Q1&Q2 分析报告

一、 问题分析

整个作业的最终目的就是根据用户的消费行为给用户进行用户画像的分类。前两问在实现方式上是一致的，都是利用 Jaccard 系数来计算用户购买记录之间的相似性，再用 Jaccard 距离作为 kmeans 聚类算法的距离公式，进行聚类分析。

区别在于 Q1 只使用最细粒度的商品编号判断两个用户是否存在相同的购买行为，而 Q2 由粗到细使用 4 级商品编号判断。最终造成的结果是 Q1 计算出来的距离普遍比较大而且不同用户之间 Jaccard 距离差距大；而 Q2 计算的距离相比 Q1 都要小些而且不同用户之间的 Jaccard 距离差距也会减小。从理论和结果上看，Q1 的距离没有 Q2 的合理，但较大的距离差异方便聚类，Q2 则刚好相反。

二、 运行结果

Q1、Q2 方法中的前两小问运行结果:

```
D:\Anaconda3\python.exe C:\Users\82621\Desktop\datamining20-HW1\Q1&Q2.py
第4级品类结构金额汇总:
{22002: 519.9599999999999, 34150: 65.8, 11054: 434.0999999999999, 27400: 2035.9599999999994, 11110: 3202.2, 15110: 7327.759999999978, 32821: 25.9, 10119
--------------------------------------------------
第3级品类结构金额汇总:
{2200: 7975.78, 3415: 94.8, 1105: 1944.1000000000001, 2740: 2506.13, 1111: 3798.6000000000001, 1511: 18412.55999999985, 3282: 31.9, 1011: 2119.0599999999
--------------------------------------------------
第3级品类结构金额汇总:
{220: 17083.029999999995, 341: 1541.1000000000006, 110: 3541.6000000000045, 274: 6093.899999999998, 111: 4698.100000000002, 151: 24783.84000000025, 328:
--------------------------------------------------
第3级品类结构金额汇总:
{22: 32141.73000000007, 34: 5140.699999999992, 11: 23739.659999999993, 27: 22244.100000000006, 15: 49231.86000000103, 32: 3322.6699999999996, 10: 17961.
--------------------------------------------------
方法1 Jaccard相似度: 0.011161824350459183
方法2 Jaccard相似度: 0.09608776217863843
```

使用 Q1 方法聚类的结果:

K = 9:

```
簇序号:  0 length:  4
------------------------------
簇序号:  1 length:  5
------------------------------
簇序号:  2 length:  3
------------------------------
簇序号:  3 length:  5
------------------------------
簇序号:  4 length:  436
------------------------------
簇序号:  5 length:  6
------------------------------
簇序号:  6 length:  8
------------------------------
簇序号:  7 length:  7
------------------------------
簇序号:  8 length:  12
------------------------------
0.9552895355552944 0.0064579251616547415
```

每一个簇的长度

CP

SC

K = 10:

```
簇序号:  0 length:  424
------------------------------
簇序号:  1 length:  7
------------------------------
簇序号:  2 length:  6
------------------------------
簇序号:  3 length:  10
------------------------------
簇序号:  4 length:  7
------------------------------
簇序号:  5 length:  5
------------------------------
簇序号:  6 length:  9
------------------------------
簇序号:  7 length:  8
------------------------------
簇序号:  8 length:  5
------------------------------
簇序号:  9 length:  5
------------------------------
0.953942429501898 0.004805815703647386
```

使用 Q2 方法聚类的结果:

K = 8:

```
簇序号:  0 length:  3
------------------------------
簇序号:  1 length:  4
------------------------------
簇序号:  2 length:  2
------------------------------
簇序号:  3 length:  1
------------------------------
簇序号:  4 length:  459
------------------------------
簇序号:  5 length:  12
------------------------------
簇序号:  6 length:  4
------------------------------
簇序号:  7 length:  1
------------------------------
0.8449359752651088 0.054816079068001004
```

K = 9:

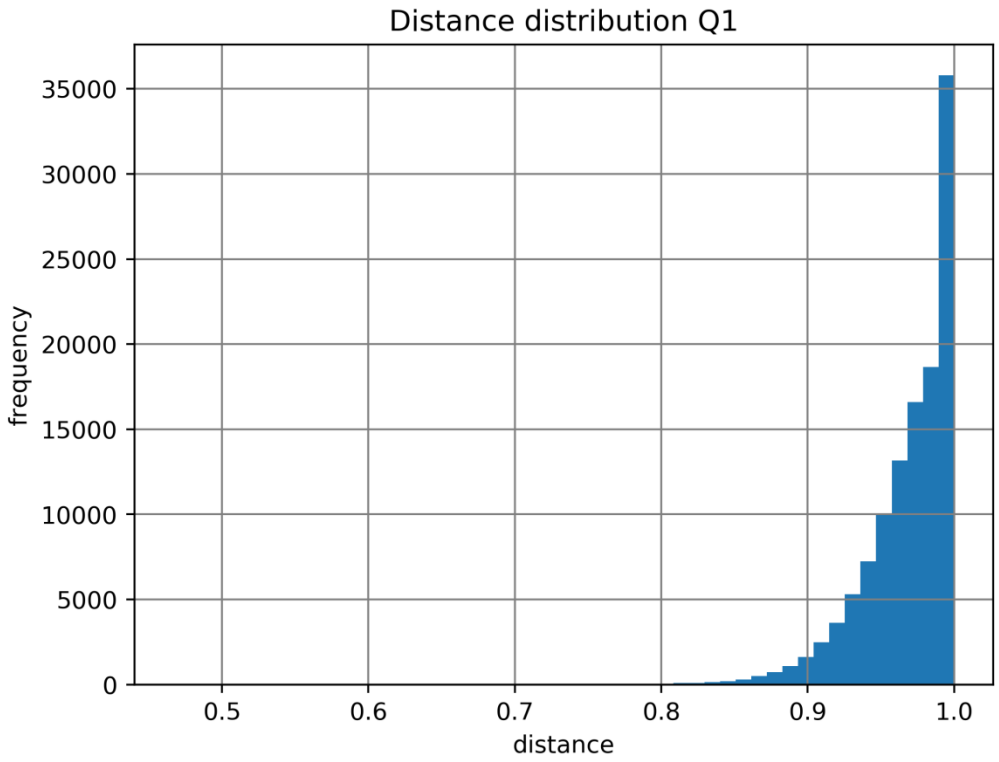
```
簇序号:  0 length:  5
------------------------------
簇序号:  1 length:  3
------------------------------
簇序号:  2 length:  3
------------------------------
簇序号:  3 length:  3
------------------------------
簇序号:  4 length:  7
------------------------------
簇序号:  5 length:  1
------------------------------
簇序号:  6 length:  5
------------------------------
簇序号:  7 length:  4
------------------------------
簇序号:  8 length:  455
------------------------------
0.8555441529043086 0.04051099840035175
```

