# q4. 对于 q1，q2，q3 3种方法的比较

1651718 方沛

## 1. 实验结果对比

| | q1 | q2 | q3 |
|---|---|---|---|
| Dist | Jaccard | Jaccard | FTC_Tree |
| SC | ≈0.013（k=25） | 0.05-0.06 （k=2） | 0.083 (k=17) |
| CP | 0.839 | 0.84 | 0.12 |

## 2. 方法概述

### 2.1. q1

> q1 借鉴了 K-means， 将用户购买的商品品类转化为向量，然后就可以用K-means的取质心方法。

### 2.2. q2

> q2 算法上相比于q1 没有什么进步，增加了分级品类取平均，考虑到了用户买了各种品类, 使得在 q1 中的两个商品产生了内在的关联。

### 2.3. q3

> q3通过FTC_tree 将用户的购买记录建成一棵树，并具备购买记录随时间退化的效果。自定义了一套取质心的方法。收获了不错的效果（虽然我在q3中并没有复现出这样好的效果）。

## 3. 方法优缺点比较

### 3.1 q1

**优点**

1. 我觉得q1没有什么优点。做法很普通，$SC$ 和 $CP$ 效果也很差。

**缺点**

1. 没有考虑到数据集的特征，包括商品的低层和高层特征。这点在q2有所改善。

### 3.2 q2

**优点**:

1. 增加了分级品类取平均，考虑到了用户买了各种品类, 使得在 q1 中的两个商品产生了内在的关联。
2. 试验出来的 $SC$ 和 $CP$ 指标相比于q1有进步。 为q3 论文中的方法提供了指导。

**缺点**:

1. 直接对每个品类取平均的做法还是太简单粗暴了。品类越细，比重应该越大比较合理。这点在q3得到了改善。
2. 没有什么创意。

## 3.3 q3

**优点**:

1. 优点很多。首先想到了将时间纳入考量。论文中提到，一个顾客的购买记录越久远，那么这条记录更加细分的商品类别会变得不准确，只能相信粗略的分类。我觉得这一点很合理, 通过$FTC_tree$成功做到了类别随时间退化。
2. 商品的各级品类权重随着并集树的深度有所不同。
3. 想法非常新颖。

**缺点**

1. 虽然 $FTC\_Tree$ 成功使得用户的距离得以区分，但是还是会有一些情况不符合逻辑。

   > 比如，用户$A$买了100件11；而用户$B$只买了1件11,3件25,他们之间的相似度会随着A购买的11的数目增加而增加，若用户B和用户A购买了统一种商品且未随之间退化。根据距离公式定义，$A$和$B$ 的距离会接近0, 这种情况需要考虑。

2. 质心树的生成完全不合理。论文中也没讲出这样生成质心树的直觉。质心的选取对论文中 $SC = 0.28$ 没有任何帮助（在q5中，我采用了简单的K-mediod算法加补充选取初始点的规则，就超过了0.28）。如果质心的选取能够变得合理，我认为效果可以更好。

   > 两个各买了100块肥皂的人，他们的质心是一个买了200块肥皂的人。这是文中的质心树算法产生的结果，事实上，这样的质心让它离簇内的点都变远了。

3. 除了模型以外，论文的写作有些瑕疵。主要表现在对方法可行性的推理不是很多；其次表述不明确，一些关键字句隐藏在整段文字中而不在推导中体现；文中对于时间的权重说的是采用交集树的深度，饶老师在群里的图片上写的是并集树；$FTC_tree$的距离计算本身就比较复杂繁琐，而论文插图中的例子缺乏代表性，无法体现距离计算的要点，给复现带来了难度。