Web Search Project 2

Student: Hsuan-Chih, Chen

ID: W1116621

Section 1

1.1

Cosine Similarity

My cosine similarity generally has the best performance than pearson and item base. I implemented K neighbors to the cosine similarity, and try 3 different K, K=20, 100, 200. there is a table shown below:

Cosine			
	k=20	**k=100**	k=200
MAE of GIVEN 5	0.884081530573965	**0.855570839064649**	0.855570839064649
MAE of GIVEN 10	0.867333333333333	**0.791666666666667**	0.7925
MAE of GIVEN 20	0.890421529854346	**0.768399729912221**	0.769846628725764
OVERALL MAE	0.88265473649647	**0.802741750123133**	0.803562633393531

I found if k is too small, the error will be too big. k equals to 100 is the best. Its MAE around 0.802. But actually k = 200 is not so different from k= 100. It shows that we should consider more training data to make our prediction better.

Pearson Correlation

Pearson Correlation		
	Use original Ave of each user	Use common term Ave of each user
MAE of GIVEN 5	0.916843816431162	0.927472802300863
MAE of GIVEN 10	0.823	0.832833333333333
MAE of GIVEN 20	0.777081122793479	0.786341275200154
OVERALL MAE	**0.834263667706452**	**0.84411426695124**

My Pearson correlation is not better than cosine. In common, Pearson should be better because it calculate relative similarity between two users. Cosine calculate absolute similarity in contrast. However, I think because the rating’s range is not big enough. It’s just 1 to 5 so relative similarity may not work very well than absolute similarity. It is also probable that data set may not be big enough for learning. Besides, I try to use common term average instead of total average. I just focus on the rating pattern of 5 , 10, 20 common movies. However, it doesn’t work as I expected.

1.2

Pearson Correlation with IUF (Inverse user frequency) and Case modification

Pearson Correlation with IUF and Case modification			
	IUF	Case(p = 1.5)	IUF & Case
MAE of GIVEN 5	0.919094660497687	0.922470926597474	0.921595598349381
MAE of GIVEN 10	0.826	0.826833333333333	0.833166666666667
MAE of GIVEN 20	0.784315616861194	0.786341275200154	0.792804089900646
OVERALL MAE	**0.838819569857166**	**0.840994910523724**	**0.845017238548678**

My IUF and Case both doesn't improve my original Pearson Correlation. And IUF & Case's performance is not better than only IUF or only Case.

Section2

Item-Based

Item Based	
MAE of GIVEN 5	1.37514067775416
MAE of GIVEN 10	1.538
MAE of GIVEN 20	1.49686505257066
OVERALL MAE	1.46704153669348

My item based has the worst performance. But I don't think I implement something wrong. In common, item based is supposed to be better when items' (movies) quantity are larger than users. However, by observing this data set, users actually don't have many common movies and lots of movies ratings are empty. Therefore, it may increase prediction difficulties. Perhaps when we use even larger data set, the item based algorithm will be better than user based algorithm.

Section3

My own aAgorithm		
	PureCos 0.6 *Peason(P=1.5) 0.4	PureCos 0.4 *Peason(P=1.5) 0.6
MAE of GIVEN 5	0.79942478429411	0.862948605727148
MAE of GIVEN 10	0.779333333333333	0.794833333333333
MAE of GIVEN 20	0.779010321211537	0.767531590624096
OVERALL MAE	0.785790510589394	0.805573797406009

My algorithm is doing different combinations of Cosine, Pearson, itemBased, IUF, and Case modification. I try IUF and Case modication in Cosine as well. Combination really works somehow. My best performance is 0.6 pure Cosine without IUF and case modification plus 0.4 Pearson with p = 1.5. I got around 0.775.