ICIC Express Letters
Part B: Applications
Volume **8**, Number **2**, February **2017**

 ISSN 2185-2766
pp. **255–261**

A CONTEXT-ENTROPY BASED COLLABORATIVE FILTERING MODEL FOR PERSONALIZED RECOMMENDATION

Xiaoyi Deng[1,2,3] and Chun Jin[3]

[1]College of Business Administration
Huaqiao University
No. 269, Chenghua North Rd., Fengze Dist., Quanzhou 362021, P. R. China
londonbell@hqu.edu.cn

[2]Research Center for Applied Statistics and Big Data
Huaqiao University
No. 668, Jimei Ave., Jimei Dist., Xiamen 361021, P. R. China

[3]Institute of Systems Engineering
Dalian University of Technology
No. 2, Linggong Rd., Ganjingzi District, Dalian 116024, P. R. China
pltdeng@dlut.edu.cn

Received July 2016; accepted October 2016

Abstract. *Collaborative filtering (CF) is one of the most successful recommendation technologies to deal with information overload problem. Traditional neighborhood-based CF models solely use user/item similarities instead of existing user preferences to form neighborhoods and predict user ratings. Therefore, prediction accuracy excessively relies on the process of selecting the nearest neighbors. Besides, the customers' interests and demands may vary with contexts in different environment. As a result, the recommendations quality of conventional CF models would suffer. To address these issues, this paper developed a novel context-entropy based CF model. The context-entropy is introduced for measuring uncertainty degree of user rating under different contexts. And the context-entropy is incorporated into user similarity calculation for gathering the most similar users. Benchmark experiments on real-world datasets are carried out to compare our method's accuracy with other conventional CF algorithms. The results show that our method outperforms other methods and improves recommendation quality effectively.*
Keywords: Collaborative filtering, Context-entropy, Nearest neighbor selection, Personalized recommendation

1. **Introduction.** For decades, recommender systems (RSs) have become the most successful application of personalized recommendation for solving the information overload problem. RSs receive information from users about items that they are interested in, and then recommend to them items that may fit their needs [1]. The core of RSs usually relies on a well-known recommending algorithm, collaborative filtering (CF). CF can generate a recommendation according to the previous ratings of the neighbor users who have the same/similar interests with the active user, without relying on any information about the items themselves other than their ratings [2]. CF has an advantage in situations where it is hard to analyze the underlying content, such as music, videos and other digital products or services. Therefore, CF has been developed over decades and widely applied in many RSs and Internet-related fields [1], such as Amazon, Netflix, and Taobao.

Despite its advances, CF suffers from several problems, such as data sparsity and cold start problem. Data sparsity is common for the user-item ratings matrix to be extremely sparse. It makes traditional CF difficult to select the nearest neighbors for identifying similar users or items, and hard to produce accurate predictions or recommendations. To solve this problem, many different dimensionality reduction approaches have been

proposed, such as singular value decomposition [3], probabilistic matrix factorization [4], collaborative topic regression [5] and user access sequences [6]. However, useful information for recommendations related to those approaches may get lost and recommendation quality may be degraded, when certain users or items are discarded [7]. Some other researches have made use of information entropy to improve recommendation performance [8-10]. For instant, information entropy has been integrated with selective predictability to estimate the relationships between the target users and active users [8]. And, both the entropy of user and item is taken into account for the measurement of the relative difference between user ratings [9,10]. However, all the rating differences and user differences are usually treated individually without considering the correlation between them when the similarity is computed. Besides, the accuracy of predicting consumer preference depends on the degree to which the relevant contextual information is integrated into a recommendation model. The context has been recognized as an important factor for recommendation. However, most CF methods have not taken context into consideration [11]. These problems severely affected the quality of CF recommendation.

This paper attempts to develop an improved CF model, called CECF, which uses context and information entropy for the sake of improving prediction quality. In CECF, the entropy of user context is introduced to measure the uncertainty degree of the user rating behaviors under different contexts, and the uncertainty can be interpreted as how users understand rating domain to distinguish their tastes. Then, the context-entropy is incorporated into the process of nearest neighbor selection to enhance prediction quality. The rest of the paper is organized as follows. Section 2 introduces the concept of context-entropy and explains CECF model. Then, experimental results are demonstrated and discussed in Section 3. Finally, the paper is concluded and future research direction is given in Section 4.

