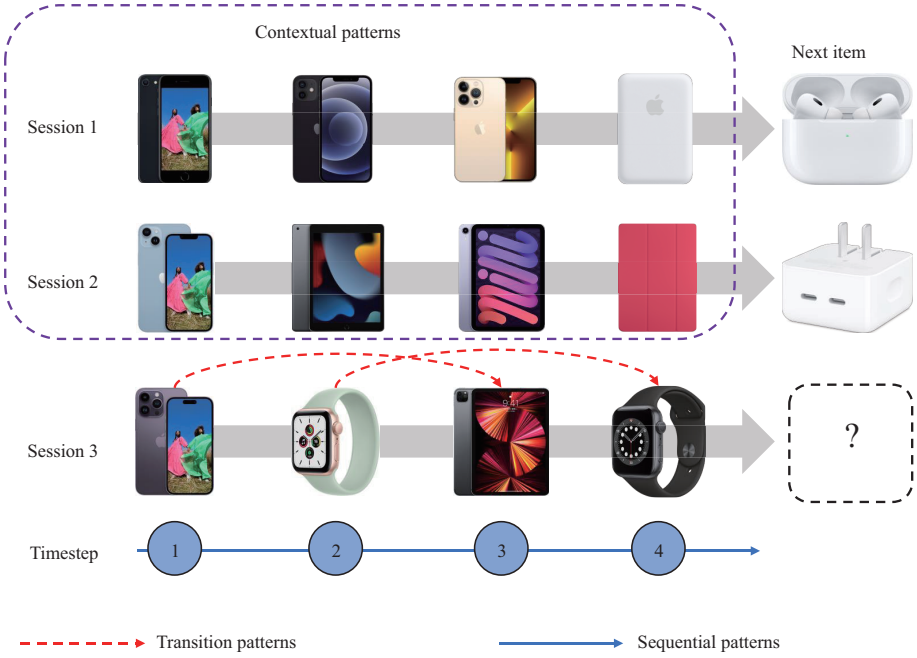

图 1 (网络版彩图) 转换模式、序列模式和上下文模式的示例

Figure 1 (Color online) Illustration of transition relation, sequential information and context information

虑 [3], 一些用户更倾向于匿名访问网站, 导致用户身份标识和历史记录不可用, 严重限制了推荐系统的性能. 近年来, 基于会话的推荐系统 (session-based recommendation, SBR) 逐渐崭露头角, 它仅根据用户的有限信息 (如用户点击) 来捕获用户偏好, 以做出精确的推荐, 有效地降低对用户身份标识和长期历史记录的依赖.

早期的研究 [4] 主要采用马尔可夫链 (Markov chain, MC) 和 K 最近邻方法 (k-nearest neighbor, KNN) 建模序列信息和物品共现模式. 但这些方法只能捕获相邻物品之间的信息, 无法探索用户的兴趣迁移模式, 因此性能受到限制. 基于循环神经网络 (recurrent neural network, RNN) [5,6] 的方法可以建模时序依赖关系, 但仅能捕获短序列中较简单的序列信息, 无法探索长序列中的远距离依赖关系. 并且某些 RNN 变体计算复杂度较高, 导致运行效率低下. 注意力机制 (attention mechanism, AM) 也可用于序列建模, 但 AM 方法 [7] 主要关注会话的序列信息, 忽略了物品之间丰富的转换关系. 而胶囊网络 (capsule network, CN) [8] 则是利用其空间感知能力, 从多个层面细粒度建模用户兴趣. 但胶囊网络的缺点是模型架构复杂, 时间复杂度高, 计算开销很大.

值得注意的是, 序列模式和转换模式对于预测用户兴趣同样重要 [9]. 如图 1 所示, 序列模式是指会话的序列位置信息和长期依赖关系, 它反映了会话的全局依赖性. 而转换模式则是物品之间丰富的转换关系, 它反映了会话的局部依赖性. 图神经网络 (graph neural network, GNN) [10] 的快速发展为捕捉物品之间的转换关系提供了有效方法, 然而, 当前的 GNN 方法 [11,12] 仅考虑相邻物品之间的一阶转换关系, 无法捕捉非相邻物品之间的高阶转换信息. 此外, 在会话图中缺乏序列位置信息的情况下, 不同的会话可能会被转换为相同的图结构, 导致序列模式难以被捕获. 此外, 现有的大多数方法都仅关注当前会话的内部信息, 忽略了邻居会话提供的外部协作信息, 即上下文模式. 如图 1 所示, 具有相似用户偏好的邻居会话提供的上下文模式有助于预测当前会话的下一个点击物品.

由上述可知, 当前的 AM 方法聚焦于捕获会话的序列模式, 忽略了物品之间丰富的转换关系. 而

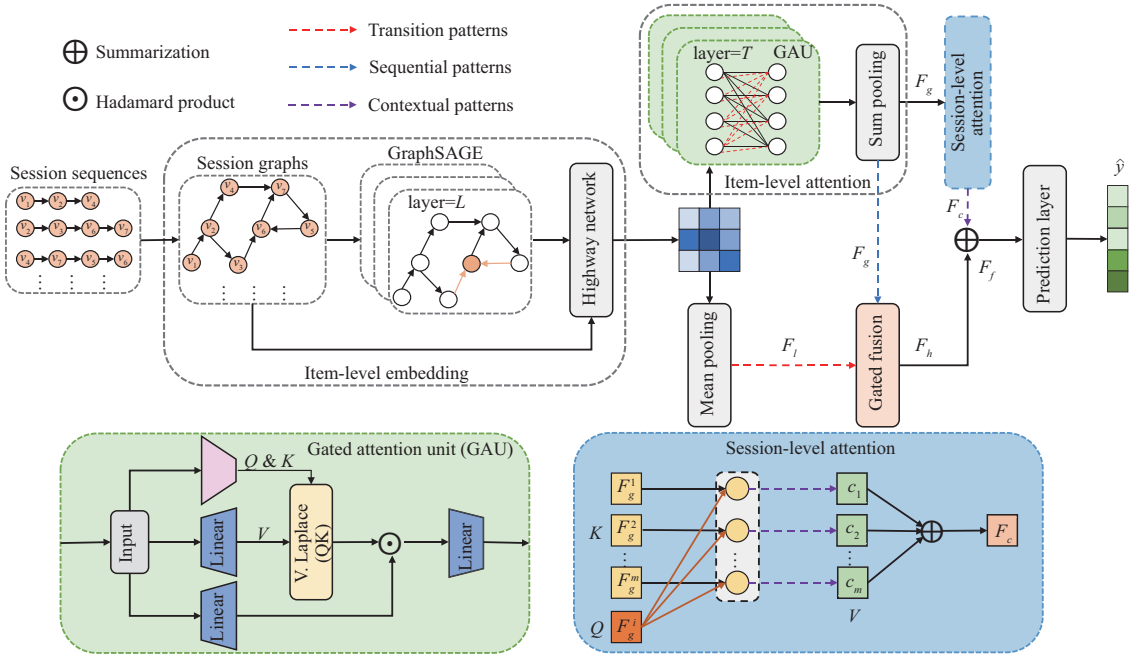

图 2　(网络版彩图) CEMA 模型架构

Figure 2　(Color online) Overall framework of CEMA

GNN 方法则主要关注物品之间的转换模式, 无法建模会话的序列位置信息和长期依赖关系. 此外, 无论是 AM 方法还是 GNN 方法, 绝大多数都忽略了邻居会话中有协作价值的上下文信息, 无法获取更准确的用户偏好表示, 进而限制了模型的推荐性能.

为了克服上述限制, 本文提出了一种新颖的基于上下文增强的多级注意力模型 (context enhanced multi-level attention model, CEMA). 不同于现有的方法, CEMA 同时结合了图神经网络和注意力机制各自的优势, 它应用图神经网络建模会话中物品之间的复杂转换关系. 特别地, 在 CEMA 模型中设计了多级注意力机制, 即物品级和会话级注意力机制, 从物品和会话这两个层级, 分别捕获会话内部的序列模式和不同会话之间的上下文模式. 如图 2 所示, CEMA 模型由 3 个部分组成, 即物品级嵌入模块 (item-level embedding module, ILE)、物品级注意力模块 (item-level attention module, ILA) 和会话级注意力模块 (session-level attention module, SLA). 首先, ILE 模块通过多层图采样聚合网络 (graph sample and aggregate network, GraphSAGE) 捕获物品之间的复杂转换关系, 以生成准确的物品嵌入表示. 接着, 在 ILA 模块中引入门控注意力单元 (gated attention unit, GAU), 通过门控注意力计算会话中不同物品的重要性, 聚焦于用户真正感兴趣的物品, 以更好地建模会话的序列模式, 获取更准确的当前会话表示. 此外, 在 SLA 模块中设计了一种简单且高效的软注意力机制, 它可以计算不同会话之间的相似性, 从而区分不同邻居会话的重要性, 并高效地从邻居会话中捕获上下文模式. 最后, 结合当前会话表示和上下文表示预测用户下一次点击的物品, 以增强模型的个性化推荐效果. 在 3 个公开的基准数据集上进行的实验结果表明, CEMA 的性能超过了现有的最佳基线方法, 充分证明了 CEMA 的有效性和优越性.

本文工作的主要贡献点如下:

• 提出了一个新颖的 CEMA 模型, 它结合了图神经网络和注意力机制各自的优势, 充分地利用并整合会话内部的转换模式和序列模式, 以及不同会话之间的上下文模式, 增强了会话推荐性能.

• 提出了一种物品级注意力机制, 通过应用强大且高效的门控注意力单元 (GAU) 计算会话中不同物品的重要性, 聚焦用户真正感兴趣的物品, 以帮助捕获用户的真实意图, 从而得到更为准确的长期序列依赖关系. 本文是第 1 个在会话推荐任务中应用 GAU 捕获会话序列信息的工作.

