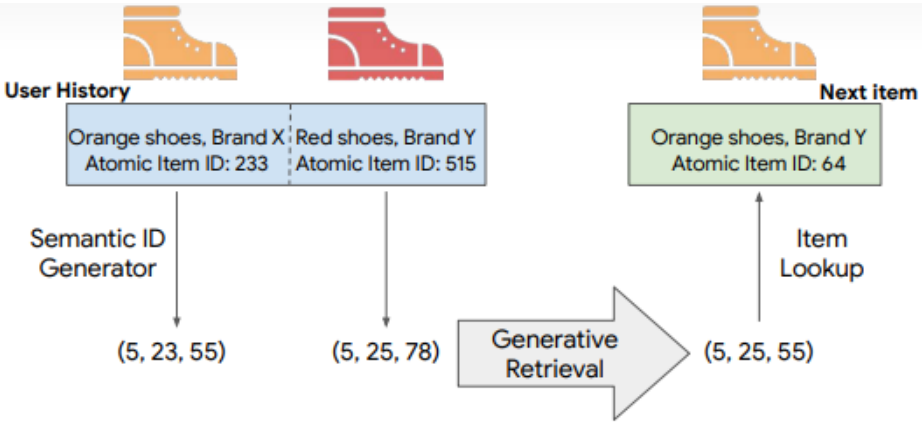


谷歌-2023：TIGER-基于生成式召回模型的推荐系统

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

Introduction

推荐系统用于帮助用户发现兴趣内容并广泛应用于各个领域，如视频、应用程序、产品和音乐。目前的推荐系统采用召回和排序策略，首先从候选集中选出若干相关性较高的内容，然后采用排序模型对这些候选集进行排序。然而，考虑到排序模型只能处理接收的候选集，因此检索阶段需要输出高度相关的候选集。

对于构建召回模型。其中，矩阵分解法可学习查询和候选内容在同一空间中的嵌入；而双塔架构则是近年来越来越受欢迎的一种方法，它将查询和候选内容的嵌入映射到同一空间，并通过内积将查询和候选内容的嵌入连接起来。为了让这些模型能够在推理期间使用，需要创建一个存储所有内容嵌入的索引，并使用候选内容的塔来获取给定查询的嵌入。随后，通过使用近似最近邻（ANN）算法对这些嵌入进行检索，即可找到相关性强的候选集。

随着研究的发展，双塔架构还被扩展到了考虑用户与物品交互顺序的序列式推荐。这有助于进一步提高推荐质量。我们提出了一个新的生成式召回模型构建方法，用于序列推荐。与传统的方法不同，我们的方法通过端到端的生成模型直接预测候选ID。我们建议将Transformer的记忆（参数）用作推荐系统的召回索引，类似于Tay等人使用Transformer内存进行文档检索。我们将其称为"Transformer Index for Generative Recommenders"（TIGER）。

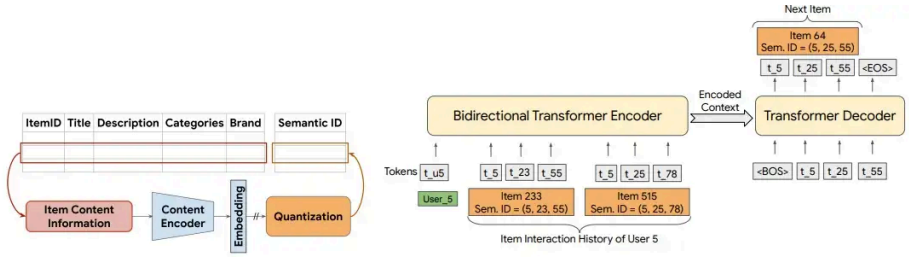
图2展示了TIGER的概览。TIGER的独特之处在于它对物品的新颖语义表示，"Semantic ID"，即每个物品的文本特征提取出的一组有序的标记或代码词，当做特征。具体来说，对于一个物品的文本特征，我们使用预训练的文本编码器（如SentenceT5）生成密集内容嵌入。然后应用量化方案对物品的嵌入进行编码以形成一组顺序的标记/代码词，我们将其称为该物品的"Semantic ID"。

