Volume 1, 2019
DOI: 10.31058/j.adp.2019.11002

Advanced Cosine Measures for Collaborative Filtering

Loc Nguyen[1*], Ali A. Amer[2]

[1] Loc Nguyen's Academic Network, Board of Advisors, Long Xuyen, Vietnam
[2] TAIZ University, Computer Science Department, Taiz, Yemen

Email Address

ng_phloc@yahoo.com (Loc Nguyen), aliaaa2004@yahoo.com (Ali A. Amer)
*Correspondence: Loc Nguyen

Received: 26 July 2019**; Accepted:** 30 August 2019**; Published:** 17 October 2019

Abstract:

Cosine similarity is an important measure to compare two vectors for many researches in data mining and information retrieval. In this research, cosine measure and its advanced variants for collaborating filtering (CF) are evaluated. 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 is proposed as an improved version of cosine measure. 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 and other advanced cosine measures are tested with other similarity measures. From experimental results, TA is not a preeminent measure but it is better than traditional cosine measures in most cases and it is also adequate to real-time application. Moreover, its formula is simple too.

Keywords:

Collaborating Filtering (CF), Cosine, Similarity Measure, nearest Neighbors (NN) Algorithm, Rating Matrix.

1. Introduction

Recommendation system is a system which recommends items to users among many existing items in database. Item is anything which users consider, such as product, book, and newspaper. There are two main approaches for recommendation such as content-based filtering (CBF) and collaborative filtering (CF). CF recommends an item to a user if her/his neighbors (other users like her/him) are interested in such item. One of popular algorithms in CF is nearest neighbors (NN) algorithm. The essence of NN algorithm [1, pp. 16-18] is to find out nearest neighbors of a regarded user (called active user) and then to recommend active user items that these neighbors may like. Let $\boldsymbol{U} = \{u_1, u_2, \ldots, u_m\}$ be the set of users and let $\boldsymbol{V} = \{v_1, v_2, \ldots, v_n\}$ be the set of items. User-based rating matrix is the matrix in which rows indicate users and columns indicate items and each cell is a rating which a user gave

to an item. In other words, each row in user-based rating matrix is a rating vector of a specified user. Rating vector of active user is called active user vector. As a convention, rating matrix implies user-based rating matrix if there is no additional explanation. Table 1 is an example of user-based rating matrix in which missing values are denoted by question masks [2, p. 218]. In Table 1, active vector is u_4 = (r_{41}=1, r_{42}=2, r_{43}=?, r_{44}=?), which is shaded.

Table 1. *User-based rating matrix.*

	Item 1	Item 2	Item 3	Item 4
User 1	r_{11} = 1	r_{12} = 2	r_{13} = 1	r_{14} = 5
User 2	r_{21} = 2	r_{22} = 1	r_{23} = 2	r_{24} = 4
User 3	r_{31} = 4	r_{32} = 1	r_{33} = 5	r_{34} = 5
User 4	r_{41} = 1	r_{42} = 2	r_{43} = ?	r_{44} = ?

User-based rating matrix can be transposed into item-based rating matrix in which each row is a rating vector of a specified item. Table 2 is the item-based rating matrix which is transposed from the user-based rating matrix shown in Table 1.

Table 2. *Item-based rating matrix.*

	User 1	User 2	User 3	User 4
Item 1	r_{11} = 1	r_{21} = 2	r_{31} = 4	r_{41} = 1
Item 2	r_{12} = 2	r_{22} = 1	r_{32} = 1	r_{42} = 2
Item 3	r_{13} = 1	r_{23} = 2	r_{33} = 5	r_{43} = ?
Item 4	r_{14} = 5	r_{24} = 4	r_{34} = 5	r_{44} = ?

In Table 2, active item vectors are v_3 = (r_{13}=1, r_{23}=2, r_{33}=5, r_{43}=?) and v_4 = (r_{14}=5, r_{24}=4, r_{34}=5, r_{44}=?), which are shaded.

In Table 1, there are four rating vectors u_1 = (1, 2, 1, 5), u_2 = (2, 1, 2, 4), u_3 = (4, 1, 5, 5), and u_4 = (1, 2, r_{43}=?, r_{44}=?). Suppose the active rating vector is u_4, NN algorithm will find out nearest neighbors of u_4 and then compute the predictive values for r_{43} and r_{44} based on similarities between these neighbors and u_4. The NN algorithm which acts on user-based rating matrix is called user-based NN algorithm and the NN algorithm which acts on item-based rating matrix is called item-based NN algorithm. Although ideology of user-based NN algorithm and item-based NN algorithm is the same, their implementations are slightly different. User-based NN algorithm is mentioned by default. In general, NN algorithm includes two steps [1, pp. 17-18]:

1. Find out nearest neighbors of the active user by calculating similarities between active vector and other vectors. The more the similarity is, the nearer two users are. Given a threshold, users whose similarities between them and active user are equal to or larger than a threshold are considered as nearest neighbors of active user.

2. Compute predictive values for missing ratings of active vector. The computation is based on ratings of nearest neighbors and similarities calculated in step 1.

The essence of NN algorithm is to use similarity measures in order to find out nearest neighbors of an active rating vector. This research focuses on similarity measures for CF. The most popular similarity measures are cosine and Pearson. Given two rating vectors u_1 = (r_{11}, r_{12},…, r_{1n}) and u_2 = (r_{21}, r_{22},…, r_{2n}) of user 1 and user 2, in which user 1 is considered as active user and some r_{ij} can be missing (empty). Let I_1 and I_2 be set of indices of items that user 1 and user 2 rated, respectively. Let $I = I_1 \cap I_2$ denote intersection set of I_1 and I_2 and let $I_1 \cup I_2$ denotes union set of I_1 and I_2. All items whose indices belong to $I_1 \cap I_2$ are rated by both user 1 and user 2. In other words, all items whose indices belong to $I_1 \cap I_2$ co-exist in vectors u_1 and u_2.

All items whose indices belong to $I_1 \cup I_2$ are rated by user 1 or user 2. Notation $|x|$ indicates absolute value of number, length of vector, length of geometric segment, or cardinality of set, which depends on context. Please pay attention to these denotations.

