Embedding技术在58商业的探索与实践

原创 刘笠熙、刘杨 58技术 3月30日

导语

本文主要介绍商业策略技术团队在Embedding技术上的探索实践。通过介绍主流的Embedding技术，结合具体业务场景，阐述该技术在广告检索的召回、Rank两个环节的实践应用，并分享对新技术落地的实践效果和心得体会。

前言

随着2013年Google的word2vec拉开Embedding技术发展的序幕，越来越多相关的研究产出和应用落地，2016年微软的item2vec将Embedding技术从NLP领域推广到推荐、广告、搜索等领域，2018年阿里对Graph Embedding针对性的改进，使得该技术在工业界获得非常成功的应用。本地生活网络平台上存在大量需要对用户、帖子、文本进行Embedding的场景，如招聘业务中求职者和岗位的匹配度计算、黄页的搜索文本相关性计算、二手车业务的用户偏好计算等。在进行商业流量变现的同时，兼顾用户体验的平衡不可或缺，如何强化对用户兴趣的考量，优化广告检索过程，同时带来用户体验与商业收入的双提升，是我们需要不断探索的问题。

本文主要介绍商业策略技术团队在Embedding技术上的探索实践，包括以下几点：

- 主流的Embedding技术简介
- 二手车应用场景简介
- Embedding技术在广告召回的应用
- Embedding 技术在广告排序的应用
- 总结和展望

主流的Embedding技术简介

简而言之，Embedding泛指用一个稠密向量来描述一个实体，这种描述不仅仅包含该实体本身的属性信息，还包含了实体之间的关系以及该描述方式和最终目标之间的关系，目前在推荐场景下，有如下几种主流的对用户、帖子进行Embedding的技术。

1. Skip-Gram Model

基于skip-gram的embedding最有代表性的就是word2vec和item2vec，推荐场景下基于用户行为序列，对用户产生行为的实体进行embedding，假定用户在短时间内产生行为的帖子具有某种关联，以产生行为的帖子为正样本，未产生行为的帖子为负样本或者全局随机负采样，然后基于极大似然和softmax函数构建优化目标，进而完成模型构建。

2. Graph-embedding

Graph-embedding里面比较有代表性的是Deep Walk，其主要思想是根据用户行为构建帖子之间的关系，然后采用随机游走生成各种物品序列，序列生成后，进而采用skip-gram进行模型构建。

3. NN-embedding

NN-embedding采用多层神经网络，将高维稀疏向量映射为低维稠密向量，是使用传统深度学习的常用方法。这种方法计算的embedding跟loss设定的目标强相关，并且随着技术迭代，可以融入Attention机制以及其它复杂网络，逐渐演变为一种灵活的适用于各种场景的有效技术手段。

上面提到的三类主流Embedding技术，各有优缺点以及其适合的场景，Graph-embedding在信息捕获及表达能力上强于skip-gram，但建模复杂度和计算复杂度提升明显。Skip-gram和NN-embedding的优点是基础模型成本低，并且对于sequence类的数据天然支持能力较好，其中NN-embedding的可拓展

性极强，后续可迭代空间大。结合商业的具体场景，我们从落地成本和后续拓展性考虑，在召回和rank场景，分别落地了基于skip-gram的embedding和基于多层神经网络的NN-embedding，都获得了不错的业务收益。

二手车应用场景简介

笔者曾在二手车业务上进行了Embedding技术实践，二手车有以下业务特点：

- 用户复杂多变：二手车购买属于大件消费，二手车用户决策周期通常偏长，用户会在站内产生很多类型的行为，这些行为直接或间接反应了用户对车的偏好。从行为数据中可以观察到，用户的偏好会发生变迁，一段时间内会呈现偏好从发散到逐步收敛，然后又发散再收敛持续往复的现象。另外用户中有正常用户、中介以及车商，这些用户行为差异很大，例如中介的偏好会极其分散，并且转化率极高等；
- 信息标准化：二手车的信息非常结构化、标准化，大部分信息都围绕车辆属性，例如出的品牌、价格、车系、车型、里程等；
- 业务场景复杂：从流量上看有listing筛选流量、推荐流量、搜索流量，从商业产品上看有精选（cpc类）、置顶（cpt类）、来电通（cpa）等多种商业产品，这些商业产品的商业模式差异较大；

二手车业务是一个重匹配效率业务，如何在这些复杂场景下，提升各种商业产品的匹配效率，达到用户体验和商业收益的共同提升，是在进行技术落地时，需要去解决的问题。

Embedding技术在广告召回的应用

1. 推荐召回的主要流程

在召回阶段，当前系统主要由AB test分流模块、推荐算法模块、结果融合模块和结果过滤模块4部分组成（如图1所示）。

- AB test分流模块，目前提供基于uv、pv、uv+时间片3种分流方式，不同的流量支持按配置调用不同的算法、过滤等逻辑。

- 推荐算法模块，通过配置可以在不同场景应用不同的推荐算法组合做广告（物品）召回。
- 结果融合模块，当配置有多种通道的推荐召回策略时，需要对结果做合并。目前支持2种方式，第一种是按配置依次取各通道的结果；第二种是给不同的通道赋不同的权重，各通道将召回物品归一化后的得分乘以该通道的权重得到最终分数，按分数取top得到推荐结果。
- 结果过滤模块，该模块包含已下线广告过滤、展示频次过滤及业务规则过滤，根据不同场景的业务需要灵活适配。

