向量召回在躺平APP的实践

原创 陈雷慧（豆苗） 淘系技术 6天前

收录于话题
#躺平APP 1 #阿里巴巴淘系技术 9 #向量召回 1

背景

躺平APP是“躺平”这个大生态中生活记录社区，记录生活记录家。

躺平
记录种种生活

该业务场景中，个性化推荐在充分利用流量实现精细化运营、促进信息流动方面有着不可或缺的地位。在业务成长初期，其内容推荐也面临着如下两个问题：

- **用户冷启动**：对于无任何行为或行为稀疏的用户，难以有效的捕捉到用户的兴趣偏好，进行个性化推荐；
- **内容冷启动**：受限于有限的流量，也存在一定程度的内容冷启动，小众内容得不到推荐展示机会；

目前，业务处在拉新促活的阶段，因此我们迫切的需要缓解上述两个问题：

- 一方面要使得内容能够精准的触达新增和低活用户，提升留存，将其转化为平台的活跃用户；
- 另一方面也要保障长尾小众社区用户的内容有推荐展示的机会，避免小众用户的流失。

一般来说，推荐系统采用召回->排序两阶段的架构。召回阶段从海量内容池中召回数千条内容生成候选集，排序阶段利用用户、内容侧丰富的特征、上下文信息和复杂的模型对候选集中的内容进行打分排序，最终为用户返回数十条内容。

因而，召回是推荐系统的基础，很大程度决定了推荐效果的上限。随着表征学习的发展，向量召回逐渐成为推荐系统中提高召回准确性和多样性的一种行之有效的方案。

鉴于此，我们在原有的个性化召回和热门召回的架构上，新增一路个性化向量召回。并且，考虑到躺平APP是一个以图片为内容主要载体的UGC社区，我们主要从网络表征学习和多模态表征学习两个角度进行了相关的探索。

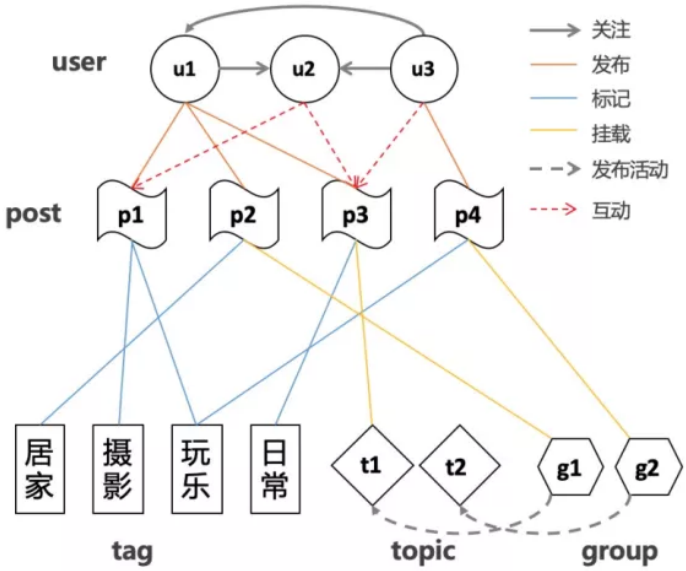
本文旨在分享向量召回在躺平APP社区上的应用与实践，内容将会分为以下三个部分：

- 第一部分为网络表征学习在召回上的落地
- 第二部分简单介绍利用多模态信息学习内容的向量表征
- 最后一部分对本文进行总结与展望

希望能对大家有所帮助与启发。

Graph Embedding的实践

用户在躺平APP社区的行为，如发布、点击、点赞、评论和收藏等都可以抽象为网络关系图。因此Graph Embedding非常自然地成为学习社区中用户与内容embedding的一项关键技术。



