Computer Science and Application 计算机科学与应用, 2026, 16(3), 593-606
Published Online March 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.163087

Hans汉斯

基于时序信息增强的图注意力网络会话推荐

孙　伟[1*]，陈平华[2]，梁秋铭[1]

[1]广州理工学院计算机科学与工程学院，广东　广州
[2]广东工业大学计算机学院，广东　广州

收稿日期：2026年1月27日；录用日期：2026年2月26日；发布日期：2026年3月5日

摘　要

针对现有基于图神经网络的会话推荐模型过度侧重图结构空间关联、忽视会话时序特性，且会话编码器依赖单一项目嵌入、难以捕捉用户长短期真实兴趣及多阶段兴趣转移的问题，本文提出基于时序信息增强的图注意力网络会话推荐模型，首先通过时序信息增强模块，融合初始项目顺序嵌入与位置嵌入，借助多头自注意力强化时序全局依赖，并引入Dropout正则化与残差连接优化特征聚合；其次构建含自连接的有向会话图，利用多层图注意力网络自适应学习邻居节点权重，实现精准多跳信息传播与噪声抑制；最后设计多级意图会话编码器，按时间顺序从近到远划分多阶段兴趣，通过多头注意力学习各阶段贡献权重并聚合，避免单一编码的片面性。实验结果显示，该模型在Tmall与Diginetica两大电商数据集的P@10、P@20、MRR@10、MRR@20四项核心指标上，均显著优于包括Transformer基方法、最新GNN变体在内的主流基线模型。消融实验进一步验证时序信息增强模块、多层图注意力机制及多级意图编码器的关键有效性，并直接证明了多级意图编码器相较于标准的“序列最后隐藏层 + 全局注意力”机制的显著优势，证明模型能深度融合时序动态与图结构关联信息，显著提升会话推荐的精准度与合理性。

关键词

会话推荐，图注意力网络，时序信息增强，多级意图编码，用户兴趣建模

Session-Based Recommendation Using Graph Attention Networks Enhanced with Temporal Information

Wei Sun[1*], Pinghua Chen[2], Qiuming Liang[1]

[1]School of Computer Science and Engineering, Guangzhou Institute of Science and Technology, Guangzhou Guangdong
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

[*]通讯作者。

文章引用: 孙伟, 陈平华, 梁秋铭. 基于时序信息增强的图注意力网络会话推荐[J]. 计算机科学与应用, 2026, 16(3): 593-606. DOI: 10.12677/csa.2026.163087

Received: January 27, 2026; accepted: February 26, 2026; published: March 5, 2026

Abstract

Existing graph neural network-based session-based recommendation models often overemphasize graph structural spatial relationships while neglecting sequential characteristics of sessions. Furthermore, their session encoders rely on single item embeddings, making it difficult to capture users' long-term and short-term interests and multi-stage interest shifts. To address these issues, this paper proposes a graph attention network-based session recommendation model enhanced with temporal information. First, a temporal information enhancement module fuses initial item sequential embeddings and positional embeddings, using multi-head self-attention to strengthen global temporal dependencies, and incorporating Dropout regularization and residual connections to optimize feature aggregation. Second, a directed session graph with self-connections is constructed, and a multi-layer graph attention network is used to adaptively learn neighbor node weights, enabling accurate multi-hop information propagation and noise suppression. Finally, a multi-level intent session encoder is designed to divide interests into multiple stages chronologically from near to far, learning the contribution weights of each stage through multi-head attention and aggregating them, avoiding the limitations of single encoding. Experimental results show that the proposed model significantly outperforms mainstream baseline models including Transformer-based methods and the latest GNN variants on four key metrics (P@10, P@20, MRR@10, and MRR@20) on two major e-commerce datasets, Tmall and Diginetica. Ablation experiments further validate the critical effectiveness of the temporal information enhancement module, multi-layer graph attention mechanism, and multi-level intent encoder, and directly prove the significant advantage of the multi-level intent encoder over the standard "last hidden layer of sequence + global attention" mechanism, demonstrating that the model can deeply integrate temporal dynamics and graph structural relational information, significantly improving the accuracy and rationality of session-based recommendations.

Keywords

Session-Based Recommendation, Graph Attention Networks, Temporal Information Enhancement, Multi-Level Intent Encoding, User Interest Modeling

1. 引言

随着电子商务与社交媒体平台的快速发展，会话推荐作为个性化推荐的核心任务之一，旨在基于用户当前会话中的交互行为序列，预测其下一个可能感兴趣的项目[1]。相较于传统推荐任务，会话推荐无需依赖用户长期历史行为，仅通过短期交互序列捕捉动态兴趣，更能适配匿名用户或兴趣快速变化的场景[2]。近年来，图神经网络(Graph Neural Networks, GNN)凭借其强大的复杂关系建模能力，在会话推荐领域得到广泛应用。现有基于 GNN 的会话推荐模型通过构建会话图捕捉项目间的多跳依赖关系，显著提升了推荐性能[3]-[5]。然而，此类模型仍存在两个关键问题：第一，过度关注图结构中的空间关联信息，忽视了会话序列固有的时序特性(如项目交互顺序、位置信息等)，导致项目嵌入无法反映用户行为的动态演变规律；第二，会话编码器设计片面，多数模型直接将最后一个项目的嵌入作为用户短期兴趣表示，

未能考虑误点击、好奇心驱动等非真实兴趣行为，且难以捕捉长会话中的多阶段兴趣转移。

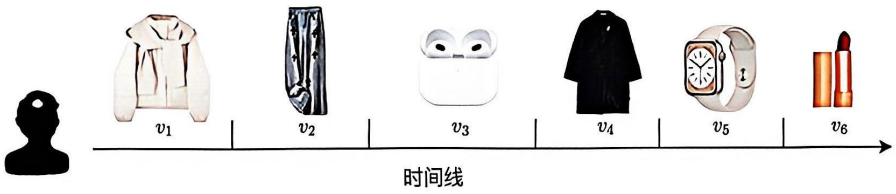


Figure 1. An example for session
图 1. 匿名用户的会话序列图

图 1 展示了一个典型的用户会话序列示例。用户在会话初始阶段浏览服装类商品，中间穿插电子类产品交互，最终点击口红类美妆产品，体现了会话中用户兴趣的多元性与跨类别转移特性。若在项目嵌入更新过程中，对与中心节点(如长裤)关联度不同的邻居节点(如棉衣与耳机)赋予同等权重，必然引入无关噪声；同时，若仅通过“序列最后隐藏层 + 全局注意力”机制编码兴趣，或单纯以最后一个项目表征短期兴趣，可能因误点击或兴趣偏移导致推荐偏差。因此，如何在 GNN 框架中有效融合时序信息，以及设计合理的会话编码器突破“序列最后隐藏层 + 全局注意力”机制的固有局限、捕捉真实长短期兴趣，成为提升会话推荐性能的关键。为解决上述问题，本文提出(Graph attention network session-based recommendation enhanced by temporal information, GAT-SRET)模型，其核心创新点如下：

