Recommender Systems For workers

(This is the Document that tells the procedure that we have used while working on recommender system)

Group Members

Asad Fiaz [P14-6079]

Mussawar Al Yasah [P14-6105]

Ussama Hassan [P14-6090]

Collaborative filtering

We are using collaborative filtering technique for Recommendations.Collaborative filtering Recommender systems are special in nature.

If we want to use KNN approach we cannot directly use KNN as in classification because we do not have any features to begin with.

The only thing given is (user ID, item ID, ratings) the interactions between items and the users on the basis of the ratings that have been provided by the users to the items (workers).

So to find nearest neighbors we have to compute the similarity between users with the help of their interactions with items.

The basic idea behind it is.

We say if I like the same thing as you do then our preferences in future will also be the same.

Find similarity between users

Lots of people have researched about which similarity measurements to use. Because no dataset works best on all the similarity measurement algorithms. Due to which, it must be checked that which similarity measurement algorithm works best on our dataset. Some of the Similarity Measurement Algorithms are stated as:

1- Euclidian distance similarity.

$$r_2\,(\mathrm{x},\,\mathrm{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} = \sqrt{\sum_{i=1}^{2}(x_i - y_i)^2}$$

2- Pearson correlation.

$$\mathrm{P,\ C(w,u)} = \frac{\sum_i (r_{w,i} - \bar{r}_w)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_i (r_{w,i} - \bar{r}_w)^2 \sum_i (r_{u,i} - \bar{r}_u)^2}}$$

3- Jacquard similarity or tanimoto

$$T(A,B) = \frac{\mathbf{A.B}}{|\mathbf{A}|^2 + |B|^2 - A.B} = \frac{\sum_{i=1}^{n} A_i x B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i x B_i}$$

4- Cosine Similarity

$$Cos(\theta) = \frac{\sum_{i=1}^{n} A_i x B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Note: We have used all the similarity measurements to conclude a difference in the results.

After we compute the similarity we get users x users similarity matrix we have now found the nearest neighbors, as if we want 10 nearest neighbors we get ten most similar users whose similarity value is maximum and who have best ratings also in our case it will be 5.

Prediction for unknown ratings

As recommender system consists of prediction and similarity measurements. So, for predicting the ratings that are unknown, we first see that which users are most similar and then on the basis of these similarities we apply the following formula to predict the ratings. These ratings basically are predicted while using the similar behaviors among the users. Lets say A and B are two users, who are similar in nature that we found out through similarity measurement and A has given an item X 4 ratings and then we will say that B will also give 4 ratings to the item X. But this can never be correct. So for that we have following formula that uses the similarity and the ratings given by the user u to item N and conclude a predicted rating.

Weighted Sum

Here Si,N represents similarity between the workers and Ru,N represents the Ratings Given by user U.

$$\frac{\sum_{all\ similar\ workers,N}(Si,N\ *Ru,N)}{\sum_{all\ similar\ items,N}(|Si,N|)}$$

Evaluating Ratings Prediction

For evaluation we use RMSE and MAE so the similarity measure whose RMSE value is minimum is the best option for this problem. We are using following formulas for evaluation in which $\hat{r}_{ui}$ represents the predicted ratings for a test set $\mathcal{T}$ of user-item (worker) pairs (u , i) for which the true ratings r_{ui} are known.

RMSE

$$\text{RMSE} = \sqrt{\frac{1}{|\mathscr{T}|} \sum_{(u,i) \in \mathscr{T}} (\hat{r}_{ui} - r_{ui})^2}$$

MAE

$$\text{MAE} = \frac{1}{|\mathscr{T}|} \sum_{(u,i) \in \mathscr{T}} |\hat{r}_{ui} - r_{ui}|$$

Now value of RMSE is too high.(minimum is with MSD and that is 0.9621) Because we have very sparse values as a user has only rated 5 to 10 workers but we will have to predict on all the workers that are say 10,000 and RMSE will be computed over all the predictions. How to reduce this. One way is to use linear algebra. As we have users x items matrix and the matrix is very sparse. To reduce sparsity in matrices linier algebra has a technique called SVD.

SVD divides the matrix into two sub matrices of (users x k) and (k x items) where k is small constant depending on sparsity. Now to predict the ratings we only multiply the vectors from both Matrices.

Note: In our case we get 0.9311 RMSE and 0.7426 MAE that is minimum from all the methods.

Recommendation:

At the end when we predict all the unknown ratings and similarities among the users. Then we select top N Items correspond to the similar users and make a list of that and show as a recommended list.

As youtube oversees that which videos you have see explicitly and implicitly and then they recommend the best similar videos to you.