Let $U = \{u_1, u_2, \ldots, u_m\}$ be the set of users and let $I = \{I_1, I_2, \ldots, I_n\}$ be the set of items, the next similarity measures are defined over item-based rating matrix, in which each row is a rating vector of a specified item, as follows.

Cosine

Cosine which is the most popular similarity measure is specified as follows:

$$\text{sim}(u_1, u_2) = \cos(u_1, u_2) = \frac{u_1 \cdot u_2}{|u_1||u_2|} = \frac{\sum_{j \in I_1 \cap I_2} r_{1j} r_{2j}}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j})^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j})^2}} \quad (2.1)$$

Where $|u_1|$ and $|u_2|$ are lengths of u_1 and u_2, respectively whereas $u_1 \bullet u_2$ is dot product (scalar product) of u_1 and u_2, respectively. If all ratings are non-negative, range of cosine measure is from 0 to 1. If it is equal to 0, two users are totally different. If it is equal to 1, two users are identical. By following the ideology of Jaccard measure, cosine measure is modified as follows:

$$\text{COJ}(u_1, u_2) = \frac{\sum_{j \in I_1 \cap I_2} r_{1j} r_{2j}}{\sqrt{\sum_{j \in I_1} (r_{1j})^2} \sqrt{\sum_{j \in I_2} (r_{2j})^2}} \quad (2.2)$$

Traditional cosine is only determined by the common set of items which are rated by both users but COJ is more general by concerning items which are rated by at least one user. The normalized cosine measure (CON) (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158) is defined as follows:

$$\text{CON}(u_1, u_2) = \text{CPC}(u_1, u_2) = \frac{\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)(r_{2j} - r_m)}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - r_m)^2}}$$

Obviously, CON measure is constrained Pearson correlation (CPC) mentioned later.

Let $v_j = (r_{1j}, r_{2j}, \ldots, r_{mj})$ be vector of rating values that item j receives from m users, for example. The mean of v_j is:

$$\bar{v}_j = \frac{1}{m} \sum_{i=1}^{m} r_{ij}$$

Adjusted cosine measure (COD) is defined as follows:

$$\text{COD}(u_1, u_2) = \frac{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{v}_j)(r_{2j} - \bar{v}_j)}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{v}_j)^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - \bar{v}_j)^2}} \quad (2.4)$$

3. Pearson

Pearson correlation is another popular similarity measure besides cosine, which is defined as follows (Sarwar, Karypis, Konstan, & Riedl, 2001, p. 290):

$$\text{Pearson}(u_1, u_2) = \frac{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{u}_1)(r_{2j} - \bar{u}_2)}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{u}_1)^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - \bar{u}_2)^2}} \quad (3.1)$$

Where $\bar{u}_1$ and $\bar{u}_2$ are mean values of u_1 and u_2, respectively.

$$\bar{u}_1 = \frac{1}{|I_1|} \sum_{j \in I_1} r_{1j}, \bar{u}_2 = \frac{1}{|I_2|} \sum_{j \in I_2} r_{2j}$$

The range of Pearson measure is from –1 to 1. If it is equal to –1, two users are totally opposite. If it is equal to 1, two users are identical. Pearson measure is sample correlation coefficient in statistics. Pearson measure has some variants. Constrained Pearson correlation (CPC) measure

considers impact of positive and negative ratings by using median r_m instead of using the means; for example, if rating values range from 1 to 5, the median is $r_m = (1+5) / 2 = 3$.

Weight Pearson correlation (WPC) measure and sigmoid Pearson correlation (SPC) measure concern how much common items are. WPC is defined as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

$$\mathrm{WPC}(u_1, u_2) = \begin{cases} \mathrm{Pearson}(u_1, u_2) * \frac{|I|}{H}, \text{if } |I| \leq H \\ \mathrm{Pearson}(u_1, u_2), \text{otherwise} \end{cases} \tag{3.3}$$

SPC is defined as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

$$\mathrm{SPC}(u_1, u_2) = \mathrm{Pearson}(u_1, u_2) * \frac{1}{1 + \exp\left(-\frac{|I|}{2}\right)} \tag{3.4}$$

Where H is a threshold and it is often set to be 50 (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158). PearsonJ is combinations of Jaccard and Pearson as follows (Amer and & Nguyen, 2021):

$$\begin{aligned} \mathrm{PearsonJ}(u_1, u_2) &= \mathrm{Pearson}(u_1, u_2) * \mathrm{Jaccard}(u_1, u_2) \\ &= \frac{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{u}_1)(r_{2j} - \bar{u}_2)}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{u}_1)^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - \bar{u}_2)^2}} * \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|} \end{aligned} \tag{3.7}$$

Amer, A. A., & Nguyen, L. (2021). Combinations of jaccard with numerical measures for collaborative filtering enhancement: Current work and future proposal. *arXiv preprint arXiv:2111.12202*.