Let $\text{sim}(u_1, u_2)$ denote the similarity of u_1 and u_2. For instance, the cosine measure of u_1 and u_2 is defined as follows [1, 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 based on ratings of nearest neighbors and similarities according to step 2 of NN algorithm [1, 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 [3, 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 [4, 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 [4, p. 158]:

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

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

Where H is a threshold and it is often set to be 50 [4, 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 [4, p. 158]:

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

Another version of Jaccard is [4, 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 [4, 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 [4, p. 158]:

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

There are some other researches related to apply similarity measures into CF. Liu et al. [4, p. 156] proposed a new similarity measure called NHMS to improve

recommendation task in which only few ratings are available. Their NHMS measure [4, p. 160] is based on sigmoid function and the improved PIP measure as PSS (Proximity – Significance – Singularity). PSS similarity is calculated as follows [4, 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 [4, 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. [4, 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. [4, 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. [4] 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. [5, 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. [5, 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 similarity for items is calculated as follows [5, 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, BC similarity concerns two items. According to Patra et al. [5, p. 5], user BC similarity is sum of products of item BC similarities 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. [5, p. 5] proposed Bhattacharyya coefficient in CF (BCF) as sum of user BC similarity and Jaccard measure as follows:

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

This research also implements the Similarity Measure for Text Processing (SMTP) for testing. SMTP was developed by Lin, Jiang, and Lee [6], 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 [6, 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 [6, 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 [6, p. 1577] defined SMTP measure based on function F as follows:

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

Over all measures, cosine is simplest, and its accuracy is also good, which is proved from our experiments later. In this research, we aim to not only improve its accuracy as much as possible but also keep its simplicity. The next section describes advanced cosine measures including our proposed triangle area (TA) measure.

2. Triangle Area (TA) Measure and Other Advanced Cosine Measures

Given two rating vectors $u_1 = (r_{11}, r_{12},\ldots)$ and $u_2 = (r_{21}, r_{22},\ldots)$ of user 1 and user 2, cosine similarity of u_1 and u_2 is cosine value of the angle α formed by u_1 and u_2. The larger cosine measure is, the more similar user 1 and user 2 are. Let $|u_1|$ and $|u_2|$ be 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. Recall that equation 1 defines the traditional cosine measure [1, p. 17]:

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

Where,

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

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

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

Note, I_1 and I_2 are two sets of items (item indices) on which user 1 and user 2 rate, respectively. So, $I_1 \cap I_2$ denotes intersection set of I_1 and I_2. It is easy to recognize that cosine measure does not concern items rated uniquely by user 1 or user 2. Conversely, Jaccard measure concerns all items, which is ratio of cardinality of intersection set $I_1 \cap I_2$ to cardinality of union set $I_1 \cup I_2$ given two item sets I_1 and I_2. Recall that Jaccard measure is calculated as follows:

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

Note, $I_1 \cup I_2$ denotes union set of items which are rated by user 1 or user 2. However, Jaccard does not concerns magnitude of rating values like cosine measure does. By following the ideology of Jaccard measure, cosine measure is modified according to equation 2.

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

As a convention, the modified cosine measure specified by equation 2 is called COJ measure. Obviously, the numerator of equation 1 is equal to or smaller than the denominator of equation 2 and hence, COJ measure is equal to or smaller than cosine measure. If the traditional cosine is calculated over all items and every item which is not rated by a user is received assumptively rating value 0 from such user then, the traditional cosine becomes COJ [4, p. 158].

Let r_m be median of rating values, for example, if rating values range from 1 to 5, the median is $r_m = (1+5) / 2 = 3$. The normalized cosine measure (CON) [4, p. 158] is defined by equation 3.

$$\mathrm{CON}(u_1, u_2) = \frac{\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)(r_{2j} - r_m)}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - r_m)^2}} \tag{3}$$

When the median r_m is replaced by mean of rating values which are rated on given item, the normalized cosine measure becomes adjusted cosine measure. CON measure is aforementioned CPC measure [4, p. 158].

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

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

Equation 4 specifies the adjusted cosine measure (COD).

$$\mathrm{COD}(u_1, u_2) = \frac{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{v}_j)(r_{2j} - \bar{v}_j)}{\sqrt{\sum_{j \in I_1 \cap I_2} (r_{1j} - \bar{v}_j)^2} \sqrt{\sum_{j \in I_1 \cap I_2} (r_{2j} - \bar{v}_j)^2}} \tag{4}$$

In general, COJ, CON, and COD specified by equations 2, 3, and 4 are advanced cosine similarities. Here the so-called triangle area (TA) measure is proposed, which is sketched by Figure 1. TA measure, which is another advanced cosine similarity, is based on triangle areas.

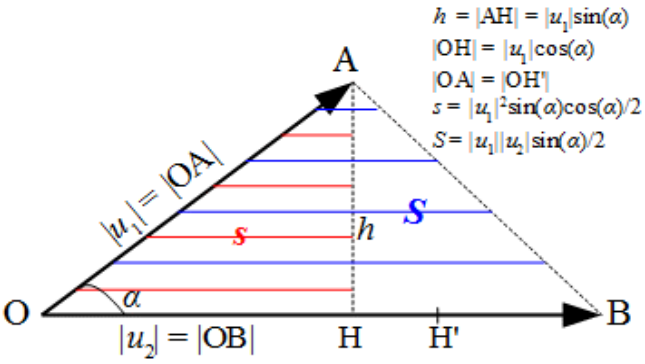

Figure 1. *Triangle area (TA) measure with $0 \leq \alpha \leq \pi/2$.*

In Figure 1, let u_1 = OA and u_2 = OB be two rating vectors and let α be the angle formed by u_1 and u_2. Of course, such two vectors form the triangle OAB. Cosine measure has a drawback that there may be two points like A and B which are far from each other according to Euclidean distance, but their cosine is high. This case is not totally wrong because cosine similarity really measures proportional correlation between two vectors. For example, user 1 and user 2 have rating vectors u_1 = (1, 1) and u_2 = (10, 10) on two items, respectively. Although distance between the two vectors are low, their cosine is highest, $\cos(u_1, u_2) = 1$. So, it is possible that user 1

and user 2 can give similar ratings on the third item. However, the Euclidean distance still affects negatively on cosine measure. In fact, the rating vector u_3 = (9, 9) is nearer to u_2 = (10, 10) than the u_1 = (1, 1) is although the cosine between u_3 = (9, 9) and u_2 = (10, 10) is also 1. Inspired from this observation, TA measure is really cosine similarity which alleviates the negative effect of Euclidean distance. On the other hand, TA measure also keeps positive effect of both cosine similarity and Euclidean distance.

