Choice of Metrics used in Collaborative Filtering and their Impact on Recommender Systems

J.L. Sánchez, F. Serradilla, E. Martínez, J. Bobadilla

Intelligent Systems Department, Computer Science University, U.P.M., Madrid, Spain

Abstract- The capacity of recommender systems to make correct predictions is essentially determined by the quality and suitability of the collaborative filtering that implements them. The common memory-based metrics are Pearson correlation and cosine, however, their use is not always the most appropriate or sufficiently justified. In this paper, we analyze these two metrics together with the less common mean squared difference (MSD) to discover their advantages and drawbacks in very important aspects such as the impact when introducing different values of k-neighborhoods, minimization of the MAE error, capacity to carry out a sufficient number of predictions, percentage of correct and incorrect predictions and behavior when attempting to recommend the n-best items. The paper lists the results and practical conclusions that have been obtained after carrying out a comparative study of the metrics based on 135 experiments on the MovieLens database of 100,000 ratios.

Index Terms- collaborative filtering, metrics, recommender systems, cosine, correlation

I. INTRODUCTION

At present, recommender systems (RS) comprise an important collaboration tool among Web 2.0 users, providing services that enable accurate recommendations to be made to internet users who have been previously defined according to their preferences in a certain field.

Nowadays, the most broadly-studied RS focus on recommending films [1][2][3], based on votes previously made by users, who have therefore defined their preferences on a limited number of these films. Nevertheless, RS have been introduced to a greater extent in the collaborative systems environment and they cover a large number of applications [4][5][6] in which it is useful to receive recommendations based on the preferences of a group of users with similar tastes or needs to those of each individual that makes use of these systems.

The basic principle of RS is the expectation that the group of users similar to one given user, (i.e. those that have rated an important number of elements in a similar way to the user) can be used to adequately predict that individual’s ratings on products the user has no knowledge of. This way, a trip to Senegal could be recommended to an individual who has rated different destinations in the Caribbean very highly, based on the positive ratings about the holiday destination of “Senegal” of an important number of individuals who also rated destinations in the Caribbean very highly. This suggestion (recommendation) will often provide the user of the service with inspiring information from the collective knowledge of all other users of the service.

In short, the internal operating core of RS is based on carrying out collaborative filtering (CF) [7]: starting from the ratings expressed by a group of users about a group of items, the aim is to select users who have the most similar ratings or tastes to those of the individual who is using the system at any one time. In general, the objective is to suggest a series of elements to the individual, on which this individual have not shown a preference but which have been very highly rated by an important proportion of the group of users with similar preferences to the individual.

The quality of the results offered by a RS greatly depends on the quality of the results provided by its CF [8] phase; i.e. it is essential to be capable of adequately selecting the group of users most similar to a given individual. Two different approaches are available to tackle this important task: memory-based methods and model based methods.

Memory-based methods [9] use metrics [8] that are directly applied to the data matrix that contains the ratings made by the set of users of the system on the set of items available. The current RS for commercial use employ memory-based methods due to their robustness, predictability and efficiency.

Model based methods use data to create a model, (Bayesian classifier [10], neural network [11], fuzzy system [12], etc.) and from this model they predict the clusters of similar users. Model based methods are generally found in the research stage, in the implementation of recommender systems for non-commercial use.

Although a great variety of metrics capable of implementing the CF have been published, most of the systems only use two ways of measuring the distance between users: cosine and Pearson correlation. This is due to the fact that these metrics have obtained the best results in various research projects and that their behavior in RS is satisfactory.

II. IMPROVABLE ASPECTS

Despite the generalized use of cosine and Pearson correlation metrics [8][9] in RS in general and commercial RS in particular, there is no detailed knowledge about the way they work in different situations.

One important factor to consider in these metrics is the choice of the number k of similar users (k-neighborhood) [13] with which the recommendations will be made.

A number k that is too small will give us very biased recommendations (e.g. recommending Senegal based on
the opinions of just 5 individuals). In addition, it is rather unlikely to be able to recommend new items (books, holiday destinations) based on few users, as these users have probably not rated enough elements that are different to those already rated by the individual that is using the RS.

On the other hand, a number \( k \) that is too large can lead to excessive generalization, due to the fact that there will be a proportion of these users that is not similar enough to the reference individual. The system’s efficiency will also be compromised with excessive values of \( k \).

Another situation that has not been sufficiently studied is the behavior of metrics when working with matrices of excessively dispersed ratings (those in which very few ratings have been made by the users).