1. Jaccard

The first measure which is described here is Jaccard because of its special feature when it does not concern magnitude of numeric rating values. Jaccard measure is ratio of cardinality of common set $I_1 \cap I_2$ to cardinality of union set $I_1 \cup I_2$. It measures how much common items both users rated, which is defined as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

$$\mathrm{Jaccard}(u_1, u_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|} \tag{1.1}$$

4. MSD

Mean squared difference (MSD) is defined as inverse of distance between two vectors. Let MAX be maximum value of ratings, MSD is calculated as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

$$\mathrm{MSD}(u_1, u_2) = 1 - \frac{\sum_{j \in I} \left(\frac{r_{1j} - r_{2j}}{\mathrm{MAX}}\right)^2}{|I|} \tag{4.1}$$

Another variant of MSD is specified by some authors as follows:

$$\mathrm{MSD}(u_1, u_2) = \frac{1}{1 + \frac{1}{|I|} \sum_{j \in I} (r_{1j} - r_{2j})^2} \tag{4.2}$$

5. SRC

When rating values are converted into ranks, Spearman's Rank Correlation (SRC) is defined as follows (Hyung, 2008, p. 39):

$$\mathrm{SRC}(u_1, u_2) = 1 - \frac{6\sum_{j \in I} d_j^2}{|I|(|I|^2 - 1)} \tag{5.1}$$

Where d_j is difference between two ranks on item j given by user 1 and user 2.

$$d_j = \mathrm{rank}_{1j} - \mathrm{rank}_{2j}$$

Note, it is easy to convert ratings values to ranks. For example, suppose rating values (bins) are 5, 6, 7, 8, 9 then, we have rank 1 (for value 9), rank 2 (for value 8) , rank 3 (for value 7), rank 4 (for value 6), and rank 5 (for value 5). If user 1 rates value 9 to item j, we have $rank_{1j}$ = 1. The larger the value is, the smaller (higher) the rank is.

7. PSS and NHMS

Liu et al. (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 156) proposed a new similarity measure called NHMS to improve recommendation task in which only few ratings are available. Their NHMS measure (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 160) is based on sigmoid function and the improved PIP measure as PSS (*Proximity – Significance – Singularity*). PSS similarity is calculated as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 160):

$$\begin{aligned}\mathrm{PSS}(u_1, u_2) = \sum_{j \in I} &\mathrm{Proximity}(r_{1j}, r_{2j}) * \mathrm{Significance}(r_{1j}, r_{2j}) \\ &* \mathrm{Singularity}(r_{1j}, r_{2j})\end{aligned} \tag{7.1}$$

Where, $I = I_1 \cap I_2$ is intersection set of I_1 and I_2. The proximity factor determines similarity of two ratings, based on distance between them; such distance is as less as better. The significance factor determines similarity of two ratings, based on distance from them to rating median; such distance is as more as better. The significance factor determines similarity of two ratings, based on difference between them and other ratings; such difference is as less as better. Followings are equations of these factors based on sigmoid function (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 161).

$$\mathrm{Proximity}(r_{1j}, r_{2j}) = 1 - \frac{1}{1 + \exp(-|r_{1j} - r_{2j}|)}$$

$$\mathrm{Significance}(r_{1j}, r_{2j}) = \frac{1}{1 + \exp(-|r_{1j} - r_m||r_{2j} - r_m|)}$$

$$\mathrm{Singularity}(r_{1j}, r_{2j}) = 1 - \frac{1}{1 + \exp\left(-\left|\frac{r_{1j} + r_{2j}}{2} - \mu_j\right|\right)}$$

Note, r_m be median of rating values, for example, if rating values range from 1 to 5, the median is r_m = (1+5) / 2 = 3 whereas μ_j is rating mean of item j. Liu et al. (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 161) also considered the similarity between two users via URP measure as follows:

$$\mathrm{URP}(u_1, u_2) = 1 - \frac{1}{1 + \exp(-|\mu_1 - \mu_2||\sigma_1 - \sigma_2|)} \tag{7.2}$$

Where μ_1 and μ_2 are rating means of user 1 and user 2, respectively and σ_1 and σ_2 are rating standard deviations of user 1 and user 2, respectively.

$$\mu_1 = \frac{1}{|I_1|}\sum_{j \in I_1} r_{1j}\,, \mu_2 = \frac{1}{|I_2|}\sum_{j \in I_2} r_{2j}$$

$$\sigma_1 = \sqrt{\frac{1}{|I_1|}\sum_{j \in I_1}(r_{1j} - \mu_1)^2}\,, \sigma_2 = \sqrt{\frac{1}{|I_2|}\sum_{j \in I_2}(r_{2j} - \mu_2)^2}$$

PSS associated with Jaccard produces a so-called PSSJ measure as follows:

$$\mathrm{PSSJ}(u_1, u_2) = \mathrm{PSS}(u_1, u_2) * \mathrm{Jaccard}(u_1, u_2) \tag{7.3}$$

Where Jaccard is specified by equation 1.1, respectively. Liu et al. (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 161) proposed a new heuristic similarity model (NHSM) as triple product of PSS measure, URP measure, and Jaccard2 measure.

$$\text{NHSM}(u_1, u_2) = \text{PSS}(u_1, u_2) * \text{URP}(u_1, u_2) * \text{Jaccard2}(u_1, u_2) \quad (7.4)$$

In general, Liu et al. (Liu, Hu, Mian, Tian, & Zhu, 2013) aim to alleviate the problem of few rated common items via their NHSM measure. From experimental result, NHSM gave out excellent estimation.

