IJACT 13-2-1

Improved Collaborative Filtering Using Entropy Weighting

Hyeong-Joon Kwon[1]

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

Abstract

In this paper, we evaluate performance of existing similarity measurement metric and propose a novel method using user's preferences information entropy to reduce MAE in memory-based collaborative recommender systems. The proposed method applies a similarity of individual inclination to traditional similarity measurement methods. We experiment on various similarity metrics under different conditions, which include an amount of data and significance weighting from n/10 to n/60, to verify the proposed method. As a result, we confirm the proposed method is robust and efficient from the viewpoint of a sparse data set, applying existing various similarity measurement methods and Significance Weighting.

Keywords: *Entropy Weighting, Collaborative Filtering, MAE*

1. Introduction

The recommendation method for recommender system is divided into three classes: content-based, collaborative filtering and hybrid recommendation [1][2][3]. Content-based methods recommend results similar to earlier user preferences. The collaborative method recommends items that people with similar tastes and preferences liked in the past. The hybrid method combines collaborative and content-based methods. The collaborative method is in the limelight, because it can predict the user's preference score. It includes two approaches: memory-based and model-based methods [4][5]. A common feature of the two approaches is to use similarity among users.

The similarity is calculated from the preferences score matrix. The final aim of collaborative filtering is to predict "?". In Collaborative Filtering, the similarity measurement method is able to define the most important element to improve performance of collaborative recommender system, because the similarity is used to predict preference score and choose the nearest neighborhood. The performance of a collaborative recommender system is contingent on a result of similarity measurement among all users. A similarity measurement metric includes the vector space model-based angle and distance, a ratio of intersection, correlation coefficient and others [4][5][6][7][8]. Each method exhibits different recommendation performance based on the total number of items and users. For this reason, many similarity measurement methods have been proposed in the past. Their performance differs according to the number of data sets and the Significance Weighting. The Significance Weighting is an additional method for similarity measurement and is only efficient for a correlation coefficient [9][10]. A novel method is in demand to improve performance of existing similarity measurement methods. Therefore, in this paper, we propose a novel method to apply user preference information entropy [11] to existing similarity measurements. This study aims to improve performance of collaborative recommendation with preference information entropy without

Manuscript Received: Sept. 08, 2013 / Revised: Dec. 1, 2013 / Accepted: Dec. 6, 2013
Corresponding Author: katsyuki@skku.edu
Tel:+82-31-290-7196, Fax: +82-31-290-7998
School of Information and Communication Engineering, Sungkyunkwan University, Korea

reference to a number of data sets. Each user in the preferences score matrix will include individual entropy. It will represent each user's inclination and their amount of information. We attempt to reduce mean absolute error with information entropy. We experiment on the measurement of mean absolute error with memory-based recommender systems. Proposed method was effective in all conditions, amounts of information, all similarity algorithms and significance weightings. That is, it is robust in every state. This approach may not be the only recommender system, but it is effective for other systems using similarity measures. Such examples include semantic similarity in ontology, a genome similarity, and similarity of music melody.

The remainder of the paper is organized as follows. We describe various similarity metrics and collaborative recommender systems in Section 2. Then, we explain a method to apply information entropy to similarity measurement for collaborative recommender systems in Section 3. In Section 4, we show diverse experimental results with the proposed method from the viewpoint of mean absolute error, number of data sets and each similarity algorithm. Section 5 concludes the paper.

2. Proposed Method Using User Preference Information Entropy

A user's information entropy in preference data includes various means that is the amount of information and impurity or disorder in the preference sequence. We can apply similarity of information entropy among users to the similarity measurement; it should be able to perform better than a collaborative recommendation using existing similarity measurement methods. Existing similarity measurement methods differ based on the data set form. Significance Weighting is only efficient in the Pearson Correlation Coefficient. We will show the experimental result of Significance Weighting in Section 4. That is, a novel method is demanded to improve performance of existing similarity measurement methods. We propose a method that applies Shannon's information entropy. The proposed method applies the similarity of information entropy among users to the result of the existing similarity measurement. When we consider users' similarity vector A ={x1, x2 … xn} and B = {y1, y2 … yn}, the proposed method SimEW(p,q) is as follows:

$$SimEW(A,B) = similarity(A,B) \times (|H(A) - H(B)| + 1)^{-1}$$

$$H(A) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \quad \text{and} \quad H(B) = -\sum_{i=1}^{n} P(y_i)\log_2 P(y_i)$$

In the equation for the proposed method, P is the probability mass function and H is Shannon's information entropy [11]. We derive the inverse of information entropy as a weight and add 1 to the information entropy for a negative number. If the user's preference sequence tends to a specified number, information entropy will be a negative number. If the information entropy of p and q are the same, the proposed method will not change the initial similarity. Table 1 shows the process changing similarity by the proposed method with virtual user A ,who includes entropy 1.292.

