浅析Faiss在推荐系统中的应用及原理

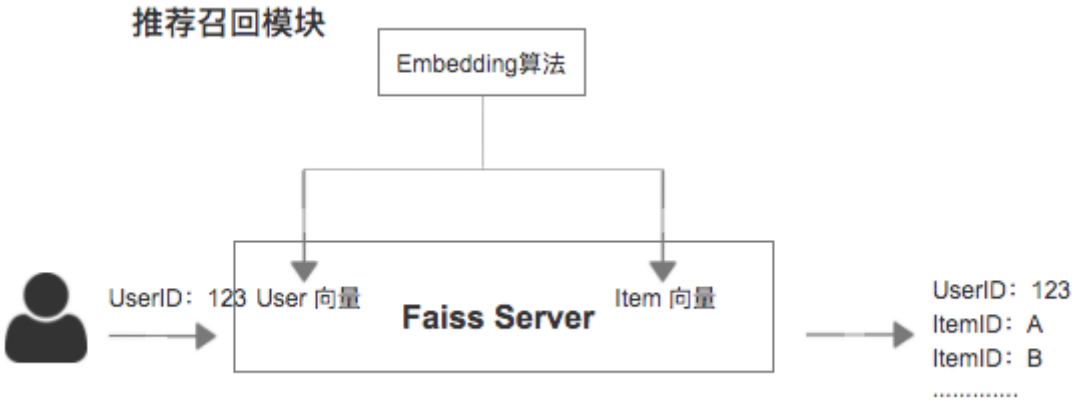
凡人机器学习 5月5日

之前在业务中应用了许多Faiss，也看了几篇关于Faiss的论文，简单记录下Faiss的一些属性和应用。Faiss是Facebook的AI团队开源的一套用于做聚类或者相似性搜索的软件库，底层是用C++实现。Faiss因为超级优越的性能，被广泛应用于推荐相关的业务当中。接下来分Faiss在推荐业务应用和Faiss的基本原理两部分进行介绍。

1

Faiss在推荐业务中的应用

在我的认知里，基本上50%以上的手机APP的推荐业务会应用到Faiss服务，可见应用之广。那Faiss究竟是在哪个模块使用呢，通过下方这个图给大家介绍：



大家都知道推荐业务包含排序和召回两个模块，Faiss比较多的应用在召回模块。召回业务中有很多是向量生成类的算法，比如Graph Embedding、ALS Embedding、FM Embedding等。ALS就是经典的矩阵分解算法，它可以将User和Item的行为数据利用矩阵分解的方式生成User向量和Item向量，这些向量分别代表User和Item的属性（工科研究生矩阵论课程学过矩阵分解，不懂的同学要补课了）。

当我们拿到了User和Item的向量，只要计算出哪些Item和User的向量距离较短（最简单的解法是算欧式距离），就可以得出User偏爱的Item。但是当User和Item的数量巨大的时候，设想下某短视频平台，每天有上百万User登录，有存量的上千万的Item短视频，怎么能快速的计算出向量距离，就成了一个亟待解决的技术难点，因为推荐业务的召回模块需要在50ms以内拿到结果。这也就是Faiss的价值所在，Faiss几乎可以在10ms内完成百万*百万以上的向量距离计算，它是怎么实现的呢？