1) 构建时序信息增强模块，将项目的顺序关系与位置特征融入初始嵌入，使项目表征同时包含空间结构与时序动态信息；

2) 设计含自连接的有向会话图，并通过多层图注意力网络自适应学习邻居节点权重，实现精准的信息聚合与噪声抑制；

3) 提出多级意图会话编码器，突破传统“序列最后隐藏层 + 全局注意力”的编码局限，通过按时间顺序从近到远硬性划分多阶段兴趣的方式，替代单一的长短期兴趣划分逻辑，通过多头注意力聚合获得全面的会话兴趣表示，更贴合用户兴趣的时序转移规律，证明硬性阶段划分相较于全局注意力无差别加权的显著优势。

2. 相关工作

2.1. 传统序列推荐

传统序列推荐模型以统计规律和简单序列依赖为核心建模思路，主要分为基于协同过滤和马尔科夫链两类[6]。Item-KNN 作为协同过滤的代表性方法，通过计算项目静态相似度进行推荐，虽实现简单但完全忽视序列关联性与跨类别兴趣转移[7]；基于用户协同过滤的方法则受限于稀疏数据下相似用户挖掘的可靠性。马尔科夫链类方法以 FPMC 为典型，融合一阶马尔科夫链与矩阵分解，首次捕捉相邻项目的直接转移关系[8]，但仅关注一阶依赖，无法处理多跳间接关联，且未结合时序动态特征，难以适配用户兴趣的快速变化。整体而言，传统模型结构简单、计算高效，但未能充分挖掘序列时序信息与复杂依赖关系，难以满足精准推荐需求。

2.2. 基于循环神经网络的会话推荐

基于循环神经网络(Recurrent Neural Networks, RNN)的模型核心在于捕捉长序列时序依赖并动态更新兴趣表示[9]。GRU4Rec 作为首个将 RNN 引入会话推荐的模型，通过门控循环单元逐次编码会话序列，

以最后一个隐藏状态作为兴趣表示，性能远超传统方法[10]。然而，该模型存在梯度易消失、难以捕捉远距离依赖及未区分长短期兴趣的局限。STAMP 强化短期兴趣优先原则，简化长短期兴趣融合方式，在短会话场景中表现优异[11]，但过度依赖最后一个项目表征短期兴趣，忽视误点击行为与全局时序结构。NARM 首次提出“序列最后隐藏层 + 全局注意力”的编码机制，通过全局注意力对所有项目嵌入加权聚合表征长期兴趣，结合最后一个隐藏层的短期兴趣实现融合，一定程度提升了编码的全面性，但该结构未对兴趣进行时间维度的阶段划分，全局注意力对不同时间节点的项目一视同仁，无法精准捕捉用户由近及远的多阶段兴趣转移，易受非真实兴趣行为的干扰，也无法体现用户兴趣的时序衰减特性[9]。此类模型虽提升了时序依赖捕捉能力，但未摆脱 RNN 的固有缺陷，且兴趣编码方式仍存在片面性。

2.3. 基于图神经网络的会话推荐

基于 GNN 的模型通过构建图结构捕捉项目间的复杂多跳依赖与非序列关联，为会话推荐提供了新的解决方案[12]。SR-GNN 作为该领域的开创性工作，将会话转化为带权有向图，借助门控 GNN 聚合邻居特征，证明了图结构在会话推荐中的有效性[13]，但完全忽视时序信息，且依赖最后一个项目表征短期兴趣。FGNN 引入加权图注意力层，自适应分配邻居权重以减少噪声[14]，却未区分边类型差异对信息传播的影响。S^2-DHCN 通过超图与线图捕捉高阶关联，结合自监督学习增强嵌入表达[15]，但模型复杂度较高且未优化时序与兴趣编码。TA-GNN 创新引入目标感知注意力，融合候选项目信息提升推荐针对性[16]，但仍未有效利用时序信息。近年来，研究人员提出一系列新型 GNN 变体模型用于会话推荐，此类模型进一步优化了图结构的建模效果，提升了推荐性能，但仍未突破传统的“最后隐藏层 + 全局注意力”兴趣编码的局限，或未将时序信息与图结构进行深度融合，制约性能的进一步提升。

2.4. 时序信息在会话推荐中的应用

时序信息是刻画用户动态兴趣的关键[17]，但现有多数模型未能将其与图结构建模深度融合。部分研究尝试引入时序特征，但设计较为简单：例如，部分 RNN 模型在项目嵌入中添加位置编码，但仅作为辅助特征，未参与核心的兴趣聚合过程[18]；部分 GNN 模型通过边权重编码时间间隔，但未考虑项目在序列中的绝对位置和全局顺序依赖[19]。最新 GNN 变体对时序的融合设计仍较为浅层，如何将时序信息与 GNN 的空间结构建模深度融合，同时突破传统兴趣编码机制的局限，仍是当前会话推荐领域亟待解决的问题。

2.5. 基于 Transformer 的会话推荐

Transformer 凭借多头自注意力机制强大的全局时序依赖捕捉能力，成为序列推荐领域的研究热点，也为会话推荐提供了新的技术思路，是当前会话推荐领域的重要基线方法。SASRec 首次将 Transformer 引入序列推荐，通过单向自注意力机制捕捉项目间的长距离时序依赖，摆脱 RNN 的梯度消失问题[20]；BERT4Rec 则借鉴 BERT 的预训练思想，采用双向自注意力机制建模序列的上下文依赖，在会话推荐中展现出优异的性能[21]。此类模型的核心优势在于能精准捕捉序列的全局时序关联，对长距离依赖的建模能力显著优于 RNN，但存在缺乏对项目间多跳结构依赖的建模能力，无法有效刻画会话中项目的复杂关联关系，且在处理短会话时易出现过拟合问题，难以兼顾时序动态与结构关联的双重需求。

