Research and Improvement of Personalized Recommendation Algorithm Based on Collaborative Filtering

Lijuan Zheng†,Yaling Wang††, Jiangang Qi†††, Dan Liu†

†School of Computer and Information Engineering, Shijiazhuang Railway Institute, Shijiazhuang, 050043 China
††Two Six Zero Hospital, Shijiazhuang 050041, China
††† TianZheng Jian li company, He Bei water project geologic reconnaissance institute, Shijiazhuang, 050021 China

Summary

Collaborative filtering is one of the most frequently used techniques in personalized recommendation systems. But currently used user-based collaborative filtering recommendation algorithm and the collaborative filtering recommendation algorithm based on item rating prediction has disadvantage in similarity computation method. Basing on this disadvantage, the paper puts forward an improved collaborative filtering recommendation algorithm. We improve it in two aspects: First, we bring in a coefficient to coordinate the problem of inexact finding and falling recommendation quality which is caused by the fewer items when weighting the user similarity. Second, we collect the users' interest words implicitly when build the user interest model. At last, we develop an online network bookshop as an example, test and analyze the three algorithms. The testing results show that in most cases, the improved algorithm that we put forward can improve recommendation quality.

Key words:

Collaborative Filtering Personalized Recommendation Algorithm, Mean Absolute Error

1. Introduction

Wide application of the Internet creates basic foundation for the rapid development of E-commerce. E-commerce provides new chances for enterprises, but at the same time, it also puts forward new challenge. How can the clients acquire needed information conveniently, quickly and accurately in the world where information is a flood has become a vital problem oriented by many enterprises, the development and utilization of personalized recommendation technique is an important way to solve it.

In traditional collaborative filtering recommendation algorithms, if we adopt cosine similarity, the rating of the items that has not been rated is all the same (which is zero). Of course, this is not accurate [1]. The paper computes the semblance among items by adjusting cosine similarity measurement method. During the process of user interest model construction, we introduce the agent. Through test and analysis, we can draw a conclusion that the improved algorithm has solved the disadvantage that lies in traditional similarity computation method when the user rating data are extremely sparse, and it apparently improves recommendation quality.

2. Traditional Collaborative Filtering

Traditional collaborative filtering algorithms include user-based collaborative filtering algorithm and collaborative filtering algorithm based on item rating prediction.

2.1 User-Based Collaborative Filtering Algorithm

User-based collaborative filtering algorithm produces recommendation list for object user according to the view of other users. It is based on these assumptions: if the ratings of some items rated by some users are similar, the rating of other items rated by these users will also be similar. Collaborative filtering recommendation system uses statistical techniques to search the nearest neighbors of the object user and then basing on the item rating rated by the nearest neighbors to predict the item rating rated by the object user, and then produce corresponding recommendation list. The process of this algorithm can be divided into three steps: data description, find the nearest neighbors, produce recommendation data set [2].

(1)Data description: user rating data can be represented by a $m*n$ matrix A(m,n), m represents the number of users, n represents the number of items, item R locates at the ith line and jth column represents the rating of the item j rated by user i. User rating data matrix is shown as Table 1:

Table 1: User-item rating data matrix

User	***Item***				
	Item$_1$	…	Item$_k$	…	Item$_n$
User$_1$	$R_{1,1}$	…	$R_{1,k}$	…	$R_{1,n}$
…	…	…	…	…	…
User$_i$	$R_{i,1}$	…	$R_{i,k}$	…	$R_{i,n}$
…	…	…	…	…	…
User$_m$	$R_{m,1}$	…	$R_{m,k}$	…	$R_{m,n}$

Manuscript received July 5, 2007
Manuscript revised July 25 2007

(2)Find the nearest neighbors. In order to find the nearest neighbors of the object user, it must measure the similarity of the users, and select several users that have the highest similarity as the nearest neighbors of the object user. We adopt cosine similarity algorithm to measure the similarity between user i and j. User rating can be treated as a vector on an n-dimensional item space. If the user does not rate the items, we can assume the rating is zero. Assuming the rating of the n-dimensional item space rated by user i and user j is respectively vector $\vec{i}$ and $\vec{j}$, the similarity between user i and user j is sim (i, j):

$$sim(i, j) = \cos(\bar{i}, \bar{j}) = \frac{\bar{i} \cdot \bar{j}}{\|\bar{i}\| \cdot \|\bar{j}\|}$$