2. **Context-Entropy Based CF Model.**

2.1. **Description of context.** The context has been identified as an essential factor in affecting users acceptance of RSs. The previous research [12] on service recommendation has suggested that the user profiles (such as age, and occupation) have a significant impact on service selection, as well as on consumer choice of purchase channel and perception that determine choice. Han et al. [13] found that context can be organized as a hierarchical directed acyclic graph with size information, which can be used to compute similarity between content and context for personalization. Mallat et al. [14] concluded that the context can be measured as a construction representing both user profiles and the conditions. The former describes personal features that may affect user's preference, and the latter represents conditions that users meet when they conduct commerce in different places and time. Therefore, context in this study is divided into two categories: user profile and environment information.

User profile is denoted by a triple C_u, which consists of user information including age, gender and occupation. Environment information is denoted by a triple C_s, which consists of three subsets: weather, time and holidays. C_u and C_s are shown in Equation (1) and Equation (2), respectively.

$$\begin{cases} C_u =< \text{Age, Gender, Occupation, Location} > \\ \text{Age} \in (A = \{A_i | i = 1, 2, \ldots, 7\}) \\ \text{Gender} \in (G = \{0, 1\}) \\ \text{Occupation} \in (O = \{O_i | i = 1, 2, \ldots, 20\}) \end{cases} \tag{1}$$

$$\begin{cases} C_s =< \text{Weather, Time, Holiday} > \\ \text{Weather} \in (W = \{W_i | i = 1, 2, \ldots, n\}) \\ \text{Time} \in (T = \{T_i | i = 1, 2, 3\}) \\ \text{Holiday} \in (H = \{0, 1\}) \end{cases} \tag{2}$$

where the set of *Age* is composed of 7 distinct sections: under 18, $18 \sim 24$, $25 \sim 34$, $35 \sim 44$, $45 \sim 49$, $50 \sim 55$ and older than 56; the set of *Gender* includes only two elements: male and female that are denoted by 1 and 0, respectively; the set of *Occupation* consists of more than 20 different occupations, such as teacher, doctor, engineer, and student. The *Weather* is denoted by set W, which contains n kinds of unique weathers; the *Time* is composed of three sections of the daytime: morning, afternoon and evening; the *Holiday* is similar to the *Gender* set which includes 0 and 1. If $H = 1$, it is a holiday; otherwise, it is a working day.

Assumed that $U = \{u_i | i = 1, 2, \ldots, m\}$ is a set of m users, for any user u_i, context can be denoted by set $C_i =< C_{ui}, C_{si} >= (A_i, G_i, O_i, W_i, T_i, H_i)$.

2.2. **Definition of context-entropy.** In Shannon's information theory, the entropy is defined as the expected value of the information contained in each event in a given message. In general, the more uncertain the event is, the more information it will contain. In other words, the entropy is a measure of unpredictability of information content. The entropy is denoted by H, defined as follows.

$$H(X) = -\sum_{i=1}^{n} (P(x_i) \cdot \log_2 P(x_i)),\ x_i \in X,\ \sum_{i=1}^{n} P(x_i) = 1 \tag{3}$$

where $P(x_i)$ is the probability of possible events x_i for message X.

In this paper, the context-entropy is defined as the uncertainty degree of user preferences under a certain context. In other words, the context-entropy is utilized to measure the uncertainty degree of user rating behaviors under different contexts. The context-entropy is denoted by $H_c(I)$, shown in the following equation.

$$\begin{aligned} H_c(u, I) &= -\sum_{i=1}^{n} (P_c(I_i) \cdot \log_2 P_c(I_i)),\ I_i \in I,\ \sum_{i=1}^{n} P_c(I_i) = 1 \\ P_c(I_i) &= \frac{\text{Number of rating on } I_i}{\text{Total number of rating on } I_i} \times \frac{\text{Types of context on } I_i}{\text{Total types of context}} \end{aligned} \tag{4}$$

where $P_c(I_i)$ represents the occurrence probability of each rating value of a user u on I_i under context C. In general, the smaller the context-entropy is, the more certain user preferences are; otherwise, the more uncertain user preferences are.

