第五部分：用户行为序列建模

1、简单平均

2、DIN模型

3、SIM模型

1、简单平均

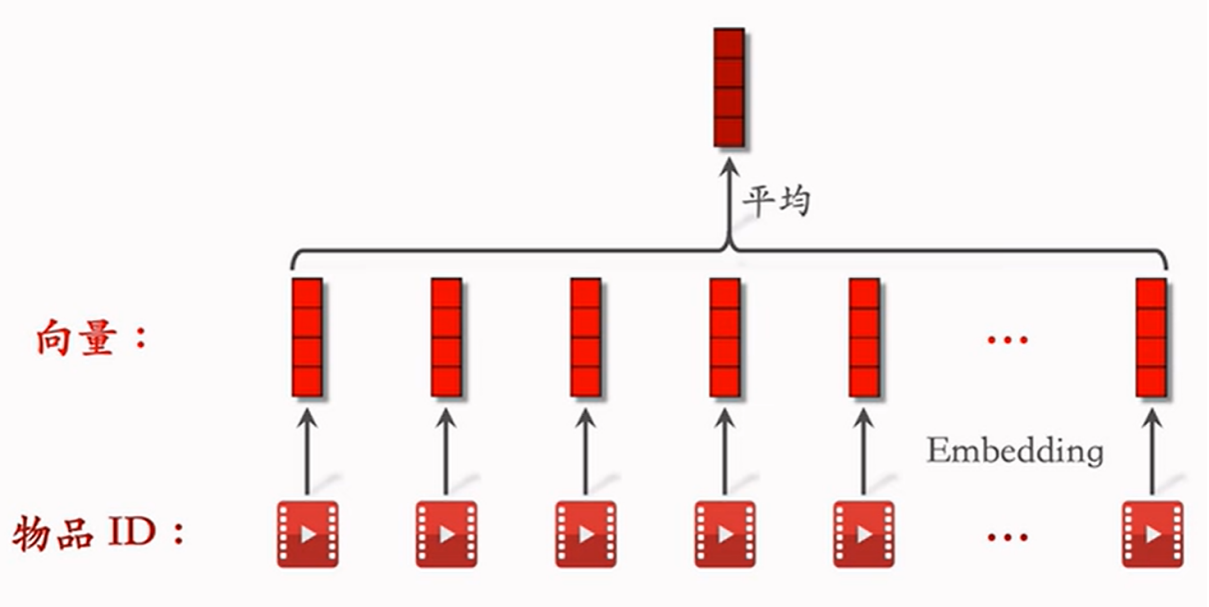
LastN特征

定义：用户最近n次交互（点击、点赞）过的物品ID

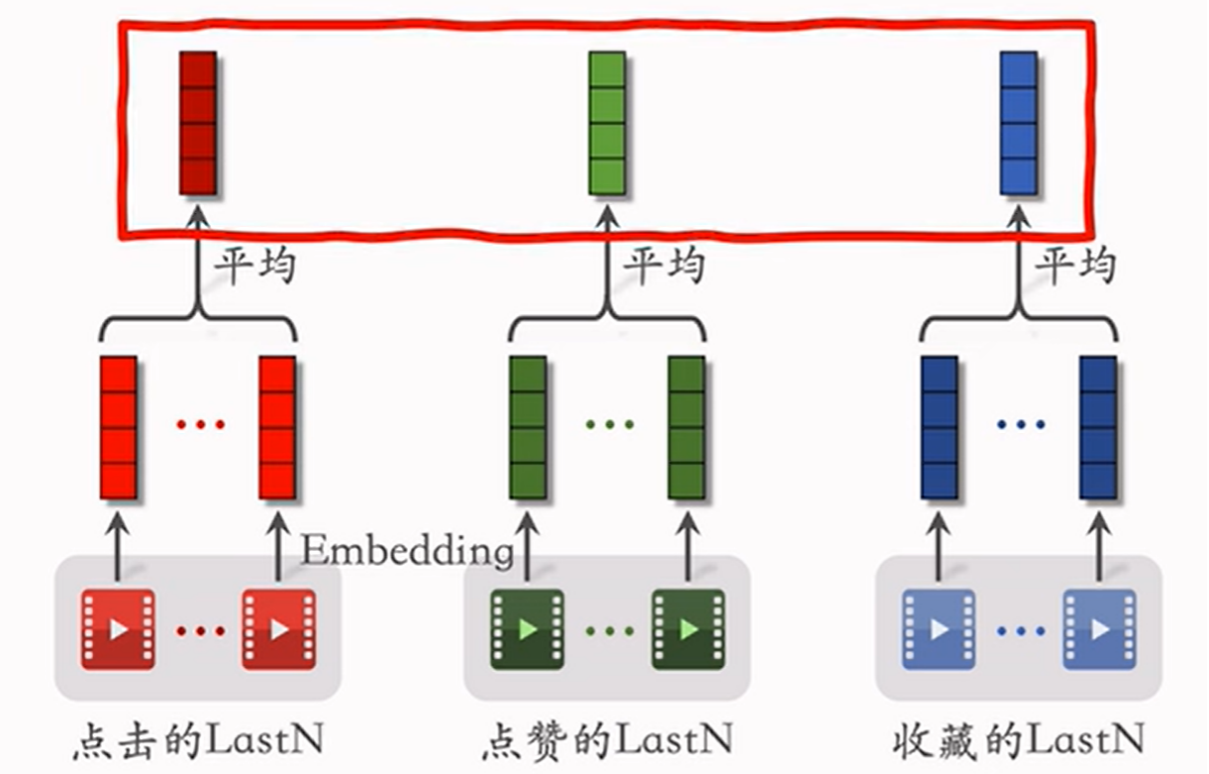
操作：对LastN物品做ID embedding，得到n个向量，再对向量取平均表示用户特征，反映用户曾经对什么样的物品感兴趣

