Third International Symposium on Intelligent Information Technology and Security Informatics

A Collaborative Filtering Algorithm based on Time Period Partition

Yuchuan Zhang, Yuzhao Liu
Institute of Chemical Defense of CPLA
Beijing, China
xiezuolunwen_zjg@163.com, liuyuzhao_zjg@163.com

***Abstract*—Today collaborative filtering is the most successful recommender system technology. However, in traditional collaborative filtering algorithms, users' interest is considered to be static. That means, in these algorithms, ratings produced at different times are weighted equally, and changes in user purchase interest are not taken into consideration. For this reason, the system may recommend unsatisfactory items when users' interest has changed. To solve this problem, the time factor has been brought into collaborative filtering. In new algorithms, we have divided users' rating history into several periods, analyzed users' interest distribution in these periods and quantize every user's interest. At the same time, we find user's recent interest by setting a time window. With these two technologies, we propose a collaborative algorithm time period partition named TPPCF. Experiments have shown that our new algorithm TPPCF substantially improves the precision of item-based collaborative filtering.**

Keywords- time factor; item-based; phase; time period partition; TPPCF

I. INTRODUCTION

Nowadays, because of the increasing information in the world, we have to spend more time searching and browsing information we need by accessing the Web, which is one of the most effective approaches. Therefore, the Search Engines like Google, Baidu, etc., become the most popular approaches helping people surf on the Internet. Though they can meet people's need to some extent, their generality make it hard to satisfy those query requests with different backgrounds, objectives at different periods. To solve these problems, the personalization technology has been put forward to serve specific users and meet their specific needs [6].

Recommender system is one of the most important forms applying in personalization technology, which combines many technologies such as search engine, data mining, machine learning and so on, and has been widely used in e-commerce system for its precise recommendation based on users' tastes and preferences. At present, various recommendation technologies have been used in many large E-commerce Websites like Amazon, CDNow, Netflix, eBay [7] and so on.

Over the years, various approaches for collaborative filtering have been developed [2,3,4]. The collaborative filtering, which has been one of the most effective recommendation technology in information, are categorized into two classes: Memory-based algorithms and model-based algorithms. It can filter useless information by comparing users' similarity and bring the users with new items without the form limitation. However, the traditional approaches was carried out on the presupposition that the users' interest is stable, which can not reflect the changes of users' interests and produce low efficiency and precisions. That is to say, ratings produced at different times are weighted equally. To solve this problem, the focus of this paper is to analyze the purchase interest history of each user, observe the interest distribution with the time, and then calculate its time weight, which will be introduced into recommendation.

The remainder of the paper is organized as follows. Section 2 briefly presents some of the research literature related to collaborative filtering. In section 3, we propose our novel improved algorithm considering time factor TPPCF. Section 4 presents our experimental work. In the final section, we make a conclusion.

II. RELATED WORK

The classic collaborative filtering is user-based algorithm, which treats a user as a vector in the item-space and offer top-N recommendation sets by searching the most similar user through comparing the similarity of users. However, with the tremendous growth in the amount of available information and the number of visitors to Web sites in recent years, the performance of these algorithms declines. To improve the scalability of collaborative filtering, item-based collaborative filtering proposed in [1], which treats an item as a vector in the user-space and offer the nearest neighbor items for the item that has not been accessed by the user by comparing the similarity of items. Contrary to user-based algorithms, item-based algorithms can further reduce scalability. So it has been the focus of research, and many improved approaches are on the base of it.

With the fast growth of e-commerce, many collaborative filtering applications have been fielded for a long time and many Websites has accumulated tens of millions user ratings, some of which are very old [3].

Since the value of these very old ratings is questionable, we should seek to develop an algorithm that will decay the influence of these [3,4,5]. Recently, there are some improved approaches to select data and trace changes in user purchase interest to alleviate the influence of old data. In [3], the author proposed an approach to predict precisely user future purchase interests by deploying time weights and setting parameter for each item cluster. If the user preference for the type of items is consistent, old ratings related to the type of items can help improve accuracy of predicting future preferences for the type of items. To solve this problem, in [5], the author proposed item similarity-based data weight into time weight algorithms. However, these algorithms was inclined to set same parameters for each item for all users. In fact, each user's interest has its own variable law which is different other's. So different user has different parameters for the same item. In this paper, our approach learn users' rating history to find appropriate personalized parameter for each item, which the parameter can reflect the item's interest category for each user.