3. 模型整体设计

GAT-SRET 模型的核心目标在于给定一个会话 S，提取出能表示会话兴趣的嵌入向量 M^*，进而生成 M^* 与所有候选项的匹配分数，通常按分数排序生成 Top-K 推荐列表，取最高分数的项目作为下一项推荐给用户。具体地，模型由时序信息增强模块、图注意力网络以及提取用户兴趣的会话编码器三部分

组成。首先，需要通过时序增强模块对会话图中的每个项目附加时序信息；其次，根据各个会话中用户与项目的交互顺序构建会话图，并通过图注意力网络生成第 $l+1$ 层的项目嵌入 $h_{v_i}^{(l+1)}\left(1\le i\le |s|\right)$；最后，经过多层网络的学习，生成最终的项目嵌入后，由多级意图会话编码器提取会话兴趣，与候选项目集合中的各项进行匹配以预测用户可能点击的下一个项目，模型整体架构如图 2 所示。

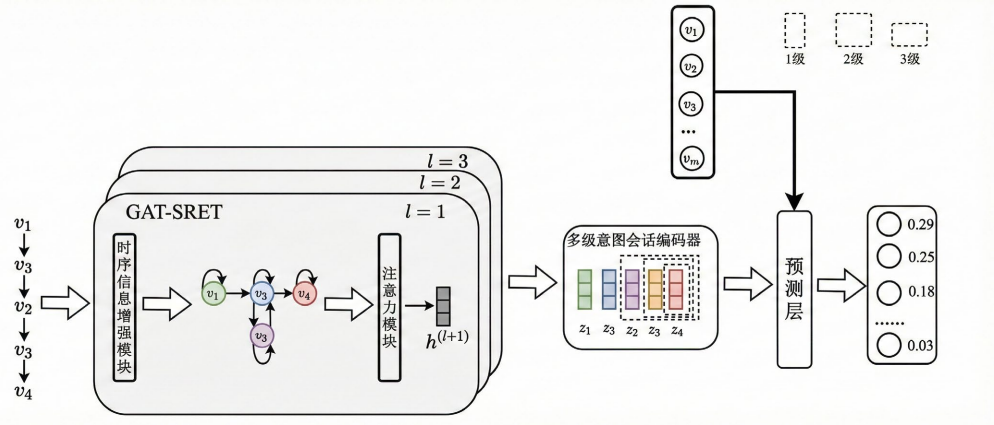


Figure 2. The framework of the GAT-SRET model
图 2. GAT-SRET 模型的整体框架

3.1. 时序增强模块

会话序列的时序特性(顺序关系与位置信息)直接反映用户兴趣的动态演变，而现有 GNN 模型的初始嵌入多为随机生成或静态向量，缺乏时序动态特征。为此，本文设计时序信息增强模块，将项目的顺序嵌入与位置嵌入融合，为后续图注意力网络提供富含时序信息的初始输入。具体实现步骤如下：

1) 嵌入初始化：给定会话序列 $S=\{v_1,v_2,\cdots,v_n\}$，其中 v_i 表示第 i 个交互项目；根据会话中的用户/项目交互顺序，从初始项目嵌入矩阵 $E\in R^{M\times d}$ 中查询得到顺序嵌入 $e_i^s=E[v_i]$，其中 M 为项目总数，d 为嵌入维度；同时，从位置嵌入矩阵 $P\in R^{L\times d}$ (L 为最大会话长度)中查询得到位置嵌入 $e_i^p=P[i]$，用于表征项目在会话中的绝对位置。

2) 嵌入融合：将顺序嵌入与位置嵌入相加得到初始时序嵌入 $e_i=e_i^s+e_i^p$，形成时序嵌入序列 $E=[e_1,e_2,\cdots,e_n]$。

3) 多头自注意力增强：为捕捉时序信息的全局依赖，采用多头自注意力机制对初始时序嵌入进行增强。将 E 输入 h 个并行的自注意力头，每个注意力头通过线性变换将嵌入映射到不同子空间，计算过程如公式(1)所示：

$$h_i^k=\text{Attention}\left(W_q^k e_i,W_k^k E,W_v^k E\right) \tag{1}$$

其中 $k=1,2,\cdots,h$，$W_q^k,W_k^k,W_v^k\in R^{d\times d/h}$ 为第 k 个注意力头的可训练参数。

4) 特征聚合与正则化：将 h 个注意力头的输出级联，通过前馈神经网络聚合特征，同时引入 Dropout 正则化缓解过拟合，最终得到时序增强后的项目嵌入 $\tilde{E}=[\tilde{e}_1,\tilde{e}_2,\cdots,\tilde{e}_n]$，计算过程如公式(2)所示：

$$\tilde{e}_i=\text{FFN}\left(\text{Concat}\left(h_i^1,h_i^2,\cdots,h_i^h\right)\right)+e_i \tag{2}$$

其中残差连接 e_i 用于缓解深度网络的梯度消失问题。

时序信息增强模块的核心优势在于：使项目嵌入同时包含空间结构信息与时序动态信息，为后续图

注意力网络的精准信息聚合奠定基础。

3.2. 多层图注意力网络

为捕捉项目间的复杂多跳依赖关系，并自适应区分邻居节点的重要性，本文构建含自连接的有向会话图，并设计多层图注意力网络实现信息的精准传播与聚合。

3.2.1. 会话图构建

针对会话序列 $S=[v_1,v_2,\cdots,v_n]$，构建有向图 $G=(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 为节点集合(对应会话中的项目)，E 为有向边集合。根据用户交互顺序，若用户在点击 v_i 后点击 v_j $(i<j)$，则添加有向边 $e_{i,j}\in E$。同时，为保留项目自身特征并支持重复交互建模，为每个节点添加自连接边 $e_{ii}\in E$。最终，会话图中包含四种边类型：入边(e_{ji})、出边(e_{ij})、双向边(e_{ij} 与 e_{ji} 同时存在)和自循环边(e_{ii})，分别对应节点间的不同关联关系，如图 3 所示。

