|  |  |
| --- | --- |
|  |  |
| 项目报告E |  |
|  |  |
|  | 课程：数据挖掘姓名：姜其升 |
|  | 学号：1752058指导教师：饶卫雄教授 |

[项目报告E 1](#_Toc39148640)

[1 改进算法 3](#_Toc39148645)

[1.1 改进思想 3](#_Toc39148646)

[1.2 改进后相似度或距离公式 3](#_Toc39148647)

[2 关键代码 4](#_Toc39148648)

[3 聚类结果及运行截图 5](#_Toc39148649)

[3.1 B题改进 5](#_Toc39148650)

[3.2 C题改进 7](#_Toc39148651)

[4 聚类结果分析 10](#_Toc39148652)

[4.1 距离分布情况 10](#_Toc39148653)

[4.1.1 B题 10](#_Toc39148654)

[4.1.2 C题 10](#_Toc39148655)

[4.2 SC/CP的值 11](#_Toc39148656)

[4.2.1 B题 11](#_Toc39148657)

[4.2.2 C题 11](#_Toc39148658)

[5 算法性能 12](#_Toc39148659)

## 1 改进算法

### 1.1 改进思想

考虑到随着对于不同的商品分类级别，如：22蔬果科、220水果、2200实果类、22002桃，随着分类级别的增加，商品的相似度也在增加，即对于顾客来讲，如果三个顾客A、B、C，其中A和B购买了更多在220这个级别品类的商品，而顾客A和C则购买了更多22001这个级别品类的商品，显然A和C的相似性要大于A和B的相似性，但是在之前提到的计算相似度或距离公式中，并没有考虑到这一点，尤其是在题目B中的相似度计算方式中，因此就有可能出现AB和AC相似度相同甚至AB大于AC的情况。因此我们在这次计算相似度时，考虑这一点，对于不同的品类级别得到的相似度进行加权计算最后的相似度，即品类级别越低（2200低于22），权重越大。

### 1.2 改进后相似度或距离公式

对于B题：

其中sim1为第一级品类（前两位），sim2为第二级品类（前三位），sim3为第三级品类（前四位），sim4为第四级品类（前五位）。

对于C题：

## 2 关键代码

1. # B题
2. **def** jaccardSeri(user1, user2):
3. sim = 0
4. **for** j **in** sum\_user2.index:
5. **if** j **in** sum\_user1.index:
6. jiao = jiao + min(sum\_user1.loc[j], sum\_user2.loc[j])
7. bing = bing + max(sum\_user1.loc[j], sum\_user2.loc[j])
8. **else**:
9. bing = bing + sum\_user2.loc[j]
10. **for** j **in** sum\_user1.index:
11. **if** j **not** **in** sum\_user2.index:
12. bing = bing + sum\_user1.loc[j]
13. sim = sim + jiao / bing \* i
14. **return** sim/4
16. # C题
17. **def** Dist(FTCTree, freq):
18. IRTree = FTCTree.intersection(freq)
19. URTree = FTCTree.union(freq)
20. **return** (1-(sim1\*wl[0] + sim2\*wl[1]\*1.2 + sim3\*wl[2]\*1.4 + sim4\*wl[3]\*1.6))

