Computer Science and Application 计算机科学与应用, 2022, 12(4), 1108-1121
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124114

Hans汉斯

融合图神经网络和注意力机制的会话推荐模型

黄浩文，陈平华

广东工业大学，计算机学院，广东 广州

收稿日期：2022年3月17日；录用日期：2022年4月18日；发布日期：2022年4月25日

摘　要

针对电子商务场景的会话推荐难以处理用户行为随机性、商品数据稀疏性和推荐结果滞后性问题，提出融合图神经网络和注意力机制的会话推荐模型。首先引入相对点击时间率改善由用户随机点击造成的推荐性能下降问题，参与生成由会话转换的商品关系依赖图结构；然后由多层门控图神经网络处理图结构，通过聚合更多节点信息输出相对稠密的商品表示；接着使用门控循环单元捕捉会话信息，并借助注意力机制强化会话靠后的项目，综合形成用户表示，最终获得实时推荐。模型在Yoochoose和Diginetica两个公开数据集上进行实验，获得了较好的表现，结果表明所提出的模型可以提高推荐准确性。

关键词

会话推荐，图神经网络，门控循环单元，注意力机制

Session-Based Recommendation Model with Graph Neural Network and Attention Mechanism

Haowen Huang, Pinghua Chen

School of Computer, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 17th, 2022; accepted: Apr. 18th, 2022; published: Apr. 25th, 2022

Abstract

Aiming at the problems that session-based recommendation in e-commerce platforms hardly solve the problems of random-like behavior from users, the sparsity from items and the real-time performance from recommendation results, a session-based recommendation model with graph

文章引用：黄浩文，陈平华. 融合图神经网络和注意力机制的会话推荐模型[J]. 计算机科学与应用，2022，12(4): 1108-1121. DOI: 10.12677/csa.2022.124114

neural network and attention mechanism is proposed. First, the relative click time rate is introduced to tackle the recommendation effect degradation caused by random clicks from users, and participates in the generation of the item relationship dependency graph transformed by sessions. Then, the multi-layer gated graph neural network is used for learning the graph structure, and the relatively dense item's representation is output by aggregating more node information. After that, each session's information is captured by gated recurrent units, and items at the rear of the session are strengthened with the help of the attention mechanism to generate each user's representation, so as to finally obtain real-time recommendation. Experiments on Yoochoose and Diginetica datasets show that the proposed model can achieve good performance and improve the accuracy of recommendation.

Keywords

Session-Based Recommendation, Graph Neural Network, Gated Recurrent Unit, Attention Mechanism

1. 引言

会话推荐是推荐系统的重要组成部分，其对用户点击交互的商品、新闻、音乐等历史序列进行分析，预测下一步点击的内容。会话推荐由于主要关注序列本身，不借助用户或项目的特征即可对该用户下一次点击的商品实现推荐或预测，有利于解决匿名用户推荐等问题，故在近几年备受关注[1]。

会话为用户在一段时间内访问电子商务等平台，依次点击商品的序列，本身不含用户或商品特征上下文，传统的协同过滤[2]、矩阵分解[3]等方法不适用处理会话信息。会话推荐的主要研究内容是如何分析用户的点击会话数据，设计合适的推荐算法，预测用户后续点击的内容。根据会话的时序特性，其建模一般使用马尔可夫链[4]或 RNN (Recurrent Neural Network，循环神经网络) [5]等方法处理，以获得会话建模。然而，单纯的 RNN 等结构难以处理用户随机点击问题；其次，各会话相对孤立，会话之间的信息未被充分挖掘。对于使用 GNN (Graph Neural Network，图神经网络) [6]的会话推荐模型，其将所有的一维会话序列整合成一个二维图结构，以刻画信息较为整体的向量表示。虽然 GNN 从整体的方式克服了会话的孤立性，但未能以个体的角度捕捉每个用户的兴趣变化，未能实现实时推荐，如某学生在考试前关注图书或文具类商品，考试后关注娱乐类商品等，且噪声商品问题仍待解决。

针对会话推荐中的用户行为随机、商品数据稀疏和推荐结果滞后等问题，本文提出一种融合了 GNN 和 AM (Attention Mechanism，注意力机制)的会话推荐模型 TGGA-SR，在充分利用会话数据的基础上实现准确推荐，并解决以下问题：

1) 用户行为随机性。为减少上述噪声商品对推荐结果的影响，模型根据用户点击相邻商品的时间间隔生成相对点击时间率，参与商品关系依赖图的生成，对很短时间连续点击的商品之间赋予较小的权重。

2) 商品数据稀疏性。针对会话推荐中用户或商品特征缺失的特点，模型首先通过生成商品关系依赖图建立商品之间的联系，然后使用多层 GNN 进行向量表示，聚合多阶邻居节点信息。

3) 推荐结果滞后性。用户在不同时期会表现不同的兴趣偏好，点击的商品会发生变化。根据这一特点，用户下一步关注的商品与最后一次点击的商品有很大关联。模型结合 GRU 和 AM，对用户会话时间靠后的商品信息进行强化，形成实时推荐。

2. 研究现状

2.1. 传统会话推荐方法

这类方法采用数据挖掘或机器学习等手段挖掘会话内数据的相关性，进行会话推荐。

Item-KNN [7] [8]是基于物品的 *k*-近邻算法，通过召回和会话余弦相似度最高的物品进行推荐。Mobasher 等人[9]提出了关联规则的方法，以条件概率的方式计算置信度返回预测项目。BPR-MF [10]结合贝叶斯个性化排序与矩阵分解，主要通过随机梯度下降法优化 Pairwise 排序损失函数。FPMC [4]则采用了马尔可夫链模型，为每个用户生成单独的转移概率矩阵，生成转移矩阵立方体，用于预测未知物品引起用户兴趣的可能性，并以此排序得出物品推荐列表。