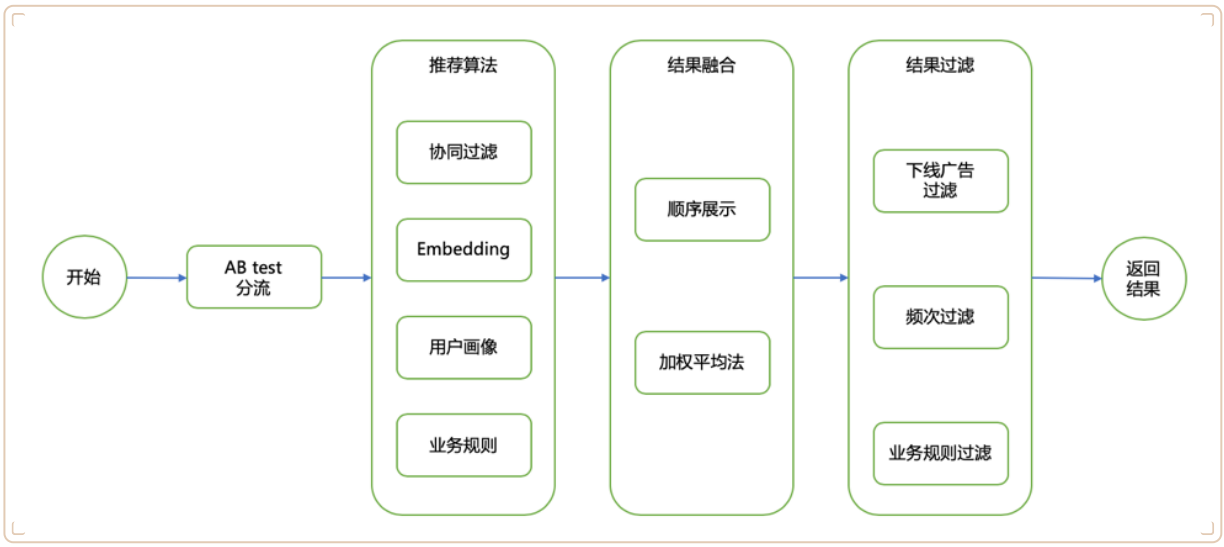

图1 推荐召回的主要流程

在推荐算法模块中，当前线上运行的有协同过滤类算法、基于embedding技术的向量召回算法、基于标签用户画像的召回方法和一些基于业务规则的召回方法（例如基于LBS的召回、基于热门度的召回）。

2. 初版召回算法：协同过滤

协同过滤是最经典的一类推荐算法，典型的有基于用户（User-Based）和基于物品（Item-Based）的协同过滤，对于一些用户量大、新客比例较高的互联网平台，我们使用的是基于物品的协同过滤。在具体实践中，我们根据实际业务特点对协同过滤做了大量优化：首先，模型融合了点击、收藏、在线沟通、转化（电话）等多种行为，并使用时间衰减策略增强用户近期行为的表达；然后对用户做了有针对性的筛选，剔除掉如爬虫等异常用户；在为用户生成推荐结果时，会考虑帖子（物品）的发布时间，对新帖子做一定加权，同时使用一些人工规则过滤掉明显的badcase等等。算法上线后，相比基于物品热门度这种非个性化的推荐方法，在点击率和转化率业务指标上有显著的提升。

协同过滤虽然取得了一定成绩，但它也有很大的缺陷——协同过滤是一种对“用户-物品”联系的记忆模型，少“泛化”能力，对稀疏行为场景的表达能力较弱。在具体实践中，这些问题突出表现为召回物品集合较少，无法处理新物品的冷启动问题等。

3. 基于Embedding的向量召回

针对以上问题，我们首先借鉴了Simrank++[1]模型的思想，通过使用随机游走的方法，捕捉更远距离的“用户-物品”相似性关系，使得更多的帖子能够有机会被推荐。该算法上线后，相比基础的协同过滤，点击率提升了约10%。但是随机游走的方法计算量巨大，即便在Spark分布式计算环境下也需要数天的时间才能收敛，不适用于短时间内有帖子（物品）大量上下线情况的业务场景在此之后，我们采用了“将物品通过Embedding向量化表达进而计算相似度”的思路来改善协同过滤泛化能力不足的问题。

在物品Embedding做法上，我们核心参考了Airbnb在2018年发表的paper《Real-time Personalization using Embeddings for Search Ranking at Airbnb》[2]里的思路：将用户的行为序列（主要是点击和电话）涉及的物品视为一系列上下文，使用word2vec的Skip-gram模型将物品向量化。具体如下：

首先是构造样本。我们将用户的一系列行为按session进行组织，session代表了一段时间内用户的连续行为，而session之间有明显的一段时间间隔。我们认为在一个session内用户的兴趣是稳定的，而在不同session内用户的兴趣是可能发生变化的。session的划分规则在不同业务场景中有所不同。

然后是训练word2vec模型。这里我们选择Skip-gram模型，虽然和CBOW模型相比需要更长的训练时间，但Skip-gram更擅长学习低频、长尾物品的embedding向量。使用负采样（negative sampling）技术后、Skip-gram模型需要优化的目标函数如下：

$$\text{argmax}_\theta \sum_{(l,c)\in\mathfrak{D}_p} \log\frac{1}{1+e^{-v'_c v_l}} + \sum_{(l,c)\in\mathfrak{D}_n} \log\frac{1}{1+e^{v'_c v_l}} + \sum_{(l,c)\in\mathfrak{D}_b} \log\frac{1}{1+e^{-v'_c v_l}}$$

其中， 表示需要被学习物品， 表示物品的上下文物品，是从session构造的正样本集合，表示从物品全局采样的负样本集合，是从session中针对有转化行为（电话/简历投递）的重采样。图2是Skip-gram模型

训练物品向量的示例图，标红的物品代表在一个session中发生了转换(电话/简历投递)，通过将发生点击转化的物品引入到所在session的每一个物品上下文中、强化了转化动作在向量模型中的影响。