The angle α shown in Figure 1 is equal to or less than $\pi/2$ and hence the first case $0 \leq \alpha \leq \pi/2$ is considered. Let OAB be the whole triangle and let S be its area. Given $|u_1| \leq |u_2|$ and $0 \leq \alpha \leq \pi/2$, we have:

$$S = \frac{1}{2}h|\mathrm{OB}| = \frac{1}{2}|u_1||u_2|\sin(\alpha)$$

Let OAH be the basic triangle and let s be its area. Given $|u_1| \leq |u_2|$ and $0 \leq \alpha \leq \pi/2$, we have:

$$s = \frac{1}{2}h|\mathrm{OH}| = \frac{1}{2}|u_1|\sin(\alpha)\,|u_1|\cos(\alpha) = \frac{1}{2}|u_1|^2\sin(\alpha)\cos(\alpha)$$

Given $|u_1| \leq |u_2|$ and $0 \leq \alpha \leq \pi/2$, reinforced factor k is defined as areal ratio of the basic triangle area s(OAH) to the whole triangle area S(OAB) as follows:

$$k = \frac{s}{S} = \frac{\frac{1}{2}|u_1|^2\sin(\alpha)\cos(\alpha)}{\frac{1}{2}|u_1||u_2|\sin(\alpha)} = \frac{|u_1|\cos(\alpha)}{|u_2|}$$

The greater the reinforced factor is, the more similar the two vectors u_1 and u_2 is. In general, in the first case $0 \leq \alpha \leq \pi/2$, reinforced factor k is determined by equation 5.

$$0 \leq \alpha \leq \frac{\pi}{2}: k = \begin{cases} \frac{|u_1|\cos(\alpha)}{|u_2|} \text{ if } |u_1| \leq |u_2| \\ \frac{|u_2|\cos(\alpha)}{|u_1|} \text{ if } |u_1| > |u_2| \end{cases} \quad (5)$$

There are many points on the two rays OA = u_1 and OB = u_2 so that their cosine is the same but only points A' and B' whose distance is shortest will obtain highest reinforced factor. In other words, if B' is the projection of A' on the ray OB or A' is the projection of B' on the ray OA then, reinforced factor will be maximal because within the same cosine value (the same angle), A' and B' are nearest. Therefore, equation 5 consolidates the viewpoint that within the same angle (the same cosine), reinforced factor is optimal if distance between two vectors (two end points) is shortest. This viewpoint is called *shortest distance viewpoint*.

However, there is another viewpoint that within the same angle, reinforced factor is optimal if two vectors have equal length. This viewpoint is called *equal vector-length viewpoint*. According to the second viewpoint, the basic triangle is OAH' where $|\mathrm{AH'}| = |\mathrm{OA}| = |u_1|$ and so given $|u_1| \leq |u_2|$ and $0 \leq \alpha \leq \pi/2$, the basic triangle area is:

$$s = \frac{1}{2}|\mathrm{OA}||\mathrm{OH'}|\sin(\alpha) = \frac{1}{2}|u_1|^2\sin(\alpha)$$

Reinforced factor k becomes:

$$k = \frac{s}{S} = \frac{\frac{1}{2}|u_1|^2\sin(\alpha)}{\frac{1}{2}|u_1||u_2|\sin(\alpha)} = \frac{|u_1|}{|u_2|}$$

Therefore, in the first case $0 \leq \alpha \leq \pi/2$, equation 6 specifies reinforced factor concerning the viewpoint of equal vector length.

$$0 \leq \alpha \leq \frac{\pi}{2}: k = \begin{cases} \frac{|u_1|}{|u_2|} \text{if } |u_1| \leq |u_2| \\ \frac{|u_2|}{|u_1|} \text{if } |u_1| > |u_2| \end{cases} \quad (6)$$

From experiments, equation 5 is better than equation 6 and so this research follows equation 5 to calculate reinforced factor in the first case $0 \leq \alpha \leq \pi/2$ with the shortest distance viewpoint.

Now the second case $\pi/2 < \alpha \leq \pi$ shown in Figure 2 is considered. In such case, the equal vector-length viewpoint is obeyed where reinforced factor is optimal if two vectors have equal length within the same angle.

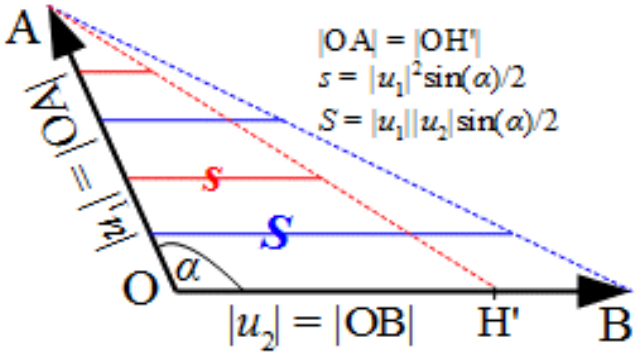

Figure 2. *Triangle area (TA) measure with $\pi/2 < \alpha \leq \pi$.*

Let OAB be the whole triangle and let S be its area. Given $|u_1| \leq |u_2|$ and $\pi/2 < \alpha \leq \pi$, we have:

$$S = \frac{1}{2}|u_1||u_2|\sin(\alpha)$$

Let OAH' be the basic triangle and let s be its area. Given $|u_1| \leq |u_2|$ and $\pi/2 < \alpha \leq \pi$, we have:

$$s = \frac{1}{2}|\mathrm{OA}||\mathrm{OH}'|\sin(\alpha) = \frac{1}{2}|u_1|^2\sin(\alpha)$$

Given $|u_1| \leq |u_2|$ and $\pi/2 < \alpha \leq \pi$, reinforced factor k is defined as areal ratio of the basic triangle area s(OAH') to the whole triangle area S(OAB) as follows:

$$k = \frac{s}{S} = \frac{\frac{1}{2}|u_1|^2\sin(\alpha)}{\frac{1}{2}|u_1||u_2|\sin(\alpha)} = \frac{|u_1|}{|u_2|}$$

Therefore, in the second case $\pi/2 < \alpha \leq \pi$, equation 7 specifies reinforced factor.

$$\frac{\pi}{2} < \alpha \leq \pi: k = \begin{cases} \frac{|u_1|}{|u_2|} \text{if } |u_1| \leq |u_2| \\ \frac{|u_2|}{|u_1|} \text{if } |u_1| > |u_2| \end{cases} \quad (7)$$

In general, if rating values are only non-negative, only equation 5 is used to calculate k. If rating values are only non-positive, only equation 6 is used to calculate k. If rating values are heterogeneous, both equations 6 and 7 are used to calculate k. Obviously, equations 6 and 7 are the same here. This research uses only equation 5 to calculate k because experimental dataset used for this research has only positive ratings. In practice, equation 5 is often used because rating values are often positive.

Over equations 5, 6, and 7, it is easy to recognize that reinforced factor is calculated by dot product and vector lengths. Changing the way to calculate dot product and vector length is to change reinforced factor.

The equal vector-length viewpoint is relative to the shortest distance viewpoint. Figure 3 is an extension of Figure 2 in which let A' and H'' be projections of A and H' on lines OB and OA, respectively.