2.2. 基于深度学习的会话推荐方法

这类方法主要结合 RNN、GNN 等手段，具有表达能力强，挖掘更多数据隐藏模式等特点。

RNN 主要用于建模序列信息，其变体 LSTM 或 GRU 因能有效记忆较长序列而获得较广泛的应用，在机器翻译[11]、语音识别[12]、视频描述[13]等领域都有很好的效果，也适合处理会话。GRU4REC [14]为经典的融合 RNN 会话推荐模型，该模型还根据物品流行程度进行采样，划分正负样本，使用基于排名的损失函数。NARM [15]在基于 RNN 的会话推荐模型基础上加入了 AM [16]，从隐状态捕获用户在当前会话的目的，并结合用户浏览时的序列行为进行推荐。STAMP [17]则结合记忆力与注意力，同时考虑用户通用兴趣和当前兴趣，并通过提高短期兴趣的权重来减轻兴趣漂移对推荐结果的影响。

GNN 为图表示学习(Graph Representation Learning)的一个分支，采用 CNN (Convolutional Neural Networks，卷积神经网络)思想表示图结构，GNN 根据节点聚合与更新方式的不同，主要模型有 GCN [18]、GraphSAGE [19]、GGNN [20]等，在推荐系统[21]、生物医学[22]、知识图谱[23]等领域皆有应用。相较于 RNN 对会话进行建模缺乏对商品的刻画，GNN 可以将所有会话合成一张图，进而学习商品表示。SR-GNN [24]是其中的经典模型，其结合 AM 考虑用户当前兴趣。TAGNN [25]模型也借助 AM 将用户的短期和长期兴趣结合推荐。SEMGNN [26]模型则从商品类别和商品本身两个不同粒度分别形成图结构，得出商品和商品类别的表示并用 AM 进行信息融合，输入到 GRU 学习会话信息，形成推荐。

3. 算法模型

3.1. 问题描述

如图 1，多个用户在一段时间内登录电子商务平台，分别先后点击了若干商品。用户依次点击的商品序列形成会话，而这些会话不包含用户或商品的特征信息。本模型需要通过分析这些会话预测各用户下一次点击的商品，实现推荐。

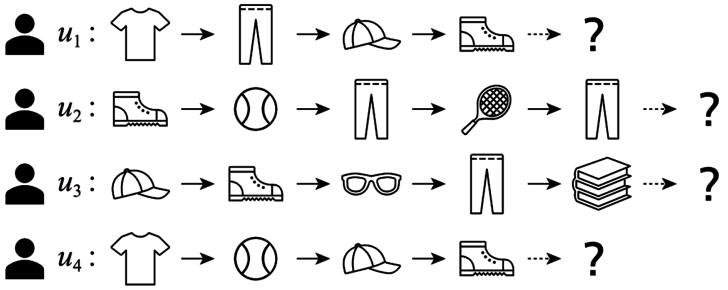

Figure 1. An example of problem description
图 1. 问题描述示例

3.2. 总体结构

本文 TGGA-SR 模型结构如图 2 所示。该图展示模型根据用户点击商品记录实现商品推荐，大量用户浏览商品行为数据经过处理后，形成用户点击商品会话(①)。这些会话将整合到商品关系依赖图的有向图结构中(②)，并由 GNN 学习商品表示(③)。处理用户表示时，将用户点击商品对应的商品表示按会话顺序排列，作为 GRU 的输入，并由结合 AM 形成最终用户表示(④)，最后生成推荐结果(⑤)。

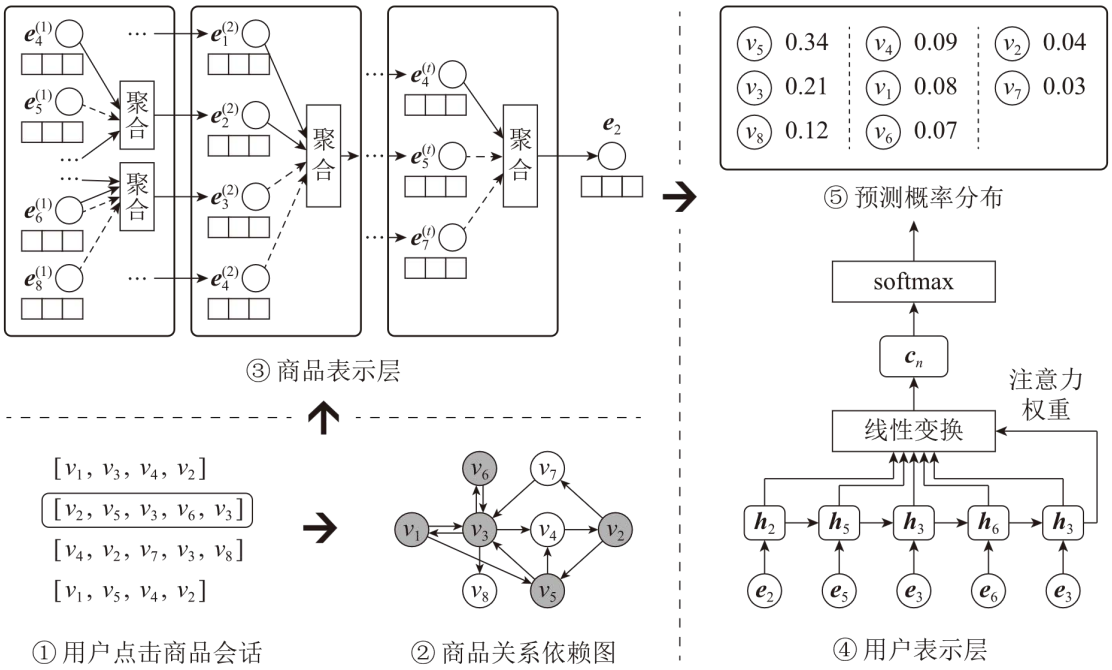

Figure 2. The architecture of TGGA-SR model
图 2. TGGA-SR 模型结构

3.3. 建立商品关系依赖图

TGGA-SR 的输入数据为电子商务平台在一段时间内用户与商品的交互日志，包括用户 ID、商品 ID 和交互时刻(用户点击商品的时刻)等关键信息。模型首先把属于同一个用户 ID 的所有日志按时序整理成会话，即会话包括某用户先后点击的所有商品。一般而言，用户 u 在一段时间依次点击了 n 件商品 $v_1,v_2,\cdots,v_j,v_k,\cdots,v_{n-1},v_n$，则相应形成会话 $s_u=\left[v_1,v_2,\cdots,v_j,v_k,\cdots,v_{n-1},v_n\right]$。会话用于用户向量建模，也参与商品关系依赖图的生成。

