Assignment 1: User-based Collaborative Filtering Recommendations

The assumptions made while completing the assignment:

- It's assumed that obtaining **50 similar users** when making the predictions will provide a reasonable recommendation.

Other important things about the code:

- When selecting 50 similar users for making the predictions, the top 50 users who have rated for a given particular movie were selected.
- When calculating the Pearson correlation between 2 users, only the commonly rated movies were selected. This is done as there are many missing values in the dataset. Also, correlation is calculated if the commonly rated movie count is greater than 2. This is done because we can't say 2 users are similar if they have similar ratings for a very small number of movie count like 1,2.

How to run the code:

The program file is a .py file. When running the code, one requires to give the user ID of the interested user, as the input to the program. Then it'll provide answers to all questions provided in the assignment.

Moreover, the code will display the top 10 similar users using the newly implemented similarity function as well.

E) Design and implement a new similarity function

The newly implemented similarity function is called Improved Heuristic Similarity. [1] This was proposed by Haifeng Liu and the group in 2014.

The widely used similarity functions in collaborative filtering techniques, such as Pearson correlation, cosine similarity, Jaccard similarity, etc., have certain shortcomings. [1] This new similarity function introduces a more extensive framework for user similarity computation, addressing the shortcomings of traditional methods. The improved Heuristic Similarity incorporates several critical elements for a user-based collaborative filtering approach:

- The similarity measure takes into account both the percentage of common ratings and the absolute ratings.
- The resemblance is determined by both the global preference of user behavior and the local circumstances.

The below-mentioned equations are used in implementing the Improved Heuristic Similarity (NHSM). All the equations are taken from [1].

$$sim\,(u,v)^{NHSM} = sim\,(u,v)^{JPSS} \cdot sim\,(u,v)^{URP}$$

$$sim\,(u,v)^{JPSS} = sim\,(u,v)^{PSS} \cdot sim\,(u,v)^{Jaccard'}$$

$$sim\,(u,v)^{URP} = 1 - \frac{1}{1+\exp(-|\mu_u - \mu_v| \cdot |\sigma_u - \sigma_v|)}$$

The user rating preference (URP) considers the preference of each user.

$$sim\,(u,v)^{Jaccard'} = \frac{|I_u \cap I_v|}{|I_u| \times |I_v|}$$

The importance of proportion of common ratings is analyzed through the Jaccard similarity.

$$sim\,(u,v)^{PSS} = \sum_{p \in I} PSS(r_{u,p}, r_{v,p})$$

$$PSS\,(r_{u,p}, r_{v,p}) = Proximity\,(r_{u,p}, r_{v,p}) \times Significance\,(r_{u,p}, r_{v,p}) \times Singularity\,(r_{u,p}, r_{v,p})$$

Here PSS ($r_{u,p}$, $r_{v,p}$) means the PSS value of user u and v.

$$Proximity\,(r_{u,p}, r_{v,p}) = 1 - \frac{1}{1+\exp(-|r_{u,p} - r_{v,p}|)}$$

$$Significance\,(r_{u,p}, r_{v,p}) = \frac{1}{1+\exp(-|r_{u,p} - r_{med}| \cdot |r_{v,p} - r_{med}|)}$$

$$Singularity\,(r_{u,p}, r_{v,p}) = 1 - \frac{1}{1+\exp\left(-\left|\frac{r_{u,p}+r_{v,p}}{2} - \mu_p\right|\right)}$$

PSS consists of *Proximity, Significance, and Singularity* which measures the the distance between two ratings, ratings distance from the median rating, and the differentness of two ratings with other ratings respectively.

[1] Z. H. A. M. H. T. X. Z. Haifeng Liu, "A new user similarity model to improve the accuracy of collaborative filtering," *Knowledge-Based Systems,* pp. 156-166, 2014.

Submitted by

iqra.nasir@tuni.fi

piyumika.herathmudiyanselage@tuni.fi