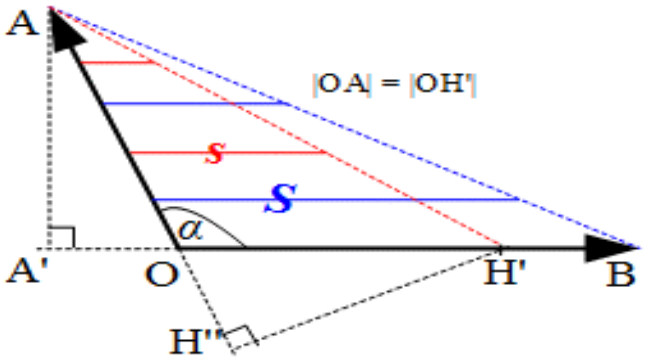

Figure 3. *Obtuse triangle with $\pi/2 < \alpha \leq \pi$.*

Of course, |AA'| is distance from A to the line OB and |H'H''| is distance from H' to the line OA with note that |OA| = |OH'|. We have |AA'| = |H'H''| due to:

$$|\mathrm{AA'}| = |\mathrm{OA}| \sin(\widehat{\mathrm{AOA'}}) = |\mathrm{OH'}| \sin(\widehat{\mathrm{AOA'}}) = |\mathrm{OH'}| \sin(\widehat{\mathrm{H'OH''}}) = |\mathrm{H'H''}|$$

Therefore, the condition of equal vector length only assures that two distances from every point to its dual line are equal (|AA'| = |H'H''|) but |AH'| is not shortest distance because it is obvious that |AA'| = |H'H''| < |AH'|. So, the condition of equal vector-length (equations 6 and 7) is weaker than the condition of shortest distance (equation 5).

TA measure, which is denoted TA(u_1, u_2), is defined as product of cosine value and reinforced factor, according to equation 8.

$$\mathrm{TA}(u_1, u_2) = k \cos(\alpha) \tag{8}$$

Where cos(α) is defined by equation 1 and reinforced factor is defined by equation 5, 6, or 7, which depends on context. Note that maximal value of TA measure is cosine value. Hence, TA measure is always equal to or less than cosine measure because reinforced factor is put in TA measure. According to the shortest distance viewpoint by equation 5, if B' is the projection of A' on the ray OB ($|u_1|\cos(\alpha) = |u_2|$) or A' is the projection of B' on the ray OA ($|u_2|\cos(\alpha) = |u_1|$) then, TA measure is equal to cosine measure because reinforced factor is 1. According to the equal vector-length viewpoint by equations 6 and 7, if $|u_1| = |u_2|$, TA measure is equal to cosine measure because reinforced factor is 1.

Let A'' the projection of A' on the ray OB and let B'' be the projection of B' on the ray OA. Focus of the shortest distance viewpoint is that equation 5 is really optimal in practice when A'' is near to B' ($|u_1|\cos(\alpha) \approx |u_2|$) or B'' is near to A' ($|u_2|\cos(\alpha) \approx |u_1|$) because if A ≡ H ($|u_1|\cos(\alpha) = |u_2|$) or B ≡ H ($|u_2|\cos(\alpha) = |u_1|$) then, there is other A' or B' so that distance |A'B'| is shortest and this continues. This recursive process of the shortest distance viewpoint makes equation 5 better than equations 6 and 7. Following is the explanation. Given Figure 4, suppose A_1 is an approximative projection of A on the ray OB so that A_1 is near to H ($|AA_1| \approx |AH| = |u_1|\cos(\alpha)$) according to the shortest distance viewpoint and we have |AH'| = $|u_1|$ according to the equal vector-length

viewpoint. When A_1 is the optimal point of equation 5 in practice and H' is the optimal point of equations 6 and 7 then, equation 5 is better than equations 6 and 7 because A_1 is between H and H' and $|OH| < |OA_1| < |OH'|$. The best A_1 occurs if and only if $H \equiv A_1 \equiv H'$ when $A \equiv B$ and the triangle OAB degrades into segment OA ($\cos(\alpha)=1$ and $|AB|=0$). Therefore, equation 5 balances length of vectors and distance between vectors (points).

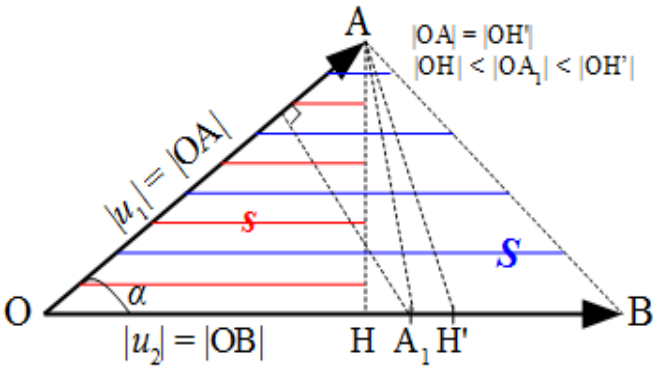

Figure 4. *Shortest distance viewpoint in practice.*

The essence of TA measure is cosine similarity. The computational cost for equation 5 is higher than traditional cosine. Because only equation 5 is used in this research when experimental dataset has only positive rating values, it is necessary to reduce equation 8 as much as possible. Recall that cosine of the angle α formed by u_1 and u_2 is calculated via dot product as $\cos(\alpha) = \frac{u_1 \cdot u_2}{|u_1||u_2|}$ and hence, the event $\cos(\alpha) \geq 0$ ($\cos(\alpha) < 0$) is equivalent to the event $u_1 \bullet u_2 \geq 0$ ($u_1 \bullet u_2 < 0$), which is also equivalent to $0 \leq \alpha \leq \pi/2$ ($\pi/2 < \alpha \leq \pi$). Suppose $u_1 \bullet u_2 \geq 0$ and $|u_1| \leq |u_2|$, according to equation 5, we have:

$$\mathrm{TA}(u_1, u_2) = \frac{|u_1|}{|u_2|}\cos^2(\alpha) = \frac{|u_1|}{|u_2|}\left(\frac{u_1 \cdot u_2}{|u_1||u_2|}\right)^2 = \frac{(u_1 \cdot u_2)^2}{|u_1|(|u_2|)^3}$$

Suppose $u_1 \bullet u_2 < 0$ and $|u_1| \leq |u_2|$, according to equations 6 and 7, we have:

$$\mathrm{TA}(u_1, u_2) = \frac{|u_1|}{|u_2|}\cos(\alpha) = \frac{|u_1|}{|u_2|}\frac{u_1 \cdot u_2}{|u_1||u_2|} = \frac{u_1 \cdot u_2}{(|u_2|)^2}$$

Equation 9 is reduced version of equation 8 for calculating TA measure.

$$\begin{aligned} u_1 \cdot u_2 \geq 0 &: \mathrm{TA}(u_1, u_2) = \begin{cases} \frac{(u_1 \cdot u_2)^2}{|u_1|(|u_2|)^3} \text{ if } |u_1| \leq |u_2| \\ \frac{(u_1 \cdot u_2)^2}{(|u_1|)^3|u_2|} \text{ if } |u_1| > |u_2| \end{cases} \\ u_1 \cdot u_2 < 0 &: \mathrm{TA}(u_1, u_2) = \begin{cases} \frac{u_1 \cdot u_2}{(|u_2|)^2} \text{ if } |u_1| \leq |u_2| \\ \frac{u_1 \cdot u_2}{(|u_1|)^2} \text{ if } |u_1| > |u_2| \end{cases} \end{aligned} \quad (9)$$

Where, as usual:

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

Where I_1 and I_2 are two sets of items on which user 1 and user 2 rate, respectively. Note that TA measure is calculated by dot product and vector lengths. Changing the way to calculate dot product and vector length is to change TA measure.

