International Journal of Internet, Broadcasting and Communication Vol.5 No.2 5-10 (2013)
http://dx.doi.org/10.7236/IJIBC.2013.5.2.5

IJIBC 13-2-2

Entropy-based Similarity Measures for Memory-based Collaborative Filtering

Hyeong-Joon Kwon[1] and Haniph Latchman[2]

[1]*School of Information and Communication Engineering, Sungkyunkwan University, South Korea*
katsyuki@skku.edu

[2]*Department of Electrical and Computer Engineering, University of Florida*
latchman@list.ufl.edu

Abstract

We proposed a novel similarity measure using weighted difference entropy (WDE) to improve the performance of the CF system. The proposed similarity metric evaluates the entropy with a preference score difference between the common rated items of two users, and normalizes it based on the Gaussian, tanh and sigmoid function. We showed significant improvement of experimental results and environments. These experiments involved changing the number of nearest neighborhoods, and we presented experimental results for two data sets with different characteristics, and results for the quality of recommendation.

Keywords: *similarity, collaborative filtering, entropy*

1. Introduction

The CF predicts the preference score of a user for items he/she had not previously evaluated. To predict the preference score, CF systems use a user-item data set which is comprised of numerical preference scores with a fixed range. Data sets for algorithmic experiments have been provided by many research institutions, Including the MovieLens data set (GroupLens project team, Minnesota University), the Jester data set (Berkeley Laboratory for Automation Engineering) and EachMovie (HP/Compaq DEC research center)[2], [3], [4]. Each data set includes its own attributes in their separate forms. For example, the range of reference scores and the genres of each movie. Many researchers and laboratories focus on improving prediction accuracy with such data sets. According to a survey paper of CF systems, CF can be divided into two types [1]. They are memory-based and model-based CFs. The memory-based CF refers to an all user-items matrix of a data set for predicting preference scores. It evaluates the similarity between each user or item, generates nearest neighborhoods, and predicts preference scores with nearest neighborhoods. The evaluation of similarity is the most essential step, and the evaluated similarity is used as a weight for predicting preference scores and as a measure for generating nearest neighborhoods [5], [6], [7], [8].

In this paper, we propose novel similarity measures using difference score entropy. The proposed measure evaluates the entropy with the difference of preference scores of common rating items between two users, and normalizes it for use as a weight. Because the proposed similarity measure exploits the advantages of information entropy, evaluation of the proposed similarity measure is simple, and it is possible to change the weight according to score difference of common rating items. Furthermore, the proposed measure shows

Manuscript Received: May. 03, 2013 / Revised: Jul. 25, 2013 / Accepted: Aug. 9, 2013
Corresponding Author: katsyuki@skku.edu,
Tel: +82-31-290-7196, Fax: +82-31-290-7196
School of Information and Communication Engineering, Sungkyunkwan University, South Korea

better prediction accuracy than existing similarity measures. The scope of applications of the proposed similarity measure is vast. For example, it could be used for finding document or image similarities.

The remainder of the paper is organized as follows. We describe various similarity metrics and memory-based CF systems in Section 2. Then, we explain the proposed similarity measure using difference score-based information entropy in Section 3. In Section 4, we present a variety of experimental results for the proposed similarity measure, in terms of the mean absolute error (MAE), number of data sets and each similarity algorithm. Section 5 concludes the paper.

2. Similarity Measures Using Deference Entropy

The proposed similarity measure is based on scores of common rating items between two users, and is implemented in the same manner as the existing similarity metric. To explain the proposed similarity measure, assume the user-item matrix such as Table 1 in Section 2. The proposed similarity measure consists of three steps.

Step 1. It evaluates the difference of common rating items between two users. Assume a common rating between U1 = {rU1,I1, rU1,I2, rU1,I3, ···, rU1,In} and U2 = {rU2,I1, rU2,I2, rU2,I3, ···, rU2,In}. The difference score set D(U1,U2) between user U1 and U2 is as follows:

$$
\begin{aligned}
D(U_1,U_2) &= \{r_{U1,I1}-r_{U2,I1}, r_{U1,I2}-r_{U2,I2}, \cdots, r_{U1,In}-r_{U2,In}\} \\
&= \{d_1, d_2, d_3, \cdots, d_n\}
\end{aligned}
$$

Step 2. It evaluates the weighted difference information entropy WDE(U1,U2). When the result is zero, the two users are perfectly similar. On other hand, if the result is higher, the two users aren't similar. For the first time, the information entropy H(D) of D is given by Equation (8) [12]. The p(di) is probability density function (PDF) of the difference score between the two users.

