An Improved Collaborative Filtering Algorithms Based on Maximum Entropy Model and Markov Process to Mine User Behavior

Guanghua Huang, Feng Yin* and Jiulin Feng

College of Computer Science and Technology, Southwest Minzu University, Chengdu, Sichuan, China

Abstract. Aiming at problems of data sparsity and poor interpretability of recommendation reasons in user-based collaborative filtering algorithm, an improved collaborative filtering algorithm based on maximum entropy model and markov process is proposed to mine user behavior habits. Firstly, the maximum entropy model is established through the user-behavior matrix of collaborative filtering algorithm, and the user behavior is fuzzily and preliminarily mined. Secondly, fuzzy behavior information obtained by the maximum entropy model is introduced as the initial state of markov process, and an optimization model is set up according to the internal relation of user-behavior matrix, thus reaching the state probability transition matrix. Once the markov process reaches a stable state, the user’s final behavior habit is excavated. Finally, this user behavior is integrated into the recommendation process, so that the interpretability of the recommendation is enhanced. The experimental results show that the improved collaborative filtering algorithm can promote the recommendation accuracy.

1. Introduction

Although collaborative filtering algorithm [1] is widely used in recommendation system, it has some inherent defects. And this paper mainly discuss the user-based collaborative filtering algorithm. It has such problems as sparse data matrix, weak recommendation interpretability, cold start, noise, etc. [2]. To solve these problems, Zhang Jia et al. proposed a Entropy-based collaborative filtering algorithm [3]. Lin yaojin et al. proposed A method of collaborative filtering recommendation based on fuzzy information entropy [4]. However, there are exist the problem that the user similarity will most conform to the real state when the maximum entropy is reached. Hou Jichang et al. proposed A collaborative filtering Algorithm Based on Entropy of Label and Score Difference Information Entropy [5]. Cai Xiongfeng et al. proposed a method to alleviate the data sparsity of collaborative filtering algorithm [6]. Zhang wei et al. proposed a collaborative filtering recommendation algorithm based on self-similarity matrix[7].However, there are many hidden uncertainties in real-time consideration such as stability of real-time interest similarity. This paper discusses a user collaborative filtering algorithm combined maximum entropy model with markov process. It enables the system to reach the maximum entropy while mining the optimal user behavior by relating the maximum entropy model with markov process. And eventually, it can optimize the data sparseness and improve recommendation stability and the accuracy of recommendation.

2. Maximum entropy and markov process
2.1. Maximum Entropy Model
Information entropy [8] is defined as formula (1). In a system, the higher the entropy, the more stable the system is. Therefore, The maximum entropy model [9] can be expressed as formula (2).

\[
H(P) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1}
\]

\[
L(P, w) = -H(P) + w_1(1 - \sum_{y} p(y | x)) + \sum_{i=1}^{n} w_i(E_i(f_i) - E_i(f_i)) \tag{2}
\]

2.2. Markov Process
In the Markov process [10], the transition from one state to another in the state space is a random process, and the transition is memoryless. The mathematical expression is shown in formula (3).

\[
P\{x_{n+1} | x_0, x_1, \ldots, x_n\} = P\{x_{n+1} | x_n\} \tag{3}
\]

Markov chain [11] is a typical markov process, and eventually it will converge to a steady state distribution \(\pi^*\). The steady state distribution refers to the stable state (formula 5) finally reached after the state jump is carried out according to the probability matrix P (formula 4). That is:

\[
\pi_{t+1}^T = \pi_t^T A = \pi_0^T A^t \tag{4}
\]

\[
\lim_{k \to \infty} \pi^T A^k = \pi^* \tag{5}
\]

3. Improvement user-based CF algorithm(MM-CF)
For the existing data set X, the entropy of the entire data set can be obtained by solving formula (1). In order to reduce the time and space complexity of computation and to dig out the hidden information in the behavior matrix, the concept of user behavior habit is introduced. Taking MovieLens data [12] set of GroupLens project group as an example, subconsciously, people have behaviors they have developed over the years, and these behaviors affect their propensity to rate movies. Since user tendency is the degree of user preference within a range, it is considered to classify the score. By archiving and counting the data, the dimension of the user matrix is reduced from \(m \times n\) to \(m \times k\), wherein \(k\) is the number of categories that are archived. When the project set is large, the sparsity problem of data matrix can be obviously solved. But sorting is not enough. Every user should have his or her own hidden behavior habits, but they are sometimes uncontrollable and unrepresentative. Therefore, we want to be able to unearth behavior that is characteristic of groups. As we all know, habits (such as film ratings) are biased. We use the weight value to express this tendency. Suppose that the weight vector of user behavior habit is \(q\). So it represents the proportion of different behavior habit states, That is:

\[
w = [w_1, w_2, \ldots, w_k], \quad \left(\sum_{q=1}^{n} w_q = 1\right) \tag{6}
\]