Because equations 8 and 9 are based on the intersection set $I_1 \cap I_2$ of user 1 and user 2, TA measure will ignore items which are rated uniquely by user 1 or user 2. Hence, TA measure can be improved by Jaccard measure. Let TAJ denote the combined measure which combines TA measure and Jaccard measure. Equation 10 specifies TAJ measure as follows:

$$\mathrm{TAJ}(u_1, u_2) = \mathrm{TA}(u_1, u_2) * Jaccard(u_1, u_2) \tag{10}$$

Where TA measure is specified by equation 8 or equation 9.

Let r_m be median of rating values, TA measure is normalized as TAN measure according to equation 11.

$$\begin{aligned} &\mathrm{TAN}(u_1, u_2) = \mathrm{TA}(u_1, u_2) \\ &u_1 \cdot u_2 = \textstyle\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)(r_{2j} - r_m) \\ &|u_1| = \sqrt{\textstyle\sum_{j \in I_1 \cap I_2} (r_{1j} - r_m)^2} \\ &|u_2| = \sqrt{\textstyle\sum_{j \in I_1 \cap I_2} (r_{2j} - r_m)^2} \end{aligned} \tag{11}$$

By combined with Jaccard measure, TAN measure becomes TANJ measure according to equation 12.

$$\mathrm{TANJ}(u_1, u_2) = \mathrm{TAN}(u_1, u_2) * Jaccard(u_1, u_2) \tag{12}$$

Equation 9 is the key in this research because it is simple. Other equations 10, 11, and 12 are based on equation 9 in this research.

For fair testing, TA measure and COJ measure are not combined. As a convention, TA family includes TA, TAJ, TAN, and TANJ with note that TAJ, TAN, and TANJ are improved versions of TA. Cosine family includes cosine, COJ, CON, and COD. Pearson family includes Pearson, WPC, and SPC. MSD family includes MSD and MSDJ. By default, all measures are calculated based on user-based rating matrix in which every vector is user rating vector. When user-based rating matrix is transposed into item-based rating matrix in which every vector is item rating vector, equations for these measures are not changed in semantics. In experimental section, these measures are tested with both user-based rating matrix and item-based rating matrix. NN algorithm for user-based rating matrix becomes user-based NN algorithm and NN algorithm for item-based rating matrix becomes item-based NN algorithm.

3. Experimental Results and Discussions

TA family (TA, TAJ, TAN, TANJ) is tested with traditional cosine family (cosine, COJ, CON, COD) and other well-known measures such as Pearson family (Pearson, WPC, SPC), Jaccard, MSD family (MSD, MSDJ), NHSM, BCF, and SMTP. Two metrics used to test measures are MAE and CC. MAE, which is abbreviation of mean absolute error, which is average absolute deviation between predictive ratings and users' true ratings. Given tested vector u_t = (1, 2, 3) having three items, it is made empty as empty vector u' = (?, ?, ?) with missing values. Later, NN algorithm is applied with measures above into predicting (estimating) missing values. As a result, predictive vector (estimated vector) is obtained, for example, u_p = (2, 3, 4) having three estimated items. Hence, MAE metric is (|2-1| + |3-2| + |4-3|) / 3 = 1. In general, MAE is calculated by equation 13 [7, pp. 20-21] in which n is the total number of estimated items while p_j and v_j are predictive rating and true rating of item j, respectively.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|p_j - v_j\right| \tag{13}$$

The smaller MAE is, the better the measures are. CC, which is abbreviation of correlation coefficient, is used to evaluate correlation between tested vector and predictive vector. It is really Pearson correlation. The larger CC is, the better the measures are. CC is calculated by equation 14 [7, p. 30] in which n is the total number of estimated items while p_j and v_j are predictive rating and true rating of item j, respectively.

$$\text{CC} = \frac{\sum_{j=1}^{n}(p_j-\bar{p})(v_j-\bar{v})}{\sqrt{\sum_{j=1}^{n}(p_j-\bar{p})^2}\sqrt{\sum_{j=1}^{n}(v_j-\bar{v})^2}} \tag{14}$$

Where $\bar{p}$ and $\bar{v}$ are mean values of tested item and predictive item, respectively.

$$\bar{p} = \frac{1}{n}\sum_{j=1}^{n} p_j$$

$$\bar{v} = \frac{1}{n}\sum_{j=1}^{n} v_j$$

MAE evaluates accuracy of measures whereas CC evaluates adequacy of measures. They are not opposite each other but biases between them can vary in tests. All measures are tested with both user-based NN algorithm (for user-based rating matrix) and item-based NN algorithm (for item-based rating matrix). Although the ideology of user-based NN algorithm and item-based NN algorithm is the same, their implementations are slightly different.

Dataset Movielens [8] is used for evaluation, which has 100,000 ratings from 943 users on 1682 movies (items). Every rating ranges from 1 to 5. In the experiments, dataset Movielens is divided into 5 folders and each folder includes training set and testing set. Training set and testing set in the same folder are disjoint sets. The ratio of testing set over the whole dataset depends on the testing parameter r. For instance, if r = 0.1, the testing set covers 10% the dataset, which means that the testing set has 10,000 = 10%*100,000 ratings and of course the training set has 90,000 ratings. In this testing, parameter r has three values 0.1, 0.5, and 0.9. The smaller r is, the more accurate measures are because training set gets large if r gets small with note that NN algorithm is executed on training set.

Table 3 shows all tested measures with r=0.1. The parameter value r=0.1 implies that testing set is here minimum. Each folder has own tested measures and so tested measures shown here are made average over 5 folders. Shaded cells indicate best values. Item-based MAE is selected as the main referred metric because it is smallest and item-based NN algorithm is better than user-based NN algorithm. Some other authors also confirm the preeminence of item-based NN algorithm. The reason is that one execution of user-based NN algorithm implies n executions of item-based NN algorithm in implementation where n is equal to the number of missing values (see Tables 1 and 2). In other words, one execution of user-based NN algorithm loop can estimates many missing values whereas one execution of item-based NN algorithm loop estimates only one missing value. As a result, item-based NN algorithm estimates more precisely than user-based NN algorithm does. Moreover, items which are goods in e-commerce are more stable than users as customers and Movielens dataset has more items than users.