最后，这些Semantic IDs被用于在序列推荐任务上训练Transformer模型。本工作提出了一种基于语义token表示的推荐系统。通过训练Transformer模型，可以实现不同内容之间的知识共享，并且不需要像传统推荐系统那样使用原子和随机内容ID作为 内容特征。另外，这种表示方式还能够帮助模型避免陷入反馈循环，从而泛化到新的 内容中。此外，使用语义token表示还能有效地解决内容语料库规模较大的问题。最后，本工作还证明了使用这种表示方式替代传统的随机哈希技术在计算效率和准确性方面具有明显的优势。

1. 提出了TIGER，一种基于生成检索的推荐框架。该框架为每个 内容分配语义标识符，并使用检索模型预测用户可能参与的内容的语义标识符。
2. 证明了TIGER在多个数据集上的表现优于当前的最先进的推荐系统，这由召回率和NDCG度量评估。
3. 本文提出了一种新的生成式检索模型，其主要优势在于能够实现两种能力：一是提高冷启动推荐效果，二是生成多样化的推荐。

Proposed Framework

1. 将内容特征编码为嵌入向量，然后将嵌入向量量化为语义代码字元组，得到 内容语义ID.
2. 训练Transformer模型在Semantic IDs上进行推荐系统生成。



(a) Semantic ID generation for items using quantization of content embeddings.

(b) Transformer based encoder-decoder setup for building the sequence-to-sequence model used for generative retrieval.

Figure 2: An overview of the modeling approach used in TIGER.

Semantic ID Generation

相似内容有重叠的语义ID，具体而言，具有ID $(10, 21, 35)$ 的内容与具有 ID $(10, 21, 40)$ 的内容更相似。

下面将讨论生成语义ID的量化方案。

使用残差量化自编码器⁺（RQ-VAE）结合多级矢量量化⁺器来生成语义ID。此方法通过更新编码词典和DNN编码解码⁺参数进行联合训练。具体步骤如下

$$\hat{\boldsymbol{z}} = \sum_{d=0}^{m-1} \boldsymbol{e}_{c_i}$$

将$\hat{\boldsymbol{z}}$传给解码器，解码器通过$\hat{\boldsymbol{z}}$尝试恢复原始输入$\boldsymbol{x}$。

RQ-VAE损失由$\mathcal{L}(\boldsymbol{x})$定义：

$$\mathcal{L}_{\text{recon}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2$$

以及

$$\mathcal{L}_{\text{rqvae}} = \sum_{d=0}^{m-1} \|\text{sg}[\boldsymbol{r}_i] - \boldsymbol{e}_{c_i}\|^2 + \beta\|\boldsymbol{r}_i - \text{sg}[\boldsymbol{e}_{c_i}]\|^2$$

$\hat{\boldsymbol{x}}$是解码器的输出，sg是停止梯度的操作。此损失函数⁺用于同时训练编码器、解码器和代码表。

初始化：使用k-means聚类算法在第一批次训练中，并使用中心点作为初始化。

其他方案：其他量化替代方案包括使用局部敏感哈希(LSH)或层次化的k-means聚类。然而，LSH失去了不同簇之间的语义意义，而VQ-VAE则失去了ID的层次结构，但在召回期间生成候内容的性能与RQ-VAE相似。在第3部分中，讨论了这两种方法的优点和缺点。

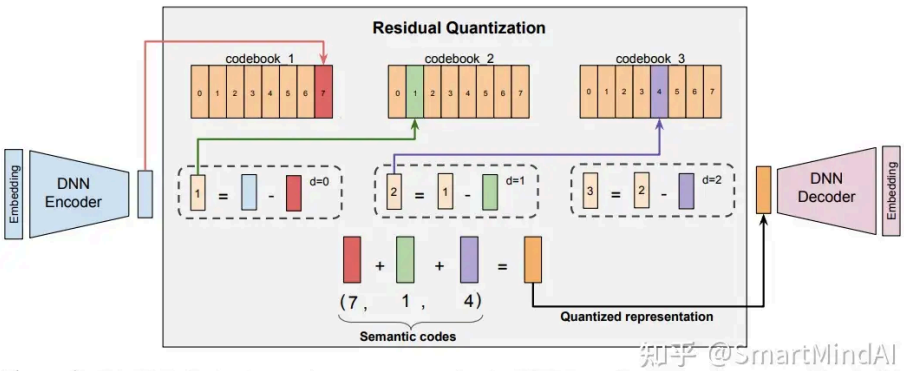


Figure 3: RQ-VAE: In the figure, the vector output by the DNN Encoder, say $\boldsymbol{r}_0$ (represented by the blue

