Collaborative Filtering: an Empirical Analysis

Miquel Perelló, E-mail: miquel.perello.nieto@est.fib.upc.edu

Abstract—This work shows the differences on parameters of different *Collaborative Filters* (CF), comparing rate errors and execution times for each combination of techniques. We will analyse user-based and item-based techniques, and two similarity functions : Pearson's and Spearman correlation coefficients.

Keywords—Recomender System, user-based, item-based, Pearson, Spearman

Introduction

Collaborative Filtering is an extended approach for recommendation systems. These techniques makes recommendations aggregating similarities of users and/or items. Although these techniques are widely used, they have some problems like *cold-start*, *latency problem*, *sparsity problem* or *gray-sheep problem*.

In this work We will focus on different techniques and their results in error rate and computation time. These techniques are user-based and item-based also different similarity functions: Pearson and Spearman. Furthermore, we will see what happens when the number of estimations or users grows.

For all experiments we will use MovieLens dataset, this is a dataset from webside movielens.umn.edu, and contains one hundred thousand ratings from nine hundred users and one thousand six hundred movies.

1 User-based and item-based

These are two collaborative filtering techniques. User-based approach follows two basic steps:

1) Search nearest neighbor user using their rating patterns.
2) Create a prediction based on some of his neighbours.

On the other side, item-based uses items to compute the distances, the steps are:

1) Compute all distances between items based in which items have the same user.
2) Create a prediction based on items of actual user and distances between items computed before.

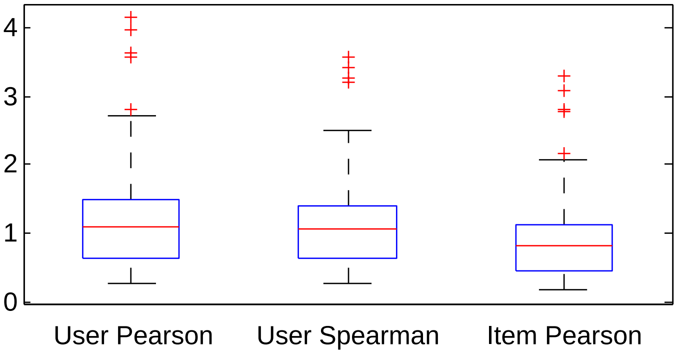

Fig. 1. Comparison of errors with different techniques (RMSE)

For datasets with large number of products and consequently more sparse in the number of similarities between users, item-based approach is usually better performing. Also, it allows to precalculate some of the results in advance.

In our example we can see the differences in error rate between different techniques (see fig. 1). Differences are not so big, but item-based seems to outperform. However this differences are not enough to affirm one is better than other.

Regarding computing time spent in each iteration (user prediction), in that case there is a big difference between user-based and item-based. In most of the cases is one point five or two times more costly for user-based (see fig. 2).

If we focus only on item-based and user-based with Pearson's similarity function we can see that time spent in user-based is always longer (see fig. 3). Concretely, item-based grows with a slope of zero point seven, and mean-

