

基于双编码器的会话型推荐模型

方军¹, 管业鹏^{1,2}

(1. 上海大学通信与信息工程学院, 200444, 上海;

2. 上海大学新型显示技术及应用集成教育部重点实验室, 200072, 上海)

摘要: 针对现有会话型推荐模型难以准确捕获物品间全局依赖的问题, 提出了一种基于双编码器的会话型推荐模型(SR-BE)。该双编码器由基于自注意力网络的全局编码器和基于图神经网络的局部编码器组成, 无论被浏览物品之间的时间间隔长还是短, 全局编码器都能够利用注意力机制自适应地捕获被浏览物品之间的全局依赖, 并将其编码为全局隐向量。为弥补自注意力网络没有结构信息而难以捕获邻近物品间局部依赖的不足, 在局部编码器中, 首先将会话序列构建成会话图, 然后通过图神经网络在会话图上捕获邻近物品间的局部依赖, 并将其编码为局部隐向量。最后将从双编码器得到的全局隐向量和局部隐向量线性组合为会话表示, 再通过预测层解码会话表示得到每个候选物品被点击的概率。实验结果表明: 将基于自注意力网络的全局编码器与基于图神经网络的局部编码器结合在一起, 比单一地使用全局编码器或局部编码器在命中率上分别提高了3.11%和6.55%。通过与同类模型客观定量比较, SR-BE模型在两个公开数据集上取得了突出的效果, 表明该模型有效、可行。

关键词: 会话型推荐; 自注意力网络; 图神经网络; 编码器; 推荐模型

中图分类号: TP391.3 **文献标志码:** A

DOI: 10.7652/xjtub202108020 **文章编号:** 0253-987X(2021)08-0166-09



OSID 码

A Session Recommendation Model Based on Dual Encoders

FANG Jun¹, GUAN Yepeng^{1,2}

(1. School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; 2. Ministry of Education Key Laboratory of Advanced Display and System Application, Shanghai University, Shanghai 200072, China)

Abstract: A session recommendation model based on dual encoders (SR-BE) is proposed to solve the problem that it is difficult to accurately capture the global dependencies between items in the existing session recommendation models. The dual encoder consists of a global encoder based on self-attention network and a local encoder based on graph neural network. The global encoder can adaptively capture the global dependencies among all items by using the attention mechanism regardless of their time intervals and encode them into global latent vectors. In order to make up for the lack of local dependencies in self-attention network, the local encoder first constructs the session sequences into a session graphs, then the graph neural networks are used to capture the local dependencies among adjacent items and encode them into local latent vectors. Finally, the linear combination of global and local latent vectors obtained from the dual encoder is used to obtain the session representations and the click probability of each candidate item is obtained by decoding the session representations through the prediction layer. Experimental results and

收稿日期: 2020-12-22。 作者简介: 方军(1989—),男,硕士生;管业鹏(通信作者),男,教授,博士生导师。 基金项目: 国家重点研发计划资助项目(2019YCF1520500,2020YFC1523004)。

网络出版时间: 2021-05-10

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1069.T.20210510.1221.002.html>

comparisons with a single global encoder and a single local encoder show that the combination of the global encoder based on self-attention networks and the local encoder based on graph neural networks improves the hit rate by 3.11% and 6.55%, respectively. Comparisons with state-of-the-art methods on two public datasets validate the effectiveness and feasibility of the proposed method.

Keywords: session-based recommendation; self-attention network; graph neural network; encoder; recommendation model

随着互联网的高速发展和通信技术的不断进步,信息过载在各种网络应用上成为了一个严重的问题,而推荐系统是缓解信息过载和提升用户体验必不可少的工具。无论用户是否注册或登录,会话都能够记录一段时间内用户的行为信息(如点击)。会话型推荐的任务是根据当前会话中已知用户的行为来预测用户的下一次行为。

由于会话型推荐有着较高的实用价值,所以吸引了很多学者进行了相关研究。马尔可夫链是一种经典的方法,它假定下一个行为取决于最后若干个行为,但这使得基于马尔可夫链的模型^[1]无法利用长序列中行为之间的依赖,并且可能会出现数据稀疏的问题。

近年来,基于深度学习的方法取得了突飞猛进的进展,其中基于循环神经网络(RNN)的研究相比传统方法取得了长足的进步。门控循环推荐模型(GRU4REC)^[2]将RNN用于建模用户的顺序行为以捕获用户当前对物品的偏好。神经注意力推荐机(NARM)^[3]采用两个RNN,其中一个用于捕获用户的顺序行为,另一个RNN与注意力机制结合以捕获主要意图。类似NARM,短期记忆优先模型(STAMP)^[4]同样利用注意力机制捕获用户的当前兴趣和总体兴趣,但它用多层感知机来替代RNN。重复网模型(RepeatNet)^[5]考虑了重复消费现象,提出重复-探索机制。协同会话推荐机(CSRM)^[6]将RNN与记忆网络相结合,将相邻会话表示存储在记忆网络之中,然后将存储在记忆网络中的协同信息与用RNN编码的当前会话表示通过融合门相结合。个性化分层循环模型(PHRM)^[7]采用3个RNN分别捕获会话级、区块级和用户级兴趣。图表示学习的会话感知模型(GESP)^[8]将RNN与图表示学习结合以缓解用户的兴趣漂移问题。但RNN难以捕获长范围的依赖。

