Let sim(u_1, u_2) denote the similarity of u_1 and u_2. For instance, the cosine measure of u_1 and u_2 is defined as follows (Torres Júnior, 2004, p. 17):

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

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. Cosine measure will be mentioned more later. The larger the similarity is, the more the user 2 is near to active user 1. Hence, the similarity is used to determine the list of neighbors of active user. Suppose NN algorithm finds out k neighbors of u_1, let N be set of indices of k neighbors of u_1. Of course, we have $|N| = k$. A missing value r_{1j} of u_1 is computed (predicted) based on ratings of nearest neighbors and similarities according to step 2 of NN algorithm (Torres Júnior, 2004, p. 18).

$$r_{1j} = \bar{u}_1 + \frac{\sum_{i \in N} (r_{ij} - \bar{u}_i) \text{sim}(u_1, u_i)}{\sum_{i \in N} |\text{sim}(u_1, u_i)|}$$

Where $\bar{u}_1$ and $\bar{u}_i$ are mean values of u_1 and u_i, respectively. The equation above is called prediction formula or estimation formula.

$$\bar{u}_i = \frac{1}{|I_i|} \sum_{j \in I_i} r_{ij}$$

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

Where I_i is the set of indices of items that user i rated. The missing value r_{1j} of u_1 can be predicted more simply as follows:

$$r_{1j} = \frac{\sum_{i \in N} r_{ij} \text{sim}(u_1, u_i)}{\sum_{i \in N} |\text{sim}(u_1, u_i)|}$$

In general, similarity measure is the heart of NN algorithm because prediction formulas are based on similarity measures. Pearson correlation is another popular similarity measure besides cosine, which is defined as follows (Sarwar, Karypis, Konstan, & Riedl, 2001, p. 290):

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

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

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

The similarity will be significant if both users rated more common items. Weight Pearson correlation (WPC) measure and sigmoid Pearson correlation (SPC) measure concern how much common items are. WPC and SPC are formulated as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

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

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

Where *H* is a threshold and it is often set to be 50 (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158).

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):

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

Another version of Jaccard is (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

$$\mathrm{Jaccard2}(u_1, u_2) = \frac{|I_1 \cap I_2|}{|I_1||I_2|}$$

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):

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

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

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

MSD measure combines with Jaccard measure, which derives MSDJ measure as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

$$\mathrm{MSDJ}(u_1, u_2) = \mathrm{MSD}(u_1, u_2) * \mathrm{Jaccard}(u_1, u_2)$$

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

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

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.

There are some other researches related to apply similarity measures into CF. Ahn (Hyung, 2008) proposed a heuristic measure to solve cold-starting problem which relates to missing data in which there is not enough information to calculate similarities between rating vectors (Hyung, 2008, p. 39). The measure called PIP measure based on concept of "agreement" in rating. If both user 1 and user 2 like or dislike the same item, it is called that they have a rating "agreement" on such item. Let r_{1j} and r_{2j} be ratings of user 1 and user 2 on item j, respectively, the agreement (Hyung, 2008, p. 43) of them is defined as follows:

$$\text{agree}(r_{1j}, r_{2j}) = \begin{cases} \text{true if } (r_{1j} > r_m \text{ and } r_{2j} > r_m) \\ \text{true if } (r_{1j} < r_m \text{ and } r_{2j} < r_m) \\ \text{false otherwise} \end{cases}$$

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. PIP measure (Hyung, 2008, p. 42) is sum of products of triples Proximity, Impact, and Popularity.

$$\text{PIP}(u_1, u_2) = \sum_{j \in I_1 \cap I_2} \text{Proximity}(r_{1j}, r_{2j}) * \text{Impact}(r_{1j}, r_{2j}) * \text{Popularity}(r_{1j}, r_{2j})$$

Proximity (Hyung, 2008, p. 43) indicates similarity of two ratings, based on agreement and distance between them. The distance is increased twice as a penalty if such two ratings are not agreed.

$$\text{Proximity}(r_{1j}, r_{2j}) = \begin{cases} \left((2(r_{max} - r_{min}) + 1) - |r_{1j} - r_{2j}|\right)^2 \\ \qquad \text{if agree}(r_{1j}, r_{2j}) = \text{true} \\ \left((2(r_{max} - r_{min}) + 1) - 2|r_{1j} - r_{2j}|\right)^2 \\ \qquad \text{if agree}(r_{1j}, r_{2j}) = \text{false} \end{cases}$$

