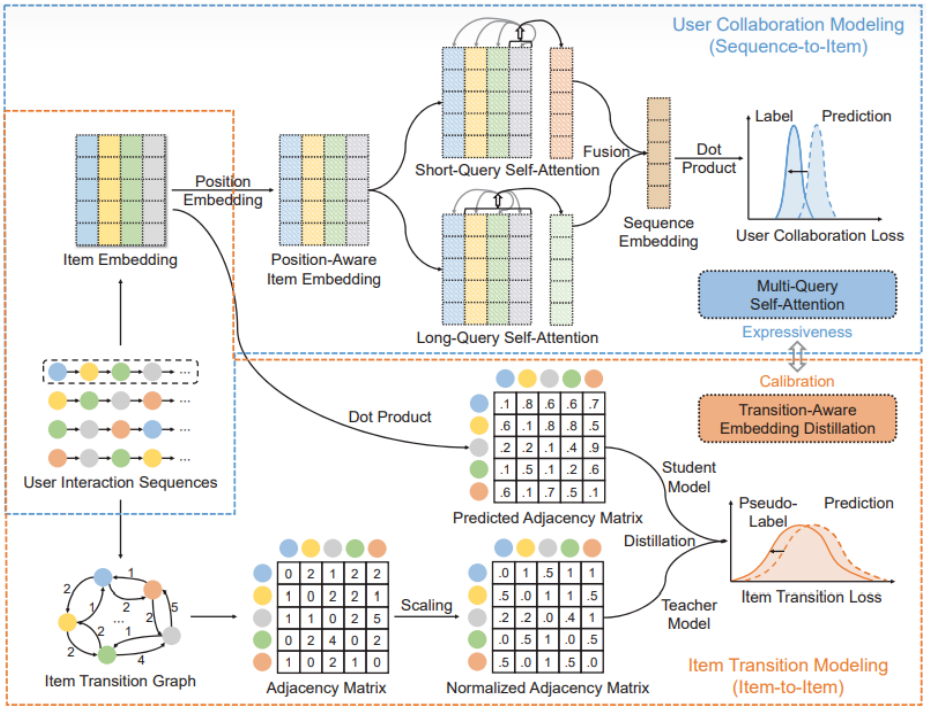


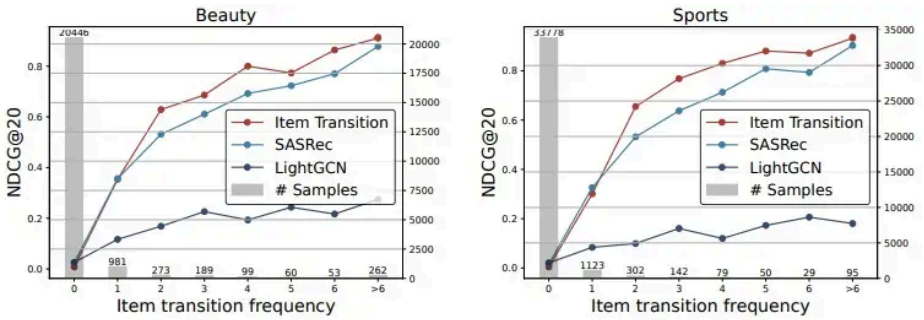
快手2023技术解析：巧妙运用Multi-Query Self-Attention实现序列推荐的高效转化

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

原文《Collaboration and Transition: Distilling Item Transitions into Multi-Query Self-Attention for Sequential Recommendation》

Introduction

在序列推荐中，协同信号和过渡信号被用来识别用户的行为模式并推荐相关内容。SASRec是一种先进的序列推荐方法，但在捕获这两种信号方面的局限性也被揭示。在实验中，对比了SASRec与两个基准方法------Item Transition和LightGCN。Item Transition是一种基于全局转换频率的记忆型非个性化方法，而LightGCN则是一种最先进的非序列推荐方法，通过线性传播在用户-物品交互图上学习用户和内容嵌入。