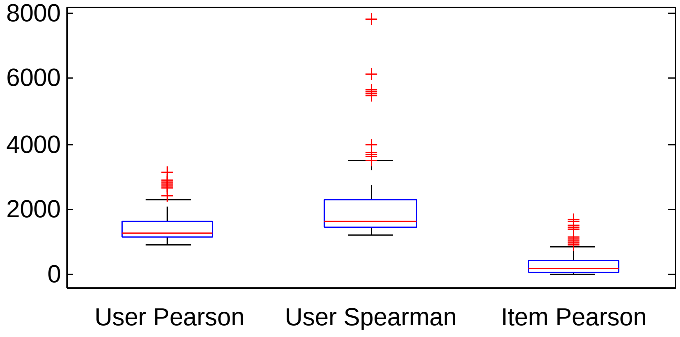

Fig. 2. Comparison of time in miliseconds with different techniques

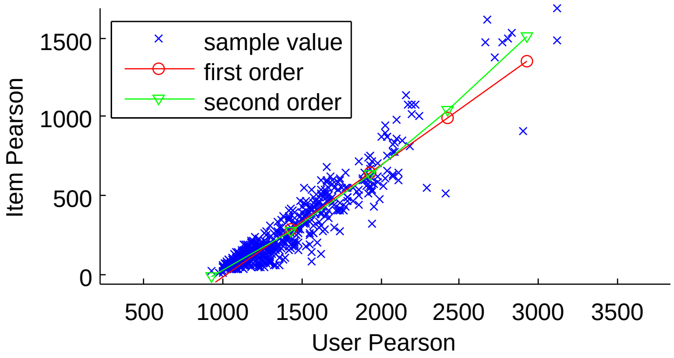

Fig. 3. Time spent in each sample and iteration between user-based and item Pearson

while in some cases item-based time is almost zero, user-based minimum time in our sample is 927 ms.

Polynomial that best fits the data is of first degree:

$$y = 0.7056x - 713.2625 \tag{1}$$

2 Error function

The errors function we are using are *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE). The unique difference between these two functions is the range and grade of errors. Both makes same errors but their value is expressed in different ways.

We can see in the figure 4 about five hundred different pair error samples. It shows the polynomial relation between both variables. It is possible to fit a second degree polynomial to convert one error to the other with no prediction changes.

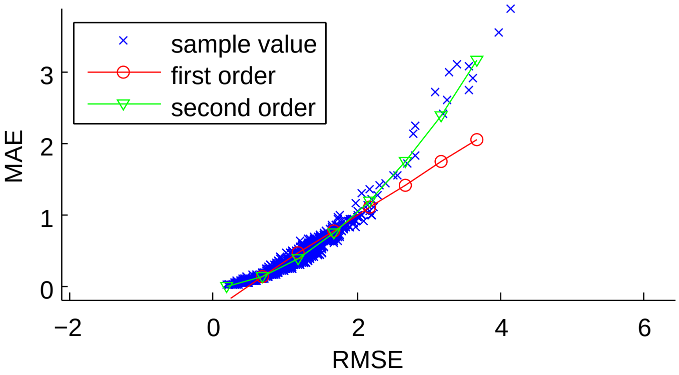

Fig. 4. Error function comparison

$$y = 0.2079x^2 + 0.0967x - 0.0193 \tag{2}$$

3 Similarity function

The two similarity functions we are comparing are *Pearson's correlation coefficient* and

In Pearson's correlation coefficient is defined as the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

For Spearman's correlation coefficient is computed with previously ranked variables with next equation:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \tag{4}$$

If we compare the results of both using their difference values ($Spearman - Pearson$ error values in fig. 5) we can see all the values are negative. It means Pearson's error is bigger in RMSE and MAE cases. We can conclude that Spearman outperforms the results.

If we do the same for time results we see in that case all values are positive (fig. 6. It means Pearson values are always smaller, therefore there is a trade-off between Spearman outperforming with error results and Pearson with fast computations.

In figure 7 all sample times are paired, the result is that Spearman is half slower than Pearsons algorithm.