Table 3. *Measures with r=0.1.*

r = 1	MAE (User-based)	MAE (Item-based)	CC (User-based)	CC (Item-based)
Cosine	0.7532	0.7427	0.4185	0.4217
COJ	0.7457	0.7317	0.3881	0.4004
CON	0.7469	0.7468	0.4339	0.4342
COD	0.8224	0.7894	0.1810	0.4403
Pearson	0.7395	0.7447	0.4355	0.4429
WPC	0.7312	0.7348	0.4519	0.4549
SPC	0.7388	0.7434	0.4378	0.4459
Jaccard	0.7465	0.7300	0.4273	0.4444
MSD	0.7529	0.7420	0.4192	0.4226
MSDJ	0.7457	0.7289	0.4288	0.4457
NHSM	0.7410	0.7219	0.4343	0.4579
BCF	0.7984	0.7822	0.3073	0.3978
SMTP	0.7533	0.7465	0.4185	0.4145
TA	0.7518	0.7399	0.4221	0.4249
TAJ	0.7449	0.7272	0.4311	0.4482
TAN	0.7467	0.7341	0.4338	0.4321
TANJ	0.7379	0.7164	0.4413	0.4656

From Table 3, TANJ is the best with lowest item-based MAE and highest item-based CC whereas WPC is the best with lowest user-based MAE and highest user-based CC. In general, top-5 measures according to item-based MAE are TANJ (1), NHSM (2), TAJ (3), MSDJ (4), and Jaccard (5). However, top-5 measures according to item-based CC are TANJ, NHSM, WPC, TAJ, and SPC. Top-5 measures according to user-based MAE and user-based CC are WPC, TANJ, SPC, Pearson, and NHSM. In TA family, TA measure does not combine with Jaccard measure and so, comparison of TA and cosine is necessary. In fact, from Table 3, TA is better than cosine with respect to both MAE and CC. Anyway, TANJ is always in top-5 measures and TA family is better than cosine family here.

Table 4 shows all tested measures with r=0.5. The parameter value r=0.5 implies that testing set is here medium.

Table 4. *Measures with r=0.5.*

r = 0.5	MAE (User-based)	MAE (Item-based)	CC (User-based)	CC (Item-based)
Cosine	0.7630	0.7541	0.3784	0.3776
COJ	0.7572	0.7450	0.3855	0.3948
CON	0.7657	0.7681	0.3525	0.3773

COD	0.8441	0.8099	0.0678	0.3672
Pearson	0.7734	0.7774	0.2905	0.3738
WPC	0.7581	0.7613	0.3265	0.3893
SPC	0.7708	0.7743	0.3013	0.3770
Jaccard	0.7583	0.7440	0.3803	0.3950
MSD	0.7627	0.7535	0.3793	0.3779
MSDJ	0.7575	0.7429	0.3822	0.3963
NHSM	0.7545	0.7376	0.3841	0.4040
BCF	0.8061	0.7934	0.2530	0.3457
SMTP	0.7629	0.7562	0.3789	0.3698
TA	0.7618	0.7517	0.3818	0.3800
TAJ	0.7568	0.7414	0.3843	0.3986
TAN	0.7654	0.7532	0.3712	0.3707
TANJ	0.7568	0.7380	0.3810	0.3968

From Table 4, NHSM is the best with lowest item-based MAE, highest item-based CC, and lowest user-based MAE, which is an incredible measure. COJ is the best with highest user-based CC. In general, top-5 measures according to item-based MAE are NHSM (1), TANJ (2), TAJ (3), MSDJ (4), and Jaccard (5). However, top-5 measures according to item-based CC are NHSM, TAJ, TANJ, MSDJ, and Jaccard. Top-5 measures according to user-based MAE are NHSM, TANJ, TAJ, cosine, and MSDJ. Top-5 measures according to user-based CC are COJ, TAJ, NHSM, MSDJ, and TA. From Table 4, pure TA is better than pure cosine with respect to both MAE and CC. Anyway, TA family is always in top-5 measures and it is better than cosine family here.

Table 5 shows all tested measures with r=0.9. The parameter value r=0.9 implies that testing set is here maximum.

Table 5. *Measures with r=0.9.*

r = 0.9	MAE (User-based)	MAE (Item-based)	CC (User-based)	CC (Item-based)
Cosine	0.8255	0.8177	0.3004	0.3154
COJ	0.8681	0.8597	0.2051	0.2620
CON	0.8985	0.8813	0.1403	0.2273
COD	0.9423	0.9098	0.0178	0.2243
Pearson	0.8473	0.8326	0.2530	0.2785
WPC	0.9202	0.9088	0.0569	0.2412
SPC	0.8490	0.8341	0.2476	0.2774
Jaccard	0.8651	0.8566	0.2058	0.2637
MSD	0.8471	0.8388	0.2335	0.2714
MSDJ	0.8538	0.8422	0.2248	0.2673
NHSM	0.8729	0.8583	0.2121	0.2547
BCF	0.8658	0.8568	0.1897	0.2815
SMTP	0.8478	0.8425	0.2328	0.2674
TA	0.8487	0.8399	0.2325	0.2702
TAJ	0.8552	0.8431	0.2243	0.2663
TAN	0.8723	0.8599	0.2148	0.2466
TANJ	0.8734	0.8593	0.2110	0.2469

Testing results from Table 5 are unexpected because the parameter value r=0.9 makes noise NN algorithm when the training set is not large enough. Cosine is the best with lowest MAE and highest CC. In general, top-5 measures according to item-based MAE are cosine (1), Pearson (2), SPC (3), MSD (4), and TA (5). However, top-5 measures according to item-based CC are cosine, BCF, Pearson, SPC, and MSD. Top-5 measures according to user-based MAE are cosine, MSD, Pearson, SMTP, and

TA. Top-5 measures according to user-based CC are cosine, Pearson, SPC, MSD, and SMTP. Measures like BCF and SMTP whose accuracy is low with *r*=0.1 and *r*=0.5 get now better. BCF aims to take advantages of statistical features and so it can resist lack of data. Similarly, Pearson, SPC, and MSD are now in top-5 measures because they also take advantages of statistical feature (correlation coefficient, sample variance). Although pure TA is in top-5 measures with user-based MAE, it is worse than pure cosine and hence, TA family is not better than cosine family here. The testing results also proved reliability of traditional measures like cosine and Pearson although they are not always preeminent in all cases.

Figure 5 shows tested measures with MAE metric and user-based rating matrix related to *r* = 0.1, 0.5, 0.9.

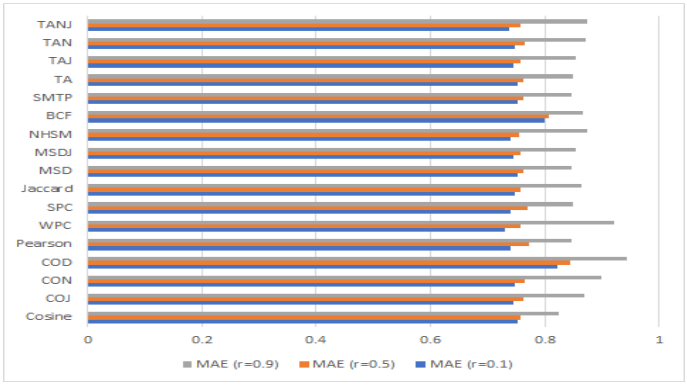

Figure 5. *Measures with user-based MAE related to r = 0.1, 0.5, 0.9.*

Figure 6 shows tested measures with MAE metric and item-based rating matrix related to *r* = 0.1, 0.5, 0.9.