8. BCF

Patra et al. (Patra, Launonen, Ollikainen, & Nandi, 2015, p. 143) proposed a new similarity measure called BCF for CF, which uses all ratings made by a pair of users. Proposed measure finds importance of each pair of rated items by exploiting Bhattacharyya (BC) similarity. The BC similarity, which is core of their own measure, measures the similarity between two distributions. So, these distributions are estimated as the number of uses rated on given item. In general, Patra et al. (Patra, Launonen, Ollikainen, & Nandi, 2015, p. 5) combined BC similarity and the local similarity where the local similarity relates to Pearson correlation. It is necessary to survey BC similarity. Bin is a terminology indicating domain of rating values, for example, if rating values range from 1 to 5, we have bins: 1, 2, 3, 4, 5. Let *m* be the number of bins, given items *i* and *j*, item BC coefficient for items is calculated as follows (Patra, Launonen, Ollikainen, & Nandi, 2015, p. 5):

$$\text{bc}(i, j) = \sum_{h=1}^{m} \sqrt{\frac{\#h_i}{\#i} \frac{\#h_j}{\#j}} \quad (8.1)$$

Note, #*i* and #*j* are the numbers of users who rated items *i* and *j*, respectively whereas $\#h_i$ and $\#h_j$ are numbers of users who gave rating value *h* on items *i* and *j*, respectively. So, item BC coefficient concerns two items. In table 0.2, rating vectors of item 3 and item 4 are v_3 = (1, 2, 5, ?) and v_4 = (5, 4, 5, ?), respectively with note that rating values range from 1 to 5 and so we have:

$$\text{bc}(v_3, v_4) = \sqrt{\frac{1}{4}\frac{0}{4}} + \sqrt{\frac{1}{4}\frac{0}{4}} + \sqrt{\frac{0}{4}\frac{0}{4}} + \sqrt{\frac{0}{4}\frac{1}{4}} + \sqrt{\frac{1}{4}\frac{2}{4}} \cong 0.35$$

According to Patra et al. (Patra, Launonen, Ollikainen, & Nandi, 2015, p. 5), user BC similarity is sum of products of item BC coefficients and local similarities as follows:

$$\text{BCF}(u_1, u_2) = \sum_{i \in I_1} \sum_{j \in I_2} \text{bc}(i, j) \text{loc}(r_{1i}, r_{2j}) \quad (8.2)$$

The local similarity is calculated as a part of constrained Pearson coefficient (CPC) as follows:

$$\text{loc}(r_{1i}, r_{2j}) = \frac{(r_{1i} - r_m)(r_{2j} - r_m)}{\sqrt{\sum_{k \in I_1} (r_{1k} - r_m)^2} \sqrt{\sum_{k \in I_2} (r_{2k} - r_m)^2}}$$

Note, r_m be median of rating values, for example, if rating values range from 1 to 5, the median is r_m = (1+5) / 2 = 3. Patra et al. (Patra, Launonen, Ollikainen, & Nandi, 2015, p. 5) proposed Bhattacharyya similarity in CF (BCF) as sum of user BC similarity and Jaccard measure as follows:

$$\text{BCFJ}(u_1, u_2) = \text{Jaccard}(u_1, u_2) + \text{BC}(u_1, u_2) \quad (8.3)$$

Singh et al. (Singh, Sinha, & Choudhury, 2022) improved BCF measure as follows:

$$\text{BCFJ2}(u_1, u_2) =$$

13. SMTP

Similarity Measure for Text Processing (SMTP) was developed by Lin, Jiang, and Lee (Lin, Jiang, & Lee, 2013), originally used for computing the similarity between two documents in text processing. Here documents are considered as rating vectors. Given two rating vectors $u_1 = (r_{11}, r_{12},\ldots, r_{1n})$ and $u_2 = (r_{21}, r_{22},\ldots, r_{2n})$, the function F of u_1 and u_2 is defined as follows (Lin, Jiang, & Lee, 2013, p. 1577):

$$F(u_1, u_2) = \frac{\sum_{j=1}^{n} A(r_{1j}, r_{2j})}{\sum_{j=1}^{n} B(r_{1j}, r_{2j})} \tag{13.1}$$

Where (Lin, Jiang, & Lee, 2013, p. 1577),

$$A(r_{1j}, r_{2j}) = \begin{cases} 0.5\left(1 + \exp\left(-\left(\frac{r_{1j} - r_{2j}}{\sigma_j}\right)^2\right)\right) \\ \text{if both } r_{1j} \text{ and } r_{2j} \text{ non} - \text{missing} \\ 0 \text{ if both } r_{1j} \text{ and } r_{2j} \text{ missing} \\ -\lambda \text{ otherwise} \end{cases}$$

$$B(r_{1j}, r_{2j}) = \begin{cases} 0 \text{ if both } r_{1j} \text{ and } r_{2j} \text{ missing} \\ 1 \text{ if both } r_{1j} \text{ and } r_{2j} \text{ non} - \text{missing} \end{cases}$$