由于图神经网络(GNN)能够捕获节点间的依赖关系和以此带来的出色性能表现,近年来越来越

多的工作采用了GNN来解决会话型推荐问题。图神经网络会话推荐模型(SR-GNN)^[9]将会话序列构建成会话图,使用GNN来捕获物品间的转移。目标注意力图神经网络模型(TAGNN)^[10]在SR-GNN模型^[9]的基础上提出一种目标注意力网络,它通过将候选物品的嵌入作为注意力机制的查询向量来发现当前会话中的物品与候选目标物品间的相关性。无损边序保留聚合及捷径图注意力会话推荐模型(LESSR)^[11]提出边顺序保留聚合层来解决基于GNN的顺序信息丢失问题。但GNN只能捕获邻节点间的局部依赖。

尽管以上方法取得了不错的推荐效果^[12-14],但是难以准确捕获远距离物品间的依赖。最近,在自然语言处理领域,出现了一个叫做Transformer的基于注意力机制的模型^[15]。它在WMT 2014^[16]翻译的任务上取得了优异的成绩。它所采用的自注意力网络是它成功的关键因素之一。自注意力网络能够有效捕获所有物品间的全局依赖,但它缺少捕获相邻物品间局部依赖性的能力,这限制了它捕获序列的上下文信息。图情境自注意力网络模型(GC-SAN)^[17]将GNN与自注意力网络级联来捕获长距离依赖。然而,节点输入自注意力网络前使用GNN对邻节点间进行的图卷积使得每个节点变得平滑,这导致输入自注意力网络后不能准确的捕获物品间的全局依赖。

综上所述,已有的会话型推荐研究并不能有效捕获物品间的全局依赖与局部依赖。对此,本文提出了SR-BE模型,该模型利用了自注意力网络和图神经网络之间的互补性,采用基于自注意力网络作为全局编码器和基于图神经网络为局部编码器的双编码器结构来捕获物品间的依赖。最后将全局编码器和局部编码器输出的线性组合作为会话表示,再通过预测层解码会话表示得到每个候选物品被点击的概率。为使实验结果可复现,源代码网址为 <https://github.com/GalaxyCruiser/SR-BE>。

1 SR-BE 会话型推荐模型

1.1 问题描述

设包含所有会话里涉及到的物品的集合为 $V = \{v_1, v_2, \dots, v_m\}$ 。 m 为集合 V 的大小。 一个匿名会话序列可以表示为 $S = [s_1, s_2, \dots, s_n]$, 其中 $s_i \in V$ 表示在会话 S 中用户点击的一个物品。 会话型推荐的任务就是为会话 S 预测用户下一个点击的物品 s_{n+1} 。 模型为 V 中每一个候选物品学习生成一个得分。 设 $\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m]$ 为模型输出的点击概率向量, 其中 $\hat{y}_i$ 为物品 v_i 对应的点击概率。 $\hat{y}$ 中点击概率最高的若干个物品作为被推荐的物品。

1.2 模型概述

SR-BE 由嵌入层, 全局编码器, 局部编码器和预测层组成。 嵌入层将输入映射到低维密集空间。 基于自注意力网络的全局编码器用于捕获全局依赖, 基于图神经网络的局部编码器用于捕获局部依赖。 全局编码器与局部编码器输出的线性的组合作为预测层的输入, 预测层输出每个物品被点击的概率。 模型的架构如图 1 所示。

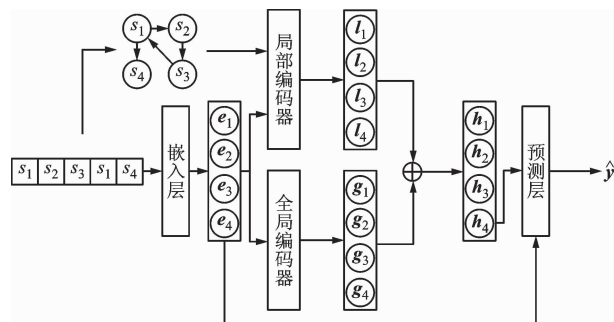


图 1 基于双编码器的会话型推荐模型的架构

Fig. 1 The architecture of the session recommendation model based on dual encoder

1.3 嵌入层

嵌入层可以理解为一个字典, 它将物品的索引映射为密集向量, 将接收物品的整数索引作为输入, 并在字典中查找整数索引, 然后返回对应的向量。 将会话序列 $S = [s_1, s_2, \dots, s_n]$ 输入嵌入层来得到对应的嵌入序列 $E = [e_1, e_2, \dots, e_n]$, 嵌入的维度为 d , 即 $e_i \in \mathbf{R}^d$ 。

1.4 基于自注意力网络的全局编码器

自注意力是注意力机制的特例, 它将同一个输入用于注意力机制的查询向量、键向量和值向量。 因此, 每一个物品对应的输出都能获取所有输入物品的信息^[15]。 这使得自注意力机制能够捕获当前

会话中所有物品间的全局依赖且不受物品间的距离影响。

将通过嵌入层得到的嵌入序列 E 输入自注意力层来捕获物品间的全局依赖^[15]

$$F = \text{softmax}\left(\frac{(EW_Q)(EW_K)^T}{\sqrt{d}}\right)(EW_V) + E \quad (1)$$

