Hybrid Recommendation Algorithm Based on Long-term and Short-term Interest and Matrix Factorization for Collaborative Filtering

Yong Xu¹,* and Ni Zhu²

¹South China University of Technology, China
²South China University of Technology, China

*Corresponding author email: xuyong@scut.edu.cn

Abstract. This paper proposes a hybrid recommendation algorithm which based on user interests and matrix factorization. It solves problems such as spares data, the difference between long-term and short-term interest, overemphasis of time and ignoring user new interests. The proposed algorithm distinguished users' interest through time window. Then it obtained the distribution of user interests. Finally, it integrated into matrix factorization to explore more new interests. The result shows that short-term interest should be recommend first. Compared with the traditional matrix factorization for collaborative filtering, forgetting curve and time window, the proposed algorithm shows superior performance on precision, recall, and F1-Score. It extracts user's interest sets, user activity, item popularity and other related indicators, which can automatically tag users and provide more options for subsequent research on the dynamic evolution of user interests or the expansion of website functions.

Keywords: Interest drift; Collaborative filtering; Matrix factorization; Long-term and short-term interest; Recommendation.

1. Introduction

With the continuous increase of the internet penetration rate, the e-commerce platform has achieved steady development. The number of users and commodities are growing exponentially, which brings a lot of information and leads to information overload. According to iiMedia Research, in the first half of 2019, China's total online retail sales have reached 19,520.97 billion-yuan, accounting for 24.7% of total social retail sales [1]. Traditional research assumes that user interests are fixed and do not change over time. In fact, interest drift occurs in many areas, such as changing the style of clothes with age. The classic solutions to handle interest drift are time window and forgetting curve. These algorithms treat user interests equally, without considering the combined effects of original and recently arrived interests. Long-term interest is relatively stable, while short-term interest is unstable, but it has high immediacy. In addition, most literature focuses on processing explicit rating data. However, in many practical situations, especially in e-commerce platform, it is necessary to focus on implicit feedback.

This paper proposes a hybrid recommendation algorithm LSIMF (Long-term and Short-term Interest and Matrix Factorization for Collaborative Filtering) based on implicit feedback. By tracking and modeling user interests, the user's long-term and short-term interest patterns are extracted, which can capture user interest drift and obtain user interest distribution, further enrich user portraits, and provide personalized recommendation strategies for different users. By fusing matrix factorization algorithm, LSIMF can realize the exploration of the user’s new interests while capturing the user’s interest drift.
2. Literature References

Static recommendation doesn’t consider time factor, which can’t accurately track the changes of user interests. Researches on dynamic interest models mainly include: time information, matrix factorization, long-term and short-term interest.

2.1. Time Information

Forgetting curve and time window and their deformations are mainly used of time information. Items of recent interest are more meaningful than items of interest a long time ago, so appropriate time decay function can be used to ensure the real situation of user preferences [2]. Fusion recommendation algorithm based on forgetting curve can improves the accuracy of recommendation and user satisfaction [3]. Using time window to form a new probability matrix factorization can improve the accuracy of prediction and reduces the complexity of time [4].

2.2. Matrix Factorization

Matrix factorization is faced with the problem of data sparsity, which can be solved by adding additional information. It can also get the potential characteristics of users and items through deep learning [5]. In social network, semantic attitude recognition can be made from user generated content. Feltoni et al. have designed a three-dimensional model. Each user's semantic attitude corresponds to a matrix factorization [6]. And matrix factorization integrates the user's social status and affinity can improve the recommendation quality [7].

2.3. Long-term and Short-term Interest

The long-term interest changes slowly over time, and the short-term interest dynamically changes in a short time [8]. In the personalized recommendation of social network, the long-term and short-term interest of users need to be considered comprehensively. The tag, user behavior and time are important factors to study user preferences [9]. In the field of personalized news recommendation, users' long-term interest can be captured by learning their complete historical behavior records, while users' short-term interest can be obtained by learning their recent reading history [10].

3. The Proposed Methodology

3.1. Identification of Long-term and Short-term Interest

The traditional recommendation algorithm only considers the user similarity or item similarity, and does not consider the user's interest changes over time. It treats the user's behavior at different times equally, resulting in the accuracy of the recommendation algorithm decreasing over time. Users are more interested in recently viewed products, while products viewed a long time ago have a relatively small impact on users' current interests. To solve data sparsity, user-item matrix is transformed into the user-category matrix. This paper divides user interests into short-term interest set S and long-term interest set L. It is controlled by time window. When the number of visits to category c within the time window exceeds a threshold, c is added to set S. Outside the time window, when the number of visits to category c is greater than the average number of visits to all item categories , c is added to set L. The specific steps are as follow:

