向量检索-大规模向量检索引擎

刘作程

2020-01-15

Outline

- 什么是向量检索?
- 向量检索有哪些算法?
- 如何实现向量检索引擎?
- 举例向量检索引擎的落地场景-拼多多拍照搜商品
- 目前向量检索有哪些问题?

什么是向量检索?

- 与普通检索的不同？ 针对向量**相似性检索/近邻搜索**

- 计算检索向量和样本向量的空间距离，选取与检索向量中距离相近的样本向量作为检索结果

$$A=[\alpha_1,\alpha_2,\ldots,\alpha_n].or.A=\begin{bmatrix}\alpha_1\\ \alpha_2\\ \ldots\\ \alpha_n\end{bmatrix}$$

向量空间距离

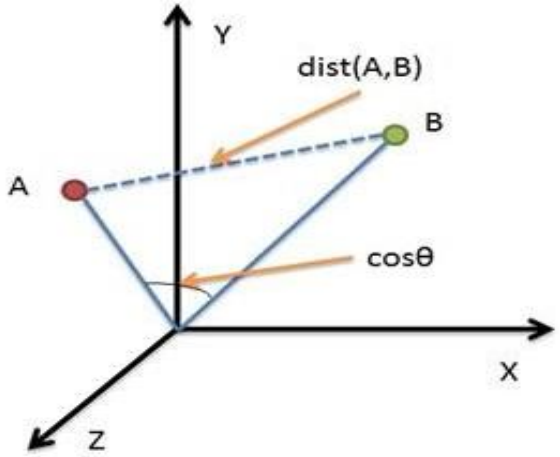
- 欧式距离
- 余弦距离
- 曼哈顿距离
- 切比雪夫距离
- 汉明距离

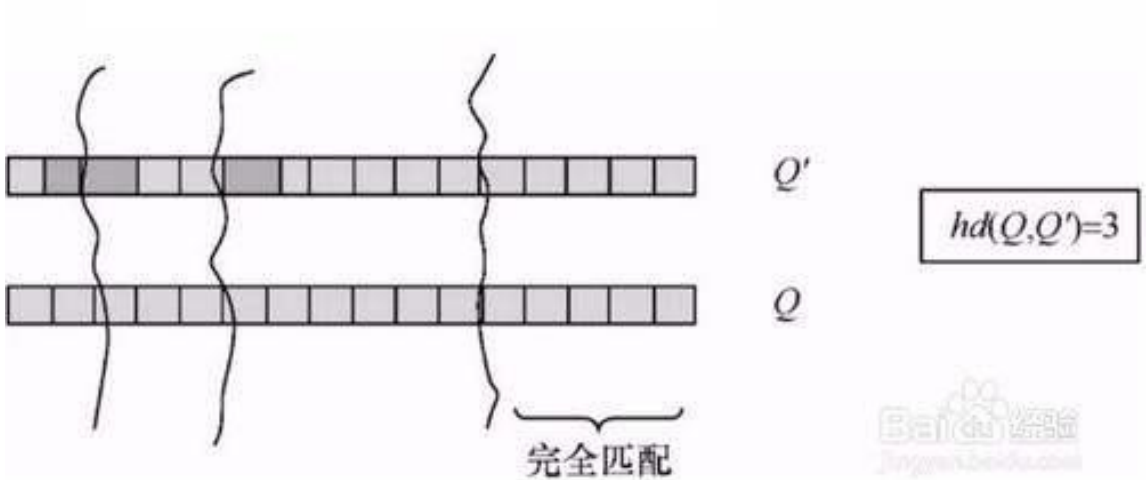
$$dist(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