式中: W_Q, W_K 和 $W_V \in \mathbf{R}^{d \times d}$ 分别为查询、键、值参数矩阵。 当嵌入维度 d 取一个较大的数时, 点积的结果可能会是一个绝对值很大的数, 为防止 softmax 函数进入梯度很小的区域^[15], 将查询向量与键向量的点积除以 $d^{1/2}$ 。 为保留低层信息, 自注意力层采用了残差连接^[15]。

然后, 将 F 按元素输入到一个前馈神经网络当中, 该网络由两个线性变换组成, 两个线性变换间包含了一个 Relu 激活函数。 前馈神经网络同样采用了残差连接^[15]

$$G = \max(0, FW_1 + b_1)W_2 + b_2 + F \quad (2)$$

式中: $G = [g_1, g_2, \dots, g_n]$ 为会话中物品的全局隐向量; $W_1 \in \mathbf{R}^{d \times f}$; $W_2 \in \mathbf{R}^{f \times d}$ 为参数矩阵; $b_1 \in \mathbf{R}^f$; $b_2 \in \mathbf{R}^d$ 为偏置向量, f 为前馈神经网络的维度; G 为全局编码器的输出。

1.5 基于图神经网络的局部编码器

1.5.1 图构建 对于给定的会话序列 $S = [s_1, s_2, \dots, s_n]$, 序列里每个物品 s_i 可以看作是一个节点, s_i 在 s_{i-1} 后的点击可以看作是两节点间的一条有向边。 因此一个会话序列可以构建成一个有向图。 对于在会话 S 中出现多于一次的物品, 边的权重通过除以节点的度来归一化。 设 $A^I, A^O \in \mathbf{R}^{n \times n}$ 分别为入矩阵和出矩阵。 入矩阵和出矩阵的每一行分别表示该节点与其他节点的入边与出边关系, 其值大小为边的权重。 当会话序列 $S = [s_1, s_2, s_3, s_1, s_4]$ 时, 其对应的会话图、入矩阵和出矩阵如图 2 所示。

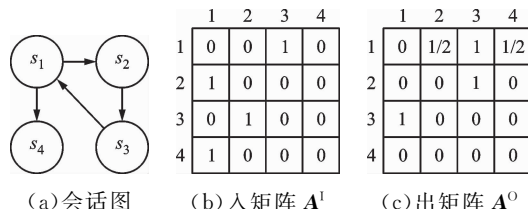


图 2 会话图及其入矩阵 A^I 和出矩阵 A^O 的样例

Fig. 2 Examples of a session graph and its incoming matrix A^I and outgoing matrix A^O

以节点 s_1 为例, 该节点仅有一条来自节点 s_3 的入边, 故入矩阵对应的第 1 行第 3 列为 1; 有两条去往节点 s_2 和 s_4 的出边, 出度为 2, 因此出矩阵对应

的第 1 行第 2 和第 4 列为 1/2。

1.5.2 局部隐向量学习 局部编码器采用门控图神经网络^[18]来学习局部隐向量。对于会话图中第 t 个节点 s_t ,其图卷积为^[9]

$$\begin{aligned} \mathbf{a}_t &= \text{concat}(\mathbf{A}_t^1([\mathbf{e}_1, \dots, \mathbf{e}_n]^T \mathbf{W}^1 + \mathbf{b}^1), \\ &\quad \mathbf{A}_t^0([\mathbf{e}_1, \dots, \mathbf{e}_n]^T \mathbf{W}^0 + \mathbf{b}^0)) \end{aligned} \quad (3)$$

式中: $\mathbf{a}_t \in \mathbf{R}^{2d}$ 为图卷积结果; $\mathbf{W}^1, \mathbf{W}^0 \in \mathbf{R}^{d \times d}$ 为参数矩阵; $\mathbf{b}^1, \mathbf{b}^0 \in \mathbf{R}^d$ 为偏置向量; $\mathbf{A}_t^1, \mathbf{A}_t^0 \in \mathbf{R}^{1 \times n}$ 为对应矩阵的第 t 行; $\text{concat}(\cdot)$ 为拼接函数,其作用是将输入的向量进行拼接。将 $\mathbf{a}_t$ 和 $\mathbf{e}_t$ 输入到门控循环单元^[9](GRU),可得

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{a}_t + \mathbf{U}_z \mathbf{e}_t) \quad (4)$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{a}_t + \mathbf{U}_r \mathbf{e}_t) \quad (5)$$

$$\mathbf{c}_t = \tanh(\mathbf{W}_o \mathbf{a}_t + \mathbf{U}_o(\mathbf{r}_t \odot \mathbf{e}_t)) \quad (6)$$

$$\mathbf{l}_t = (1 - \mathbf{z}_t) \odot \mathbf{e}_t + \mathbf{z}_t \odot \mathbf{c}_t \quad (7)$$

式中: $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W}_o \in \mathbf{R}^{d \times 2d}$; $\mathbf{U}_z, \mathbf{U}_r, \mathbf{U}_o \in \mathbf{R}^{d \times d}$ 为参数矩阵; $\sigma(\cdot)$ 为 sigmoid 函数; $\odot$ 表示按对应元素相乘; $\mathbf{z}_t$ 和 $\mathbf{r}_t$ 是更新门和重置门,分别用于控制信息的保留和遗弃; $\mathbf{c}_t$ 为候选态; $\mathbf{l}_t$ 为节点 t 的局部隐向量。会话中物品的局部隐向量 $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_n]$ 为局部编码器的输出。