目前落地的模型大致分为两类：直接优化节点的浅层网络模型和基于GNN的深层网络模型。

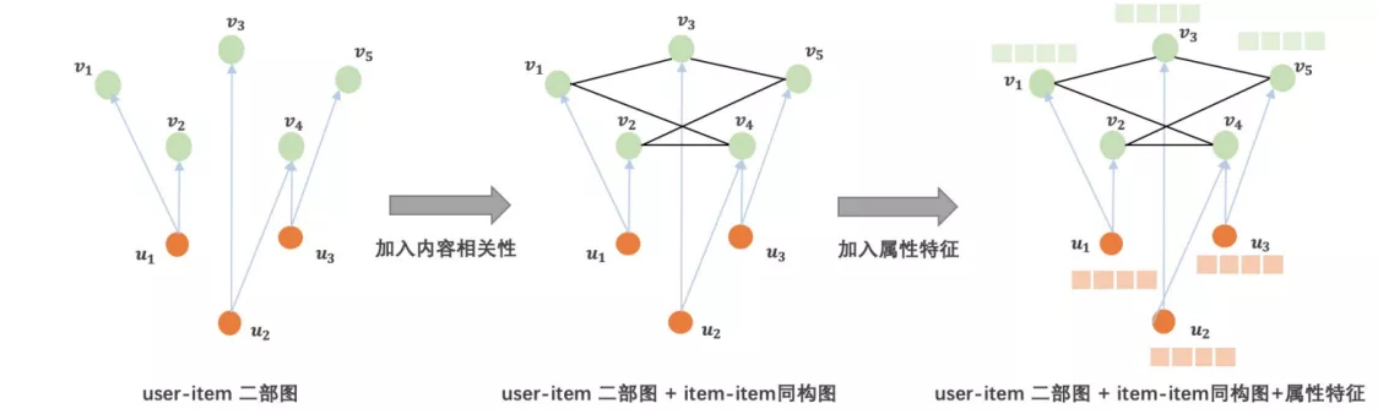
前者包括基于用户行为理解内容，学习内容向量表征的item2vec[1]，用于扩充i2i召回；同时学习用户与内容向量表征的异构网络表示学习模型metapath2vec[2]，用于提高内容召回的多样性；以群体行为代替个体行为的userCluster2vec[3]，缓解新用户冷启动问题。

后者包括采用邻域聚合的思想，同时融入节点属性信息，通过多层节点聚合使得网络中的拓扑结构能够有效捕获的GraphSAGE[4]，以及将attention机制运用邻域信息聚合阶段的GAT[5]，对用户与内容向量表征进行更加细致化学习。

本节主要介绍GraphSAGE和GAT在社区内容推荐的应用与实践。

网络构图

我们将用户与内容的交互行为构建为user-item二部图。考虑社区用户行为数据的稀疏性，我们在user-item二部图网络中加入描述item-item相关性的同构网络，避免孤立节点无法学习有效向量表征问题。



user-item 二部图网络

user-item二部图网络描述用户对内容的偏好，包含两类节点：user（用户）节点$U = \{u_1, u_2, \dots\}$和item（内容）节点$V = \{v_1, v_2, \dots\}$，以用户近7天的历史行为数据为一次session，构建网络的边。

以用户u_i与内容v_j为例，节点间边e_{ij}的权重w_{ij}的计算方式为：

$$w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{|V|} c_{ij}}$$

其中，c_{ij}为用户u_i对内容v_j互动次数，一次点击、点赞、评论、收藏均为一次互动。

item-item 同构网络

item-item同构网络衡量内容之间的相关性，仅包含一类节点：item（内容）节点$V = \{v_1, v_2, \dots\}$。从全站用户互动的内容序列挖掘得到内容之间的相关性。

属性特征

本次训练主要使用的是id类属性特征。用户侧使用如年龄、性别、地区等特征，内容侧使用标签、挂载的圈子和话题等特征。

技术实现

相比较于采用浅层网络直接优化节点的Graph Embedding模型，如DeepWalk、node2vec和metaPath2vec等，基于GNN的深层网络通过卷积(GCN)、注意力机制（GAT）和门控网络（GaAN[6]）等技术，对图的全局及局部拓扑结构进行更细致化的学习，并能够自然得融入节点的属性特征。

GraphSAGE

GNN模型在向量召回应用的第一阶段，我们主要选择归纳学习——GraphSAGE来学习用户与内容的向量表征。该模型在对邻居节点进行聚合前，先对邻居节点进行采样得到一个子图，并用该子图代替原图进行邻域聚合计算，从而有效地避免了GCN由于需要融合所有邻居节点而无法完成大规模数据向量表征学习的弊端。伪代码如下：