## 3 聚类结果及运行截图

### 3.1 B题改进

K = 16

SC = 0.338

CP = -0.351

具体结果：

[2900000383913, 13325038116, 13854627199, 15954688237, 15963883482, 18654692914, 1590120464497, 1590130640102, 1590130817948, 1590140304506, 1590140305107, 1590140306678, 1590140307286, 1590140307415, 1590140307767, 1590140307859, 1590140308030, 1590140308214, 1590142128124, 1590142129763, 1590142148962, 1590142149853, 1590142151689, 1590142179621, 1590142191722, 1590142194327, 1590142197076, 1590142201162, 1590142203104, 1590142203180, 1590142203401, 1590142205993, 1590142206747, 1590142210195, 1590142213189, 1590142213356, 1590142213981, 1590142215619, 1590142242370, 1590142477895, 1590142519632, 1590142520300, 1590142648356, 1590142751995, 1590151076256, 1590151321240, 1590151363066, 1590151464053, 1590151472027, 1590151825175, 1590151832586, 1590151832593, 1590151835808, 1590151843124, 1591013471721, 1591013766650, 1591013866800, 1591013877134, 1591014338276, 1591014523610, 1591014680184, 1591015097233, 1591015269418, 1591015327590, 1591015408602, 1591015419448, 1591015452933, 1591015454210, 1591015454562, 1591015454814, 1591015456955, 1591015476700, 1591015478124, 1591015480745, 1591015483951, 1591015519216, 1591016154157, 1591016162466, 1591016170300, 1591016174353, 1591016174957, 1591016411137, 1591016440328, 1591016517778, 1591016519031, 1591020742401, 1591030217500, 1591030411823, 1591040087452, 1591040416658, 1591040462983, 1592010017509, 1592014016171, 1592015013285, 1592015018662, 1592015046818, 1592015053267, 1592015056299, 1592015059993, 1593141148827, 1593150463515, 1594140460286, 1595140489499, 1595150403423, 1595150770945, 1595150771652, 1595151204906, 1595151442803, 1595151616471, 1595151629143, 1595151630446, 1595151630774, 1595151866807, 1595151915963, 1595160035249, 1595160353664, 1596140586607, 1596140587802, 1596160006376, 1596160024097, 1598140107578, 2900000040663, 2900000072022, 2900000074163, 2900000124974, 2900000125261, 2900000138056, 2900000164567, 2900000179417, 2900000203457, 2900000239876, 2900000245396, 2900000245914, 2900000246638, 2900000268692, 2900000308312, 2900000337633, 2900000342651, 2900000345300, 2900000350175, 2900000350458, 2900000353848, 2900000372849, 2900000394940, 2900000404731, 2900000416307, 2900000418073, 2900000448520, 2900000449732, 2900000452756, 2900000476042, 2900000485990, 2900000557161, 2900000562165, 2900000566378, 2900000587649, 2900000604360, 2900000781603, 2900000840164, 2900000840805, 2900000847224, 2900000849785, 2900000863880, 2900000874145, 2900000874930, 2900000884007, 2900000890688, 2900000934474, 2900000936997, 2900000941830, 2900001050500, 2900001050791, 2900001067911, 2900001094702, 2900001211239, 2900001211321, 2900001215688, 2900001220255, 2900001234399, 2900001240246, 2900001270199, 2900001307901, 2900001308168, 2900001421652, 2900001432658, 2900001436700, 2900001463379, 2900001465960, 2900001467353, 2900001497336, 2900001509459, 2900001540117, 2900001550246, 2900001568630, 2900001575201, 2900001578363, 2900001661508, 2900001663823, 2900001671408, 2900002927733, 2900002930092, 2900002932614, 2900002933024, 2900002933918, 2900002934304, 2900002934892, 2900002936520, 2900002936940, 2900002937428, 2900002969559, 2900003106281, 2900003106502, 2900003110479, 2900003123479, 6222021615010082737, 6227002180901670266]  
[2900001053938, 1595150738976, 2900001466509, 2900002532630]  
[1591015091286]  
[1591015454142, 1590151829753, 2900000060012, 2900000856585]  
[1595151786686, 1591015480455, 1591015598785, 1591016350504, 1594140121125, 2900000047648, 2900000384187, 2900000401495, 2900000468030, 2900001137843, 2900001355049]  