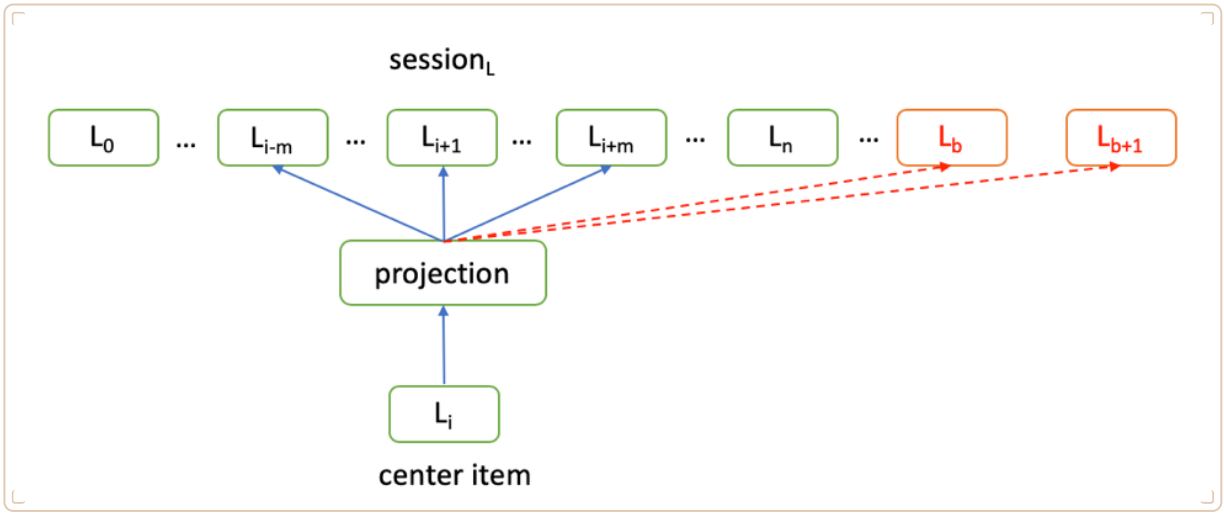

图2 Skip-gram模型训练物品向量

在新物品的冷启动问题上，我们选取和新物品具有相同属性的其它物品（如处于相同价格区间、相同售卖城市、相同车型车系、使用年限等属性的二手车）的embedding向量、经过average操作后做为新物品的向量。

4. 向量召回的问题与优化

在具体实践中，为了保证物品向量的质量，我们在session和模型训练样本上做了一些优化。用户的行为session关键在于"质量"。以二手车平台为例，C端的用户即包括普通消费者也有二手车商，车商的行为非常发散（即便在同一个session内），表现在各类价格区间、各类车型等都会有点击。因此我们会通过一些人工规则排除掉明显的车商的行为session，同时根据一些业务经验，将一个session中拆分成多个session、或者过滤掉session中显著的离群行为。其次在负样本的选择上，因为二手车平台的业务场景有很强的本地属性，为了使模型能够学习到这一点，我们会增加异地物品在负样本中的比例。

在获得物品embedding向量后，我们沿用基于物品的协同过滤思路——取用户最近N次的点击物品，然后从全库中检索，找出每个点击物品最相似（余弦相似度）的topK个物品作为推荐结果。我们使用Facebook的Faiss技术做全库检索，系统设计如图3所示：系统接收用户的实时点击消息，在Faiss索引中完成检索后，将结果更新到数据库中，供线上服务查询。在这里需要特别注意，系统还需要接收新物品的

上线消息实时更新到Faiss索引中，同时由于Faiss索引暂不支持删除操作，我们会对Faiss索引在每天凌晨时做全量更新，避免召回的结果包含大量下线物品。

基于Skip-gram模型的物品向量上线ab测后，和基线协同过滤相比，在单uv的转化率指标上带来了10%的提升。

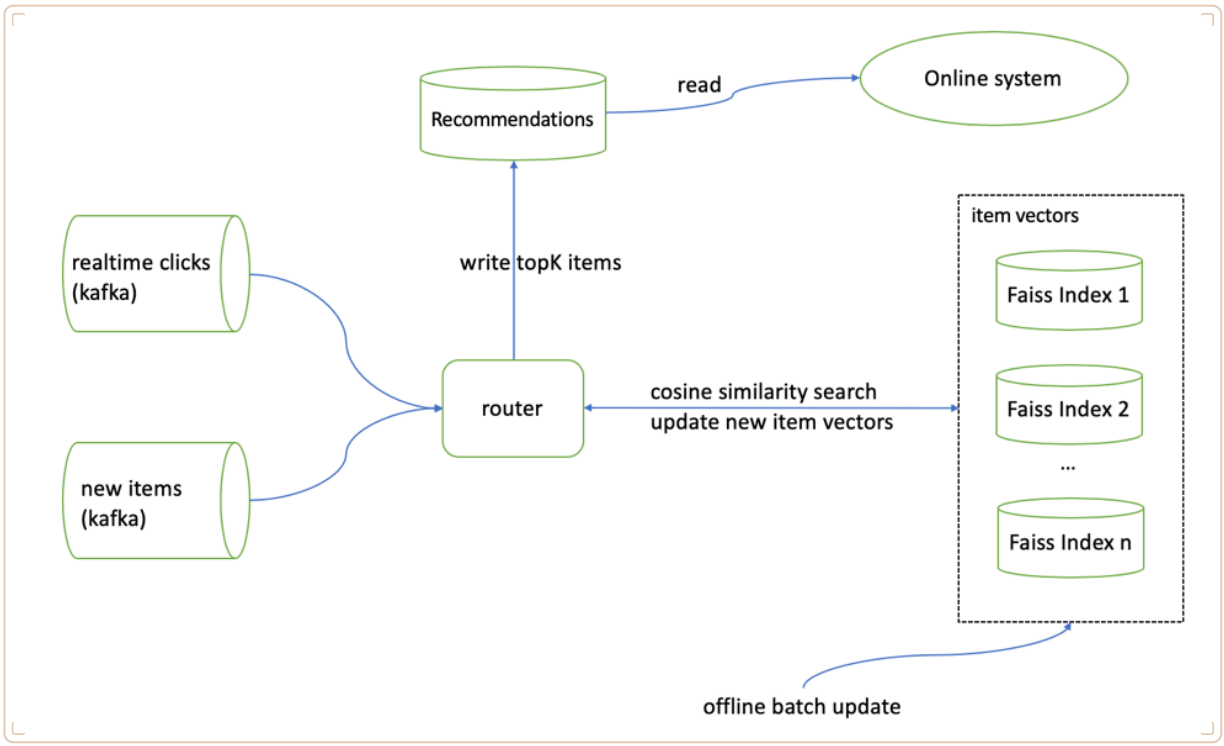

图3 基于Faiss的全库检索流程

考虑到二手车的业务偏低频，新用户的冷启动问题比较突出，我们后续计划借鉴Youtube的DNN模型的思想，在用户的行为类信息的基础上，入上下文信息来缓解新用户的冷启动问题。除此之外，通过将推荐问题转变成一个基于DNN模型的二分类问题，用户向量能够以一种有监督的方式代替无监督的方式学习得到。

