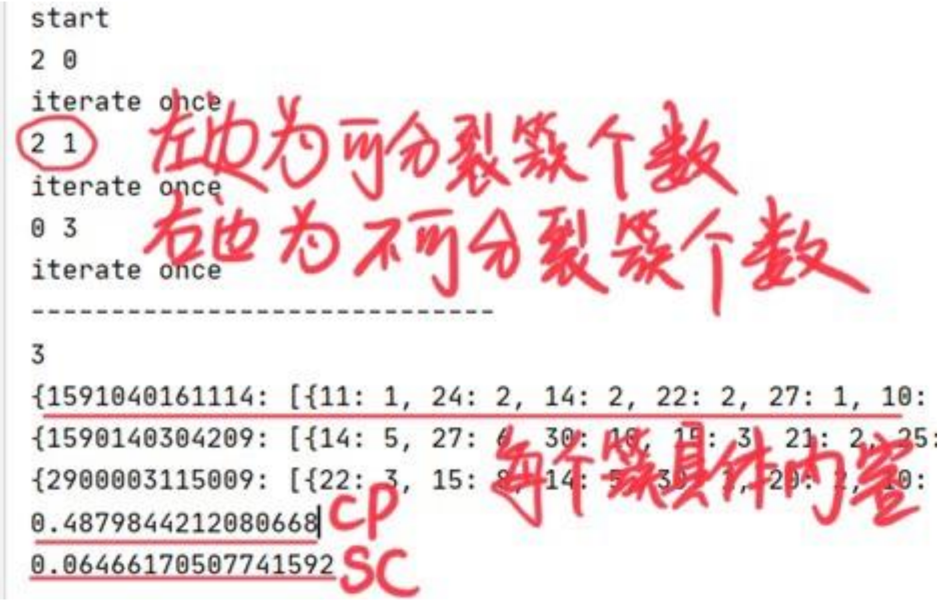
Q3 分析报告

一、 问题分析

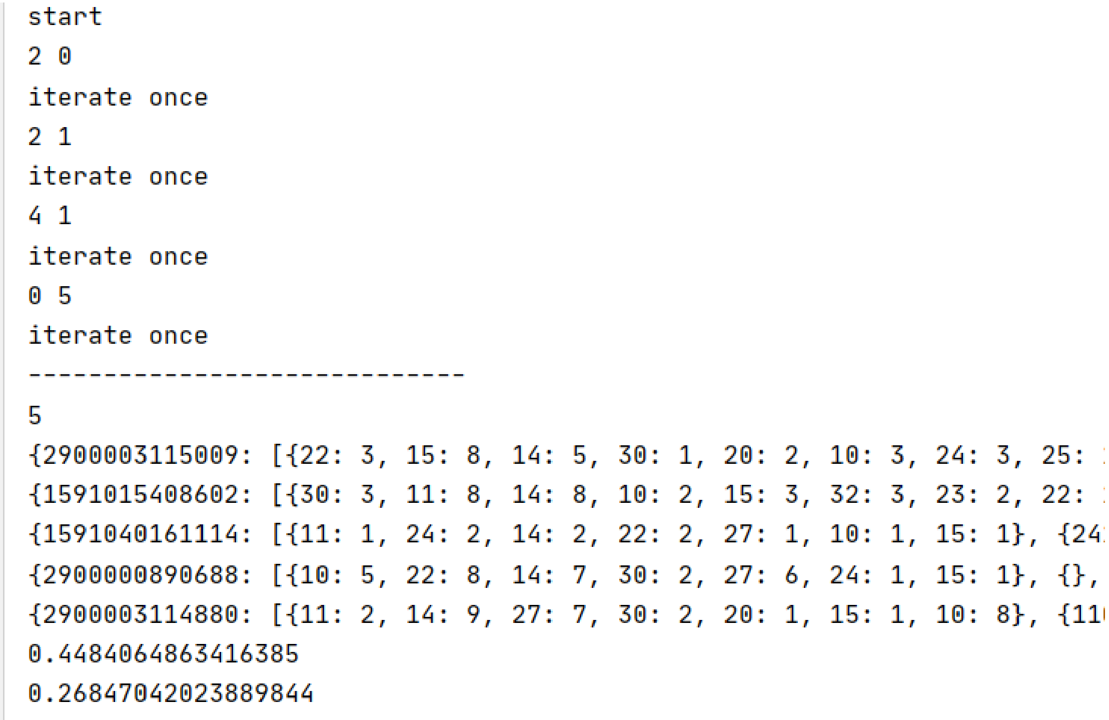
cluster.pdf 给出了一个以树结构存储用户购买记录，并通过树的深度反映购买记录的时效性，同时结合其他方法完成新型 kmeans 算法达到更好地划分用户画像种类的作用。