978-0-7695-4020-7/10 $26.00 © 2010 IEEE
DOI 10.1109/IITSI.2010.161

III. IMPROVED ALGORITHMS:TPPCF

A. Item-based Collaborative Filtering Algorithms

In classic algorithms, the input data is expressed as a user-item M×N matrix R(M,N), where item numbers noted by M and user numbers noted by N, R_{ij} represents the j-th user's opinion on the i-th item. According to the type of item, different systems have different forms of user ratings. In some system, 1 represents that some user is interested in some item, while 0 represents not. In other systems, discrete values 0,1,2,3,4,5 represent the preferences of user for some item.

The classic item-based collaborative filtering algorithms have two phases:

Phase 1 Similarity Computation. There are three main approaches to compute the similarity between two items: Cosine similarity, Pearson correlation coefficient and Conditional Probability-Based Similarity.

Cosine similarity：

$$sim(Ia,Ib)=\cos(\vec{a},\vec{b})=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}=\frac{\sum_i^m p_{ia}\times p_{ib}}{\sqrt{\sum_i^m p_{ia}^2}\sqrt{\sum_i^m p_{ib}^2}} \quad (1)$$

Where Ia, Ib identifies the a-th item and b-th item in the system, P_{ia}, P_{ib} represents the i-th user opinion on the a-th item.

Pearson correlation coefficient:

$$sim(Ia,Ib)=\frac{\sum_i^m (p_{ia}-\overline{p}_i)\times(p_{ib}-\overline{p}_i)}{\sqrt{\sum_i^m (p_{ia}-\overline{p}_i)^2}\sqrt{\sum_i^m (p_{ib}-\overline{p}_i)^2}} \quad (2)$$

Where $\overline{p}_i$ is the average of the i-th user's ratings.

Conditional Probability-Based Similarity:

$$sim(Ia,Ib)=\frac{p(i\mid j)}{Freq(i)^c}=\frac{Freq(ij)}{Freq(j)\times Freq(i)^c} \quad (3)$$

Where Freq (i) is the number of users that have purchased the i-th item, c is a value between 0 and 1 called zoom factor, which weaken the impact of the item that has been calculated many times.

Phase 2 Preference prediction. Compute the prediction of the preference for a given object.

$$p_{ij}=\frac{\sum_{c=1}^k p_{ic}\cdot sim(I_j,I_c)\cdot f(t_{ic})}{\sum_{c=1}^k sim(I_j,I_c)\cdot f(t_{ic})} \quad (4)$$

Where I_j identifies the j-thitem, I_c represents the nearest neighbors of the j-th item, p_{ij} represents the i-th user's opinion on the j-th item.

B. Time Factor

Just like in time weight algorithms [2, 3], we also assume that the user purchase interest is sensitive to time and assign a greater level of importance to recent data in the phase of the recommendation process. The time weight algorithm is described as:

$$p_{ij}=\frac{\sum_{c=1}^k p_{ic}\cdot sim(I_j,I_c)\cdot f(t_{ic})}{\sum_{c=1}^k sim(I_j,I_c)\cdot f(t_{ic})} \quad (5)$$

Where t_{ic} represents the time the user's p_{ic} was produced, f(t_{ic}) is a weight function to the time t.

We assume f(t_{ic}) is a monotonic decreasing function. We defines it as:

$$f(t)=0.5+0.5\times e^{-\lambda t} \quad (6)$$

Where t is the time span from the most recent rating for one item that is belong to one interest to now, λ is a personalized parameter that is learned from each user's rating history.

C. λ Parameter

Since users has long interests and short interests and the time user's preference for a specific item lasts is decided by them, the value of old ratings is questionable In other words, the same old ratings have different influence on different users, and the discount rate which is a constant rate to alleviate the influence of the old data depends on the duration of user preferences for items. This means the discount rate varies with different users and different items. The longer the target user's preference for a specific item lasts, the lower the discount rate is.