Dataset	Trans. Freq.	Item Trans.	SASRec	LightGCN	# Samples
Beauty	0	0.0081	0.0243	0.0261	20,446
	>0	0.5520	0.5180	0.1672	1,917
	All	0.0547	0.0666	0.0382	22,363
Sports	0	0.0045	0.0161	0.0209	33,778
	>0	0.4772	0.4523	0.1034	1,820
	All	0.0287	0.0384	0.0251	35,598

知乎

SASRec对测试样本的通用能力有限。这是由于SASRec使用最新的内容嵌入作为自我注意力模块的查询，这是一种固有的限制，使其无法充分利用协同信号。在第二部分中，SASRec的方法在处理跨时序内容转移方面存在局限性。

为了解决这个问题，我们提出了一个名为MQSA-TED的新方法。MQSA-TED由两个主要组件组成：自注意力和嵌入蒸馏。首先，我们提出了一种具有灵活大小窗口的L-查询自注意力模块，它可以在滑动窗口内捕获多内容的协同信息。然后，我们开发了一个过渡敏感的嵌入蒸馏模块，它可以将全局的内容对内容过渡模式转化为嵌入，以便模型更好地记忆和利用过渡信号进行推荐。最后，我们的方法通过双监督实现了用户协同建模和内容转换建模的内在解耦，从而使得模型能够更准确地捕捉到跨时序内容的转换关系。

Problem Formulation

序列推荐任务旨在预测用户接下来可能喜欢的物品，假设用户集合为$\mathcal{U}$，物品集合为$\mathcal{I}$，用户交互序列为$s^{(u)}$，其中

$$s^{(u)} = [i_1^{(u)}, i_2^{(u)}, \cdots, i_{n_u}^{(u)}]$$

其中n_u代表序列长度。问题可以表示为：给出用户历史交互记录，计算下一个要被互动的物品的可能性。

$$p\left(i_{n_u+1}^{(u)} | S^{(u)}\right)$$

最后，根据排序后的概率，为用户u推荐前N个物品。

SASRec

$$[\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \cdots, \hat{\mathbf{e}}_n] = \mathbf{Transformer}([\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \cdots, \hat{\mathbf{e}}_n])$$

这个模型采用多层自注意力机制和全连接网络，用于捕捉序列中的长期依赖关系，具体实现形式如下：

$$\mathbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$

$$\mathbf{Q} = \hat{\mathbf{E}}\mathbf{W}^Q,\ \mathbf{K} = \hat{\mathbf{E}}\mathbf{W}^K,\ \mathbf{V} = \hat{\mathbf{E}}\mathbf{W}^V$$

首先，$\mathbf{Q}$、$\mathbf{K}$、$\mathbf{V}$分别表示查询向量、键向量和值向量$\mathbf{W}^Q$、$\mathbf{W}^K$、$\mathbf{W}^V$是查询、键和值对应的投影矩阵。然后，通过对序列嵌入与候选内容嵌入的点积来预测排名得分