除了在召回场景，我们同样将embedding技术应用在了精排和粗排场景，并基于不同业务特性进行了针对性优化。

Embedding技术在广告Rank的应用

对于本地生活平台业务中，有主列表listing、推荐、搜索多种流量场景，仍然以二手车业务线为例，对于推荐流量，召回和排序都需要重点考虑用户偏好与帖子的相关性；对于listing下的筛选流量，虽然带了用户的强筛选意图，但在部分筛选条件下，例如价格区间，筛选后帖子候选集依旧比较充分，这里依旧可以

引入用户除价格区间外其他的偏好，进一步增加用户偏好相关性。为了进一步提升用户个性化表达，基于多因子排序框架，我们引入了用户偏好相关性因子，下面将具体介绍如何基于Embedding技术构建用户兴趣相关性因子，以及针对本地生活场景特性的优化方案。

1. 相关性预估整体方案

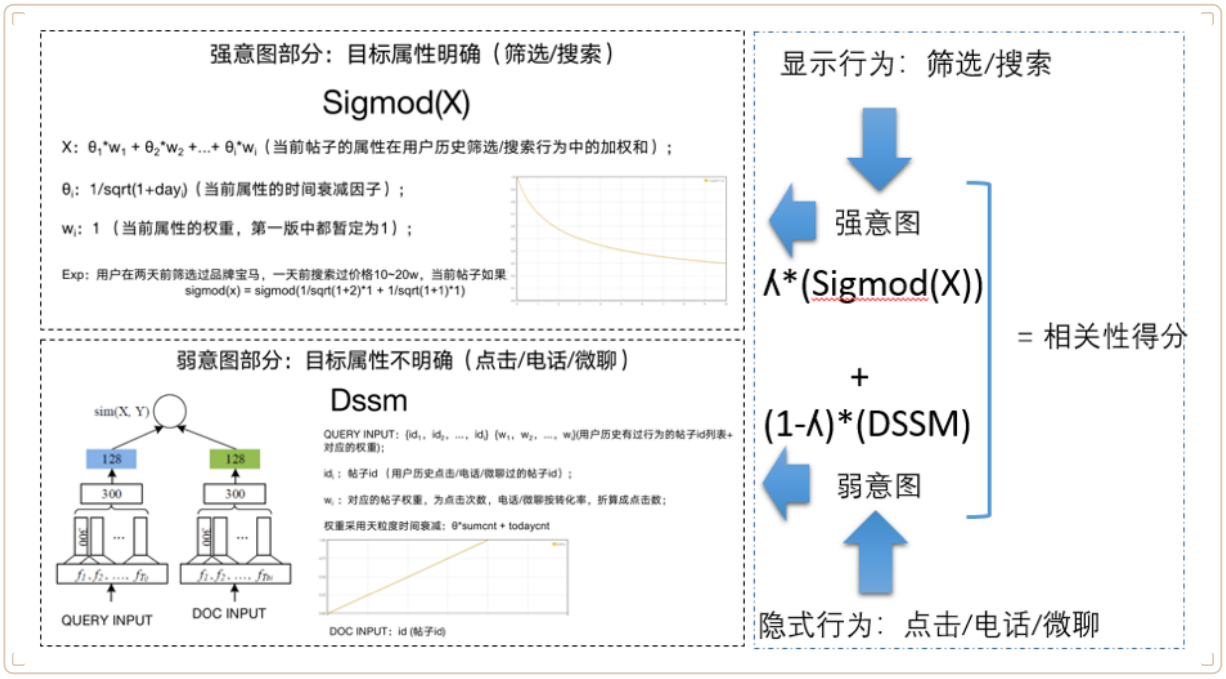

图4 相关性预估整体方案

这里我们将用户的行为拆成两种类型，一种是显示行为：筛选/搜索，因为这种行为会直接映射到具体属性值上，例如二手车场景，用户筛选品牌宝马，是一种用户对宝马品牌表达的明确的意图；另一种是隐式行为：点击/电话/微聊，这类行为落到帖子粒度，一个帖子包含多种属性信息，无法直接判断用户是因为哪些属性产生的行为，需要大量的类似信息，才能挖掘出用户的潜在意图。我们基于词袋模型构建强意图表达显示行为，使用Embedding技术构建弱意图表达隐式行为。最后对用户的强意图和弱意图进行融合，得到用户的兴趣偏好相关性得分。这里采用全站（商业数据+非商业数据），全用户行为（搜索/筛选/点击/电话/微聊等）去进行用户兴趣向量和帖子向量生成，能够全面的、完整的描述用户偏好。对外暴露生成好的用户向量和帖子向量，生成环节和使用环节解耦，可以方便各个环节快速使用。