Table 1. Change of similarity and rank by the proposed method in a virtual data set

Before	User	Similarity	Entropy	Proposed	After
rank 1	*B*	0.951	2.128	0.517	rank 3
rank 2	*C*	0.942	0.948	0.700	rank 2
rank 3	*D*	0.872	1.245	0.832	rank 1

User B (existing rank 1) was changed to rank 3 and User D (existing rank 3) was changed to rank 1 by the proposed method using information entropy. For user B, it multiplied 0.951 by 0.544, as the inverse of |1.292-2.128|+1. It becomes 0.517, by the proposed method, as the final similarity. The similarity of user B decreases the most, down 0.434. Table 1 shows an extreme state. In reality, if the similarity between two users is similar, information entropy is usually similar. If information entropy between two users is similar, the similarity between two users is changed delicately. We prepared Table 2 to prove this assertion. Table 2 represents a fragment of the MovieLens data set. It consists of 100,000 data sets with 943 users and 1682 movies [23]. User preference ranges from 1 to 5. Let us consider Table 2, a fragment (user 931, entropy 1.829) of a real data set.

Table 2. Change of similarity and rank by the proposed method in a real data set

Before	User No.	Similarity	Entropy	Proposed	After
rank 1	845	0.777	2.062	0.674	rank 1
rank 2	821	0.763	1.545	0.593	rank 3
rank 3	928	0.738	0.868	0.375	rank 4
rank 4	890	0.708	2.006	0.601	rank 2

User number 928 decreases the most, down 0.363. However, the information entropy of user number 928 is one of the lowest values in the MovieLens data set. Lowered values in the calculation result of the 943 users included 0.497 for user number 688, 0.558 for user number 849 and 0.868 for user number 928. Conversely, 863 gained the most, 2.311. Preference data for user numbers 688 and 863are shown in Table 3. In Table 3, user number 688 excessively leans toward preference 5 and the data is sparse. User number 821 also leans toward preference 5. However, the preference is as not as excessive as for user number 688, and the amount of data is greater. Let us examine user number 863's data set. An amount of preference data is great and the preference score is divided evenly among all items. The inclination of the three users in Table 3 differs. Accordingly, information entropy differs from each other by my assertion.

Table 3. Information entropy for user number in the MovieLens data set

User No.	No. of 1	No. of 2	No. of 3	No. of 4	No. of 5	Entropy
688	0	1	0	1	22	0.497
621	0	2	10	15	35	1.545
863	18	18	21	26	19	2.311

As described above, the user's information entropy suggests a user's inclination and the amount of data and the change of similarity occur delicately in the majority of cases. However, some changes cause great changes in prediction score. We prove the proposed method effective in Section 4.

3. Experiments and Results

We accessed the MovieLens data set of GroupLens to experiment with the proposed method using information entropy. This data set consists of 943 users, 1682 movies and 100,000 preference values. Preference ranges from 1 to 5 at an interval of 1.0 [22]. We used the latest 10,000 counts as the input data in the 100,000 preference data observations and used 90,000 as the similarity measurement. The simulator for an experiment was developed using the C# language and Microsoft Visual Studio 2008 based on the console environment. It performs user-based collaborative recommendation and supports selection of the similarity measurement method, a use of Significance Weighting and a use of the proposed method. Similarity measurement methods for the experiment include Cosine Coefficient (CC), Euclidean Distance (ED), Person Correlation Coefficient (PCC), Tanimoto Coefficient (TC), Significance Weighting and the proposed method, similarity using entropy weighting. We used the Weighted Mean to predict preference score. The performance evaluation used Mean Absolute Error (MAE). We experiment varying the data set size. The first experimental result shows Figure 3 with 30,000 data.

Figure 3 shows the relative performance of the existing methods and proposed method for a sparse data set. The Significance Weighting does not improve performance except in the case of the Pearson Correlation Coefficient. However, a weight using information entropy improves performance in every similarity measurement method. The result improves MAE by 0.0028 in the case of the Cosine Coefficient in seven states (non ~ significance weighting 60), Tanomoto Coefficient is 0.0017, Euclidean Distance is 0.0012 and Pearson Correlation Coefficient is 0.0014. This result in the sparse data set shows its robustness from one viewpoint of collaborative recommendation. Although the improved value seems small, the results are an average error and proposed method improves MAE in every state; it was robust for the sparse data set.