$$\hat{\mathbf{r}}_t = \tilde{\mathbf{e}}_t\mathbf{E}^T$$

使用累积交叉熵损失进行模型训练，公式为

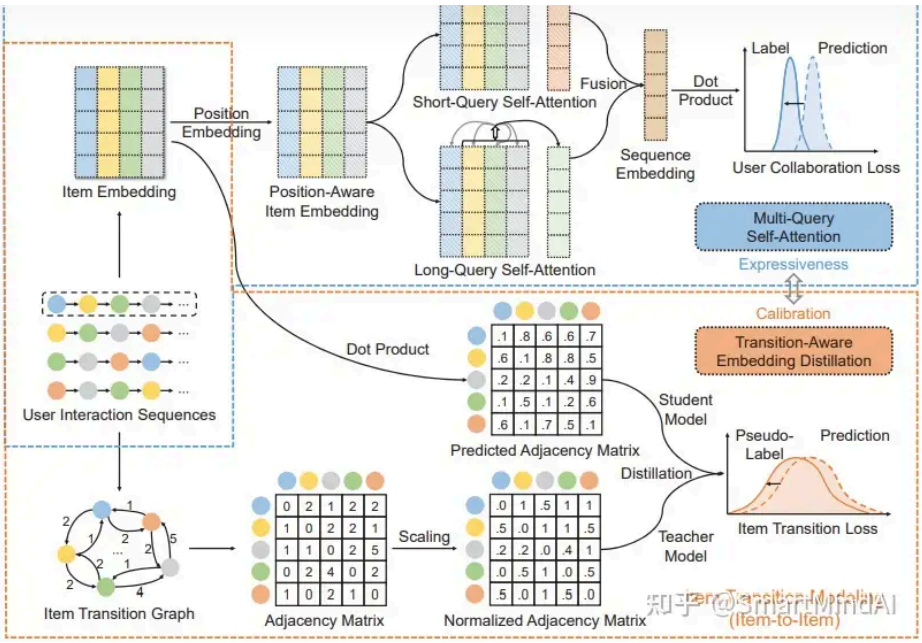
$$\mathcal{L} = -\sum_{t=1}^{n} \log\left(\frac{\exp(\hat{\mathbf{r}}_t)}{\sum_{j=1}^{N}\exp(\hat{\mathbf{r}}_j)}\right)$$

其中N为候选内容的数量。

$$\mathcal{L}_{rec} = -\sum_{t=1}^{n} \mathbf{r}_t \log \mathbf{softmax}(\hat{\mathbf{r}}_t)$$

将$\mathbf{r}_t$视作一维向量，每个元素均为0，除了位置t处为1，其他所有位置均为0。

Methodology



1）对于多查询协同建模，可以使用多查询自注意力机制。

2）为了更好地理解内容的状态，可以使用状态✦感知嵌入传递技术。

Multi-Query Self-Attention for User Collaboration Modeling

$$\mathbf{q}_t = \hat{\mathbf{e}}_t \mathbf{W}^Q$$

在时间戳t上，给定历史内容的时间戳$i_1, i_2, \cdots, i_t$，通过计算查询嵌入与历史内容的嵌入之间的归一化✦点积来确定注意力权重，注意力权重被单个在时间戳t上的内容所主导，形成短查询自我注意力。但是，这种短查询自我注意力模型在处理协作信息方面存在局限性，特别是在时间和用户的偏好不一致的情况下。

在SASRec模型中，推荐结果会受到用户交互过的内容序列顺序改变的影响，比如将最后两个内容的位置交换。这意味着SASRec可能无法很好地处理没有观察到的内容过渡的测试样例。而在现实世界的推荐场景中，用户的兴趣相对稳定且对最近几个选择的顺序不太敏感，这对于SASRec来说是一个挑战。 为解决这个问题，我们提出了L查询自注意力✦方法。

L-查询自注意力是一种自注意力模块，它使用最近L个时间戳内容的嵌入或其变换表示作为注意力查询。

在这里，我们使用最近L个时间戳t内容的嵌入的简单均值池化作为查询嵌入。

$$\tilde{\mathbf{q}}_t = \text{mean-pooling}(\hat{\mathbf{e}}_{t-L+1}, \hat{\mathbf{e}}_{t-L+2}, \cdots, \hat{\mathbf{e}}_t)\tilde{\mathbf{W}}^Q$$

其中，L是超参数✦，控制了关注查询的范围。也可以使用其他函数来生成查询嵌入，例如带有时间衰减的加权求和。注意L控制了自我注意力的历史上下文范围。使用大值L意味着模型依赖于长距离历史内容来表示用户兴趣，这有助于捕获协作信号，但也可能导致累积偏见，因为用户的兴趣可能会随着时间的推移而变化。

相反，使用小值L意味着模型采用最新的交互过的内容来表示用户兴趣，但可能会引入方差，因为使用的内容数量较少。为了平衡偏差-方差权衡，我们提出了一种名为Multi Query Self Attention (MQSA)的方法，该方法结合了Short Query Self Attention (SASRec)（$L = 1$，类似于SASRec）与Long Query Self Attention（L较大）使用一个超参数α。

$$\tilde{\mathbf{e}}_t = \alpha \cdot \tilde{\mathbf{e}}_t^{short} + (1 - \alpha) \cdot \tilde{\mathbf{e}}_t^{long}$$