(3)Produce recommendation data set: we use NBS_u to represent the nearest neighbor set of user u, the predicted rating of item i rated by user u is $P_{u,i}$ which can be gained by the rating of nearest neighbors set NBS_u rated by user u, the computation method is as the following:

$$P_{u,i} = \overline{R_u} + \frac{\sum_{n \in NBS_u} sim(u,n) * (R_{n,i} - \overline{R_n})}{\sum_{n \in NBS_u} (|sim(u,n)|)}$$

The meaning of the signals in the formula is as the following:

$sim(u,n) - -$the semblance between user u and n;

$R_{n,i} - -$the rating of item i rated by user n;

$\overline{R}_u - -$the average rating of items rated by user u;

$\overline{R}_n - -$the average rating of items rated by user u;

According to the rating of items, we select N items that have the highest rating to compose recommendation set and recommend them to object user.

2.2 Collaborative Filtering Recommendation Algorithm Based on Item Rating Prediction

The recommendation process of this algorithm is the same as user-based collaborative filtering recommendation algorithm [3]. It can be divided into three phases: data description, find the nearest neighbors, produce recommendation data set. Compared to user-based collaborative filtering algorithm, the improvement lies in the following: when we calculate user similarity, first we calculate the similarity of the items, select several items that have the highest semblance as the nearest neighbors of the object item, then according to the rating of similar items rated by the user to predict the rating of object item, through adjusting each item in the item set has the rating rated by the two users, the rating of item p rated by user i is as the following:

$$R_{i,p} = \begin{cases} r_{i,p} \\ p_{i,p} \end{cases}$$

$r_{i,p}$ is the rating of item p rated by user i, $P_{i,p}$ is the predicted rating of unrated item p rated by user i.

Item similarity computation method of this algorithm is similar to user similarity computation method, first we need acquire ratings of all the users aiming at item i and item j, and then use similarity measurement method to calculate the similarity between item i and item j. We use the following formula to compute the rating $P_{i,p}$ of item p rated by user i [4]:

$$P_{i,p} = \frac{\sum_{n \in M_p} sim_{p,n} * R_{i,n}}{\sum_{n \in M_p} (|sim_{p,n}|)}$$

The meaning of the signals in the formula is as the following:

$M_p - -$the neighbor item set of item p;

$sim_{p,n} - -$the semblance between item p and n;

$R_{i,n} - -$the rating of item n rated by user i.

3. The Improved Algorithm

Through the analysis of the disadvantage of traditional similarity measurement method and the problem of user interest modeling. We improve the algorithm in the following aspects:

(1) Semblance computation: In most cases, the number of items that jointly rated by the two users is few, usually one or two items. Even the rating of these items rated by the two users has high similarity, according to common sense we can not judge the two users are similar; but the semblance of the two users is very high if we use traditional similarity measurement method. In order to solve this problem, we introduce a coefficient: the coefficient is large if there are many items that the two users jointly rate; on the contrary, the coefficient is small. We suppose that the coefficient is K, and $K = \frac{\bigcap(i,j)}{\bigcup(i,j)}$, $\bigcap(i, j)$ represents the number of items in the intersection set that rated both by user i and j, $\bigcup(i, j)$ represents the number of items in the union set that rated both by user i and j, the range of the coefficient is between 0 and 1.

After introducing coefficient the semblance becomes:

$$sim(a,u_i)=\cos(\overline{a},\overline{u}_i)=\frac{\overline{a}\cdot\overline{u}_i}{\|\overline{a}\|\cdot\|\overline{u}_i\|}*k$$

$$K=\frac{\cap(a,u_i)}{\cup(a,u_i)}$$

Thus we can guarantee the following facts: only on the condition that the users take part in majority rating and the rating items are almost the same can the user have the most possibility to become similar user. On the other hand, the users that take part in a few items rating, even though these rating are similar, in fact the users are not similar. But by using traditional similarity measurement method we can acquire large similarity, this is not accurate. After we multiply a proportion coefficient K, the final value of semblance becomes small, and so it loses the possibility of becoming the nearest neighbors.

(2)Construction of user interest model: User interest model directly influences the recommendation quality of the recommendation system [5]. During the construction of user interest model in personalized recommendation system agents are used, which include user tracking agent, feedback agent and recommendation filtering agent. Since the agent can learn the interest and hobby information of the user, can further study based on the behavior and feedback of the user and update user interest library immediately, so the recommendation results will more suit the requirement of the user.

We use the following methods when we construct user interest model:

When the user search information, the system use the search key words that the user fills out as the representation of user interest key words, store them in the user interest table, and then assign them weight value. When the user collect some information, we can infer that the user is interested in such information, we take out some words as the user interest key words and store them in the user interest table, assign them weight value, we use the number of times to represent the weight value of the key words.

