**COEN 272 Web Search and Information Retrieval**

**Project II**

Student Name: Nitya Navali

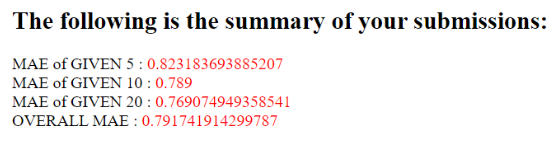
SCU ID : W1556691

**1. User-Based Collaborative Filtering Algorithms**

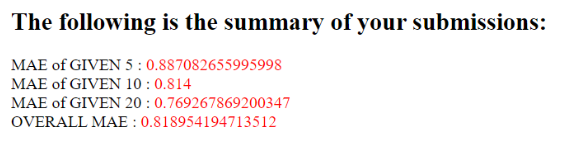
**1.1 Implement the basic user-based collaborative filtering algorithms**

Implemented 2 basic user-based collaborative filtering algorithm with the idea behind that “Similar users rate similarly”. Tested the algorithm for cases 100,80,50 neighbors. And observed that 50 neighbors gave best results. Cosine with 50 neighbors gave better results than person with 50 neighbors. While calculating similarity we consider only the neighbors who have rated that particular movie. If a user does not have neighbors, then average rating of the user is considered as rating for a movie.

**Cosine with 50 neighbors**



**Pearson with 50 neighbors**



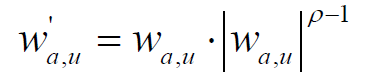
There is one disadvantage of using cosine, if the vector is one dimensional the similarity will always be 1. This can produce incorrect results. So further we shall explore options with Pearson to see if it gives better results.

**1.2 Extensions to the basic user-based collaborative filtering algorithms**

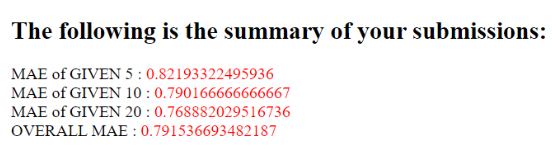
Improving the predictions by applying case amplification and IUF

**Case Amplification**

Case amplification refers to a transform applied to the weights used in the basic collaborative filtering prediction. The transform emphasizes high weights and punishes low weights. Case amplification power value chosen is **2.5**

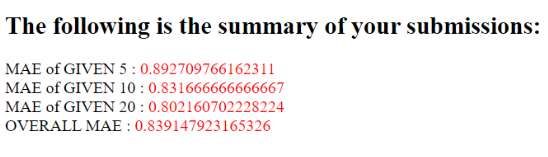


**Cosine with Case Amplification considering 50 neighbors**



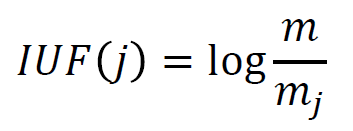
A small improvement can be observed from cosine (**MAE :0.791741914299787**) to cosine with case amp (**MAE: 0.791536693482187**) . A decreased error rate of **0.0002052208176** isobserved.

**Pearson with Case amplification considering 50 neighbors**



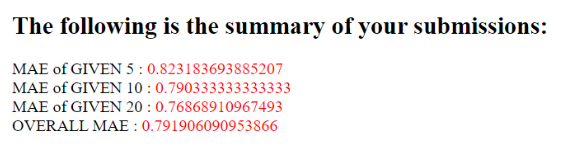
**Inverse User Frequency (IUF)**

Based on the fact that Universally rated movies are not as useful in capturing similarity as less common movies. One can penalize by multiplying the original ratings with factor IUF which is

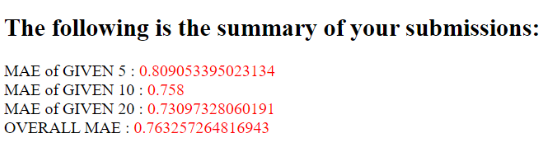


Where m is total no of users and mj is no of users who have rated movie j.