Note that λ is the pre-defined number and σ_j is the standard deviation of rating values belonging to field j (item j). In this research, λ is set to be 0.5. Lin, Jiang, and Lee (Lin, Jiang, & Lee, 2013, p. 1577) defined SMTP measure based on function F as follows:

$$\text{SMTP}(u_1, u_2) = \frac{F(u_1, u_2) + \lambda}{1 + \lambda} \tag{13.2}$$

PCC

Choi and Suh (Choi & Suh, 2013) proposed a so-called PC measure which is Pearson measure weighted by similarities of items. In other words, PC measure combines similarities of users and items (Patra, Launonen, Ollikainen, & Nandi, 2015, p. 4). The ideology is excellent. PC measure can be applied into any foundation measures. Each factor in PC measure is weighted by a similarity of active item and another item. Suppose it is necessary to estimate rating values of active item k, PC measure (Choi & Suh, 2013, p. 148) is defined as follows:

$$\text{PC}_k(u_1, u_2) = \frac{\sum_{j \in I}\left(\left(\text{sim}(v_k, v_j)\right)^2 (r_{1j} - \bar{u}_1)(r_{2j} - \bar{u}_2)\right)}{\sqrt{\sum_{j \in I_1}\left(\text{sim}(v_k, v_j)(r_{1j} - \bar{u}_1)\right)^2}\sqrt{\sum_{j \in I_2}\left(\text{sim}(v_k, v_j)(r_{2j} - \bar{u}_2)\right)^2}} \tag{3.8}$$

Where sim(v_k, v_j) is similarity of the active item k and item j. Note, sim(v_k, v_j) can be calculated by any measures here. The $\bar{u}_1$ and $\bar{u}_2$ are mean values of u_1 and u_2, respectively.

$$\bar{u}_1 = \frac{1}{|I_1|}\sum_{j \in I_1} r_{1j}, \bar{u}_2 = \frac{1}{|I_2|}\sum_{j \in I_2} r_{2j}$$

Experimental results proved that PC is an effective measure.

9. MMD

Suryakant and Mahara (Suryakant & Mahara, 2016) proposed a so-called Cosine-Jaccard-Mean Measure of Divergence (CjacMD) based on Mean Measure of Divergence (MMD) to solve the problem of sparse rating matrix. Because MMD measure takes advantages of statistical aspects, it can alleviate sparsity. MMD focuses on personal habits which are ignored

by nonstatistical measures (Suryakant & Mahara, 2016, p. 453). Recall that bin is a terminology indicating domain of rating values, for example, if rating values range from 1 to 5, we have bins: 1, 2, 3, 4, 5. Let $X = (x_1, x_2,\ldots, x_b)$ and $Y = (y_1, y_2,\ldots, y_b)$ be count vectors of user 1 and user 2, respectively where x_j (y_j) is the number of items to which user 1 (user 2) gives bin j with note that b is the number of bins. For example, rating vectors of user 1 and user 2 in table 0.1 are $u_1 = (1, 2, 1, 5)$ and $u_2 = (2, 1, 2, 4)$, respectively with note that rating values ranges from 1 to 5. We have:

$$X = (x_1 = 2, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1)$$
$$Y = (y_1 = 1, y_2 = 2, y_3 = 0, y_4 = 1, y_5 = 0)$$

MMD measure is defined as follows (Suryakant & Mahara, 2016, p. 453), (Harris & Sjøvold, 2018, p. 87):

$$\text{MMD}(u_1, u_2) = \frac{1}{1 + \frac{1}{b}\sum_{j=1}^{b}\left(\left(\theta_{1j} - \theta_{2j}\right)^2 - \frac{1}{0.5 + x_j} - \frac{1}{0.5 + y_j}\right)} \tag{9.1}$$

Where θ_{1*} and θ_{2*} are Grewal's transformations (Harris & Sjøvold, 2018, p. 85) of X and Y, respectively.

$$\theta_{1j} = \frac{1}{\sin\left(1 - \frac{2x_j}{|I_1|}\right)}$$
$$\theta_{2j} = \frac{1}{\sin\left(1 - \frac{2y_j}{|I_2|}\right)}$$

In fact, CjacMD (Suryakant & Mahara, 2016, p. 453) combines three other measures such as cosine, Jaccard, and MMD together.

$$\text{CjacMD} = \cos(u_1, u_2) + \text{Jaccard}(u_1, u_2) + \text{MMD}(u_1, u_2) \tag{9.2}$$

Experimental result proved that CjacMD model is effective similarity model.