商品关系依赖图集合了所有会话的节点和边信息。商品关系依赖图 G 以会话内商品为节点，相邻商品之间的商品节点建立有向边，由先点击商品指向后点击商品。以图 2 的会话 $s_2=\left[v_2,v_5,v_3,v_6,v_3\right]$ 为例，G 的子图 G_2' 则存在有向边 $\langle v_2,v_5\rangle$，$\langle v_5,v_3\rangle$，$\langle v_3,v_6\rangle$，$\langle v_6,v_3\rangle$，形成的子图 G_2' 结构如图 3 所示。所有会话的商品关系依赖图 G_i' 节点与有向边分别取并集，形成最终的商品关系依赖图 G。这种将原本一维的会话整合成二维的图结构方法，可以进一步挖掘商品之间的联系，减少数据稀疏。

根据用户自身偏好，用户在点击浏览不同的商品时会呈现不同的关注程度。由于会话不含用户和商品特征，用户对商品的关注程度仅能依据时间信息间接判断。如果用户对某商品的关注程度较高，则该商品的浏览时间相对较长，且在会话内所有商品浏览总时长占比较高。故 TGGA-SR 引入会话相对点击时间率(Relative Time Ratio) [27]，用来衡量会话内商品的重要程度。

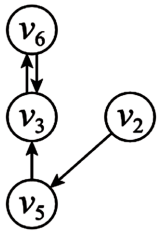

Figure 3. Subgraph of item relationship dependency graph
图 3. 商品关系依赖图的子图

一般地，在会话 $s_u=\left[v_1,v_2,\cdots,v_j,v_k,\cdots,v_{n-1},v_n\right]$ 中，用户 u 先后在 $T_{u,j}$ 和 $T_{u,k}$ 时刻分别点击商品 v_j 和 v_k，则两者点击时间间隔 $T_u\left(v_j,v_k\right)=T_{u,k}-T_{u,j}$ 相当于用户 u 在电子商务平台展示商品 v_j 的时长，以 s 为单位。若 v_j 和 v_k 商品未依次点击，即有向边 $\left\langle v_j,v_k\right\rangle\notin G_u'$，则 $T_u\left(v_j,v_k\right)=0$。时间间隔定义见式(1)。

$$T_u\left(v_j,v_k\right)=\begin{cases}T_{u,k}-T_{u,j}, & \left\langle v_j,v_k\right\rangle\in G_u',\\ 0, & \left\langle v_j,v_k\right\rangle\notin G_u'\end{cases} \tag{1}$$

较短的商品展示时间反映用户对该商品的关注程度较低，或随机性点击商品[27]。而在适当情况下展示时间越长的商品，反映该商品越能引起用户的关注，越符合用户的期望。但展示时间如果过长，并不一定意味着该商品符合期望，因为用户可能处理其它事务或离开屏幕，暂停浏览商品。故 TGGA-SR 模型根据真实数据集的统计规律，设置停留时间上限。

会话相对点击时间率用来衡量会话内的商品展示时长在整个会话时长的比例，反映该商品占整个会话的关注程度。如会话 s_u 内商品 v_j 的会话相对点击时间率如式(2)所示。

$$\mathrm{TR}_u\left(v_j,v_k\right)=\frac{T_u\left(v_j,v_k\right)}{\sum_{i=1}^{n-1}T_u\left(v_i,v_{i+1}\right)} \tag{2}$$

会话相对点击时间率参与商品关系依赖图的有向边权重计算，以设定权重的方式加强或削弱会话中商品之间的联系，如较低的会话相对点击时间率反映两种商品切换时间相对较短，关联度较小，以此减少了因随机性造成的误差。整个商品关系依赖图 G 的出度矩阵与入度矩阵分别由式(3)和(4)定义。

$$a_{(\mathrm{out})}\left(v_j,v_k\right)=\begin{cases}\dfrac{\sum_i \mathrm{TR}_i\left(v_j,v_k\right)}{d_{(\mathrm{out})}\left(v_j\right)}, & d_{(\mathrm{out})}\left(v_j\right)>0,\\ 0, & d_{(\mathrm{out})}\left(v_j\right)=0\end{cases} \tag{3}$$

$$a_{(\mathrm{in})}\left(v_j,v_k\right)=\begin{cases}\dfrac{\sum_i \mathrm{TR}_i\left(v_j,v_k\right)}{d_{(\mathrm{in})}\left(v_k\right)}, & d_{(\mathrm{in})}\left(v_k\right)>0,\\ 0, & d_{(\mathrm{in})}\left(v_k\right)=0\end{cases} \tag{4}$$

其中，i 表示会话序号；$d_{(\mathrm{out})}\left(v_j\right)$ 为商品节点 v_j 的出度，表示整个平台一段时间内先点击商品 v_j、后点击其它商品的行为次数；$d_{(\mathrm{in})}\left(v_k\right)$ 为节点 v_k 的入度，表示先点击其它商品、后点击商品 v_k 的行为次数。建立的商品关系依赖图为后续商品基于图神经网络的向量表示提供依据。

接续会话 $s_2=\left[v_2,v_5,v_3,v_6,v_3\right]$ 的例子，假设 5 件商品点击时间间隔依次为 10 s、50 s、200 s、140 s，则子图 G_2' 出度矩阵 $\boldsymbol{A}_{2(\mathrm{out})}$ 和入度矩阵 $\boldsymbol{A}_{2(\mathrm{in})}$ 分别如图 4 的(a)、(b)所示。

	v_2	v_3	v_5	v_6
v_2	0	0	0.025	0
v_3	0	0	0	0.500
v_5	0	0.125	0	0
v_6	0	0.350	0	0

(a)

	v_2	v_3	v_5	v_6
v_2	0	0	0	0
v_3	0	0	0.063	0.175
v_5	0.025	0	0	0
v_6	0	0.500	0	0

(b)