In this paper, λ is a personalized parameter representing the discount rate. Learning λ becomes the key issue in our improved algorithms. In our method, to classify the items, we need to suppose that each item already has its own predefined category (e.g. ,Family, Heath, Mangement, and etc.). All items purchased by user 0 shown in Table1 are used as an example. Table 1 shows the periods in which the purchased items appeared for user 0.That is to say, if all users of the system have n interests, we classify all items into n categories. We define c （c=1,2,…,m） as the i-th interest in the specific system. Then we divide each user's ratings history into d equal time periods. The λ is defined as:

$$\lambda_c=\alpha\sum_{i=1}^d d(i)\cdot\frac{N_c}{n} \quad (7)$$

Where λ_c identifies the parameter of the c-th interest of the target user, $d(i)=\begin{cases}0, norating\\ 1, rating\end{cases}$ i = 1,2…d identifies whether the item belonging to the type of the c-th interest during the d-th time stage. $\frac{N_c}{n}$ represents the probability, Nc identifies the number of the c-th interest, n is the total number of the user's rating, and α is a constant value between 0 and 1.

Illustrated as above, we can describe the distributions of each interest of user to time.

It can quantify the influence of interest in user. The value of the λ reflects whether the interest is long interest or short interest. The bigger λ probably describes the shorter interest, while the smaller λ, the longer interest. To be more effective, we assume that d is one or several times the number of the users' interest category.

TABLE I. INTERESTS DISTRIBUTION FOR USER 0

month	CategoryA	CategoryB	CategoryC	CategoryD
1-2	★			★
2-3	★			★
3-4	★			★
4-5				★
5-6	★	★		★
6-7	★	★		★
7-8	★	★	★	
8-9		★	★	
9-10		★	★	
10-11	★	★	★	
11-12	★	★	★	

D. *Finding Use's Recent Interests*

Our approach starts with applying a general time window method. A time window three-months wide is used for clustering. For our first run, items from months 1 through 3 inclusive are classified into each category.The time window is then advanced one month, and the next three months worth of items are classified (months 2 through 4 inclusive) . This is repeated throughout the time period.

Table 2 like Table 1 shows the periods in which the purchased items appeared for user 0. We can see there are three categories A, B, C appearing in the most recent time period. And then we go back to the lengths of these three. Category B shows the largest continuous most-recent interest. Category A appears in 7 time periods while category B appears in 6 periods, so it has an great impact on predict. To show these influences , we choose two largest continuous most-recent interests and two appears most in time periods as recent interests.

TABLE II. INTEREST CHANGING OVER TIME FOR USER 0

month	CategoryA	CategoryB	CategoryC	CategoryD
1-3	★			★
2-4	★			★
3-5	★			★
4-6				★
5-7	★	★		★
6-8	★	★		★
7-9		★	★	
8-10		★	★	
9-11	★	★	★	
10-12	★	★	★	

E. *Time Period Partition Collaborative Filtering*

Algorithm: Time Period Partition Collaborative Filtering

Input: Item-User Matrix R(i,j), the target user ratings vector u_i

Output: Top-N recommendation set for user u_i

Steps:

1. Computer the similarity of every two items using Formular (1);
2. Select the k nearest neighbors of each item;
3. Calculate λ the of the target user using Formular (7);
4. Dectect the target user's recent interests with the method above.
5. Compute the ratings of items of the target user using Formular (6);
6. Select the top-N recommendation set of the target user.

IV. EXPERIMENT EVALUATION

A. *Data Sets*

We use data from the MovieLens recommender system. MovieLens has been the most widely used common datasets in collaborative filtering research projects. Movielens consisted of 1,000,209 ratings for 3900 movies by 6040 users. In the datasets, each user has rated at least 50 movies. At the same time, some statistics such as movie genres, ages, etc are included in the datasets.

The time users' ratings last in the datasets is about 2 year.

We cleaned the MovieLens data set to retain users who continued to participate in MovieLens lasting at least one year

Table 3 is the statistic o the cleaned data set

We set each time period has 30 days, and the time window has three time periods. so it has 90 days.