[1590142240482, 1590142201513, 1590142202190, 1590142516600, 1590151457185, 1590151459721, 1590151682891, 1591014943968, 1591015027858, 1591015088262, 1592015026841, 2900000307377, 2900000516892, 2900001413886, 2900001436120, 2900002542837, 6227002180921663895]  
[1595150997786, 1591015501716, 1592015016880, 1592015058781, 1595150263683, 2900000330863, 2900000408586, 2900001157902, 2900002937503]  
[2900000574274, 18764569988, 1591015034924, 1595151355660, 1596130424193, 2900000776104, 2900001470780, 2900003105161]  
[1590151977348, 1590142149112, 1590142201148, 1590142518307, 1590151103907, 1590151427355, 1590151428581, 1591012978559, 1591014083046, 1591014214433, 1591015097516, 1591015184407, 1591015218485, 1591015420000, 1591015454692, 1591015521639, 1591015600815, 1591015604417, 1591016172045, 1591016310287, 1591016517921, 1591020377344, 1592013087615, 1595151631146, 1596140594374, 1596160058337, 2900000050785, 2900000071742, 2900000072701, 2900000177079, 2900000178106, 2900000224001, 2900000255463, 2900000256637, 2900000271814, 2900000284333, 2900000304376, 2900000386440, 2900000386495, 2900000390256, 2900000451452, 2900000490130, 2900000588691, 2900000623040, 2900000784918, 2900000845121, 2900000852525, 2900000863286, 2900000864498, 2900000883475, 2900001047739, 2900001050364, 2900001067881, 2900001068499, 2900001068833, 2900001090605, 2900001093019, 2900001095952, 2900001216616, 2900001222099, 2900001238632, 2900001255080, 2900001270205, 2900001331616, 2900001364058, 2900001398121, 2900001419826, 2900001431651, 2900001451932, 2900001452366, 2900001465380, 2900001465816, 2900001559720, 2900001645768, 2900001668996, 2900002548457, 2900002701746, 2900002932416, 2900002934342, 2900002934915, 2900002944471, 2900002944495, 2900003105154, 2900003107363, 2900003108773, 2900003114002, 2900003114170, 2900003115009, 2900003120607]  
[1595151375125, 1591015506575, 1592015056343, 1595151609213, 2900000210790, 2900000317307, 2900000339637, 2900000780835, 2900000883819, 2900001078658, 2900001270014, 2900001436366, 2900001437165, 2900001540971, 2900001582933, 2900002934762]  
[1592015058774, 1590120718170, 1590142517836, 1591013227632, 1591040203906, 1595151644528, 2900000048812]  
[1591020742449, 1591014130276, 1591040356657, 1595130081566, 2900000224315, 2900000348561, 2900000417663, 2900000780040, 2900000855502, 2900003117386]  
[2900000254503, 781924, 13864739266, 15963885355, 18554652702, 1590141216259, 1590141414228, 1590142152440, 1590142456470, 1590142514866, 1590150321449, 1590151500423, 1591013167365, 1591014278527, 1591014637324, 1591015074944, 1591015136338, 1591015456993, 1591016159916, 1591016320026, 1591016382857, 1591016443121, 1591016494765, 1591020695158, 1591030092053, 1591040155274, 1592015054592, 1593160622155, 1595150992095, 1595151168727, 1595151738968, 1598140129341, 2900000194298, 2900000216952, 2900000680036, 2900000780743, 2900001109598, 2900001309639, 2900001423069, 2900001435116, 2900001436410, 2900001437424, 2900001538596, 2900001684507, 2900002944402, 2900003114880, 2900003116822]  
[1591040497398, 15954611837, 1590140304209, 1590142150507, 1590142176354, 1590142204569, 1590142230179, 1590142516068, 1590151210391, 1590151432557, 1590151467962, 1591014133086, 1591015112424, 1591015457273, 1591015500993, 1591016439575, 1591040161114, 1593140971419, 1595112593216, 1595140612422, 1595150738747, 1595151123344, 1596140628703, 1598140055923, 2900000050419, 2900000138223, 2900000253230, 2900000385627, 2900000880467, 2900001020541, 2900003114613, 2900003116341]  
[2900000870994, 1590142176521, 1590151472782, 1590151842899, 1591014255948, 1591014577972, 1591020269281, 1595150883423, 2900000192492, 2900000544765, 2900001092012, 2900001432672, 2900001462648, 2900003122076]  
[2900000598973]



