Advanced Similarity Measures for Collaborative Filtering by Combination of Jaccard Measure and Other Numerical Measures

**Abstract**

Collaborative filtering (CF) is an important approach for recommendation system. One of popular algorithms in CF is nearest neighbors (NN) algorithm, in which similarity measures are used to determine nearest neighbors of a user. Jaccard is a similarity measure which concerns existence of ratings whereas other numeric measures like cosine and Pearson concern magnitude of ratings. Each of them has own strong points and drawbacks. Jaccard is itself not a dominant measure but it is an important factor to improve any measure. Therefore, the research focuses on combination of Jaccard and other numeric measures in order to derive advanced similarities which take advantages of both. Experimental results show that combined measures are preeminent ones.

**Keywords:** collaborating filtering (CF), similarity measure, Jaccard, nearest neighbors (NN) algorithm.

**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. NN algorithm (Torres Júnior, 2004, pp. 16-18) aims to find out nearest neighbors of a regarded user (called active user) and then to recommend active user items that these neighbors may like. Hence, 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*1*n*) and *u*2 = (*r*21, *r*22,…, *r*2*n*) of user 1 and user 2, in which user 1 is considered as active user and some *rij* 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 denote intersection set of *I*1 and *I*2 and let denotes union set of *I*1 and *I*2. Notation |x| indicates absolute value of number, length of vector, length of geometric segment, or cardinality of set, which depends on context.

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):

|  |  |
| --- | --- |
|  | (1) |

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

|  |  |
| --- | --- |
|  | (2) |

Where and are mean values of *u*1 and *u*2, respectively.

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):

|  |  |
| --- | --- |
|  | (3) |

Followings are equations of factors such as proximity, significance, and singularity based on sigmoid function (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 161).

Note, *rm* be median of rating values.

Jaccard measure is ratio of cardinality of common set to cardinality of union set . It measures how much common items both users rated, which is defined as follows (Liu, Hu, Mian, Tian, & Zhu, 2013, p. 158):

|  |  |
| --- | --- |
|  | (4) |

Liang, Ma, and Yuan (Liang, Ma, & Yuan, 2015) proposed an improved variant of Jaccard based on the concept of singularity. Their measure IJ is specified as follows (Liang, Ma, & Yuan, 2015, p. 1634):

|  |  |
| --- | --- |
|  | (5) |

Where *SjP*, *SjN*, and *SjE* are positive singularity, negative singularity, and empty singularity of item j, respectively whereas *PA*, *NA*, *D*, *PO*, and *NO* are sets of agreed positive ratings, agreed negative ratings, disagreed ratings, positive ratings, and negative ratings, respectively.

Jaccard can be combined with any measure. For example, CosineJ is combination of Jaccard and cosine as follows:

|  |  |
| --- | --- |
|  | (6) |

In general, cosine, Pearson, and PSS are numeric measures because they are calculated based on real rating values but they cannot solve the problem of missing values. At the other hand, Jaccard only focuses on existence of rating values but it ignores magnitude of ratings. The combination of Jaccard and a numeric measure takes advantages of both, which is described in next section.

**2. Methodology**

Jaccard improves especially accuracy of similarity of rating vectors in rating matrix because rating dataset has a lot of missing values whereas other measures depend on existence of ratings. Actually, Jaccard concerns both existence and inexistence of ratings when missing values in incomplete rating dataset imply inexistence of ratings. The more missing values the dataset has, the more accurate the Jaccard is. However, Jaccard does not concern magnitude of ratings. In other words, Jaccard does not concern real numbers. As a result, if rating dataset has enough rating values, Jaccard will be less accurate than other numeric measures. In general, Jaccard is itself not a dominant measure but it is an important factor to improve any measure. Therefore, I combine Jaccard measures and other numeric measures in order to taking advantages of both existence and quantity of rating values. The combination is mutual and not resonant and hence, I use multiplicative combination of Jaccard and other numeric measures such as cosine, Pearson, PSS, and TA. These numeric measures are typical with many variants. The advanced version of Jaccard which is developed by is also combined with cosine, Pearson, PSS, and TA in comparison with Jaccard. Such advanced variant of Jaccard is denoted IJ measure (Liang, Ma, & Yuan, 2015). Table 1 show the combined measures in this research.

|  |  |  |  |
| --- | --- | --- | --- |
| *Combined* | *Jaccard* | *Numeric* | *Formula* |
| CosineJ | Jaccard | Cosine | Jaccard \* cosine |
| PearsonJ | Jaccard | Pearson | Jaccard \* Pearson |
| PSSJ | Jaccard | PSS | Jaccard \* Pearson |
| TAJ | Jaccard | TA | Jaccard \* TA |
| CosineIJ | IJ | Cosine | IJ \* cosine |
| PearsonIJ | IJ | Pearson | IJ \* Pearson |
| PSSIJ | IJ | PSS | IJ \* Pearson |
| TAIJ | IJ | TA | IJ \* TA |

**Table 1.** Combination of Jaccard and other numeric measures

In this research, the 8 combined measures such as CosineJ, PearsonJ, PSSJ, TAJ, cosineIJ, PearsonIJ, PSSIJ, and TAIJ shown in table 1 are tested and compared with single and numeric measures cosine, Pearson, PSS, and TA.