然后，通过点积将序列嵌入$\tilde{\mathbf{e}}_t$和候选内容的嵌入用于预测它们的排名得分。值得注意的是，我们也允许模型学习最优的α。但是，同时学习权重和嵌入是具有挑战性的，因为其内在复杂性。我们还可以纳入更多的Ls。

组合。与Fossil 使用所有交互过的内容不同，MQSA引入了最近L个时间戳内容的灵活窗口大小，以控制偏差-方差权衡。此外，MQSA使用自我注意力模块增强表达力，导致性能优于在Fossil中使用纯内容嵌入。

Transition-Aware Embedding Distillation for Item Transition Modeling

序列推荐模型通过捕捉长期用户兴趣来提高推荐准确性，但可能无法充分利用全局内容到内容的过渡信号。特别是大多数现有的方法遵循自回归框架。对于每个用户，他们的偏好在时间戳t根据他们之间至t为止的所有交互的内容被学习，然后用于预测时间戳$t+1$的内容。然而，这个框架未能使模型能够学习全局内容到内容的模式。换句话说，未与用户互动的内容被视为平等的，而没有考虑当前内容i_t可能导致的潜在内容。

为了解决这一限制，我们提出了一种基于内容过渡的启发式推荐器，然后开发了一个知识蒸馏方法，将这些全局内容过渡模式集成到序列模型中。具体来说，我们构造了一个全局内容过渡图

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

其中$\mathcal{V}$表示内容节点$\mathcal{E}$表示内容之间的过渡边。$\mathcal{G}$是一个加权和有向图[+]，其中每条边的权重代表在一个时间间隔内的两物品之间的过渡频率，基于所有用户的交互序列。请注意，时间间隔超参数k被用来控制长期的内容过渡模式，并默认设置为1（即仅考虑直接相邻的物品之间的过渡）。我们使用邻接矩阵[+]

$$\mathbf{A} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$$

其中$a_{i,j}$是内容i到内容j的过渡频率，如图所示。它是一种基于记忆的非个性化方法，根据从当前内容到候选内容的过渡频率推荐内容。首先，我们使用行归一化方法对转换频率进行标准化，得到

$$\bar{a}_{i,j} = \frac{a_{i,j}}{\max_j a_{i,j}}$$

然后，我们使用温度τ的softmax函数生成知识蒸馏的伪标签：

$$\bar{a}_{i,j} = \frac{a_{i,j}}{\max_j a_{i,j}}$$

接着，我们使用温度τ的softmax函数生成知识蒸馏的伪标签。

$$\tilde{\mathbf{a}}_i = \mathbf{softmax}(\bar{\mathbf{a}}_i / \tau)$$

τ值越高，生成的内容过渡概率分布越软。我们的学生模型采用了一个简单的因子分解模型，在自注意力层之前使用向量$\mathbf{e}_i$与嵌入矩阵$\mathbf{E}$的点积来预测内容i的内容过渡分布。此外，我们还采用了dropout策略以增强学习。最后，我们使用温度τ的softmax函数来获得预测的内容过渡概率。

$$\hat{\mathbf{a}}_i = \mathbf{softmax}(\mathbf{e}_i \mathbf{E}^T / \tau)$$

我们将内容过渡融入到序贯模型中，通过比较预测和伪标签内容过渡概率来使用交叉熵损失法。

$$\mathcal{L}_{kd} = -\sum_{i \in \mathcal{I}} \tilde{\mathbf{a}}_i \log \hat{\mathbf{a}}_i$$

因子分解[+]模型可以通过学习内容过渡模型来让内容嵌入记住内容过渡模式。整个模型的整体损失函数[+]为:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{kd} \mathcal{L}_{kd} + \lambda_{\Theta} ||\Theta||_2^2$$