2. Embedding：DSSM/Item2vec

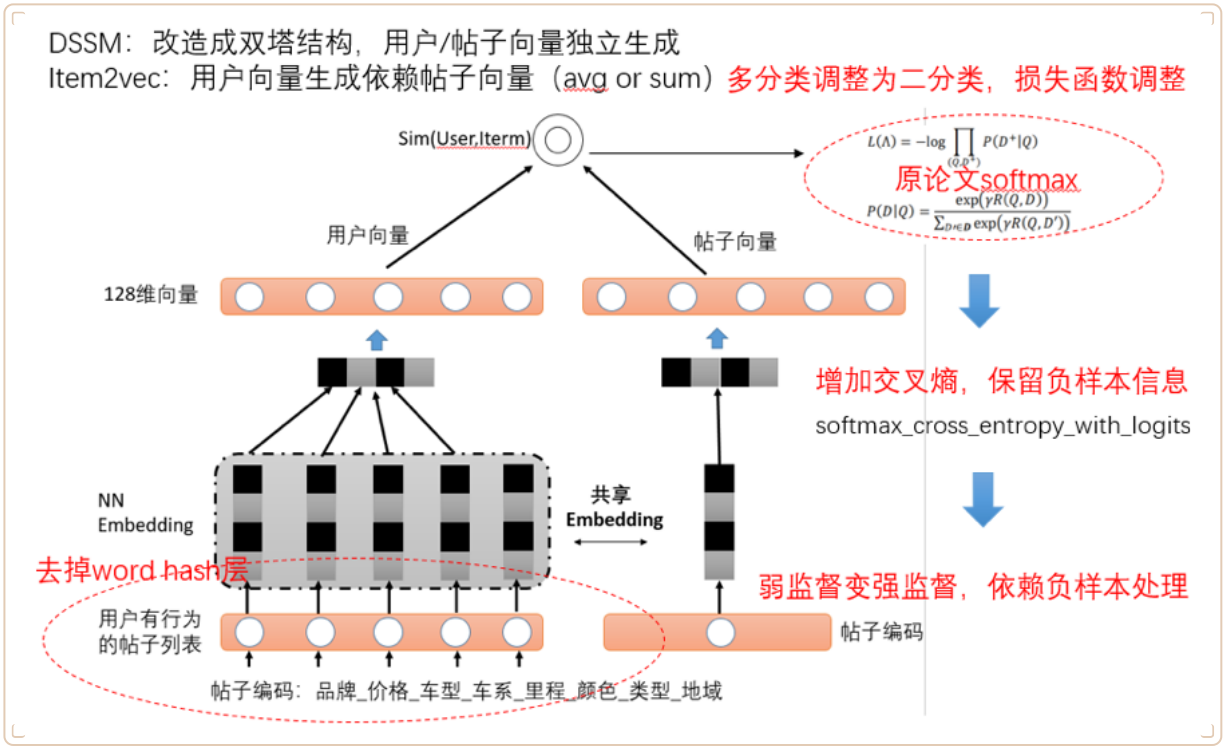

图5 定制化DSSM模型

在生成用户向量和帖子向量时，我们尝试了两种主流模型，基于skip-gram的item2vec[4]以及基于NN-embedding的dssm[3]模型，这里主要介绍我们如何使用dssm来进行向量生成的。

（1）样本生成：

样本基于用户的点击/转化数据构建。用户对某个帖子点击一次，就组成一个（用户行为列表，帖子id，1）的正样本元组，其中“用户行为列表”是一个行为序列：用户点击/转化过的帖子序列例如“{(帖子id1,w1),(帖子id2, w2),(帖子id3, w3)}”，wi表示权重，权重计算会考虑用户对该帖子的点击/转化次数；负样本需要满足几个条件，1.该用户未对该帖子进行点击/转化行为；2.该帖子跟正样本中的帖子属性差异较大，该部分会通过一些规则来进行控制，例如正负样本帖子信息中的价格不在同一价格区间段，品牌不在同一档次等；在满足上述两个条件的帖子中，随机抽取n条，生成（用户行为列表，帖子id，0）这种负样本元组。这样我们就完成了正负样本的生成。

（2）模型输入层：

原论文中dssm为多分支形态，这里我们将其调整为双塔结构，其中左半部分用于学习用户向量，右半部分用于学习帖子向量。这里，为了简化模型提升训练效率，去除了word hash层。

（3）Embedding层：

这里使用传统的NN-embedding，使用一个双层神经网络进行embedding，同时借鉴了阿里ESMM模型中共享embedding的思想，共享用户行为序列中帖子的embedding，能够让其虽然是双塔结构，但是能保持在一个向量空间。

（4）损失函数调整：

原论文中dssm的损失函数使用的正样本的softmax结果，丢弃了负样本对损失的贡献，其原因是在原论文的场景下，其正负样本包含的信息是弱监督信息，但我们的正负样本进行过特殊处理，已经包含了强监督信息。所以我们对损失函数进行了调整，由原论文正样本的极大似然+softmax损失，调整为余弦相似性+交叉熵损失，使得负样本也能对embedding结果进行信息补充。

