Postprint: An Improved Collaborative Filtering Recommendation Algorithm Based on User Rating Time

Authors: Li Daoguo, Jet Li, Shen Enping

Date: 2017-11-08T00:00:00+00:00

Abstract

[Purpose] To improve the user-based collaborative filtering algorithm to alleviate the problems caused by data sparsity and scarcity of co-rated items, thereby enhancing the accuracy of rating prediction.

[Method] We propose to incorporate user rating timestamps to identify users with similar rating behaviors, and integrate the similarity of user rating variance into the similarity calculation, making the nearest neighbor selection for target users more reasonable.

[Results] Experimental results demonstrate that, compared with the user-based collaborative filtering algorithm, the Mean Absolute Error (MAE) of the proposed algorithm decreases by approximately 2%, which improves the recommendation effectiveness of the system to a certain extent.

[Limitations] The algorithm has only been experimentally tested on the MovieLens dataset and needs to be validated on other datasets.

Conclusion The proposed algorithm can effectively improve recommendation accuracy and possesses certain feasibility and practical significance.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issue 274, 2016, No. 9

Improved Collaborative Filtering Recommendation Based on User Rating Time

Li Daoguo[1], Li Lianjie[2], Shen Enping[2]
[1](School of Information Engineering, Hangzhou Dianzi University, Hangzhou

310018, China)
[2](School of Management, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract

[**Objective**] This study aims to improve the user-based collaborative filtering algorithm to mitigate problems caused by data sparsity and scarce common ratings among users, thereby enhancing rating prediction accuracy. [**Methods**] We propose an approach that identifies users with similar rating behaviors by incorporating rating timestamps, and integrates user rating variance similarity into similarity computation to enable more rational selection of nearest neighbors for target users. [**Results**] Experimental results demonstrate that the new algorithm reduces the Mean Absolute Error (MAE) by approximately 2% compared to traditional user-based collaborative filtering, thereby improving recommendation effectiveness to some extent. [**Limitations**] The algorithm was only tested on the MovieLens dataset and requires validation on additional datasets. [**Conclusions**] The proposed algorithm can effectively improve recommendation accuracy and possesses certain feasibility and practical significance.

Keywords: Collaborative filtering; Data sparsity; Similar rating; User rating variance similarity; Nearest neighbor

Introduction

The rapid development of the Internet has ushered humanity into a new information era, with increasingly abundant information resources available online. When users confront massive amounts of data, how to quickly and accurately locate needed information becomes a critical concern, often leading to the loss of potential users—this is the so-called "information overload" phenomenon [**?**, **?**]. To help users efficiently find required information within vast datasets, personalized recommendation systems have emerged. Collaborative filtering recommendation algorithms represent the most widely applied technique in this domain, offering the advantages of having no special requirements for recommended items and being capable of handling unstructured, complex objects such as articles, movies, and books. These algorithms analyze user-item rating matrices to filter out large amounts of unnecessary information and ultimately identify items of interest to users [**?**].

Although collaborative filtering recommendation algorithms demonstrate unique advantages in many aspects, their primary drawback is excessive dependence on rating matrices. As websites experience rapid growth in both user and product numbers, the number of items users actually rate in the rating matrix becomes extremely small, typically below 1%. When data becomes overly sparse, common rating items among users in the recommendation system become extremely scarce, resulting in inaccurate similarity calculations between users and consequently degraded recommendation quality. Therefore, this paper proposes a

collaborative filtering recommendation algorithm improved based on user rating time, which can effectively alleviate problems caused by extremely sparse data and scarce common ratings among users. By optimizing the nearest neighbor identification method, the algorithm enhances recommendation accuracy.