$$sim(X,Y)=cos\theta=\frac{\vec{x}\cdot\vec{y}}{\|x\|\cdot\|y\|}$$

$$c=|x_1-x_2|+|y_1-y_2|$$

$$d_{ab}=max(|x_{1i}-x_{2i}|)$$





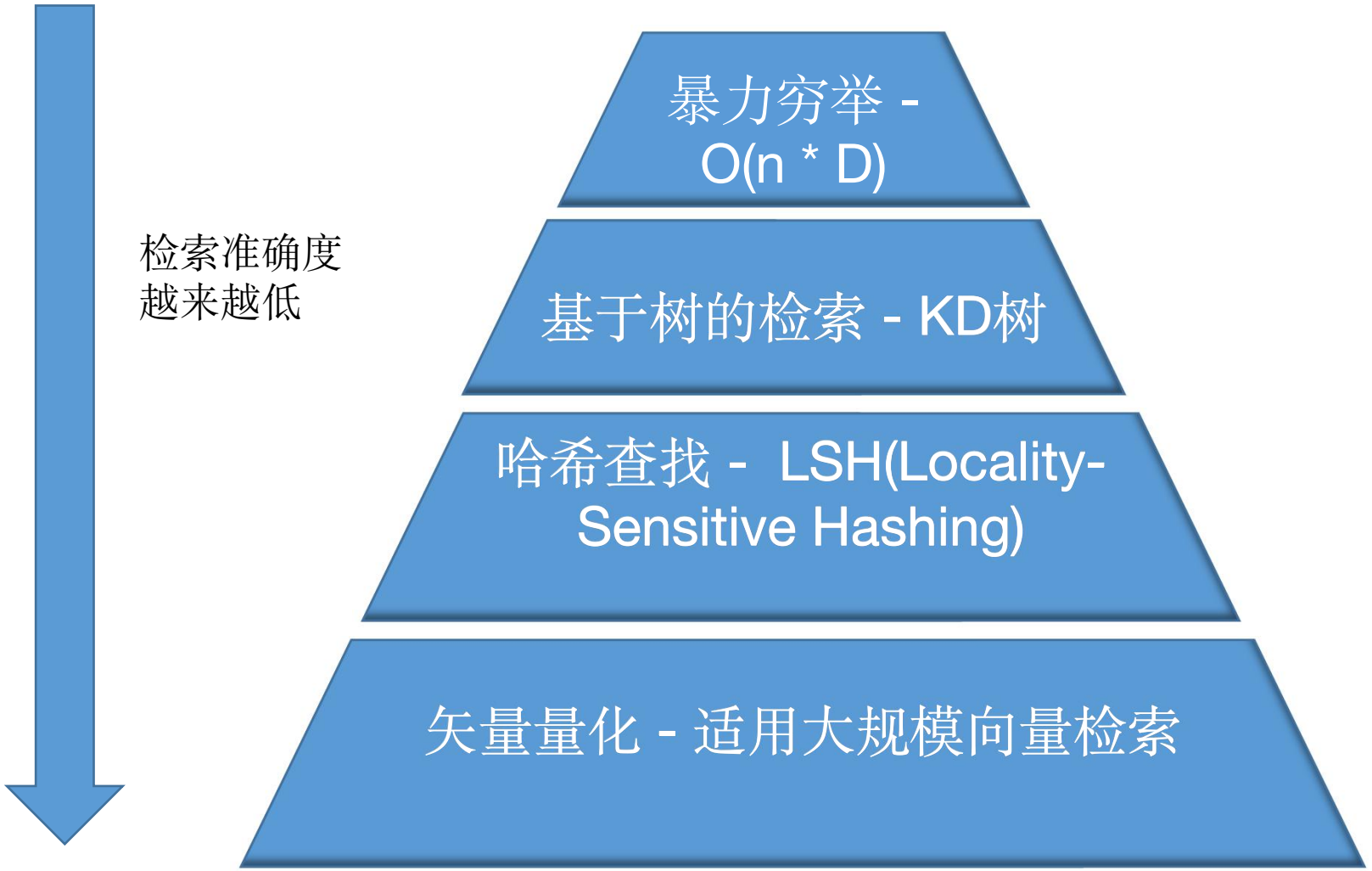
向量检索的使用领域和落地场景

- 图像检索 - 以图搜图（拍图搜商品）
- 视频检索 - 以视频搜视频 （版权判定、视频去重、视频打标）
- 推荐系统 - 依据用户特征、上下文检索 （电商推荐、广告投放）
- NLP - 文本相似度检索 （文本过滤、去重）