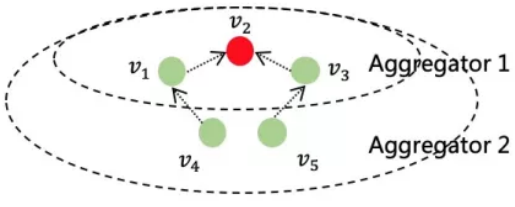
Algorithm 1: GraphSAGE embedding generation (i.e., forward propagation) algorithm

Input : Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$; input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$; depth K; weight matrices $\mathbf{W}^k, \forall k \in \{1, ..., K\}$; non-linearity σ; differentiable aggregator functions $\text{AGGREGATE}_k, \forall k \in \{1, ..., K\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

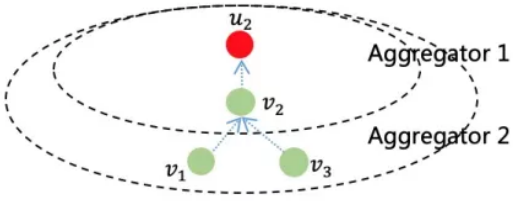
Output : Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

1 $\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$;
2 **for** $k = 1...K$ **do**
3 **for** $v \in \mathcal{V}$ **do**
4 $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
5 $\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$
6 **end**
7 $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2, \forall v \in \mathcal{V}$
8 **end**
9 $\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$

伪代码4-5行为邻域聚合层。以生成节点v第k层向量表征h_v^k为例，首先聚合采样得到的节点v所有近邻节点$N(v)$的第k-1层向量表示，从而得到节点v第k层的邻域聚合特征$h_{N(v)}^k$，并与节点v的k-1层向量表征h_v^{k-1}拼接起来，最终通过一层全连接转换得到节点v在第k层的向量表征h_v^k。