1.6 预测层

会话表示 $\mathbf{h}$ 是会话中第 n 个物品的全局隐向量 $\mathbf{g}_n$ 和局部隐向量 $\mathbf{l}_n$ 的线性组合

$$\mathbf{h} = \omega \mathbf{g}_n + (1 - \omega) \mathbf{l}_n \quad (8)$$

式中, ω 为权重因子,控制全局编码器和局部编码器在会话表示中所占的比重。

预测层由采用点积的计算得分和 softmax 函数组成。对于会话 $\mathbf{S}$ 中每个候选物品 $v_i \in V$,其对应的得分 z_i 为

$$z_i = \mathbf{h}^T \mathbf{e}_i \quad (9)$$

模型的输出向量 $\hat{\mathbf{y}}$ 通过将 $\mathbf{z} = [z_1, z_2, \dots, z_n]$ 输入 softmax 函数得到

$$\mathbf{y} = \text{softmax}(\mathbf{z}) \quad (10)$$

输出向量 $\hat{\mathbf{y}}$ 表示候选物品被点击的概率。

1.7 模型训练

损失函数定义为真值与预测值之间的交叉熵

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

式中, $\mathbf{y}$ 表示真值物品的独热编码,该损失函数通过随时间反向传播(back-propagation through time)算法来优化。

2 实验结果与分析

2.1 数据集

为验证所提模型的有效性,本文在两个代表性的公开数据集 Yoochoose^[19] 和 Diginetica^[20] 上进行了评估实验。Yoochoose^[19] 是来自 RecSys Challenge 2015 的一个包含电商网站 6 个月点击记录的数据集。由于 Yoochoose 数据集十分庞大,本文遵循对比方法将其最近一段时间的 1/64 数据用作实验,记为 Yoochoose 1/64。Diginetica^[20] 是来自 CIKM cup 2016 比赛的数据集,其中的 transaction 数据适用于本次实验。为公平比较,将长度为 1 的会话和在数据集中出现次数少于 5 次的物品滤除^[2-11,17],并且采用相同的数据增强方法^[20]。采用 Yoochoose 1/64 数据集^[19] 最后一周的数据和 Diginetica^[20] 最后一天的数据分别作为各自的测试集。数据集的统计信息如表 1 所示。

2.2 对比模型

为客观定量评价本文模型的有效性,将本文模型与下述代表性模型进行比较。因子分解个性化马尔可夫链(FPMC)^[1] 是基于马尔可夫链的下一篮(basket)推荐,为适用于会话型推荐,下一个物品被看作下一篮。基于物品的近邻方法(Item-KNN)^[22] 推荐基于当前会话已点击物品的余弦相似度。基于矩阵分解的贝叶斯个性化排序(BPR-MF)^[23] 利用随机梯度下降法优化对排序目标函数。GRU4REC 模型^[2] 首次将 RNN 用于会话型推荐。NARM 模型^[3] 将 RNN 与注意力机制结合以捕获用户的主要意图和顺序行为。STAMP 模型^[4] 利用注意力机制和多层感知机捕获会话的总体兴趣和当前兴趣。RepeatNet 模型^[5] 考虑了重复消费现象,提出重复-

表 1 实验数据集统计数据

Table 1 Statistical data of experimental datasets

数据集	点击数	训练会话数	测试会话数	物品数	平均长度/次
Yoochoose 1/64 ^[19]	557 248	369 859	55 898	16 766	6.16
Diginetica ^[20]	982 961	719 470	60 858	43 097	5.12

探索机制。CSRM 模型^[6]将 RNN 与记忆网络相结合,使得模型在推荐时可以利用协同信息。SR-GNN^[9]将会话序列构建成会话图,使用 GNN 来捕获物品间的转移。TAGNN 模型^[10]提出一种目标注意力网络,能够发现当前会话物品与目标物品间的相关性。LESSR 模型^[11]采用边顺序保留聚合层来解决基于图神经网络的顺序信息丢失问题。GC-SAN 模型^[17]将图神经网络与自注意力网络级联来捕获长距离依赖。

2.3 评价指标

由于推荐系统每次只能推荐少量的物品,故采用以下在相关工作中广泛使用的评价指标。

命中率(hit rate, HR)表示在所有测试样例中正确推荐的比例,目标标签出现在推荐列表的前 20 个物品中即为命中,其表达式为

$$R_{H,20} = \frac{n_{hit}}{N}$$
 (12)

式中: N 表示测试集中会话的数量; n_{hit} 表示命中的数量。

排名倒数均值(mean reciprocal rank, MRR)表示目标标签在推荐列表中排名倒数的平均值,通常用百分数表示,当排名超过 20 时,排名倒数设为 0。设一组数据含有 4 个会话,正确结果出现在各自推荐列表的排名分别为 4、1、2、4,则其对应的排名倒数分别为 1/4、1、1/2、1/4,这组数据的平均倒排名为 $(1/4+1+1/2+1/4)/4$,等于 0.5。该指标能够反

映出推荐的精度,推荐的结果排名数值越小,则排名倒数越大,推荐精度越高。排名倒数均值的计算公式为

$$R_{MR,20} = \frac{1}{N} \sum_{v_{label} \in S_{test}} \frac{1}{rank(v_{label})}$$
 (13)