Figure 4 demonstrates the excellent performance of the proposed method. In this result, with 60,000 data sets, the proposed method is as improved as for the former experiment (Figure 4), in every method and significance weighting. With data increases, such amelioration was observed in the Pearson Correlation Coefficient, but distinctively decreased in the Tanimoto Coefficient that does not consider user preference score. The results improve MAE by 0.0004 for the Cosine Coefficient, 0.0009 for the Tanimoto Coefficient, 0.0008 for the Euclidean Distance and 0.0006 for the Pearson Correlation Coefficient. Although the improvement in the results is slightly lower than in the earlier experiment, we were able to confirm the same merits as in the earlier experiment. The merits include robustness in a sparse data set and improvement in every state.

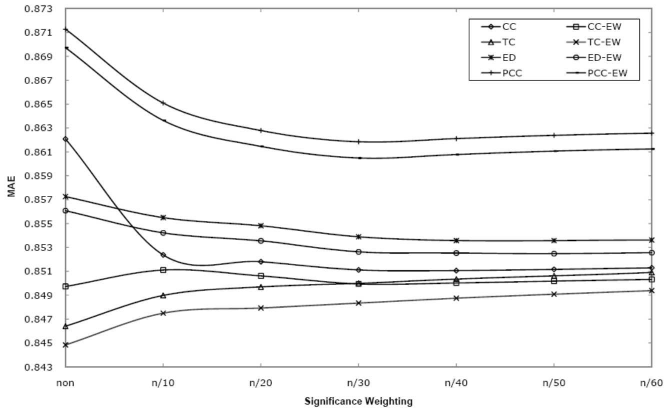

Figure 3. The experimental result with 30,000 training data

Taken altogether, the more the data count is increased, the more the performance difference among the similarity measurement methods comes into focus. The Pearson Correlation Coefficient holds a dominant position in Figure 5. Conversely, the Tanimoto Coefficient fell relatively. This experiment also shows that proposed method is very efficient. The result improves MAE by 0.0015 for the Cosine Coefficient, 0.0018 for the Tanimoto Coefficient, 0.0015 for Euclidean Distance and 0.0015 for Pearson Correlation Coefficient. We were convinced of the robustness of the proposed method concerning the data set size. When The Significance Weighting is applied, the proposed method displayed the performance and it excelled in a sparse data set. The proposed method using information entropy was able to extract regular merit and we proved it via experimentation.

