Contents

1 Research Analysis

http://www.contentwise.tv/files/An-evaluation-Methodology-for-Collaborative-Recommender-Systems.pdf

The pair (p_i, a_i) refers to the prediction on the i-th test instance and the corresponding actual value given by the active user

1.1 Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i}^{n}(p_i - a_i)^2$$

1.2 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(p_i - a_i)^2}$$

1.3 Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i}^{n}|p_i - a_i|$$

With classification metrics we can classify each recommendation such as: a) true positive (TP, an interesting item is recommended to the user) b) true negative (TN, an uninteresting item is not recommended to the user) c) false negative (FN, an interesting item is not recommended to the user) d) false positive (FP, an uninteresting item is recommended to the user)

Precision and *recall* are the most popular metrics in the information retrieval field. They have been adopted, among the others, by Sarwar et al. [17, 18] and Billsus and Pazzani [5]. According to [1], information retrieval applications are characterized by a large amount of negative data so it could be suitable to measure the performance of the model by ignoring instances which are correctly “not recommended” (i.e., TN).

It is possible to compute the metrics as follows:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

These metrics are suitable for evaluating tasks such as top-N recommendation.

When a recommender algorithm predicts the top-N items that a user is expected to find interesting, by using the *recall* we can compute the percentage of known relevant items from the test set that appear in the N predicted items. Basu et al. [2] describe a better way to approximate *precision* and *recall* in top-N recommendation by considering only rated items. According to Herlocker et al. [12], we must consider that:

- usually the number of items rated by each user is much smaller than the items available in the entire dataset
- the number of relevant items in the test set may be much smaller than that one in the whole dataset

Therefore, the value of the *precision* and the *recall* depend heavily on the number of rated items per user and, thus, their values should not be interpreted as absolute measures, but only to compare different algorithms on the same dataset.

1.4 F-mesure

F-measure, used in [18, 17], allows a single measure that combines precision and recall by means of the following relations:

$$F - mesure = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

1.5 Receiver Operating Characteristic

Receiver Operating Characteristic (ROC) is a graphical technique that uses two metrics, true positive rate (TPR) and false positive rate (FPR) defined as

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

to visualize the trade-off between TPR and FPR by varying the length N of the list returned to the user. On the vertical axis the ROC curves plot the TPR, i.e., the number of the instances recommended related to the total number of relevant ones, against the FPR, i.e., the ratio between positively misclassified instances and all the not relevant instances. Thus, by gradually varing the threshold and by repeating the classification process, we can obtain the ROC curve which visualizes the continuous trade-off between TPR and FPR.

Garcia:2012:PET:2109241.2109644

Description of the measures for evaluating our GRS Recommender systems research has used several types of measures for evaluating the quality of the recommendations offered to individuals, such as *precision*, *recall*, or *mean absolute error* (MAE). However, to the best of our knowledge, there is not a widely accepted measure for evaluating GRSs.

For this reason, we have adapted MAE to the group recommendation context because it is the most commonly used and is the easiest to interpret directly [56] when used for evaluating RSs.

First, we define MAE_u, which gives a measure of the deviation of the recommendation for the group with respect to the estimated values for a group member on his own.

Given a recommendation list of N items for a group G such that $u \in G$, the mean absolute error for the user u is defined as follows:

$$MAE^u = \frac{\sum_{i=1}^{N} |d^{ui} - d^{Gi}|}{N}$$

where d^{ui} is the estimated degree of interest of the user u in the item i. This value is obtained by a single-user RS [21,60].

Therefore, MAE^u indicates how adequate the group recommendation is for user u. The lower the MAE^u is, the more accurate the group recommendation is for this user. Unlike individual recommendations, when dealing with groups, a very important issue is to obtain recommendations that are as satisfactory as possible for all the group members, that is, the avoidance of misery.