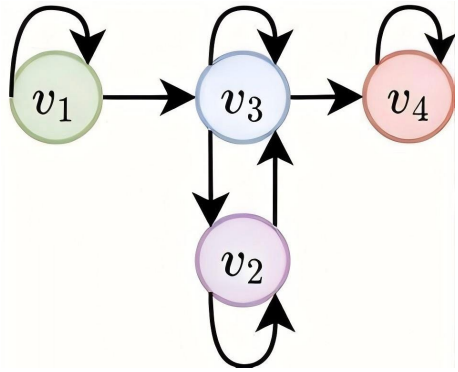


Figure 3. Construction of session graph
图 3. 会话图构建

3.2.2. 图注意力信息聚合

图注意力网络的核心是自适应学习邻居节点的注意力权重，实现精准的特征聚合。对于第 l 层的节点嵌入 h_i^l (初始层为时序增强后的嵌入 $\tilde{e}_i$)，其第 $l+1$ 层嵌入的计算过程如下：

1) 注意力系数计算：对于中心节点 v_i 的邻居节点 v_j (含自身)，首先通过线性变换将嵌入映射到同一特征空间，再通过逐元素乘积与非线性变换计算原始注意力系数，如公式(3)所示：

$$\alpha_{ij}^l=\mathrm{LeakReLU}\left(a_r^{\mathrm{T}}\left(W^l h_i^l\odot W^l h_j^l\right)\right) \tag{3}$$

其中 $W^l\in R^{d\times d}$ 为第 l 层的共享线性变换矩阵，$a_r\in R^d$ 为对应边类型 r 的注意力参数向量，$\odot$表示逐元素乘积，LeakyReLU 为非线性激活函数。

2) 注意力系数归一化：为使同一节点的不同邻居注意力系数具有可比性，通过 softmax 函数对原始注意力系数进行归一化，如公式(4)所示：

$$\hat{\alpha}_{ij}^l=\frac{\exp\left(\alpha_{ij}^l\right)}{\sum_{v_k\in N(v_i)}\exp\left(\alpha_{ik}^l\right)} \tag{4}$$

其中 $N(v_i)$ 表示节点 v_i 的一阶邻居集合(含自身)。

3) 加权特征聚合：根据归一化后的注意力系数，对邻居节点的嵌入进行加权线性组合，得到中心节点的下一层嵌入，如公式(5)所示：

$$h_i^{l+1} = \sigma\left(\sum_{v_j \in N(v_i)} \hat{\alpha}_{ij}^l W^l h_j^l\right) \tag{5}$$

其中 σ 为 sigmoid 激活函数。

通过多层图注意力网络的迭代传播，节点嵌入能够逐步融合多跳邻居的特征信息，同时通过自适应注意力权重抑制无关噪声，提升项目表征的准确性。图 4 展示了两层图注意力网络的嵌入信息传播过程。

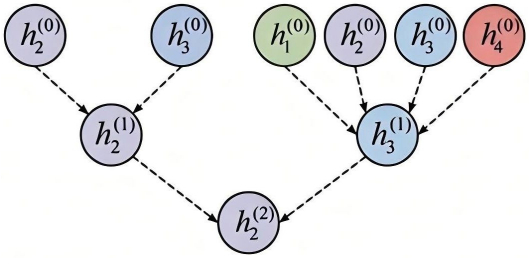


Figure 4. Embedding information propagation process
图 4. 嵌入信息传播过程

3.3. 多级意图会话编码器

现有会话编码器中，以 NARM 为代表的主流模型采用“序列最后隐藏层 + 全局注意力”的经典结构进行兴趣编码，这也是当前会话推荐中应用最广泛的编码方式，该结构通过最后隐藏层表征短期兴趣、全局注意力加权聚合所有项目表征长期兴趣，虽实现了长短期兴趣的融合，但存在明显缺陷：一是全局注意力对会话中不同时间节点的项目一视同仁，未按时间远近进行阶段划分，无法精准捕捉用户由近及远的多阶段兴趣转移特性，也无法体现兴趣的时序衰减规律；二是该结构易受误点击、好奇心驱动等非真实兴趣行为的干扰，全局注意力可能为无关项目分配较高权重，导致兴趣表示偏离用户真实需求；三是仅通过单一维度的长短期划分，难以适配长会话中复杂的兴趣演变规律。这也是本文设计多级意图编码器的核心动机，通过硬性的时间阶段划分替代全局注意力的无差别加权，解决上述固有问题。为此，本文设计多级意图会话编码器，突破传统编码结构的局限，通过按时间顺序硬性划分多阶段兴趣并自适应聚合，捕捉用户的全面长短期真实兴趣，如图 5 所示，具体实现步骤如下：

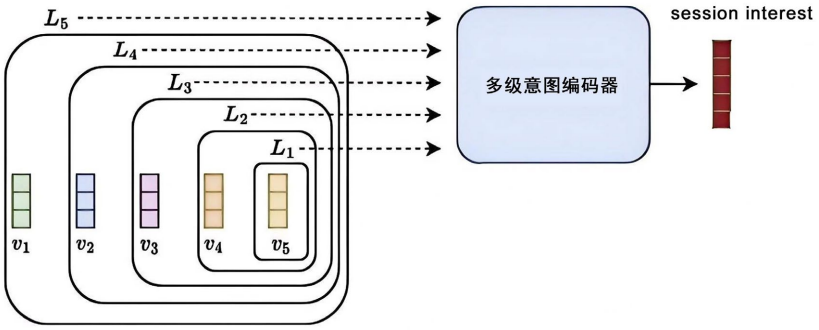


Figure 5. Session encoder
图 5. 会话编码器

1) 多阶段兴趣划分：设经过 L 层图注意力网络后，项目嵌入序列为 $H = \left[h_1^L, h_2^L, \cdots, h_n^L\right]$。按时间顺序(从近到远)划分 K 个阶段兴趣，第 k 阶段兴趣包含最近的 k 个项目嵌入，如公式(6)所示：

$$I_k = \left[h^L_{n-k+1}, h^2_{n-k+2}, \cdots, h^L_n \right] \ (k = 1,2,\cdots,K) \tag{6}$$

其中 K 为阶段数，阶段数越高，对应的兴趣感受野越广，能够捕捉更长期的兴趣偏好。该硬性按时间划分的方式，区别于 NARM 中全局注意力的无差别加权逻辑，能够精准贴合用户兴趣的时序演变与衰减规律，使近阶段兴趣更贴合用户当前的真实需求，远阶段兴趣保留用户的历史偏好，有效区分不同时间维度的兴趣贡献，这也是该模块相较于“序列最后隐藏层 + 全局注意力”机制的核心优势。