In order to make the comparison of users having different numbers of ratings easier, the value of context-entropy is normalized into interval $[0, 1]$. And, the value of context-entropy varies inversely as the uncertain degree of user preferences. Therefore, the normalized context-entropy can be calculated in Equation (5).

$$H_c^n(u, I) = 1 - \frac{H_c(u, I)}{\text{Total number of rating on } I}, \quad I_i \in I \tag{5}$$

2.3. **Context-entropy based nearest neighbor selection.** After the context-entropy value is obtained, the process of neighbor selection begins. To find the nearest neighbor of user u, the user similarity values between u and other users are computed by using user rating and context-entropy.

The similarity of user rating is measured by pearson correlation coefficient, as shown in Equation (6), where $r_{u,I}$ is the rating of item I by user u; $\overline{r_u}$ is the average rating of

user u, and $I(u_i, u_j)$ represents the items co-rated by users u_i and u_j.

$$Sim_{UR}^{PCC}(u_i, u_j) = \frac{\sum\limits_{i \in I(u_i,u_j)} (r_{u_i,I} - \overline{r_{u_i}}) \cdot \left(r_{u_j,I} - \overline{r_{u_j}}\right)}{\sqrt{\sum\limits_{i \in I(u_i,u_j)} (r_{u_i,I} - \overline{r_{u_i}})^2} \sqrt{\sum\limits_{i \in I(u_i,u_j)} \left(r_{u_j,I} - \overline{r_{u_j}}\right)^2}} \tag{6}$$

Then, the user similarity is calculated based on $Sim_{UR}^{PCC}(u_i, u_j)$, and the context-entropy is regarded as the rating weight of the users, as shown in Equation (7).

$$Sim(u_i, u_j) = \frac{\sum\limits_{i \in I(u_i,u_j)} (H_c^n(u, I) \cdot r_{u_i,I} - \overline{r_{u_i}}) \cdot \left(H_c^n(u, I) \cdot r_{u_j,I} - \overline{r_{u_j}}\right)}{\sqrt{\sum\limits_{i \in I(u_i,u_j)} (H_c^n(u, I) \cdot r_{u_i,I} - \overline{r_{u_i}})^2} \sqrt{\sum\limits_{i \in I(u_i,u_j)} H_c^n(u, I) \cdot \left(r_{u_j,I} - \overline{r_{u_j}}\right)^2}} \tag{7}$$

As the calculation of user similarity has been done, the rating prediction of itemset I by user u_i can be obtained by Equation (8).

$$PR(u_i, I) = \overline{r_{u_i}} + \frac{\sum\limits_{j \in I(u_i,u_j)} Sim(u_i, u_j) \cdot \left(r_{u_i,I} - \overline{r_{u_j}}\right)}{\sum\limits_{j \in I(u_i,u_j)} |Sim\,(u_i, u_j)\,|} \tag{8}$$

2.4. **Computational complexity analysis.** If additional information of users is discovered during recommending processes, extra off-line and online computation costs should be dealt with. Usually, the off-line computations do not affect the recommendation performance. Therefore, the online calculations are critical to the performance of prediction.

Conventional user based CF algorithms only have online computations with $O(mn)$. The time complexity of our nearest neighbor selection approach in CECF is $O(mn)$, and CECF methods require extra costs in the order of $O(n)$ for context-entropy calculation. Thus, the computational cost of our approach is $O(mn) + O(n)$, and it does not affect the performance of recommending process.

3. **Experimental Results.** In this section, the experiments are designed to test and evaluate our method. The experiments were carried out on two real world datasets provided by GroupLens Research Group at University of Minnesota and Netflix Company. The details about these datasets are shown in Table 1.

Table 1. Characteristics of two real world datasets

Datasets	*User*	*Movie*	*Rating*	*Sparsity Level*
MovieLens-100*K*	943	1682	100000	6.30%
NetFlix-100*M*	117000	8500	19000000	1.91%

To evaluate the performance of CECF, two different metrics are selected, mean absolute error (MAE) and root mean square error (RMSE). MAE is the most widely used metric for measuring the deviation of predictions generated by RSs from the user rating. The lower the MAE is, the better prediction performance is. RMSE is a statistical accuracy metric representing the accuracy of predicted rating for customers. Similar to MAE, the lower the RMSE is, the higher the accuracy is. MAE and RMSE are defined in Equations (9) and (10).

$$MAE = \frac{\sum\limits_{i=1}^{N} |P_i - Q_i|}{N} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (P_i - Q_i)^2}{N}} \tag{10}$$

where P_i is the rating prediction, Q_i is corresponding real rating, and N is the total number of user rating in user-item rating matrix.