式中: N 表示测试集中会话的数量; v_{label} 是会话的标签; S_{label} 为测试集的标签集; $rank(\cdot)$ 为排名函数,用于获取标签在推荐列表中的排名。

2.4 实验设置

SR-BE 的所有参数都采用平均值为 0、标准差为 0.1 的高斯分布初始化。采用初始学习率为 0.001、每 3 个 epoch 衰减 0.1 的 Adam 优化器。batch 大小设为 100。Yoochoose 1/64^[19] 和 Diginetica^[20] 数据集的嵌入维度分别设为 96 和 72。自注意力网络中的前馈神经网络的大小 f 设为嵌入维度的 4 倍^[15]。L2 正则化设为 10^{-5} 以缓解过拟合。

2.5 与对比模型比较

为客观定量评价本文模型的有效性。选取以上对比模型在两个公开数据集 Yoochoose 1/64^[19] 和 Diginetica^[20] 以命中率和排名倒数均值为指标进行比较,实验结果如表 2 所示,最优结果以加粗字体突出。从表 2 可以看出,本文模型在两个评价指标上都优于其他对比模型。

传统方法^[1,22-23] 性能较为一般,这些方法仅仅基于物品相似度或者物品间的转移,忽略了会话顺序上的信息。基于神经网络的方法显著超过传统方

表 2 本文 SR-BE 模型与对比模型的性能比较

Table 2 A performance comparison between the proposed SR-BE model and the baseline models

模型	排名倒数均值/%		排名倒数均值/%	
	Yoochoose 1/64 ^[19]	Diginetica ^[20]	Yoochoose 1/64 ^[19]	Diginetica ^[20]
Item-KNN ^[22]	51.60	35.75	21.81	11.57
BPR-MF ^[23]	31.31	5.24	12.08	1.98
FPMC ^[1]	45.62	26.53	15.01	6.95
GRU4REC ^[2]	60.64	29.45	22.89	8.33
NARM ^[3]	68.32	49.70	28.63	16.17
STAMP ^[4]	68.74	45.64	29.67	14.32
RepeatNet ^[5]	70.71	47.79	31.03	17.66
CSRM ^[6]	71.45	50.55	30.36	16.38
SR-GNN ^[9]	70.57	50.73	30.94	17.59
GC-SAN ^[17]	70.66	51.70	30.04	17.61
TAGNN ^[10]	71.02	51.31	31.12	18.03
LESSR ^[11]	70.64	51.71	30.97	18.15
SR-BE	71.62	52.72	31.54	18.24

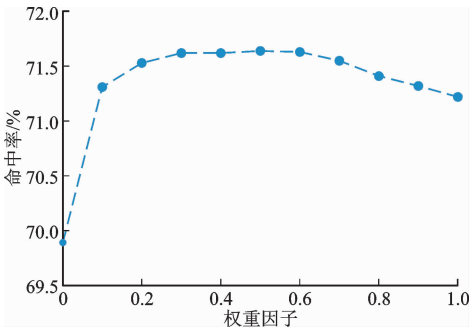
法,说明了深度学习的方法在会话型推荐领域的有效性。GRU4REC^[2]是首个采用 RNN 的会话型推荐模型,尽管在 Diginetica 数据集^[20]上的性能不如 Item-KNN^[22],但依然能够体现出 RNN 在构建序列模型的能力。NARM^[3]和 STAMP^[4]都利用注意力机制来捕获全局偏好,性能都显著优于 GRU4REC^[2],表明全局偏好对于提高推荐不可忽视。RepeatNet^[5]根据重复消费现象,提出重复-探索机制,显著提升了目标标签出现在当前会话物品中的情形下的性能,总体上优于 NARM^[3]和 STAMP^[4],然而在非重复物品场景下效果有所下降。CSRM^[6]性能优于 NARM^[3],说明了利用其他会话里的协同信息的有效性,但 CSRM^[6]的存储网络所存储相邻会话是基于顺序的,没有区分这些会话与当前会话的关联性。基于 RNN 的方法,当物品距离远时会产生梯度消失或者梯度爆炸,难以捕获到远距离物品间的依赖。

基于 GNN 网络的方法总体上优于基于 RNN 网络的方法。SR-GNN^[9]将会话序列构建成图,再使用 GNN 网络能够捕获到邻节点之间的依赖,性能总体优于多数基于 RNN 的方法。TAGNN^[10]在 SR-GNN^[9]基础上提出了目标注意力模块,使得模型能够发现当前会话物品与目标物品之间的关联性,性能上超越了 SR-GNN^[9],但由于需要对每个候选项做注意力计算,使得 TAGNN^[10]在训练过程中十分缓慢。由于 LESSR^[11]未提供在 Yoochoose 数据集上的结果,本文利用 LESSR^[11]所提供的源代码和默认设置完成了在 Yoochoose 1/64 数据集^[19]上的实验。LESSR^[11]通过解决会话序列转化为图网络时顺序信息丢失的问题,使得性能较 SR-GNN^[9]更佳,表明保留会话的序列信息的有效性,但它仅考虑了节点的入边依赖,忽略了节点的出边依赖。GC-SAN^[17]性能与 SR-GNN^[9]相近。虽然本文与 GC-SAN^[17]都采用了自注意力网络来捕获物品间的全局依赖关系,但是 GC-SAN^[17]捕获全局依赖关系时,自注意力网络的查询向量是在 GNN 网络下聚合了邻节点信息的节点隐向量,这会压制其他非邻节点的注意力权重。因此,采用这种方式不能准确捕获会话的全局依赖关系。而本文模型中的自注意力网络的查询向量中完整保留了对应物品的特征,能够有效捕获物品间的全局依赖。SR-BE 在两个数据集上以命中率和排名倒数均值为评价指标都取得了最佳效果。实验结果表明 SR-BE 采用自注意力网络捕获物品的全局依赖,图神经网络捕

