Mini Project Report

1. Our suggested improvement of the algorithem is using the basic algorithem and add a condition of only clustering if 2 movies have one or more generes in common.
we used the fact that ratings data files are real rating by real people, so that its more likely that people who like a certain movie genere will watch a number of movies from the same genere.

2.The improvment can be seen at input.txt
results:

Movie numbers are:
3165, 1514, 2090, 99, 997, 68, 423, 3514, 3266, 742, 2220, 1499, 403, 238, 1751, 928, 267, 1420, 3105, 3213, 3038, 3011, 692, 1208, 577, 514, 1414, 3915, 2528, 1999, 2607, 1380, 1850, 3333, 3237, 2123, 306, 106, 3824, 1867, 700, 3382, 1592, 2042, 3666, 1316, 1884, 2187, 1250, 2930, 1611, 241, 1880, 2972, 216, 2560, 2925, 24, 2948, 1566, 1998, 1650, 216, 1053, 2201, 2923, 23, 3081, 1512, 3020, 2347, 611, 527, 3600, 2468, 370, 1219, 3359, 3326, 1791, 3783, 914, 129, 403, 668, 2238, 1112, 1456, 3492, 3121, 3344, 2483, 2471, 232, 715, 2877, 2180, 3560, 1888, 600,

the clusters of the basic algorithm are:
cluster number 1:
3165 Boiling Point , 997 Caught , 423 Blown Away , 2220 Manxman, The , 1499 Anaconda , 403 Two Crimes , 267 Major Payne , 692 Solo , 1999 Exorcist III, The , 1867 Tarzan and the Lost City , 1592 Air Bud , 2187 Stage Fright , 1998 Exorcist II: The Heretic , 23 Assassins , 3020 Falling Down , 2347 Pope of Greenwich Village, The , 611 Hellraiser: Bloodline , 2468 Jumpin' Jack Flash , 1791 Twilight , 403 Two Crimes , 1112 Palookaville , 2180 Torn Curtain , 3560 Phantom Love ,
cluster number 2:
1514 Temptress Moon , 68 French Twist , 1208 Apocalypse Now , 3237 Kestrel's Eye , 106 Nobody Loves Me , 1250 Bridge on the River Kwai, The , 1611 My Own Private Idaho , 1880 Lawn Dogs , 2972 Red Sorghum , 232 Eat Drink Man Woman , 715 Horseman on the Roof, The ,
cluster number 3:
2090 Rescuers, The , 238 Far From Home: The Adventures of Yellow Dog , 3213 Batman: Mask of the Phantasm , 577 Andre , 2123 All Dogs Go to Heaven , 2042 D2: The Mighty Ducks , 241 Fluke , 2925 Conformist, The , 1566 Hercules , 2238 Seven Beauties ,
cluster number 4:
99 Heidi Fleiss: Hollywood Madam , 3266 Man Bites Dog , 1420 Message to Love: The Isle of Wight Festival , 3011 They Shoot Horses, Don't They? , 3333 Killing of Sister George, The , 1650 Washington Square , 1053 Normal Life , 914 My Fair Lady ,
cluster number 5:
3514 Joe Gould's Secret , 3915 Girlfight , 1850 I Love You, Don't Touch Me! , 2930 Return with Honor , 2948 From Russia with Love , 3326 What Planet Are You From? , 3783 Croupier , 129 Pie in the Sky , 600 Love and a .45 ,
cluster number 6:
742 Thinner , 2560 Ravenous , 3081 Sleepy Hollow , 3344 Blood Feast , 2483 Day of the Beast, The ,
cluster number 7:
928 Rebecca , 3038 Face in the Crowd, A , 2201 Paradine Case, The , 1219 Psycho , 3359 Breaking Away , 668 Pather Panchali , 3492 Son of the Sheik, The ,
cluster number 8:
3105 Awakenings ,
cluster number 9:
514 Ref, The , 1414 Mother , 3824 Autumn in New York , 1888 Hope Floats ,