在这个模型中，参数Θ、控制蒸馏权重的λ_{kd}和控制l_2正则化的λ_{Θ}都是重要的。

Discussion

Relationship Between Two Modules

这里我们讨论了用户协作与项转换模块之间的关系，并探讨了它们是如何相互补充的。

，采用自注意力来捕捉长期用户的偏好，并根据历史项选择最有可能的下一个项，从而具有更强的泛化能力，但对项到项的过渡模式的记忆和利用能力有限。因此，用户协作模型需要项转换模型来校准其预测。

解耦学习。 用户协作与项转换模块本质上是解耦的，因为我们使用双监督，其中原始项嵌入捕获项到项的转换信号，而经过自注意力的项嵌入捕获序列到项的协同信号。

检索与重排名。 项转换和用户协作模块可以被视为检索模型和重排名模型，分别提供生成潜在候选项的见解以及基于各自的交互历史为用户提供最相关项的见解。

Comparison with Existing Methods

本文提出了一种名为过渡感知嵌入蒸馏（TED）的新型校准器，该校准器将基于内容过渡图的概念应用于实际的推荐系统⁺设计。 **图形正则化（GraReg）** 是一个使用欧几里得距离⁺为基础的在嵌入层上使用的k最近邻（k-NN）图形上的正则化术语：

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{reg} \sum_{(i,j)\in\mathcal{E}} ||\mathbf{e}_i - \mathbf{e}_j||^2$$

其中λ_{reg}是正则化系数⁺超参数$\mathcal{E}$是k-NN图中的边。我们可以在这里将过渡频率作为边的权重。因此，GraReg使用最相关的前k个物品进行正则化⁺，从而学习局部过渡模式。此外，GraReg引入了一种对齐损失，但缺乏一种一致性损失，即相关项应彼此接近，而无关项应彼此分离。相比之下，TED使用全局内容过渡作为教师模型，使内容嵌入能够记住并利用过渡信号。

Transition-aware Embedding Smoothing (GES) 基于全局内容过渡图对序列推荐系统中嵌入进行平滑：

$$\mathbf{E}^{(l+1)} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{E}^{(l)}$$

是内容过渡图的邻接矩阵，其中包含自环$\tilde{\mathbf{D}}$是$\tilde{\mathbf{A}}$的度矩阵l是图卷积层的数量。但是，堆叠多个图卷积层可能会导致过度平滑问题，可能降低模型性能。相比之下，TED引入了一个超参数来控制内容过渡蒸馏的力量，从而使模型能够在不同的推荐场景中具有灵活性。

Model Complexity

在这里，我们分析了提出的模型的空间和时间复杂性。在SASRec中，可学习的参数包括内容嵌入、位置嵌入、自我注意力层、前馈网络和层归一化。这个模型中的总参数数量为

$$\mathcal{O}(|\mathcal{I}|d + nd + d^2)$$

我们的提出模型引入了长查询自注意力，它添加了$\mathcal{O}(d^2)$用于投影矩阵、前馈网络和层归一化的元素。特征蒸馏模块没有增加任何额外参数。因此，我们的提出模型与SASRec的空间复杂性相同。在SASRec中，自注意力层和前馈层的计算复杂性⁺为$\mathcal{O}(n^2d + nd^2)$。累计交叉熵损失的计算复杂性为

$$\mathcal{O}(|\mathcal{I}|nd)$$

因此，SASRec的总计算复杂性为

$$\mathcal{O}(|\mathcal{I}|nd + n^2d + nd^2)$$

在我们的提出模型中，自注意力模块与SASRec中的自注意力模块具有相同的复杂性。特征蒸馏模块的计算复杂性为

$$\mathcal{O}(|\mathcal{I}|nd)$$

因此，提出模型的时间复杂性与SASRec相同，但增加了累计交叉熵损失的计算复杂性。

Experimental Settings

知乎

我们在实验中使用了来自亚马逊评论数据集（cseweb.ucsd.edu/\~jmcauley/datasets.html）中的四个数据集：美丽、运动、玩具以及Yelp。每个数据集都遵循"留一评估"协议，其中第一个内容作为测试数据，第二个最后一个内容作为验证数据，其余内容作为每个用户的训练数据。具体的数据集统计信息请参考表 。

