DSSM召回模型在小米收音机个性化业务的应用

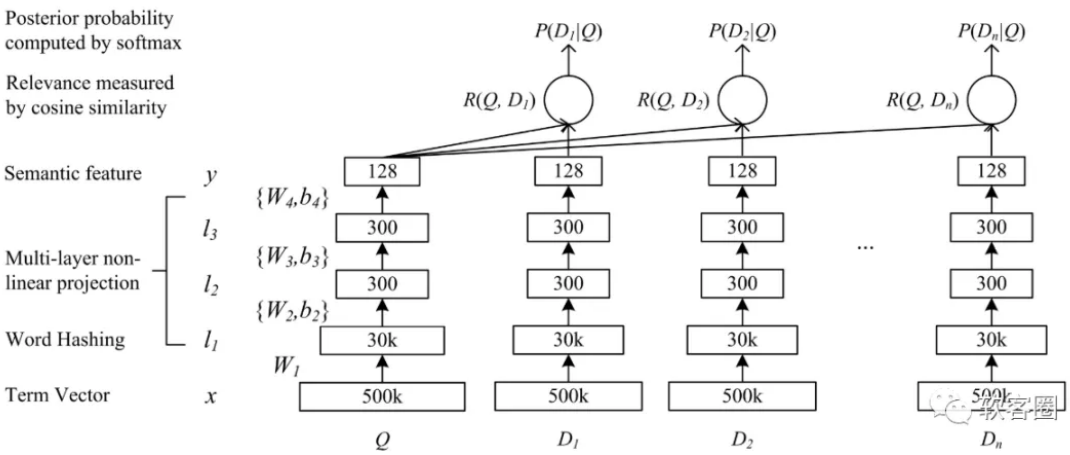
原创 mrchor 软客圈 2020-09-14

前言

召回作为推荐系统的一部分，为个性化业务在海量数据中的粗选提供了很好的支持。传统的召回大多是基于策略计算的召回结果，例如，根据用户的分类偏好去召回相应分类下的优质结果。然而，这种召回方式从某种角度上讲，对于用户来说具有一定的趋同性，即我们可能更容易把热门的内容召回给用户，这可能会破坏个性化召回的多样性，也使得我们的业务越推荐范围越窄，不利于平台内容的分发。为提高内容的分发效率以及满足用户个性化需求，小米收音机业务尝试使用召回模型为业务的召回阶段保驾护航！

本期我们首先介绍一种双塔模型——Deep Structured Semantic Models（DSSM）[1]，深度结构语义模型。

原理



DSSM模型最初来源于微软（Microsoft）在搜索业务的应用，其实际的原理很简单，即将用户搜索词（query）与文档（document）先进行多层全连接（MLP），然后将他们转化为同一尺度的embedding，从而映射到同一语义空间，然后再根据距离公式（这里使用的是余弦相似度）计算搜索词与

文档的相关性，从而得出用户搜索的展示结果。那么为什么DSSM模型能很好地在搜索业务应用呢，这是因为，大量用户在使用搜索引擎时，其输入关键字以及结果页点击行为，会自动地为我们的模型训练导出一份优秀的样本数据，即搜索词与结果的匹配以便模型的训练。
而在原文中训练集的构建也十分简单，例如一个搜索词下对应15个关联文档，那么可以根据用户日志，聚合这15个文档的点击统计，然后使用softmax，去计算后验概率：

$$\mathrm{P}(\{D_1, D_2, \cdots, D_{15}\} \mid Query)$$

DSSM模型在小米收音机业务的应用

为什么DSSM在个性化业务有效?

前面先简单叙述了DSSM模型的原理，我相信算法相关岗位的同学对于原理是很好理解的，但是之前的应用只在搜索业务上，而我们本期讲到的是在个性化推荐业务的应用，那为什么DSSM模型能在个性化召回中有效呢？这个问题，我们分以下几个方面解释：