Numerous scholars have conducted extensive research on mitigating the impact of data sparsity on recommendation systems, with approaches generally falling into two categories: utilizing certain methods to reduce data sparsity, and improving recommendation algorithms to enhance recommendation quality. Regarding algorithm improvement research, since identifying target users' nearest neighbors constitutes the core of collaborative filtering algorithms and plays a crucial role in recommendation effectiveness [**?**], the accuracy of similarity calculations between users becomes critically important. Traditional similarity calculation methods primarily include cosine similarity, adjusted cosine similarity, and Pearson correlation similarity [**?**, **?**]. The cosine similarity calculation process fails to fully utilize rating timestamp information, yet temporal information plays a key role in determining the usefulness of item ratings. Although adjusted cosine similarity and Pearson correlation similarity mitigate the impact of rating values, they do not account for a relatively special situation: when the rating matrix is exceptionally sparse, common rating items between two users become extremely few, making similarity calculations using these methods similarly inaccurate and resulting in poor recommendation quality.

To address these issues, this paper proposes incorporating user rating timestamps to identify users with similar rating behaviors, thereby improving the nearest neighbor identification method in traditional collaborative filtering algorithms. Building upon this foundation, we integrate user rating variance similarity to more comprehensively utilize user rating information for improving similarity calculations. This approach enables relatively accurate computation of user similarities even under conditions of extremely sparse data and scarce common ratings, achieving the goal of improving recommendation accuracy.

Algorithm Description

Definition 1: Similar Rating Item

Let uiT denote the time when user u rated item i. If both users u and v have rated item i, and the difference between uiT and viT is smaller than a predefined time interval, then item i is identified as a similar rating item uvS for users u and v.

Definition 2: Similar User Behavior

If the number of similar rating items between two users is greater than or equal to a specified threshold λ, these two users are considered to have similar user behavior, as expressed by the following formula:

Definition 3: User Rating Variance Similarity

This paper introduces user rating variance into similarity calculation to measure differences between users, proposing the User Rating Variance Similarity (URVS) theory. The calculation method is shown in Formula (2):

$$\mathrm{URVS}(u,v) = \frac{1}{1+|\mathrm{Var}_u - \mathrm{Var}_v|}$$

where Var_u and Var_v represent the rating variances of users u and v, respectively. For example, if the variances of users a, b, and c are 1, 3, and 5, respectively, then $\mathrm{URVS}(a,b) > \mathrm{URVS}(a,c)$. Larger rating variance indicates greater user rating controversy, while smaller variance indicates less controversy.

Improved Similarity Calculation

This paper uses adjusted cosine similarity as an example. The similarity calculation method incorporating user rating time is shown in Formula (3):

$$\mathrm{sim}_{\mathrm{time}}(u,v) = \frac{\sum_{i\in uvS}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{\sqrt{\sum_{i\in uvS}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i\in uvS}(r_{vi}-\bar{r}_v)^2}}$$

Building upon the incorporation of user rating time, the similarity calculation introducing user rating variance similarity is shown in Formula (4):

$$\mathrm{URVS\_CF}(u,v) = \alpha \cdot \mathrm{sim}_{\mathrm{time}}(u,v) + (1-\alpha)\cdot \mathrm{URVS}(u,v)$$

where α is the weighting coefficient. The advantage of the collaborative filtering recommendation algorithm improved based on user rating time lies in its ability to fully utilize user rating information and relatively accurately calculate similarities between users even when system data is extremely sparse and common ratings among users are scarce, thereby improving recommendation accuracy.

Main Steps of the Improved Algorithm

Input: User-item rating matrix, target user u
Output: TOP-N item recommendation list for target user u

When $uvS < \lambda$, it indicates that users u and v have few similar rating behaviors, and such pairs should be discarded in similarity calculations to avoid affecting accuracy. The threshold value λ for user similar rating items directly influences algorithmic accuracy in experiments, and its value should be determined based on specific experimental conditions to obtain optimal results.

1. Based on the user-item rating matrix R, calculate the similarity between user u and other users using the improved adjusted cosine similarity calculation method (Formula (4)). If the number of movies rated within a certain time period is too small, set the similarity Sim to 0.