适用模型：召回双塔模型、粗排三塔模型、精排模型

图例：



完整模型（小红书的实践）



得到的几个平均向量进行串接，作为用户特征用于召回、排序模型

除此之外物品还有其他特征（如类目）进行embedding，和ID embedding拼接在一起，比单纯ID的embedding效果更好

2、DIN模型

DIN内容

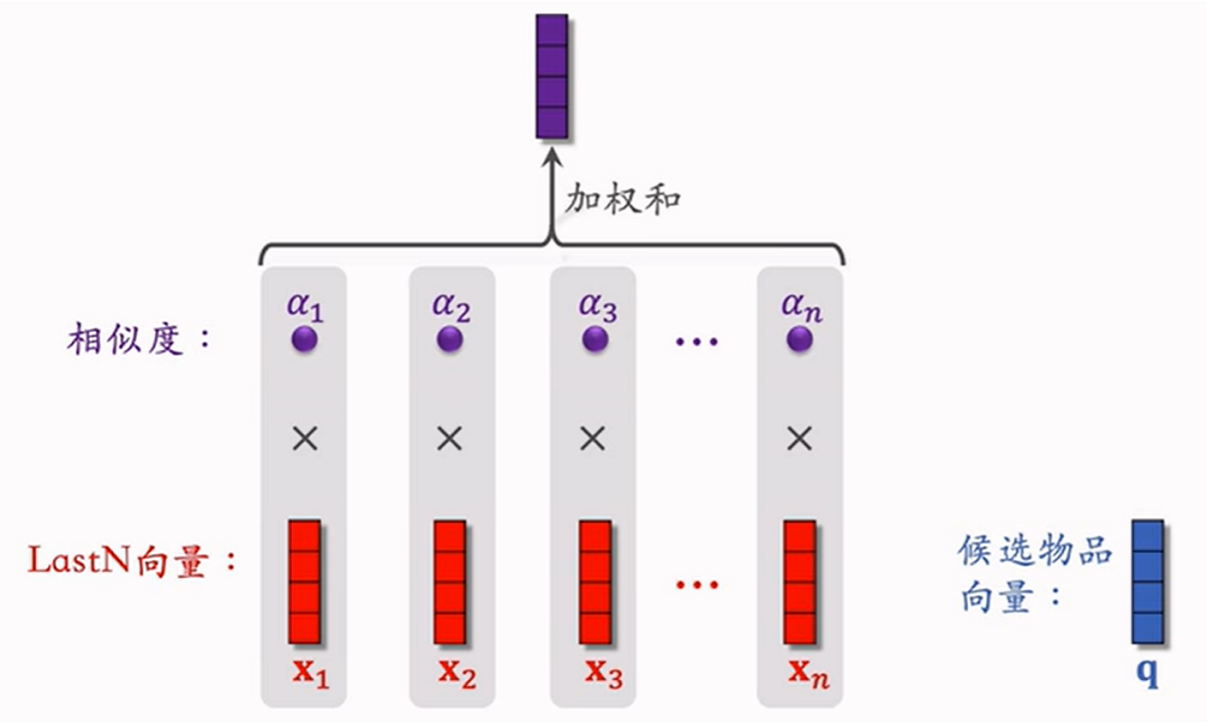
DIN用加权平均代替平均，即注意力机制（attention）

权重：候选物品与用户LastN物品的相似度

候选物品的解释：如粗排500个物品就作为精排的候选物品，精排模型要对每个候选物品打分，反映用户对候选物品的兴趣，然后将这些候选物品打分排序，选择分数最高的返回给用户