The key words in the user interest table reflect the interest and requirement of the user, weight value reflects the degree of the preference, if the weight value is very large, it shows that the user is more interested in this information; on the other hand, the preference degree is small. When the number of the key words in the user interest table reach a certain value, we delete the key words which had low weight value, thus keep the capacity of the user key words at a fixed level, so the key words in the user interest table can approach the preference of the user more accurately.

4. The Description of the Improved Algorithm

Input: user-item rating table, object user a;
Output: recommendation set;
Method:
(1)Retrieve all the items in user-item rating table and save them in set I, I= $\{i_1, i_2, \cdots, i_n\}$;
(2) Retrieve all the users in user-item rating table and save them in set U, U= $\{u_1, u_2, \cdots, u_m\}$;
(3)The object user is a, compute the semblance between user a and all the other m-1 users:
For each user $u_i \in U$ and $u_i \neq a$
Step1: Compute the union set V, the items in which have been rated by user a and u_i;
Step2: Compute the item set M in set V which has not been rated by a, the number of items in M is k;
Step3: For each item, $p \in M$: compute the semblance between p and other k-1 items, and store them in array sim1 [k-1];
Select several items in array sim1 [k-1] in descendant order as the nearest neighbor items L of item p;
Compute the predicted rating of item p rated by user a:

$$P_{a,p}=\frac{\sum_{n\in L} sim_{p,n} * R_{a,n}}{\sum_{n\in L}\left(\left|sim_{p,n}\right|\right)}$$

End for
Repeat Step2~Step3, compute the predicted rating of items unrated in V by user u_i;
Compute the semblance between user a and u_i based on set V, and store them in array sim2:

$$sim(a,u_i)=\cos(\overline{a},\overline{u}_i)=\frac{\overline{a}\cdot\overline{u}_i}{\|\overline{a}\|\cdot\|\overline{u}_i\|}*k$$

$$K=\frac{\cap(a,u)}{\cup(a,u)}$$

End for
(4)Select several users in array sim2 in descendant order as the nearest neighbor NBS_u of object user a;
(5)For each item $p \in I$:
Predict the rating of all the n items rated by user a, and store them in array $p_{[n]}$:

$$P_{a,i}=\overline{R}_a+\frac{\sum_{u\in NBS_u} sim(a,u)*(R_{u,i}-\overline{R}_u)}{\sum_{u\in NBS_u}(|sim(a,u)|)}$$

End for
(6)Select all the items in array $p_{[n]}$ as the recommendation set;
(7)Select several interest key words with the highest weight value in user interest library;
(8)Select suited books depending on user interest key words; store them in user recommendation list.

5. Comparison Analysis between Traditional Collaborative Filtering Algorithm and Improved Algorithm

In order to effectively carry out the research and analysis of personalized recommendation system, we design an online book system. On the basis of this system, we respectively use user-based collaborative filtering recommendation algorithm, collaborative filtering recommendation algorithm based on item rating prediction and the improved collaborative filtering algorithm to realize personalized recommendation, the method adopted and analysis results are as the following.

5.1 Data Set

The data we test are downloaded from MovieLens (http://movielens.umn.edu), MovieLens is a research recommendation system based on Web. It provides the rating data of movies or books for free. We randomly select 500 rating data in the database, which include rating of 514 books rated by thirty users. The rating value is an integer between one and five. If the number is large, it shows that the user is more interested in the book.

In order to measure the sparsity of the data set, we introduce the concept of sparseness rank, it is defined as the percentage of unrated items in the user rating data matrix. Through calculation we get the sparsity rank of the tested data set is: 1-500/ (30*514) = 0.9676.

5.2 Measurement Criteria

The measurement method of evaluating the recommendation quality of recommendation system mainly includes statistical precision measurement method and decision supporting precision measurement method.

Statistical precision measurement method adopts MAE (Mean Absolute Error) to measure the recommendation quality [6]. MAE is a commonly used recommendation quality measurement method. So we use MAE as the measurement criteria.

MAE calculates the irrelevance between the recommendation value predicted by the system and the actual evaluation value rated by the user. We represent each pair of interest predicted rank as <p_i, q_i>, p_i is the system predicted value, q_i is the user evaluation value. Basing on the entire <p_i, q_i> pairs, MAE calculates the absolute error value $|p_i-q_i|$ and the sum of all the absolute error value, and then calculates their average value. If the MAE value is small, it indicates good recommendation quality.

