**Thinking Hw Sat Week6**

**郭想**

Thinking 1 什么是近似最近邻查找，常用的方法有哪些

=====

近似最近邻查找是指根据数据的相似性，从数据库中寻找距离目标数据最近数据的一种算法。一般来说，在使用合适的特征和维度下，距离接近的数据意味着它们的特征越接近，那么两个数据本身也就越相似。

一般使用的算法有KNN (K最近邻查找)，ANN(近似最近邻检索)，，而局部近似敏感哈希（LSH）则是ANN的一种。另外一种是矢量量化，其代表是乘积量化（PQ）。它的主要思想是将特征向量进行正交分解，在分解后的低维正交子空间上进行量化，由于低维空间可以采用较小的码本进行编码，因此可以降低数据存储空间 。

当然，在数据维度较低的情况下可以使用穷举的方法，即在数据库中依次计算其中样本与所查询数据之间的距离，抽取出所计算出来的距离最小的样本即为所要查找的最近邻。这种方法一般基于树状结构，比如k-d树方法，R树和M树。

Thinking 2 为什么两个集合的minhash值相同的概率等于这两个集合的Jaccard相似度

=====

首先，Jaccard的计算公式为下：

两个集合的Jaccard相似度为集合的交集除于两个集合的并集。

MinHash算法则是将特征矩阵进行随机排列之后，第一个列值为1的行号。那么假设有两个集合Ci和Cj，对应的行有一共三种可能：

1. 两列的值都为1
2. 其中一列的值为1，另外一列的值是0
3. 两列的值都为0

它们MinHash相等的概率即为两列值都为1的行数除于总的行数。两列值都为1的行数相当于两个集合的交集因为交集的定义就是两个集合都必须包含这个元素。而总的行数是任意一列值是1的行数，这个数值相当于是两个集合的并集。所以MinHash值相等的概率和jaccard相似度一样。

Thinking 3 SimHash在计算文档相似度的作用是怎样的？

=====

SimHash是一种LSH，可以帮助降维数据使得计算量没有那么大。以下是SimHash的计算过程：

1. 首先设置SimHash的位数
2. 初始化SimHash，将所有位数的值都设为0
3. 提取文本中的特征，可以采用k-Shingles的方法
4. 使用传统的hash函数计算每一个特征word的hash值。
5. 然后计算每一个word的权重（一般情况下是出现的频率）。如果该位为1，则simhash相对应的值加它的权重，否则减去它的权重。
6. 计算最后得到的32位的SimHash值，如果大于0的话则设为1，小于都等于0就设为0.

以上就能得到一个文档的simhash值。在计算两个文档的相似度时可以对它们的simhash值进行汉明距离计算，如果汉明距离小于某一个阈值就判定不相似，大于阈值就定为相似。

Thinking 4 为什么YouTube采用期望观看时间作为评估指标

=====

这主要是基于一个商业化目标的考虑，我自己的理解是因为观看时间比起CTR来说更能反应用户对于某个视频的喜好程度。资料显示CTR对于视频搜索还具有欺骗性，我的理解是一些不相关的词语可能会导致无效展示。比如用户在一些网站上搜索巴塞罗那，本意是查找巴塞罗那的旅游攻略，但是有可能网站会展示巴塞罗那足球队的信息导致无效展示使得展示量偏大，从而导致CTR偏小。