It is necessary to describe TA measure here. 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 1 illustrates TA measure.

A close up of a map

Description automatically generated

**Figure 1.** Triangle area (TA) measure with 0 ≤ *α* ≤ *π*/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 as follows (Nguyen & Amer, 2019):

|  |  |
| --- | --- |
|  | (7) |

Where |*u*1| and |*u*2| are lengths of *u*1 and *u*2, respectively whereas *u*1•*u*2 is dot product (scalar product) of *u*1 and *u*2, respectively. The next section mentions experimental design and results.

**3. Results and discussions**

Dataset Movielens (GroupLens, 1998) 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 the experimental design, parameter *r* has nine values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 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. Popular metrics to assess CF algorithms are mean absolute error (MAE), recall, and precision. MAE indicates accuracy of measures. The smaller MAE is, the more accurate the measures are and so the better the algorithm is. Precision and recall are quality metrics that measure quality of recommended list – how much the recommendation list reflects user’s preferences. The larger quality metric is, the better the algorithm is. MAE, recall, and precision are mentioned in (Herlocker, Konstan, Terveen, & Riedl, 2004). Quality of a CF algorithm like NN algorithm depends on both estimation and recommendation. Estimation ability is ability to estimate or predict exactly missing values. Recommendation is ability to provide list of recommended items which is as suitable as possible to users. Here, different metrics (MAE, recall, precision) are used for different evaluation processes (estimation and recommendation). This independent evaluation allows use to test measures more objectively, in which estimation process focused on accuracy of CF algorithm and recommendation process focuses on quality of CF algorithm. In general, MAE is used for estimation whereas recall and precision are used for recommendation process.

The problem in recommendation is how to determine the number of recommended items called recommendation count which is the length of recommended vector. I propose a technique to calculate the recommendation count based on sparse-relevant ratio. Sparse-relevant ratio denoted *sr* is specified as follows:

*sr* = the-count-of-relevant-ratings / (|***U***| \* |***V***|)

Note, |***U***| is the number of users and |***V***| is the number of items. We calculate recommendation count dynamically according to both dataset and each rating vector *ui*. Let *C*(*ui*) be the recommendation count for user *i*, which means that NN algorithms will recommend at least *C*(*ui*) items to user *i*. The recommendation count *C*(*ui*) is specified as follows:

|  |  |
| --- | --- |
|  | (8) |

Where *T* is the number of items with note that every item included in *T* is rated by at least one user.

Table 2 shows MAE metric of all tested measures about all *r* = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 within estimation process. The last column shows average MAE metrics over all values of *r* and shaded cells indicate top-3 good values. As a convention, we define that preeminent measures (dominant measures) are ones in top-3 lists.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *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  (MAE) |
| Jaccard | 0.7465 | 0.7491 | 0.7502 | 0.7543 | 0.7583 | 0.7620 | 0.7717 | 0.7939 | 0.8651 | 0.7723 |
| IJ |  |  |  |  |  |  |  |  |  |  |
| Cosine | 0.7532 | 0.7551 | 0.7560 | 0.7593 | 0.7630 | 0.7654 | 0.7736 | 0.7905 | 0.8255 | 0.7713 |
| Pearson | 0.7395 | 0.7462 | 0.7519 | 0.7611 | 0.7734 | 0.7882 | 0.8091 | 0.8435 | 0.8473 | 0.7845 |
| PSS | 0.7452 | 0.7479 | 0.7490 | 0.7529 | 0.7568 | 0.7606 | 0.7708 | 0.7929 | 0.8591 | 0.7706 |
| TA | 0.7518 | 0.7538 | 0.7547 | 0.7581 | 0.7618 | 0.7643 | 0.7726 | 0.7901 | 0.8487 | 0.7729 |
| CosineJ | 0.7459 | 0.7485 | 0.7496 | 0.7537 | 0.7577 | 0.7615 | 0.7712 | 0.7921 | 0.8537 | 0.7704 |
| PearsonJ | 0.7311 | 0.7375 | 0.7427 | 0.7510 | 0.7624 | 0.7766 | 0.7992 | 0.8379 | 0.9173 | 0.7840 |
| PSSJ |  |  |  |  |  |  |  |  |  |  |
| TAJ | 0.7449 | 0.7475 | 0.7486 | 0.7527 | 0.7568 | 0.7606 | 0.7704 | 0.7920 | 0.8552 | 0.7699 |
| CosineIJ |  |  |  |  |  |  |  |  |  |  |
| PearsonIJ |  |  |  |  |  |  |  |  |  |  |
| PSSIJ |  |  |  |  |  |  |  |  |  |  |
| TAIJ |  |  |  |  |  |  |  |  |  |  |

**Table 2.** MAE metric within estimation process

Top-3 measures according to MAE metric within estimation process are TAJ, CosineJ, and PSSJ whose average MAE metrics are 0.7699, 0.7704, and 0.7706, respectively.