### 3.2 C题改进

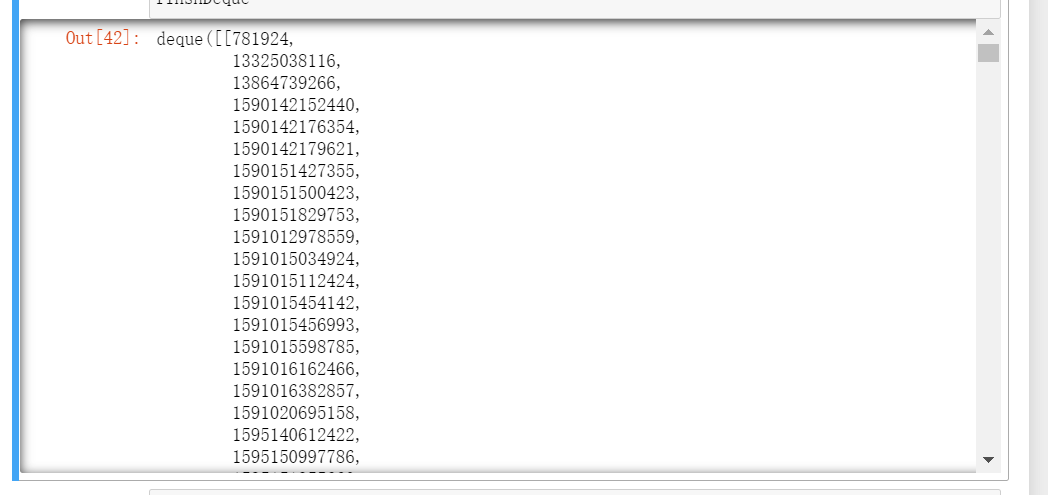
K = 8

SC = 0.319

CP = 0.189

具体结果：

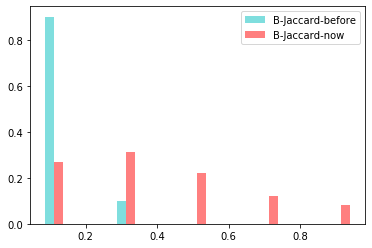
[781924,13325038116,13864739266,1590142152440,1590142176354,1590142179621,1590151427355,1590151500423,1590151829753,1591012978559,1591015034924,1591015112424,1591015454142,1591015456993,1591015598785,1591016162466,1591016382857,1591020695158,1595140612422,1595150997786,1595151355660,1595151630774,1598140055923,2900000060012,2900000138223,2900000254503,2900000304376,2900000384187,2900000574274,2900000623040,2900000776104,2900001222099,2900001240246,2900001582933,2900001645768,2900002944495,2900003116341]  
[2900001137843,2900000780835,2900000855502]  
[1593140971419,1595151375125,1595151644528,2900000050419,2900000072701,2900000192492,2900000348561,2900000385627,2900000390256,2900000780040]  
[1590142201513,13854627199,15954611837,15954688237,15963885355,18654692914,18764569988,1590120464497,1590120718170,1590130640102,1590130817948,1590140304209,1590140304506,1590140305107,1590140306678,1590140307286,1590140307415,1590140307767,1590140307859,1590140308030,1590140308214,1590141216259,1590141414228,1590142128124,1590142129763,1590142148962,1590142149112,1590142149853,1590142150507,1590142151689,1590142176521,1590142191722,1590142194327,1590142197076,1590142201148,1590142201162,1590142202190,1590142203104,1590142203180,1590142203401,1590142204569,1590142205993,1590142206747,1590142210195,1590142213189,1590142213356,1590142213981,1590142215619,1590142230179,1590142242370,1590142456470,1590142477895,1590142514866,1590142516068,1590142516600,1590142517836,1590142518307,1590142519632,1590142648356,1590142751995,1590150321449,1590151076256,1590151103907,1590151210391,1590151321240,1590151363066,1590151428581,1590151432557,1590151459721,1590151464053,1590151467962,1590151472027,1590151472782,1590151682891,1590151825175,1590151832586,1590151832593,1590151835808,1590151842899,1590151843124,1590151977348,1591013167365,1591013227632,1591013471721,1591013766650,1591013866800,1591013877134,1591014083046,1591014130276,1591014133086,1591014214433,1591014255948,1591014278527,1591014338276,1591014523610,1591014577972,1591014637324,1591014680184,1591014943968,1591015027858,1591015074944,1591015088262,1591015097233,1591015097516,1591015136338,1591015184407,1591015218485,1591015269418,1591015327590,1591015408602,1591015419448,1591015420000,1591015452933,1591015454210,1591015454562,1591015454692,1591015454814,1591015456955,1591015457273,1591015476700,1591015478124,1591015480455,1591015480745,1591015483951,1591015500993,1591015501716,1591015506575,1591015519216,1591015521639,1591015600815,1591015604417,1591016154157,1591016159916,1591016170300,1591016172045,1591016174353,1591016174957,1591016310287,1591016320026,1591016350504,1591016411137,1591016439575,1591016440328,1591016443121,1591016494765,1591016517778,1591016517921,1591016519031,1591020269281,1591020377344,1591020742401,1591020742449,1591030092053,1591030217500,1591030411823,1591040087452,1591040155274,1591040161114,1591040203906,1591040356657,1591040416658,1591040462983,1591040497398,1592010017509,1592013087615,1592014016171,1592015013285,1592015016880,1592015018662,1592015026841,1592015046818,1592015053267,1592015054592,1592015056299,1592015056343,1592015058774,1592015058781,1592015059993,1593141148827,1593150463515,1593160622155,1594140460286,1595112593216,1595130081566,1595140489499,1595150263683,1595150403423,1595150738976,1595150770945,1595150771652