10. TriangleJ

Sun et al. (Sun, et al., 2017) proposed a so-called Triangle similarity measure which considers both angle and lengths of rating vectors. For instance, given two user vectors u_1 and u_2 are considered as two vector OA = u_1 and OB = u_2 and hence, OAB forms a triangle. TS measure is ratio of the length |AB| to the sum of lengths |OA| + |OB|. Of course, |AB| is always less than or equal to |OA| + |OB| according to triangle inequality. The idea is excellent. TS measure (Sun, et al., 2017, p. 6) is defined as follows:

$$\begin{aligned}\text{Triangle}(u_1, u_2) &= 1 - \frac{|\text{AB}|}{|\text{OA}| + |\text{OB}|} = 1 - \frac{|u_1 - u_2|}{|u_1| + |u_2|} \\ &= 1 - \frac{\sqrt{\sum_{j\in I}\left(r_{1j} - r_{2j}\right)^2}}{\sqrt{\sum_{j\in I} {r_{1j}}^2} + \sqrt{\sum_{j\in I} {r_{2j}}^2}}\end{aligned} \tag{10.1}$$

Sun et al. also combined Triangle measure and Jaccard measure to form a new measure called Triangle multiplying Jaccard (TMJ) measure. The integrated TMJ (Sun, et al., 2017, p. 6) is defined as follows:

$$\text{TMJ}(u_1, u_2) = \text{Triangle}(u_1, u_2) * \text{Jaccard}(u_1, u_2) \tag{10.2}$$

Experimental result proved that TMJ is effective measure.

11. Feng

To solve the problem of sparse rating matrix, Feng et al. (Feng, Fengs, Zhang, & Peng, 2018) proposed a new model of similarity which includes three parts such as S_1, S_2, and S_3. The S_1 (Feng, Fengs, Zhang, & Peng, 2018, p. 6) is normal similarity and they choose cosine as S_1.

$$S_1(u_1, u_2) = \begin{cases} \text{cosine}(u_1, u_2) \text{ if sparsity} < \rho \\ \text{COJ}(u_1, u_2) \text{ otherwise} \end{cases}$$

Where ρ is sparsity threshold which is proposed by Feng et al. The S_2 (Feng, Fengs, Zhang, & Peng, 2018, p. 6) punishes user pairs whose co-rated items are few.

$$S_2(u_1, u_2) = \frac{1}{1 + \exp\left(-\frac{|I_1 \cap I_2|^2}{|I_1||I_2|}\right)}$$

The S_3 (Feng, Fengs, Zhang, & Peng, 2018, p. 6) focuses on statistical feature of user ratings, which reflects essential user favorites. S_3 is aforementioned URP measure.

$$S_3(u_1, u_2) = \text{URP}(u_1, u_2) = 1 - \frac{1}{1 + \exp(-|\mu_1 - \mu_2||\sigma_1 - \sigma_2|)}$$

Where μ_1 and μ_2 are rating means of user 1 and user 2, respectively and σ_1 and σ_2 are rating standard deviations of user 1 and user 2, respectively. The similarity model of Feng et al. (Feng, Fengs, Zhang, & Peng, 2018, p. 5) is product of S_1, S_2, and S_3 as follows:

$$\text{Feng}(u_1, u_2) = S_1(u_1, u_2) * S_2(u_1, u_2) * S_3(u_1, u_2) \quad (11.1)$$

Experimental result proved that Feng model is effective similarity model.

12. Mu

Mu et al. (Mu, Xiao, Tang, Luo, & Yin, 2019) combined local measures (Pearson and Jaccard) with global measure to solve the problem of sparse rating matrix. The global measure is Hellinger (Hg) distance which estimates similarity of two probabilistic distributions. In fact, Hg is inverse of BC coefficient in discrete distributions as follows (Mu, Xiao, Tang, Luo, & Yin, 2019, p. 419):

$$\text{Hg}(u_1, u_2) = 1 - \text{bc}(u_1, u_2) = 1 - \sum_{h=1}^{m} \sqrt{\frac{\#h_1}{\#1}\frac{\#h_2}{\#2}} \quad (12.1)$$

Note, #1 and #2 are the numbers of item which are rated by user 1 and user 2, respectively whereas $\#h_1$ and $\#h_2$ are numbers of items which receive rating value h from user 1 and user 2, respectively. For example, rating vectors of user 1 and user 2 in table 0.1 are u_1 = (1, 2, 1, 5) and u_2 = (2, 1, 2, 4), respectively with note that rating values range from 1 to 5 and so we have:

$$\text{Hg}(u_1, u_2) = 1 - \left(\sqrt{\frac{2}{4}\frac{1}{4}} + \sqrt{\frac{1}{4}\frac{2}{4}} + \sqrt{\frac{0}{4}\frac{0}{4}} + \sqrt{\frac{0}{4}\frac{1}{4}} + \sqrt{\frac{1}{4}\frac{0}{4}}\right) \cong 0.29$$

Given weight α, the Mu measure (Mu, Xiao, Tang, Luo, & Yin, 2019, p. 419) combines Pearson, Jaccard, and Hg as follows:

$$\text{Mu}(u_1, u_2) = \alpha * \text{Pearson}(u_1, u_2) + (1 - \alpha) * \left(\text{Hg}(u_1, u_2) + \text{Jaccard}(u_1, u_2)\right) \quad (12.2)$$

Experimental result proved that Mu measure is effective similarity model.