Therefore, our interest is to measure two aspects:

1. The *satisfaction of the group* as a whole, that is, the *accuracy* of the group recommendation for all the group members. This is achieved by unifying the MAE^u for each group member into a single measure; specifically, we define MAE^G as the average of MAE^u for all the members of the group. Therefore, a low MAE^G indicates that the group as a whole is highly satisfied with the recommendation.

2. The degree to which the group members are equally satisfied with the recommendation. This is achieved by calculating the standard deviation (distance) on MAE^u over all the group members, which we denote as D^G. A low distance represents that all the group members are equally satisfied. That is, this measure could be interpreted as the difference between the satisfaction of each group member with respect to the satisfaction of the other group members

---

Baltrunas, L., Makcinskas, T., & Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. … of the fourth ACM conference on … . Retrieved from http://dl.acm.org/citation.cfm?id=1864733

For evaluating the goodness of a ranked list of recommendations we use Normalized Discounted Cumulative Gain (nDCG), a standard Information Retrieval (IR) measure.

Let $p_1, ..., p_l$ be a ranked list of items produced as an individual or group recommendation. Let u be a user and r_{up_i} the true rating of the user u for the item p_i (ranked in position i, i.e., $\sigma_u(p_i) = i$).

Discounted Cumulative Gain (DCG) and normalized DCG (nDCG) at rank k are defined respectively as:

$$DCG^u_k = r_{up1} + \sum_{i=2}^{k} \frac{r_{up1}}{log_2(i)}$$

$$nDCG^u_k = \frac{DCG^u_k}{IDCG^u_k}$$

where IDCG is the maximum possible gain value for user u that is obtained with the optimal re-order of the k items in $p_1, ..., p_k$.

To compute nDCG we need to know the true user rating for all the items in the recommendation list. Actually, when the test set (items rated by the users) contains only some of the items ranked in the recommendation list one must update the above definition.

In our experiments we computed nDCG on all the items in the test set of the user sorted according to the ranking computed by the recommendation algorithm (individual or group recommendations).

In other words, we compute nDCG on the projection of the recommendation list on the test set of the users.

For example, imagine that $r = [1, 4, 5, 8, 3, 7, 6, 2, 9]$ is a ranked list of recommendations for a group. Since this is a group recommendation list, as we observed above, none of the items in this list occurs in the training set of any group member. Moreover, suppose that the user u test set consists of eight items $1, 4, 7, 8, 9, 12, 14, 20$. In such case, we would compute nDCG on the ranked list $[1, 4, 8, 7, 9]$.

Discounted Cumulative Gain (DCG) is a measure for ranking quality and measures the usefulness (gain) of an item based on its relevance and position in the provided list.

For comparing different lists of recommendations with various lengths, normalized Discounted Cumulative Gain (nDCG/NDCG) is used. It is computed by dividing the DCG by the Ideal Discounted Cumulative Gain or IDCG. The higher the nDCG, the better ranked list.

Normalized discounted cumulative gain (NDCG) measures the performance of a recommendation system based on the graded relevance of the recommended entities. It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities. This metric is commonly used in information retrieval and to evaluate the performance of web search engines.

Important! http://en.wikipedia.org/wiki/Discounted_cumulative_gain

https://github.com/zenogantner/MyMediaLite/blob/master/src/MyMediaLite/Eval/Measures/NDCG.cs
https://www.kaggle.com/wiki/NormalizedDiscountedCumulativeGain

---

Carvalho, L. A. M. C., Cristóvão, S., & Macedo, H. T. (2013). Users ' Satisfaction in Recommendation Systems for Groups : an Approach Based on Noncooperative Games, 951–958.

Let g be the group and R the set of items rated within the group.

A prediction function for group satisfaction for recommended items has been used to evaluate the result.

$$S(g, R) = \frac{\sum_{u \in g} S(u, R)}{|g|}$$

This function is constructed from the average of individual satisfactions of each group member to the list of recommended items.

$$S(u, R) = \frac{\sum_{i \in R} p(u, i)}{|R|}$$

The maximization of $S(g, R)$ means maximizing average satisfaction of the group members to the list of recommended items.