Where r_{min} and r_{max} are minimum rating value and maximum rating value, respectively. If two ratings are agreed, their impact (Hyung, 2008, p. 43) is proportional to difference between them and rating median. If two ratings are disagreed, their impact is inverse of such difference.

$$\text{Impact}(r_{1j}, r_{2j}) = \begin{cases} (|r_{1j} - r_m| + 1) * (|r_{2j} - r_m| + 1) \\ \qquad \text{if agree}(r_{1j}, r_{2j}) = \text{true} \\ \dfrac{1}{(|r_{1j} - r_m| + 1) * (|r_{2j} - r_m| + 1)} \\ \qquad \text{if agree}(r_{1j}, r_{2j}) = \text{false} \end{cases}$$

Popularity (Hyung, 2008, p. 43) indicates difference between ratings given by active users and the average rating.

$$\text{Popularity}(r_{1j}, r_{2j}) = \begin{cases} 1 + \left(\frac{r_{1j} + r_{2j}}{2} - \mu_j\right)^2 \\ \text{if } (r_{1j} > \mu_j \text{ and } r_{2j} > \mu_j) \\ 1 + \left(\frac{r_{1j} + r_{2j}}{2} - \mu_j\right)^2 \\ \text{if } (r_{1j} < \mu_j \text{ and } r_{2j} < \mu_j) \\ 1 \\ \text{otherwise} \end{cases}$$

Note, μ_j is average rating of item *j*, which is same mean of rating values of item *j*. Experimental results proved that cold-starting problem is solved well by PIP measure (Hyung, 2008, p. 47).

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:

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

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.

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):

$$\text{PSS}(u_1, u_2) = \sum_{j \in I} \text{Proximity}(r_{1j}, r_{2j}) * \text{Significance}(r_{1j}, r_{2j}) * \text{Singularity}(r_{1j}, r_{2j})$$

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).

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

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

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

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:

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

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.

$$\mu_1 = \frac{1}{|I_1|} \sum_{j \in I_1} r_{1j}$$

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

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

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

Liu et al. (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 161) proposed NHMS as triple product of PSS measure, URP measure, and Jaccard2 measure.

$$\text{NHMS}(u_1, u_2) = \text{PSS}(u_1, u_2) * \text{URP}(u_1, u_2) * \text{Jaccard2}(u_1, u_2)$$

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

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):

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

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 1.2, 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:

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

~~The item BC similarity is negative logarithm of item BC coefficient as follows:~~

$$\sout{\mathrm{bc}(i,j) = -\ln(\mathrm{bcc}(i,j))}$$

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:

$$\mathrm{BC}(u_1, u_2) = \sum_{i \in I_1} \sum_{j \in I_2} \mathrm{bc}(i,j)\mathrm{loc}(r_{1i}, r_{2j})$$

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

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

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:

$$\mathrm{BCF}(u_1, u_2) = \mathrm{Jaccard}(u_1, u_2) + \mathrm{BC}(u_1, u_2)$$

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,…, x_b) and Y = (y_1, y_2,…, 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 1.1 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):

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

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.

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

Experimental result proved that CjacMD model is effective similarity model.

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:

$$\text{Triangle}(u_1, u_2) = 1 - \frac{|\text{AB}|}{|\text{OA}| + |\text{OB}|} = 1 - \frac{|u_1 - u_2|}{|u_1| + |u_2|} = 1 - \frac{\sqrt{\sum_{j \in I}(r_{1j} - r_{2j})^2}}{\sqrt{\sum_{j \in I} {r_{1j}}^2} + \sqrt{\sum_{j \in I} {r_{2j}}^2}}$$

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

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

Experimental result proved that TMJ is effective measure.

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:

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

Experimental result proved that Feng model is effective similarity model.

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):

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

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 1.1 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:

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

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

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

Experimental result proved that Mu measure is effective similarity model.

This research also implements the Similarity Measure for Text Processing (SMTP) for testing. SMTP was developed by Lin, Jiang, and Lee (Lin, Jiang, & Lee, 2013), which as 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},…, r_{1n}) and u_2 = (r_{21}, r_{22},…, r_{2n}), the function F of u_1 and u_2 is defined as follows (Lin, Jiang, & Lee, 2013, p. 1577):

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

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

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

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

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:

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

In this research, we proposed combined measures for CF.