16. TA

Cosine measure is effective but it has a drawback that there may be two end points of two vectors which are far from each other according to Euclidean distance, but their cosine is high. This is negative effect of Euclidean distance which decreases accuracy of cosine similarity. Therefore, a so-called triangle area (TA) measure (Nguyen & Amer, 2019) is proposed as an improved version of cosine measure. Figure 17.1 illustrates TA measure.

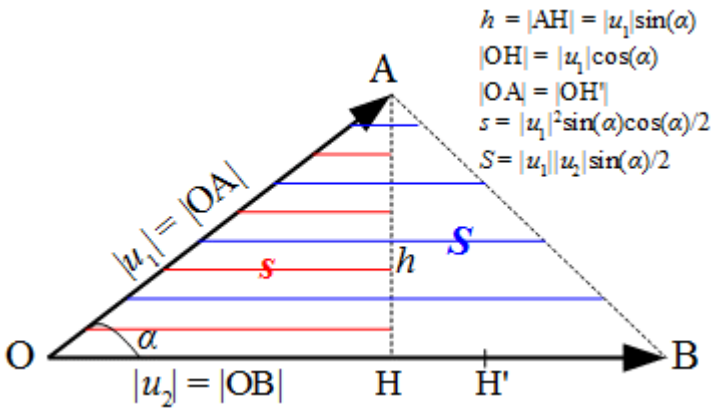

Figure 17.1. Triangle area (TA) measure with $0 \leq \alpha \leq \pi/2$

TA measure uses ratio of basic triangle area to whole triangle area as reinforced factor for Euclidean distance so that it can alleviate negative effect of Euclidean distance whereas it keeps simplicity and effectiveness of both cosine measure and Euclidean distance in making similarity of two vectors. TA is considered as an advanced cosine measure. TA is defined by equation 16.1 (Nguyen & Amer, 2019):

$$u_1 \cdot u_2 \geq 0: \mathrm{TA}(u_1, u_2) = \begin{cases} \dfrac{(u_1 \cdot u_2)^2}{|u_1|(|u_2|)^3} \text{ if } |u_1| \leq |u_2| \\ \dfrac{(u_1 \cdot u_2)^2}{(|u_1|)^3 |u_2|} \text{ if } |u_1| > |u_2| \end{cases}$$

$$u_1 \cdot u_2 < 0: \mathrm{TA}(u_1, u_2) = \begin{cases} \dfrac{u_1 \cdot u_2}{(|u_2|)^2} \text{ if } |u_1| \leq |u_2| \\ \dfrac{u_1 \cdot u_2}{(|u_1|)^2} \text{ if } |u_1| > |u_2| \end{cases} \quad (16.1)$$

Where, as usual:

$$u_1 \cdot u_2 = \sum_{j \in I_1 \cap I_2} r_{1j} r_{2j}$$

$$|u_1| = \sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j})^2}$$

$$|u_2| = \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j})^2}$$

Let r_m be median of rating values, TA measure is normalized as TAN measure as follows:

$$\mathrm{TAN}(u_1, u_2) = \mathrm{TA}(u_1, u_2)$$

$$u_1 \cdot u_2 = \sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)(r_{2j} - r_m)$$

$$|u_1| = \sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)^2} \quad (16.3)$$

$$|u_2| = \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - r_m)^2}$$

By combined with Jaccard measure, TAN measure becomes TANJ measure as follows:

$$\mathrm{TANJ}(u_1, u_2) = \mathrm{TAN}(u_1, u_2) * Jaccard(u_1, u_2) \quad (16.4)$$

As a convention, TA family includes TA, TAJ, TAN, and TAJ. Hence, equation 16.1 is the key of TA family.

14. SMD

Given two rating vectors $u_1 = (r_{11}, r_{12},\ldots, r_{1n})$ and $u_2 = (r_{21}, r_{22},\ldots, r_{2n})$ of user 1 and user 2, respectively, in which some r_{ij} can be missing (empty). In binary representation, r_{ij} is converted into 1 if it is non-missing (rated) and otherwise, r_{ij} is converted into 0 if it is missing (not rated). Let N_{12} be the number of common values "1" in both u_1 and u_2. Let N be the total number of all items under consideration; in this case, $N = n$. Let N_1 and N_2 be the numbers of values "1" of u_1 and u_2, respectively. Let F be the number of differences between u_1 and u_2; for example, the fact that $r_{11} = 0$ and $r_{21} = 1$ contributes one difference to F. Amer defined a so-called *SMD measure* in binary representation as follows:

$$\text{SMD} = \frac{\left(1 - \frac{F}{N}\right) + \left(\frac{2N_{12}}{N_1 + N_2}\right)}{2} \tag{14.1}$$

Let I_1 or I_2 be sets of indices of items that user 1 or user 2 rates, respectively. Amer also defined another so-called *HSMD measure* in numerical representation in which values r_{ij} are kept in numerical values as rating values, as follows:

$$\text{HSMD} = 1 - \frac{R_1 R_2 + 1}{G} \tag{14.2}$$

Where, R_1 (R_2) is the sum of non-missing values r_{1j} (r_{2j}) of u_1 (u_2) such that respective values r_{2j} (r_{1j}) are missing.