**Cosine with IUF considering 50 neighbors**



**Pearson with IUF considering 50 neighbors**

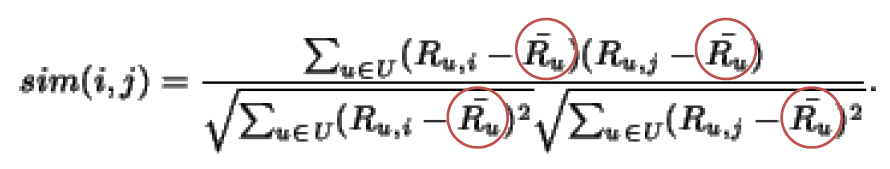


Pearson showed much improvement than cosine. As in Pearson we consider preferences than actual rating i.e a user with similarity -1.0 is as good as user with similarity 1.0 as we consider reverse relation also to be good case to predict rating. Whereas in cosine the similarity with -1.0 is considered least similarity.

**2. Item-Based Collaborative Filtering Algorithm with adjusted cosine similarity**

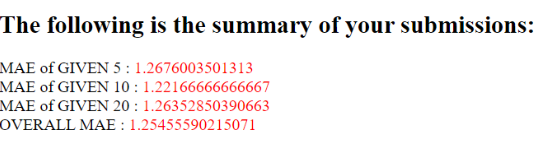
Implemented item-based collaborative filtering algorithm with the idea behind that “Similar items are rated similarly”. Predictions is made on the basis of “Rather than matching the active user to similar customers, finding items that get similar ratings”.

Similarity is calculated using the formula



And prediction is done using weighted average formula.

The results are as follows



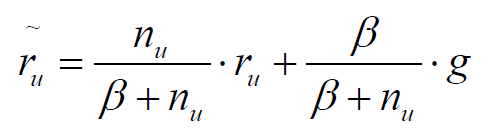
Item based gave high error rate due to lack of enough data points to give good estimate. As there are only 200 users the algorithm has only 200 records to provide estimate and since the matrix is very sparse the data available is not enough. Whereas in user based 1000 ratings per user is available to provide estimate.

Item based algorithm will give good results if there are more users and user based algorithm will provide betters results if items are more.

**3. Implement your own algorithm**

I tried different ways to improve the accuracy of results

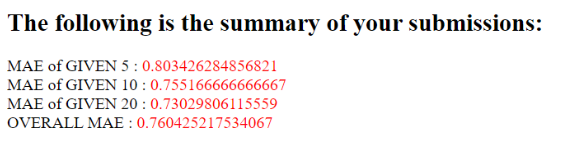
1. Using Cosine user-based algorithm, the similarity for only one dimensional vector is calculated by Euclidean distance and for rest using regular cosine similarity formula.
2. In pearson when calculating users average rating. If user has rated very few movies then the average tends to be noisier. So centering technique is applied in this case. The average user rating is calculated using the formula where g is global rating and ru is users average rating.



1. Ensemble approach is used where results from different techniques is used to predict rating. After trying different combinations, the best results were produced for following combination. Weighted average is used while predicting the ratings. Since Pearson with IUF had best results so far, its given highest weights and for rest weights are assigned as per the accuracy they produced.

|  |  |
| --- | --- |
| Method | Weights |
| Pearson with IUF | 5 |
| Cosine with case amplification | 2 |
| Cosine with 50 neighbors | 1 |
| Cosine with IUF | 1 |

Among these 3 methods, ensemble produced best results which are as follows



**Overall Summary**

|  |  |
| --- | --- |
| **Algortihm** | **Overall MAE** |
| Item Based Adjusted Cosine | 1.25455590215071 |
| Pearson Case Amplification | 0.839147923165326 |
| Pearson with 50 neighbors | 0.818954194713512 |
| Cosine -IUF | 0.791906090953866 |
| Cosine with 50 neighbors | 0.791741914299787 |
| Cosine Case Amplification | 0.791536693482187 |
| Pearson -IUF | 0.764734854703661 |
| **Custom Ensemble Algorithm** | **0.760425217534067** |

**Ensemble approach gave best results compared to rest as it takes average of top 4 best performing algorithms.**