After introducing this, The calculation of entropy in the hidden behavior habit should be defined as:

\[
H(x) = -\sum_{q=1}^{n} w_q p(x_q) \log [w_q p(x_q)] \tag{7}
\]

formula (6) can be obtained through the maximum entropy model. At this point, we preliminary get the hidden user behavior tendency that conforms to the characteristics of the group. But we all know that users' behaviors and hobbies are not static. We hope that other factors will not affect hidden user behavior habits that are unearthed. Therefore, we make equation (6) reach the steady-state distribution through the markov process, and finally obtain the behavior habit that is consistent with the group characteristics that is not affected by other factors.
Obviously, the probability transfer matrix of a single user is not representative. So we need to get the probability transfer matrix that conforms to the group characteristics. According to formula (4), the estimated probability after \( k \)-step transition can be obtained. If the true value of the probability after \( k \)-step transfer is \( \hat{\pi}_k \), the fitting error is formula (8):

\[
    e_r(k) = \hat{\pi}_k - \pi_k
\]

We can obtain the state probability transition matrix from this data set through the optimization model [9], then the model is shown in formula (9):

\[
    \min Q = \sum_{i,j} (u_r(k) + v_r(k))

    \begin{cases}
        u_r(k) = \frac{|e_r(k)| - e_r(k)}{2} \\
        v_r(k) = \frac{|e_r(k)| + e_r(k)}{2}
    \end{cases}

    \sum_{j=1}^{n} p_j = 1, i = 1, 2, \ldots n

    p_j \geq 0, i, j = 1, 2, \ldots n
\]

The probability transfer matrix is solved by formula (7), then the steady-state distribution reached by the final markov chain can be solved through formula (4) and formula (5). In this way, we finally find the user behavior habits that are not affected by external factors and conform to the characteristics of the group.

After digging out the final user behavior, for the public, the real score value in line with the behavior habits of the public should be the result obtained by multiplying the value shown in the user behavior matrix and the weight obtained since we're solving for weights in grades. So, the new formula for calculating cosine similarity and prediction formula for scoring are shown as formula (10) and (11):

\[
    \text{sim}(u_i, u_j) = \frac{\sum_{q=1}^{m} (w_{i,q} - \bar{r}_q)(w_{j,q} - \bar{r}_q)}{\sqrt{\sum_{q=1}^{m} (w_{i,q} - \bar{r}_q)^2 \sum_{q=1}^{m} (w_{j,q} - \bar{r}_q)^2}}
\]

\[
    r_{ij} = \bar{r}_q + \frac{\sum_{q=1}^{m} \text{sim}(u_i, u_j)(w_{j,q} - \bar{r}_q)}{\sum_{q=1}^{m} \text{sim}(u_i, u_j)}
\]

The idea of the MM-CF algorithm is illustrated in Figure 1.

![Figure 1. Idea of the MM-CF algorithm](image)

4. Experimental
4.1. Evaluation Index
MAE (Mean Absolute Error) is an effective method to detect fitting problems. One point needs to be noted that since the original user-behavior matrix has sparse items (items with a value of 0), MAE’s calculation should eliminate these sparse items. MAE is calculated as shown in References [13].

4.2. Results and Analysis

4.2.1. Mining User Behavior.
According to Movielens' dataset, users are rated in three levels, with 0-2 for low, 2-4 for normal and 4-5 for high. So the dimension of the user matrix is reduced from m×n to m×3. By formula (2), the weight vector of initial user behavior habit can be obtained as equations (12):
\[ \pi_v = [0.2802, 0.3498, 0.37] \]
According to the this values, the state probability transfer matrix of Movielens data can be gained as follows:
\[ P = \begin{pmatrix} 0.116425 & 0.662153 & 0.221421 \\ 0.161311 & 0.629433 & 0.209245 \\ 0.128783 & 0.617747 & 0.25347 \end{pmatrix} \]
(13)
After the values are obtained, we can dig out the stable user behavior habit with group characteristics of the data set according to equations (4) and (5), and the value is:
\[ \pi = [0.1477, 0.6312, 0.2211] \]
So, prediction and algorithm evaluation are carried out. It show as equations (14).