Table 3 shows precision metric of all tested measures about all *r* = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 within recommendation process given precision metric. The last column shows average precision metrics over all values of *r* and shaded cells indicate top-3 good values.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *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  (MAE) |
| Jaccard | 0.0056 | 0.0105 | 0.0155 | 0.0207 | 0.0261 | 0.0317 | 0.0377 | 0.0438 | 0.0511 | 0.0270 |
| IJ |  |  |  |  |  |  |  |  |  |  |
| Cosine | 0.0055 | 0.0104 | 0.0154 | 0.0207 | 0.0262 | 0.0324 | 0.0396 | 0.0508 | 0.0836 | 0.0316 |
| Pearson | 0.0051 | 0.0095 | 0.0141 | 0.0187 | 0.0237 | 0.0291 | 0.0359 | 0.0467 | 0.0803 | 0.0292 |
| PSS | 0.0057 | 0.0106 | 0.0157 | 0.0210 | 0.0266 | 0.0329 | 0.0401 | 0.0512 | 0.0834 | 0.0319 |
| TA | 0.0055 | 0.0104 | 0.0155 | 0.0207 | 0.0263 | 0.0325 | 0.0397 | 0.0509 | 0.0836 | 0.0317 |
| CosineJ | 0.0056 | 0.0105 | 0.0156 | 0.0209 | 0.0265 | 0.0327 | 0.0399 | 0.0510 | 0.0835 | 0.0318 |
| PearsonJ | 0.0052 | 0.0097 | 0.0143 | 0.0190 | 0.0240 | 0.0295 | 0.0362 | 0.0471 | 0.0804 | 0.0295 |
| PSSJ |  |  |  |  |  |  |  |  |  |  |
| TAJ | 0.0056 | 0.0106 | 0.0157 | 0.0209 | 0.0265 | 0.0328 | 0.0400 | 0.0511 | 0.0835 | 0.0319 |
| CosineIJ |  |  |  |  |  |  |  |  |  |  |
| PearsonIJ |  |  |  |  |  |  |  |  |  |  |
| PSSIJ |  |  |  |  |  |  |  |  |  |  |
| TAIJ |  |  |  |  |  |  |  |  |  |  |

**Table 3.** Precision metric within recommendation process

Top-3 measures according to precision metric within recommendation process are TAJ, PSSJ, and CosineJ, whose average precision metrics are 0.0319, 0.0319, and 0.0318, respectively.

Table 4 shows recall metric of all tested measures about all *r* = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 within recommendation process given recall metric. The last column shows average recall metrics over all values of *r* and shaded cells indicate top-3 good values.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *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  (MAE) |
| Jaccard | 0.9266 | 0.9230 | 0.9221 | 0.9191 | 0.9158 | 0.9155 | 0.9073 | 0.8947 | 0.8496 | 0.9082 |
| IJ |  |  |  |  |  |  |  |  |  |  |
| Cosine | 0.9241 | 0.9208 | 0.9211 | 0.9177 | 0.9150 | 0.9147 | 0.9066 | 0.8937 | 0.8021 | 0.9018 |
| Pearson | 0.9439 | 0.9402 | 0.9388 | 0.9359 | 0.9331 | 0.9309 | 0.9190 | 0.8948 | 0.7834 | 0.9133 |
| PSS | 0.9248 | 0.9219 | 0.9215 | 0.9179 | 0.9152 | 0.9143 | 0.9055 | 0.8903 | 0.7936 | 0.9006 |
| TA | 0.9242 | 0.9211 | 0.9211 | 0.9177 | 0.9149 | 0.9145 | 0.9060 | 0.8928 | 0.8005 | 0.9014 |
| CosineJ | 0.9266 | 0.9232 | 0.9223 | 0.9193 | 0.9159 | 0.9153 | 0.9066 | 0.8914 | 0.7970 | 0.9020 |
| PearsonJ | 0.9429 | 0.9440 | 0.9373 | 0.9351 | 0.9323 | 0.9309 | 0.9186 | 0.8948 | 0.7814 | 0.9130 |
| PSSJ |  |  |  |  |  |  |  |  |  |  |
| TAJ | 0.9265 | 0.9229 | 0.9224 | 0.9191 | 0.9154 | 0.9149 | 0.9060 | 0.8907 | 0.7957 | 0.9015 |
| CosineIJ |  |  |  |  |  |  |  |  |  |  |
| PearsonIJ |  |  |  |  |  |  |  |  |  |  |
| PSSIJ |  |  |  |  |  |  |  |  |  |  |
| TAIJ |  |  |  |  |  |  |  |  |  |  |

**Table 4.** Recall metric within recommendation process

Top-3 measures according to recall metric within recommendation process are Pearson, PearsonJ, Jaccard, whose average recall metrics are 0.9133, 0.9130, and 0.9082, respectively.