$$
\begin{aligned}
H(D) &= \sum_{i=1}^{n} p(d_i)\log_2\left(\frac{1}{p(d_i)}\right) \\
&= -\sum_{i=1}^{n} p(d_i)\log_2 p(d_i)
\end{aligned}
\tag{8}
$$

When the information entropy H(D) of D is calculated, it can be weighted according to each score difference of common rating items. It can be generalized as follows:

$$
WDE(U_1,U_2) = -\sum_{i=1}^{n} p(d_i)\log_2 p(d_i)\times|\,d_i\,| \tag{9}
$$

It is further extended by a weighting parameter, such as di2, di3 … din or the square root of di. These flexible weighting methods can result in better results than the absolute weighting of Equation (10). This extensional possibility is a major advantage of the proposed similarity measure.

Step 3. It normalizes WDE(U1,U2) to [0 , 1]. The reason this process is needed is that WDE(U1,U2) ranges from 0 to infinity. To normalize, we can consider various functions, such as tanh of SVM, and the sigmoid, Fuzzy and Gaussian function. This is the other advantage of the proposed similarity measure. The performance of the CF system using WDE can be improved according to the change in the normalization method. Equation (11), (12) and (13) represents the Guassian, Sigmoid and tanh function, respectively, where x is the similarity.

$$
G(x) = \mathrm{e}^{-x^2/2\sigma^2} \tag{10}
$$

$$
P(x) = \frac{1}{1+e^{-x}} \tag{11}
$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{12}$$

It is a result for the user ID 1of the MovieLens data set. Each normalization method shows a similar rank but the similarity value is different. In Fig. 4, users similar to user ID 1 are clearly shown according to various weightings. Sensible users can be applied to generation of nearest neighborhoods.

We compared the Cosine similarity with the proposed similarity measure, which was weighted as a square and normalized as a Gaussian function, as shown in Fig. 4. The x-axis is the user ID and the y-axis is the similarity. We confirmed that the Cosine similarity is assembled. This result for the Cosine similarity shows that similarity assortment between two users is difficult, and it is ambiguous. The proposed similarity measure shows similar users more clearly than the Cosine similarity and Person Correlation Coefficient.

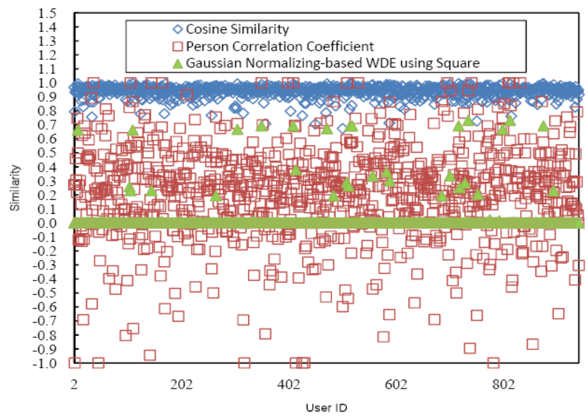

Figure 4. Similarity distribution WDE and existing measures

The algorithm of the proposed similarity measurement is as follows. Because the mechanism is simple, its application is easy. The proposed similarity measure can be applied to various fields. For example, it can be used in comparison between images of the same size, detecting similar signals, document similarity measurement, and evaluating the similarity between two variables. In Section 4, we perform verification via its application to the CF system.

```
BEGIN

    // algorithm name: weighted difference entropy
       // author: Hyeong-Joon Kwon(katsyuki@skku.edu)

    // length of co-rating items
    constant length;

    // preference array
    preference array x[length];
    preference array y[length];

    // deference set
    difference set D[length];

    // calculation entropy and weighting
    for(int i=1;i<sizeof(x);i++)
        D[i] = abs(x[i] – y[i]);

    for(int i=1;i<sizeof(D);i++)
        WDE += p(d[i])*log(p(d[i]),2)*d[i];

   // return inverse value as result
   return normalization(-WDE);

END
```

3. Experimental Results

We considered the MovieLens data set (GropuLens project) to verify the proposed similarity measure [2]. This data set contains 100,000 ratings for 1,682 movies by 943 users. The rating scale for this data set is from 1 to 5, {1, 2, 3, 4, 5}. The proposed similarity measure was normalized as a Gaussian function for this experiment. The reason is that the inverse function is generally used as a weight in the tanh and sigmoid function. The Gaussian function doesn't need an inverse. We used the kNN method for generation of nearest neighborhoods. To evaluate the performance of the proposed similarity measure, we experimented with various numbers of neighborhoods. Then, we used the Weighted Mean to predict the preference score, as explained in Section 2. This is the most generic method for predicting the score.