Dataset	# Users	# Items	# Actions	Density	Avg. Len.
Beauty	22,363	12,101	198,502	0.073%	8.88
Sports	25,598	18,357	296,337	0.063%	8.32
Toys	19,412	11,924	167,597	0.072%	8.63
Yelp	30,431	20,033	316,354	0.052%	10.40

Baselines

我们在序列推荐方面将提出的方法与其他几种最先进的基准方法进行比较：

- 非个性化方法：这种方法是基于物品流行度排名来进行推荐的。
- 图卷积网络（GCN）方法：这种方法是基于图卷积神经网络（GCN）学习用户和物品嵌入的线性传播方法。
- 马尔科夫链（MC）方法：这种方法是结合矩阵分解和因子化马尔科夫链算法。
- 卷积神经网络（CNN）方法：这种方法是使用水平和垂直卷积来学习序列模式的算法。
- 单向$Transformer$方法：这种方法是使用自注意力模块在Transformer中模型用户兴趣的方法。
- 双向$Transformer$方法：这种方法是使用Bert中的自注意力模块模型用户兴趣的方法。
- 多层感知机（MLP）方法：这种方法是当前基于滤波增强多层感知机的最先进序列推荐模型。

Evaluation Metrics and Protocols

我们采用了HR\@N和NDCG\@N来评估比较方法在序列推荐任务上的性能，并且根据N=5、10和20设置默认参数，同时对每个用户对其训练或验证数据中所有非正内容的所有内容进行排名，并且为确保结果的稳健性，我们进行了五次随机初始化每个模型的操作，并报告平均性能。

Implementation and Hyperparameter Settings

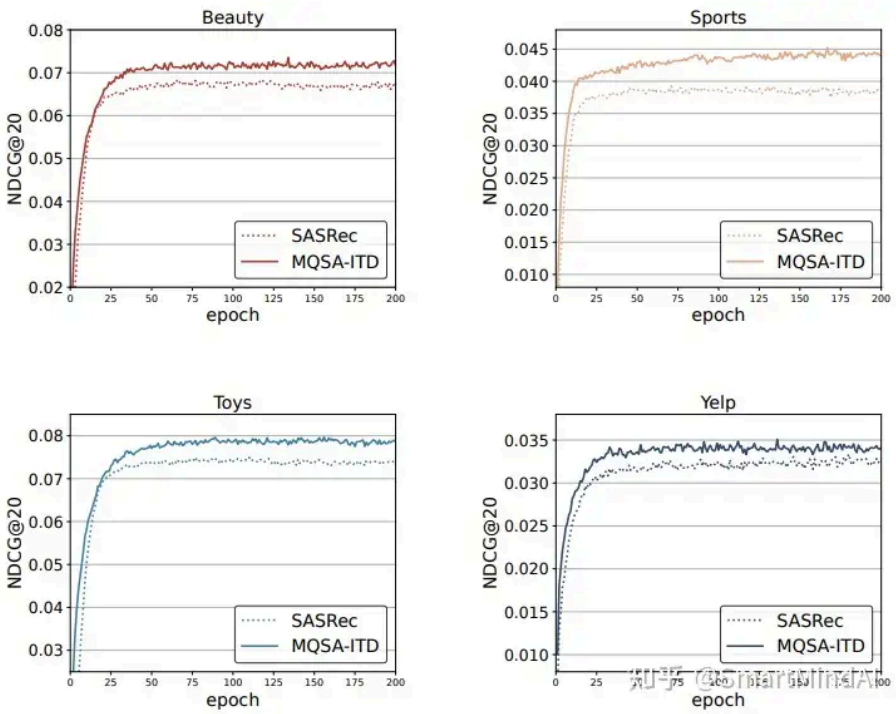
我们使用TensorFlow实现并比较模型，并使用交叉熵损失进行优化。模型参数包括最大序列长度50、嵌入大小64、学习率为5e-3-1e-4，L2正则化为0-1e-6。我们使用小批量Adam进行训练，批次大小为256，并将不同模型的其他参数根据论文建议在验证集中进行调整。。