1、本质上都是匹配问题

a)搜索业务的本质就是根据用户的搜索词去匹配跟搜索词（或搜索意图）最为匹配的文档；

b)个性化召回就是在海量数据中为用户圈定大概会感兴趣的内容，而这种问题可以转化为与用户最匹配的内容的search。

2、DSSM中不存在query与doc的“feature interaction”，便于模型拆解

综上所述，DSSM可以在个性化业务中充当召回角色。

DSSM在小米收音机个性化业务中的改造

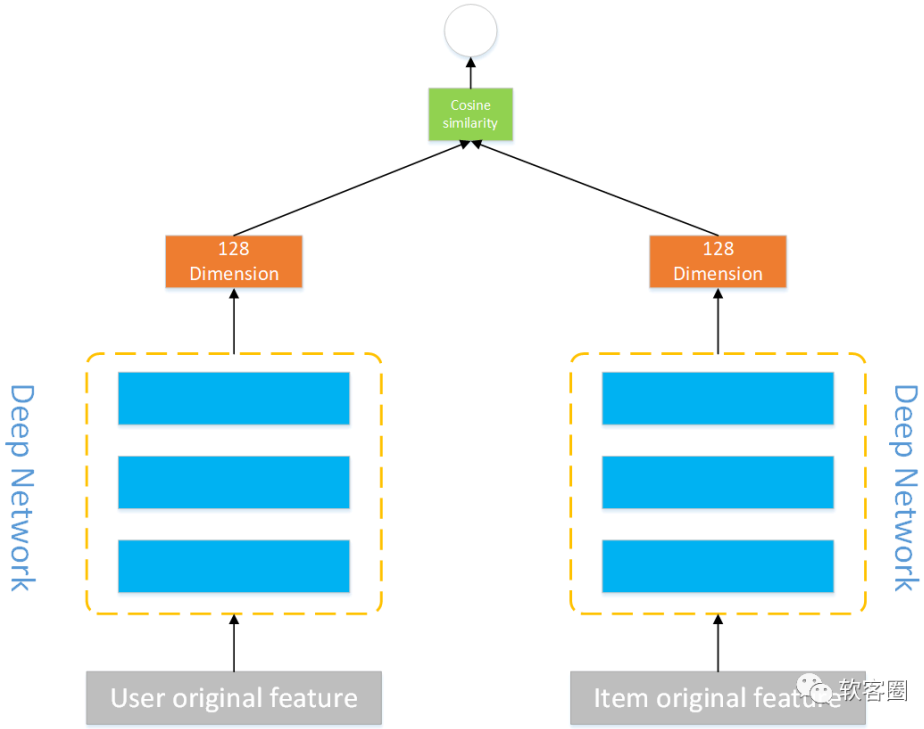
经过上一小节的介绍，虽然DSSM模型可以应用到召回中，但是相比于搜索业务还有需要需要改造的地方，本小节我们通过对DSSM的改造，使其符合召回模型的要求。

模型架构改造

在搜索中，DSSM主要是为了搜索词与文档的匹配，所以模型构造成一个softmax的概率分布，来确定最终文档的排序。而在个性化业务，我们实际上是为了扩大用户点击内容的概率，因此我们需要以CTR模型的思维去对原生DSSM模型进行改造：

1、多文档匹配改为单item，变成CTR模型架构；

2、Softmax的后验概率改为余弦相似度计算，与CTR问题做对应。



样本构造策略修改

召回模型≠CTR模型

在样本构造中，初始时，由于我很少做召回的模型。因此，对于DSSM模型，我按照排序的思路去求解这个问题，因此对于样本构造，等同于CTR模型的样本，然后进行了一版模型的迭代，结果发现，为我召回出来的数据完全不是我喜欢的内容，甚至有一些召回内容是风马牛不相及的，简直差到了极点。后来我查找了很多双塔模型关于召回的样本构造策略，发现大多都是基于随机采样做的负样本构造，于是我们重新对样本数据进行构造，发现效果好了很多，那么产生这种结果是为什么呢？这其实涉及到两个误区[2]：

1、**错误认定负样本：我们一开始使用CTR思维构造样本时，模型其实是将曝光未点击的内容作为负样本给出，然后这部分样本其实是经过之前的排序模型给出的，也即他们只是用户相对于点击的内容不太喜欢的，而相对于未曝光的内容其实是更容易被用户点击的，因此，这部分内容我们不能将其视为负样本给出数据。**
2、**冰山理论：由于我们给出的负样本是曝光未点击，可以认为就像冰山理论一样，我们只是给模型看到了冰山的一角，但是想让模型画出整个冰山，这其实是不现实的，因此我们需要让模型看到冰山海面下的部分，这就是我们需要以随机采样作为负样本采集的策略的原因。**