*注意区分候选物品和LastN物品的区别，前者是参与某过程（如精排）打分的物品，后者是用户最近交互的物品

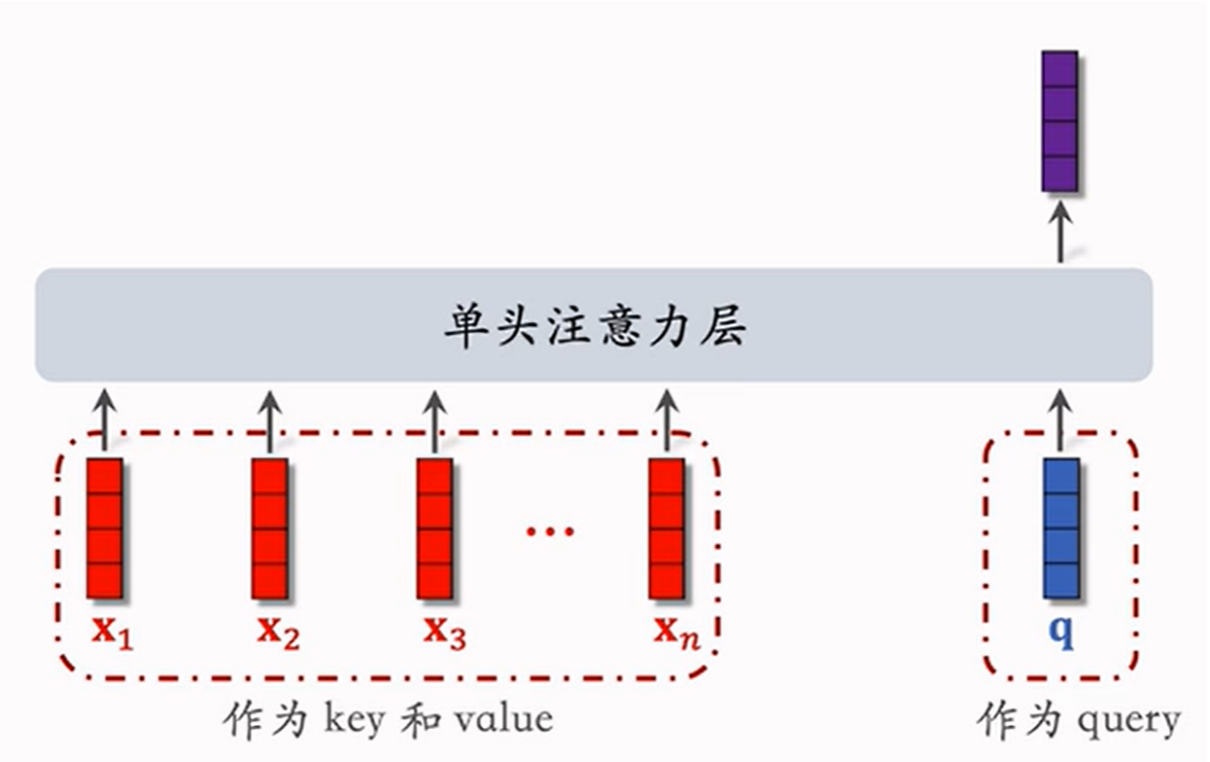
图例：



相似度是LastN向量分别与候选物品向量计算的结果（内积、余弦相似度等方式）

加权得到的向量作为用户表征输入排序模型，预估（用户，候选物品）的点击率、点赞率等指标

本质：注意力机制



简单平均 VS 注意力机制

相同

简单平均和注意力机制都适用于精排模型

不同

（1）简单平均适用于双塔模型、三塔模型

*简单平均只需要用到LastN，属于用户自身特征，把LastN向量的平均作为用户塔的输入

（2）注意力机制不适用于双塔模型、三塔模型

*注意力机制需要用到LastN+候选物品，然而用户塔看不到候选物品，不能把注意力机制用在用户塔

3、SIM模型

DIN模型的缺点

缺点

1、注意力层计算量正比于n（用户行为序列长度，即LastN交互物品数量），因此只能记录最近几百个物品，否则计算量过大

2、关注短期兴趣，遗忘长期兴趣

改进DIN具体内容