向量检索算法

数据规模越来越大

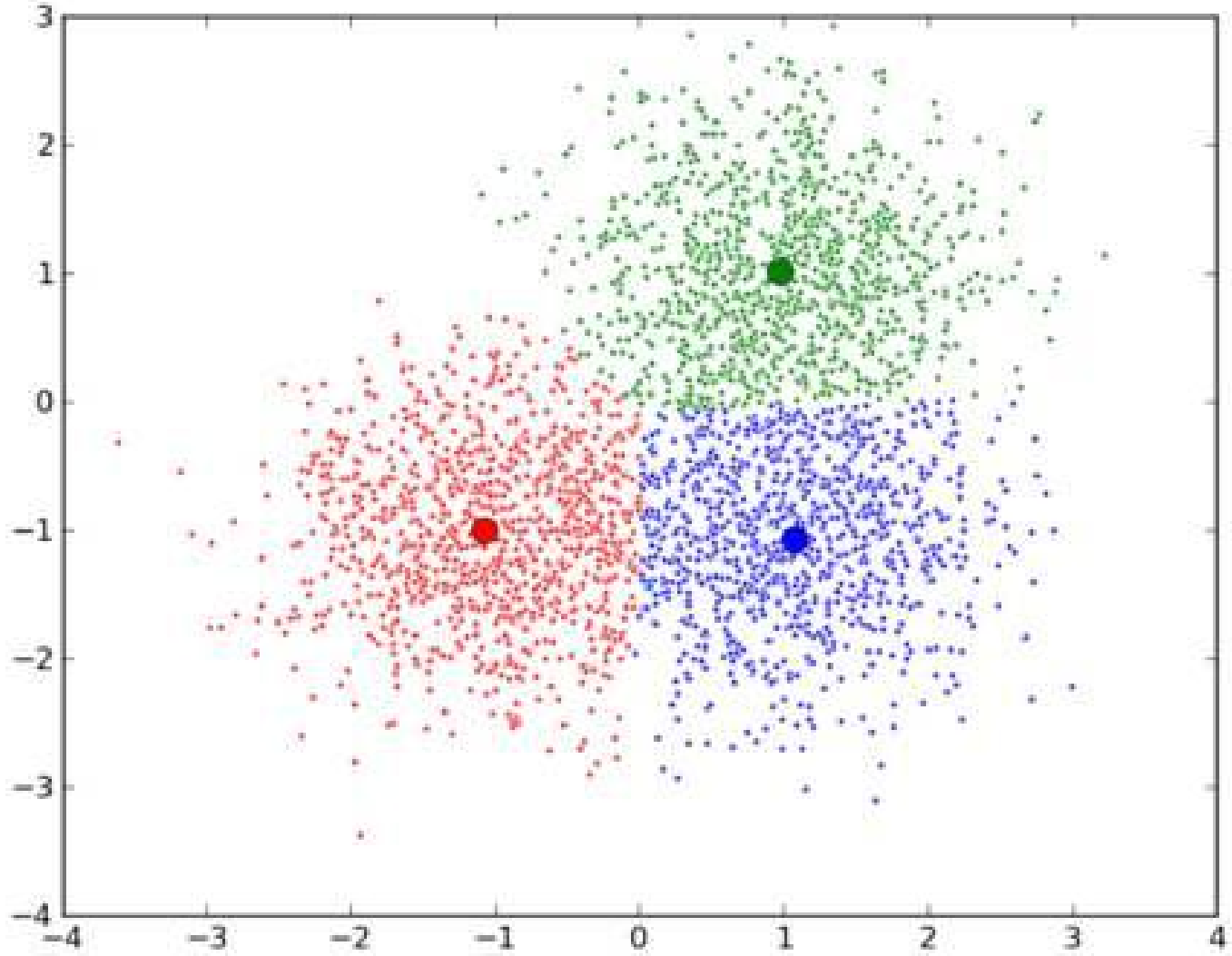
检索准确度越来越低



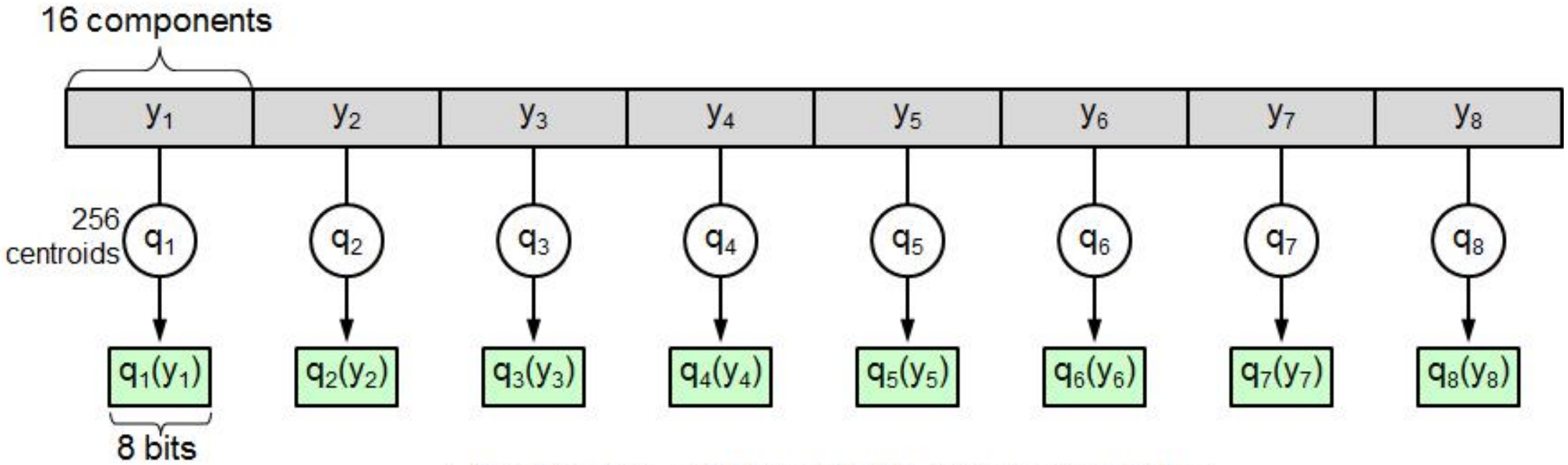
大规模向量检索的挑战

- 高维向量，512维、4096维
- 数据量大，1亿条以上，数据存储在3T以上

Kmeans 聚类



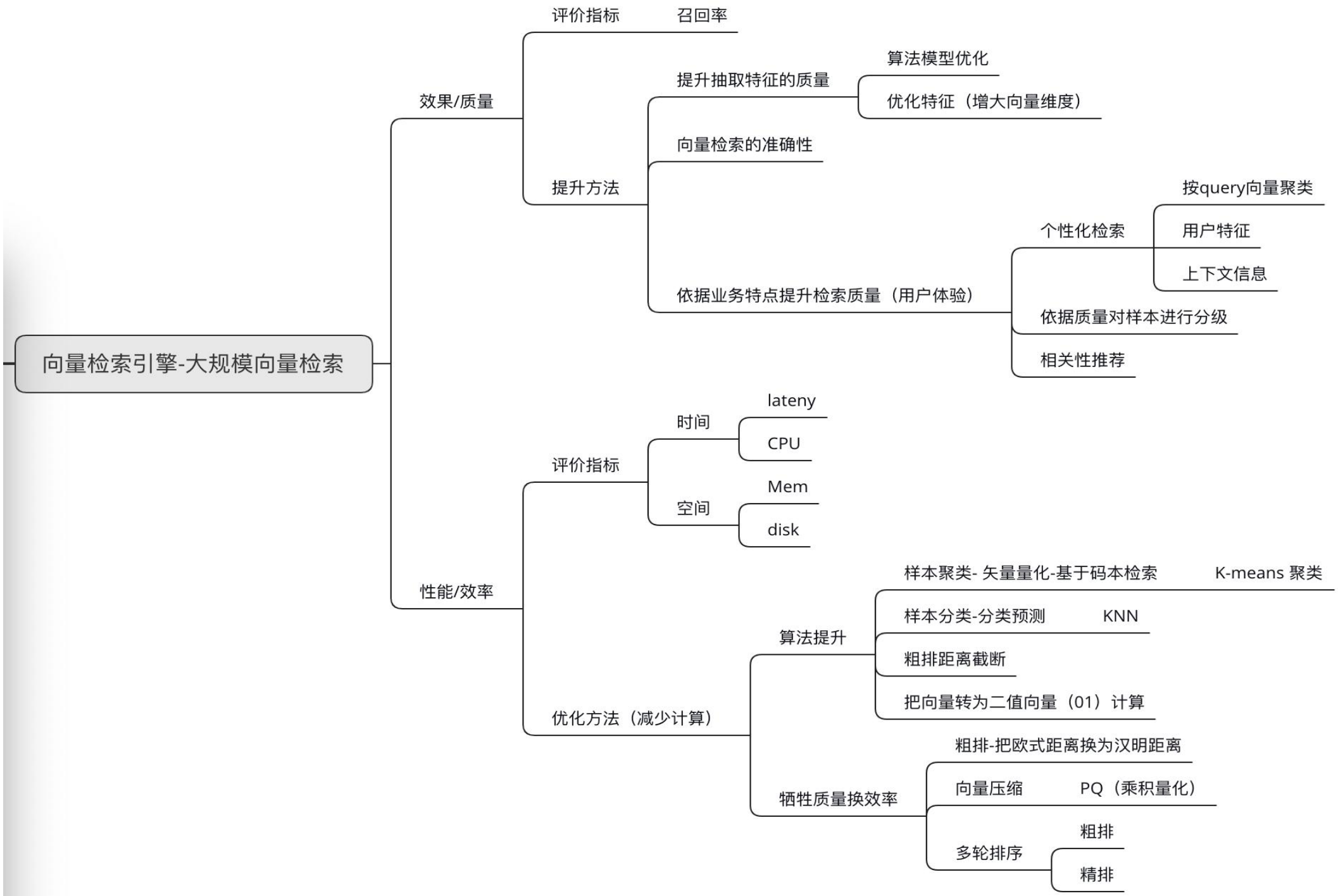
PQ乘积量化（Product Quantization）

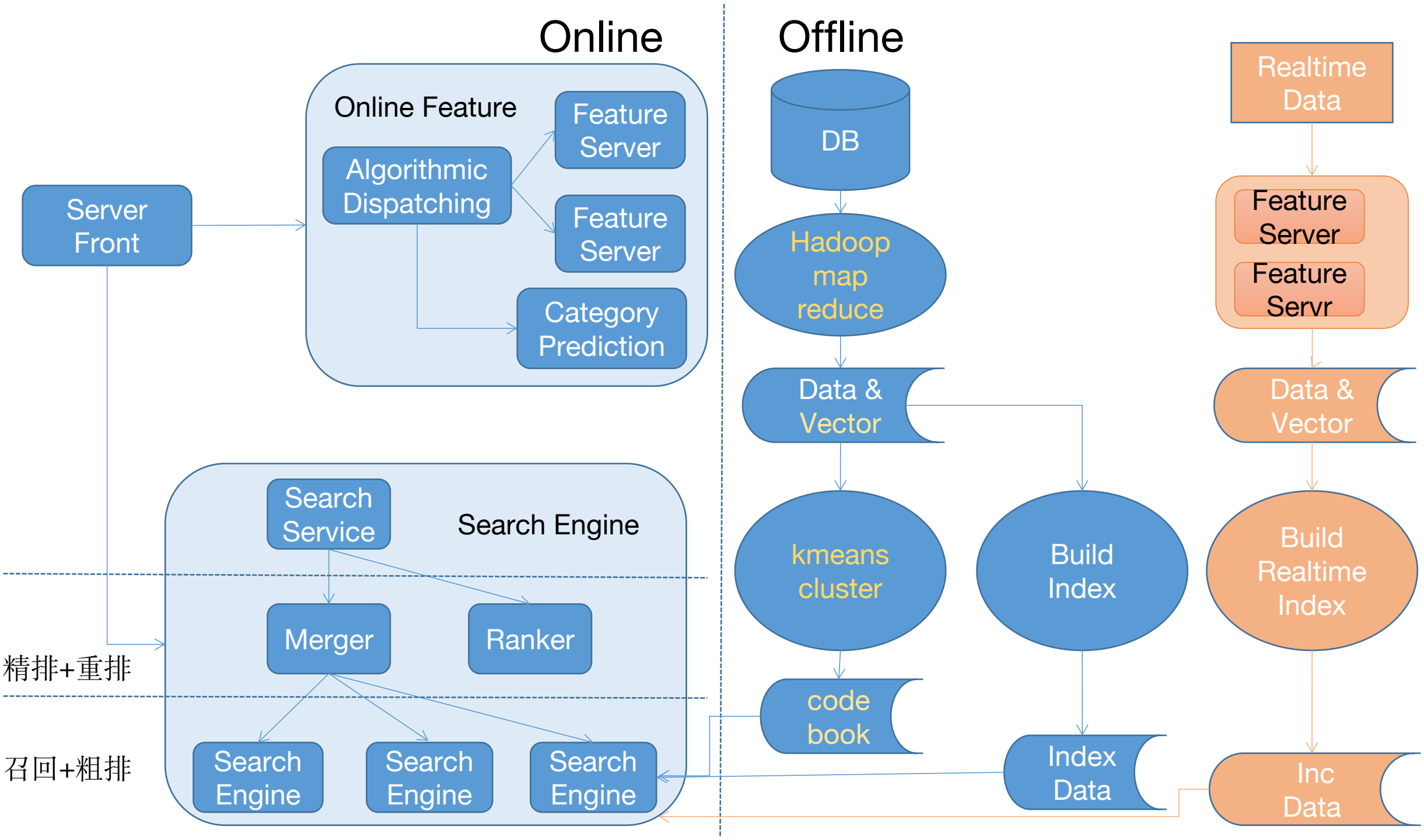


⇒ 8 subvectors x 8 bits = 64-bit quantization index

PQ乘积量化-码本表