To compare the performance of CECF, other two CF algorithms are employed: one is an item-based CF algorithm [2] (denoted by KNN), and the other is an entropy-based CF algorithm (EBCF) [8]. The experiments were done in Mablab environment.

The experimental results from MAE/RMSE comparisons of three algorithms on MovieLens-100K and Netflix-100M are shown in Figures 1 to 4, respectively.

In Figure 1 and Figure 2, the MAE and RMSE values of three algorithms on MovieLens-100K are presented respectively. On the one hand, the minimum MAE value of CECF is 0.7216 with $k = 60$. Both of KNN and EBCF obtain their best accuracy values as 0.7323

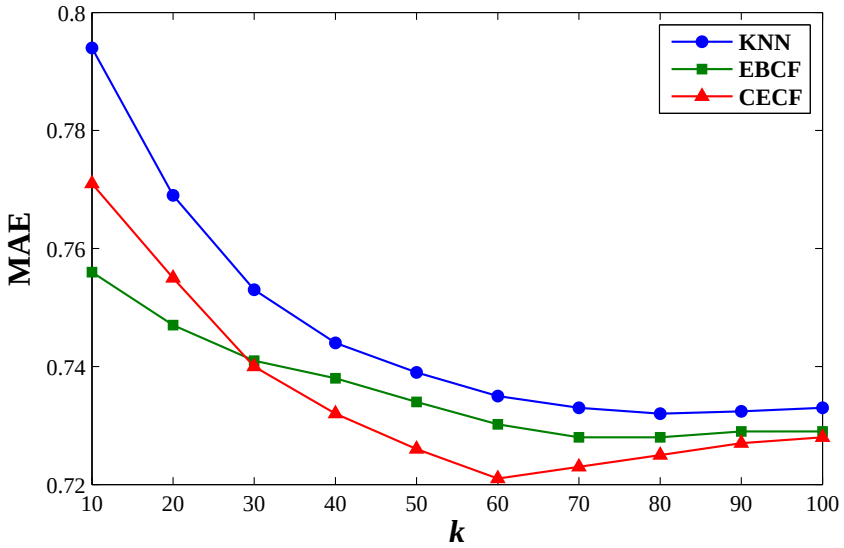

FIGURE 1. Comparisons of three algorithms' MAE results on MovieLens-100K

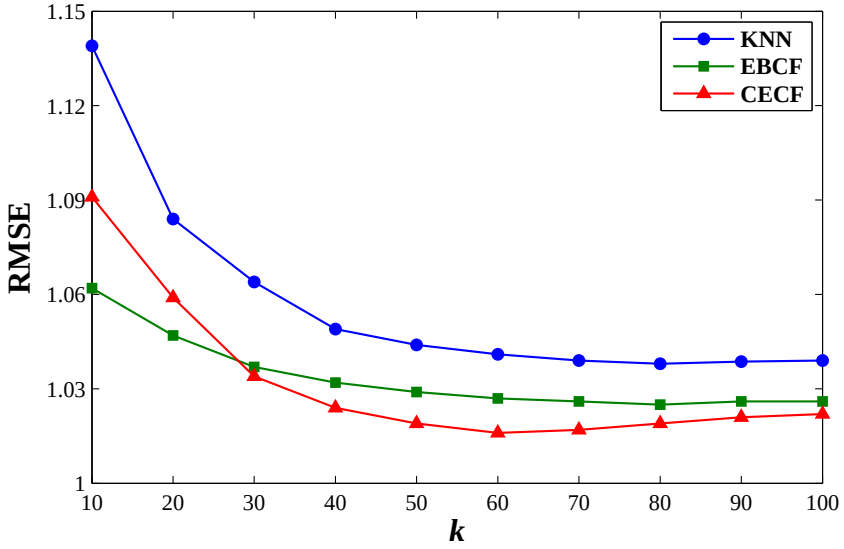

