query改写

# 1、问题背景

　检索的主要的问题还是在于用户query和doc之间存在GAP，特别是中长尾query。把问题分成以下几种类型：

* 多种描述：划痕笔/补漆笔/修补笔/点漆笔，又或者苹果、apple
* 信息冗余: 冰箱温控器温度控制==冰箱温控器
* 属性检索：118冰箱、60寸液晶电视机4k高清智能60曲面
* 宽泛意图：超美吊灯、大容量冰箱

Query改写本质上是要找到和原始Query相似的候选Query，候选Query来自用户搜索query日志清洗过滤得到。在简单依赖搜索引擎的es中，BM25是基于共现词来判定query和doc相似程度的，比如苹果多少钱与apple什么价可能相似度为0.这种情况下，我们需要对query进行改写。

# 2、常用方法

**基于Query内容：**

1. 基于文本相似度，基于编辑距离（字或词），基于拼音
2. 基于同义词

**Co-Click基于用户点击行为：**query-doc协同过滤

我理解这种场景是指用户输入了一个query，虽然和输入没有完全匹配，但可能根据点击数据或者推荐等将相关doc返回，用户就直接点击了，这种情况下，我们认为query和doc存在一种映射关系。建立Query下点击行为矩阵；采用协同过滤的思想计算Q1和Q2相似度，具体计算相似度的方法可以有很多种，常用的入cosine余弦相似度。在query的点击行为比较稀疏时，还可以通过在query和商品的点击二部图上游走来扩充query的行为向量。

基于随机游走的方法，simrank，simrank++

**Co-Session基于用户同一个时间段的连续操作：**基于session embedding

纯粹从用query session出发：在同一个session中用户输入多个query，我们以为同一个session中的query都是有关系的，我们通过query编辑距离或者相似度的计算，可以挖掘相关的query序列出来.

这样我们可以得到一次query session数据:

user1: query1,query2,query3…….,

user1: query1,query2,query3…….,

我们在考虑用户共同兴趣的时候，联想到用户群体可能存在的相似性。推荐相关的query给不用的用户。

1.query之间的相似度.将上面的user->query矩阵变化成query->user矩阵,计算不同的query的相似度.

2.query聚类计算.直接通过层次聚类算法.