2

Faiss原理

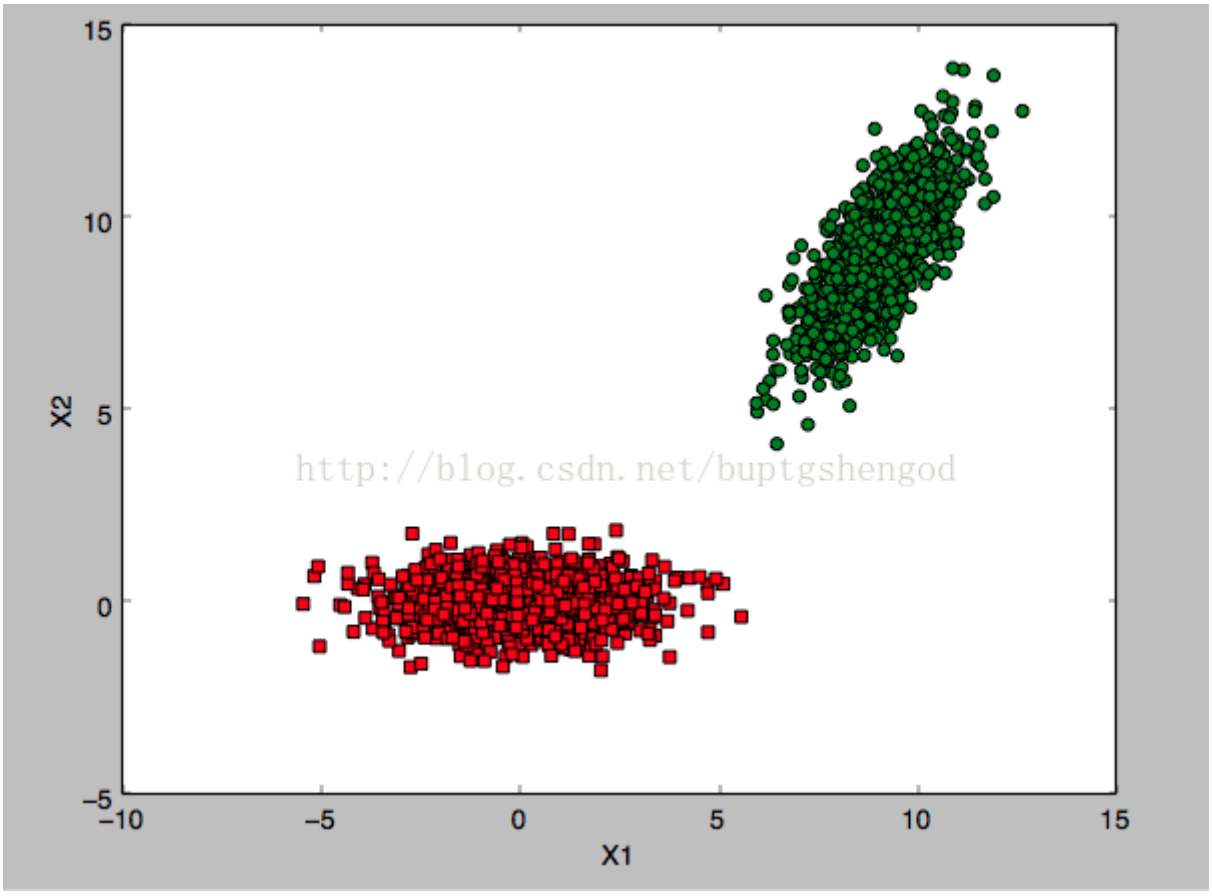
向量计算是一个最经典的时空优化问题，在查询过程中建立更多的索引固然可以提升查询速度，但是却有占据了存储空间，我们希望系统可以即减少索引又能提升查询性能。

为了得到时间和空间的最优，Faiss使用了PCA和PQ两个手段进行向量压缩和编码，当然还有其它的一些优化手段，但是PCA和PQ是最为核心的。

PCA降维

PCA是一种降维手段，简单理解就是将高维向量变为低维，这样就可以有效的节省存储空间，PCA我之前介绍过，今天就不多说了。有兴趣可以看下我的博客：
我的博客-PCA

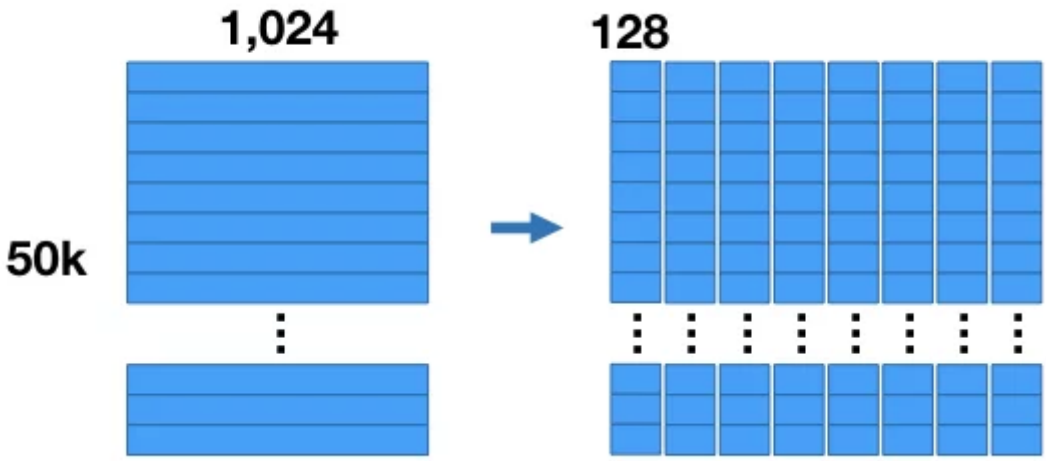
大家看下图绿色的点，它其实是二维的，既有纵向坐标的属性也有横向坐标的属性，可以用PCA方式让它变为一维，这样就成了红色这样的点簇，这就是PCA的价值，通过压低维度降低向量存储空间。