• 设计了一种会话级注意力机制, 它是一种简单高效的软注意力机制, 可以快速地计算不同会话之间的相似性, 并有效地从邻居会话中提取上下文信息.

• 在 Diginetica, Retailrocket 和 Tmall 这 3 个公开的基准数据集上进行了详细的实验和分析, 实验结果表明 CEMA 的性能超过了现有的最佳会话推荐方法, 证明了 CEMA 的有效性和优越性.

2 相关工作

会话推荐算法主要分为传统方法、基于循环神经网络的方法、基于注意力机制的方法和基于图神经网络的方法.

2.1 传统方法

传统的会话推荐方法主要包括基于马尔可夫链 (MC) 的推荐方法和基于 K 最近邻 (KNN) 的推荐方法.

一些较早推荐方法, 如 S-POP [13], 主要根据全局范围中最流行的物品为用户做推荐. 而基于 MC 的方法则把会话序列映射到一个马尔可夫链中, 然后根据用户的前一个点击行为推断出用户的下一个点击行为. Rendle 等 [14] 提出了一种把矩阵分解和一阶马尔可夫链相结合的推荐算法 (FPMC), 以捕获序列依赖关系和用户的长期偏好. 基于马尔可夫链的方法通常侧重于对两个相邻物品的序列转换关系进行建模, 这导致了它难以捕获长距离的序列依赖关系.

基于 KNN 的方法通过推荐与当前会话的最后一个物品最相似的物品来扩展到基于会话的推荐场景. Item-KNN [15] 从全局范围内筛选出与当前会话的最后一个物品最相似的 K 个物品, 作为候选物品推荐给用户. 最近几年, 出现了利用整个会话来推荐的扩展型 KNN 方法. Jannach 等 [16] 提出的 SKNN 模型, 通过探索那些含有当前会话物品的邻居会话为用户做推荐. 随后, Garg 等 [17] 在 SKNN 的基础进行了改进, 提出了序列和时间感知的 STAN 模型, 额外考虑了位置信息等因素. 但这些方法没有考虑会话中的序列依赖关系, 导致它们的性能仍然受限.

2.2 基于循环神经网络的方法

由于会话推荐任务可以看成是一个序列预测任务, 于是研究者把 RNN 引入到会话推荐任务, 提出了基于 RNN 的会话推荐方法. Hidasi 等 [18] 提出了 GRU4Rec, 它是第 1 个把 RNN 应用于会话推荐任务的模型. GRU4Rec 模型采用多层门控循环单元 (gated recurrent unit, GRU) 建模物品交互序列. Li 等 [19] 提出了 NARM 模型, 它将注意力集中整合到多层堆叠的 GRU 网络中, 以捕获会话推荐任务中更有表征意义的物品转换信息. Wang 等 [20] 提出了一个混合架构 CSRM, 它同时考虑当前会话信息和邻居会话信息, 并利用 RNN 预测当前会话的用户偏好. Wang 等 [21] 提出了循环记忆网络 (RMN), 通过刻画用户长期和局部偏好进行推荐. 然而, 大多数基于 RNN 的方法只能粗略学习短会话的序列信息, 无法在长会话中捕获远距离依赖关系. 此外, RNN 是串行结构, 无法并行计算, 容易带来较高的计算开销, 导致模型训练效率低下.

2.3 基于注意力机制的方法

最近, 注意力机制, 特别是 Transformer [22] 架构中的自注意力机制, 在序列建模方面表现出了强

大的能力并且取得了良好的效果. Liu 等 [23] 提出了一种基于注意力的短期记忆网络 (STAMP), 它可以在不使用 RNN 的情况下捕获用户的当前兴趣. 受到预训练模型 BERT (bidirectional encoder representations from transformers) [24] 的启发, Sun 等 [25] 提出了 BERT4Rec, 利用深度双向自注意力建模用户的行为序列. Luo 等 [26] 提出了协作自注意力网络 (CoSAN), 通过探索邻居会话预测当前会话中的用户意图. Zhou 等 [27] 基于自注意力网络架构提出了 S^3-Rec 模型, 它利用内在的数据相关性获得自监督信号, 并通过预训练的方法增强数据表示. Yuan 等 [28] 设计了一个双稀疏注意力网络 (DSAN), 结合稀疏自注意力和目标注意力捕获用户兴趣并消除无关物品的影响. Zhao 等 [29] 提出了 RecBole 推荐系统库, 它是一个统一、综合且高效的框架, 可用于开发和复现研究目的的推荐算法. Zhang 等 [30] 提出了一种多层次的注意力网络 (Atten-Mixer), 它利用多层次的用户意图来提高推荐性能. 尽管基于 AM 的方法取得了非常不错的成果, 但它们缺乏建模物品之间复杂转换关系的能力, 从而限制了推荐模型的性能.

2.4 基于图神经网络的方法

最近, 一些学者建议把图神经网络引入到会话推荐任务中, 把会话序列构建为会话图结构, 然后使用图神经网络学习会话的物品嵌入特征, 以建模用户偏好表示. Wu 等 [31] 提出了 SR-GNN 模型, 把每一个会话序列都转换为一个会话图, 并应用门控图神经网络捕获会话图中物品之间的转换关系. 随后, Xu 等 [32] 提出了 GC-SAN 模型, 把自注意力机制和 GNN 相结合, 取得了更好的效果. Wang 等 [33] 提出了 GCE-GNN 模型, 通过学习所有会话的物品转换关系捕获用户兴趣. Xia 等 [34] 提出了 S^2-DHCN 模型, 利用超图卷积网络建模会话物品之间复杂的高阶转换信息, 并使用自监督学习增强超图建模效果. Xia 等 [35] 提出了 COTREC 模型, 结合自监督学习和协同训练进行数据增强, 提升会话推荐效果. Zhang 等 [36] 提出了图邻域路由随机游走模型 (GNRRW), 通过物品在所有会话中的出现频率来聚合物品的局部嵌入和全局嵌入. Yan 等 [37] 提出了 IHGCN 模型, 通过分析会话的商品交互序列, 建模物品之间的转换关系, 同时探索了历史交互信息和相邻交互信息. Feng 等 [38] 提出了 GNN-GNF 模型, 通过过滤噪声数据并以一种更全面合理的方式捕获物品之间的转换模式. Chen 等 [39] 设计了一种基于图神经网络的自动搜索框架 AutoGSR, 它可以自动搜索并找到基于图神经网络的最优会话推荐模型. Tang 等 [40] 利用时间增强的图神经网络 (TE-GNN) 学习物品嵌入, 以此捕获会话内复杂的用户兴趣转换模式. Yang 等 [41] 提出了一种去偏差对比学习范式的推荐框架 (DCRec), 它通过自适应的一致性感知增强, 将序列模式编码与全局协同关系建模统一起来. 尽管基于 GNN 的推荐方法已经取得了良好的进展, 但它们主要关注当前会话中相邻物品之间的一阶转换关系, 忽略了非相邻物品之间的高阶转换信息和来自相邻会话的上下文信息的影响.

为了克服上述限制, 本文提出一种新颖的架构 CEMA, 它不仅极大地受益于物品之间的复杂转换关系和每个会话中的序列模式, 而且还可以有效地利用不同会话之间的拓扑上下文信息, 以更好地预测用户的下一个点击物品.

3 模型方法

本小节主要介绍会话推荐任务的相关概念, 以及 CEMA 模型的组件架构和运行原理. 如图 2 所示, CEMA 模型由 3 个模块组成: 物品级嵌入模块 (ILE)、物品级注意力模块 (ILA) 和会话级注意力模块 (SLA), 这 3 个模块分别用于捕获转换模式、序列模式和上下文模式.

表 1 符号汇总说明表

Table 1 Summary and description of notations

Notation	Description
V	Item set
$\|V\|$	Number of all unique items
S	Session set
$\|S\|$	Number of all sessions
s	Session
v_i	The i-th item in the session
n	The length of session
$\hat{y}_i$	Click probability of the i-th item in the item set
y_i	Ground truth label of the i-th item in the item set
$\mathcal{G}_s$	Session graph
$\mathcal{V}_s$	Node set of session graph
$\mathcal{E}_s$	Edge set of session graph
$\boldsymbol{A}_{\text{in}}$	Incoming matrix
$\boldsymbol{A}_{\text{out}}$	Outgoing matrix
L	Number of layers in GraphSAGE
T	Number of layers in GAU
F_l	User's local preference representation
F_g	User's global preference representation
F_h	User's hybrid preference representation
F_c	Context representation of session

3.1 符号定义及问题描述

本文使用粗体大写字母表示矩阵, 小写字母表示向量. 表 1 汇总了一些重要的数学符号.

会话推荐的主要目的是仅根据会话序列数据预测用户最可能点击的下一个物品, 而不依赖于任何的用户身份标识和长期历史交互记录. 令 $V = \{v_i\}_{i=1}^{|V|}$ 表示所有会话中包含的全部物品组成的集合, 即物品集, 其中 $|V|$ 是全部物品的数量. $S = \{s_i\}_{i=1}^{|S|}$ 表示所有会话组成的集合, 即会话集, 其中 $|S|$ 是所有会话的数量. 对每一个会话序列 $s \in S$, 都可以表示为一个列表 $s = [v_1, v_2, \ldots, v_n]$. 其中, v_i 表示用户点击的第 i 个物品. 物品 $v_{n+1} \in V$ 表示要预测的用户下一次可能点击的物品. 对于每个会话 s, 所有物品的点击概率设为 $\hat{y} = \{\hat{y}_i\}_{i=1}^{|V|}$. 其中, 向量 $\hat{y}_i$ 的元素值是物品集 V 中第 i 个物品的推荐分数. 在本文的实验中, 分数最高的 K 个物品将作为候选物品推荐给用户.