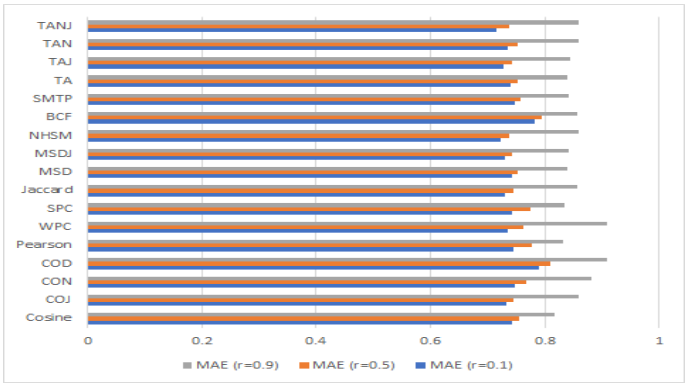

Figure 6. *Measures with item-based MAE related to r = 0.1, 0.5, 0.9.*

Figures 5 and 6 show the correlation of measures within MAE when parameter *r* is changed. The accuracy of measures decreases unexpectedly when *r* approaches 0.9. This is reasonable because the parameter value *r*=0.9 implies the training set is too small to train NN algorithm. Conversely, the parameter value *r*=0.1 is adequate to real-time application which has large rating database. The parameter value *r*=0.5 is adequate to testing application. Therefore, the lower the parameter value *r* is, the more adequate to real-time application the measures are.

Figure 7 shows tested measures with CC metric and user-based rating matrix related to *r* = 0.1, 0.5, 0.9.

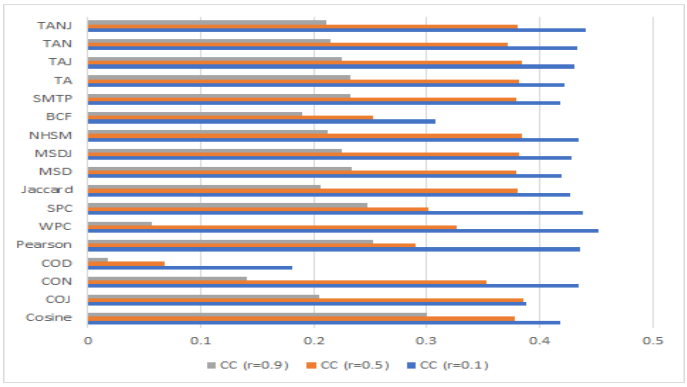

Figure 7. *Measures with user-based CC related to r = 0.1, 0.5, 0.9.*

Figure 8 shows tested measures with CC metric and item-based rating matrix related to r = 0.1, 0.5, 0.9.

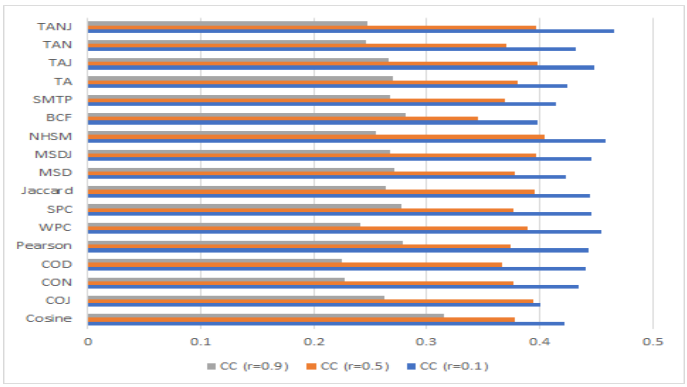

Figure 8. *Measures with item-based CC related to r = 0.1, 0.5, 0.9.*

Figure 7 and Figure 8 show the correlation of measures within CC when the parameter r is changed. The chart confirms the unexpected decrease of adequacy when r approaches 0.9. The reason was explained above.

With two parameter values r=0.1 and r=0.5, TA family is better than cosine family and NHSM is the best measure when MAE is selected as the main referred metric because item-based NN algorithm is always better than user-based NN algorithm. Three values of r such as 0.1, 0.5, and 0.9 are enough for us to survey all measures because these values are key values. For instance, the minimum value r=0.1 implies large real-time database, the medium value r=0.5 implies testing database, and the maximum value r=0.9 implies unexpectedly small database. However, it is not possible to draw which measures are the best in general yet. So, all measures are tested with all values of r: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and then average item-based MAE for each measure is calculated in order to determine best measures. Table 6 shows item-based MAE values of all measures over all values of r. Best values are shaded.

Table 6. *Item-based MAE of all measures over all values of r.*

	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	Average
TANJ	0.7164	0.7200	0.7230	0.7294	0.7380	0.7467	0.7614	0.7888	0.8593	0.7537
NHSM	0.7219	0.7251	0.7260	0.7309	0.7376	0.7436	0.7566	0.7843	0.8583	0.7538
TAJ	0.727	0.730	0.731	0.735	0.741	0.746	0.756	0.777	0.843	0.7544

	2	2	2	8	4	2	5	9	1	
MSDJ	0.7289	0.7320	0.7329	0.7375	0.7429	0.7474	0.7576	0.7782	0.8422	0.7555
Jaccard	0.7300	0.7330	0.7339	0.7385	0.7440	0.7485	0.7592	0.7815	0.8566	0.7584
COJ	0.7317	0.7346	0.7354	0.7398	0.7450	0.7495	0.7600	0.7831	0.8597	0.7599
Cosine	0.7427	0.7449	0.7456	0.7497	0.7541	0.7573	0.7653	0.7818	0.8177	0.7621
TA	0.7399	0.7421	0.7429	0.7471	0.7517	0.7551	0.7634	0.7809	0.8399	0.7626
MSD	0.7420	0.7443	0.7450	0.7491	0.7535	0.7566	0.7646	0.7811	0.8388	0.7639
TAN	0.7341	0.7365	0.7397	0.7454	0.7532	0.7601	0.7724	0.7959	0.8599	0.7664
SMTP	0.7465	0.7487	0.7488	0.7517	0.7562	0.7589	0.7667	0.7833	0.8425	0.7670
WPC	0.7348	0.7393	0.7438	0.7514	0.7613	0.7744	0.7933	0.8245	0.9088	0.7813
SPC	0.7434	0.7511	0.7569	0.7649	0.7743	0.7862	0.8022	0.8287	0.8341	0.7824
CON	0.7468	0.7505	0.7539	0.7604	0.7681	0.7772	0.7905	0.8188	0.8813	0.7831
Pearson	0.7447	0.7531	0.7591	0.7676	0.7774	0.7897	0.8057	0.8314	0.8326	0.7846
BCF	0.7822	0.7832	0.7845	0.7884	0.7934	0.7972	0.8034	0.8193	0.8568	0.8009
COD	0.7894	0.7937	0.7980	0.8031	0.8099	0.8183	0.8290	0.8494	0.9098	0.8223

The last column in Table 6 is average MAE values of measures. Hence, measures are sorted according to descending order of their average MAE values. Figure 9 shows comparable chart of all measures about their average MAE values. Note, the lower the column in Figure 9 is, the better the measure is.