,1595150883423,1595150992095,1595151123344,1595151168727,1595151204906,1595151442803,1595151609213,1595151616471,1595151629143,1595151630446,1595151631146,1595151738968,1595151786686,1595151866807,1595151915963,1595160035249,1595160353664,1596130424193,1596140586607,1596140587802,1596140594374,1596140628703,1596160006376,1596160024097,1596160058337,1598140107578,1598140129341,2900000040663,2900000047648,2900000048812,2900000050785,2900000071742,2900000072022,2900000074163,2900000124974,2900000125261,2900000138056,2900000164567,2900000177079,2900000178106,2900000179417,2900000194298,2900000203457,2900000210790,2900000216952,2900000224001,2900000224315,2900000239876,2900000245396,2900000245914,2900000246638,2900000253230,2900000255463,2900000256637,2900000268692,2900000271814,2900000284333,2900000307377,2900000308312,2900000317307,2900000330863,2900000337633,2900000339637,2900000342651,2900000345300,2900000350175,2900000350458,2900000353848,2900000372849,2900000383913,2900000386440,2900000386495,2900000394940,2900000401495,2900000404731,2900000408586,2900000416307,2900000417663,2900000418073,2900000448520,2900000449732,2900000451452,2900000452756,2900000468030,2900000476042,2900000485990,2900000490130,2900000544765,2900000557161,2900000562165,2900000566378,2900000587649,2900000588691,2900000604360,2900000680036,2900000780743,2900000781603,2900000784918,2900000840164,2900000840805,2900000845121,2900000847224,2900000849785,2900000852525,2900000863286,2900000863880,2900000864498,2900000870994,2900000874145,2900000874930,2900000880467,2900000883475,2900000883819,2900000884007,2900000890688,2900000934474,2900000936997,2900000941830,2900001020541,2900001047739,2900001050364,2900001050500,2900001050791,2900001067881,2900001067911,2900001068499,2900001068833,2900001078658,2900001090605,2900001092012,2900001093019,2900001094702,2900001095952,2900001109598,2900001157902,2900001211239,2900001211321,2900001215688,2900001216616,2900001220255,2900001234399,2900001238632,2900001255080,2900001270199,2900001270205,2900001307901,2900001308168,2900001309639,2900001331616,2900001355049,2900001364058,2900001398121,2900001413886,2900001419826,2900001421652,2900001423069,2900001431651,2900001432658,2900001432672,2900001435116,2900001436120,2900001436366,2900001436410,2900001436700,2900001437165,2900001437424,2900001451932,2900001452366,2900001462648,2900001463379,2900001465380,2900001465816,2900001465960,2900001466509,2900001467353,2900001470780,2900001497336,2900001509459,2900001538596,2900001540117,2900001540971,2900001550246,2900001559720,2900001568630,2900001575201,2900001578363,2900001661508,2900001663823,2900001668996,2900001671408,2900001684507,2900002532630,2900002542837,2900002548457,2900002701746,2900002927733,2900002930092,2900002932416,2900002932614,2900002933024,2900002933918,2900002934304,2900002934342,2900002934762,2900002934892,2900002934915,2900002936520,2900002936940,2900002937428,2900002937503,2900002944402,2900002944471,2900002969559,2900003105154,2900003105161,2900003106281,2900003106502,2900003107363,2900003108773,2900003110479,2900003114002,2900003114170,2900003114613,2900003114880,2900003115009,2900003116822,2900003117386,2900003120607,2900003122076,2900003123479,6222021615010082737,6227002180901670266,6227002180921663895]  
[1590142520300,15963883482,18554652702]  
[1590142240482,1590151457185,2900000516892]  
[1591015091286,2900000598973,2900001270014]  
[1594140121125,1595150738747,2900000856585,2900001053938]