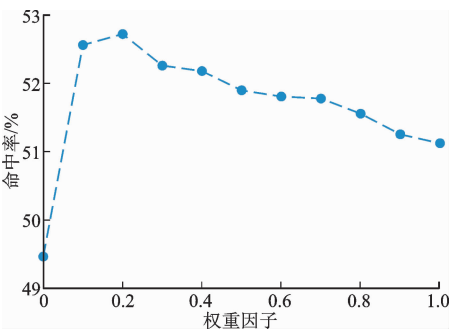
获邻近节点间的局部依赖,能够得到更加精准的会话表示,从而有效的提高了会话推荐的准确度。

2.6 权重因子的影响

权重因子 ω 控制了自注意力网络和图神经网络的对会话表示的影响力。图 3、图 4 显示了权重因子对命中率和排名倒数均值的影响。由图 3 和图 4 可知,4 幅曲线图都呈现出两头低、中间高、右端点高于左端点的态势。这表明仅仅采用全局编码器 ($\omega=1$) 比仅仅采用局部编码器 ($\omega=0$) 性能更佳;将基于自注意力网络的全局编码器与基于图神经网络的局部编码器结合在一起,能够比单一使用其中任何一个编码器得到更好的效果。这说明全局编码器能够有效的捕获物品间的全局依赖,但是局部依赖性对于提高推荐性能同样必不可少。对于 Yoo-



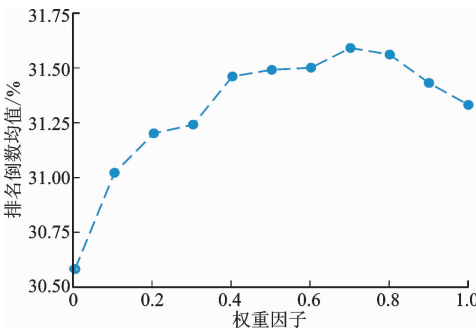
(a) 在 Yoochoose 1/64 数据集的影响



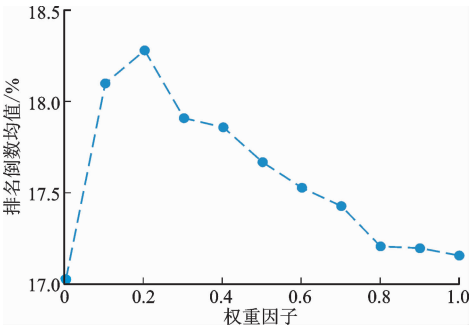
(b) 在 Diginetica 数据集的影响

图 3 权重因子对命中率的影响

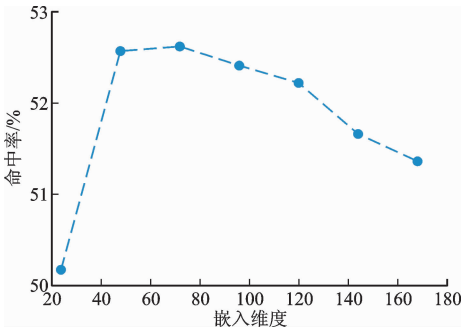
Fig. 3 The impact of weight factor on hit rate



(a) 在 Yoochoose 1/64 数据集的影响



(b)在 Diginetica 数据集的影响



(a)对命中率的影响

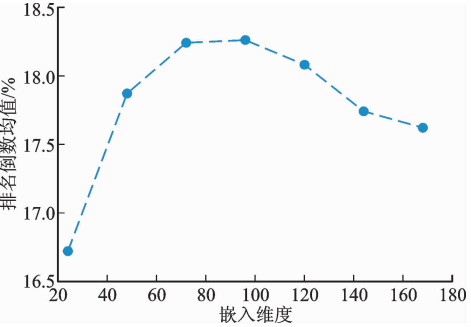
图 4 权重因子对排名倒数均值的影响

Fig. 4 The impact of weight factor on mean reciprocal rank

choose 1/64 数据集^[19],在命中率指标上的最佳权重因子在 0.5 附近,在排名倒数均值上最佳权重在 0.7 附近。对于 Diginetica^[20]数据集,在命中率和排名倒数均值上的最佳权重因子在 0.2 附近。

2.7 嵌入维度的影响

图 5、图 6 分别展示了嵌入维度 d 在 24~168 范围内对 Yoochoose 1/64 数据集^[19]和 Diginetica 数据集^[20]的影响,其中 Yoochoose 1/64 数据集^[19]的权重因子设为 0.5,Diginetica 数据集^[20]的权重因子设为 0.2。由图 5、图 6 可知,增大嵌入维度并非