1. Set the size of the time window T;
2. Calculate the number of visits \( N_{uc} \) of user u for item category c;
3. Set the short-term interest threshold \( \alpha \) in the time window, when \( N_{uc} > \alpha \), add c to the set S;
4. Outside the time window, calculate the average number of visits \( \text{avg}_u \) of user u to all item categories. When \( N_{uc} > \text{avg}_u \) , add c to set L;
5. Adjust the parameters T, \( \alpha \) according to the hit rate of long-term and short-term interest sets by comparing what user actual buy on the forecast day;

3.2. Matrix Factorization for Collaborative Filtering Based on Implicit Feedback

Implicit feedback data can be converted into two dimensions, the degree of preference and confidence [11]. First introduce a set of binary variables that represent the user's preference for the item, as in Eq.
\[
\begin{align*}
    p_{ui} &= \begin{cases} 
        1 & r_{ui} > 0 \\
        0 & r_{ui} = 0
    \end{cases} 
\end{align*}
\]

When user u has consumed item i \((r_{ui} > 0)\), it is assumed that user u likes i \((p_{ui} = 1)\). However, when the user u consumed item i, it may not mean that u like i. Therefore, a set of variables \(c_{ui}\) can be introduced to express our confidence in the prediction result, as in Eq. 2:

\[
c_{ui} = 1 + \alpha r_{ui}
\]

The goal is to find a set of vectors \(x_u \in \mathbb{R}^f\) for each user u and a set of vectors \(y_i \in \mathbb{R}^f\) for each item i, so that the two sets of vectors can affect user preferences. The loss function is shown in Eq. 3:

\[
\min_{x_u,y_i} \sum_{u,i} c_{ui} (p_{ui} - x_u^T y_i)^2 + \lambda \left( ||x_u||^2 + ||y_i||^2 \right)
\]

Compared with explicit feedback data, implicit feedback data has a wider range and is easier to collect. And recommendation systems based on implicit feedback data are more time-sensitive. In addition, the data scale of implicit feedback data is a multiple of explicit feedback data. In order to apply user implicit feedback data to traditional recommendation algorithms, some scholars have tried to convert implicit feedback data into explicit data for research. This paper converts user behavior data into a score between 1 and 5. The input data includes user id, item id, and the sum of user's score, solving by ALS algorithm and the output is the corresponding user's preference for item.

### 3.3. Hybrid Recommendation

The TOP N recommendations based on user interests set have a high accuracy rate, but the elements in set are limited because of the limited interests of user. On the other hand, the TOP N recommendation set based on matrix factorization collaborative filtering is sufficient, but it can't capture user's interest drift. In this paper, linear weighting is used for Hybrid, as in Eq. 4:

\[
LSIMF\_Score = (1 - \alpha) \cdot LSI\_Score + \alpha \cdot MF\_Score
\]

Collaborative filtering based on matrix factorization can directly obtain the confidence weight value of user u interested in item i, which is defined as MF\_Score, and user interests set needs to further quantify the size of the interest value, which is defined as LSI\_Score.

### 4. Experiment

#### 4.1. Dataset

The experimental data in this paper uses the Taobao 2015 Double 11 transaction data set from the Xiamen University Database Lab, which includes user id, item id, item category, and type of user behavior (including clicks, follow, add to cart, purchase) and time. Which can extract six dimension of indicators: item category, time factor, frequency, whether purchased, user activity, and item popularity to enrich user description files.

#### 4.2. User Interests Distribution

Interest set is extracted from the user's historical data, then actual purchase behavior of the user on the prediction day is compared to obtain the accuracy of the prediction. According to the long and short interest extraction rules, the time window size T and the short-term interest judgment threshold \(\alpha\) can affect the experimental results. Therefore, a comparison experiment is performed based on the different values of the time window T and the short-term interest judgment threshold \(\alpha\). As in Fig.1.
According to the experiment, we get the best parameters to divide long-term and short-term interest. As shown in Fig.2, in the user’s large historical behavior records, 69% of user behaviors are marked as uninteresting. 20% of user behaviors are marked as long-term interest. 11% of user behaviors are marked as short-term interest. Compared with the item purchased on the double 11 day of 2015, as shown in Fig.3. 24% of the items have been marked as short-term interest, and 20% of the items have been marked as long-term interest. This shows that it is unscientific to recommend short-term interest only. In addition, 37% of the items have no historical behavior records. This part represents the accidental interests of users. And it can’t be predicted by the user's historical behavior data. However, users can be recommended through other algorithms, such as collaborative filtering.

4.3. Recommendation Process
First, implement TOP N recommendation based on user interests set. When making top N recommendations, the ranking of elements in the recommendation set needs to be considered. Taking the top 5 recommendation as an example, the confusion matrix is shown in Table 1.

| Confusion matrix | Predict |  |
|------------------|---------|------------------|
|                  | Positive| Short-term set:TP S=368, Long-term set:TP L=502 |  |
|                  | Negative| Short-term set:FP S=497, Long-term set:FP L=2437 | FN=1462  |
| True             |         | TN=19344         |  |

According to the calculation in Table 1, for short-term interest, the precision is 42.54%, recall is 20.11%, and F1-Score is 27.21%; For long-term interest, the precision is 17.08%, and recall is 25.56% , F1-Score is 20.47%. The F1-Score of the short-term interest is about 6% higher than long-term interest, so short-term interest should be recommended first, then long-term interest.