按上述这些步骤，我们构建了一个能支持用户向量和帖子向量生成的定制化dssm模型，在实际落地过程中，还遇到了一些问题，将在下一节中进行相应阐述。

3. 相关性预估的问题与优化

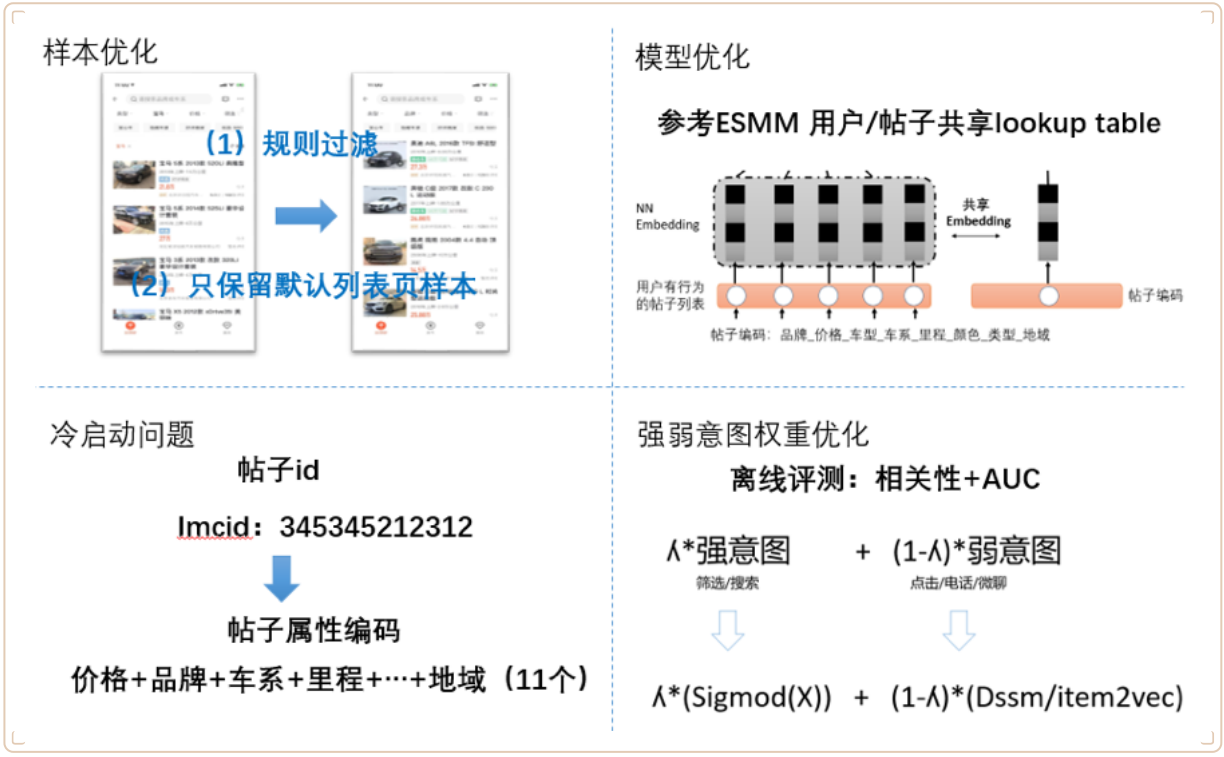

图6 相关性预估-问题和优化

实际落地过程中，遇到了不少问题，主要汇总为图6中四大方面：

（1）样本优化：

最开始我们使用的是点击作为正样本，未点击的作为负样本，这里会有2个问题，1.线上流量中有筛选场景，筛选场景下的帖子相似性本身都相对较高，比如筛选了品牌宝马，出来的都是宝马，在这种场景下去进行样本挑选，会产生很多不靠谱的负样本；2.用户对于帖子点击或者不点击，并不一定完全是跟用户兴趣不相关，有可能在列表中，有两个相似的帖子，用户点了第一个，就不想点第二个，如果不进行过滤，这样原本是相似的帖子，反而会变成负样本，影响模型效果。最后我们只保留了无筛选条件的样本，并且进行帖子相关性的过滤，如果某条帖子出现在某个用户的正样本中，那么跟这条帖子相似的帖子都不会出现在该用户的负样本中，这里采用一些规则进行过滤；

（2）模型优化：

双塔共享embedding这个在前文中简单提到过，主要的目的是为了保证最后生成的用户向量和帖子向量在同一个向量空间，能有效提升embedding的效果；

（3）冷启动问题：

这个问题需要重点提一下，基于帖子id进行embedding有两个天然的问题，1.稀疏性问题，最后进行结果评测时，能发现展现点击量大的帖子，embedding的效果较好，展现点击量小的帖子，embedding结果差强人意；2.新帖子问题，对于模型没见过的新帖子id，无法进行embedding；针对上述这两个问题，我们没有直接使用帖子id进行embedding，而是使用帖子的属性编码，以二手车为例，我们采用帖子对应的车辆属性编码，例如价格+品牌+车系+...+地域等11个关键信息进行编码，类似于提前进行了帖子的细粒度分类，然后模型去学习这些属性编码的embedding，一个是一种编码对应的数据量更多，稀疏性得到了解决，另一种是当新帖子来的时候，直接去找对应属性编码的embedding结果，较好的解决了上述两个问题。

（4）强弱意图权重调整：

上文提到我们将用户兴趣拆解成了强意图部分和弱意图部分，那么在这两部分得分进行融合的时候，如何来进行权重分配。这里我们搭建了一套离线评测环境，人工按相关性精细标注了一批样本，然后采用一个简单的线性回归调整这两部分的权重，去达到在这批样本上的AUC最佳，从而得到这两部分线上最后使用的权重。

4. 相关性预估的效果评估

我们进行了三个环节的评估，分别是帖子embedding结果评估，用户embedding结果评估，线上业务评估。

（1）帖子embedding结果评估，帖子embedding结果评估我们采用了两种方式，第一种是人工评估，随机抽取一批帖子的embedding结果，按余弦相似度找出top帖子，人工评估top帖子跟该帖子的相关性，如图7所示可以看出，输入帖子和找到的top相关性（计算得出）帖子的真实相关性较高，说明帖子embedding结果较好。第二种方法是对帖子向量进行降维，然后随机抽取一批帖子进行点图绘制，查看关键维度情况。如图8所示，我们使用PCA将帖子向量降低到2维，然后将其中相同品牌的点绘制成相同颜色，可以看出宝马（蓝色）和保时捷（红色）有部分重合，因为同属高端品牌，同时宝马和大众（绿色）则区分较大。说明在品牌这个维度，帖子embedding的结果较好；