2) 多头注意力聚合：将各阶段兴趣作为查询向量，通过多头注意力学习不同阶段兴趣对整体会话兴趣的贡献权重。首先，将阶段兴趣矩阵 $I = [I_1, I_2, \cdots, I_K]$ 输入 h 个并行的注意力头，计算每个注意力头的权重系数，如公式(7)所示：

$$w^m_k = \mathrm{Softmax}\left(\frac{\left(Q^m I_k\right)\left(K^m H\right)^{\mathrm{T}}}{\sqrt{d/h}} \right) \quad (m = 1,2,\cdots,h) \tag{7}$$

其中 $Q^m, K^m \in R^{(d/h)\times d}$ 为第 m 个注意力头的查询与键矩阵。

3) 权重融合与兴趣生成：对每个注意力头的权重系数进行平均池化，如公式(8)所示：

$$\bar{w}_k = \frac{1}{h} \sum\nolimits_{m=1}^{h} w^m_k \tag{8}$$

最后，通过加权聚合各阶段兴趣与项目嵌入序列，生成最终的会话兴趣表示 s，如公式(9)所示：

$$s = \mathrm{LayerNorm}\left(W_s \sum\nolimits_{k=1}^{K} \bar{w}_k \left(I_k V^m \right) + b_s \right) \tag{9}$$

其中 $W_s \in R^{d\times d}$ 、$V^m \in R^{d\times(d/h)}$ 为训练参数，$b_s \in R^d$ 为偏置向量，LayerNorm 用于归一化加速模型收敛。

3.4. 预测层

为预测用户下一个可能交互的项目，将会话兴趣表示 s 与候选项目嵌入矩阵 E 进行匹配计算。对于候选项目 v，其推荐概率通过点积相似度与 softmax 函数计算，如公式(10)所示：

$$p\left(v \middle| S\right) = \frac{\exp\left(s^{\mathrm{T}} E[v]\right)}{\sum_{v' \in V} \exp\left(s^{\mathrm{T}} E[v']\right)} \tag{10}$$

模型训练采用交叉熵损失函数，最小化预测概率与真实标签的差距，损失函数如公式(11)所示：

$$L = -\frac{1}{N} \sum\nolimits_{S \in D} \log p\left(v^* \middle| S\right) \tag{11}$$

其中 D 为训练集，为会话 S 中真实的下一个交互项目，N 为训练集会话总数。

4. 实验结果与分析

4.1. 数据集

为全面验证模型的有效性与泛化能力，选取电商领域两大经典数据集 Tmall 与 Diginetica 进行实验，覆盖不同用户行为特征与数据分布场景，确保实验结果的可靠性与模型的泛化性能。Tmall：记录匿名用户在双十一购物狂欢节前 6 个月内的交互行为，数据包含丰富的跨类别购物场景，用户兴趣转移特征明显，平均会话长度较长。Diginetica：涵盖电商平台用户的商品浏览与购买交互记录，数据量更大，项目种类更丰富，用户会话长度相对更短，对短期兴趣捕捉精度要求更高。为保证实验公平性与数据质量，两个数据集采用统一且规范的预处理规则，具体流程如下：1) 丢弃出现次数少于 5 的项目，以减少噪声干扰；2) 保留长度大于 1 的会话，确保存在下一个交互项目的预测目标；3) 将最后一周的会话作为测试

集，其余作为训练集；4) 通过序列截断扩充训练数据：对于长度为 n 的会话 $[v_1,v_2,\cdots,v_n]$，生成 $n-1$ 个子序列 $[v_1,\cdots,v_i](i=2,\cdots,n)$，其中 v_{i+1} 为预测目标，预处理后的数据集统计信息如表 1 所示。

Table 1. Dataset details introduction
表 1. 数据集详情介绍

数据集	点击数目	训练数据数	测试数据数	项目个数	平均会话长度
Tmall	818,479	351,268	25,898	40,728	6.69
Diginetica	982,961	719,470	60,858	43,097	5.12

4.2. 基线方法

为了客观地评价本模型的推荐性能，选取 13 种主流会话推荐模型作为基线，涵盖传统序列推荐、基于 RNN 的推荐、基于 GNN 的推荐和基于 Transformer 的推荐四类，具体如下：

1) 传统序列推荐：Item-KNN (基于项目协同过滤)、FPMC (融合一阶马尔科夫链与矩阵分解)；

2) 基于 RNN 的推荐：GRU4Rec (基于 GRU 的序列编码)、NARM (“序列最后隐藏层 + 全局注意力”经典模型)、STAMP (短期兴趣优先的注意力模型)；

3) 基于 GNN 的推荐：经典模型(SR-GNN、FGNN、S^2-DHCN、TA-GNN)、最新变体(GCE-GNN、COTREC)；

4) 基于 Transformer 的推荐：SASRec (单向自注意力序列推荐模型)、BERT4Rec (双向自注意力预训练序列推荐模型)。

4.3. 评价指标

实验采用推荐领域常用的排名类指标 P@K 和 MRR@K 评估模型性能，其中 $K=10$、20：

1) P@K：衡量推荐列表前 K 个项目中包含真实目标项目的比例，反映推荐精准度，如公式(12)所示：

$$\mathrm{P@K}=\frac{1}{N}\sum_{S\in D_{test}}\coprod\left(v^{*}\in \text{Top-K}(S)\right) \tag{12}$$

其中 D_{test} 为测试集，$\coprod(\cdot)$ 为指示函数，$\text{Top-K}(S)$ 为模型为会话 S 生成的前 K 个推荐项目集合。

2) MRR@K：衡量真实目标项目在推荐列表中的排名位置，反映推荐排序合理性，如公式(13)所示：

$$\mathrm{MRR@K}=\frac{1}{N}\sum_{S\in D_{test}}\frac{1}{\text{rank}\left(v^{*}\middle|S\right)} \tag{13}$$

其中 $\text{rank}\left(v^{*}\middle|S\right)$ 为真实目标项目 v^{*} 在推荐列表中的排名(若不在前 K 则计为 0)。