Item邻域聚合

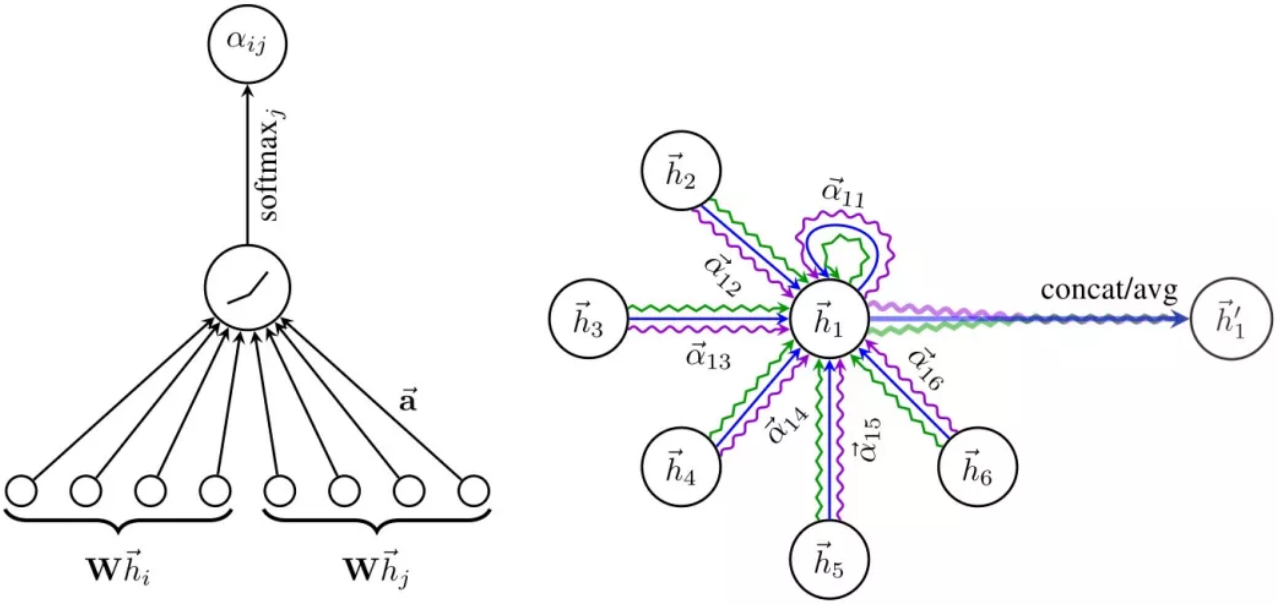


user邻域聚合

为了在充分学习到网络的拓扑结构信息的同时控制训练成本，我们进行了多组离线实验，最终发现仅采用二阶的邻域聚合就能获得比较理想的效果。具体地，通过聚合2阶邻域节点的embedding，生成一阶邻域节点embedding，再聚合一阶邻居embedding，并融合自己本身embedding，完成自身embedding的更新。需要注意的是，内容侧从item-item同构网络中采样得到邻居节点的子图，用户侧从user-item二部图中采样得到邻居节点。采用Cross Entropy的分类损失函数进行模型训练。

GAT

考虑用户行为序列中存在许多噪声，如误点击等，且用户兴趣在具有多样性同时也具有侧重性。在GNN模型在向量召回应用的第二阶段，我们进一步引入attention机制到邻域信息聚合阶段，借鉴GAT[5]思想, 对节点邻域节点进行差异化的信息组合，从而对用户兴趣进行更准确捕捉，提升召回效果。



针对图结构数据，GAT（graph attention networks）使用masked self-attention层使得图中的每个节点可以根据邻节点的特征，为其分配不同的权值，聚焦在相关输入，且GAT无需使用预先构建好的图。

实验证明，GAT模型可以有效地适用于（基于图的）归纳学习问题与转导学习问题。

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$

我们基本沿用原始论文中的聚合方案，在各节点采样完成的邻居节点上进行Attention操作，详细原理参考[5]。

对于user，通过$u-i-i-i...$的路径进行邻居采样；对于item，则仅通过$i-i-i-i...$的路径进行邻居采样用于后续embedding聚合阶段。在正负样本采样阶段，对于user节点，我们沿着path进行随机采样；对于item节点，借助item间的相似性系数，沿着path按照权重高低进行采样。

效果

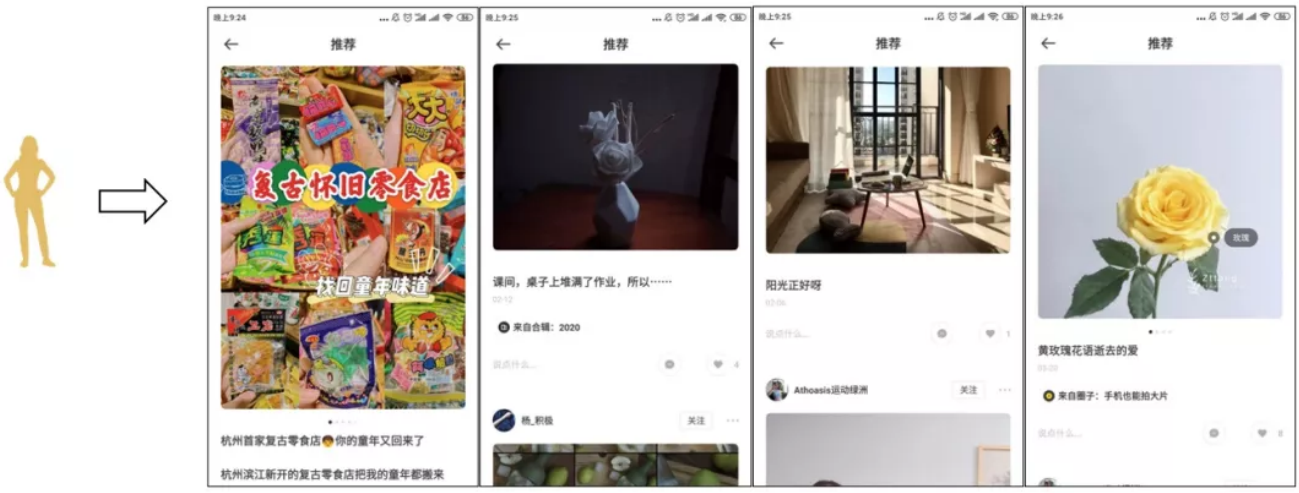
我们将训练得到的用户与内容embedding用于躺平APP内容推荐的召回的阶段，以用户的embedding作为trigger，寻找与其相似的内容，并送入后续的精排模型，进行打分。