映射到相应的 内容。这种方法比使用高维嵌入更有效率，因为查找表在存储方面效率更高。

Generative Retrieval with Semantic IDs

为每个用户构建序列，方法是按时间顺序对他们已交互过的项进行排序。然后，在形式为$(\text{item}_1, \ldots, \text{item}_n)$的序列中，推荐系统的任务是预测下一个项$\text{item}_{n+1}$。我们提出了一种生成式的方法，直接预测下一个项的语义ID。让$(c_{i,0}, \ldots, c_{i,m-1})$是对$\text{item}_i$的$m$长度的语义ID。然后我们将序列转换为

$$(c_{1,0}, \ldots, c_{1,m-1}, c_{2,0}, \ldots, c_{2,m-1}, \ldots, c_{n,0}, \ldots, c_{n,m-1})$$

然后，该序列到序列模型被训练来预测item_{n+1}的语义ID，即

$$(c_{n+1,0}, \ldots, c_{n+1,m-1})$$

由于我们的框架具有生成式，因此生成的语义ID可能与推荐内容不匹配。

Experiments

使用 Amazon Product Reviews 数据集中的用户评论和商品元数据来完成序列推荐任务。评估方法包括Recall\@K和NDCG\@K，分别设定了K值*为5和10。

我们使用了4个含有6个注意力头、每个头维度为64的每层Transformer模型，并且所有的层都应用了ReLU激活函数*。MLP和输入维度分别为1024和128。我们设置了0.1的Dropout，并总共设置了约1300万个参数。 我们对 Beauty 和 Sports and Outdoors 数据集进行了20万步的训练，而 Toys and Games 数据集只有10万步。我们使用批大小为256，初始学习率为0.01，在前10k步内进行指数根衰减。

Performance on Sequential Recommendation

GRU4Rec, Caser, HGN, SASRec, BERT4Rec, FDSA, S^3-Rec和P5都是推荐系统的基线方法。它们都使用双塔器来学习高维向量空间，编码用户的过去交互项和候选项。所有方法除了P5外，都使用多实例学习（MIPS）技术来检索用户可能的下一个候选项。我们的新颖框架则直接预测每个 内容的语义ID，使用了序列到序列（Seq2Seq）模型。

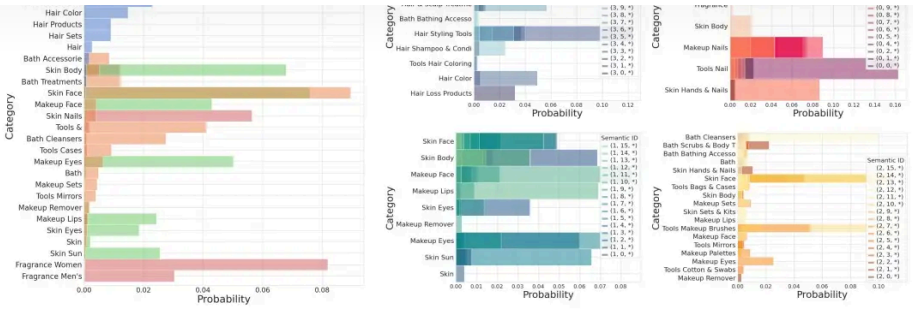
我们发现TIGER始终优于现有基准。我们观测到所有考虑的三个基准的重大改进。特别是在Beauty基准上，TIGER比排序第二的基准有了高达29%的NDCG\@5提高，并且比SASRec提高了17.3%的召回\@5。在Toys和Games数据集上，TIGER在NDCG\@5和NDCG\@10上的表现分别比其他基准好21%和15%。

Table 1: Performance comparison on sequential recommendation. The last row depicts % improvement with TIGER relative to the best baseline. Bold (underline) are used to denote the best (second-best) metric.