$$y = 1.9613x - 830.1996 \tag{5}$$

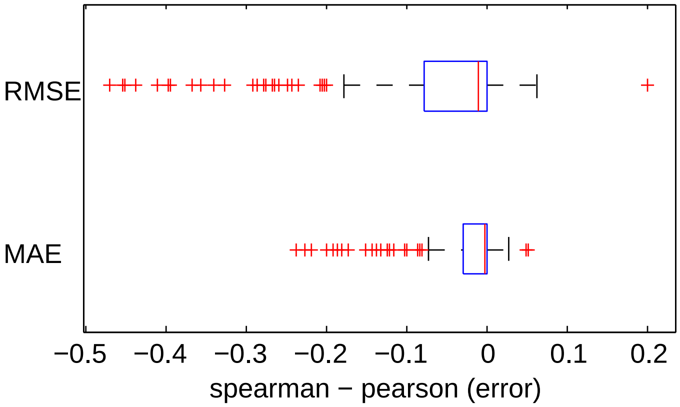

Fig. 5. Difference of errors for user-based Spearman and Pearson

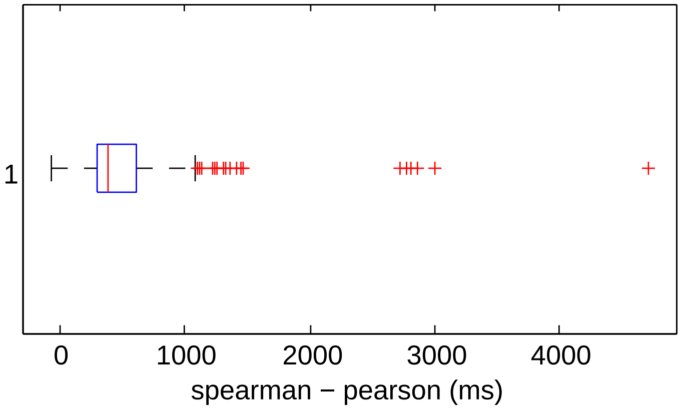

Fig. 6. Difference of time for user-based Spearman and Pearson

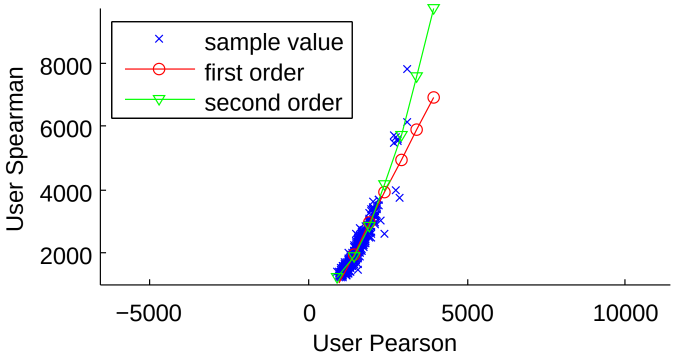

Fig. 7. Time spent in each sample and iteration between user-based Pearson and Spearman

4 Estimations number

This is the number of items removed from users and tried to predict. In all these cases there are twenty users and we increase the estimation number from five to twenty. All error values are extracted from Pearson, Spearman, user-based and item-based. Results shows error rate increasing linearly with a little slope and outliers growing with a bigger slope (see fig. 8).

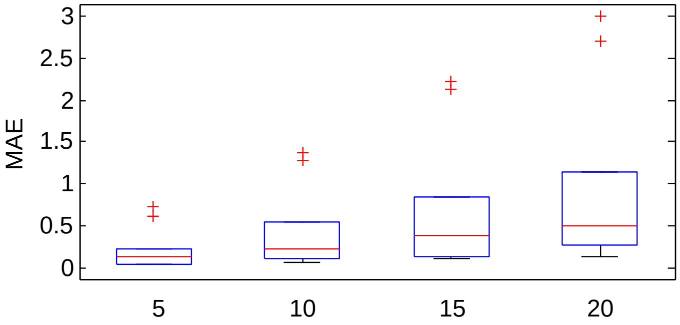

Fig. 8. Error differences increasing number of estimations for twenty users

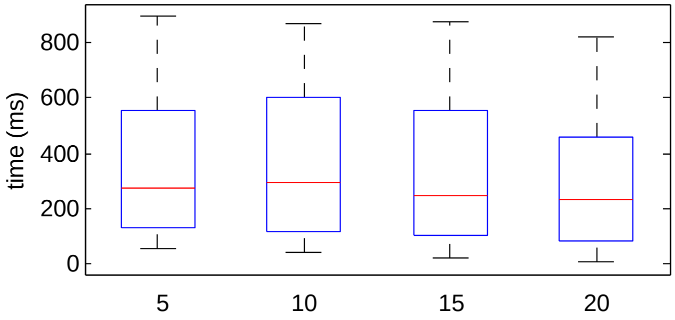

Fig. 9. Time differences increasing number of estimations for twenty users

Concerning the time execution in all the cases time per iteration remains constant (see fig. 9). And the number of iterations is the same, because it is one per user and users remains constant.

