4.6 局部敏感哈希——让Embedding插上翅膀的快速搜索方法

Embedding最重要的用法之一是作为推荐系统的召回层，解决相似物品的召回问题。推荐系统召回层的主要功能是快速地将待推荐物品的候选集从十万、百万量级的规模减小到几千甚至几百量级的规模，避免将全部候选物品直接输入深度学习模型造成的计算资源浪费和预测延迟问题。

Embedding技术凭借其能够综合多种信息和特征的能力，相比传统的基于规则的召回方法，更适于解决推荐系统的召回问题。在实际工程中，能否应用Embedding的关键就在于能否使用Embedding技术“快速”处理几十万甚至上百万候选集，避免增大整个推荐系统的响应延迟。

4.6.1 “快速”Embedding最近邻搜索

传统的Embedding相似度的计算方法是Embedding向量间的内积运算，这就意味着为了筛选某个用户的候选物品，需要对候选集合中的所有物品进行遍历。在k维的Embedding空间中，物品总数为n，那么遍历计算用户和物品向量相似度的时间复杂度是 $O(kn)$。在物品总数 n 动辄达到几百万量级的推荐系统中，这样的时间复杂度是承受不了的，会导致线上模型服务过程的巨大延迟。

换一个角度思考这个问题。由于用户和物品的Embedding同处于一个向量空间内，所以召回与用户向量最相似的物品Embedding向量的过程其实是一个在向量空间内搜索最近邻的过程。如果能够找到高维空间快速搜索最近邻点的方法，那么相似Embedding的快速搜索问题就迎刃而解了。

通过建立 kd（k-dimension）树索引结构进行最近邻搜索是常用的快速最近邻搜索方法，时间复杂度可以降低到$O(\log_2 n)$。一方面，kd树的结构较复杂，而且在进行最近邻搜索时往往还要进行回溯，确保最近邻的结果，导致搜索过程较复杂；另一方面，$O(\log_2 n)$的时间复杂度并不是完全理想的状态。那么，有没有时间复杂度更低，操作更简便