FIGURE 2. Comparisons of three algorithms' RMSE results on MovieLens-100K

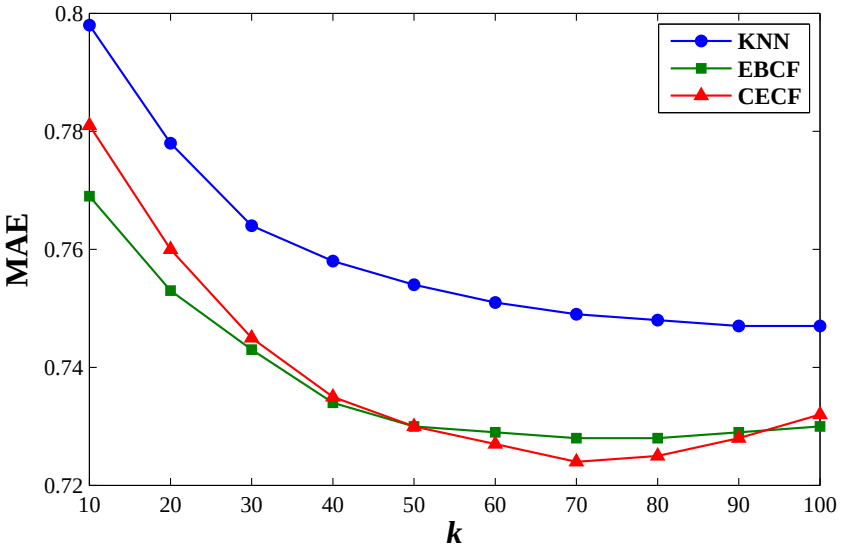

FIGURE 3. Comparisons of three algorithms' MAE results on Netflix-100M

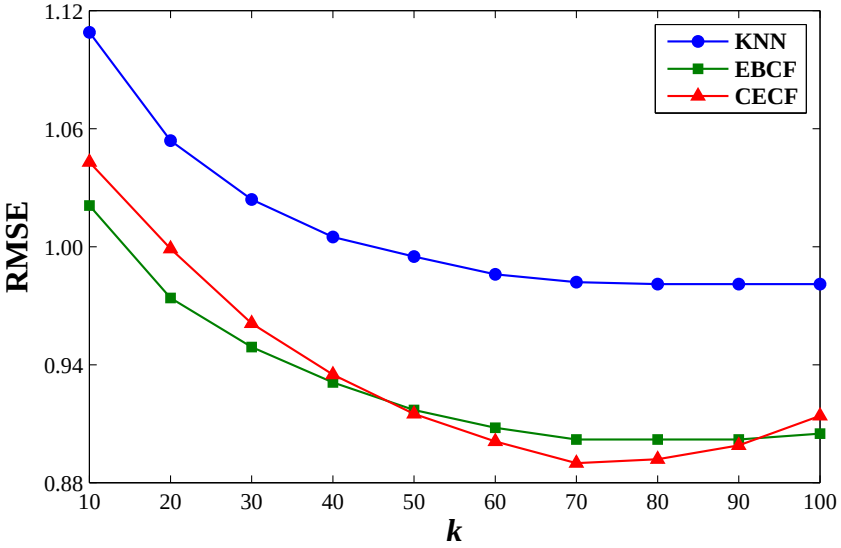

FIGURE 4. Comparisons of three algorithms' RMSE results on Netflix-100M

and 0.7285 with $k = 80$, respectively. The best MAE of CECF is 98.54% of that of KNN, and 99.05% of that of EBCF. Meanwhile, the lowest RMSE value of CECF is 1.0149 with $k = 60$. KNN and EBCF acquire their best RMSE values as 1.0386 and 1.0254 with $k = 80$, respectively. The optimal RMSE of CECF is 97.72% and 98.98% of that of KNN and EBCF, respectively. The results show that CECF has the minimum MAE and RMSE on MovieLens-100K, when $k \in [40, 100]$.