A/B Test

用户点击率和曝光深度均有正向提升，加入属性信息的效果比未加入的更优。

case study

随机挑选一名女性和男性用户，观察其召回效果。其中，女性用户倾向于消费美食、手工DIY和居家生活类内容，男性用户则偏好潮玩、摄影和居家生活类内容。两位用户召回结果如下：

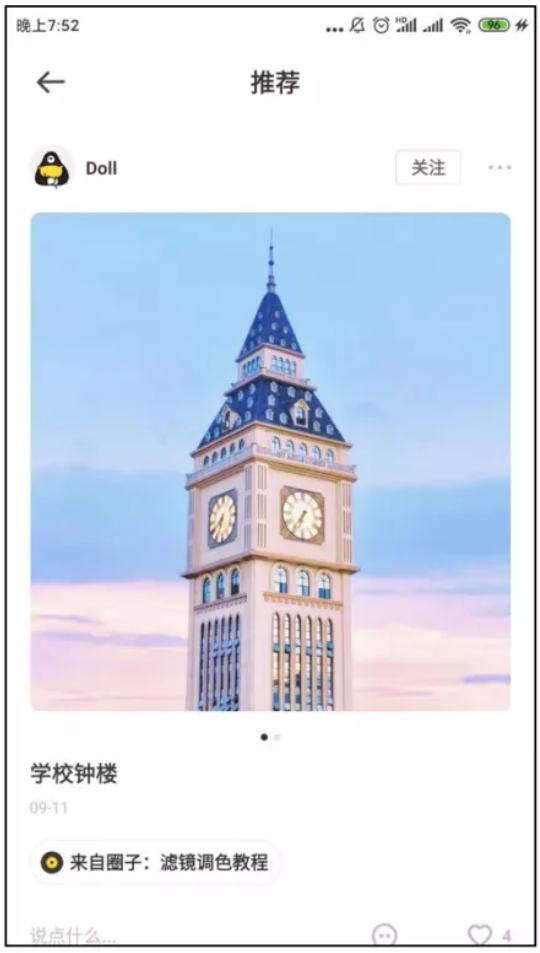


基本可命中用户的兴趣偏好。

多模态表征学习的实践

利用多模态信息学习内容的向量表征，是缓解由用户行为为主导的推荐系统中内容冷启动问题行之有效的方法之一，工业界也有很多相关的落地实践。受此启发，我们也希望从内容本身所包含的信息出发，挖掘内容之间的相关性。

主feeds双列流



详情页单列流

通过对社区内容调研发现，整体内容上存在以下两个明显的特点：

- 短文本随意性较大，字数较少，以阐述心情、描述生活状态为主；
- 图片质量较高，大都能清晰地展现出人、物或景，但也存在多图且内容发散的问题，即一条内容的多张图之间的关联性较弱；
- 无论是主feeds双列流还详情单列流，都是以用户发布的内容首图来进行展示。

综合考虑，我们尝试通过从内容首图的图片理解角度出发，学习到腰部和长尾内容的图片向量表征，并基于向量召回的技术，提高小众内容召回效果，缓解内容冷启动问题，提升推荐效果。

技术实现

关于图片特征的提取的方法有很多种，其中最为典型的就是利于基于ImageNet训练的ResNet[7]网络去提取图片特征向量。

但这种方法也有一定的局限性，首先，用于预训练的ImageNet图像数据的类别信息与躺平APP的内容类目体系存在较大差异；其次，ResNet网络学习出的图片向量维度较高，影响向量召回的效率。

因此，利用躺平的内容类目体系对基于ImageNet训练的ResNet网络进行fine-tuning，进而得到躺平内容的图片向量。

我们在预训练ResNet网络上添加了两层全连接对图片进行有监督的标签分类。该模型在产出标签的同时，也可以从中间网络层中提取出面向内容和业务理解的图片向量特征副产。