Figure 4. An example of out-degree matrix and in-degree matrix
图 4. 出度矩阵和入度矩阵示例

3.4. 商品向量表示

TGGA-SR 使用 GNN 学习商品关系依赖图的节点，形成商品的向量表示。由于图为节点和边表示的结构，相邻节点以边联结，GNN 形成节点向量的过程则由其相邻节点向量的信息传播到该节点，并作一定聚合获得。

商品关系依赖图商品节点采用 GGNN (Gated Graph Neural Network，门控图神经网络)学习。GGNN 结合 RNN 思想学习节点向量，节点信息随单位时间传播。图 5 为 G_2' 按时序展开传播信息的示意图，各节点向量信息在当前时刻以同样的规则传播、聚合形成下一时刻的节点信息，以 $t=1$ 为初始单位时刻。传播规则为，对于尾节点指向头节点的有向边，每个头节点接收尾节点的信息，也向尾节点反向发送当前信息。以图 5 的有向边 $\langle v_2, v_5 \rangle \in G_2'$ 为例，在 $t=1$ 时刻，节点 v_2 对应的信息 $\boldsymbol{e}_2^{(1)}$ 向 v_5 传播(实线箭头)，参与形成 $t=2$ 时刻的信息 $\boldsymbol{e}_5^{(2)}$；同时 $\boldsymbol{e}_5^{(1)}$ 反向传播到 v_2 (虚线箭头)，参与形成 $\boldsymbol{e}_2^{(2)}$。

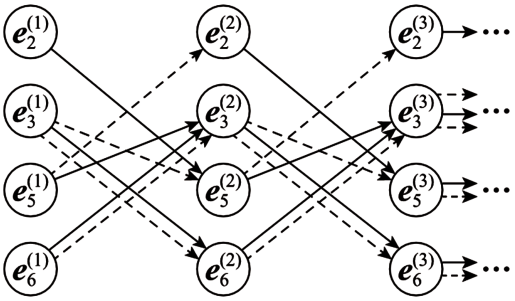

Figure 5. The propagation of GGNN per time unit
图 5. GGNN 单位时间传播示意

对于整个商品关系依赖图，TGGA-SR 设定多层 GGNN，聚合更多相邻商品的信息，使商品之间的联系更紧密。学习过程如式(5)至式(10)所示。

$$\boldsymbol{e}_i^{(1)} = \left[\boldsymbol{v}_i^{\mathrm{T}}, \boldsymbol{0}\right]^{\mathrm{T}} \left(i \in 1, \cdots, n\right) \tag{5}$$

$$\boldsymbol{a}_i^{(t)} = \boldsymbol{A}_{i:}^{\mathrm{T}} \left[\boldsymbol{e}_1^{(t-1)\mathrm{T}}, \cdots, \boldsymbol{e}_n^{(t-1)\mathrm{T}}\right]^{\mathrm{T}} \boldsymbol{H} + \boldsymbol{b} \tag{6}$$

$$\boldsymbol{z}_i^{(t)} = \sigma\left(\boldsymbol{W}_z \boldsymbol{a}_i^{(t)} + \boldsymbol{U}_z \boldsymbol{e}_i^{(t-1)}\right) \tag{7}$$

$$\boldsymbol{r}_i^{(t)} = \sigma\left(\boldsymbol{W}_r \boldsymbol{a}_i^{(t)} + \boldsymbol{U}_r \boldsymbol{e}_i^{(t-1)}\right) \tag{8}$$

$$\tilde{\boldsymbol{e}}_i^{(t)} = \tanh\left(\boldsymbol{W}_o \boldsymbol{a}_i^{(t)} + \boldsymbol{U}_o \left(\boldsymbol{r}_i^{(t)} \odot \boldsymbol{e}_i^{(t-1)}\right)\right) \tag{9}$$

$$\boldsymbol{e}_i^{(t)} = \left(1-\boldsymbol{z}_i^t\right)\odot \boldsymbol{e}_i^{(t-1)} + \boldsymbol{z}_i^t \odot \tilde{\boldsymbol{e}}_i^{(t)} \tag{10}$$

上述式中，式(5)的 $\boldsymbol{e}_i^{(1)}$ 为商品节点的初始向量表示。式(6)的 $\left[\boldsymbol{e}_1^{(t-1)},\cdots,\boldsymbol{e}_n^{(t-1)}\right]^{\mathrm{T}}$ 是单位时刻 $(t-1)$ 各商品节点向量表示；$\boldsymbol{A}$ 是商品关系依赖图 G 的出度矩阵 $\boldsymbol{A}_{(\mathrm{out})}$ 和入度矩阵 $\boldsymbol{A}_{(\mathrm{in})}$ 的拼接，$\boldsymbol{A}=\left[\boldsymbol{A}_{(\mathrm{out})},\boldsymbol{A}_{(\mathrm{in})}\right]\in\mathbb{R}^{n\times 2n}$；$\boldsymbol{A}_{i:}\in\mathbb{R}^{n\times 2}$ 表示从 $\boldsymbol{A}$ 中选择对应节点 v_i 的两列；$\boldsymbol{H}$ 为权重；$\boldsymbol{b}$ 为偏置。

式(7)至式(10)类似于 GRU (Gated Recurrent Unit，门控循环单元)正向传播过程，其中 $\sigma(\bullet)$ 为 sigmoid 激活函数，式(7)相当于更新门，用于控制上一时刻的信息的去留；式(8)相当于重置门，用于控制新信息的产生；式(9)为候选新向量表示，其中 $\odot$ 表示哈达玛积(Hadamard Product)，即对应元素乘积；式(10)通过对新旧向量表示的控制，形成下一时刻商品向量表示。

3.5. 用户向量表示

TGGA-SR 用 GRU 和 AM 生成用户向量，以表示用户偏好程度。用户向量表示的输入为会话的商品序列，其中的商品为上一步学习的商品表示。图 6 以会话 $s_2=[v_2,v_5,v_3,v_6,v_3]$ 为例进行用户向量表示的过程。