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	2
∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘
0	0	0	0	0	0	0	255
∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘	∘ ∘ ∘
255	255	255	255	255	255	255	254
255	255	255	255	255	255	255	255


向量检索引擎-大规模向量检索
效果/质量
评价指标
召回率
提升方法
提升抽取特征的质量
算法模型优化
优化特征（增大向量维度）
向量检索的准确性
依据业务特点提升检索质量（用户体验）
个性化检索
按query向量聚类
用户特征
上下文信息
依据质量对样本进行分级
相关性推荐
性能/效率
评价指标
时间
lateny
CPU
空间
Mem
disk
优化方法（减少计算）
算法提升
样本聚类- 矢量量化-基于码本检索
K-means 聚类
样本分类-分类预测
KNN
粗排距离截断
把向量转为二值向量（01）计算
牺牲质量换效率
粗排-把欧式距离换为汉明距离
向量压缩
PQ（乘积量化）
多轮排序
粗排
精排


Online
Offline
Online Feature
Feature Server
Algorithmic Dispatching
Feature Server
Server Front
Category Prediction
DB
Realtime Data
Hadoop map reduce
Feature Server
Feature Servr
Data & Vector
Data & Vector
Search Service
Search Engine
kmeans cluster
Build Index
Build Realtime Index
Merger
Ranker
精排+重排
code book
召回+粗排
Search Engine
Search Engine
Search Engine
Index Data
Inc Data

向量检索引擎的问题

- 软件成熟度不够
- 向量检索融入普通搜索引擎
- 全样本分类后做类目预测差

向量检索在推荐系统中的应用

- 推荐系统中的数据量有多大 - 中等
- 深度树模型（阿里妈妈）
- 触发服务

Q & A