3.2 会话图的构建

每一个会话序列 s 都可以转换为一个有向图 $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, 其中 $\mathcal{V}_s = \{v_1, v_2, \ldots, v_m\}$ 表示节点集合, 即会话 s 中出现的物品. 因为会话序列中可能由物品重复出现, 而节点集合中的物品都是唯一不重复的, 所以节点数量 m 小于等于会话序列的物品数量 n. 更多地, 每一条边 $(v_{i-1}, v_i) \in \mathcal{E}_s$ 表示用户先点击了物品 v_{i-1} 后, 接着点击了 v_i. 考虑到会话中有重复出现的物品, 需要对每一条边进行归一化处理. 通过每一条边的起始节点的出度除以该边的出现次数, 得到该边对应的权重. 每一个物品 $v \in V$ 都会被嵌入到一个统一的嵌入空间中, 以此得到物品节点的向量表 $x \in \mathbb{R}^d$, 其中 d 是嵌入维度.

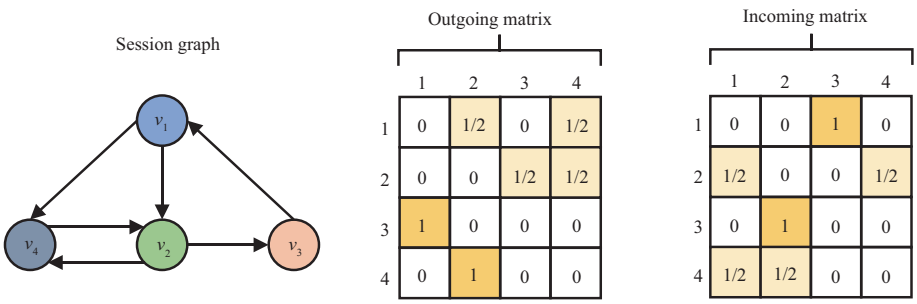

图 3 (网络版彩图) 会话图及其邻接矩阵示例

Figure 3 (Color online) Example of a session graph and the adjacency matrix

然后, 每个会话图 $\mathcal{G}_s$ 都可以表示如下:

$$\boldsymbol{X}_s^{(0)} = [x_1, x_2, \ldots, x_m], \tag{1}$$

其中, $\boldsymbol{X}_s^{(0)} \in \mathbb{R}^{m\times d}$ 表示会话图的初始嵌入表示.

3.3 物品级嵌入模块

3.3.1 图采样聚合网络

物品之间复杂的转换关系, 通常反映着用户的动态兴趣迁移模式, 亦即会话的局部依赖关系. 在 ILE 模块中, 通过堆叠多层 GraphSAGE 捕获物品之间的一阶转换关系和高阶转换关系, 以此学习物品的图嵌入特征表示. 由于原始的 GraphSAGE 在邻居采样后是直接等权聚合邻居特征来更新当前节点的特征, 这导致 GraphSAGE 无法处理加权图, 从而无法进一步探索复杂的加权边连接关系. 为了解决这个问题, 如图 3 所示, 本文把会话物品图中的边分为入边和出边两种类型, 则对应的邻接矩阵被分为入度矩阵和出度矩阵, 以此探索更丰富的物品转换关系.

以会话物品图为输入, 则多层 GraphSAGE 的逐层节点更新机制如下:

$$\boldsymbol{X}_s^{(l)} = \boldsymbol{X}_s^{(l-1)}\boldsymbol{W}_1 + \text{mean}([(\boldsymbol{A}_{\text{in}} + \boldsymbol{A}_{\text{out}})\boldsymbol{X}_s^{(l-1)}]\boldsymbol{W}_2), \tag{2}$$

其中, $\boldsymbol{X}_s^{(l)} \in \mathbb{R}^{m\times d}$ 表示第 l 层节点特征, $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{d\times m}$ 表示可训练的参数矩阵, $\boldsymbol{A}_{\text{in}}, \boldsymbol{A}_{\text{out}} \in \mathbb{R}^{m\times m}$ 分别表示会话图的入度矩阵和出度矩阵, mean 表示平均池化. 为了提升模型的非线性表达能力和稳定性, 应用 ReLU 激活函数和 Dropout 函数到每一层 GraphSAGE 的输出

$$\widehat{\boldsymbol{X}_s^{(l)}} = \text{Dropout}(\text{ReLU}(\boldsymbol{X}_s^{(l)})), \tag{3}$$

接着, 通过堆叠 L 层 GraphSAGE 捕获物品之间的高阶转换关系, 以生成更准确的物品嵌入表示

$$\boldsymbol{X}_s^{(L)} = f^L_{\text{GraphSAGE}}(\boldsymbol{A}_s, \boldsymbol{X}_s^{(0)}), \tag{4}$$

其中, $f^L_{\text{GraphSAGE}}$ 表示 L 层 GraphSAGE 的特征学习器, $\boldsymbol{X}_s^{(L)}$ 表示最后一层 GraphSAGE 的输出, $\boldsymbol{A}_s = [\boldsymbol{A}_{\text{in}}||\boldsymbol{A}_{\text{out}}]$ 表示会话图的邻接矩阵.

3.3.2 高速公路网络

多层图神经网络在学习图结构数据时很容易发生过平滑问题. 为了避免这个问题, 本文在 L 层 GraphSAGE 之后引入了高速公路网络 (highway network, HN). HN 首先根据未经 GNN 学习的初始

物品嵌入表示 $\boldsymbol{X}_s^{(0)}$ 和经过 GNN 学习到的物品嵌入表示 $\boldsymbol{X}_s^{(L)}$ 得到门控向量 $\boldsymbol{G}_x$, 然后再通过门控向量自适应地从 $\boldsymbol{X}_s^{(0)}$ 和 $\boldsymbol{X}_s^{(L)}$ 中选择有效的信息特征. HN 的计算过程如下:

$$\boldsymbol{G}_x = \sigma([\boldsymbol{X}_s^{(0)}||\boldsymbol{X}_s^{(L)}]\boldsymbol{W}_g), \tag{5}$$

$$\widehat{\boldsymbol{X}}_s^{(L)} = \boldsymbol{G}_x \odot \boldsymbol{X}_s^{(0)} + (1-\boldsymbol{G}_x) \odot \boldsymbol{X}_s^{(L)}, \tag{6}$$

其中, $[\cdot||\cdot]$ 表示拼接操作, $\boldsymbol{W}_g \in \mathbb{R}^{2d\times d}$ 表示可训练的参数矩阵, 用于把拼接后的向量维度从 $2d$ 转换为 d. $\sigma(\cdot)$ 表示 sigmoid 激活函数, $\odot$ 表示哈达玛乘积 (Hadamard product). 最后, 基于 GraphSAGE 学习到的物品节点向量 $\widehat{x}_i \in \widehat{\boldsymbol{X}}_s^{(L)}$, 每个会话序列 s 可表示为

$$\widehat{\boldsymbol{X}}_s = [\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n], \tag{7}$$

其中, $\widehat{\boldsymbol{X}}_s \in \mathbb{R}^{n\times d}$ 表示经过 GraphSAGE 学习后的会话序列的嵌入表示.

3.4 物品级注意力模块

会话中的序列位置信息和全局依赖关系, 通常反映了用户在当前会话中相对稳定的全局偏好. 而用户在会话序列中点击过的物品, 有些是用户真正感兴趣的重要物品, 有些则是用户随意点击甚至是误点的物品. 在捕获会话的序列模式时, 除了需要关注物品的顺序位置, 还需特别区分不同物品的重要性, 以获取更为准确的序列依赖关系. 本文设计了一个物品级注意力模块 (ILA), 通过应用门控注意力单元 (GAU) 计算会话中不同物品的重要性, 聚焦于用户真正感兴趣的重要物品, 忽略那些无关紧要的物品, 以更好捕获会话中的长期序列依赖关系, 生成更准确的用户全局偏好表示.

3.4.1 门控注意力单元

GAU 是 Hua 等 [42] 提出的用于代替 Transformer 模型中的多头自注意力 (multi-head self-attention, MHSA) 和多层感知机 (multilayer perceptron, MLP) 的新范例, 它由高效的门控线性单元 (gated linear unit, GLU) 和简化的单头自注意力构成, 不仅减少了对多头注意力的依赖, 还降低了计算参数量, 同时极大地提高了性能. GLU 是 Dauphin 等 [43] 提出的利用门控机制进行性能增强的改进版 MLP, 其计算公式如下:

$$\boldsymbol{U} = \phi_u(\widehat{\boldsymbol{X}}_s\boldsymbol{W}_u), \tag{8}$$

$$\boldsymbol{V} = \phi_v(\widehat{\boldsymbol{X}}_s\boldsymbol{W}_v), \tag{9}$$

$$\boldsymbol{H} = (\boldsymbol{U} \odot \boldsymbol{V})\boldsymbol{W}_h, \tag{10}$$

其中, $\phi_u(\cdot), \phi_v(\cdot)$ 表示激活函数, $\boldsymbol{W}_u, \boldsymbol{W}_v, \boldsymbol{W}_h \in \mathbb{R}^{d\times n}$ 表示可训练的参数矩阵, $\odot$ 表示逐位对应相乘 (哈达玛乘积).