新型 kmeans 算法是在传统算法的基础上创新每个环节的方法使其更有针对性。首先改进的是距离公式，树形结构的存储将用户的购买记录考虑在内，使得结果更加合理。质心树生成算法通过不断删除出现频率小的节点找到一个离簇内所有节点最近的树，现实的意义就是剔除具体的商品考虑商品所属的大类，这也符合现实逻辑。最后采用针对性的 BIC 公式判断一个簇是否能够继续分裂，解决了 kmeans 算法 k 参数调试的问题。

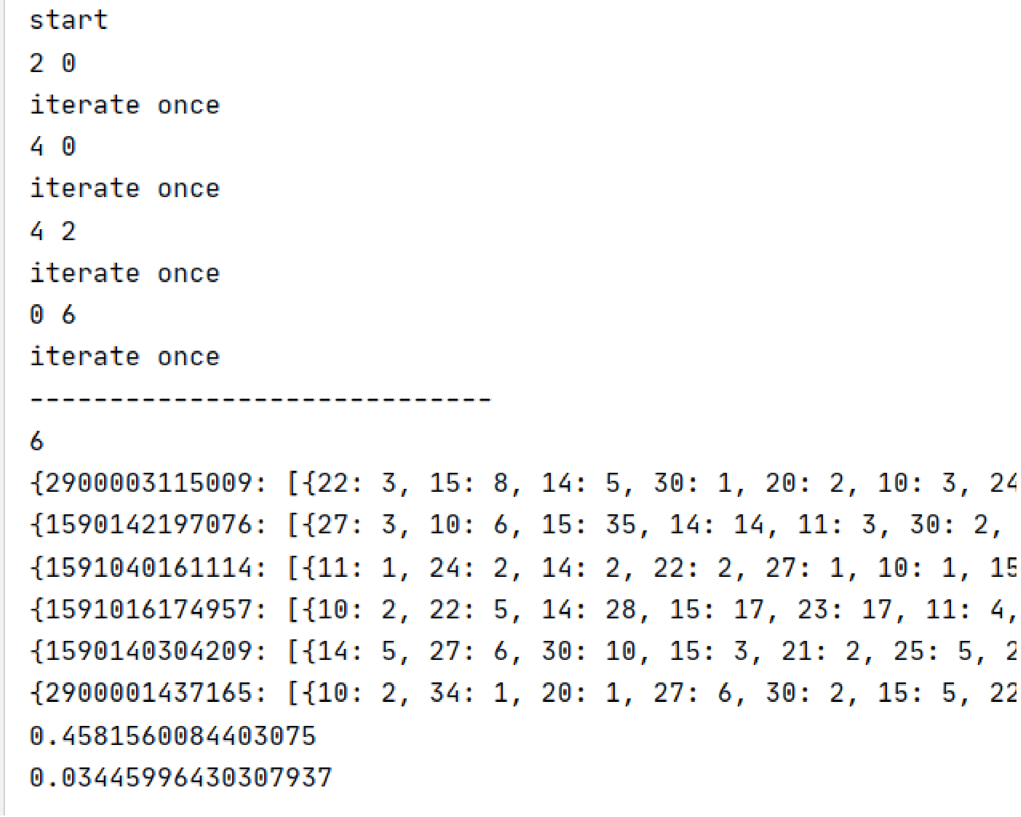
二、 运行结果



```
start
2 0
iterate once
2 1
iterate once
0 3
iterate once
----------------------------
3
{1591040161114: [{11: 1, 24: 2, 14: 2, 22: 2, 27: 1, 10:
{1590140304209: [{14: 5, 27: 6, 30: 10, 19: 3, 21: 2, 25:
{2900003115009: [{22: 3, 15: 8, 14: 5, 30: 1, 20: 2, 10:
0.4879844212080668
0.06466170507741592
```