In Figure 3 and Figure 4, the MAE and RMSE values of three algorithms on Netflix-100M are shown respectively. The minimum MAE values of EBCF and CECF are 0.7273 and 0.7242 with $k = 70$, respectively, and KNN gains its minimum MAE as 0.7469 with $k = 90$. The MAE value of EBCF and CECF is much lower than that of KNN. On the other hand, EBCF and CECF obtain their minimum RMSE values as 0.9020 and 0.8945 with $k = 70$ respectively, and KNN achieves the highest accuracy with RMSE of 0.9811 as $k = 90$. It is clear that MAE and RMSE of CECF are smaller than those of KNN and

EBCF as $k = 70$ on Netflix-100M. Therefore, our proposed approach CECF outperforms the other two CF models on both datasets.

4. **Conclusions.** This paper proposed a context-entropy based CF model to improve the prediction quality of personalized recommendation. CECF employs entropy theory to measure the contextual uncertainty of users rating behaviors in RSs, which can describe users' different preferences under certain contexts. And the context-entropy is utilized for calculating user similarity to form better neighborhoods for improving the prediction accuracy. The experimental results have shown that CECF succeeds in advancing the quality of rating prediction, which reveals the potential in dealing with the sparsity issue better than traditional CF approaches. In the future, we will integrate uncertainty information of users into trust based CF methods.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 71401058), the Program for New Century Excellent Talents in Fujian Province University, NCETFJ (No. Z1625110), and the Project of Science and Technology Plan of Fujian Province of China (No. 2017H01010065).

REFERENCES

[1] Y. Shi, M. Larson and A. Hanjalic, Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges, *ACM Computing Surveys*, vol.47, no.1, pp.1-45, 2014.

[2] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, Item based collaborative filtering recommendation algorithms, *Proc. of the 10th International Conference on World Wide Web*, Hong Kong, pp.285-295, 2001.

[3] D. Ben-Shimon, L. Rokach and B. Shapira, An ensemble method for top-N recommendations from the SVD, *Expert Systems with Applications*, vol.64, pp.84-92, 2016.

[4] H. Bao, Q. Li, S. S. Liao, S. Song and H. Gao, A new temporal and social PMF-based method to predict users' interests in micro-blogging, *Decision Support Systems*, vol.55, no.3, pp.698-709, 2013.

[5] C. Chen, X. Zheng, Y. Wang, F. Hong and Z. Lin, Context-aware collaborative topic regression with social matrix factorization for recommender systems, *Proc. of the 28th AAAI Conference on Artificial Intelligence*, Quibec, pp.9-15, 2014.

[6] X. Deng and C. Jin, A novel recommendation model to mitigate new user cold start problem in mobile e-commerce, *ICIC Express Letters, Part B: Applications*, vol.6, no.7, pp.1829-1836, 2015.

[7] C. He, D. Parra and K. Verbert, Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities, *Expert Systems with Applications*, vol.56, pp.9-27, 2016.

[8] H. Chandrashekhar and B. Bhasker, Personalized recommender system using entropy based collaborative filtering technique, *Journal of Electronic Commerce Research*, vol.12, no.3, pp.214-237, 2011.

[9] F. Zhang, H. Liu and Y. Cui, Rating information entropy for cold-start recommendation, *Journal of Information & Computational Science*, vol.8, no.1, pp.16-22, 2011.

[10] W. Wang, G. Zhang and J. Lu, Collaborative filtering with entropy-driven user similarity in recommender systems, *International Journal of Intelligent Systems*, vol.30, no.8, pp.854-870, 2015.

[11] X. Deng, C. Jin, J. C. Han and Y. Higuchi, Improved collaborative filtering model based on context clustering and user ranking, *System Engineering – Theory & Practice*, vol.33, no.11, pp.2945-2953, 2013.

[12] W. Lee and K. Lee, Making smartphone service recommendations by predicting users' intentions: A context-aware approach, *Information Sciences*, vol.277, pp.21-35, 2014.

[13] J. Han, H. R. Schmidtke, X. Xie and W. Woo, Adaptive content recommendation for mobile users: Ordering recommendations using a hierarchical context model with granularity, *Pervasive and Mobile Computing*, vol.13, pp.85-98, 2014.

[14] N. Mallat, M. Rossi, V. K. Tuunainen and A. Oorni, The impact of use context on mobile services acceptance: The case of mobile ticketing, *Information & Management*, vol.46, no.3, pp.190-195, 2009.