Methods	Sports and Outdoors				Beauty				Toys and Games			
	Recall @5	NDCG @5	Recall @10	NDCG @10	Recall @5	NDCG @5	Recall @10	NDCG @10	Recall @5	NDCG @5	Recall @10	NDCG @10
P5 [8]	0.0061	0.0041	0.0095	0.0052	0.0163	0.0107	0.0254	0.0136	0.0070	0.0050	0.0121	0.0066
Caser [33]	0.0116	0.0072	0.0194	0.0097	0.0205	0.0131	0.0347	0.0176	0.0166	0.0107	0.0270	0.0141
HGN [25]	0.0189	0.0120	0.0313	0.0159	0.0325	0.0206	0.0512	0.0266	0.0321	0.0221	0.0497	0.0277
GRU4Rec [11]	0.0129	0.0086	0.0204	0.0110	0.0164	0.0099	0.0283	0.0137	0.0097	0.0059	0.0176	0.0084
BERT4Rec [32]	0.0115	0.0075	0.0191	0.0099	0.0203	0.0124	0.0347	0.0170	0.0116	0.0071	0.0203	0.0099
FDSA [42]	0.0182	0.0122	0.0288	0.0156	0.0267	0.0163	0.0407	0.0208	0.0228	0.0140	0.0381	0.0189
SASRec [17]	0.0233	0.0154	0.0350	0.0192	0.0387	0.0249	0.0605	0.0318	0.0463	0.0306	0.0675	0.0374
S^3-Rec [44]	0.0251	0.0161	0.0385	0.0204	0.0387	0.0244	0.0647	0.0327	0.0443	0.0294	0.0700	0.0376
TIGER [Ours]	**0.0264**	**0.0181**	**0.0400**	**0.0225**	**0.0454**	**0.0321**	**0.0648**	**0.0384**	**0.0521**	**0.0371**	**0.0712**	**0.0432**
	+5.22%	+12.55%	+3.90%	+10.29%	+17.31%	+29.04%	+0.15%	+17.43%	+12.53%	+21.24%	+1.71%	+14.97%

Item Representation

我们对RQ-VAE语义ID进行了定性和定量分析。我们首先使用定性分析观察了语义ID的层次结构，并展示了亚马逊*美容数据集上RQ-VAE的级别数量设置为3，代码书大小分别为4，16和256的情况（图）。此外，我们还使用c_1来标记每个 内容的类别，并在总体类别分布中可视化c_1特定的类别。结果显示，c_1能够捕捉 内容的高级别类别，如当$c_1 = 3$时，含有大量与"头发"相关的产品；而大多数带有$c_1 = 1$的 内容是"化妆"和"皮肤"产品，用于面部、嘴唇和眼睛。图1显示了当c_1固定时，所有可能的c_2值的层次结构。



(a) The ground-truth category distribution for all the items in the dataset colored by the value of the first codeword c_1.

(b) The category distributions for items having the Semantic ID as $(c_1, *, *)$, where $c_1 \in \{1, 2, 3, 4\}$. The categories are color-coded based on the second semantic token c_2.

Figure 4: Qualitative study of RQ-VAE Semantic IDs (c_1, c_2, c_3, c_4) on the Amazon Beauty dataset. We show that the ground-truth categories are distributed across different Semantic tokens. Moreover, the RQVAE semantic IDs form a hierarchy of items, where the first semantic token (c_1) corresponds to coarse-level category, while second/third semantic token (c_2/c_3) correspond to fine-grained categories.

通过观察，我们可以发现第二编码词c_2能更精细地捕捉高阶语义信息[+]。RQ-VAE学习的语义ID层次性为我们开辟了许多新的研究方向，并在第 § 3进行了详细探讨。相较于基于随机原子ID的学习推荐系统，TIGER采用共享知识的语义ID，使语义相似的项具有重叠的编码词，进而有效利用了数据集中蕴含的语义相似知识。

表中比较了RQ-VAE与LHS语义ID的性能。对于LHS语义ID，我们采用8个随机平面和m=4来保证与RQ-VAE的基数一致。随机平面的参数从标准正态分布[+]中随机抽取，以保证其球对称性。结果显示RQ-VAE始终优于LHS。这表明，相较于使用随机投影产生的量化效果，非线性深度学习架构学习的语义ID更好。

此外，我们也比较了在生成检索推荐系统中的语义ID与随机ID的重要程度。我们使用m=4个随机码字来生成随机ID基准，每个 内容分配一个随机ID。随机ID基准的基数与RQ-VAE的语义ID相似。结果显示，语义ID始终优于随机ID基准，进一步强调了利用内容型语义信息的重要性。

New Capabilities

我们描述了RQ-VAE生成检索框架的新特性：冷启动推荐和推荐多样性。这些特性是由于RQ-VAE与我们的框架协同作用产生的，现有序列推荐模型无法直接实现这些应用需求。在下文我们将讨论如何使用TIGER在这些场景中应用。

本文研究了TIGER框架的冷启动推荐能力。由于现实世界中的推荐语料库的快速变化性质，新项不断引入。然而，现有模型无法处理新 内容，因为它们缺乏用户印象在训练语料库中。因此，TIGER框架通过利用 内容语义在预测下一个 内容时，可以轻松进行冷启动推荐。