5 Users number

Finally the number of users to predict with five estimations for each. We started observing results for all techniques and aggregating the results for each number of users (in range [10, 20, 30, 40, 50, 80, 120, 150] in figure 10). The results on picture are all iteration results (not the sum of iterations that grows linearly) Observing the results we do not see differences between different techniques and we expect all to behave the same for larger numbers. These decision is based in more results and plots that are not added in this document for space reasons, but all of them seems to behave similarly.

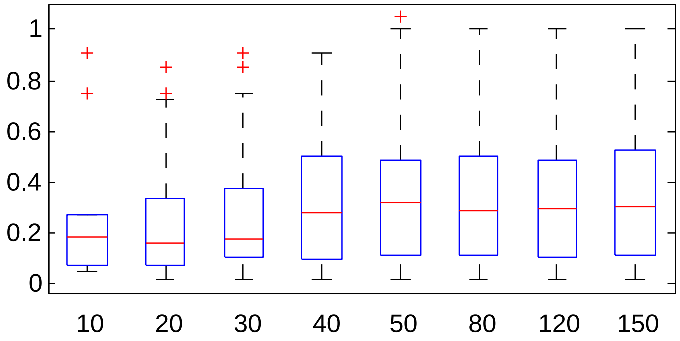

Fig. 10. Mean Absolute Error increasing number of users for five estimations

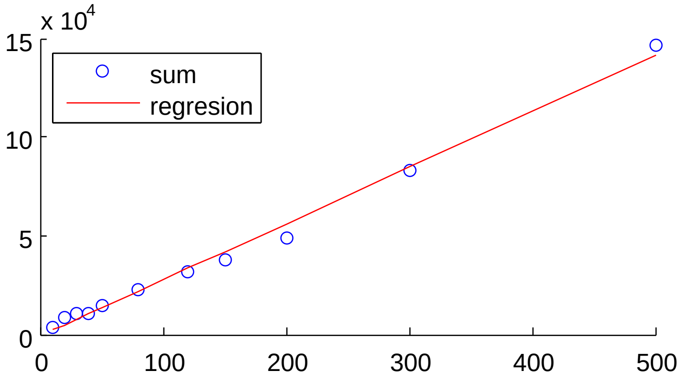

Fig. 11. Total time spent for different number of users and five estimations

We performed larger tests only for item-based with Pearsons similarities. This is because computation time starts growing. Figure 11 shows total computation time for different number of users. It grows linearly with the equation:

$$y = 285.3496x - 792.0340 \tag{6}$$

6 Conclusion

We have seen different collaborative filters and their results changing their error function, similarity function, number of estimations and number of users. Results in this work has shown that item-based with Pearsons similarity function outperforms other techniques. Furthermore we have not seen any difference with error function for prediction, it seems only a linear transformation of their values.

There are other techniques that are not compared in this work, but it will be interesting to test in future works. These are *Content-based* (CN) that uses the products information, *Demographic* (DM) that uses information of users location and *Knowledge-based* (KB) that uses discrimination trees for products.

References

[1] Breese, John S., David Heckerman, and Carl Kadie. ”Empirical analysis of predictive algorithms for collaborative filtering.” Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998.

[2] Sarwar, Badrul, et al. ”Item-based collaborative filtering recommendation algorithms.” Proceedings of the 10th international conference on World Wide Web. ACM, 2001.

[3] Benesty, Jacob, et al. ”Pearson correlation coefficient.” Noise reduction in speech processing. Springer Berlin Heidelberg, 2009. 1-4.

[4] Thornton, GEORGE R. ”The significance of rank difference coefficients of correlation.” Psychometrika 8.4 (1943): 211-222.