4.4. 实验设置

本实验超参数设置如下：项目嵌入维度 $d=100$，批量大小 $=100$，训练轮次 $=64$，采用早停策略(连续 10 轮性能无提升则停止训练)；多头注意力头数 $=6$，初始学习率 $=0.001$，学习率衰减率 $=0.1$；时序信息增强模块中 Dropout 丢弃率 $=0.6$；初始项目嵌入与位置嵌入由均值为 0、标准差为 0.1 的高斯分布随机生成；采用 Adam 优化器更新模型参数。

4.5. 实验结果

4.5.1. 推荐性能对比

表 2 的实验结果显示，在 Tmall 和 Diginetica 两个电商数据集的四项核心指标(P@10、MRR@10、

P@20、MRR@20)上，本文提出的 GAT-SRET 模型均显著优于传统序列推荐、基于 RNN、基于经典 GNN、基于最新 GNN 变体、基于 Transformer 的 13 类基线模型，展现出稳定且卓越的推荐性能。

传统序列推荐模型中，Item-KNN 和 FPMC 表现最差，如 Item-KNN 在 Tmall 的 P@10 仅 6.65、Diginetica 的 MRR@20 仅 2.95，印证了其忽视序列关联性与复杂依赖的局限性；基于 RNN 的模型(GRU4Rec、NARM、STAMP)性能优于传统模型，其中 NARM 作为“序列最后隐藏层 + 全局注意力”的经典代表，在 Tmall 的 P@10 达 20.89，虽优于 GRU4Rec 但仍受限于 RNN 梯度易消失及编码结构的固有缺陷，整体表现不及基于 GNN 和 Transformer 的模型，也直接证明了该编码机制的局限性；基于 Transformer 的模型(SASRec、BERT4Rec)凭借多头自注意力的全局时序建模能力，性能优于多数基于 RNN 的模型，BERT4Rec 在 Tmall 的 P@10 达 25.73，但其缺乏图结构建模能力，无法捕捉项目间的多跳依赖，在 Diginetica 的短会话场景中表现受限；基于 GNN 的模型凭借图结构对多跳依赖的建模优势表现突出，其中最新 GNN 变体(GCE-GNN、COTREC)性能优于经典 GNN 模型，COTREC 作为 GNN 基线最优模型，在 Tmall 的 P@10 和 MRR@20 分别达 29.87 和 16.53，在 Diginetica 的 P@20 达 31.85，该类模型虽优化 GNN 的建模能力，但未将时序信息与图结构进行深度融合，且仍采用传统的兴趣编码方式。

GAT-SRET 模型通过融合时序信息、多层图注意力机制与多级意图编码，同时兼顾时序动态建模与图结构多跳依赖捕捉，在 Tmall 的 P@10、MRR@20 分别达 31.18、17.21 (较最新 GNN 变体 COTREC 分别提升 4.39%、4.17%，较 BERT4Rec 提升 21.18%、9.53%，较 NARM 提升 49.26%、48.87%)，在 Diginetica 的 P@10、MRR@10 分别达 29.45、15.79，各项指标均居首位，充分证明其在捕捉时序动态、抑制噪声及聚合多阶段兴趣方面的有效性，同时验证了多级意图编码器相较于传统“序列最后隐藏层 + 全局注意力”结构的显著优势，以及融合时序与图结构相较于纯 Transformer、纯 GNN 模型的技术优势。

Table 2. Comparison of experimental results
表 2. 对比实验结果

模型	Tmall				Diginetica			
	P@10	MRR@10	P@20	MRR@20	P@10	MRR@10	P@20	MRR@20
Item-KNN	6.65	3.11	9.15	3.31	5.82	2.76	8.03	2.95
FPMC	13.10	7.12	16.06	7.32	11.85	6.43	14.52	6.61
GRU4Rec	9.47	5.78	10.93	5.89	8.63	5.12	9.87	5.23
NARM	20.89	11.25	24.98	11.56	18.97	10.12	22.65	10.38
STAMP	22.63	13.12	26.47	13.36	20.15	11.68	23.82	11.91
SR-GNN	23.41	13.45	27.57	13.72	21.03	11.97	24.76	12.19
FGNN	20.67	10.07	25.24	10.39	18.92	9.05	22.53	9.28
S^2-DHCN	26.22	14.60	31.42	15.05	24.07	13.21	28.54	13.56
TAGNN	28.54	15.28	32.75	15.87	26.31	14.39	30.18	14.48
GCE-GNN	28.15	14.96	31.98	16.02	25.87	14.05	30.96	14.12
COTREC	29.87	15.93	34.26	16.53	27.96	15.02	31.85	15.16
SASRec	23.58	12.89	27.86	13.15	22.16	11.89	25.97	12.15
BERT4Rec	25.73	14.02	30.15	15.73	24.89	13.56	29.03	13.89
GAT-SRET	**31.18**	**16.80**	**35.82**	**17.21**	**29.45**	**15.79**	**33.96**	**15.83**

4.5.2. 消融实验

为了进一步评估 GAT-SRET 的时序信息增强模块、图注意力机制对模型推荐性能的贡献，以及验证多级意图会话编码器的有效性，设计三组消融变体模型进行对比实验：

1) GAT-SRET-I：去除时序信息增强模块，将初始嵌入直接输入图注意力网络，且后续多层网络均不经过时序信息增强模块。

2) GAT-SRET-II：删除图注意力机制，采用等权重平均聚合邻居节点特征。

3) GAT-SRET-II：将多级意图会话编码器替换为 NARM 中的“序列最后隐藏层 + 全局注意力”编码器，即通过最后一个项目嵌入表征短期兴趣、全局注意力加权聚合所有项目嵌入表征长期兴趣，融合后生成最终兴趣表示。这三个 GAT-SRET 的变体版本的性能结果如表 3 所示。

Table 3. Performance evaluation of different GAT-SRET variant models
表 3. 不同 GAT-SRET 变体模型的性能评估

模型	Tmall		Diginetica	
	P@20	MRR@20	P@20	MRR@20
GAT-SRET-I	29.51	14,38	26.89	12.76
GAT-SRET-II	34.58	16.97	32.47	15.51
GAT-SRET-III	32.70	16.84	30.65	15.03
GAT-SRET	**35.82**	**17.21**	**33.96**	**15.83**