cluster number 10:
2528 Logan's Run ,
cluster number 11:
2607 Get Real , 700 Angus , 24 Powder ,
cluster number 12:
1380 Grease , 3600 Blue Hawaii , 2877 Tommy ,
cluster number 13:
306 Three Colors: Red ,
cluster number 14:
3382 Song of Freedom ,
cluster number 15:
3666 Retro Puppetmaster ,
cluster number 16:
1316 Anna ,
cluster number 17:
1884 Fear and Loathing in Las Vegas ,
cluster number 18:
216 Billy Madison , 216 Billy Madison , 370 Naked Gun 33 1/3: The Final Insult , 1456 Pest, The ,
cluster number 19:
2923 Citizen's Band ,
cluster number 20:
527 Schindler's List ,
cluster number 21:
3121 Hitch-Hiker, The ,
cluster number 22:
2471 Crocodile Dundee II ,
Total Cost: 894.5876097560888

the clusters of the improved algorithm are:
cluster number 1:
3165 Boiling Point , 997 Caught , 423 Blown Away , 2220 Manxman, The , 1499 Anaconda , 403 Two Crimes , 692 Solo , 1867 Tarzan and the Lost City , 3020 Falling Down , 2347 Pope of Greenwich Village, The , 611 Hellraiser: Bloodline , 2468 Jumpin' Jack Flash , 1791 Twilight , 403 Two Crimes , 1112 Palookaville , 3560 Phantom Love ,
cluster number 2:
1514 Temptress Moon , 68 French Twist ,
cluster number 3:
2090 Rescuers, The , 238 Far From Home: The Adventures of Yellow Dog , 3213 Batman: Mask of the Phantasm , 577 Andre , 2123 All Dogs Go to Heaven , 1592 Air Bud , 2042 D2: The Mighty Ducks , 241 Fluke , 1566 Hercules ,
cluster number 4:
99 Heidi Fleiss: Hollywood Madam , 1420 Message to Love: The Isle of Wight Festival , 3237 Kestrel's Eye ,
cluster number 5:
3514 Joe Gould's Secret , 3915 Girlfight , 1850 I Love You, Don't Touch Me! , 1650 Washington Square , 1053 Normal Life , 3783 Croupier ,
cluster number 6:
3266 Man Bites Dog , 3011 They Shoot Horses, Don't They? , 306 Three Colors: Red , 106 Nobody Loves Me , 2972 Red Sorghum , 2925 Conformist, The , 2201 Paradine Case, The , 715 Horseman on the Roof, The ,

cluster number 7:
742 Thinner , 1999 Exorcist III, The , 2560 Ravenous , 1998 Exorcist II: The Heretic , 23 Assassins , 3081 Sleepy Hollow , 3344 Blood Feast , 2483 Day of the Beast, The , 600 Love and a .45 ,
cluster number 8:
928 Rebecca , 2187 Stage Fright , 1219 Psycho , 914 My Fair Lady , 2180 Torn Curtain ,
cluster number 9:
267 Major Payne , 514 Ref, The , 700 Angus , 1884 Fear and Loathing in Las Vegas , 216 Billy Madison , 216 Billy Madison , 370 Naked Gun 33 1/3: The Final Insult , 1456 Pest, The ,
cluster number 10:
3105 Awakenings , 1611 My Own Private Idaho , 1880 Lawn Dogs ,
cluster number 11:
3038 Face in the Crowd, A , 1250 Bridge on the River Kwai, The , 3359 Breaking Away , 668 Pather Panchali , 232 Eat Drink Man Woman ,
cluster number 12:
1208 Apocalypse Now , 2607 Get Real ,
cluster number 13:
1414 Mother , 1888 Hope Floats ,
cluster number 14:
2528 Logan's Run ,
cluster number 15:
1380 Grease , 3600 Blue Hawaii , 2877 Tommy ,
cluster number 16:
3333 Killing of Sister George, The , 2238 Seven Beauties ,
cluster number 17:
3824 Autumn in New York , 527 Schindler's List , 129 Pie in the Sky ,
cluster number 18:
3382 Song of Freedom ,
cluster number 19:
3666 Retro Puppetmaster ,
cluster number 20:
1316 Anna ,
cluster number 21:
2930 Return with Honor ,
cluster number 22:
24 Powder ,
cluster number 23:
2948 From Russia with Love ,
cluster number 24:
2923 Citizen's Band ,
cluster number 25:
3326 What Planet Are You From? ,
cluster number 26:
3492 Son of the Sheik, The ,
cluster number 27:
3121 Hitch-Hiker, The ,
cluster number 28:
2471 Crocodile Dundee II ,
Total Cost: 872.4589212887149

3. We had a big challange finding an algorithm that will prefrom better than the original algorithm for the vast majorty of the times.