From metrics MAE, precision, and recall shown in tables 2, 3, 4, respectively, it is not easy to determine which measures are preeminent. F1 metric (Herlocker, Konstan, Terveen, & Riedl, 2004, p. 25) is the way to assembling precision and recall together specifies F1 metric. The larger F1 is, the better measures are.

|  |  |
| --- | --- |
|  | (9) |

Shortly, MAE is used to evaluate estimation process and F1 is used to evaluate recommendation process. Table 5 which is derived from tables 2, 3, and 4 shows average MAE values and F1 values of all measures. Shaded cells indicate good values.

|  |  |  |
| --- | --- | --- |
|  | MAE | F1 |
| Jaccard | 0.7723 | 0.026189 |
| IJ |  |  |
| Cosine | 0.7713 | 0.030551 |
| Pearson | 0.7845 | 0.028327 |
| PSS | 0.7706 | 0.030819 |
| TA | 0.7729 | 0.030602 |
| CosineJ | 0.7704 | 0.030717 |
| PearsonJ | 0.7840 | 0.028566 |
| PSSJ |  |  |
| TAJ | 0.7699 | 0.030768 |
| CosineIJ |  |  |
| PearsonIJ |  |  |
| PSSIJ |  |  |
| TAIJ |  |  |

**Table 5.** General MAE and F1 over all measures

Top-3 measures according to F1 metric within recommendation process are PSSJ, TAJ, and CosineJ whose average recall metrics are 0.030819, 0.030768, and 0.030717, respectively.

In general, two top-3 sets of good measures are {TAJ, CosineJ, PSSJ}, and {PSSJ, TAJ, CosineJ}. The preeminent measures are determined as members of the intersection of such three sets which are PSSJ, TAJ, and CosineJ. It is useful to compare PSSJ, TAJ, and CosineJ but it is impossible to unify metrics MAE and precision / recall together. However, we can compare them by radar chart but some transformations are necessary. Let IMAE be inverse of normalized MAE. The larger the IMAE is, the better the measures are.

|  |  |
| --- | --- |
|  | (10) |

Note, *m* is the maximum rating value. Table 6 lists metrics IMAE, precision, and recall of preeminent measures PSSJ, TAJ, and CosineJ.

|  |  |  |  |
| --- | --- | --- | --- |
|  | IMAE | Precision | Recall |
| PSSJ |  |  |  |
| TAJ |  |  |  |
| CosineJ |  |  |  |

**Table 6.** Comparison of PSSJ, TAJ, CosineJ with IMAE, precision, and recall

Figure 1 shows radar chart of preeminent measures PSSJ, TAJ, and CosineJ regarding IMAE, precision, and recall.

**Figure 2.** Comparison of PSSJ, TAJ, CosineJ with IMAE, precision, and recall

As seen in figure 4, lines of PSSJ, TAJ, and CosineJ are nearly overlapped. Hence, it is not necessary to determine which one is the best over all metrics.

**4. Conclusions**

From experimental results, there is no doubt that combined measures are preeminent ones which takes advantages of data existence and data magnitude. A combined measure has two built-in measures such as Jaccard and a numeric measure. With the built-in Jaccard, missing values are not ignored and they also contribute to accuracy of recommendation process. With the built-in numeric measure, real number values reflex exactly user favorites. Given numeric measure A, suppose the combined measure of A and Jaccard is called A+. Of course, A+ is better than A but it is not asserted that A+ is better than another numeric measure like B, for example. In other words, whether A+ is better than B is dependent on measure A itself. It is possible that we develop a numeric measure as well as possible and then combine it with Jaccard or variants of Jaccard. However, in practice when A and B are not far different in accuracy, A+ is often better than B. Especially, rating data is always incomplete, in which Jaccard is proved with its good accuracy. Therefore, research on improvement of Jaccard is necessary as aforementioned that Jaccard is an important factor to enhance any numeric measures.

**References**

GroupLens. (1998, April 22). *MovieLens datasets*. (GroupLens Research Project, University of Minnesota, USA) Retrieved August 3, 2012, from GroupLens Research website: http://grouplens.org/datasets/movielens

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems (TOIS), 22*(1), 5-53. doi:10.1145/963770.963772

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

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

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

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

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