从表 3 可观察到各核心模块对模型性能的影响，具体分析如下，GAT-SRET-I 在两个数据集上的性能降幅最为显著，Tmall 的 P@20 从原模型的 35.82 降至 29.51 (下降 17.62%)，MRR@20 从 17.21 降至 14.38；Diginetica 的 P@20 从 33.96 降至 26.89(下降 20.82%)，这表明时序特征与空间结构特征的深度融合，是提升项目表征有效性的关键，尤其在 Diginetica 这类短会话数据集上，时序模块对动态兴趣的刻画作用更突出；GAT-SRET-II 的 Tmall P@20 降至 34.58 (较原模型下降约 3.46%)，Diginetica P@20 降至 32.47，说明图注意力机制能够自适应区分邻居节点的重要性，有效过滤无关噪声，提升信息聚合的精准度；GAT-SRET-III 的 Tmall P@20 降至 32.70 (较原模型下降约 8.71%)，Diginetica P@20 降至 30.65，验证了多级意图编码器相较于 NARM 中“序列最后隐藏层 + 全局注意力”编码结构的显著优势，证明通过时间维度的多阶段硬性划分，能够更精准地捕捉用户多阶段兴趣转移特性，有效抑制非真实兴趣行为的干扰，避免单一兴趣编码的片面性。

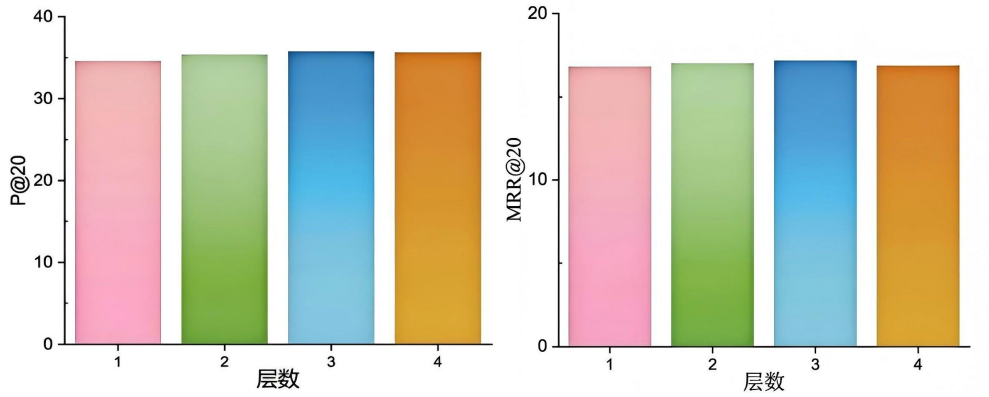
4.5.3. 网络层数的影响分析

图神经网络的层数直接影响节点特征的传播范围与融合效果，为探究层数对 GAT-SRET 性能的影响，实验在 Tmall 数据集上开展，层数设置从 1 到 4，对比不同层数下的 P@20 和 MRR@20 指标，结果如图 6 所示。可以观察到：随着层数从 1 增加到 3，模型性能持续提升，当层数为 3 时达到最优(P@20 = 35.82, MRR@20 = 17.21)，适当增加层数能够让节点融合更多跳邻居的有效信息；当层数增加到 4 时，性能出现轻微下降(P@20 = 35.12, MRR@20 = 16.93)，这是由于过多层数导致的“过平滑”问题，使得节点嵌入趋于同质化，丧失区分度。

4.5.4. 超参数的影响分析

在模型训练过程中，超参数的不同设置会对实验结果产生不同影响。本文在不改变其他超参数的情况下，研究了不同学习率对实验结果的影响，实验在 Diginetica 数据集上开展，学习率分别设置为 0.0001、

0.0005、0.001、0.0015、0.002、0.0025 和 0.003，测试结果如图 7 所示。由图可知，当学习率为 0.001 时，P@K 和 MRR@K 的数值大多高于其他学习率下的对应数值，因此将学习率设置为 0.001 有助于提高模型的准确率。



(a)不同传播层数在 P@20 指标下的对比　(b) 不同传播层数在 MRR@20 指标下的对比

Figure 6. Comparison chart of propagation layers
图 6. 传播层数对比图

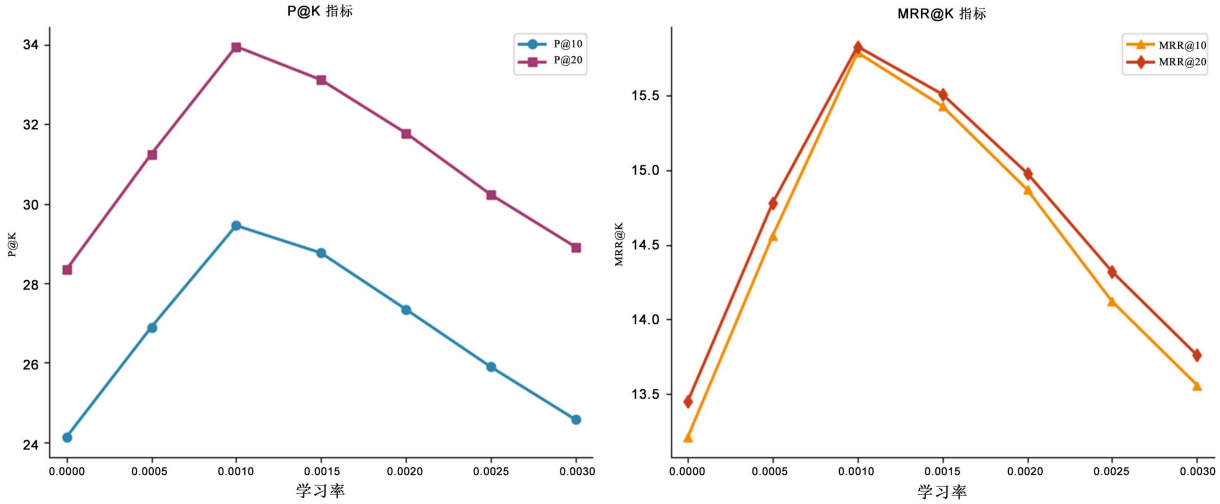


Figure 7. Effect of different learning rates on model results
图 7. 不同学习率模型性能的影响

5. 结论

本文提出 GAT-SRET 模型，通过时序信息增强模块融合项目的顺序与位置特征，使项目嵌入兼具空间结构与时序动态信息；利用多层图注意力网络自适应学习邻居节点权重，实现精准信息聚合与噪声抑制；针对多级意图会话编码器突破传统“序列最后隐藏层 + 全局注意力”编码结构的局限，按时间顺序硬性划分多阶段兴趣，全面捕捉用户长短期真实兴趣，证明硬性阶段划分相较于全局注意力无差别加权的显著优势。在 Tmall 与 Diginetica 数据集上的实验表明，GAT-SRET 在各项核心指标上均显著优于包括 Transformer 基方法、最新 GNN 变体在内的主流基线模型，消融实验不仅验证了时序信息增强模块、多层图注意力机制的有效性，更直接证明了多级意图编码器相较于传统“序列最后隐藏层 + 全局注意力”