$$R_1 = \sum_{\substack{r_{1j}\,\text{non-missing} \\ r_{2j}\,\text{missing}}} r_{1j} = \sum_{j \in I_1 \setminus I_2} r_{1j}$$

$$R_2 = \sum_{\substack{r_{2j}\,\text{non-missing} \\ r_{1j}\,\text{missing}}} r_{2j} = \sum_{j \in I_2 \setminus I_1} r_{2j}$$

Note, notation "\" denote complement operator in set theory. G is product of two sums of non-missing values for both r_1 and r_2.

$$G = \left(\sum_{r_{1j}\,\text{non-missing}} r_{1j}\right)\left(\sum_{r_{2j}\,\text{non-missing}} r_{2j}\right) = \left(\sum_{j \in I_1} r_{1j}\right)\left(\sum_{j \in I_2} r_{2j}\right)$$

In general, measures SMD and HSMD are defined firstly for weight vectors of documents in information retrieval, in which every element of a vector is a weight which is product of term frequency (TF) and inverse document frequency (IDF). Here they are applied into CF. For example, given two rating vectors $u_1 = (r_{11}=2, r_{12}=5, r_{13}=7, r_{14}=8, r_{15}=?, r_{16}=9)$ and $u_2 = (r_{21}=9,$

References

Al-Shamri, M. H. (2021, October 30). Similarity modifers for enhancing the recommender system performance. *Applied Intelligence*. doi:10.1007/s10489-021-02900-7

Ayub, M., Ghazanfar, A. M., Khan, T., & Saleem, A. (2020, May 12). An Efective Model for Jaccard Coefcient to Increase the Performance of Collaborative Filtering. (B. M. Ali, Ed.) *Arabian Journal for Science and Engineering, 45*(12), 9997 - 10017. doi:10.1007/s13369-020-04568-6

Ayub, M., Ghazanfar, M. A., Mehmood, Z., Saba, T., Alharbey, R., Munshi, A. M., & Alrige, M. A. (2019, August 1). Modeling user rating preference behavior to improve the performance of the collaborative filtering based recommender systems. *PLoS ONE, 14*(8), e0220129. doi:10.1371/journal.pone.0220129

Bag, S., Kumar, S. K., & Tiwari, M. K. (2019, January 11). An Efective Model for Jaccard Coefcient to Increase the Performance of Collaborative Filtering. (W. Pedrycz, Ed.) *Information Sciences, 483*(May 2019), 53 - 64. doi:10.1016/j.ins.2019.01.023

Bobadilla, J., Ortega, F., & Hernando, A. (2012, March). A collaborative filtering similarity measure based on singularities. *Information Processing and Management, 48*(2), 204-217. doi:10.1016/j.ipm.2011.03.007

Chen, L.-J., Zhang, Z.-K., Liu, J.-H., Gao, J., & Zhou, T. (2016, October 3). A vertex similarity index for better personalized recommendation. (H. W. Capel, Ed.) *Physica A, 466*(2017), 607 - 615. doi:10.1016/j.physa.2016.09.057

Choi, K., & Suh, Y. (2013, January). A new similarity function for selecting neighbors for each target item in collaborative filtering. (H. Fujita, & J. Lu, Eds.) *Knowledge-Based Systems, 37*(2013), 146-153. doi:10.1016/j.knosys.2012.07.019

Deng, J., Wang, Y., Guo, J., Deng, Y., Gao, J., & Park, Y. (2018, October 29). A similarity measure based on Kullback-Leibler divergence for collaborative filtering in sparse data. *Journal of Information Science, 45*(5), 656-675. doi:10.1177/0165551518808188

Feng, J., Fengs, X., Zhang, N., & Peng, J. (2018, September 24). An improved collaborative filtering method based on similarity. (H. Wang, Ed.) *PLos ONE, 13*(10), 1-18. doi:10.1371/journal.pone.0204003

Gazdar, A., & Hidri, L. (2019, September 25). A new similarity measure for collaborative filtering based recommender systems. (J. Lu, Ed.) *Knowledge-Based Systems, 188*(2020). doi:10.1016/j.knosys.2019.105058

Harris, E. F., & Sjøvold, T. (2018, September 7). Calculation of Smith's Mean Measure of Divergence for Intergroup Comparisons Using Nonmetric Data. (E. F. Harris, Ed.) *Dental Anthropology: A Publication of the Dental Anthropology Association, 17*(3), 83-93. doi:10.26575/daj.v17i3.152

Hyung, J. A. (2008, January 2). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. (W. Pedrycz, Ed.) *Information Sciences, 178*(1), 37-51. doi:10.1016/j.ins.2007.07.024

Jin, Q., Zhang, Y., Cai, W., & Zhang, Y. (2020, January 10). A New Similarity Computing Model of Collaborative Filtering. *IEEE Access, 8*, 17594 - 17604. doi:10.1109/ACCESS.2020.2965595