Second, implement TOP N recommendation based on matrix factorization for collaborative filtering according to Eq.3. When making hybrid recommendation, in order to recommend short-term interest first, it can give short-term interest a larger weight and long-term interest a smaller weight. Third, calculate the user interests value according to Eq.4. Finally, select the top N items with the highest scores to recommend for users.
4.4. Contrast Algorithm

With the in-depth study of interest drift by scholars, more and more scholars have applied the forgetting curve to the field of recommendation. After experimental simulation, the function of the forgetting curve is shown in Eq.5. \(x\) represents the time interval, that is, the length of the time interval between the first operation and the last operation.

\[
f(x) = 1 - 0.658 \times x^{0.06685}
\]  \hspace{1cm} (5)

Time window is another classic method for solving interest drift. The basic idea is to divide the time period into multiple time windows. Different time windows correspond to different weights. The closer the time window near to the current time, the greater the weight of the time window.

In order to illustrate the effectiveness of integrating of user interests, LSIMF is compared with a traditional matrix factorization for collaborative filtering model(MFCF). In order to compare the effect of handle interest drift between the proposed algorithm and traditional algorithm, LSIMF is compared with the traditional forgetting curve and time window.

![Figure 4. Effect of N on Precision.](image1)

![Figure 5. Effect of N on Recall.](image2)

![Figure 6. Effect of N on F1-Score.](image3)

The Precision, Recall and F1-Score of LSIMF is higher than that of matrix factorization for collaborative filtering, forgetting curve and time window under different N values, especially when N is less than 20. When N>20, the F1-Score of LSIMF is only slightly higher than matrix factorization for collaborative filtering, which shows the limitation of integrating user interests. Because the number of user interests of users is limited, LSIMF has obvious advantages when the TOP N recommendation scale is small. Comparing with the forgetting curve and time window algorithm, the F1-Score of LSIMF is higher under any values of N, which shows LSIMF can capture user interest drift better.

5. Summary

The accuracy of interest set is very high, but the number of elements in set is limited because user has limited interests. In addition, the elements in set are extracted based on historical user behavior data, so new interests can’t be recommended for user. Matrix factorization for collaborative filtering can not only use implicit feedback data, but also provide users with new interests, and there have enough elements, however they have the disadvantage of low accuracy. Therefore, this paper proposes a hybrid recommendation algorithm that considers interest drift and matrix factorization for collaborative filtering. It can obtain the distribution of user interests to enrich user portraits. The extracted tags including interest set, user activity, item popularity can automatically tag users and provide more options for subsequent research on dynamic evolution or extended function for website.

References

[1] iiMedia new retail industry research center, 2019 China E-commerce Semi-Annual Development Report, R. China: iiMedia network. (2019).

[2] Wang Yang, Han Lixin, Personalized recommendation via network-based inference with time, J. Physica A: Statistical Mechanics and its Applications. 123917 (2020) 1-6.

[3] Li Taoying, Jin Linlin, Wu Zebin, et al. Combined recommendation algorithm based on improved similarity and forgetting curve, J. Information (Switzerland). 10 (2019) 130-131.
[4] Zhang Jianyun, Lu Xianling, A Multi-Trans Matrix Factorization Model with Improved Time Weight in Temporal Recommender Systems, J. IEEE Access. 8 (2020) 2408-2416.

[5] Zhang Weina, Zhang Xingming, Wang Haoxiang, et al. A deep variational matrix factorization method for recommendation on large scale sparse dataset, J. Neurocomputing. 334 (2019) 206-218.

[6] Feltoni Gurini Davide, Gasparetti Fabio, Micarelli Alessandro, et al. Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization, J. Future Generation Computer Systems. 78 (2018) 430-439.

[7] Chen Rui, Hua Qingyi, Wang Bo, et al. A Novel Social Recommendation Method Fusing User's Social Status and Homophily Based on Matrix Factorization Technique, J. IEEE Access. 7 (2019) 18783-18798.

[8] Zhao Wei, Wang Benyou, Ye Jianbo, et al. PLASTIC: Prioritize long and short-term information in top-n recommendation using adversarial training, J. IJCAI International Joint Conference on Artificial Intelligence. (2018) 3676-3682.

[9] Yu Jianjun, Zhu Tongyu, Combining long-term and short-term user interest for personalized hashtag recommendation, J. Frontiers of Computer Science. 9 (2015) 608-622.

[10] Hu Linmei, Li Chen, Shi Chuan, Graph neural news recommendation with long-term and short-term interest modeling, J. Information Processing and Management. 57(2020).

[11] Hu Y, Koren Y, Volinsky C, Collaborative Filtering for Implicit Feedback Datasets, C.// Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy. IEEE, 2008.