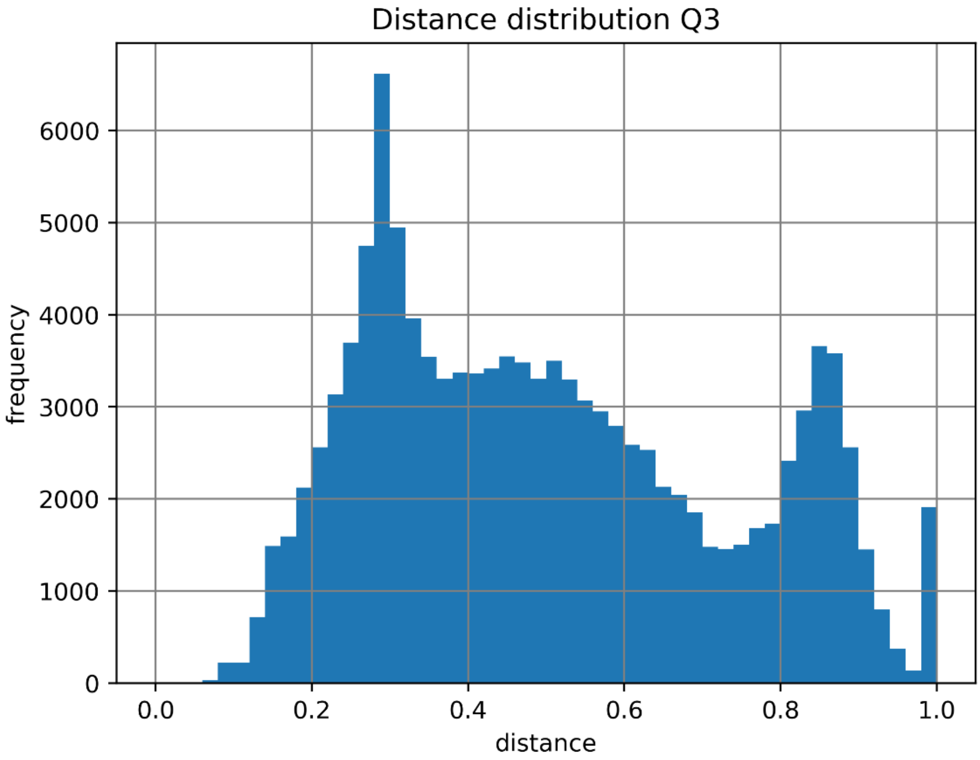
```
start
2 0
iterate once
2 1
iterate once
4 1
iterate once
0 5
iterate once
-----------------------------
5
{2900003115009: [{22: 3, 15: 8, 14: 5, 30: 1, 20: 2, 10: 3, 24: 3, 25:
{1591015408602: [{30: 3, 11: 8, 14: 8, 10: 2, 15: 3, 32: 3, 23: 2, 22:
{1591040161114: [{11: 1, 24: 2, 14: 2, 22: 2, 27: 1, 10: 1, 15: 1}, {24
{2900000890688: [{10: 5, 22: 8, 14: 7, 30: 2, 27: 6, 24: 1, 15: 1}, {},
{2900003114880: [{11: 2, 14: 9, 27: 7, 30: 2, 20: 1, 15: 1, 10: 8}, {11
0.4484064863416385
0.26847042023889844
```



```
start
2 0
iterate once
4 0
iterate once
4 2
iterate once
0 6
iterate once
-----------------------------
6
{2900003115009: [{22: 3, 15: 8, 14: 5, 30: 1, 20: 2, 10: 3, 24
{1590142197076: [{27: 3, 10: 6, 15: 35, 14: 14, 11: 3, 30: 2,
{1591040161114: [{11: 1, 24: 2, 14: 2, 22: 2, 27: 1, 10: 1, 15
{1591016174957: [{10: 2, 22: 5, 14: 28, 15: 17, 23: 17, 11: 4,
{1590140304209: [{14: 5, 27: 6, 30: 10, 15: 3, 21: 2, 25: 5, 2
{2900001437165: [{10: 2, 34: 1, 20: 1, 27: 6, 30: 2, 15: 5, 22
0.4581560084403075
0.03445996430307937
```