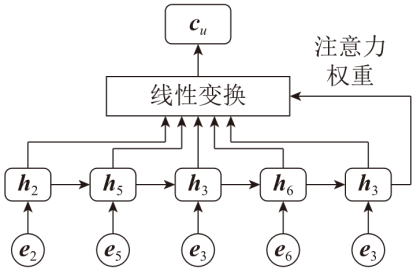

Figure 6. User vector representation
图 6. 用户向量表示

3.5.1. 用户会话处理

TGGA-SR 使用 GRU 处理会话，因其可避免传统 RNN 在长序列训练过程出现的梯度消失问题，也较 LSTM (Long Short-Term Memory，长短期记忆)使用更少的参数。GRU 模型如图 7 所示。

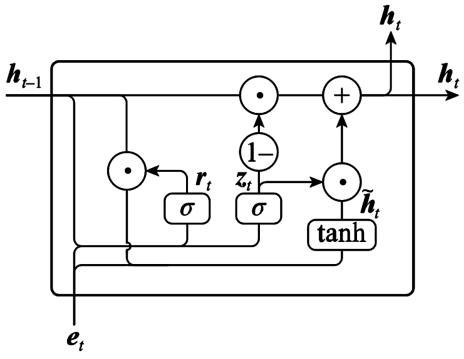

Figure 7. The GRU unit
图 7. GRU 单元结构

GRU 的前向传播如式(11)至式(14)所示。

$$\boldsymbol{z}_t = \sigma\left(\boldsymbol{W}_z\cdot\left[\boldsymbol{h}_{t-1},\boldsymbol{e}_t\right]\right)+\boldsymbol{b}_z \tag{11}$$

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{W}_r \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{e}_t\right]\right) + \boldsymbol{b}_r \tag{12}$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(\boldsymbol{W}_r \cdot \left[\boldsymbol{r} \odot \boldsymbol{h}_{t-1}, \boldsymbol{e}_t\right]\right) \tag{13}$$

$$\boldsymbol{h}_t = \left(1 - \boldsymbol{z}_t\right) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t \tag{14}$$

其中，$\boldsymbol{e}_t$ 为当前时刻商品向量输入，$\boldsymbol{h}_{t-1}$ 为上一时刻输出。式(11)和式(12)分别为更新门和重置门的处理过程。式(13)为候选新状态生成，并在式(14)与旧状态结合，形成新时刻状态。

3.5.2. 注意力机制

注意力机制(AM)源于人类对事物的观察，即人的目光聚焦在事物的关键区域，以获取所需信息，而其它无关区域会被忽略。推荐模型应用 AM 可将相关特征进行强化或抑制，以得到更准确的推荐结果。TGGA-SR 利用 AM 先计算每个特征的权值，再对特征进行加权求和，权值越大，对当前的结果影响就越大。

考虑到用户的兴趣具有阶段性变化，点击的商品与上一项有关联，即会话越靠后的商品对用户兴趣建模的贡献度越大，故应用 AM 对会话靠后的商品信息进行加强，同时也能对噪声类商品信息进一步抑制，实现实时推荐。设置注意力权重因子 $\alpha_{ni}\left(i = 1, \cdots, n\right)$，其中 n 为会话的末端商品，用于衡量两个 GRU 输出位置之间的关联程度。权重因子计算如式(15)所示。

$$\alpha_{ni} = q\left(\boldsymbol{h}_n, \boldsymbol{h}_i\right) \tag{15}$$

其展开如式(16)所示。

$$q\left(\boldsymbol{h}_n, \boldsymbol{h}_i\right) = \boldsymbol{V}^{\mathrm{T}} \sigma\left(\boldsymbol{A}_1 \boldsymbol{h}_n + \boldsymbol{A}_2 \boldsymbol{h}_i\right) \tag{16}$$

其中，参数矩阵 $\boldsymbol{A}_1$，$\boldsymbol{A}_2$ 分别将 LSTM 隐层状态 $\boldsymbol{h}_n$，$\boldsymbol{h}_i$ 进一步转化为隐向量表示。通过将隐层状态转换求和，得到的结果再经过 sigmoid 激活函数，获得新的隐向量表示。将新隐向量表示与参数矩阵 $\boldsymbol{V}^{\mathrm{T}}$ 进行矩阵乘法，最终得到权重因子。

将权重因子分别作用于各 GRU 单元输出，并进行线性变换，最终得到用户表示 $\boldsymbol{c}_u$，该过程如式(17)所示。

$$\boldsymbol{c}_u = \sum_{i=1}^{n} \alpha_{ni} \boldsymbol{h}_i \tag{17}$$

3.6. 推荐结果预测

在获得用户表示与商品表示后，TGGA-SR 先进行用户与商品的余弦相似度计算，再进行 softmax 归一化形成预测商品的概率分布，进行 Top-k 推荐。此过程如式(18)和(19)所示。

$$\hat{r}_i = \frac{\boldsymbol{c}_u \cdot \boldsymbol{e}_i}{\left\|\boldsymbol{c}_u\right\| \left\|\boldsymbol{e}_i\right\|} \tag{18}$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}\left(\hat{\boldsymbol{r}}\right) \tag{19}$$

由于用户对商品浏览行为只有浏览和不浏览两种，因此选择二分类交叉熵作为损失函数，如式(20)所示。

$$L = -\sum_{i=1}^{m} \left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right) \right] \tag{20}$$

其中，$\hat{y}_i$ 为预测概率分布，y_i 为真实分布。

4. 实验

4.1. 数据集与处理

本文使用 Yoochoose 和 Diginetica 两个数据集进行实验。Yoochoose 数据集是由英国在线零售商提供的用户在 6 个月内，含 9,249,729 件商品的共计 33,003,994 条用户点击记录，每条记录分别包含会话 ID、时间戳、商品 ID、商品类别信息。Diginetica 数据集记录了另一个电子商务平台用户在 5 个月内，含 43,097 件商品的共计 204,771 条用户点击记录，每条记录分别包含会话 ID、用户 ID、商品 ID、点击发生的时间帧、点击发生的日期信息。

本文对数据集进行拆分。对于 Yoochoose 数据集，将最后 1 天数据作测试集，其余作训练集；对于 Diginetica 数据集，将最后 7 天数据作测试集，其余作训练集。实验根据 Li 等人[15]和 Liu 等人[17]的工作进行数据预处理。两个数据集长度为 1 的会话和出现次数少于 5 次的商品将被过滤，同时过滤测试集中未曾出现在训练集中的商品。过滤后，Yoochoose 数据集有 7,981,580 个会话及 37,483 件商品，Diginetica 数据集则有 202,633 个会话及 43,097 件商品。

