Recsys Data Analysis

February 17, 2017

1 Results

Table 1: LastFM results

Algorithm	Avg. of RMSE.ByUser	Avg. of RMSE.ByRating	Avg. of Predict.nDCG	Avg. of MRR
ItemItem	1455.425755	4469.689135	0.734481196	0.001738894
PersMean	1225.316632	3052.545397	0.811768122	0.000650303
UserUser	1226.077364	3251.001417	0.743536709	0.001728617

Table 2: Movielens results

Algorithm	Avg. of RMSE.ByUser	Avg. of RMSE.ByRating	Avg. of Predict.nDCG	Avg. of MRR
ItemItem	0.890329159	0.896982711	0.955260097	0.095015162
PersMean	0.920828036	0.93823614	0.949940793	0.002643017
UserUser	0.915197144	0.924923196	0.953311329	0.003773996

Table 3: Jester results

Algorithm	Avg. of RMSE.ByUser	Avg. of RMSE.ByRating	Avg. of Predict.nDCG	Avg. of MRR
ItemItem	0,790645	0,834773	0,951061	0,611854
PersMean	0,826521	0,87686	0,944675	0,61771
UserUser	0,796074	0,836992	0,951285	0,712654

2 Datasets

2.1 LastFM

#Users: 1892 users
#Items: 17632 artists
This dataset contains 92.834 user-artist relations. Each relation represents the number of times a user listened to an artist. In order to run the recommender system algorithms, we decided to consider this counter as a kind of rating.

2.2 Movielens

#Users: 943 users
#Items: 1682 movies
This dataset contains 100.000 user-movie relations. Each relation represents the evaluation of a movie by a user (in a scale from 1 to 5).

2.3 Jester

#Users: 73.495 users
#Items: 100 jokes
This dataset contains 4.1 million of user-joke relations. Each relation represents the evaluation of a

movie by a user (in a scale from -10.0 to 10.0). As the scale is different to the used in Movielens, we decided to normalized the evaluations to a scale of [1,5]. Thus, the results of running the algorithm with Jester and Movielens can be comparable.

2.4 Bookcrossing

#Users: 278.858 users
#Items: 271.379 books
This dataset contains 1.149.780 of user-book relations. Each relation represents the evaluation of a book by a user. Also with this dataset, we normalized the ratings to a scale of [1,5].

3 Metrics

3.1 RMSE

RMSE reflects the difference between the real values and those predicted by the algorithm. It is calculated in two ways:

- Grouped by user: RMSE.byUser

$$\frac{\sum_{\forall\ user\ j} \sqrt{\frac{\sum_{\forall\ rating\ i} err_{ij}^2}{Total\ ratings\ user\ i}}}{Total\ users} \tag{1}$$

- Globally: RMSE.byRating

$$\sqrt{\frac{\sum_{\forall\ user\ j}\sum_{\forall\ rating\ i} err_{ij}^2}{Total\ ratings}} \tag{2}$$

In general, both ways gives similar results.

In the case of LastFM, while grouping by users, the results are approximately a third of those obtained using the Global expression. And it is interesting to see that the values are enormous in comparison to those obtained with Movielens and Jester.

3.2 MRR

This metric measures the ranking quality of the algorithm. This metric gives particular good results for Jester. That seems to be caused by the reduced number of items in the dataset (only 100 jokes). We could say that it is easier to "guess" the correct ranking in Jester than in the other datasets because of it has less items to order. Also it's curious that MRR result for Movielens using CF item-item are approximately 40 times better than the outcome for the same metric, the same dataset, but with other algorithms.

3.3 nDCG

nDCG also reflects the ranking performance of an algorithm. For Jester and Movielens, we can see that it has a similar behavior to that of MRR. But in LastFM dataset, its worst result coincides with the best results of MRR.

4 Algorithms

4.1 CF Item-item

For the rating evaluations (MRR), CF Item-item is the algorithm with the best performance in LastFM and Movielens. But with Jester it is worse than CF User-user.

4.2 PersMean

This algorithm has the best RMSE performance for Jester and Movielens, while in LastFM the other algorithms perform better.

4.3 CF User-user

In general CF User-user has good, but not outstanding, RMSE results with Jester and Movielens. Also, in Jester the best MRR results are obtained by CF User-user.

5 Conclusions

Although we didn't do enough analysis to deduce solid conclusions, we can see that the behavior of the algorithms with LastFM is notoriously different to the one with the other datasets. This is probably a consequence of the fact that the relation between user-item is not an explicit evaluation, like in the other datasets, it is the number of times the user listened to the item (artist). Nevertheless, our intention is to perform grid-search, in order to obtained stronger conclusions.