值得指出的是, GLU 是改进版的 MLP, 因此它的各个 token 之间没有进行交互, 也就是矩阵 U, V 的每一行都是独立计算的. 而 GAU 在 GLU 的基础上, 对式 (10) 进行了修改, 引入了注意力权重矩阵, 把每个 token 之间的交互补充到 $\boldsymbol{U}, \boldsymbol{V}$ 上去:

$$\boldsymbol{H} = (\boldsymbol{U} \odot \boldsymbol{A}\boldsymbol{V})\boldsymbol{W}_h, \tag{11}$$

其中, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 表示注意力权重矩阵. GLU 中的门控机制也可以有效降低对多头注意力的依赖, 从而减少参数和计算量. 因此, GAU 可以只使用一个简化的单头注意力, 而性能几乎不会有损失:

$$\boldsymbol{Z} = \phi_z(\widehat{\boldsymbol{X}}_s\boldsymbol{W}_z), \tag{12}$$

$$\boldsymbol{B} = \mathcal{Q}(\boldsymbol{Z})\mathcal{K}(\boldsymbol{Z})^{\mathrm{T}} + p, \tag{13}$$

其中, $\phi_z(\cdot)$ 是激活函数, 在 GAU 的实现中, 使用的激活函数是 Swish. $\boldsymbol{W}_z \in \mathbb{R}^{d\times n}$ 表示可训练的参数矩阵, $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ 是一个共享的表示. $\mathcal{Q}(\cdot)$, $\mathcal{K}(\cdot)$ 是两个简单的变换函数, 用于对 $\boldsymbol{Z}$ 进行维度缩放和偏移. $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ 表示注意力分数, p 表示相对位置编码. 在本文所提出 CEMA 模型的具体实现中, 使用的是 Su 等 [44] 提出的旋转位置编码 (rotary position embedding, RoPE), 它采用了复数运算的形式, 比普通的相对位置编码和绝对位置编码取得了更好的效果. RoPE 的计算公式如下:

$$p = f_{\mathrm{RoPE}}(z_m, z_n, m-n) = \mathrm{Re}[\mathcal{Q}(z_m)\mathcal{K}(z_n)^* \mathrm{e}^{\mathrm{i}(m-n)\theta}], \tag{14}$$

其中, $z_m, z_n \in \boldsymbol{Z}$, 分别表示 $\boldsymbol{Z}$ 中的第 m 个行向量和第 n 个行向量, 亦即会话序列中第 m 个物品和第 n 个物品的向量表示, $m-n$ 表示它们的欧氏距离. $\mathrm{Re}[\cdot]$ 表示取复数的实部, $\mathcal{K}(z_n)^*$ 表示 $\mathcal{K}(z_n)$ 的共轭复数, i 为虚数单位, e 是自然常数, θ 表示预设的非零常数.

接着, 需要对注意力分数 $\boldsymbol{B}$ 进行权重归一化. 在原始的 GAU 架构中, 使用了 So 等 [45] 提出的 relu^2 函数代替 softmax 函数对注意力分数进行归一化, 并且取得了比 softmax 函数更好的效果. 而在本文所提出的 CEMA 模型中, 则使用了 Ma 等 [46] 提出的 laplace 函数对注意力分数进行归一化, 因为 laplace 函数的性能比 relu^2 更稳定, 更利于模型的训练:

$$\boldsymbol{A} = f_{\mathrm{laplace}}(\boldsymbol{B}; \mu, \sigma) = 0.5 \times \left[1 + \mathrm{erf}\left(\frac{\boldsymbol{B}-\mu}{\sigma\sqrt{2}}\right)\right], \tag{15}$$

其中, $\mathrm{erf}(\cdot)$ 表示误差函数, μ, σ 是两个系数, 令 $\mu = \sqrt{1/2}$, $\sigma = \sqrt{1/4\pi}$ 即可使 laplace 函数近似于 relu^2 函数. 在 CEMA 中, 通过堆叠 T 层 GAU 来捕获会话序列中的全局依赖关系, 并在每一层 GAU 的输入前使用层归一化 LayerNorm, 在每一层输出后使用残差连接:

$$\widehat{\boldsymbol{H}}_s = \mathrm{GAU}(\mathrm{LayerNorm}(\widehat{\boldsymbol{X}}_s) + \widehat{\boldsymbol{X}}_s), \tag{16}$$

其中, $\widehat{\boldsymbol{H}}_s \in \mathbb{R}^{n\times d}$ 表示通过 GAU 学习到的序列信息表示, 即会话序列中的全局依赖关系.

3.4.2 门控融合机制

转换模式更多地反映了用户在短时间内的局部偏好, 而序列模式则更多地反映了用户在整个会话中的全局偏好. 用户在会话中的交互行为, 通常会受到用户局部和全局偏好的共同影响. 为了获得用户的局部和全局偏好表示, 分别对 GraphSAGE 学习到的局部依赖关系和 GAU 学习到全局依赖关系进行池化操作. 具体来说, 对 GraphSAGE 学习到的局部依赖关系 $\widehat{\boldsymbol{X}}_s$ 进行平均池化, 以得到用户的局部偏好表示

$$\boldsymbol{F}_l = \frac{1}{n}\sum_{i=1}^{n} \widehat{x}_i, \tag{17}$$

其中, $\widehat{x}_i \in \widehat{\boldsymbol{X}}_s$, $\boldsymbol{F}_l \in \mathbb{R}^d$ 是用户的局部偏好表示. 接着, 对 GAU 提取到的全局依赖关系 $\widehat{\boldsymbol{H}}_s$ 进行加和池化, 以获取用户的全局偏好表示

$$\boldsymbol{F}_g = \sum_{i=1}^{n} \widehat{h}_i, \tag{18}$$

其中, $\widehat{h}_i \in \widehat{\boldsymbol{H}}_s$, $\boldsymbol{F}_g \in \mathbb{R}^d$ 表示用户的全局偏好表示. 为了综合考虑用户局部和全局偏好的重要性, 本文应用门控机制来自适应地融合局部偏好表示和全局偏好表示, 进而得到用户的混合偏好表示

$$\boldsymbol{G}_f = \sigma(\boldsymbol{F}_l\boldsymbol{W}_l + \boldsymbol{F}_g\boldsymbol{W}_g), \tag{19}$$

$$\boldsymbol{F}_h = \boldsymbol{G}_f \odot \boldsymbol{F}_l + (1 - \boldsymbol{G}_f) \odot \boldsymbol{F}_g, \tag{20}$$

其中, $\boldsymbol{W}_l, \boldsymbol{W}_g \in \mathbb{R}^{d\times d}$ 表示可训练的参数矩阵, $\sigma(\cdot)$ 表示 sigmoid 激活函数, $\boldsymbol{G}_f \in \mathbb{R}^d$ 表示由 $\boldsymbol{F}_l$ 和 $\boldsymbol{F}_g$ 共同决定的门控向量, $\odot$ 表示哈达玛乘积, $\boldsymbol{F}_h \in \mathbb{R}^d$ 表示用户的混合偏好.

3.5 会话级注意力模块

对于会话推荐任务来说, 那些与当前会话有着相似用户行为的邻居会话, 其用户意图和偏好也是相似的. 这意味着可以通过邻居会话的用户偏好预测当前会话的用户偏好. 现有的方法大都只关注当前会话内部的信息, 忽略了来自邻居会话中有价值的协作信息, 即上下文模式.

为了解决上述问题, 本文在会话级注意力模块中设计了一种简单高效的上下文感知的软注意力机制 (context-aware soft attention, CA-SA), 它通过一个简单的神经网络来动态计算不同会话之间的注意力权重 (即相似度), 以此区分不同邻居会话对于当前会话的重要性. 然后, 通过根据注意力权重进行加权求和以得到上下文信息. CA-SA 可以自适应地给那些与当前会话有着相似用户行为和意图的邻居会话分配更高的注意力权重. 因为越相似的邻居会话, 其重要程度越高, 越能帮助模型预测用户在当前会话中的下一次点击行为. CA-SA 的计算公式如下:

$$\alpha_{ij} = \frac{\exp(q^{\mathrm{T}} \cdot \mathrm{LeakyReLU}(\boldsymbol{W}_a(\boldsymbol{F}_g^i \odot \boldsymbol{F}_g^j)))}{\sum_{i=1}^{b} \exp(q^{\mathrm{T}} \cdot \mathrm{LeakyReLU}(\boldsymbol{W}_a(\boldsymbol{F}_g^i \odot \boldsymbol{F}_g^k)))}, \tag{21}$$

$$\boldsymbol{F}_c^i = \sum_{i=1}^{b} \alpha_{ij} \boldsymbol{F}_g^j, \tag{22}$$

其中, $\boldsymbol{W}_a \in \mathbb{R}^{d\times d}$ 表示可训练的参数矩阵, $\boldsymbol{F}_g^i, \boldsymbol{F}_g^j$ 分别表示会话 i 和会话 j 的全局表示, LeakyReLU 是激活函数, $q \in \mathbb{R}^d$ 表示线性转换向量, α_{ij} 表示会话 i 和会话 j 之间的注意力权重, b 表示训练批次大小, $\boldsymbol{F}_c^i$ 表示会话 i 的上下文信息.

3.6 预测层

用户的混合偏好表示属于会话内部的信息, 而上下文信息则提取自邻居会话, 属于会话外部的信息. 为了增强会话推荐的性能, 需要进一步结合用户的混合偏好表示和上下文信息. 这里, 使用了简单的加和池化聚合方式

$$\boldsymbol{F}_f = \boldsymbol{F}_h + \boldsymbol{F}_c, \tag{23}$$

其中, $\boldsymbol{F}_f$ 表示聚合后的最终会话表示. 接着, 把每个物品的初始嵌入与最终会话表示相乘来得到对应的推荐分数, 再应用 softmax 函数进行归一化, 即可得到每个物品对应的点击概率

$$\widehat{y}_i = \mathrm{softmax}(\boldsymbol{F}_f^{\mathrm{T}} \cdot x_i), \tag{24}$$

其中, $\widehat{y}_i$ 表示物品 i 作为用户下一次点击的物品出现的概率. 损失函数被定义为预测值和真值的交叉熵

$$\mathcal{L}(\widehat{y}) = -\sum_{i=1}^{|V|} (y_i \log(\widehat{y}_i) + (1 - y_i) \log(1 - \widehat{y}_i)), \tag{25}$$

其中, $y_i \in y$ 是真值的独热 (one-hot) 编码向量. 若 $y_i = 1$ 表示物品集中的第 i 个物品是用户下一次点击的真实物品, 若 $y_i = 0$ 则不是.