本文根据 Tan 等人[28]的工作对实验数据进行扩充。具体扩充一般方法为，原会话 $s_u=[v_1,v_2,\cdots,v_n]$ 拆分成一系列子会话和标签 $[(v_1),v_2]$，$[(v_1,v_2),v_3]$，……，$[(v_1,v_2,\cdots,v_{n-1}),v_n]$，其中 $(\bullet)$ 内的子会话为输入的会话，最后一项为结果标签。由于 Yoochoose 数据集数据量大，根据 Li 等人[15]和 Liu 等人[17]的工作，仅使用该数据集的靠后部分进行训练就可以获得更好的结果，故分别取该数据集最后的 1/64 和 1/4 数据进行实验。表 1 为实验数据集统计信息。

Table 1. Statistics of datasets
表 1. 数据集统计信息

数据集	Yoochoose 1/64	Yoochoose 1/4	Diginetica
点击次数	557,248	8,326,407	982,961
训练集会话数	369,859	5,917,745	719,470
测试集会话数	55,898	55,898	60,858
物品数	16,766	29,618	43,097
平均会话跳数	6.16	5.71	5.12

由于 Yoochoose 和 Diginetica 原始数据集均记录了商品点击时间帧，故模型可以记录会话内点击相邻商品之间的时间间隔。图 8 和图 9 分别展示了 Yoochoose 和 Diginetica 数据集相应点击时间间隔的频次分布。

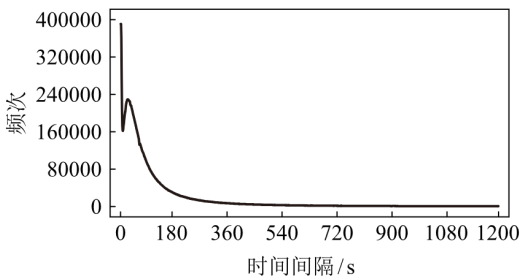

Figure 8. The distribution of click-time interval from Yoochoose datasets
图 8. Yoochoose 数据集点击时间间隔分布

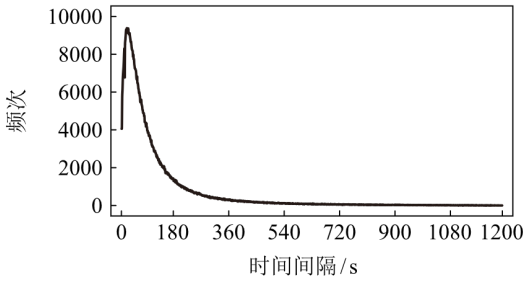

Figure 9. The distribution of click-time interval from Diginetica datasets
图 9. Diginetica 数据集点击时间间隔分布

需要说明的是，Yoochoose 数据集的最长时间间隔达 3600 s，Diginetica 数据集则长达 1200 s。表 2 以分段的形式完整给出了不同时间间隔区间的频次占比。

Table 2. Statistics of time interval from datasets
表 2. 数据集时间间隔统计信息

时间间隔区间	Yoochoose 数据集占比	Diginetica 数据集占比
0~100 s	67.96%	67.79%
101~200 s	16.30%	18.35%
201~300 s	5.78%	6.38%
301~400 s	2.80%	3.00%
401~500 s	1.62%	1.67%
501~600 s	1.04%	1.02%
601~700 s	0.73%	0.65%
701~800 s	0.54%	0.46%
801~900 s	0.42%	0.31%
901~1000 s	0.34%	0.20%
1001~1100 s	0.27%	0.12%
1101~1200 s	0.23%	0.04%
1201~3600 s	1.97%	-

虽然两个数据集组成不同，但分别处理和统计后可从统计图表看出，除了 0 s 附近时间间隔有差异，两个数据集的用户点击行为整体相近，大多数时间间隔集中在 180 s 以内，此后的频次随时间间隔增加呈现递减趋势。

由于 1200 s 以上的时间间隔频次少，较长的时间间隔可能由处理其它事务或离开屏幕而暂停浏览商品造成。为保持数据的完整性以及减少其它商品的会话相对点击时间率过小的影响，模型设置时间间隔上限为 1200 s。另外，Yoochoose 数据集频次最高的点击时间间隔集中在 5 s 附近，用户在如此短的时间内点击，往往未能有效阅读平台展示的商品信息。这些商品被视为噪声商品，不利于形成用户兴趣，给推荐结果带来不利影响。相对点击时间率的引入试图削弱短时间间隔浏览商品之间的联系。

4.2. 基准方法

为了评估本文所提出 TGGA-SR 模型的有效性，将该模型与以下模型进行对比。

第一类为传统会话推荐模型，包括 POP [14]、S-POP [14]、Item-KNN [7] [8]、BPR-MF [10]、FPMC [4]，部分模型已在前文讨论。POP 是简单的流行度预测方法，将数据集中出现频率最高的商品进行推荐；S-POP 是基于当前会话的流行度预测方法，为用户推荐当前会话下出现频率最高的商品。

第二类为基于深度学习的会话推荐模型，包括 GRU4REC [14]、NARM [15]、STAMP [17]、SR-GNN [24]、SEMGNN [26]，已在前文讨论。

4.3. 评价指标

本文采用会话推荐场景中最常用的 Precision@k 和 MRR@k (Mean Reciprocal Ranks, MRR)两种评价指标评估模型。

Precision@k 用于衡量基于会话的推荐系统的预测准确性，表示推荐结果列表中排在前 k 个推荐物品中，有正确物品的样本所占的比例。指标定义如式(21)所示。

$$\text{Precision}@k = \frac{n_{\text{hit}}}{N} \tag{21}$$

其中，n_{hit} 表示前 k 个推荐物品中有正确物品的样本数量，N 表示测试集的总样本数。

MRR@k 表示平均倒数排名，是在 Precision@k 方法的基础上，加入了商品位置的影响。在推荐物品列表中物品的位置越靠前，则其值越大，反之越小，当物品不在前 k 个推荐物品中时，该值为 0。指标定义如式(22)所示。

$$\text{MRR}@k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}(i)} \tag{22}$$