结构的显著优势，回答了多级编码设计的核心动机与改进价值。未来工作可从以下方向展开：1) 针对“过平滑”问题，引入残差连接或图归一化技术，增强深层网络的特征区分度；2) 将模型扩展到多模态会话推荐场景，融合文本、图像等多源信息提升推荐性能。

基金项目

广东省重点领域研发计划项目(2023B1111050010)。

参考文献

[1] Kumar, C. and Kumar, M. (2024) Session-Based Recommendations with Sequential Context Using Attention-Driven LSTM. *Computers and Electrical Engineering*, **115**, Article ID: 109138. https://doi.org/10.1016/j.compeleceng.2024.109138

[2] Liu, F., Cheng, Z., Zhu, L., Liu, C. and Nie, L. (2020) A2-GCN: An Attribute Aware Attentive GCN Model for Recommendation. IEEE Transactions on Knowledge and Data Engineering.

[3] Li, Z., Yang, C., Chen, Y., Wang, X., Chen, H., Xu, G., *et al.* (2024) Graph and Sequential Neural Networks in Session-Based Recommendation: A Survey. *ACM Computing Surveys*, **57**, 1-37. https://doi.org/10.1145/3696413

[4] Hasan, E., Rahman, M., Ding, C., Huang, J.X. and Raza, S. (2025) Review-Based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives. *ACM Computing Surveys*, **58**, 1-41. https://doi.org/10.1145/3742421

[5] Chang, J., Gao, C., Zheng, Y., Hui, Y., Niu, Y., Song, Y., *et al.* (2021) Sequential Recommendation with Graph Neural Networks. *Proceedings of the* 44*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11-15 July 2021, 378-387. https://doi.org/10.1145/3404835.3462968

[6] Su, X. and Khoshgoftaar, T.M. (2009) A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, **2009**, Article ID: 421425. https://doi.org/10.1155/2009/421425

[7] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001) Item-Based Collaborative Filtering Recommendation Algorithms. *Proceedings of the* 10*th International Conference on World Wide Web*, Hong Kong, 1-5 May 2001, 285-295. https://doi.org/10.1145/371920.372071

[8] Rendle, S., Freudenthaler, C. and Schmidt-Thieme, L. (2010) Factorizing Personalized Markov Chains for Next-Basket Recommendation. *Proceedings of the* 19*th International Conference on World Wide Web*, Raleigh, 26-30 April 2010, 811-820. https://doi.org/10.1145/1772690.1772773

[9] Graves, A. (2012) Long Short-Term Memory. In: Graves, A., Ed., *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, 37-45. https://doi.org/10.1007/978-3-642-24797-2_4

[10] Fan, Z., Liu, Z., Wang, Y., Wang, A., Nazari, Z., Zheng, L., *et al.* (2022) Sequential Recommendation via Stochastic Self-Attention. *Proceedings of the ACM Web Conference* 2022, 25-29 April 2022, 3036-3047. https://doi.org/10.1145/3485447.3512077

[11] Liu, Q., Zeng, Y., Mokhosi, R. and Zhang, H. (2018) STAMP: Short-Term Attention/Memory Priority Model for Session-Based Recommendation. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1831-1839. https://doi.org/10.1145/3219819.3219950

[12] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S. (2021) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24. https://doi.org/10.1109/tnnls.2020.2978386

[13] Chang, J., Gao, C., He, X., Jin, D. and Li, Y. (2021) Bundle Recommendation and Generation with Graph Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 2326-2340. https://doi.org/10.1109/tkde.2021.3114586

[14] Gao, C., Wang, X., He, X. and Li, Y. (2022) Graph Neural Networks for Recommender System. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 21-25 February 2022, 1623-1625. https://doi.org/10.1145/3488560.3501396

[15] Xia, X., Yin, H., Yu, J., Wang, Q., Cui, L. and Zhang, X. (2021) Self-Supervised Hypergraph Convolutional Networks for Session-Based Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 4503-4511. https://doi.org/10.1609/aaai.v35i5.16578

[16] Yu, F., Zhu, Y., Liu, Q., Wu, S., Wang, L. and Tan, T. (2020) TAGNN: Target Attentive Graph Neural Networks for Session-Based Recommendation. *Proceedings of the* 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 25-30 July 2020, 1921-1924. https://doi.org/10.1145/3397271.3401319

[17] Wei, P., Shu, H., Gan, J., Deng, X., Liu, Y., Sun, W., *et al.* (2025) Sequential Recommendation System Based on Deep

Learning: A Survey. *Electronics*, **14**, Article 2134. https://doi.org/10.3390/electronics14112134

[18] Ma, Y. and Gan, M. (2021) DeepAssociate: A Deep Learning Model Exploring Sequential Influence and History-Candidate Association for Sequence Recommendation. *Expert Systems with Applications*, **185**, Article ID: 115587. https://doi.org/10.1016/j.eswa.2021.115587

[19] Wang, R., Lou, J. and Jiang, Y. (2022) Session-Based Recommendation with Time-Aware Neural Attention Network. *Expert Systems with Applications*, **210**, Article ID: 118395. https://doi.org/10.1016/j.eswa.2022.118395

[20] Kang, W. and McAuley, J. (2018) Self-Attentive Sequential Recommendation. 2018 *IEEE International Conference on Data Mining* (*ICDM*), Singapore, 17-20 November 2018, 197-206. https://doi.org/10.1109/icdm.2018.00035

[21] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., *et al.* (2019) BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, 3-7 November 2019, 1441-1450. https://doi.org/10.1145/3357384.3357895