At the beginning we tried some complex crossed data conditions which didnt preduce the results we wanted for most of our random files.

So we chose to go "simple" and simply improve the condition in the original algorithm and to cross it with the genre, and we found it was successful in all of our 20 random files.

second of all we found certain difficulties in reading certain files and it took us a while until we figured out a good way to read and parse the data files.

4. Our 3 "special" subset files are:

1. subset1.txt which contains only movies that atleast one of their genres is Comedy, in this example the improved algorithm will produce the same result as the basic one. // output:

Movie numbers are:

1, 3, 4, 5, 7, 11, 12, 19, 21, 34, 38, 39, 45, 52, 54, 63, 64, 65, 68, 69, 70, 72, 75, 84, 87, 88, 93, 96, 101, 102, 104, 106, 107, 109, 115, 118, 119, 122, 125, 129, 133, 135, 141, 144, 153, 156, 157, 166, 171, 174, 176, 178, 180, 186, 187, 189, 195, 203, 205, 212, 216, 218, 223, 224, 228, 231, 232, 234, 235, 236, 237, 239, 243,

the clusters of the basic algorithm are:

cluster number 1:

1 Toy Story , 11 American President, The , 34 Babe , 39 Clueless , 68 French Twist , 96 In the Bleak Midwinter , 106 Nobody Loves Me , 125 Flirting With Disaster , 135 Down Periscope , 239 Goofy Movie, A ,

cluster number 2:

3 Grumpier Old Men , 5 Father of the Bride Part II , 7 Sabrina , 38 It Takes Two , 64 Two if by Sea , 70 From Dusk Till Dawn , 84 Last Summer in the Hamptons , 93 Vampire in Brooklyn , 118 If Lucy Fell , 122 Boomerang , 141 Birdcage, The , 144 Brothers McMullen, The , 166 Doom Generation, The , 186 Nine Months , 195 Something to Talk About , 228 Destiny Turns on the Radio , 236 French Kiss , 237 Forget Paris ,

cluster number 3:

4 Waiting to Exhale , 45 To Die For , 54 Big Green, The , 63 Don't Be a Menace to South Central While Drinking Your Juice in the Hood , 65 Bio-Dome , 69 Friday , 87 Dunston Checks In , 88 Black Sheep , 102 Mr. Wrong , 109 Headless Body in Topless Bar , 115 Happiness Is in the Field , 157 Canadian Bacon , 174 Jury Duty , 187 Party Girl , 203 To Wong Foo, Thanks for Everything! Julie Newmar , 205 Unstrung Heroes , 218 Boys on the Side , 224 Don Juan DeMarco , 231 Dumb & Dumber , 234 Exit to Eden ,

cluster number 4:

12 Dracula: Dead and Loving It , 19 Ace Ventura: When Nature Calls , 75 Big Bully , 104 Happy Gilmore , 189 Reckless , 216 Billy Madison ,

cluster number 5:

21 Get Shorty , 52 Mighty Aphrodite , 129 Pie in the Sky , 235 Ed Wood ,

cluster number 6:

72 Kicking and Screaming , 101 Bottle Rocket , 156 Blue in the Face , 171 Jeffrey , 176 Living in Oblivion , 178 Love & Human Remains , 180 Mallrats ,

cluster number 7:

107 Muppet Treasure Island , 153 Batman Forever , 212 Bushwhacked ,

cluster number 8:

119 Steal Big, Steal Little ,

cluster number 9:

133 Nueba Yol ,

cluster number 10:

223 Clerks ,

cluster number 11:
232 Eat Drink Man Woman ,
cluster number 12:
243 Gordy ,
Total Cost: 631.9162488905495

the clusters of the improved algorithm are:
cluster number 1:
1 Toy Story , 11 American President, The , 34 Babe , 39 Clueless , 68 French Twist , 96 In the Bleak Midwinter , 106 Nobody Loves Me , 125 Flirting With Disaster , 135 Down Periscope , 239 Goofy Movie, A ,
cluster number 2:
3 Grumpier Old Men , 5 Father of the Bride Part II , 7 Sabrina , 38 It Takes Two , 64 Two if by Sea , 70 From Dusk Till Dawn , 84 Last Summer in the Hamptons , 93 Vampire in Brooklyn , 118 If Lucy Fell , 122 Boomerang , 141 Birdcage, The , 144 Brothers McMullen, The , 166 Doom Generation, The , 186 Nine Months , 195 Something to Talk About , 228 Destiny Turns on the Radio , 236 French Kiss , 237 Forget Paris ,
cluster number 3:
4 Waiting to Exhale , 45 To Die For , 54 Big Green, The , 63 Don't Be a Menace to South Central While Drinking Your Juice in the Hood , 65 Bio-Dome , 69 Friday , 87 Dunston Checks In , 88 Black Sheep , 102 Mr. Wrong , 109 Headless Body in Topless Bar , 115 Happiness Is in the Field , 157 Canadian Bacon , 174 Jury Duty , 187 Party Girl , 203 To Wong Foo, Thanks for Everything! Julie Newmar , 205 Unstrung Heroes , 218 Boys on the Side , 224 Don Juan DeMarco , 231 Dumb & Dumber , 234 Exit to Eden ,
cluster number 4:
12 Dracula: Dead and Loving It , 19 Ace Ventura: When Nature Calls , 75 Big Bully , 104 Happy Gilmore , 189 Reckless , 216 Billy Madison ,
cluster number 5:
21 Get Shorty , 52 Mighty Aphrodite , 129 Pie in the Sky , 235 Ed Wood ,
cluster number 6:
72 Kicking and Screaming , 101 Bottle Rocket , 156 Blue in the Face , 171 Jeffrey , 176 Living in Oblivion , 178 Love & Human Remains , 180 Mallrats ,
cluster number 7:
107 Muppet Treasure Island , 153 Batman Forever , 212 Bushwhacked ,
cluster number 8:
119 Steal Big, Steal Little ,
cluster number 9:
133 Nueba Yol ,
cluster number 10:
223 Clerks ,
cluster number 11:
232 Eat Drink Man Woman ,
cluster number 12:
243 Gordy ,
Total Cost: 631.9162488905495

2. subset2.txt which contains 20 movies that every pair of movies from the subset dont share the same genre. in this extreme example the improved algorithm fail to improve the result of the basic algorithm, by one point. // output:

Movie numbers are:
5, 9, 14, 18, 28, 37, 152, 210, 681, 720, 792, 918, 941, 1426, 1464, 1450, 1570, 2727,
the clusters of the basic algorithm are:
cluster number 1:
5 Father of the Bride Part II ,
cluster number 2:
9 Sudden Death ,
cluster number 3:
14 Nixon , 18 Four Rooms ,
cluster number 4:
28 Persuasion ,
cluster number 5:
37 Across the Sea of Time , 792 Hungarian Fairy Tale, A , 1464 Lost Highway ,
cluster number 6:
152 Addiction, The ,
cluster number 7:
210 Wild Bill , 941 Mark of Zorro, The ,
cluster number 8:
681 Clean Slate ,
cluster number 9:
720 Wallace & Gromit: The Best of Aardman Animation ,
cluster number 10:
918 Meet Me in St. Louis ,
cluster number 11:
1426 Zeus and Roxanne ,
cluster number 12:
1450 Prisoner of the Mountains , 1570 Tetsuo II: Body Hammer ,
cluster number 13:
2727 Killer's Kiss ,
Total Cost: 167.12615011310737