图7 帖子embedding结果评估

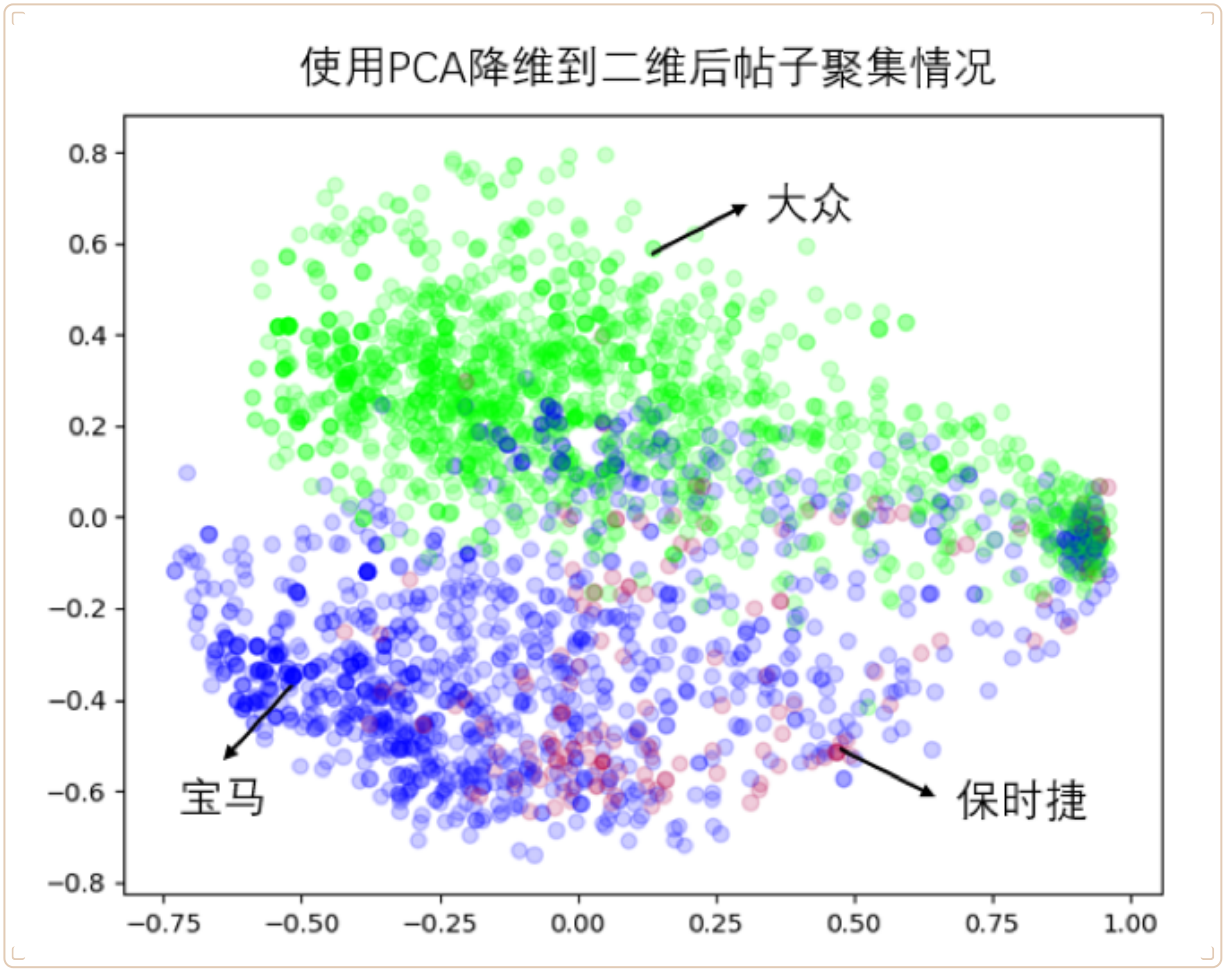

图8 帖子使用PCA降低到二维后帖子聚集情况

（2）用户embedding结果评估，随机抽取用户，拉取用户历史行为数据，使用这些历史行为数据生成用户embedding结果，然后按余弦相似性找top帖子，人工评估top帖子跟用户历史行为的相关性，如图9所示，可以看出用户embedding的效果较好；

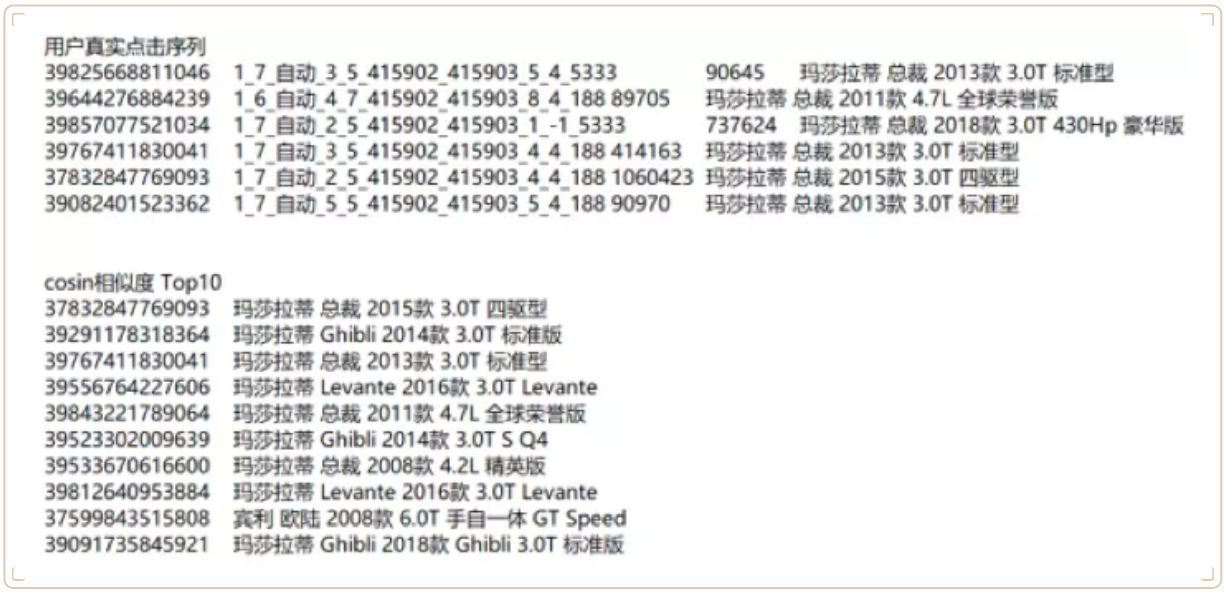

图9用户embedding结果评估

（3）线上业务评估，最终上线后，该相关性因子作为一个权重较大的因子，生效在精排环节，我们人工评估了实验和基线的NDCG得分，相比基线，NDCG得分提升12.36%；对比线上ab测实验数据，在原有转化率预估模型的基础上，单用户转化率获得了10%以上的提升；