表 2 数据集统计信息

Table 2 Statistics of the used datasets

	Diginetica	Retailrocket	Tmall
Items	43097	36968	40728
Train	719470	433648	35268
Test	60858	15132	25898
Avg.len	5.12	5.40	6.69

3.7 模型时间复杂度分析

CEMA 模型的运行时间成本主要来自物品级嵌入 (ILE) 模块、物品级注意力 (ILA) 模块和会话级注意力 (SLA) 模块. 其中, ILE 模块的时间计算成本主要来自 GraphSAGE, 其图计算主要涉及线性变换和平均池化, 所以 GraphSAGE 的时间复杂度是线性的, 即 $\mathcal{O}(N)$. ILA 模块的主要时间消耗来自于门控注意力单元 (GAU). GAU 跟标准自注意力类似, 其注意力矩阵 A 仍然是通过执行 Q 和 K 的内积, 然后除以嵌入维度的平方根得到的, 因此时间复杂度仍然是二次方, 即 $\mathcal{O}(N^2)$. 需要注意的是, 与标准自注意力不同的是, GAU 简化了 Q 和 K 的来源变换, 并且采用了 laplace 作为激活函数. 而对于 SLA 模块来说, 其时间消耗主要来自上下文感知的软注意力机制 (CA-SA), 而 CA-SA 主要涉及线性变换和求和运算, 所以其时间复杂度为 $\mathcal{O}(N)$. 因此, CEMA 模型的整体时间复杂度约为 $\mathcal{O}(N)+\mathcal{O}(N^2)+\mathcal{O}(N)\approx\mathcal{O}(N^2)$.

4 实验

4.1 数据集

本文在 3 个公开的基准数据集上评估了所提出的 CEMA 和相关对比方法的性能, 这 3 个数据集分别是 Diginetica[1)], Retailrocket[2)] 和 Tmall[3)]. Diginetica 数据集来自于 CIKM Cup 2016 比赛, 包含典型的用户交易数据. Retailrocket 是由一家电子商务公司发布的 Kaggle 竞赛数据集, 其中包含了用户在六个月内的浏览活动. Tmall 数据集来自于 IJCAI-15 比赛, 其中包含天猫在线购物平台上匿名用户的购物日志. 这 3 个数据集的统计分析结果如表 2 所示.

跟之前的研究工作 [4, 31] 一样, 本文首先对 3 个数据集进行预处理. 具体来说, 在实验前过滤掉数据集中所有长度小于 3 的会话和出现次数少于 5 次的物品. 然后, 将最近一周的会话 (最新数据) 作为测试集, 其他会话作为训练集. 此外, 对于每一个会话 $s=\{v_1,v_2,\ldots,v_n\}$, 使用序列分割方法生成多个序列和相应的标签: $([v_1,v_2],v_3),([v_1,v_2,v_3],v_4),\ldots,([v_1,v_2,\ldots,v_{n-1}],v_n)$. 值得注意的是, 每个序列中最后点击的物品则是相应的真值标签.

4.2 评估指标

本文采用了会话推荐任务中常用的两个评估指标: HR@K (命中率) 和 MRR@K (平均倒数排名). HR@K 是用户真实点击的物品在推荐列表前 K 位中出现的概率, 用来衡量会话推荐模型的推荐准确性. MRR@K 是用户真实点击物品在推荐列表中的位置排名倒数的平均值, 不仅考虑了推荐准确性,

1) CIKM Cup 2016. https://competitions.codalab.org/competitions/11161.

2) Kaggle. https://www.kaggle.com/retailrocket/ecommerce-dataset.

3) IJCAI-15 competition. https://tianchi.aliyun.com/dataset/dataDetail?dataId=42.

还额外考虑了用户真实点击物品在推荐列表中的位置排名. 与之前的工作 [31,36,37] 一样, 本文主要考虑 HR@20 和 MRR@20 这两个指标, 用于评估会话推荐模型的 Top-20 推荐性能.

4.3 实验设置

在实验中, 本文使用了均值为 0、标准差为 0.1 的高斯 (Gauss) 分布初始化所提出的 CEMA 模型的所有参数. 此外, 本文使用 Adam 优化器优化模型参数, 其中初始学习率为 10^{-3}, 3 个批次后衰减 0.1. 对于 3 个数据集, 向量的嵌入维数统一设置为 100, dropout 率设置为 0.5. 每批样本大小设置为 100, L2 惩罚项设置为 10^{-5}, 所有实验都是使用 PyTorch 实现的, 并在两块 24 GB 显存的英伟达 3090 GPU 上进行训练.

4.4 对比方法

本文将所提出的方法 CEMA 与一系列最先进的基线方法进行了比较, 这些方法如下.

• Item-KNN [15] 推荐与用户先前点击过的物品相似的其他物品, 它使用了会话向量之间的余弦相似度.

• S-POP [13] 是基于流行度的方法的改进版本, 它根据物品的流行度为当前会话中的用户推荐最受欢迎的物品.

• FPMC [14] 是一种基于一阶马尔可夫链和正态矩阵分解的混合方法.

• SKNN [16] 是一种基于启发式的最近邻会话方法, 它采用了高效的内存数据结构和邻域采样, 可以有效保证模型的可扩展性.

• STAN [17] 是 SKNN 方法的扩展, 考虑了会话的序列和时间信息, 并从相似的邻居会话中捕获协作信息.

• GRU4Rec [18] 是一个基于 RNN 的会话推荐模型, 它使用 GRU 对用户点击序列进行建模, 以捕获用户偏好.

• NARM [19] 旨在对用户的序列行为进行建模, 并通过注意力机制和 GRU 捕获用户在当前会话中的主要意图.

• CSRM [20] 将用户在当前会话中的兴趣信息和来自邻居会话的协作信息结合起来, 以在 RNN 的帮助下更好地预测用户的意图.

• STAMP [23] 是一种短期注意力/记忆优先级模型, 它可以从会话上下文中捕获用户的一般偏好, 并从最后一次点击中探索用户的当前兴趣.

• BERT4Rec [25] 采用深度双向自注意力对用户的行为序列进行建模.

• CoSAN [26] 是一种基于自注意力网络的模型, 它从邻居会话探索协作信息, 以获取会话表示并预测当前会话的用户偏好.

• DSAN [28] 是一种双稀疏注意力网络模型, 它使用可学习的目标物品嵌入以对用户当前的兴趣进行建模, 并采用自适应稀疏变换函数消除不相关物品的影响.

• Atten-Mixer [30] 是一种多层次的注意力混合网络, 它利用多层次的用户意图增强模型的推理能力.

• SR-GNN [31] 采用门控图神经网络捕获物品之间的复杂转换关系, 并通过使用软注意力网络聚合全局兴趣和当前兴趣计算会话表示.

• GC-SAN [32] 利用图神经网络和自注意力机制分别捕获会话的局部依赖关系和全局依赖关系, 以获取当前会话表示.

• GCE-GNN [33] 是一种基于图注意力网络的模型, 它旨在将会话序列转换为具有本地表示和全局表示的会话图, 以增强会话推荐的性能.

• COTREC [35] 是一种基于图卷积网络的模型, 它将自监督学习与联合训练相结合以进行数据增强, 从而提升会话推荐的效果.

• S^2-DHCN [34] 是一种双通道超图卷积网络模型, 它通过超图建模物品之间复杂的高阶信息, 并利用自监督学习增强模型性能.

• GRRNW [36] 提出了图邻域路由算法和图随机游走算法, 根据物品在所有会话中的共现情况学习每个物品的局部嵌入和全局嵌入.

• GNN-GNF [38] 是一种基于图神经网络和注意力网络的模型, 它通过过滤会话的噪声数据构建全局图, 并借助 GNN 在全局图中探索物品的转换模式.

• AutoGSR [39] 是一种基于图的自动搜索神经网络框架, 该框架提供了一个实用且通用的解决方案自动找到基于 GNN 的最优会话推荐模型.

• TE-GNN [40] 利用时间图卷积网络来学习基于时间增强会话图的物品嵌入, 以此捕获会话内复杂的用户兴趣转换模式.

4.5 对比实验

为了验证本文所提出的 CEMA 模型的优越性, 将 CEMA 与现有的方法在 Diginetica, Retailrocket 和 Tmall 3 个数据集上进行了对比实验, 实验结果如表 3 所示. 最好的结果以粗体显示, 次优的结果以下划线显示. 最后一行为 CEMA 方法相比现有的最佳方法所提升的性能分数点. 根据表 3 的实验结果, 经过观察后得到以下结论.

(1) 在所有的传统方法中, FPMC 的效果最差, 因为它只利用会话中的最后一个物品来预测用户的下一次点击, 没有考虑整体会话的全局偏好. Item-KNN 是基于 K 个最相似物品进行推荐的方法, STAN 和 SKNN 则是基于 K 个最相似会话预测用户兴趣的方法, 它们都取得了较为不错的结果. 特别地, STAN 和 SKNN 利用了多个不同会话上的用户行为数据, 其性能甚至优于一些基于深度学习的方法, 比如 GRU4Rec, NARM, STAMP 和 CoSAN. 这证明了从不同的邻居会话中提取上下文依赖信息以帮助预测用户偏好的重要性.