Main Results (RQ1)

Dataset	Metric	POP	LightGCN	FPMC	Caser	GRU4Rec	SASRec	BERT4Rec	FMLP-Rec	MQSA-TED	Improv.
Beauty	HR@5	0.0077	0.0374	0.0596	0.0359	0.0489	0.0694	0.0419	0.0698	**0.0752***	7.23%
	NDCG@5	0.0042	0.0247	0.0419	0.0241	0.0342	0.0492	0.0275	0.0488	**0.0534***	8.58%
	HR@10	0.0135	0.0571	0.0838	0.0511	0.0695	0.0932	0.0647	0.0995	**0.1039***	4.44%
	NDCG@10	0.0061	0.0311	0.0497	0.0290	0.0408	0.0568	0.0349	0.0583	**0.0627***	7.48%
	HR@20	0.0217	0.0841	0.1151	0.0720	0.0998	0.1286	0.0992	0.1361	**0.1435***	5.40%
	NDCG@20	0.0081	0.0379	0.0576	0.0343	0.0484	0.0657	0.0435	0.0675	**0.0726***	7.62%
Sports	HR@5	0.0057	0.0252	0.0337	0.0195	0.0221	0.0380	0.0241	0.0415	**0.0455***	9.52%
	NDCG@5	0.0041	0.0170	0.0234	0.0128	0.0143	0.0267	0.0161	0.0287	**0.0320***	11.34%
	HR@10	0.0091	0.0384	0.0499	0.0290	0.0357	0.0541	0.0380	0.0598	**0.0643***	7.48%
	NDCG@10	0.0052	0.0212	0.0286	0.0159	0.0187	0.0318	0.0206	0.0346	**0.0380***	9.85%
	HR@20	0.0175	0.0576	0.0703	0.0431	0.0548	0.0752	0.0583	0.0847	**0.0906***	6.93%
	NDCG@20	0.0073	0.0260	0.0337	0.0195	0.0235	0.0371	0.0257	0.0409	**0.0446***	9.09%
Toys	HR@5	0.0065	0.0378	0.0664	0.0307	0.0420	0.0736	0.0379	0.0785	**0.0834***	6.24%
	NDCG@5	0.0044	0.0251	0.0463	0.0224	0.0297	0.0533	0.0244	0.0570	**0.0600***	5.31%
	HR@10	0.0090	0.0564	0.0925	0.0420	0.0597	0.0989	0.0589	0.1062	**0.1130***	6.42%
	NDCG@10	0.0052	0.0311	0.0547	0.0260	0.0354	0.0615	0.0312	0.0659	**0.0696***	5.56%
	HR@20	0.0143	0.0795	0.1212	0.0597	0.0834	0.1299	0.0857	0.1399	**0.1503***	7.41%
	NDCG@20	0.0065	0.0370	0.0619	0.0305	0.0414	0.0693	0.0379	0.0743	**0.0789***	6.23%
Yelp	HR@5	0.0056	0.0290	0.0272	0.0199	0.0211	0.0232	0.0264	0.0270	**0.0320***	10.18%
	NDCG@5	0.0036	0.0184	0.0173	0.0129	0.0134	0.0151	0.0169	0.0169	**0.0205***	11.74%
	HR@10	0.0096	0.0486	0.0433	0.0334	0.0367	0.0379	0.0441	0.0446	**0.0517***	6.36%
	NDCG@10	0.0049	0.0246	0.0224	0.0172	0.0184	0.0198	0.0226	0.0225	**0.0269***	8.95%
	HR@20	0.0158	0.0790	0.0695	0.0535	0.0603	0.0623	0.0737	0.0721	**0.0832***	5.26%
	NDCG@20	0.0065	0.0323	0.0290	0.0222	0.0244	0.0259	0.0300	0.0294	**0.0348***	7.62%