Jindal, A., Sharma, N., & Verma, V. (2022). Joyful Jaccard: An Analysis of Jaccard-Based Similarity Measures in Collaborative Recommendations. *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020. 1*. Springer Nature. doi:10.1007/978-981-16-8542-2_3

Lee, S. (2017). Improving Jaccard Index for Measuring Similarity in Collaborative Filtering. In K. Kim, & N. Joukov (Ed.), *Information Science and Applications 2017 (ICISA 2017). 424*, pp. 799 - 806. Macau: Springer. doi:10.1007/978-981-10-4154-9_93

Lee, S. (2018). Entropy-weighted similarity measures for collaborative recommender systems. *AIP Conference Proceedings. 1982*, pp. 1-6. AIP Publishing. doi:10.1063/1.5045417

Liang, S., Ma, L., & Yuan, F. (2015, 10). A singularity-based user similarity measure for recommender systems. *International Journal of Innovative Computing Information and Control, 11*(5), 1629-1638. doi:10.24507/ijicic.11.05.1629

Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2013, January 25). A Similarity Measure for Text Classification and Clustering. (J. Pei, Ed.) *IEEE Transactions on Knowledge and Data Engineering, 26*(7), 1575-1590. doi:10.1109/TKDE.2013.19

Liu, H., Hu, Z., Mian, A., Tian, H., & Zhu, X. (2013, November 20). A new user similarity model to improve the accuracy of collaborative filtering. (H. Fujita, & J. Lu, Eds.) *Knowledge-Based Systems, 56*(2014), 156-166. doi:10.1016/j.knosys.2013.11.006

Manochandar, S., & Punniyamoorthy, M. (2020, August 22). A new user similarity measure in a new prediction model for collaborative filtering. (M. Ali, & H. Fujita, Eds.) *Applied Intelligence, 51*(1), 586 - 615. doi:10.1007/s10489-020-01811-3

Moghadam, P. H., Heidari, V., Moeini, A., & Kamandi, A. (2019, September 6). An exponential similarity measure for collaborative filtering. *SN Applied Sciences, 1*(10), 1-4. doi:10.1007/s42452-019-1142-8

Mu, Y., Xiao, N., Tang, R., Luo, L., & Yin, X. (2019, January). An Efficient Similarity Measure for Collaborative Filtering. (R. Bie, Y. Sun, & J. Yu, Eds.) *Procedia Computer Science, 147*(2019), 416-421. doi:10.1016/j.procs.2019.01.258

Nguyen, L., & Amer, A. A. (2019, October 17). Advanced Cosine Measures for Collaborative Filtering. (ITS, Ed.) *Adaptation and Personalization (ADP), 1*(1), 21-41. doi:10.31058/j.adp.2019.11002

Patra, B. K., Launonen, R., Ollikainen, V., & Nandi, S. (2015, March 1). A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. (H. Fujita, & J. Lu, Eds.) *Knowledge-Based Systems, 82*, 163-177. doi:10.1016/j.knosys.2015.03.001

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295). Hong Kong: ACM. doi:10.1145/371920.372071

Singh, P. K., Sinha, S., & Choudhury, P. (2022, January 25). An improved item-based collaborative filtering using a modified Bhattacharyya coefficient and user–user similarity as weight. *Knowledge and Information Systems, 64*, 665-701. doi:10.1007/s10115-021-01651-8

Sun, S.-B., Zhang, Z.-H., Dong, X.-L., Zhang, H.-R., Li, T.-J., Zhang, L., & Min, F. (2017, August 17). Integrating Triangle and Jaccard similarities for recommendation. (Q. Zou, Ed.) *PLoS ONE, 12*(8), 1-16. doi:10.1371/journal.pone.0183570

Suryakant, & Mahara, T. (2016, January). A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment. (R. K. Venugopal, R. Buyya, & M. L. Patnaik, Eds.) *Procedia Computer Science, 89*(2016), 450-456. doi:10.1016/j.procs.2016.06.099

Tan, Z., & He, L. (2017, November 29). An Efficient Similarity Measure for User-Based Collaborative Filtering Recommender Systems Inspired by the Physical Resonance Principle. *IEEE Access, 5*, 27211 - 27228. doi:10.1109/ACCESS.2017.2778424

Torres Júnior, R. (2004). *Combining Collaborative and Content-based Filtering to Recommend Research Paper.* Universidade Federal do Rio Grande do Sul, Programa de Pós Graduação em Educação. Porto Alegre: Universidade Federal do Rio Grande do Sul. Retrieved from http://www.lume.ufrgs.br/bitstream/handle/10183/5887/000432990.pdf;sequence=1

Verma, V., & Aggarwal, R. K. (2019). A New Similarity Measure Based on Simple Matching Coefficient for Improving the Accuracy of Collaborative Recommendations. (M. He, & Y. V. Bodyanskiy, Eds.) *International Journal of Information Technology and Computer Science, 11*(6), 37-49. doi:10.5815/ijitcs.2019.06.05