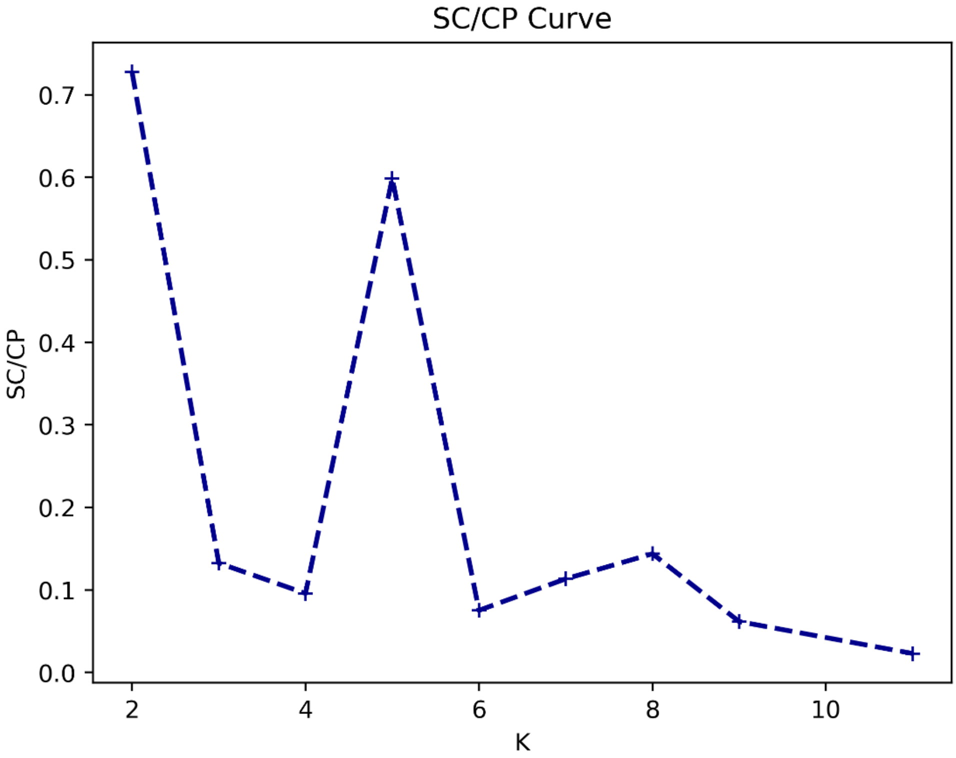
三、　结果分析

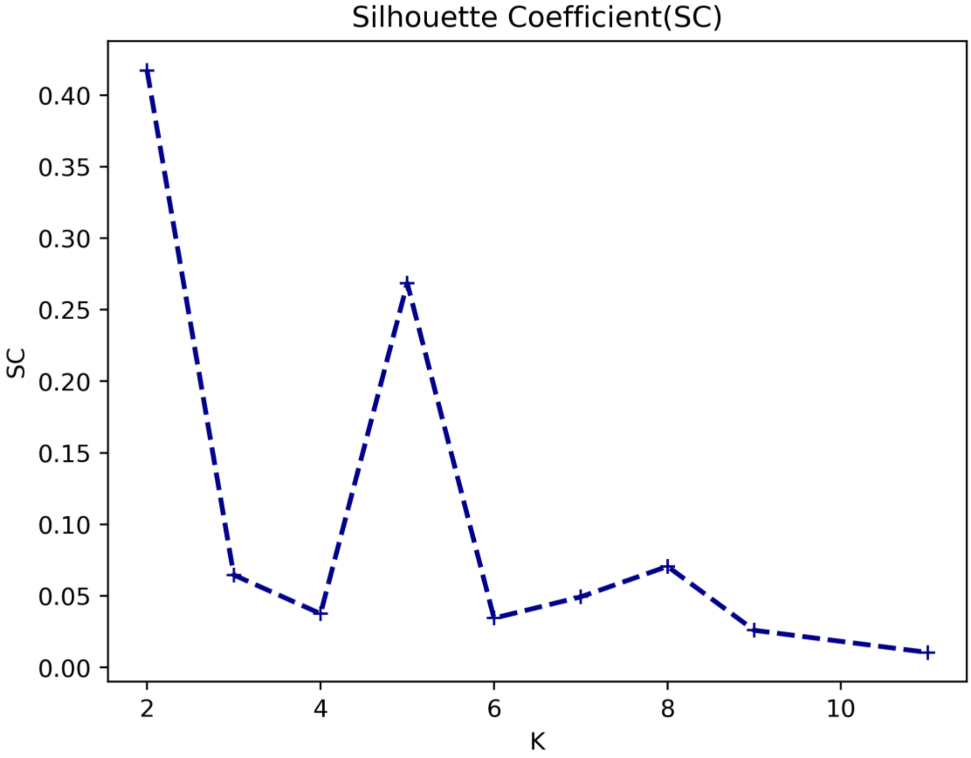
整个数据集的距离分布直方图:

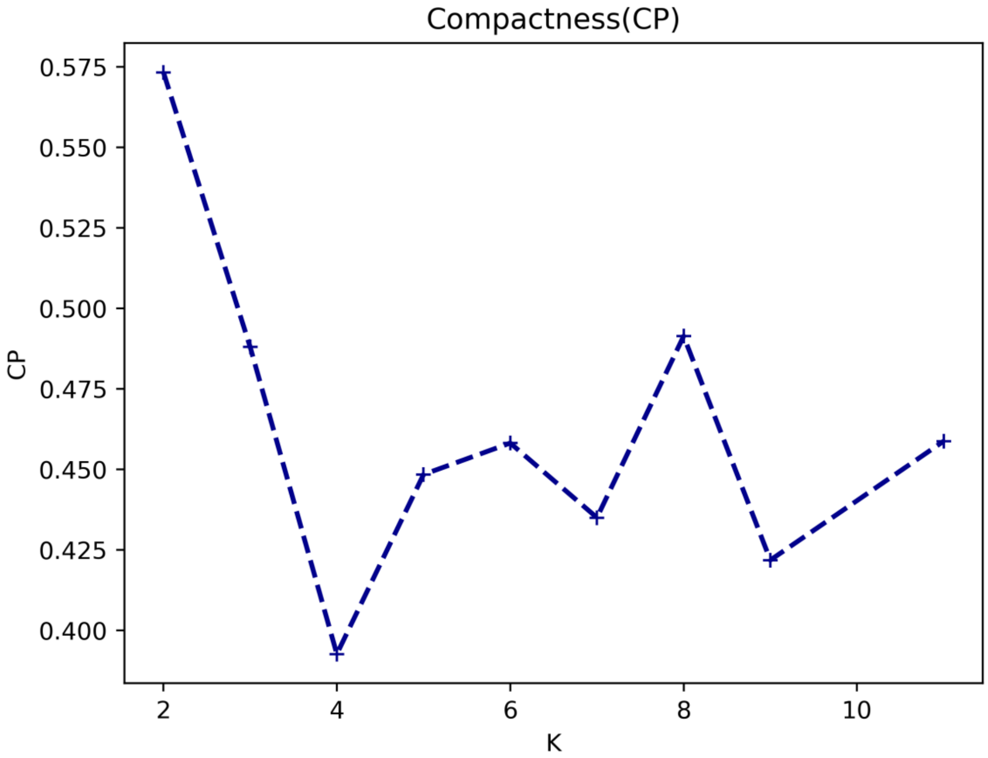


通过比较与前两问的距离分布直方图可见，采用树形结果将时间维度纳入测量范围后，距离不再集中在某段区域内，这说明对于不用用户的购买记录区分效果更加好，同时也说明这种新的判断用户画像相似度的方法是更好的。

下面三幅图依次是聚类得出的不同 K 值下的 SC/CP 曲线图、SC 曲线图、CP 曲线图:


SC/CP Curve
SC/CP
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2
4
6
8
10
K


Silhouette Coefficient(SC)
SC
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
2
4
6
8
10
K



因为是随机选择初始点，所以经常聚类的结果不理想，图中的点是经过有限次重复实验后选取的较好一点的样本，并不能完全代表最正确的聚类结果。但大体趋势与前两问是统一的，即 SC 和 CP 值随着 K 增大总体上在下降，而 SC/CP 的值则取决于二者。