the clusters of the improved algorithm are:
cluster number 1:
5 Father of the Bride Part II ,
cluster number 2:
9 Sudden Death ,
cluster number 3:
14 Nixon ,
cluster number 4:
18 Four Rooms ,
cluster number 5:
28 Persuasion ,
cluster number 6:
37 Across the Sea of Time ,
cluster number 7:
152 Addiction, The ,
cluster number 8:
210 Wild Bill ,
cluster number 9:
681 Clean Slate ,
cluster number 10:
720 Wallace & Gromit: The Best of Aardman Animation ,

cluster number 11:
792 Hungarian Fairy Tale, A ,
cluster number 12:
918 Meet Me in St. Louis ,
cluster number 13:
941 Mark of Zorro, The ,
cluster number 14:
1426 Zeus and Roxanne ,
cluster number 15:
1464 Lost Highway ,
cluster number 16:
1450 Prisoner of the Mountains ,
cluster number 17:
1570 Tetsuo II: Body Hammer ,
cluster number 18:
2727 Killer's Kiss ,
Total Cost: 168.75584759018935

3. subset3.txt this file refers to the original algorithm. in this file every movie, along with the first movie in the file as a pivot, will pass the condition for the clustering and therefore the original algorithm will produce a single
cluster.

Movie numbers are:

1, 11, 13, 34, 39, 48, 68, 96, 99, 106, 120, 125, 135, 137, 824, 837, 844, 876, 888, 917, 963, 977, 985, 986,

the clusters of the basic algorithm are:

cluster number 1:
1 Toy Story , 11 American President, The , 13 Balto , 34 Babe , 39 Clueless , 48 Pocahontas , 68 French Twist , 96 In the Bleak Midwinter , 99 Heidi Fleiss: Hollywood Madam , 106 Nobody Loves Me , 120 Race the Sun , 125 Flirting With Disaster , 135 Down Periscope , 137 Man of the Year , 824 Kaspar Hauser , 837 Matilda , 844 Story of Xinghua, The , 876 Police Story 4: Project S , 888 Land Before Time III: The Time of the Great Giving , 917 Little Princess, The , 963 Inspector General, The , 977 Moonlight Murder , 985 Small Wonders , 986 Fly Away Home ,
Total Cost: 238.05355074715564

the clusters of the improved algorithm are:

cluster number 1:
1 Toy Story , 11 American President, The , 13 Balto , 34 Babe , 39 Clueless , 48 Pocahontas , 68 French Twist , 96 In the Bleak Midwinter , 106 Nobody Loves Me , 125 Flirting With Disaster , 135 Down Periscope , 837 Matilda , 888 Land Before Time III: The Time of the Great Giving , 917 Little Princess, The , 986 Fly Away Home ,

cluster number 2:
99 Heidi Fleiss: Hollywood Madam , 137 Man of the Year ,
cluster number 3:
120 Race the Sun ,
cluster number 4:
824 Kaspar Hauser ,
cluster number 5:
844 Story of Xinghua, The ,
cluster number 6:
876 Police Story 4: Project S ,
cluster number 7:
963 Inspector General, The ,
cluster number 8:
977 Moonlight Murder ,
cluster number 9:
985 Small Wonders ,
Total Cost: 220.97689509025892

5.

algorithm: random subset#	Basic algorithm	Improved algorithm
1	923.2670350853801	901.5541087334125
2	914.8518552258155	893.1787771558539
3	893.9923887819472	866.3387282649425
4	945.8370262612322	930.8674784617855
5	885.718568255704	874.7363094249524
6	885.6669766965671	860.1619027860992
7	907.1022332191676	880.5461275895756
8	905.8602875443955	877.5572681940237
9	921.590712338925	891.7081536025072
10	924.2312032717609	893.7582351992992
11	949.0085958820782	901.4849106071335
12	872.6817472944222	858.3729259163543
13	932.2223593459893	926.11738954937
14	933.281350419605	919.3800483192769
15	894.5210369292489	872.2685877730514
16	890.7436039775055	869.5308359881072
17	901.7901969982008	883.4214349182254
18	920.3055407319802	884.2619790687536
19	889.8231381119302	873.7883711412801
20	909.9980658911556	882.0795667817796
average	910.1258963214798	887.0158723421349