其中，N 表示测试集的总样本数，$\text{rank}(i)$ 表示第 i 个测试样本的推荐列表中正确物品所在的排列位置。

实验取 $k=20$，即前 20 件候选商品，这些商品在现实中往往出现在电子商务平台的第一页，受到大多数用户关注。

4.4. 参数设置

本文模型参数设定如下。商品向量嵌入维度 $d=100$，学习率 $\eta=0.001$，学习率衰减 $\lambda=0.1$，商品图神经网络嵌入表示层及 GRU 层的隐层节点个数均设置为 100。训练批次设定方面，Batch 大小为 100，迭代次数为 10。所有权重矩阵均采用服从 $N(0,0.1)$ 的正态分布随机初始化，L2 惩罚系数为 10^{-5}，算法使用 Adam 优化方法对模型参数进行求解。

4.5. 实验结果及分析

本文所提出 TGGA-SR 模型与其它模型的对比见表 3。

其中，FPMC 在处理 Yoochoose 1/4 数据集时出现内存不足问题，未给出结果。从表 3 可以看出，传统会话推荐模型缺乏对会话转移信息的有效捕捉，效果总体较基于深度学习的会话推荐模型差。在基于深度学习的会话推荐模型中，引入 RNN、GNN 等可以较好地表示会话和商品所隐含的信息，从而带来相对较好的推荐结果。本文的 TGGA-SR 则通过引入会话相对点击时间率对噪声商品的信息有所抑制，同时使用 AM 获得相对实时的推荐，提高了推荐性能，在 Yoochoose 1/64 和 Diginetica 数据集均取得相对显著的提升。

Table 3. Experimental results
表 3. 实验结果

模型	Yoochoose1/64		Yoochoose 1/4		Diginetica	
	Precision@20	MRR@20	Precision@20	MRR@20	Precision@20	MRR@20
POP	6.71	1.65	1.33	0.30	0.89	0.20
S-POP	30.44	18.35	27.08	17.75	21.06	13.68
Item-KNN	51.60	21.81	52.31	21.70	35.75	11.57
BPR-MF	31.31	12.08	3.40	1.57	5.24	1.98
FPMC	45.62	15.01	-	-	26.53	6.95
GRU4REC	60.64	22.89	59.53	22.60	29.45	8.33
NARM	68.32	28.63	69.73	29.23	49.70	16.17
STAMP	68.74	29.67	70.44	30.00	45.64	14.32
SR-GNN	68.69	29.36	69.90	30.30	49.56	16.92
SEMGNN	69.16	29.80	70.62	31.28	50.16	17.04
TGGA-SR	**70.44**	**30.85**	**71.33**	**32.09**	**51.80**	**18.29**

4.6. 消融实验

消融实验(ablation study)的作用是验证 TGGA-SR 模型的各个部分对推荐性能的影响，本文设置了四种模型加以对照。相对于完整的 TGGA-SR，–T 为仅去除相对点击时间率，保留商品关系依赖图，但其出度矩阵和入度矩阵的构造与 SR-GNN [24]相同；–TG 为仅去除商品关系依赖图和 GNN 的商品表示部分，直接学习会话信息；–GA 为仅去除 GRU 和 AM 的用户表示部分，直接将用户会话涉及商品表示的平均值作为用户表示；–A 为仅去除 AM，直接使用 GRU 建模会话形成用户表示。实验结果如图 10 和图 11 所示。

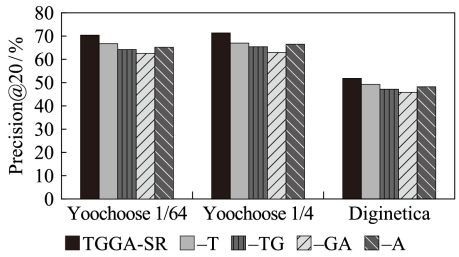

Figure 10. Results of ablation study (Precision@20)
图 10. 消融实验结果(Precision@20)

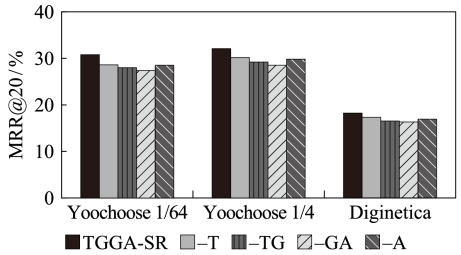

Figure 11. Results of ablation study (MRR@20)
图 11. 消融实验结果(MRR@20)

从图 10 和图 11 可以看出，对于仅去除用户表示部分(–GA)的模型，由于会话潜在信息不能被 GRU 和 AM 有效挖掘，造成用户表示不理想，推荐效果显著降低。对于仅去除商品表示部分(–TG)的模型，商品之间的联系由于分散在一维会话结构中而不能有效聚合，形成的推荐结果不一定很好地符合用户需求。对于仅去除 AM (–A)的模型，其缺少实时性修正，推荐的结果未必反映用户兴趣阶段性变化特点，效果不如完整模型。而仅去除相对点击时间率(–T)的模型则受到噪声商品影响，尤其是 Yoochoose 数据集出现较多随机点击行为，推荐性能较完整模型下降相对明显。

消融实验结果表明，本文 TGGA-SR 模型结合各模块的优势，具有更好的推荐效果。

5. 结束语

在基于会话的推荐算法中，目前大多数方法都只是使用会话中商品的信息，而潜在的用户点击商品时间间隔等信息没有利用，导致算法的学习不够充分。针对此问题，本文提出的 TGGA-SR 模型可以学到更充分的信息，模型的推荐性能用于其它模型。

但是，用户的长期兴趣和短期兴趣对推荐结果也有很大的影响，至于长期兴趣的表示，以及如何权衡长期兴趣与短期兴趣综合推荐，本文未作研究。后续将进一步研究用户的长短期兴趣对模型泛化能力的影响。

基金项目

广东省重点领域研发计划(2021B0101200002，2019B01018001，2020B0101100001)；广东省科技计划项目(2020B1010010010，2019B101001021)。

参考文献

[1] Wang, S., Cao, L., Wang, Y., *et al*. (2022) A Survey on Session-Based Recommender Systems. *ACM Computing Surveys*, **54**, Article No. 154. https://doi.org/10.1145/3465401

[2] Nilashi, M., Bagherifard, K., Ibrahim, O., *et al*. (2013) Collaborative Filtering Recommender Systems. *Research Journal of Applied Sciences, Engineering and Technology*, **5**, 4168-4182. https://doi.org/10.19026/rjaset.5.4644

[3] Koren, Y., Bell, R. and Volinsky, C. (2009) Matrix Factorization Techniques for Recommender Systems. *IEEE Computer Journal*, **42**, 30-37. https://doi.org/10.1109/MC.2009.263

[4] 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

[5] 黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述[J]. 计算机学报, 2018, 41(7): 1619-1647.