为了进行这项分析，我们从亚马逊评论数据集Beauty中删除了5%的测试项，这些移除的项被称为"未见项"。我们在训练分割上训练RQ-VAE量化器和序列到序列模型，然后使用RQ-VAE模型为数据集中所有 内容生成Semantic IDs，包括任何不在 内容文集中出现的未见项。 根据模型预测的Semantic ID(c_1, c_2, c_3, c_4)，我们可以获取具有相同对应ID的已知 内容。

需要注意的是，每个模型预测的Semantic ID最多只能匹配数据集中的一个 内容。此外，包含前三个同名的语义令牌，即(c_1, c_2, c_3)的未见项也被包含到候选列表中。 最后，当我们在搜索一组前K个候选时，我们引入了一个超参数[+]ϵ，该参数指定了我们的框架选择未见项的最大比例。

在推荐系统评估中，Recall和NDCG是主要的评价指标。推荐系统缺乏多样性可能会损害用户长期参与度。本文提出使用生成检索框架来预测推荐系统的多样性，并且展示了如何在解码过程中通过温度采样来实现这一目标。虽然温度采样适用于任何现有的推荐模型，但是TIGER（Total Information Genetic Regularized VAE）模型具有更多的灵活性，因为它能够从多个层次上进行采样，包括从粗粒度类别检索物品以及在类别内检索物品。

Table 4: Recommendation diversity with temperature-based decoding.

Target Category	Most-common Categories for top-10 predicted items	
	T = 1.0	T = 2.0
Hair Styling Products	Hair Styling Products	Hair Styling Products, Hair Styling Tools, Skin Face
Tools Nail	Tools Nail	Tools Nail, Makeup Nails
Makeup Nails	Makeup Nails	Makeup Nails, Skin Hands & Nails, Tools Nail
Skin Eyes	Skin Eyes	Hair Relaxers, Skin Face, Hair Styling Products, Skin Eyes
Makeup Face	Tools Makeup Brushes,Makeup Face	Tools Makeup Brushes, Makeup Face,Skin Face, Makeup Sets, Hair Styling Tools
Hair Loss Products	Hair Loss Products,Skin Face, Skin Body	Skin Face, Hair Loss Products, Hair Shampoos,Hair & Scalp Treatments, Hair Conditioners

知乎

Table 3: The entropy of the category distribution predicted by the model for the Beauty dataset. A higher entropy corresponds more diverse items predicted by the model.

Temperature	Entropy@10	Entropy@20	Entropy@50
T = 1.0	0.76	1.14	1.70
T = 1.5	1.14	1.52	2.06
T = 2.0	1.38	1.76	2.28

Ablation Study

研究了序列到序列模型层数变化对性能影响。表显示随着网络增大，指标略有提升。同时评估了用户信息提供效果，结果在附录表中给出。

Table 5: Recall and NDCG metrics for different number layers.

Number of Layers	Recall@5	NDCG@5	Recall@10	NDCG@10
3	0.04499	0.03062	0.06699	0.03768
4	0.0454	0.0321	0.0648	0.0384
5	0.04633	0.03206	0.06596	0.03834

Invalid IDs

由于模型通过递归解码目标语义ID来预测可能无效的ID（这些ID不能映射到推荐数据集中的任何项），因此其可能性为256^4，约为4万亿。然而，我们使用的数据集数量为10K-20K。尽管模型可能生成无效的ID，但在检索项数量增加的情况下，它的表现几乎没有变化。我们可以通过增加束尺寸或过滤无效ID来防止这种情况的发生。

Conclusion

本文提出的TIGER是一种推荐系统检索候选人的方法，使用RQ-VAE生成语义ID表示，无需创建索引即可训练和使用。与需要线性增长*嵌入表基数的系统相比，我们的嵌入表基数与 内容空间基数不成比例。实验在三个数据集上进行，结果表明我们的模型实现了SOTA检索性能并能泛化到新 内容。

论文原文《Recommender Systems with Generative Retrieval》

编辑于 2024-02-20 19:46 · IP 属地北京

工业级推荐系统　生成式　Transformer

理性发言，友善互动

3 条评论　默认　最新

魂殇

MIPS是Maximum Inner Product Search，不是啥多实例...

02-05 · 浙江　回复　1

oner

有代码吗

04-25 · 广东　回复　喜欢