## 4 聚类结果分析

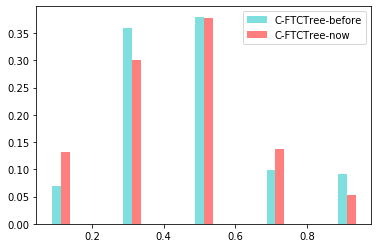
### 4.1 距离分布情况

#### 4.1.1 B题



上图为Jaccard系数的分布情况，我们可以看到，经过调整后的分布更加均匀。

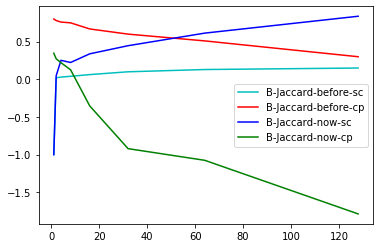
#### 4.1.2 C题



上图为FTCTree距离加权后的变化情况，变化并不显著。

### 4.2 SC/CP的值

#### 4.2.1 B题



我们看到的是数据从[1, 2, 4, 8, 16, 32, 64, 128]变化的情况图，我们可以看到青色曲线为改变前SC的值随K变化的情况，红色曲线为改变前CP的值随K值变化的情况，蓝色曲线为改变后SC的值随K变化的情况，绿色曲线为改变后CP的值随K值变化的情况。改变后曲线的变动幅度明显增强，寻找最优值时也更加清楚。

#### 4.2.2 C题

|  |  |  |
| --- | --- | --- |
|  | 改变前 | 改变后 |
| K | 4 | 8 |
| SC | 0.189 | 0.319 |
| CP | 0.289 | 0.189 |

对比SC和CP的值，我们发现改变后SC的值和CP值均向更好的方向发展。

## 5 算法性能

性能上和之前并无太大差异。

但从指标结果上看，改进后时距离确实有所优化，更利于聚类。