4.2.2. MM-CF Prediction and Evaluation.
The Movielens dataset was predicted by using a modified cosine similarity. By selecting different similar user sets, the prediction and evaluation results with different similar user sets are obtained. And further, by comparing the traditional user-based CF algorithm [1], Entropy-based CF algorithm [3], A method of CF recommendation based on fuzzy information entropy[4], CF recommendation algorithm based on the self-similarity matrix[7], A CF Algorithm Based on Entropy of Label and Score Difference Information Entropy [5], [1] exist problem of data sparsity and weak recommendation interpretability. [3] and [4] did not consider that the system is the most stable when the entropy reaches the maximum. [5] did not consider there are many hidden uncertainties factors in real-time(such as the change of the user-interest). [7] did not consider that the data will still be affected if the initial matrix is too sparse. The MM-CF algorithm solved those problems. And experimental results are shown in Figure 2. Maybe the Figure 2 can verify the correctness of the algorithm.

As can be seen in the Figure2, when k number increases, the accuracy of the algorithm increases continuously. Owing to constantly decreasing of the MAE of the algorithm, the recommendation accuracy of the algorithm is constantly improving, Because the smaller the error, then, the more
accurate the recommendation. The experimental results verify that the improved algorithm proposed in this paper is obviously superior to other algorithms. To sum up, those improvement is feasible.

5. Conclusion
This paper proposes an improved collaborative filtering algorithm based on maximum entropy model and markov process to mine user behavior and habits in view of the problem of data sparsity problem and poor interpretability of recommendation reasons problem that existing in the traditional user-based collaborative filtering algorithm. And the final Experimental results show that algorithm proposed in this paper can effectively solve the problem of data sparsity and enhance the interpretability of recommendation reasons. However, the algorithm mainly compares the entropy-related algorithms, and has less contrast with other types of algorithms. The next work focus will be transferred to the comparison of related Markov algorithms.

6. References
[1] M. Elahi, F. Ricci, and N. Rubens, A survey of active learning in collaborative filtering recommender systems. Comput. Sci. Rev. vol. 20, pp. 29–50, May 2016.
[2] Dheeraj Bokde, Sheetal Girase, Debajyoti Mukhopadhyay. Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey[J]. Procedia Computer Science, 2015, 49.
[3] ZHANG jia, LIN Yaojin, Lin meng lei, et al. Entropy-based collaborative filtering algorithm. Journal of Shandong University (Engineering Science), 2016, 46(02):43-50.
[4] LIN Yaojin, ZHANG jia, LIN Menglei, et al. A method of collaborative filtering recommendation based on fuzzy information entropy. Journal of Shandong University (Engineering Science), 2016, 46 (05):13-20.
[5] HOU Jichang, CHEN Jiaqi. A Collaborative Filtering Algorithm Based on Entropy of Label and Score Difference Information Entropy. Electr Sci Technol, 2018, 31(05):57-61, 65.
[6] CAI Xiong-feng, AI Li-hua, DING Ding. An method for alleviating data sparsity in collaborative filtering algorithm. Computer engineering, 2015, 36 (03): 41-47.
[7] ZHANG Wei, ZHENG Jun, et al. Collaborative filtering recommendation algorithm based on the self-similarity matrix. Journal of East China Normal University (Natural Science), 2018(04):120-128, 146.
[8] Liu W, Fan M, Wang G. Adaptive flow abnormality identification based on information entropy. Concurrency Computat Pract Exper. 2019;31:e4713.DOI= https://doi.org/10.1002/cpe.4713
[9] A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39–71, 1996.
[10] Markov, A. A. (1971). Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Dynamic Probabilistic Systems, volume 1: Markov Chains. Chichester: John Wiley and Sons.
[11] Xu Zhihui. Research of Method to Estimate Markov State Transition Probability Matrix. Northeast Agricultural University, 2013.
[12] M. Sadeghian and M. Khansari, "A Recommender Systems Based on Similarity Networks: MovieLens Case Study," 2018 9th International Symposium on Telecommunications (IST), Tehran, Iran, J. J. Davidson B. Liebald J. Liu P. Nandy T. Van Vleet U. Gargi S. Gupta Y. He M. Lambert B. Livingston et al. "The youtube video recommendation system " in Proceedings of the fourth ACM conference on Recommender systems. ACM 2010 pp. 293-296. 2018, pp. 705-709.
[13] Kaleli C. An Entropy-Based Neighbor Selection Approach for Collaborative Filtering. Knowledge-Based Systems, 2014, 56: 273–280.

Acknowledgments
Thanks to my Supervisor Yin Feng for his patient guidance. And This paper was funded by Southwest Minzu University postgraduate innovative research project (CX2019SZ62), And also funded by The National Natural Science Foundation of China (61105061), Sichuan science and technology support project (14ZC1767). Express my heartfelt thanks to them.