5. 粗排引入相关性

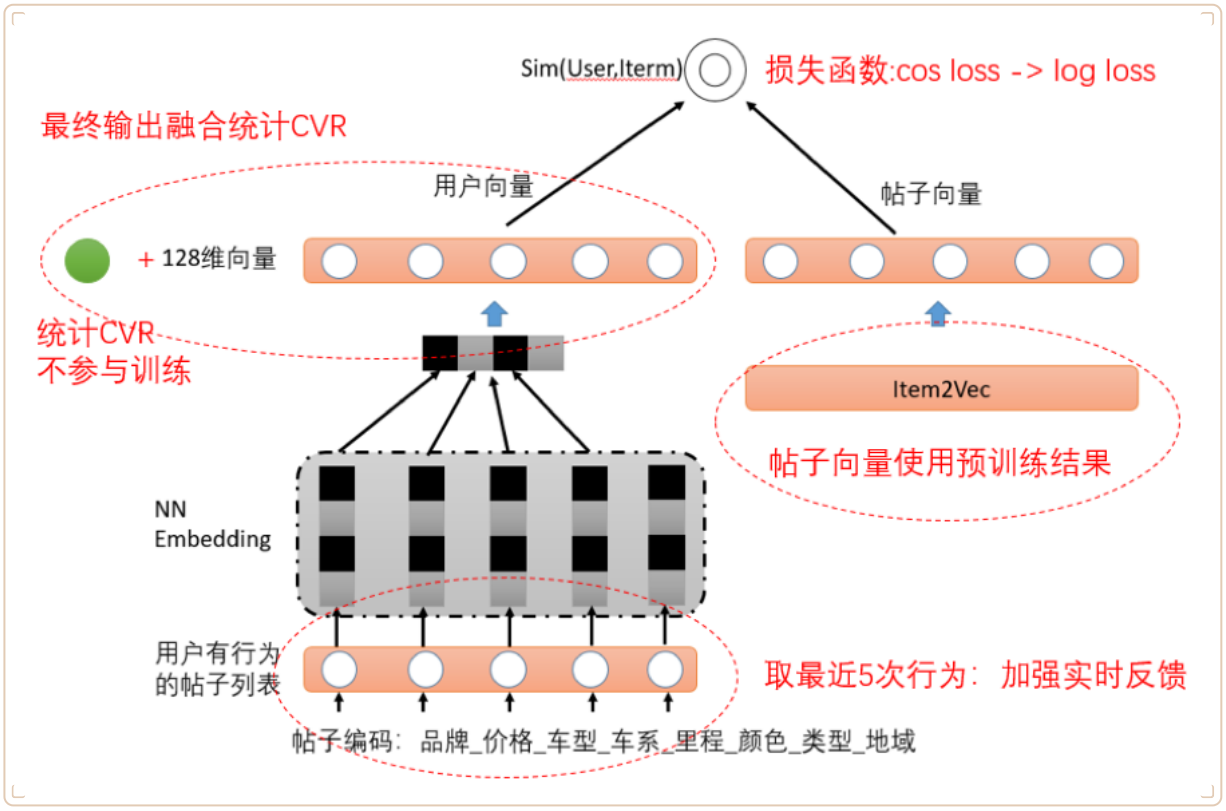

图10 相关性在粗排环节的应用

相关性因子在精排环节获得一些收益后，我们又尝试将其应用在粗排环节，粗排主要是为精排提供候选的环节，这里选择应用在粗排环节是出于3个方面的考虑：

（1）多个业务线部分推荐/筛选场景下的候选集充足，目前精排的复杂模型的排序能力是“百”这个量级的，对出超过量级（千，万）的情况，需要进行粗选，增加进入到精排候选集的质量；

（2）产品形态的升级带来候选集增加，目前正在建设的feed流+多商业产品混排场景会带来候选集的暴涨；

（3）目前的粗筛使用的是历史统计值，无法进行用户个性化，更无法捕获用户的实时偏好变化，存在较大优化空间。

基于上述几点，我们在粗排环节引入了相关性因子，实际落地过程中，经过数次迭代优化，目前能看到一些业务收益，具体优化路径如下：

（1）加强实效性：在生成用户向量时，在模型输入层，只保留用户最新最近的5次行为，这样可以更快的捕获用户的实时兴趣变化，对用户兴趣变迁做出快速反应；

（2）帖子向量预训练：帖子向量使用item2Vec进行预训练，一方面可以加快模型收敛速度，提升训练效率，另一方面能有效提高帖子的embedding结果；

（3）修改优化目标：原先相关性模型的优化目标是相关性，损失函数为余弦相似度，为了使得embedding结果能针对业务目标，我们修改损失函数为交叉熵，同时调整样本的生成方式，使得用户向量和帖子向量的生成奔着点击率/转化率提升的目标去调整；

（4）多样性增加：粗排因为是为精排提供候选，除了考虑相关性外，还需要考虑多样性，这里我们增加了一些其他维度的业务因子跟相关性因子融合，能增加一些多样性；

（5）用户向量生成：除了使用dssm双塔结构进行用户向量的生成，同时也尝试了使用帖子embedding的结果进行avg或者sum进行用户向量的生成，通过线上实验对比，发现使用dssm直接训练得出的用户向量效果要更好一些。

在粗排上的尝试迭代至今，获得了二手车商业产品单用户连接率5%的提升，目前还在持续进行优化中。

总结和展望

结合二手车商业变现业务场景，我们在广告的召回、粗排、精排三个环节，较好的落地了对embedding技术的应用，并针对具体情况进行了针对性的优化和调整，提升了召回和rank质量，达到了用户体验和收入共赢的目标。后续将会考虑使用带时序信息或者基于图结构的embedding，使向量能够表征更多信息，不局限在文字、用户、帖子、图片。Embedding技术能力将作为一种问题解决思路，应用到未来可能遇到的问题上，万物皆可Embedding。

参考文献

1. Antonellis I, Molina H G, Chang C C. Simrank++: query rewriting through link analysis of the click graph[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 408-421.

2. 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. ACM, 2018: 311-320.

3. Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013: 2333-2338.

4. 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.

作者简介

刘杨，商业产品技术部-策略技术团队算法资深开发工程师；

刘笠熙，商业产品技术部-策略技术团队算法架构师；

END

阅读推荐

1. 开源 | dl_inference：通用深度学习推理服务

2. 开源 | LPA-Detector：基于GraphX 的LPA算法改进

3. 开源 | Zucker：Android APP模块化大小自动分析统计工具

4. 开源 | WBBlades：基于Mach-O文件解析的APP分析工具

5. 开源 | wwto：小程序跨端迁移解决方案——微信转其他小程序