In this experiment, we randomly extracted 20% of user ratings from the MovieLens data set, and we predicted extracted ratings. And then, we measured the change of MAE according to the number of nearest neighborhoods. MAE decreases until the number of nearest neighborhoods reaches the arbitrary point where it starts to increase. The Pearson correlation is generally known as one of the most useful metrics in CF, and the performance of the Spearman correlation is also known to be comparable [5]. In Fig. 6, WDE using absolute weighting shows that the improvement of the prediction accuracy was remarkable for the proposed similarity metric, compared with existing similarity metrics.

In addition, this was suggested by the difference entropy according to Equation (10) in Section 3. The results for the various weighting methods are shown in Fig. 7. We confirmed that changing the weighting reduces MAE, as expected. But the weighting using a square root increases MAE. Although square root weighting shows bad results, its performance differs according to the format of the data set. For example, the MovieLens data set consists of a range [1, 5]. But the Jester data set is [-10.00, 10.00] and the Book-Crossing data set is [1, 10]. The problem of a sparse data set is one of the major issues of CF systems. In case of a data set containing only a few user ratings, the recommendation quality and prediction accuracy is reduced. All CF systems suffer from the sparse data set problem. To evaluate the performance for a sparse data set, we conducted two types of experiments reflecting the differing numbers of ratings available to the recommenders, All But One and Given 5. In the first type, All But One, we extracted a single randomly selected rating for each user in the test set, and tried to predict the given value based on all the other ratings the user submitted. In the second type Given 5 we randomly selected five ratings from each test user as observed ratings, and then attempted to predict the remaining ratings. This is a similar experimental method to those used in many other studies.

Fig. 7 shows the results for the sparse data experiment. The x-axis is MAE, and the y-axis is the experimental conditions. In Fig. 6, WDE showed better results than existing metrics. The MAE in the CF system was significantly decreased by WDE using various weighting methods. As shown in the results of this experiment, the most important point is that the proposed method showed better performance and prediction accuracy compared to existing methods.