后期优化改造

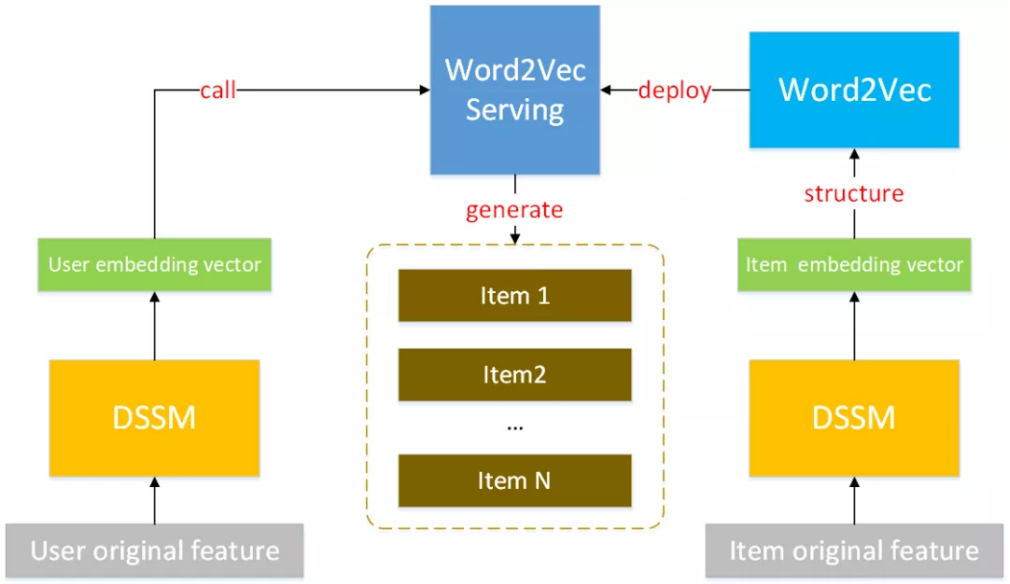
对于召回模型来说，随机采样具有一定的概率可能使用户喜欢的内容被划入到负样本中，这其实是业务无法容忍的，为了减轻这种情况的影响，再加之我们上文中对曝光未点击内容的描述，我们将DSSM的二分类问题做成一个回归问题，即我们认为曝光未点击内容作为CTR模型筛选的结果，可以认定为用户潜在喜欢的内容，因此，我们对把样本的标签设置为：点击-1、曝光-0.5（**更可以根据用户点击内容的相对位置，动态设置曝光内容的分值**）、随机采样-0。
在经过这一步优化后，DSSM模型的召回结果会有一个更好的提升。

DSSM召回的线上部署(DSSM on Word2Vec)

根据了解，业界在模型召回的问题上几乎都采用了Facebook提出的向量索引框架——Faiss，但这对于业务的召回部署带来一定的挑战，需要增加一些服务计算成本。在小米收音机业务中，我们采用了另外一种更为容易轻量的部署方式。
DSSM召回中，我们是根据用户转换后的embedding向量去匹配计算最相近的内容的距离，这与word2vec模型计算最相似word的初衷是一致的。因此我们借助了word2vec模型的哈夫曼搜索树的方式计算用户的召回内容，我们使用word2vec on spark的API做了上述处理：

```
val map: Map[String, Array[Float]] = paidAlbumVector
  .collect()
  .toMap
val word2VecModel = new Word2VecModel(map)
word2VecModel.save(sc, args(4))
```


以下是具体的DSSM on word2vec召回流程图：



DSSM on Word2Vec 软客圈

总结

召回模型在模型复杂度部分是比较简单的，但是想要做好却是比较困难的，很容易陷入致命问题。本文简单介绍了DSSM模型在小米收音机业务的召回使用经验，期待我们的经验能抛砖引入，为你的业务带来提升。如果你有更好的想法，可以私聊我一起交流~

参考文献

[1] P. Sen Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 2333–2338, 2013, doi: 10.1145/2505515.2505665.

[2] "负样本为王：评Facebook的向量化召回算法," pp. 1–9, 2020.

如果你喜欢我的文章，欢迎关注我的微信公众号【软客圈】（ID：recoquan）

纯手工打造，实属不易，欢迎大家分享和转发~

原创内容，转载需注明出处，否则视为侵权并将被追诉！