(b)对排名倒数均值的影响

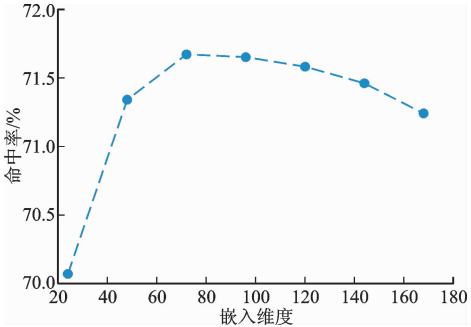
图 6 嵌入维度在 Diginetica 数据集的影响

Fig. 6 The impact of embedding size on Diginetica dataset

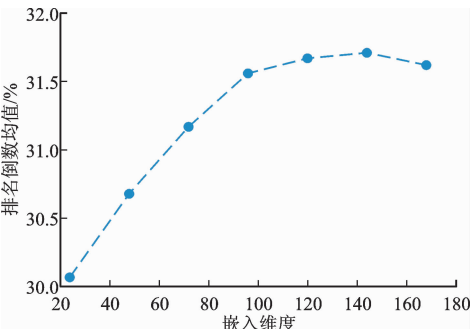
据集^[19],它在命中率指标上的最佳值在 72 附近,它在排名倒数均值上的最佳值在 144 附近。对于 Diginetica 数据集^[20],它在命中率指标上的最佳值在 72 附近,它在排名倒数均值上的最佳值在 96 附近。当嵌入维度超过最佳值后,模型的性能开始下降,这是因为模型进入了过拟合的状态;当嵌入维度低于最佳值时,模型的性能随着嵌入维度的增加而提高,这是因为更大的嵌入维度为模型提供了更大的学习容量。

2.8 消融实验

为进一步说明局部编码器和全局编码器对会话型推荐的影响,本文比较了 SR-BE 模型和它的两种变种的性能。其中 SR-BE-L 表示只使用局部编码器的 SR-BE 模型,它只捕获邻近物品间的局部依赖关系。SR-BE-G 代表只使用全局编码器的 SR-BE 模型。SR-BE、SR-BE-L 和 SR-BE-G 的实验结果如表 3 所示。从表 3 中可以得出以下结论:首先,SR-BE-G 性能优于 SR-BE-L,这说明基于自注意力网络的全局编码器能够有效捕获物品间的全局依赖,单独使用时较基于图神经网络的局部编码器准确度更高;其次,SR-BE 模型在两个数据集的所有评价指标上都优于 SR-BE-L 和 SR-BE-G,这说明将基于自注意力网络的全局编码器与基于图神经网络的局



(a)对命中率的影响



(b)对排名倒数均值的影响

图 5 嵌入维度在 Yoochoose 1/64 数据集的影响

Fig. 5 The impact of embedding size on Yoochoose 1/64 dataset

总是能够提升模型性能。对于 Yoochoose 1/64 数

表 3 不同编码器对模型的影响

Table 3 Performance comparison of the SR-BE model with different encoders

模型	命中率/%		排名倒数均值/%	
	Yoochoose 1/64 ^[19]	Diginetica ^[20]	Yoochoose 1/64 ^[19]	Diginetica ^[20]
SR-BE-L	69.89	49.48	30.58	17.03
SR-BE-G	71.22	51.13	31.33	17.16
SR-BE	71.62	52.72	31.54	18.24

部编码器结合在一起,能够比单一地使用其中一个编码器取得更好的推荐效果。以 Diginetica 数据集^[20]为例,与 SR-BE-L 和 SR-BE-G 相比,SR-BE 模型在命中率评价指标上的性能提升分别是 6.55%和 3.11%,而在排名倒数均值评价指标上的性能提升分别是 7.11%和 6.29%。

2.9 自注意力可视化分析

为了直观体现自注意力网络在捕获物品间全局依赖时发挥的作用,本文在图 7 中展示了一些会话实例。这些实例是从 Yoochoose 1/64 数据集^[19]中随机选取的。每个长条代表一个会话,长条的左边标号为会话的 ID,长条中每一个方块代表会话中的一个物品。颜色的深度代表着物品对于全局表示 g_t 的重要程度。从图中可以得出以下结论:①对于全局表示重要的物品较多的出现在会话尾部附近。这种现象符合人们的行为模式:用户的下一个行为通常与最近的若干行为高度相关②在某些实例中,权重最大的物品出现在会话的头部或中间。这说明无论这些物品出现在会话中的何种位置,基于自注意力网络的全局编码器能够有效捕获物品间的全局依赖。

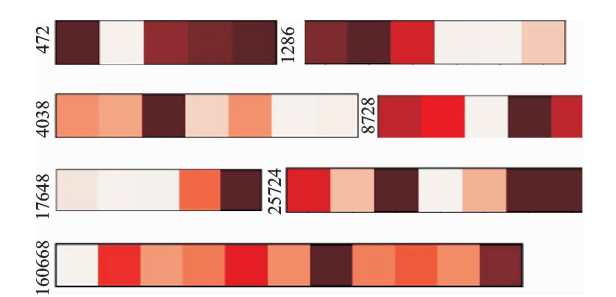


图 7 自注意力权重可视化示意图

Fig. 7 A visualization diagram of self-attention weight

3 结 论

本文提出了 SR-BE 会话型推荐模型。SR-BE 采用自注意力网络作为全局编码器来捕获所有物品间的全局依赖,将会话序列构建成会话图,采用图神

经网络作为局部编码器来捕获到邻节点间的局部依赖。最后,将通过全局编码器得到的全局隐向量和通过局部编码器得到的局部隐向量的线性组合用来更加准确的表示会话序列。

(1)基于自注意力网络的全局编码器能够有效捕获物品间的全局依赖,单独使用时较基于图神经网络的局部编码器准确度更高。

(2)将基于自注意力网络的全局编码器与基于图神经网络的局部编码器结合在一起,能够比单一地使用其中一个编码器取得更好的推荐效果。

(3)本文模型能够有效捕获物品间的全局依赖与局部依赖,提高推荐准确度。该模型在公开数据集上验证了性能,实验结果表明 SR-BE 模型性能明显高于相关对比模型。

参考文献:

[1] RENDLE S, FREUDENTHALER C, SCHMIDT T L. Factorizing personalized Markov chains for next-basket recommendation [C]// Proceedings of the 19th International World Wide Web Conference. New York, USA: ACM, 2010: 811-820.

[2] HIDASI B, KARATZOGLOU A, BALTRUNAS L, et al. Session-based recommendations with recurrent neural networks [EB/OL]. [2020-03-28]. <https://arxiv.org/abs/1511.06939>.

[3] LI Jing, REN Pengjie, CHEN Zhumin, et al. Neural attentive session-based recommendation [C]// Proceeding of the 26th ACM Conference on Information and Knowledge Management. New York, USA: ACM, 2017: 1419-1428.

[4] LIU Qiao, ZENG Yifu, MOKHOSI R, et al. Stamp: short-term attention/memory priority model for session-based recommendation [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2018: 1831-1839.

[5] REN Pengjie, CHEN Zhumin, LI Jing, et al. Repeat-Net: a repeat aware neural recommendation machine

- for session-based recommendation [C] // Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI, 2019: 4806-4813.
- [6] WANG Meirui, REN Pengjie, MEI Lei, et al. A collaborative session-based recommendation approach with parallel memory modules [C] // Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2019: 345-354.
- [7] 王雅青, 郭彩丽, 楚云霏, 等. 面向会话型推荐系统的个性化分层循环模型 [J]. 北京邮电大学学报, 2019, 42(6): 142-148.
- WANG Yaqing, GUO Caili, CHU Yunfei, et al. Personalized hierarchical recurrent model for session-based recommendation systems [J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(6): 142-148.
- [8] 曾义夫, 牟其林, 周乐, 等. 基于图表示学习的会话感知推荐模型 [J]. 计算机研究与发展, 2020, 57(3): 590-603.
- ZENG Yifu, MU Qilin, ZHOU Le, et al. Graph embedding based session perception model for next-click recommendation [J]. Journal of Computer Research and Development, 2020, 57(3): 590-603.
- [9] WU Shu, TANG Yuyuan, ZHU Yanqiao, et al. Session-based recommendation with graph neural networks [C] // Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI, 2019: 346-353.
- [10] YU Feng, ZHU Yanqiao, LIU Qiang, et al. TAGNN: target attentive graph neural networks for session-based recommendation [C] // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2020: 1921-1924.
- [11] CHEN Tianwen, WONG R C. Handling information loss of graph neural networks for session-based recommendation [C] // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2020: 1172-1180.
- [12] 赵海燕, 赵佳斌, 陈庆奎, 等. 会话推荐系统 [J]. 小型微型计算机系统, 2019, 40(9): 1869-1875.
- ZHAO Haiyan, ZHAO Jiabin, CHEN Qingkui, et al. Session-based recommendation system [J]. Journal of Chinese Computer Systems, 2019, 40(9): 1869-1875.
- [13] 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.
- HUANG Liwei, JIANG Bitao, LÜ Shouye, et al. Survey on deep learning based recommender systems [J]. Chinese Journal of Computers, 2018, 41(7): 1619-1647.
- [14] FANG Hui, GUO Guibing, ZHANG Danning, et al. Deep learning-based sequential recommender systems: concepts, algorithms, and evaluations [C] // 19th International Conference on Web Engineering. Berlin, Germany: Springer Verlag, 2019: 574-577.
- [15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] // Red Hook, NY, USA: Neural Information Processing Systems Foundation, 2017: 5999-6009.
- [16] ONDŘEJ B. The WMT 2014 translation task dataset [DB/OL]. [2020-06-10]. <http://www.statmt.org/wmt14/translation-task.html>.
- [17] XU Chengfeng, ZHAO Pengpeng, LIU Yanchi, et al. Graph contextualized self-attention network for session-based recommendation [C] // 28th International Joint Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2019: 3940-3946.
- [18] LI Yujia, TARLOW D, BROCKSCHMIDT M, et al. Gated graph sequence neural networks [EB/OL]. [2020-04-11]. <https://arxiv.org/abs/1511.05493>.
- [19] BEN-SHIMON D. The yoochoose dataset [DB/OL]. [2020-06-01]. <https://2015.recsyschallenge.com/challenge.html>.
- [20] SPIRIN N. The diginetica dataset [DB/OL]. [2020-06-01]. <https://competitions.codalab.org/competitions/11161>.
- [21] TAN Y K, XU Xinxing, LIU Yong. Improved recurrent neural networks for session-based recommendations [C] // Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. New York, USA: ACM, 2016: 17-22.
- [22] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms [C] // Proceedings of the 10th International World Wide Web Conference. New York, USA: ACM, 2001: 285-295.
- [23] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback [C] // Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Arlington, VA, USA: AUAI Press, 2009: 452-461.

(编辑 刘杨)