2. Based on the similarities calculated in step 1, identify k nearest neighbors for target user u. Let the nearest neighbor set be $\{v_1, v_2, ..., v_k\}$, and let the similarity between user u and its nearest neighbors be $\text{sim}_1, \text{sim}_2, ..., \text{sim}_k$.

3. Identify items rated by target user u and similar neighbors, denoted as uI and $\{i_1, i_2, ..., i_m\}$, respectively. Take the union of all iI, then remove items already present in set uI to generate candidate set Z.

4. For each item $j \in Z$ in the candidate set, predict user u' s rating for item j using Formula (6):

$$P_{uj} = \bar{r}_u + \frac{\sum_{v \in N(u)} \text{URVS\_CF}(u, v) \cdot (r_{vj} - \bar{r}_v)}{\sum_{v \in N(u)} |\text{URVS\_CF}(u, v)|}$$

where $N(u)$ represents the nearest neighbor set of user u.

5. Sort the predicted ratings for items from step 4 in descending order and select the top n items with the highest ratings to recommend to user u.

Experimental Data and Environment

This study employs the 100K dataset from the MovieLens dataset created by the GroupLens research group at the University of Minnesota [**?**]. The dataset contains 100,000 ratings from 943 users on 1,682 movies, with rating values ranging from 0 to 5, where higher values indicate greater user preference [**?**]. We randomly selected 80% of the dataset as the training set and the remaining 20% as the test set [**?**]. The sparsity of the dataset is calculated as follows:

$$\text{Sparsity} = 1 - \frac{\text{Number of rated entries}}{\text{Total number of users} \times \text{Total number of items}} = 1 - \frac{100,000}{943 \times 1,682} \approx 93.7\%$$

This demonstrates that the rating matrix of the selected dataset is extremely sparse.

The experimental environment consists of an Intel(R) Core(TM) i3-2310M 2.10GHz CPU, 2GB RAM, Microsoft Windows 7 operating system, with algorithms implemented in MATLAB.

Evaluation Metrics

We adopt Mean Absolute Error (MAE) to evaluate the recommendation quality of the system. MAE measures algorithm performance by calculating the differ-

ence between actual ratings and predicted ratings. Smaller MAE values indicate better algorithm performance. The MAE calculation is shown in Formula (7):

$$\text{MAE} = \frac{\sum_{(u,i)\in T} |P_{u,i} - R_{u,i}|}{n}$$

where $P_{u,i}$ represents the predicted rating of user u for movie i, $R_{u,i}$ represents the actual rating of user u for movie i, and n denotes the number of rating pairs in the test set.

Experimental Analysis

Experiment 1: Impact of Parameter λ on Recommendation System Performance

We calculate similarities between users using Formula (3) and predict ratings for unrated items based on the rating prediction formula. Since this experiment primarily tests the impact of parameter λ on MAE values, we control the number of nearest neighbors at 30 while varying the threshold values at 3, 5, 8, 10, 12, and 14. The variation in MAE values is shown in Figure 1 [Figure 1: see original paper].

As illustrated in Figure 1, the MAE value reaches its minimum and recommendation accuracy peaks when threshold λ is set to 10. Therefore, we set $\lambda = 10$ in subsequent experiments.

Experiment 2: Impact of Parameter α on Recommendation System Performance

This experiment examines the effect of parameter α on MAE values, again controlling the number of nearest neighbors at 30. The variation of MAE values with different α values is shown in Figure 2 [Figure 2: see original paper].

Figure 2 reveals that the MAE value is minimized and recommendation results are optimal when $\alpha = 0.2$. As α gradually increases to 0.2, the MAE value decreases progressively; however, as α continues to increase beyond 0.2, the MAE value begins to rise slowly. Experimental results demonstrate that α plays a crucial role in collaborative filtering, and only by selecting an appropriate α can we obtain optimal recommendation targets and achieve the best recommendation results with minimal MAE values. Consequently, we set $\alpha = 0.2$ in Experiment 3.