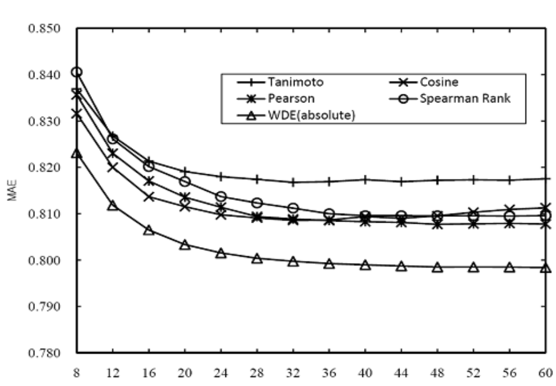

Figure 5. Experimental result for a given number for neighborhoods

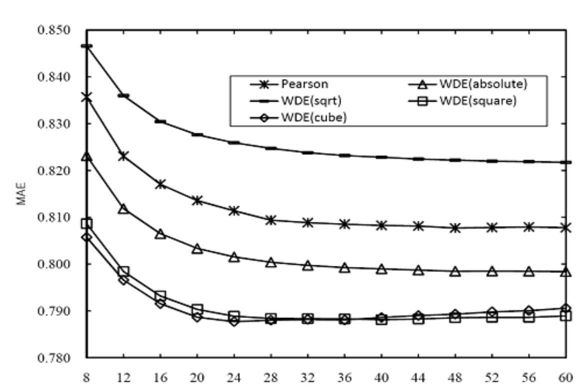

Figure 6. Weighting result for a given number for neighborhoods

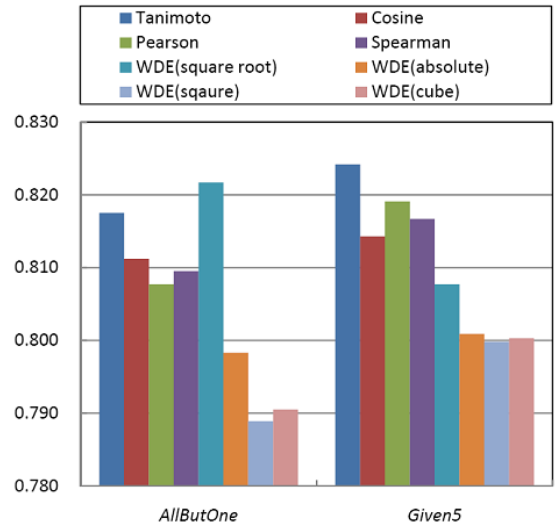

Figure 7. Experimental result for sparse data set

To evaluate the quality of recommendation, various characteristics have been considered, such as ROC-4, precision and recall. The quality of recommendation varies according to the generation method of the nearest neighborhood. It is affected by the similarity metric, but the effect is insignificant. The aim of this experiment was not improvement of the quality of recommendation; the proposed method is about the deterioration of the quality of recommendation. Although the proposed WDE improves the prediction accuracy, if it lowers the quality it will not be useful.

TABLE I. Experimental Result for Recommendation Quality

Similarity Measures		Precision	Recall	F-measure
Tanimoto		0.6764	0.2594	0.3749
Cosine		0.6817	0.2612	0.3776
Pearson		0.7039	0.2687	0.3889
Spearman Rank		0.6951	0.2766	0.3957
WDE	$\sqrt{d}$	0.6703	0.2744	0.3893
	$\lvert d \rvert$	0.7021	0.2797	0.4000
	d^2	*0.7102*	*0.2884*	*0.4102*
	d^3	0.7098	0.2810	0.4026

To improve the precision and recall, it must consider the generation of the nearest neighborhood in memory-based CF, because of factors affecting the recommendation quality. Based on Table 1, WDE also improves the quality of recommendation together with MAE. But this result is merely due to the reduction of MAE. The aim of this experiment was to investigate the deterioration of quality, rather than its improvement.

4. Conclusion

We proposed a novel similarity measure using weighted difference entropy (WDE) to improve the performance of the CF system. The proposed similarity metric evaluates the entropy with a preference score difference between the common rated items of two users, and normalizes it based on the Gaussian, tanh and sigmoid function. We showed significant improvement of experimental results and environments. These experiments involved changing the number of nearest neighborhoods, and we presented experimental results for two data sets with different characteristics, and results for the quality of recommendation.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 6, June 2005 , pp. 734-749.

[2] Bradley N. Miller, Istvan Albert, Shyong K. Lam, Joseph A. Konstan, "MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System", ACM 8th Int. Conf. on Intelligent user interfaces, 2003, pp. 263 – 266.

[3] Dhruv Gupta, Mark Digiovanni, Berkeley, Hiro Narita, Ken Goldberg, "Jester 2.0: Collaborative Filtering to Retrieve Jokes", ACM SIGIR 22nd Int. Conf. on Research in Information Retrieval, 1999, pp. 333.

[4] Steve Glassman, "EachMovie Collaborative Filtering Data Set", DEC Research Center, http://research.compaq.com/SRC/eachmovie/, 1997.

[5] Herlocker J.L., Konstan J.A., Borchers A. and Riedl, J., "An Algorithmic Framework for Performing Collaborative Filtering" ACM SIGIR 22nd Int. Conf. on Research and Development in Information Retrieval, 1999, pp.230-237.

[6] Toby S., "Programming Collective Intelligence: Building Smart Web 2.0 Applications", O'reilly, 2007.

[7] E. Garcia, "Cosine Similarity and Term Weight Tutorial", http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html#Cosim. 2006.

[8] Fu Lee Wang, "Improvements to Collaborative Filtering Algorithm", Lecture Note in Computer Science, Vol. 3314, 2004, pp. 975-981.

[9] Sarwar B., Karypis G., Konstan J. and Riedl J., "Item-Based Collaborative Filtering Recommendation Algorithms", ACM 10th Int. Conf. on World Wide Web, 2001, pp.285-295.

[10] Taek-Hun Kim and Sung-Bong Yang, "An Improved Recommendation Algorithm in Collaborative Filtering", Lecture Note in Computer Science, vol. 2455, 2005, pp. 254-261.

[11] D. Billsus and M.J. Pazzani, "Learning Collaborative Information Filters," Proc. 15th Int. Conf. Machine Learning, pp. 46-54, 1998.

[12] Shannon, Claude E.: Prediction and entropy of printed English, The Bell System Technical Journal, Vol. 30, 1951, pp.50-64.