We utilize the protocol, All But One. In All But One, the newest rated items for each user are used for testing. The evaluation metric used in our experiments is the mean absolute error (MAE). It is computes the average absolute deviation of recommendations from their true user-specified values. The MAE can be computed as:

$$MAE = \frac{1}{m_a} \sum_{j=1}^{m_a} | p_{a,j} - r_{a,j} | \quad (8)$$

Where m_a identifies the number of the user's ratings, $p_{a,j}$ identifies the predicted rating for the j-th item, $r_{a,j}$ identifies the true rating for the j-th item.

TABLE III. DATA SET STATIC

Cleaned Data Set	
User	686
Movies	3528
Ratings	247577
Ratings/User	360

B. *Experimental Results*

In our experiment, we compare the new algorithm (TPPCF) to time weight collaborative filtering algorithms (TWCF) in [3] and traditional collaborative filtering algorithms(CF).We select five groups with 10, 20 30, 40, 50 neighbors, do five experiments with each algorithms and compute the average value of the five results. The results are demonstrated in Table IV and Figure.1. The table and figure tells us our new algorithms and TWCF are both able to boost the prediction precision.. Therefore, our algorithm is appropriate for the Website, which last long time.

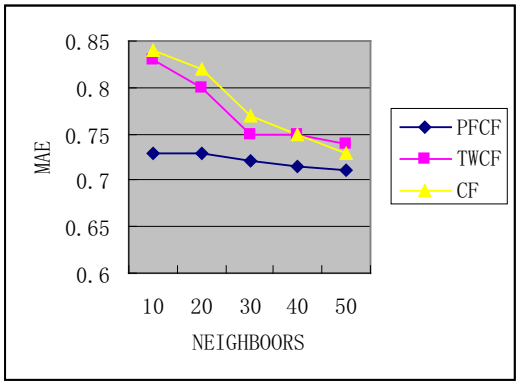

Figure.1. MAE using different algorithms on in All But One in three time stages.

TABLE IV. MAE USING DIFFERENT ALGORITHMS ON MOVIELENS IN ALL BUT ONE.

Neighbors / Algorithm	10	20	30	40	50
CF	0.73	0.73	0.72	0.71	0.71
TWCF	0.83	0.80	0.75	0.75	0.74
TPPCF	0.84	0.82	0.77	0.75	0.73

V. CONCLUSION

In this paper, we present a new collaborative filtering algorithms namely time period partition collaborative filtering. We have divided users' rating history into several time periods, and analyzed users' interest distribution in these time periods. In this new algorithm, we can quantify the preference of each user for items and get personalized parameter for each interest of each user. Experiments have shown our new algorithms can improve the predication precision more than classic algorithms, especially when the users' ratings last long time.

REFERENCES

[1] B. Sarwar, G..Karypis, J. Konstan, and J. Riedl. Item-Based Collaborative Filtering Recommendatin Algorithms. In International World Wide Web Conference, pp. 285-295,2001.

[2] L.T. Weng, Y. Xu, Y.F. Li, and R.C. Nayak. Improving Recommendation Novelty Based on Topic Taxonnmy.In the workshop of International Conferences on Web Intelligence and Intelligent Agent Technology,pp.115-119,2007.

[3] Y. Ding, and X. Li. Time Weight Collaborative Filtering. ACMCIKM, 2005.

[4] Y. Ding, X. Li., M.E. Orlowska. Recency-Based Collaborative Filtering.ADC.2006

[5] C.X. Xing, F.R. Gao., S.N. Zhan, and L.Z. Zhou. A Collaborative Filtering Recommendation Algorithms Incorporated with User Interest Change.Journal of Computer Research and Development.Vol.44.No.2. pp.296-301,2007.

[6] C. Zeng., C.X. Xing, and L.Z. Zhou. A Survey of Personalization Technology. Journal of Software. Vol.13(10),pp.1952-1962,2003.

[7] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-Item collaborative filtering. IEEE Internet Computing,Vol.7(1),pp. 76 – 80,2001.