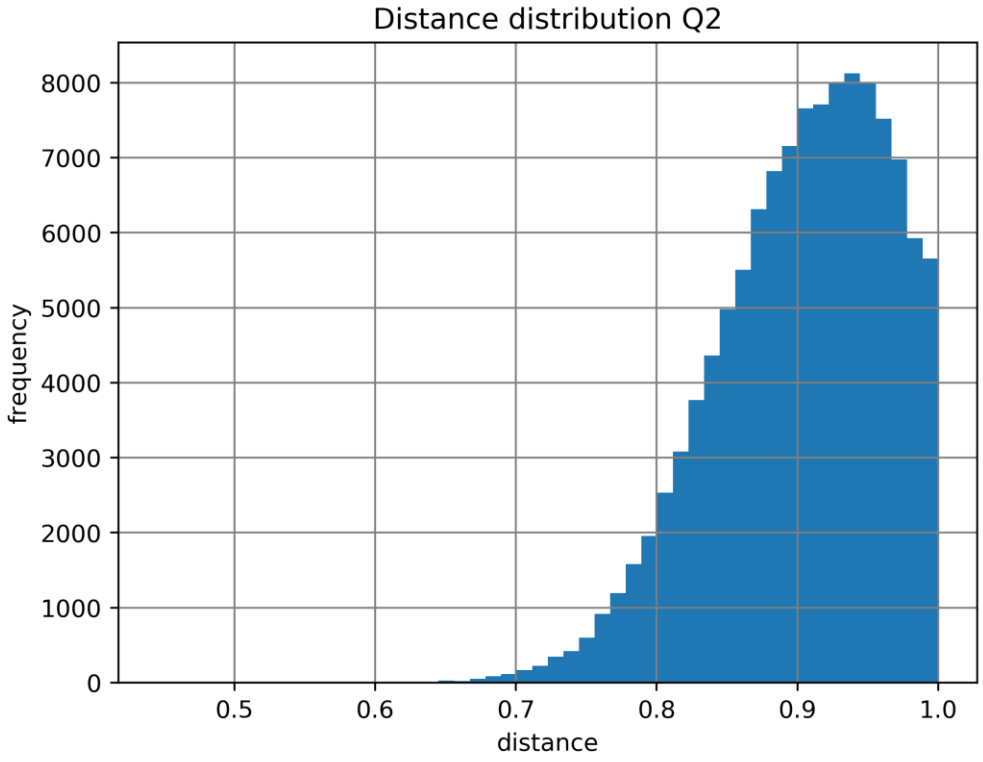
K = 10:

```
簇序号:  0 length:  1
------------------------------
簇序号:  1 length:  1
------------------------------
簇序号:  2 length:  4
------------------------------
簇序号:  3 length:  2
------------------------------
簇序号:  4 length:  5
------------------------------
簇序号:  5 length:  3
------------------------------
簇序号:  6 length:  8
------------------------------
簇序号:  7 length:  2
------------------------------
簇序号:  8 length:  1
------------------------------
簇序号:  9 length:  459
------------------------------
0.7412770924813799 0.051915062573300474
```

三、 结果分析

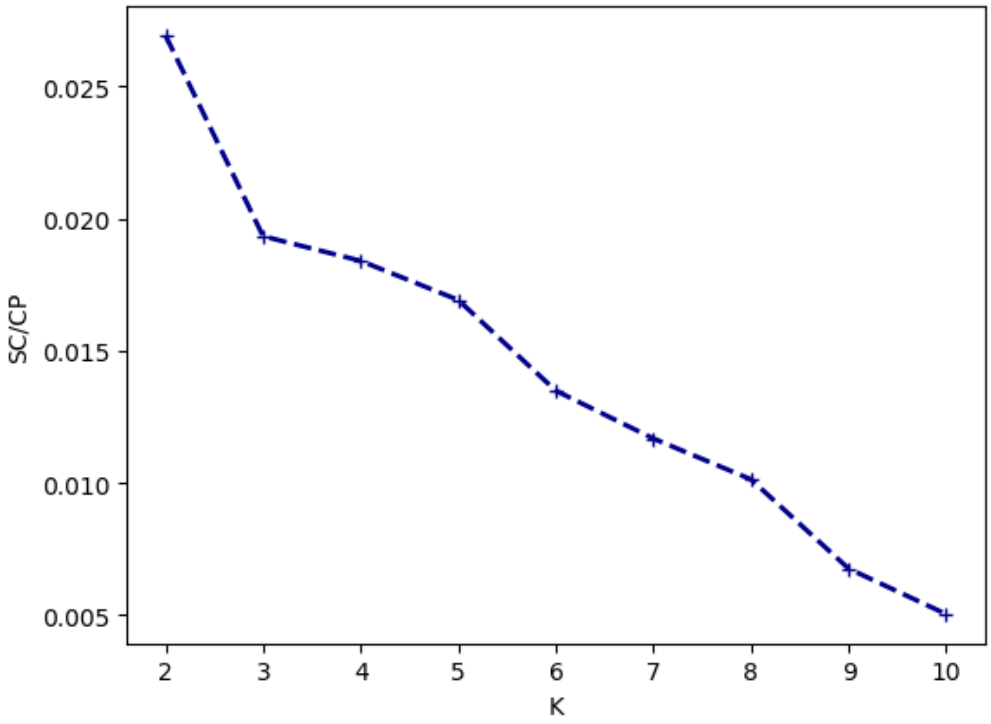
下面分别是使用 Q1、Q2 中的距离公式计算的距离分布图:

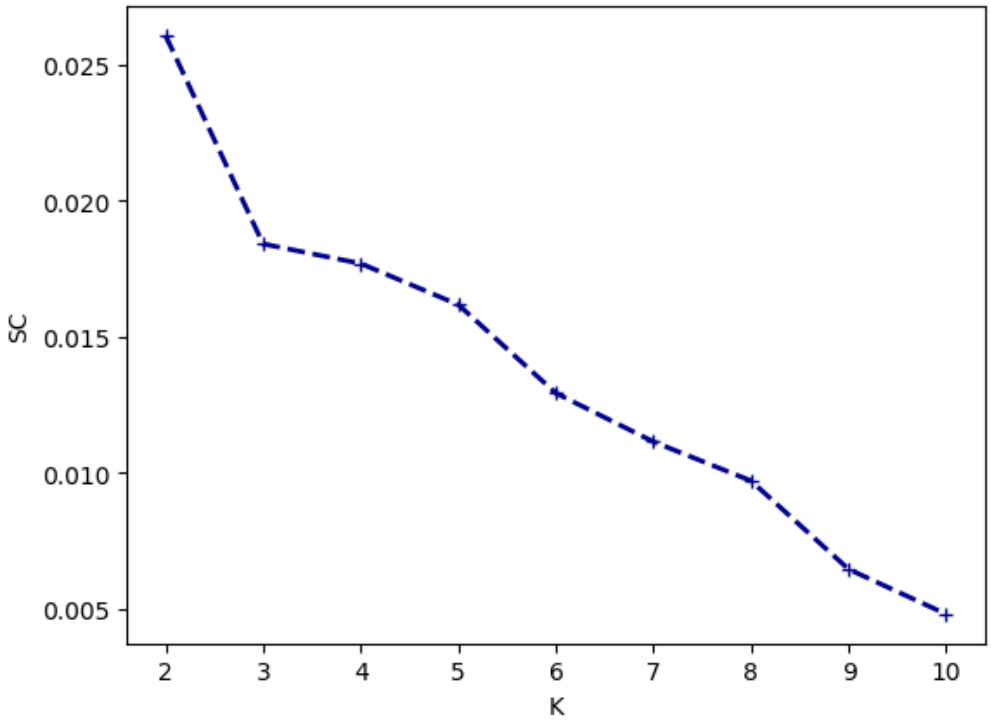


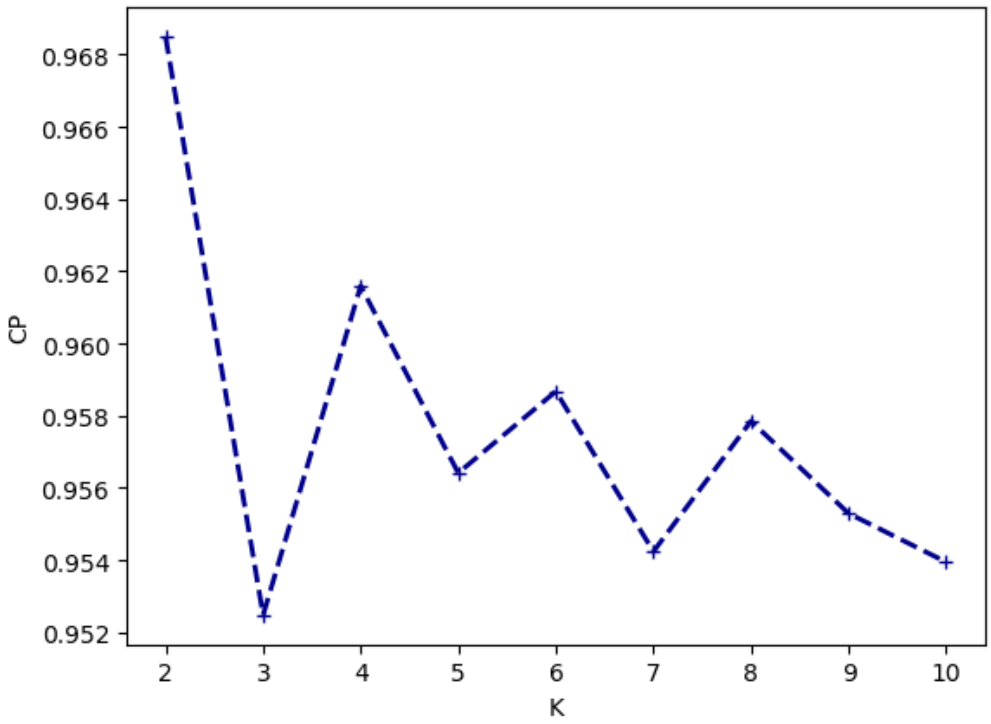


结果显示使用一个品级计算得到的距离要比使用四个品级的更大，这与距离公式代表的现实意义也是相符合的，使用越细粒度的距离公式会导致计算所得距离越大。

下面三幅折线图依次是按照 Q1 方法得出的 SC/CP - K 曲线图，SC - K 曲线图，CP - K 曲线图:

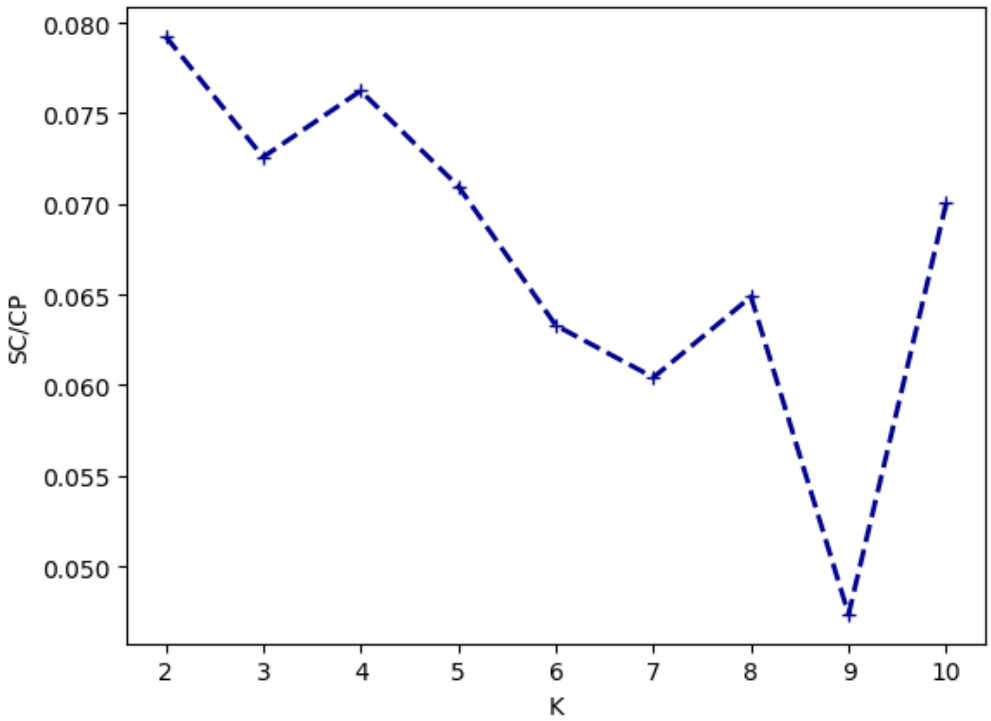


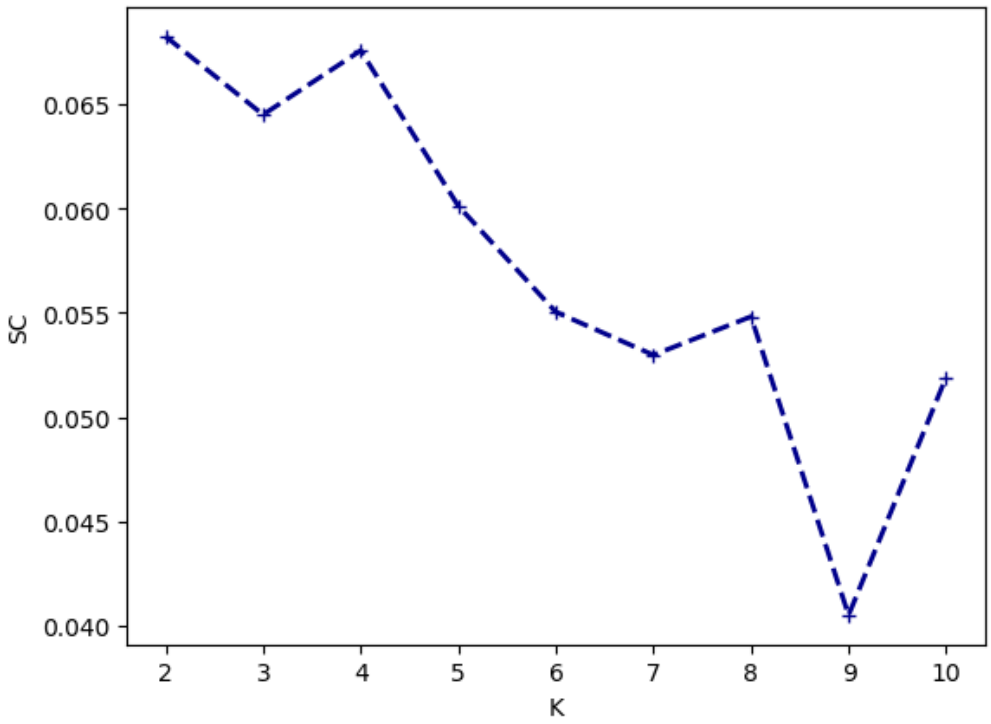


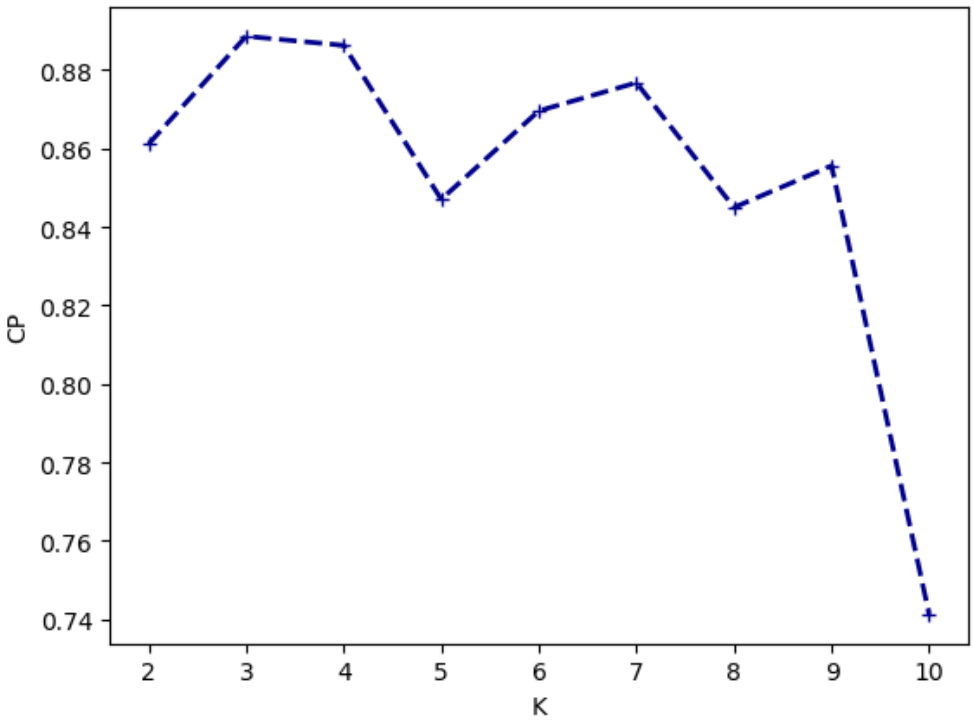


可以看到随着 K 值的增大，SC、CP、SC/CP 的值整体趋势都在下降。

同样按照 Q2 方法得出的结果图，依次是 SC/CP - K 曲线图，SC - K 曲线图，CP - K 曲线图:


0.080
0.075
0.070
0.065
0.060
0.055
0.050
SC/CP
2
3
4
5
6
7
8
9
10
K


0.065
0.060
0.055
0.050
0.045
0.040
SC
2
3
4
5
6
7
8
9
10
K

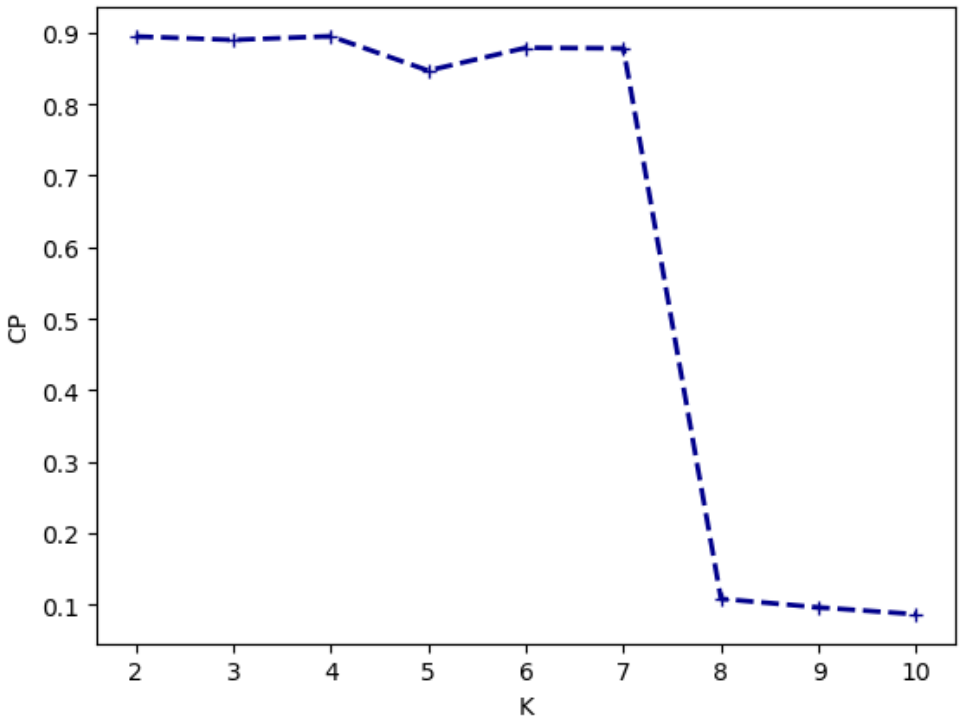


同样的 SC、CP、SC/CP 的值随着 K 增大减小，像 K=10 的反弹情况可能因为起始点选择相较前面更好从而聚类质量提升。

聚类过程中发现的问题:

但在聚类的过程中发现，K 值越大时越容易聚出只有一个类的情况，首先与初始点的选择有关系，经过多次重复实验后可以得到相应结果，同时也与距离公式有关，使用四个品级的距离公式一定程度上增大了不同用户记录之间的相似度，因为判断标准由细粒度变为较粗粒度，这就导致更多不同的点属于同一类。

下图为 Q2 方法得到的不同 K 值下的 cp 值，可以看到 K 在 2 到 7 之间时在 0.9 附近属于正常范围，而 K=8、9、10 时 cp 骤降到 0.1 附近其实是聚类时将所有点分为一类导致的结果。


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
CP
2
3
4
5
6
7
8
9
10
K