The predicted user rating set can be represented as {p_1, p_2, …, p_N}, its corresponding actual user rating set can be represented as {q_1, q_2, …, q_N}, the MAE can be defined as the following [7]:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

5.3 Experiment Result

In order to test the effectiveness of the algorithm, we compare the improved algorithm, user-based collaborative filtering recommendation algorithm and the collaborative filtering recommendation algorithm based on item rating prediction. Let us examine the influence of various nearest neighbor set on predictive validity. We gradually increase the number of neighbors; the experiment result is shown in Table 2:

Table2: Influence of various size of nearest neighbor set on predictive validity

A	*MAE*		
	B	C	D
4	1.3624	1.3073	1.2256
8	1.2652	1.2595	1.2180
12	1.2670	1.2615	1.2150
16	1.2705	1.2480	1.2203
20	1.2617	1.2578	1.2192
24	1.2617	1.2572	1.2192

In Table2, A represents size of neighbor set, B represents user-based collaborative filtering algorithm, C represents collaborative filtering recommendation algorithm based on item rating prediction, D represents improved collaborative filtering algorithm.

As Fig.1 shown, the improved algorithm has smaller MAE value than traditional user-based collaborative filtering recommendation algorithm and collaborative filtering recommendation algorithm based on item rating prediction in most cases.

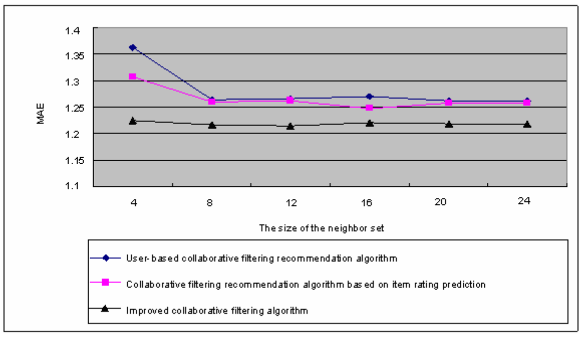

Fig. 1 The comparison sketch map of three collaborative filtering algorithms

5.4 The Analysis of Experiment Result

The improved algorithm put forward in this paper uses a coefficient when calculating user similarity. This coefficient can exclude the following possibility: when the user takes part in a few items rating, we can acquire large semblance by using traditional similarity measurement method, but in fact, the two users are not similar.

We can see through Fig.1, in most cases, the value of MAE of the improved algorithm we put forward is smaller, so the recommendation precision is high.

6. Conclusion

Basing on the analysis of user-based collaborative filtering algorithm and the collaborative filtering recommendation algorithm based on item rating prediction, we put forward an improved algorithm. Through the experiment analysis we can draw a conclusion that the improved algorithm has good recommendation effect.

References

[1] Chun Zeng, Chunxiao Xing, and Lizhu Zhou, “A Survey of Personalization Technology”,Journal of Software, vol.13, pp.1952 -1961,2002.

[2] Yan Zeng, Yonghao Mai, “Collaborative Filtering Recommen- dation Based on Content and Item Rating Prediction”, Computer Applications, vol.24, pp.111-113, 2004.

[3] Ailin Deng, Yangyong Zhu, and Baile Shi, “A Collaborative Filtering Recommendation Algorithm Based on Item Rating Prediction”, Journal of Software, vol.13, pp. 1621-1628, 2003.

[4] Li YU, “Research on personalized recommendations in E-business”, Computer Integrated Manufacturing Systems, vol.10, pp. 1306-1313, 2004.

[5] Xiuyan Gu, Linfeng Jiang, and Ziyi Zhang, “Study on User’s Browse Behavior to Measure the User’s Browse Interest”, Network and Communication, vol.15, pp.43-45, 2005.

[6] Huihong Zhou, Yijun Liu, Weiqing Zhang, and Junyuan Xie, “A Survey of Recommender System Applied in E-commerce”, Computer Application Research, vol.1, pp.8-12, 2004.

[7] Zhi Zhao, Bing Shi, “An Adaptive Algorithm for Personal Recommendation”, Journal of Changchun University, vol.15, pp.26-29, 2005.

Lijuan Zheng received the bachelor's degree and master's degree, from North China Electric Power University in 2000 and 2003, respectively. After working as an assistant(from 2003), an instructor (from 2005) in the School of Computer and information engineering of Shijiazhuang Railway Institute. Her research interest includes information security, applied cryptography, and E-commerce.