(2) 在基于 RNN 的方法中, GRU4Rec 表现最差, 因为 RNN 无法有效地建模会话序列中的长距离依赖关系. 此外, CSRM 同时考虑当前会话中的用户偏好信息和邻居会话中的协作信息, 其性能明显优于 GRU4Rec 和 NARM. 这验证了只关注当前会话中的序列信息会限制模型的表达能力, 还需进一步从邻居会话中捕获有价值的协作信息, 即上下文信息.

(3) 各种基于 AM 的方法, 例如 DSAN 和 Atten-Mixer, 都取得了非常有竞争力的性能表现, 这证明了注意力机制能够有效地区分会话中不同物品的重要程度, 从而更好地提取会话中的全局序列依赖关系. 然而, 由于缺少物品之间复杂的转换关系, 基于 AM 的方法无法像基于 GNN 的方法那样捕获到用户的兴趣迁移模式. 因此, 基于 AM 的方法整体上要弱于基于 GNN 的方法.

(4) 绝大多数基于 GNN 的方法的性能表现都好于上述方法, 这验证了利用物品之间的转换关系来学习会话的局部依赖关系的有效性. 更多地, GC-SAN 和 TE-GNN 都综合考虑物品之间的转换关系和会话中的序列依赖关系, 而 GCE-GNN 和 GRRNW 则额外利用了邻居会话中的协作信息, 这些方法都取得了更好的性能. 这进一步证明了结合多种会话模式特征预测用户偏好的重要性.

(5) 总体而言, 本文所提出的 CEMA 方法在所有数据集的所有评估指标上都明显优于所有的对比方法, 以 Diginetica 和 Retailrocket 数据集为例, CEMA 的 HR@20 和 MRR@20 分别为 63.37%, 26.00%,

表 3 所提出方法 CEMA 与对比方法在 3 个数据集上的实验结果

Table 3 Results of the proposed CEMA and the compared methods on three datasets

		Diginetica		Retailrocket		Tmall	
Method	Year	HR@20	MRR@20	HR@20	MRR@20	HR@20	MRR@20
Item-KNN	2001	35.75	11.57	–	–	9.15	3.31
S-POP	2005	24.09	13.93	38.03	24.81	–	–
FPMC	2010	25.71	7.65	32.37	13.82	16.06	7.32
SKNN	2017	48.35	16.49	54.28	26.46	–	–
STAN	2019	49.93	17.59	53.48	26.81	–	–
GRU4Rec	2016	39.27	10.59	38.55	23.64	10.93	5.89
NARM	2017	49.70	16.17	50.22	24.88	23.30	10.70
CSRM	2019	52.56	17.16	55.04	–	29.46	13.96
STAMP	2018	46.47	14.89	50.96	25.17	26.47	13.36
BERT4Rec	2019	48.78	14.25	54.19	26.42	–	–
CoSAN	2020	48.34	15.22	48.34	15.22	32.68	14.09
DSAN	2021	53.76	18.99	56.54	<u>30.74</u>	–	–
Atten-Mixer	2023	<u>55.66</u>	18.96	–	–	–	–
SR-GNN	2019	50.50	17.63	50.32	26.57	27.57	13.72
GC-SAN	2019	50.84	17.79	51.18	27.40	–	–
GCE-GNN	2020	53.11	18.81	54.72	28.54	33.42	15.91
S^2-DHCN	2021	53.18	18.44	53.66	27.30	31.42	15.05
COTREC	2021	54.18	19.07	56.17	29.97	36.35	18.04
GRRNW	2021	55.64	19.30	<u>57.73</u>	30.21	–	–
GNN-GNF	2022	51.61	17.77	54.64	27.52	–	–
AutoGSR	2022	54.56	19.20	–	–	33.71	15.87
TE-GNN	2022	54.78	<u>19.35</u>	–	–	<u>39.01</u>	<u>18.47</u>
CEMA	–	**63.37**	**26.00**	**61.07**	**34.78**	**40.22**	**20.92**
Improv.	–	7.71 ↑	6.65 ↑	3.34 ↑	4.04 ↑	1.21 ↑	2.45 ↑

61.07% 和 34.78%, 与次优的方法相比, 分别提升了 7.71%, 6.65%, 3.34% 和 4.04%, 性能提升显著. 同时, CEMA 也明显优于其他基于 GNN 的方法, 如 CRRNW, AutoGSR 和 TE-GNN. 上述实验结果充分证明了将每个会话内部的转换模式和序列模式, 以及来自邻居会话的上下文模式相结合以增强模型性能的有效性和优越性.

4.6 消融实验

为了验证 GAU 和 CA-SA 这两个组件对 CEMA 模型性能的影响, 本文在 3 个数据集上进行了消融实验. 具体来说, 本文设计了两种 CEMA 的变体方法: "CEMA w/o GAU" 和 "CEMA w/o CA-SA". "CEMA w/o GAU" 表示从 ILA 模块中删除 GAU, "CEMA w/o CA-SA" 表示从 SLA 模块中删除 CA-SA. 消融实验结果如图 4 所示, 经过分析后有以下结论.

(1) 当从 ILA 模块中删除 GAU 时, 即 "CEMA w/o GAU", 会导致其在所有数据集上出现明显的性能下降, 这证明了序列模式对于会话推荐任务来说是至关重要的, 也验证了 GAU 在捕获会话序列

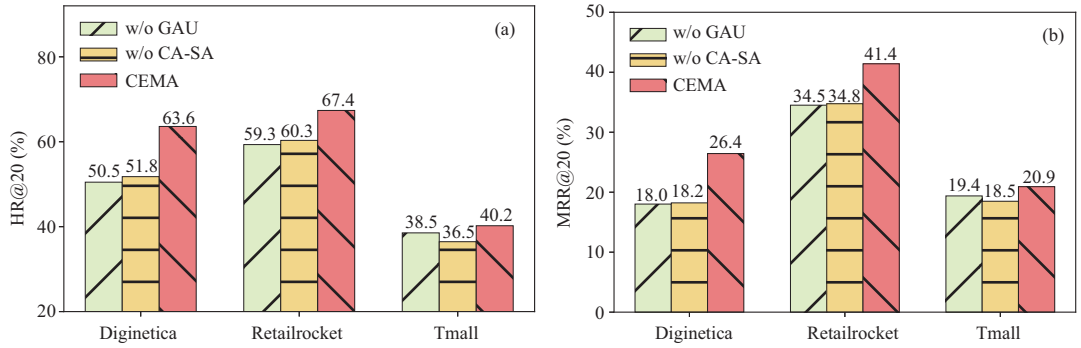

图 4 (网络版彩图) 在 3 个基准数据集上的消融实验结果

Figure 4 (Color online) Ablation experiment results of the proposed method on three benchmark datasets. (a) HR@20; (b) MRR@20

依赖信息方面的有效性.

(2) 当从 SLA 模块中删除 CA-SA 时, 即 "CEMA w/o CA-SA", 它在所有数据集上的表现都比 CEMA 要差得多, 这进一步证明了来自邻居会话的上下文信息对于预测用户下一次点击的重要性. 同时, 也证明了 CA-SA 在探索上下文信息方面的有效性.

(3) 不同于 Diginetica 和 Retialrocket 这两个数据集上的结果, "CEMA w/o CA-SA" 在 Tmall 数据集上的性能要比 "CEMA w/o GAU" 更差, 即缺少上下文信息时, 模型在 Tmall 数据集上的性能要下降的更多. 这可能是因为 Tmall 数据集上拥有更多相似用户行为和用户兴趣的邻居会话. 在这种情况下, 从邻居会话中提取的上下文信息更有助于增强会话推荐的性能. 相比之下, 在 Diginetica 和 Retailrocket 这两个数据集上, "CEMA w/o CA-SA" 的表现比 "CEMA w/o GAU" 更差. 这说明在缺少序列依赖信息的情况下, 模型在 Diginetica 和 Retailrocket 这两个数据集上的性能更差. 这可能是 Diginetica 和 Retailrocket 这两个数据集上的用户偏好, 受会话中的序列依赖关系的影响更多.

4.7 GraphSAGE 层数的影响

为了进一步探索 ILE 模块中 GraphSAGE 层数对 CEMA 模型性能的影响, 本文进行了 GraphSAGE 层数的超参数实验. 在实验中, 使用了不同的 GraphSAGE 层数, 即 $L \in \{1,2,3,4,5,6\}$, 以评估 GraphSAGE 从会话图中学习物品嵌入表示的效果. GraphSAGE 层数对 CEMA 模型性能影响的实验结果如图 5 所示. 经分析后发现, 当 GraphSAGE 层数增加到 2 层时, CEMA 在 Retailrocket 和 Tmall 这两个数据集上取得了最佳的性能. 而对于 Diginetica 数据集来说, GraphSAGE 层数增加到 3 层时, CEMA 实现了最佳的性能. 继续增大 GraphSAGE 层数, 则 CEMA 在 3 个数据集上都会出现性能下降. 这说明适当的 GraphSAGE 层数设置, 可以有效地建模物品之间的一阶转换关系和高阶转换关系, 从而生成更准确的物品嵌入表示. 同时也证明了物品之间复杂的转换关系对于会话推荐任务的重要性. 而过多的 GraphSAGE 层数 (层数大于 3) 则可能会使得消息传播过程中的误差逐渐积累扩大, 使得模型更容易过拟合训练数据, 导致模型性能下降.