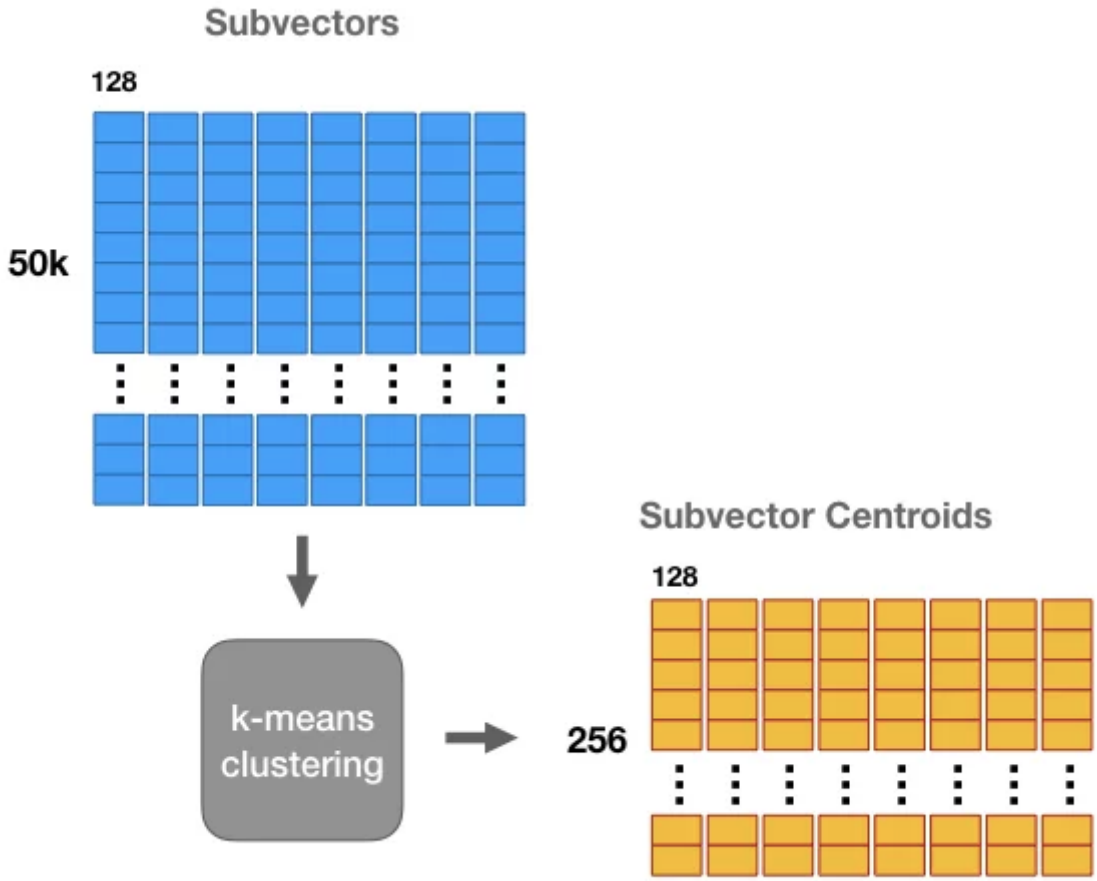
PQ编码

Product quantization(乘积量化PQ)，PQ是一种建立索引的方式。这里参考这篇文章为大家说明：http://www.fabwrite.com/productquantization

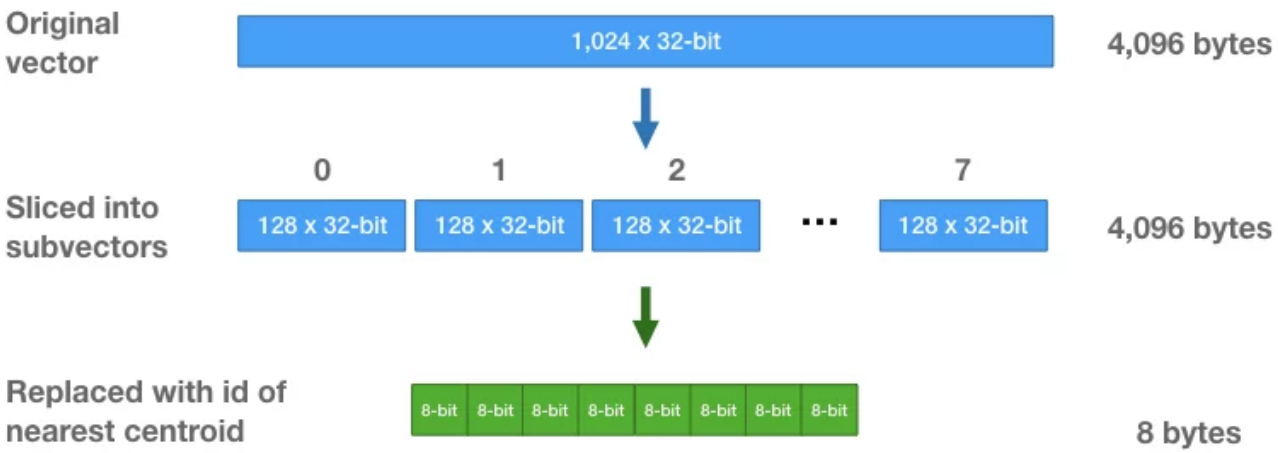
假设原始向量是1024维，可以把它拆解成8个子向量，每个子向量128维。



然后对每个字向量的全部50k数据分别作Kmeans计算，假设设置Kmeans的K为256。就得到了8组，每组256个中心点这样的码本，这个码本可以对50k个向量进行编码。



也就是说把编码从原始的1024维向量表示需要10bit，压缩成了只需要log(256)，8bit来表示。这样每个向量的索引就减少了许多。



参考文档（衷心感谢以下老师们的贡献）：

（1）http://www.fabwrite.com/productquantization

（2）https://github.com/facebookresearch/faiss