Experiment 3: Variation of Recommendation Performance with Number of Nearest Neighbors

To validate the effectiveness of the proposed collaborative filtering recommendation algorithm improved based on user rating time, we conduct comparative

experiments. We compare the recommendation accuracy of the improved algorithm with traditional user-based collaborative filtering on the MovieLens dataset. Based on results from Experiments 1 and 2, we set $\lambda = 10$ and $\alpha = 0.2$. The variation of MAE values with different numbers of nearest neighbors is shown in Figure 3 [Figure 3: see original paper].

Figure 3 demonstrates that the MAE values of the improved collaborative filtering recommendation algorithm based on user rating time are lower than those of traditional user-based collaborative filtering across nearest neighbor counts of 10, 20, 30, 40, and 50, with an average MAE reduction of 2%. This indicates that considering user rating time in similarity calculations and introducing user rating variance similarity significantly improves recommendation accuracy.

Conclusion

This paper addresses the shortcomings of traditional user-based collaborative filtering recommendation algorithms by proposing an improved collaborative filtering recommendation algorithm based on user rating time. The new algorithm considers scenarios where the user-item rating matrix is extremely sparse and common rating items between two users are extremely scarce, which leads to inaccurate similarity calculations and degraded recommendation accuracy. To tackle this problem, we identify users with similar rating behaviors by incorporating rating timestamps and fuse user rating variance similarity to improve traditional similarity calculations between users, thereby optimizing the nearest neighbor identification method for target users. Experimental results show that the proposed algorithm can effectively improve recommendation accuracy and possesses certain feasibility and practical significance.

References

[1] Zhang Li, Qin Tao, Teng Piqiang. An Improved Collaborative Filtering Recommendation Algorithm Based on User Clustering[J]. Information Science, 2014, 32(10): 24-27.

[2] Fang Yaoning, Guo Yunfei, Ding Xuetao, et al. An Improved Singular Value Decomposition Recommender Algorithm Based on Local Structures[J]. Journal of Electronics & Information Technology, 2013, 35(6): 1284-1289.

[3] Sun Hui, Ma Yue, Yang Haibo, et al. Collaborative Filtering Recommendation Algorithm by Optimizing Similarity and Clustering Users[J]. Journal of Chinese Computer Systems, 2014, 35(9): 1967-1970.

[4] Gao Xiang. Research of Collaborative Filtering on Recommendation Systems for E-Commerce[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2011.

[5] Xu Zhihong, Wang Baoying. Collaborative Filtering Recommendation Algorithm Based on Item Complex Similarity[J]. Application Research of Computers,

2014, 31(2): 398-400.

[6] Wen Junhao, Shu Shan. Improves Collaborative Filtering Recommendation Algorithm of Similarity Measure[J]. Computer Science, 2014, 41(5): 68-71.

[7] Yan Dongmei, Lu Chenghua. Optimized Collaborative Filtering Recommendation Algorithm Based on Users' Interest Degree and Feature[J]. Application Research of Computers, 2012, 29(2): 497-500.

[8] Zhao Xue. The Personalized Collaborative Filtering Recommendation Algorithm Based on User Interest[D]. Qinhuangdao: Yanshan University, 2014.

Conflict of Interest Statement

All authors declare no conflict of interest.

Author Contributions

Li Daoguo: Conceived research ideas and designed research methodology; Li Lianjie, Shen Enping: Analyzed data and conducted experiments; Li Daoguo, Li Lianjie: Drafted manuscript and revised final version.

Supporting Data

Supporting data is available in the online version of the journal at http://www.infotech.ac.cn.

[1] Li Daoguo, Li Lianjie, Shen Enping. base.base. Basic dataset.
[2] Li Daoguo, Li Lianjie, Shen Enping. test.test. Test dataset.

Received: 2016-04-22
Revised: 2016-05-24

Note: Figure translations are in progress. See original paper for figures.