Most of the research carried out on memory-based metrics has been approached from one sole objective: to maximize the accuracy of the results, however, there are other important objectives that should also be taken into consideration:

- To choose a number \( k \) in accordance with the metrics, the amount of data and its nature.
- To maximize the number of recommendations that can be made.
- To avoid overspecialization (recommending items that are too well-known).
- To maximize the results of the recommendation of the n-best items.
- To maximize the number of perfect recommendations (those which are close to the user’s tastes or needs).
- To minimize the number of bad recommendations (the recommendations that turn out to be contrary to the user’s tastes or needs).

The last two objectives lead us to search for an average value of accuracy [14][15][16][17] that allows for the fact that the distribution of the variance does not have to be even, but rather a greater number of errors of average type can be obtained in exchange for avoiding errors of greater magnitude (bad recommendations) or for encouraging good choices (perfect recommendations), for example.

Finally, we must consider that the real difficulty in carrying out research on the aspects described in this section lies in the fact that there is an interrelation between them. In this way, as an example, parameter \( k \) affects the different metrics in the objectives in a different way: accuracy, bad predictions, perfect predictions, and n-best items.

III. INITIAL ANALYSIS OF THE METRICS

The metrics that are tested in this article are: cosine, Pearson correlation and mean squared difference (MSD) [7][9]. The first two are test because they are the most commonly used by RS and the third because it shows individual characteristics that make it more recommendable in certain circumstances.

In this section, we define the metrics and explain the weak points we have detected in them in order to be able to compare these weak points with the empirical tests carried out.

The similarity of two users \( x \) and \( y \) can be calculated:

\[
\text{cosine:} \quad \text{sim}(x, y) = \frac{\sum_{i} x_{i} y_{i}}{\sqrt{\sum_{i} x_{i}^{2}} \sqrt{\sum_{i} y_{i}^{2}}} \quad (1)
\]

\[i \in I \quad r_{x,i} \neq \phi \text{ and } r_{y,i} \neq \phi\]

\[
\text{correlation:} \quad \text{sim}(x, y) = \frac{\sum (r_{x,i} - \bar{r}_{x})(r_{y,i} - \bar{r}_{y})}{\sqrt{\sum (r_{x,i} - \bar{r}_{x})^{2}} \sum (r_{y,i} - \bar{r}_{y})^{2}} \quad (2)
\]

\[
\text{MSD:} \quad \text{sim}(x, y) = \frac{1}{I} \sum_{i=1}^{I} (r_{x,i} - r_{y,i})^{2} \quad (3)
\]

In the cosine metric, each user \((x \text{ and } y)\) is considered as a vector in a multidimensional space, subsequently the similarity is measured as the cosine of the angle between two users.

In Fig. 1a, the graph shows how the cosine metric uses the angle between 2 vectors to calculate the similarity between two users. If we assume that there are only 2 items (dimension 2), one user will have rated the first item 1 (x axis) and the second item 2 (y axis), whilst the second user will have rated them 2 and 1.

The most serious problem presented by this metric is that users who have rated very different values can be considered as similar users. Fig. 1b shows this anomalous behavior, whereby we can observe that the angle between the vectors will be minimum whenever the values rated between the 2 users are approximately proportional (and proportional is in no way equivalent to similar in the field of SR). In this way, on a scale of 1 to 10, if one user has rated 2 films as very bad (1, 1) and the other user has as excellent (10, 10), the cosine metric will completely fail as it gives the maximum value of similarity between these very different users (Fig. 1b).

The cosine metric works reasonably well in RS because as the number of items increases, the probability of maintaining the proportionality in the ratios of 2 users reduces. However, the problem is still latent and its influence will not be negligible when the ratios’ matrices are very sparse.

The Pearson correlation metric carries out a normalization process with the average of each user’s ratios, which is then adjusted, as there are individuals with a tendency to rate very positively and others with a tendency to rate very negatively.
This normalization process has the effect of reducing the probability of the problem shown in the cosine being presented. However, it does not cancel out it: in the following case, for example: a user who has rated 2 films as (4, 2), in a range of 1 to 20, will obtain the same similarity with a user who has rated (8, 4) as with another who has rated (20, 10).

The MSD metric focuses on achieving a good level of accuracy, however, its greatest vulnerability lies in its poor capacity to make recommendations. More specifically, the problem with this metric lies in the fact that users who have rated very few items easily show a high level of similarity with almost all other users: Let us take, for example, a user who out of 500 cinema films has only rated the last 4 box office hits. The user will probably have similar values in those 4 items to most of the other users; however, the user’s capacity to recommend films is practically non-existent. This harmful effect should reduce as the parameter $k$ increases and/or when only users with a minimum number of rated items are considered (as is the case in many SR).

IV. EMPIRICAL TESTS PERFORMED.
EXPIMENT DESIGN

In order to discover the behavior of each of the three metrics analyzed, we used the “MovieLens” database ¹ [18] as a base, which has been a reference for many years in research carried out in the area of CF.