值得注意的是, 在 3 个数据集上, 当 GraphSAGE 增加到 5 层时, CEMA 模型会有一个较小幅度的性能提升. 这可能是由于 5 阶物品转换关系给模型带来的性能增益大于 5 层 GraphSAGE 带来的消息传播误差. 具体来说, 增加网络层数有助于模型学习更复杂、更抽象的特征表示. 而每一层的 GraphSAGE 都可以看作是对输入特征的一次非线性变换, 通过多层次的变换, 模型更有可能学到更高阶次的特征表示, 可以更好地捕捉数据中的复杂关系和交互模式. 在本文的超参数实验中, 当

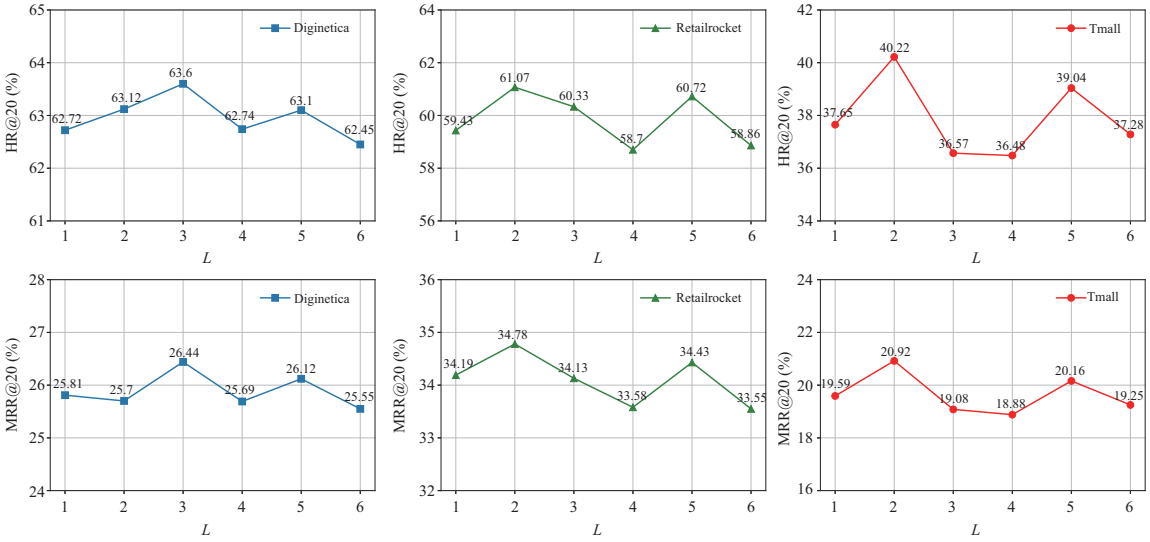

图 5 (网络版彩图) 不同 GraphSAGE 层数的实验结果

Figure 5 (Color online) Results of the proposed method with different depth of GraphSAGE layers

GraphSAGE 在增加到 5 层的时候, 模型在学习物品特征表示时, 考虑了相对较长的物品转换路径, 捕获到了物品之间更复杂的转换关系, 即 5 阶物品转换关系. 这可能包括了更丰富的用户交互信息, 也反映了更深层次的用户兴趣特征, 从而帮助模型更好地预测用户的真实意图和兴趣偏好. 因此, 当 GraphSAGE 层数增加 5 层的时候, 模型性能在 3 个数据集上会出现一个显著的性能上升.

4.8 GAU 层数的影响

为了进一步探索 ILA 模块中 GAU 层数对 CEMA 性能的影响, 本文进行了 GAU 层数的超参数实验. 具体来说, 设置了不同的 GAU 层数, 即 $T \in \{1,2,3,4,5,6\}$, 以评估 GAU 组件在学习会话的序列依赖信息方面的敏感性. GAU 层数的超参数实验结果如图 6 所示. 根据观察, 当 GAU 层数为 2 时, CEMA 在 Diginetica 和 Retailrocket 这两个数据集上取得了最好的性能表现. 而对于 Tmall 数据集来说, 当 GAU 层数增加到 3 时, CEMA 取得了最好的效果. 对于 3 个数据集来说, 当层数大于 3 时, 则会出现较为明显的性能下降. 这些结果表明适当的 GAU 层数可以有效提取会话中的序列依赖信息, 从而帮助模型更好的预测用户的下一次点击, 以增强模型的个性化推荐性能. 但过多 GAU 层数也会带来额外的误差信息, 使得模型发生过拟合, 导致性能下降.

4.9 序列建模方式的影响

为了进一步验证 ILA 模块中的 GAU 在序列建模方面的优越性, 本文特别设计了另一种 CEMA 的变体: CEMA-SAN. CEMA-SAN 表示用自注意力网络 (SAN) 替换 ILA 模块中的 GAU, 即用 SAN 建模会话的序列信息. 在 3 个数据集上进行了 CEMA 和 CEMA-SAN 的对比实验, 实验结果如图 7 所示. 从实验结果来看, CEMA-SAN 在 3 个数据集上的性能表现都明显差于 CEMA. 这是因为 SAN 过于依赖于多头注意力 (MHSA) 和多层感知机 (MLP), 参数量较大, 计算开销较高, 容易导致模型过拟合, 从而导致了模型性能下降. 而 GAU 结合了 GLU 和单头自注意力来区分会话中不同物品的重要性, 聚焦于用户真正感兴趣的物品, 以更好地建模会话的全局序列依赖关系, 进而生成更准确的用户全局偏好表示. 这就证明了 GAU 在序列建模方面的有效性和优越性.

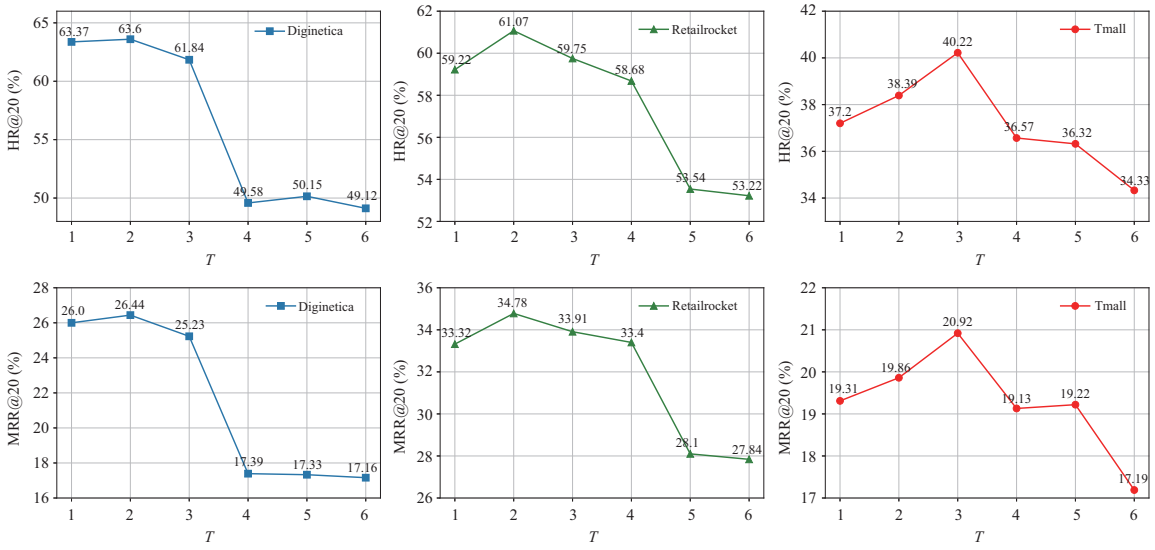

图 6　(网络版彩图) 不同 GAU 层数的实验结果

Figure 6　(Color online) Results of the proposed method with different depth of GAU layers

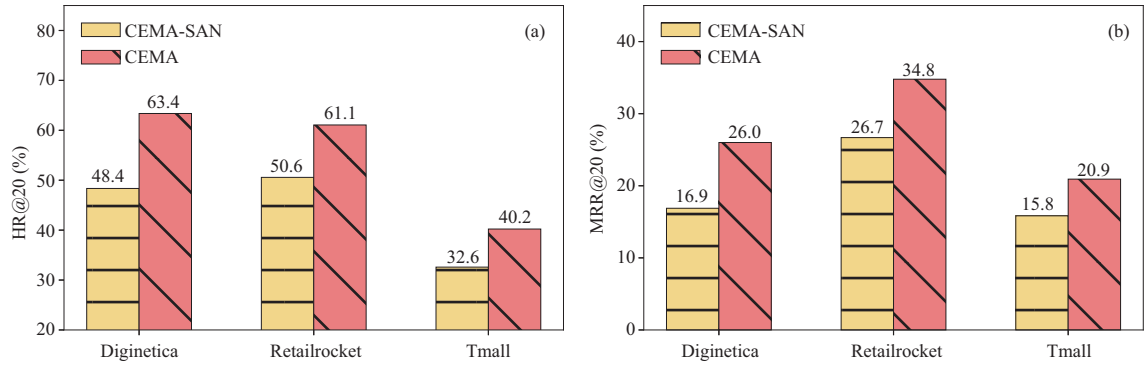

图 7　(网络版彩图) 不同序列建模方式的实验结果

Figure 7　(Color online) Results of the proposed method with different sequence modeling methods. (a) HR@20; (b) MRR@20

5　结论与未来工作

本文提出了基于上下文增强的多级注意力会话推荐模型 (CEMA), 很好地解决了现有方法存在的问题, 进一步提升了会话推荐的性能. 具体来说, CEMA 包括 3 个模块: 物品级嵌入模块 (ILE)、物品级注意力模块 (ILA) 和会话级注意力模块 (SLA). 首先, ILE 模块利用多层 GraphSAGE 捕获物品之间复杂的转换关系, 以此学习用户的动态兴趣迁移模式, 即会话的转换模式. 其次, ILA 模块通过 GAU 计算会话中不同物品的重要性, 特别关注那些用户真正感兴趣的物品, 以更好地提取会话中的序列模式. 此外, 在 SLA 模块中设计了一种上下文感知的软注意力机制 (CA-SA), 该机制可以快速计算不同会话之间的相似性, 从而给那些与当前会话有着相似用户行为和意图的邻居会话分配更高的注意力权重, 并通过加权求和的方式得到上下文模式. 最后, 结合多种模式特征一起预测用户的下一次点击, 增强了模型的个性化推荐能力. 在 3 个公开的基准数据集上的实验结果表明所提出方法比现有的最佳

方法取得了更好的性能, 验证了 CEMA 模型的有效性和优越性.