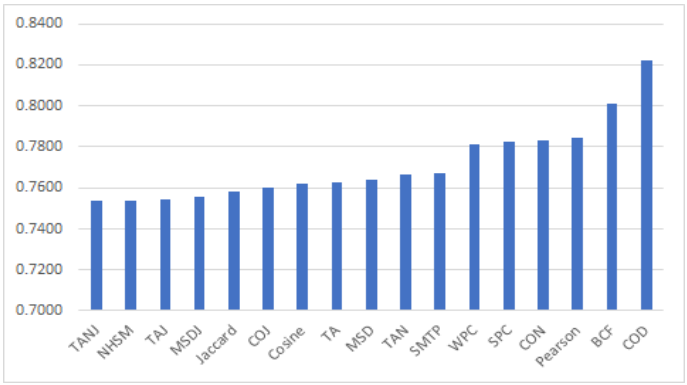

Figure 9. *Comparison of all measures regarding average MAE.*

From Table 6 and Figure 9, general top-5 measures are TANJ, NHSM, TAJ, MSDJ, and Jaccard whose average item-based MAE values are 0.7537, 0.7538, 0.7544, 0.7555, and 0.7584, respectively in which TANJ is the dominant measure. Because TA measure in TA family does not combine with Jaccard measure, comparison of pure TA and pure cosine is necessary. In fact, from Table 6, TA (average MAE = 0.7626) is worse than cosine (average MAE = 0.7621). However, it is necessary to survey carefully this situation. For most r = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8,

TA whose MAE values are 0.7399, 0.7421, 0.7429, 0.7471, 0.7517, 0.7551, 0.7634, and 0.7809, respectively is better than cosine whose MAE values are 0.7427, 0.7449, 0.7456, 0.7497, 0.7541, 0.7573, 0.7653, 0.7818, and 0.8177, respectively. With only r=0.9, TA (MAE=0.8399) is unexpectedly worse than cosine (MAE=0.8177). Figure 10 shows comparison of TA and cosine, in which MAE line of TA is always under MAE line of cosine from r=0.1 to r=0.8 but the MAE line of TA increases suddenly at r=0.9.

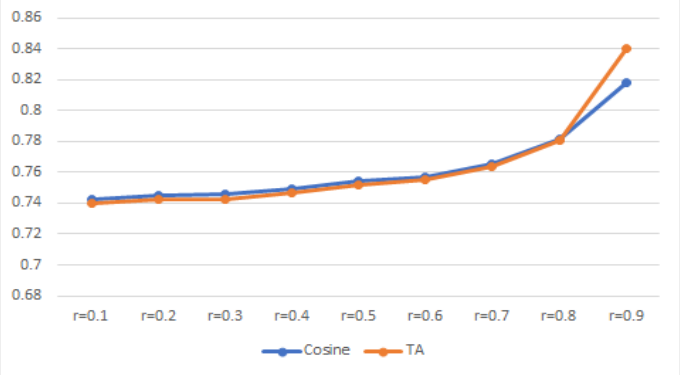

Figure 10. *Comparison of TA and cosine regarding MAE.*

This situation is easily explained because the parameter value r=0.9 implies the training set is too small to train NN algorithm and so, the strong point of TA which alleviates negative effect of Euclidean distance is broken by lack of information in small training set. On the other hand, pure cosine proved itself the stability over lack or disturbance of data. Anyway, TA is better and more suitable to real applications than cosine because values of r which are smaller 0.5 indicate large rating databases in real applications. Moreover, formula of TA is still simple.

4. Conclusions

There is no doubt that TA family is better than traditional cosine family in most cases (r = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) although pure TA measure is not preeminent like TANJ measure. Moreover, the general top-5 measures are TANJ, NHSM, TAJ, MSDJ, and Jaccard (see Table 8), in which variants of TA as TANJ and TAJ are in such top-5 list. There is an interesting observation that Jaccard is itself not a preeminent measure like TANJ or NHSM, but it is an important factor to improve any measure. In fact, good measures such as TANJ, NHSM, TAJ, MSDJ combine themselves with Jaccard. As another example, cosine is not dominant measure but COJ measure which follows the ideology of Jaccard measure gets better than cosine. It is surprising that Jaccard does not concern magnitude of rating values. The reason is that numerical measures except Jaccard are calculated with respect to rating values of common items on which both users rated and hence, these numerical measures ignore items which are rated uniquely by each user whereas Jaccard implies accuracy of these measures because Jaccard is the ratio of the number of common items to the number of all items. As a result, putting Jaccard into another measure is to adjust accuracy of such measure. In future trend, we try our best to improve TA so that it follows the ideology of Jaccard measure instead of combining TA and Jaccard as usual.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] R.D.; Torres Júnior. Combining Collaborative and Content-based Filtering to Recommend Research Paper. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

[2] M.P.T.; Do, D.V. Nguyen; L. Nguyen, "Model-based Approach for Collaborative Filtering," in Proceedings of The 6th International Conference on Information Technology for Education (IT@EDU2010), Ho Chi Minh, 2010.

[3] B. Sarwar; G. Karypis, J.; Konstan; J. Riedl. Item-based Collaborative Filtering Recommendation Algorithms. In Proceedings of the 10th international conference on World Wide Web, Hong Kong, 2001.

[4] H. Liu; Z. Hu; A. Mian; H. Tian; X. Zhu. A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems,* 2014, 56, 156-166.

[5] B.K. Patra; R. Launonen; V. Ollikainen; S. Nandi. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. *Knowledge-Based Systems,* 2015, 82, 163-177.

[6] Y.S. Lin, J.Y. Jiang; S.J. Lee. A Similarity Measure for Text Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering,* 2013, 26(7), 1575-1590.

[7] J.L. Herlocker; J.A. Konstan; L.G. Terveen; J.T. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems (TOIS),* 2004, 22(1), 5-53.

[8] GroupLens. MovieLens datasets. GroupLens Research Project, University of Minnesota, USA, 22 April 1998. Available online: http://grouplens.org/datasets/movielens (accessed on 3 August 2012).