利用向量召回技术，将提取出的图片向量作为一路召回，添加进召回环节，用以缓解内容冷启动问题。

显然，标签的选择，对图片向量的效果有着至关重要的作用。

细粒度的标签，可以提高召回的准确性，但内容同质化现象较为明显，无法彰显社区的多样性。

粗粒度的标签，则无法保障召回的准确性。

我们依次从图片叶子标签、一级标签和二级标签等不同粒度的标签着手，对图片进行分类实验，最终确定以图片二级标签作为分类网络的优化目标，得以在保证产出的图片在拥有较高召回准确性的基础上兼具一定的泛化能力，对用户的兴趣偏好，能够进行发现性的探索。

效果

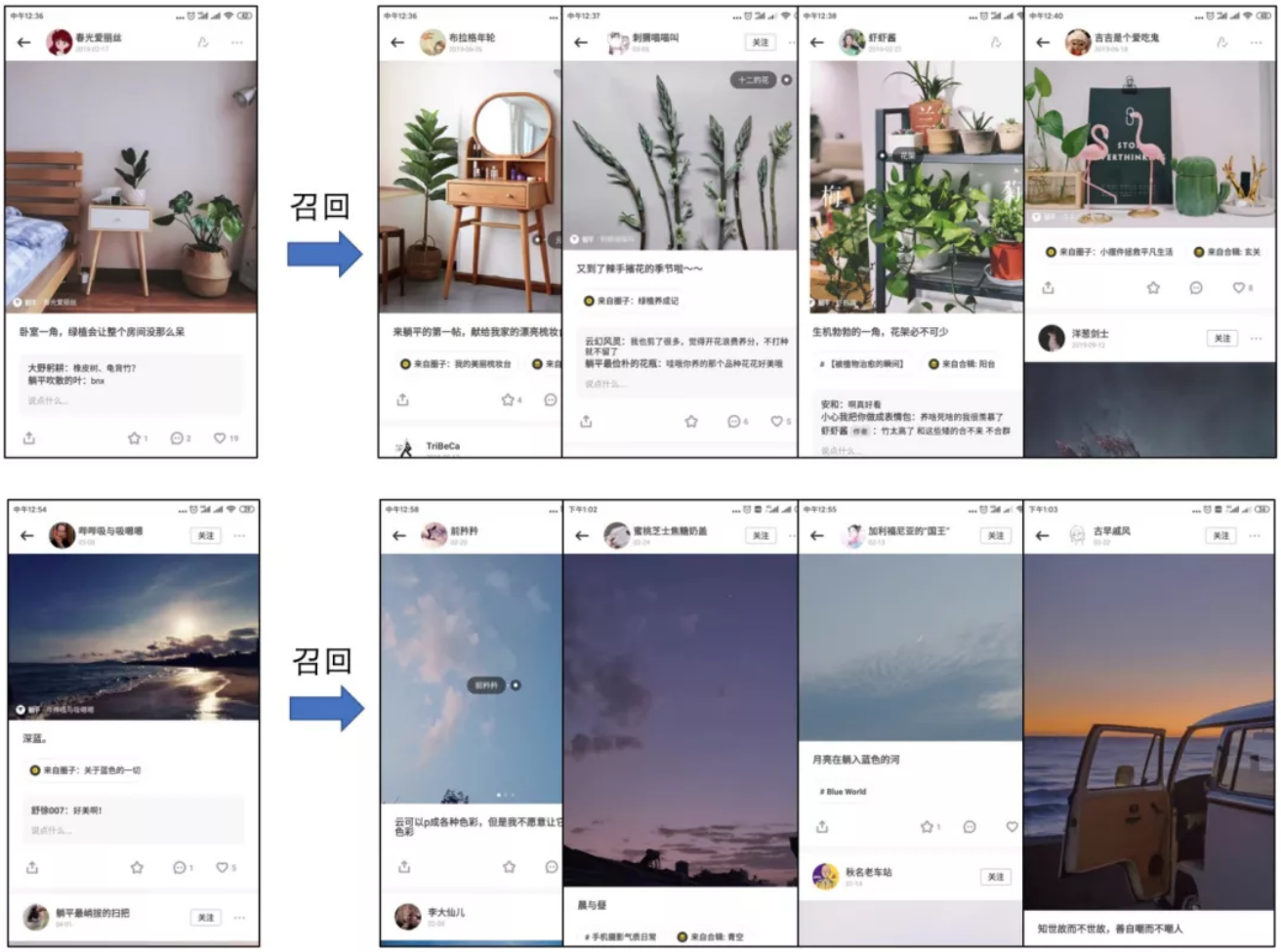
我们将训练得到的图片向量用于躺平APP内容推荐的召回的阶段，以用户的历史偏好以及实时点击内容的图片embedding作为trigger，寻找与其相似的内容，并送入后续的精排模型，进行打分。