在未来的研究工作, 我们将在物品 ID 的基础上, 额外引入物品类型和物品价格等特征信息, 以此构造异构图 (物品 ID 节点、物品类型节点和物品价格节点). 然后, 我们将通过异构图神经网络 (heterogeneous graph network, HGN) 捕获异构图中丰富多样的连接关系, 以更准确地建模用户偏好, 从而进一步提升会话推荐方法的性能.

参考文献

1 Liang F, Yang E Y, Pan W K, et al. Survey of recommender systems based on federated learning. Sci Sin Inform, 2022, 52: 713–741 [梁锋, 羊恩跃, 潘微科, 等. 基于联邦学习的推荐系统综述. 中国科学: 信息科学, 2022, 52: 713–741]

2 Liu H, Jing L, Yu J, et al. Social recommendation with learning personal and social latent factors. IEEE Trans Knowl Data Eng, 2021, 33: 2956–2970

3 Bao T, Xu L, Zhu L H, et al. Optimized setting of privacy budget in a recommendation system with local differential privacy. Sci Sin Inform, 2022, 52: 1481–1499 [暴婷, 徐蕾, 祝烈煌, 等. 满足本地化差分隐私的推荐系统中隐私预算的优化设置. 中国科学: 信息科学, 2022, 52: 1481–1499]

4 Hariri N, Mobasher B, Burke R D. Context-aware music recommendation based on latenttopic sequential patterns. In: Proceedings of the 6th ACM Conference on Recommender Systems, 2012. 131–138

5 Cui Q, Wu S, Liu Q, et al. MV-RNN: a multi-view recurrent neural network for sequential recommendation. IEEE Trans Knowl Data Eng, 2020, 32: 317–331

6 Zhao P, Luo A, Liu Y, et al. Where to go next: a spatio-temporal gated network for next POI recommendation. IEEE Trans Knowl Data Eng, 2022, 34: 2512–2524

7 Wu L, Chen L, Hong R, et al. A hierarchical attention model for social contextual image recommendation. IEEE Trans Knowl Data Eng, 2020, 32: 1854–1867

8 Zhang Q, Wu B, Sun Z C, et al. Fine-grained modeling of user interests for sequential recommendation. Sci Sin Inform, 2022, 52: 1775–1791 [张麒, 吴宾, 孙中川, 等. 细粒度建模用户兴趣的序列化推荐方法. 中国科学: 信息科学, 2022, 52: 1775–1791]

9 Pan Z, Cai F, Chen W, et al. Star graph neural networks for session-based recommendation. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 2020. 1195–1204

10 Hamilton W L, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems, 2021. 1024–1034

11 Wang W, Zhang W, Liu S, et al. Incorporating link prediction into multi-relational item graph modeling for session-based recommendation. IEEE Trans Knowl Data Eng, 2021, 35: 2683–2696

12 Yin Z, Han K, Wang P, et al. Multi global information assisted streaming session-based recommendation system. IEEE Trans Knowl Data Eng, 2022, 35: 8245–8256

13 Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng, 2005, 17: 734–749

14 Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation. In: Proceedings of the 19th International Conference on World Wide Web, 2010. 811–820

15 Sarwar B M, Karypis G, Konstan J A, et al. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 19th International Conference on World Wide Web, 2001. 285–295

16 Jannach D, Ludewig M. When recurrent neural networks meet the neighborhood for session-based recommendation. In: Proceedings of the 11th ACM Conference on Recommender Systems, 2017. 306–310

17 Garg D, Gupta P, Malhotra P, et al. Sequence and time aware neighborhood for session-based recommendations: STAN. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019. 1069–1072

18 Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks. In: Proceedings of the 4th International Conference on Learning Representations, 2016

19 Li J, Ren P, Chen Z, et al. Neural attentive session-based recommendation. In: Proceedings of the ACM on

Conference on Information and Knowledge Management, 2017. 1419–1428
20 Wang M, Ren P, Mei L, et al. A collaborative session-based recommendation approach with parallel memory modules. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019. 345–354
21 Wang H W, Guo M Y. Recurrent memory networks: modeling long short-term user preferences for session-based recommendation. Sci Sin Inform, 2020, 50: 1867–1881 [王鸿伟, 过敏意. 刻画长短期用户兴趣的基于会话的推荐系统. 中国科学: 信息科学, 2020, 50: 1867–1881]
22 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems, 2017. 5998–6008
23 Liu Q, Zeng Y, Mokhosi R, et al. STAMP: short-term attention/memory priority model for session-based recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. 1831–1839
24 Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186
25 Sun F, Liu J, Wu J, et al. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019. 1441–1450
26 Luo A, Zhao P, Liu Y, et al. Collaborative self-attention network for session-based recommendation. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020. 2591–2597
27 Zhou K, Wang H, Zhao W X, et al. S3-Rec: self-supervised learning for sequential recommendation with mutual information maximization. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 2020. 1893–1902
28 Yuan J, Song Z, Sun M, et al. Dual sparse attention network for session-based recommendation. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021. 4635–4643
29 Zhao W X, Mu S, Hou Y, et al. RecBole: towards a unified, comprehensive and efficient framework for recommendation algorithms. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 2021. 4653–4664
30 Zhang P, Guo J, Li C, et al. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In: Proceedings of the 16th ACM International Conference on Web Search and Data Mining, 2023. 168–176
31 Wu S, Tang Y, Zhu Y, et al. Session-based recommendation with graph neural networks. In: Proceedings of the 33th AAAI Conference on Artificial Intelligence, 2019. 346–353
32 Xu C, Zhao P, Liu Y, et al. Graph contextualized self-attention network for session-based recommendation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019. 3940–3946
33 Wang Z, Wei W, Cong G, et al. Global context enhanced graph neural networks for session-based recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020. 169–178
34 Xia X, Yin H, Yu J, et al. Self-supervised hypergraph convolutional networks for session-based recommendation. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021. 4503–4511
35 Xia X, Yin H, Yu J, et al. Self-supervised graph co-training for session-based recommendation. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 2021. 2180–2190
36 Zhang Z, Wang B. Graph neighborhood routing and random walk for session-based recommendation. In: Proceedings of the IEEE International Conference on Data Mining, 2021. 1517–1522
37 Yan S, Xiang X G, Li Z C. Item correlation modeling in interaction sequence for graph convolutional session recommendation. Sci Sin Inform, 2022, 52: 1069–1082 [闫昭, 项欣光, 李泽超. 基于交互序列商品相关性建模的图卷积会话推荐. 中国科学: 信息科学, 2022, 52: 1069–1082]
38 Feng L, Cai Y, Wei E, et al. Graph neural networks with global noise filtering for session-based recommendation. Neurocomputing, 2022, 472: 113–123
39 Chen J, Zhu G, Hou H, et al. AutoGSR: neural architecture search for graph-based session recommendation. In:

Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022. 1694–1704
40 Tang G, Zhu X, Guo J, et al. Time enhanced graph neural networks for session-based recommendation. Knowl-Based Syst, 2022, 251: 109204
41 Yang Y, Huang C, Xia L, et al. Debiased contrastive learning for sequential recommendation. In: Proceedings of the ACM International World Wide Web Conference, 2023. 1063–1073
42 Hua W, Dai Z, Liu H, et al. Transformer quality in linear time. In: Proceedings of International Conference on Machine Learning, 2022. 9099–9117
43 Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 933–941
44 Su J, Ahmed M, Lu Y, et al. Roformer: enhanced transformer with rotary position embedding. Neurocomputing, 2024, 568: 127063
45 So D R, Manke W, Liu H, et al. Searching for efficient transformers for language modeling. In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems, 2021. 6010–6022
46 Ma X, Zhou C, Kong X, et al. MEGA: moving average equipped gated attention. In: Proceedings of the 11th International Conference on Learning Representations, 2023

Context enhanced multi-level attention model for session-based recommendation

Biqing ZENG*, Junlong CHI, Jiatao CHEN & Liangqi XIE

School of Software, South China Normal University, Foshan 528225*, China*
* Corresponding author. E-mail: zengbiqing@scnu.edu.cn

Abstract Session-based recommendation aims to predict users' next click based on their interactions in anonymous sessions. While graph neural network (GNN)-based methods have shown promising results, they still have limitations. GNN-based methods overlook the session's sequential patterns and only consider transition patterns between items. Furthermore, most methods focus solely on the internal information of the current session and ignore external collaborative information from neighboring sessions, i.e., contextual patterns. To address these issues, we propose a context enhanced multi-level attention model (CEMA), which uses multi-level attention mechanisms to learn item features and model user preferences at both item and session levels. CEMA applies a multi-layer GraphSAGE to learn complex transition patterns between items to capture local user preferences. The item-level attention mechanism in CEMA employs a gate attention unit to calculate item importance, identify user interests, and avoid noise interference. This helps capture the sequential patterns in the session to model users' global preferences. Moreover, a session-level attention mechanism is designed to efficiently calculate the similarity between sessions and extract contextual patterns to predict users' next click. Experiments conducted on three public benchmark datasets demonstrate CEMA's superior performance compared to existing methods.

Keywords session-based recommendation, multi-level attention mechanism, graph neural network, sequential patterns, contextual patterns