[6] Scarselli, F., Gori, M., Tsoi, A.C., *et al*. (2009) The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, **20**, 61-80. https://doi.org/10.1109/TNN.2008.2005605

[7] Sarwar, B., Karypis, G., Konstan, J., *et al*. (2001) Item-Based Collaborative Filtering Recommendation. In: *Proceedings of the 10th International Conference on World Wide Web*, Association for Computing Machinery, New York, 285-295. https://doi.org/10.1145/371920.372071

[8] Davidson, J., Liebald, B., Liu, J., *et al*. (2010) The YouTube Video Recommendation System. *RecSys*'10: *Proceedings of the* 4*th ACM conference on Recommender Systems*, Association for Computing Machinery, New York, 293-296. https://doi.org/10.1145/1864708.1864770

[9] Bamshad, M., Honghua, D., Tao, L., *et al*. (2001) Effective Personalization Based on Association Rule Discovery from Web Usage Data. *Proceedings of the* 3*rd ACM Workshop on Web Information and Data Management*, Atlanta, November 2001, 9-15.

[10] Rendle, S., Freudenthaler, C., Gantner, Z., *et al*. (2009) BPR: Bayesian Personalized Ranking from Implicit Feedback. In: *UAI*'09: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, Arlington, 452-461.

[11] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. Computer Science.

[12] He, Y., Sainath, T.N., Prabhavalkar, R., *et al*. (2018) Streaming End-to-End Speech Recognition for Mobile Devices. 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 6381-6385. https://doi.org/10.21437/Interspeech.2019-1341

[13] Zhou, L., Zhou, Y., Corso, J., *et al*. (2018) End-to-End Dense Video Captioning with Masked Transformer. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Salt Lake City, 18-23 June 2018, 8739-8748. https://doi.org/10.1109/CVPR.2018.00911

[14] Balázs, H., Alexandros, K., *et al*. (2016) Session-Based Recommendations with Recurrent Neural Networks. *Proceedings of the* 4*th International Conference on Learning Representations* (*ICLR* 2016), San Juan, 2-4 May 2016, 1-10. https://doi.org/10.48550/arXiv.1511.06939

[15] Li, J., Ren, P., Chen, Z., *et al*. (2017) Neural Attentive Session-Based Recommendation. In: *Proceeding of the* 11*th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, 1418. https://doi.org/10.1145/3132847.3132926

[16] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 6000-6010.

[17] Liu, Q., Zeng, Y., Mokhosi, R., *et al*. (2018) STAMP: Short-Term Attention/Memory Priority Model for Session-Based Recommendation. In: *KDD*'18: *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, 1831-1839. https://doi.org/10.1145/3219819.3219950

[18] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. 5*th International Conference on Learning Representations*, *ICLR* 2017—*Conference Track Proceedings*, Toulon, 24-26 April 2017, 1-14. https://doi.org/10.48550/arXiv.1609.02907

[19] Hamilton, W., Ying, Z. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. In: 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Curran Associates, Inc., Long Beach, 1-11. https://doi.org/10.48550/arXiv.1706.02216

[20] Li, Y., Tarlow, D., Brockschmidt, M., *et al*. (2016) Gated Graph Sequence Neural Networks. *Proceedings of the* 4*th International Conference on Learning Representations* (*ICLR* 2016), Caribe Hilton, 2-4 May 2016, 1-20. https://doi.org/10.48550/arXiv.1511.05493

[21] Yehuda, K., Robert, B. and Chris, V. (2009) Matrix Factorization Techniques for Recommender Systems. *IEEE Computer Journal*, **42**, 30-37. https://doi.org/10.1109/MC.2009.263

[22] Goyal, P. and Ferrara, E. (2017) Graph Embedding Techniques, Applications, and Performance: A Survey. *Knowledge Based Systems*, **151**, 78-94. https://doi.org/10.1016/j.knosys.2018.03.022

[23] Dong, X., Gabrilovich, E. and Heitz, G. (2014) Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In: *Proceedings of the* 20*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, 601-610. https://doi.org/10.1145/2623330.2623623

[24] Wu, S., Tang, Y., Zhu, Y., *et al*. (2019) Session-Based Recommendation with Graph Neural Networks. In: *Proceedings of the* 2019 33*rd AAAI Conference on Artificial Intelligence*, AAAI Press, Palo Alto, 346-353. https://doi.org/10.1609/aaai.v33i01.3301346

[25] Yu, F., Zhu, Y.Q., Liu, Q., *et al*. (2020) TAGNN: Target Attentive Graph Neural Networks for Session-Based Recommendation. In: *Proceedings of the* 2020 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, 1921-1924. https://doi.org/10.1145/3397271.3401319

[26] 任俊伟, 曾诚, 肖丝雨, 等. 基于会话的多粒度图神经网络推荐模型[J]. 计算机应用, 2021, 41(11): 3164-3170.

[27] 刘洪伟, 高鸿铭, 陈丽, 等. 基于用户浏览行为的兴趣识别管理模型[J]. 数据分析与知识发现, 2018, 2(2): 74-85.

[28] Tan, Y.K., Xu, X. and Liu, Y. (2016) Improved Recurrent Neural Networks for Session-Based Recommendations. In: *DLRS* 2016: *Proceedings of the* 1*st Workshop on Deep Learning for Recommender Systems*, Association for Computing Machinery, New York, 17-22. https://doi.org/10.1145/2988450.2988452