DIN模型缺点：

注意力层的计算量正比于n，n越大，计算量越大

通常只记录用户最近几百个物品，否则计算量太大

所以DIN只关注短期兴趣，忽略长期兴趣。（暴力增加序列长度会提升指标，但是提升的同时会带来高额的计算量导致系统响应变慢，得不偿失）

如何改进呢？

我们目标是保留用户的长期序列（n很大），但是不希望计算量过大。

DIN对lastN向量作加权平均，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

快速排除与候选物品无关的lastN物品，降低注意力层的计算量

SIM模型:

保留用户的长期行为记录，n的大小可以是几千

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。

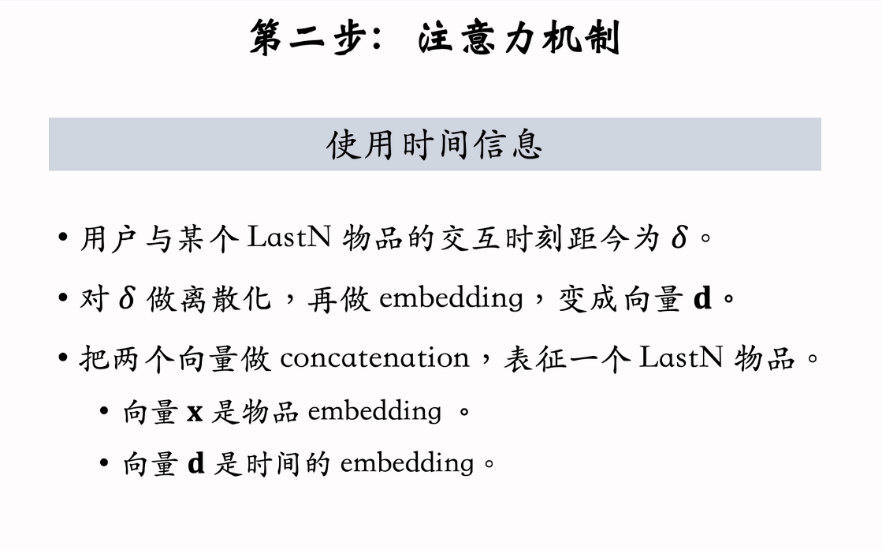
把LastN变成TopK，然后输入到注意力层。

SIM模型减少计算量（从n到K）

实验表明用hard search

图片包含 图表

描述已自动生成



图示

低可信度描述已自动生成

为什么SIM模型用到时间信息？

DIN用到的lastN物品都是近期的交互物品，时间差别不大。加上时间信息冗余了。

SIM是考虑了长期行为信息，时间差别大，时间信息对于这种长期行为也是一种很有效的补充信息。时间越久重要性越低