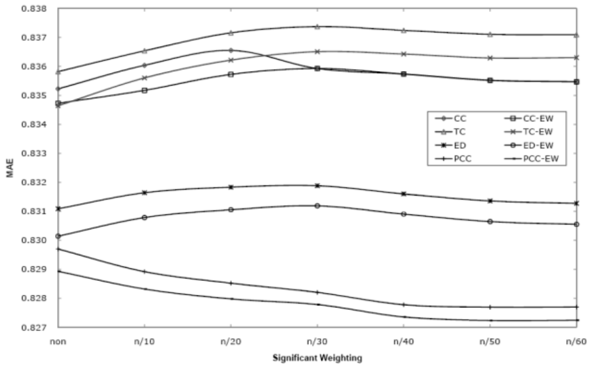

Figure 4. The experimental result with 60,000 training data

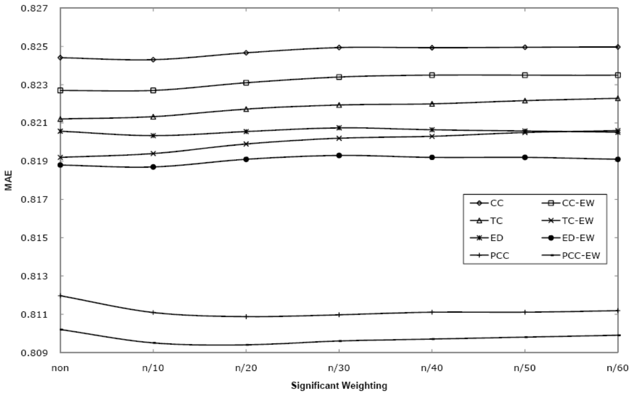

Figure 5. The experimental result with 90,000 training data

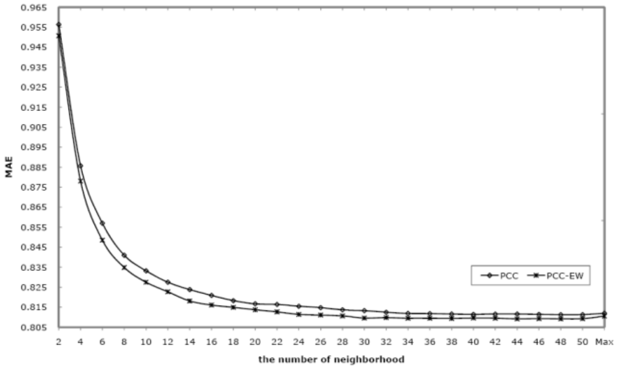

Figure 6. Experimental Result with a change of the number of neighborhood

4. Conclusion

In this paper, we survey existing similarity measurement methods and evaluate performance in collaborative recommender systems. We proposed a novel similarity weighting method using Shannon's information entropy that includes user inclination and the amount of information. The proposed method was robust from the viewpoint of a sparse data set, applying existing various similarity measurement methods and Significance Weightings. Information entropy can be useful to reduce prediction error in collaborative recommendation systems. If information entropy is applied to a formation of nearest neighborhood and predicting preference score, recommendation error will decrease.

References

[1] Sung-Ho Ha, "Digital Content Recommender on the Internet", IEEE Intelligent Systems 21-2 (2006) pp.70-77

[2] Baumann S., and Hummel O., "Enhancing Music Recommendation Algorithms Using Cultural Metadata", Journal of New Music Research 34-2 (2005), pp.161– 172

[3] Krulwich B. and Burkey C., "Learning user information interests through extraction of semantically significance phrases", Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access (2006)

[4] Hanhoon Kang and Seong Joon Yoo, "SVM and Collaborative Filtering-based Prediction of User Preference for Digital Fashion Recommendation Systems", IEICE Transactions on Information and Systems E90-D-12 (2007), pp.2100-2102

[5] J.S. Breese, D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", Proceedings of 14th Conference on Uncertainty in Artificial Intelligence (1998)

[6] Konstan J.A., Miller B.N, Maltz, D, Herlocker J.L., Gordon L.R. and Riedl J., “GroupLens: Applying Collaborative Filtering to Usenet News”, Communications of the Association for Computing Machinery 40-3 (1997), pp.77-87

[7] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, “Introduction to Data Mining”, Addison Wesley (2005)

[8] Toby S., Programming Collective Intelligence: Building Smart Web 2.0 Applications, O'REILLY (2007)

[9] Dengsheng Zhang and Guojun Lu, “Evaluation of similarity measurement for image retrieval”, Proceedings of the International Conference on Neural Networks and Signal Processing (2003), pp.14-17

[10] Good N., Schafer J.B., Konstan J.A., Borchers A., Sarwar B., Herlocker J. and Riedl J., “Combining Collaborative Filtering with Personal Agents for Better Recommendations”, Proceedings of American Association for Artificial Intelligence (1999)

[11] Shannon, Claude E., “Prediction and entropy of printed English”, The Bell System Technical Journal 30 (1950), pp.50-64

[12] Sarwar B., Karypis G., Konstan J. and Riedl J., “Item-Based Collaborative Filtering Recommendation Algorithms”, Proceedings of the 10th international conference on World Wide Web (2001), pp.285-295

[13] Linden G., Smith B. and York J., “Amazon.com recommendations: item-to-item collaborative filtering”, IEEE Internet Computing , (2003), pp.76-80

[14] Bevington, Philip R. and Robinson D. Keith., “Data Reduction and Error Analysis for the Physical Sciences”, McGraw-Hill (2002)

[15] Hua-Mei Chen, Pramod K. Varshney, Manoj K. Arora, “Performance of Mutual Information Similarity Measure for Registration of Multitemporal Remote Sensing Images”, IEEE Transaction on Geoscience and Remote Sensing 41-11 (2003), pp.2445- 2454

[16] Euclidean Distance, http://en.wikipedia.org/wiki/Euclidean_distance.

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

[18] E. Garcia, “C-Indices and Measures of Associations”, http://www.miislita.com/semantics/c-index-2.html (2008)

[19] Herlocker J.L., Konstan J.A., Borchers A. and Riedl, J., “An Algorithmic Framework for Performing Collaborative Filtering” Proceedings of the 22nd annual international SIGIR conference on Research and development in information retrieval (1999), pp.230-237

[20] Sarwar B., Karypis G., Konstan J. and Riedl, J., “Analysis of Recommendation Algorithms for E-Commerce”, Proceedings of the 2nd ACM conference on Electronic commerce (2000), pp.158-167

[21] Miller B.N., Albert I., Lam S.K., Konstan J.A. and Riedl J. “MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System”, Proceedings of the 8th international conference on Intelligent user interfaces (2003), pp.263-266

[22] GroupLens, “MovieLens Data Set”, http://grouplens.org/node/73 (2006)