目标：保留用户长期行为序列（n很大），但是计算量不会过大

方法：

DIN对LastN向量做加权平均，权重是相似度

如果某LastN物品与候选物品差异很大，则权重接近零

把这些与候选物品差异很大，几乎没有关系的LastN快速排除，降低注意力层计算量

SIM模型的内容

目标

SIM模型是对DIN模型的改进，它能保留用户长期兴趣

步骤

1、保留用户长期行为记录，n的大小可以是几千

2、对于每个候选物品，在用户LastN记录中做快速查找，找到k个相似物品

3、把LastN变成TopK，然后输入到注意力层

4、用这种方式，SIM模型减小了计算量（从n降到k）

步骤解析

1、查找

方法一：hard search

根据候选物品的类目，保留LastN物品中类目相同的，这种方式简单快速，无需训练

方法二：soft search

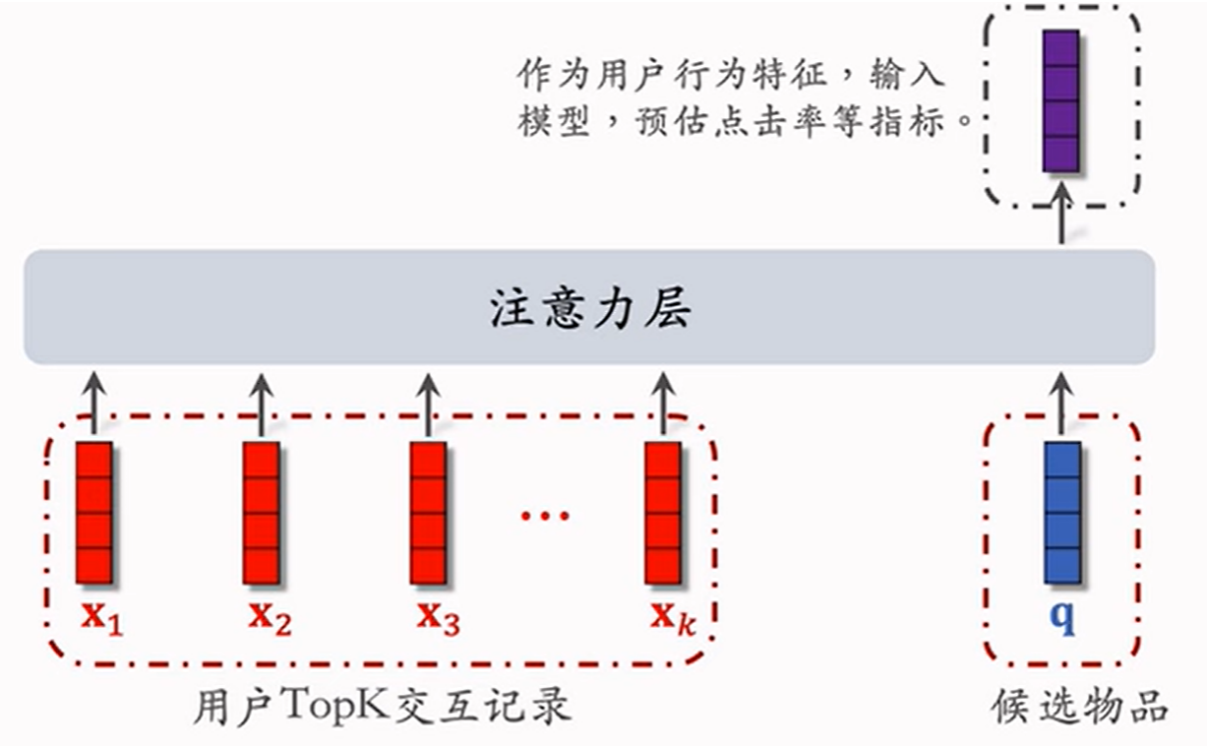
把物品做embedding变成向量，把候选物品向量作为query，做k近邻查找，保留LastN物品中最接近的k个，这种方法效果更好，编程实现更复杂，预估指标auc更高