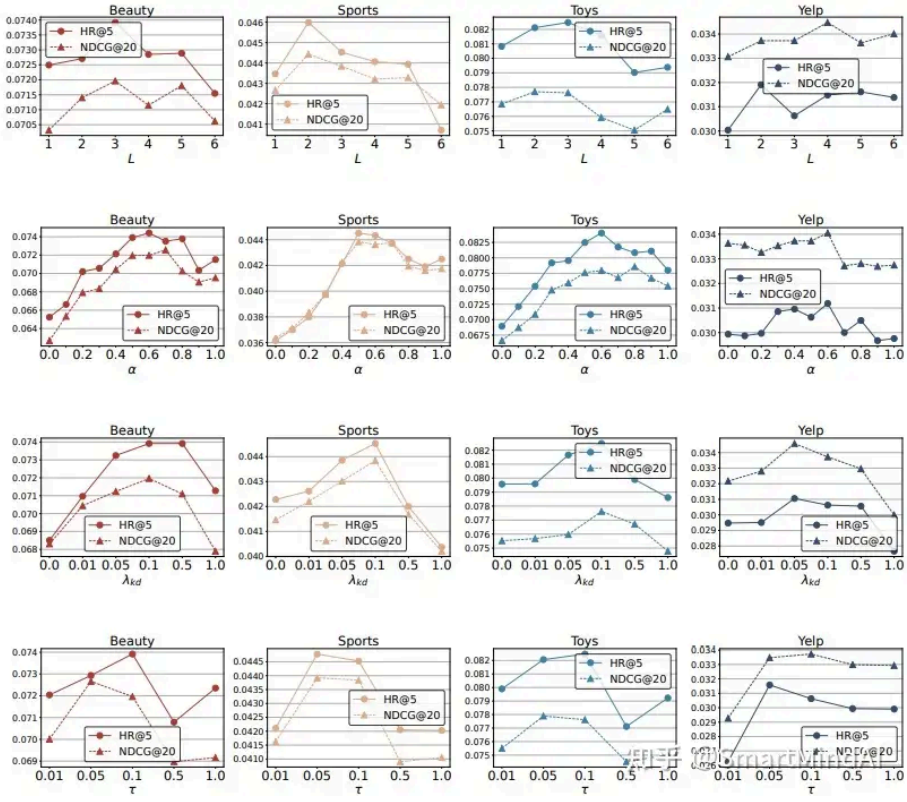
知乎

Yelp数据集上，由于Yelp上的用户交互序列较弱，LightGCN超过了顺序方法。此外，我们提出的方案显著优于所有基线方法，比最好的基线方法平均提高了6.24%在Hit Ratio\@20和7.64%在NDCG\@20。图2显示了SASRec和我们提出的方案与训练epoch的关系。可以观察到我们的提出的方法始终优于SASRec一个明显的优势，显示出提出模块的有效性。



Hyperparameter and Ablation Studies (RQ2)

图2展示了我们提出的方案以及它与其他超参数和模块的关系。



Length of Long-Query Self-Attention L.

我们可以发现最佳的L取决于数据集，并且模型在$[2, 4]$的L范围内表现最佳，这表明长查询自注意力的有效性在于捕捉协作信号。

知乎

实验结果显示，当α约为0.5时，模型表现最佳，这意味着模型在平衡偏差和方差方面取得了适当的效果。另外，当α等于1时，模型退化为SASRec的TED版本。因此，我们提出的大规模查询自注意力明显优于SASRec所采用的小规模查询自注意力，并具有合适的α。

Weight of Embedding Distillation λ_{kd}.

通过观察，可以发现当λ_{kd}约为0.1时，模型表现最好，这表明了TED模块的有效性。值得注意的是，当$\lambda_{kd} = 0$时，我们提出的模型退化为没有TED的MQSA模型，这会导致性能大幅度降低。

Temperature of Embedding Distillation τ.

实验结果显示，为了有效实现知识蒸馏，模型需要使用相对较难的伪标签来表示内容转移分布，这是因为当τ取值为0.05或0.1时，可以获得最优性能。

发布于 2024-03-18 11:37 · IP 属地北京

快手 序列推荐 Transformer