A/B Test

在上线环节，我们对用户近期交互的内容首图向量进行向量召回，最终点击率有正向提效。

case study

随机挑选内容，离线观测召回效果，示例case如下图所示。从结果上来看，一些用户行为稀疏的相似内容可以被召回，起到缓解内容冷启动的作用。



总结展望

本文对躺平APP拉新促活阶段所面临的用户和内容冷启动问题的缓解方案进行了讨论。在综合考虑社区内容本身特点和业务的实际需要，在原有的召回架构上增加一路向量召回，期望提高召回结果的准确性和多样性。

首先，从社区属性出发，基于用户交互行为构建图网络，并利用基于归纳学习的GraphSAGE，对用户和内容进行表征，通过引入attention机制，改善向量表征学习的侧重点。相比较于传统的基于统计的召回，归纳学习的GNN模型在进一步提高行为稀疏用户召回内容精准性的情况下也增强了多样性。

其次，从图片内容社区角度出发，学习到腰部和长尾内容的图片向量表征，并通过向量召回的技术，提高小众内容召回效果，彰显社区的多样性，避免展示在推荐位的内容永远是热门的大众内容。期望促进信息流动，优化小众群体的生产与消费内容的体验，从而提高这部分用户的留存。

但是，在Graph Embedding向量召回建模方面，目前我们只利用了点击数据，在处理好数据不足和不平衡的情况下，后续可以融入对不同用户行为的建模，从而进一步挖掘用户行为的隐含兴趣偏好；也可以在保留行为顺序基础上结合GNN进行用户长短周期兴趣建模；在多模态内容表征学习方面，整体设计较为基础，后续会融入挂载圈子、话题和内容文本等信息，探索出更为丰富的内容表征，也欢迎大家一起交流。

参考文献：

[1] Barkan O, Koenigstein N. Item2vec: neural item embedding for collaborative filtering[C]//2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2016: 1-6.

[2] Dong Y, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 135-144.

[3] Grbovic M, Cheng H. Real-time personalization using embeddings for search ranking at airbnb[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 311-320.

[4] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[5] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[6] Zhang J, Shi X, Xie J, et al. Gaan: Gated attention networks for learning on large and spatiotemporal graphs[J]. arXiv preprint arXiv:1803.07294, 2018.

[7] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2015.

淘系技术部-商业机器智能团队

商业机器智能部是一支数据和算法一体的团队，服务于淘宝、天猫、聚划算、闲鱼和躺平等业务线的二十余个业务场景，提供线上零售、内容社区、3D智能设计和端上智能等数据和算法服务。我们通过机器学习、强化学习、数据挖掘、机器视觉、NLP、运筹学、3D算法、搜索和推荐算法，为千万商家寻找商机，为平台运营提供智能化方案，

为用户提高使用体验，为设计师提供自动搭配和布局，从而促进平台和生态的供给繁荣和用户增长，不断拓展商业边界。

这是一支快速成长中的学习型团队。在创造业务价值的同时，我们不断输出学术成果，在KDD、ICCV、Management Science等国际会议和杂志上发表数篇学术论文。团队学习氛围浓厚，每年组织上百场技术分享交流，互相学习和启发。真诚邀请海内外相关方向的优秀人才加入我们，在这里成长并贡献才智。

如果您有兴趣可将简历发至**leihui.clh@alibaba-inc.com**，期待您的加入！

✿ 拓展阅读

作者|陈雷慧（豆苗）

编辑|橙子君

出品|阿里巴巴新零售淘系技术