补充：关于auc可以看机器学习（三十五）— AUC 原理及计算方式 - 深度机器学习 - 博客园

*soft search和hard search的选择取决于公司的工程基建

2、注意力机制

本质和DIN没有区别，只是LastN向量变成了TopK向量



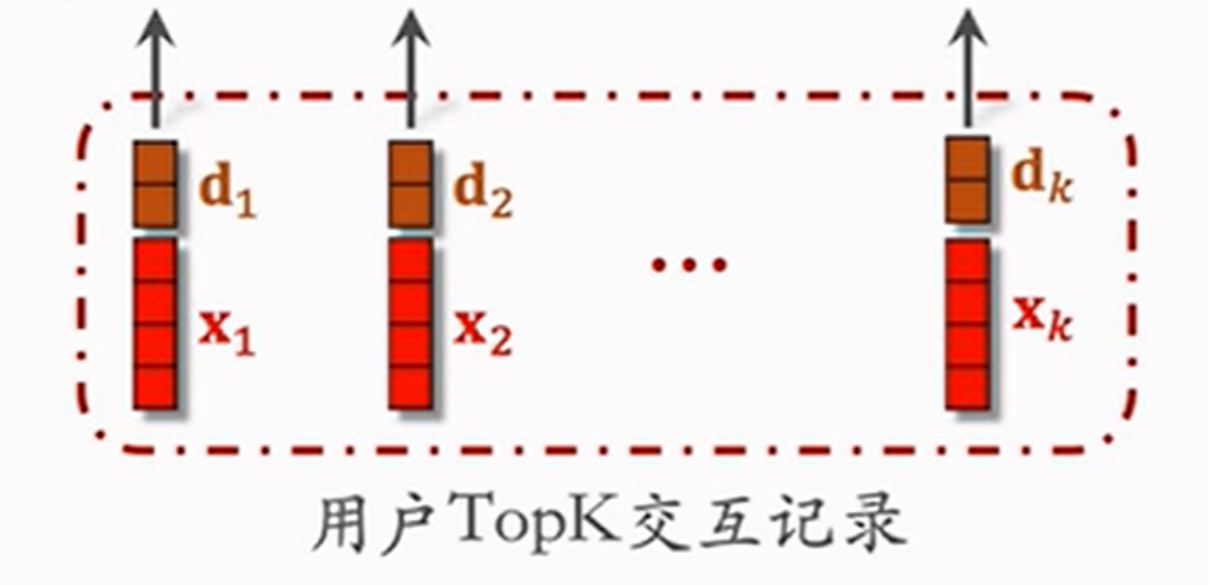
技巧：使用时间信息

（SIM序列长，这样便于记录用户长期行为）

用户与某个LastN物品的交互时刻距今为δ

对δ做离散化，再做embedding，变成向量d

把两个向量x（物品embedding）和d（时间embedding）做串接，表征一个LastN物品



*实践表明采用时间信息对系统性能有提升