The database contains 943 users, 1682 items and 100,000 ratings, with a minimum of 20 items rated per user. The items represent cinema films and the rating ranges vary from 1 to 5 stars.

In all the experiments carried out, for each film that each user has rated (100,000 ratings in total), the average value of the ratios given by their $k$-neighborhoods for that film has been calculated and the prediction has been compared with the value rated by the user, thus obtaining the calculation of the mean absolute error (MAE) [8].

\[
\left| \bar{E}_{\text{user}} \right| = \frac{1}{I} \sum_{i=1}^{I} \left| \frac{1}{K} \sum_{k \in \tilde{K}} r_{k,i} - r_{\text{user},i} \right|
\]  

(4)

Where $\tilde{K}$ denotes the set of $k$-neighborhoods on a table of $U$ users who have rated $I$ items.

The previous process was carried out for each of the following $k$ values: 15, 30, 60, 90, 120, 150, 180, 210 and 240, covering from 1.6% to 25% of the total number of users.

In turn, all of these calculations have been carried out 3 times (one for each metric included in the study). The number of experiments carried out is 135, grouped in such a way that the following can be determined:

- Accuracy.
- Number of recommendations made.
- Number of perfect predictions.
- Number of bad predictions.
- Accuracy in predicting the 10 best-items.

We consider a perfect prediction to be each situation in which the prediction of the number of stars recommended for one user in one film matches the value rated by that user for that film.

We consider a bad prediction to be each situation in which the prediction of the number of stars recommended for one user in one film is different by more than 2 stars from the value rated by that user for that film.

We consider a recommendation made to be each situation in which a user has rated an item and at least one of the user’s $k$-neighborhoods has also rated it, in such a way that a prediction could be made and an MAE error obtained.

V. RESULTS

A. Accuracy results

Fig. 2 shows the evolution of the MAE (on a scale of 0 to 4) for each of the three metrics used in the study. As we can see, all the metrics show an error of less than 1 (i.e. on average, the accuracy of the RS offers prediction errors of less than 1 star).

The levels of accuracy of cosine and correlation are very similar: the accuracy of cosine is slightly better (less error), and in both cases, as we had predicted, the error is greater in small values of $k$ (number of users used to make the predictions). The accuracy of the MSD metric is much better than that of the other two metrics, particularly with low values of $k$, where, in all probability, the number of items rated by the k-neighborhoods will be too small.

B. Total number of recommendations made

Fig. 3 shows the percentage of recommendations (predictions) that have been made as regards the total number of rated items in the database (100,000).

In this case, there is no distinction between correct or incorrect predictions: it shows the total prediction capacity of each of the metrics.

1 Our acknowledgement to the GroupLens Research Group
From this perspective, the Pearson correlation offers a huge improvement compared to the other metrics, except when the value of \( k \) is high.

**C. Number of perfect predictions**

The results shown in Fig. 4 are very interesting, as Pearson correlation shows the poorest values of the 3 metrics, leading us to make a critical assessment of the positive results in Fig. 3. Indeed, the real objective is not to be able to make many predictions, but rather to suitably recommend, and in this case MSD is a great improvement on cosine and correlation, particularly at low values of \( k \).

**D. Number of bad predictions**

Fig. 5 shows the percentage of predictions that have proven to be very incorrect. This information is critical as the aim is to minimize these situations in order to prevent the users of the RS from losing faith.

MSD offers very good results, particularly for low and medium values of \( k \), whilst cosine and correlation present their worst values in the lower band of \( k \).

**VI. CONCLUSIONS**

There is no one metric that exceeds all the others in all desirable aspects within a RS.

Despite being the most widely used metrics, Cosine and Pearson correlation present serious deficiencies in some important objectives to be carefully considered in CF-based systems.

In order to be able to choose a specific metric we must first prioritize the type of results we wish to obtain, i.e. we must establish an order of importance among the following contrasting characteristics:

- To maximize the global accuracy (minimize the MAE error).
- To obtain a large number of possible recommendations.
- To give priority to obtaining accurate results (perfect predictions).
Once the prioritization has been established, the combination (metric, k value) that best adapts to obtaining the desired results can be chosen.

The Pearson correlation metric is particularly appropriate when you require a large number of possible predictions with an acceptable level of quality (few incorrect predictions and an acceptable number of correct predictions). It does not work well with small values of k, and its total accuracy is the lowest (MAE is the highest) of the three metrics studied.

The cosine metric shows similar behavior to the Pearson correlation but, as we had predicted, with worse results for low values of k. Its MAE is lower than that of the correlation metric; however, it shows worse results in the number and quality of the predictions.

The MSD metric offers very interesting results due to its different behavior compared to the other two studied and its good results in all aspects except in the number of recommendations it is capable of making, which is undoubtedly its weak point.

VII. REFERENCES


