IBCF Improved Algorithm Based on the Tag

Qiuyue Xu, Shangzhi Zheng,Min Cai
Department of computer, Chaohu University, Hefei, china
E-mail:xqyyue@mail.ustc.edu.cn

Abstract - When the traditional Item - Based collaborative filtering recommendation is computing the nearest neighbor set of items, relying on scoring matrix similarity degree evaluation standard is onefold. In order to calculate program similarity better, the improved algorithm brings in project-Tag matrix, which can get comprehensive similarity combined with the score matrix. The experimental results show that IBCF-TAG recommendation algorithm based on the Tag can achieve better accuracy of recommendation than traditional recommendation algorithm based on the item.

Index Terms – IBCF, Tag, Collaborative Filtering

1. Introduction

The "personalized recommendation" has been put forward since the 1990s, and recommend system has been further applied in a social network, e-commerce, etc[1]. How to achieve more accurate recommend for different users, is the key to measure a recommendation system quality stand or fall[2]. The essence of collaborative filtering recommendation algorithm is "looking for someone who has the same interests and hobbies with it, and taking their recommendations as its choice"[3][4][5]. On that basis it divide into the collaborative filtering recommendation based on the User (User-Based Collaborative Filtering, UBCF) and the collaborative filtering recommendation based on the item (Item-Based Collaborative Filtering, IBCF)[6]. In some cases performance of Item - Based Collaborative Filtering recommendation is better, and recommendation result is more accurate. The practical application of the electronic commerce system also shows that if there is relationship between items, the relationship is generally stable. And the stable similarity between the items can be offline calculated, so Item - Based collaborative filtering recommendation can save running time.

How to find the internal similarity relation between the items is the key research of Item-Based collaborative filtering. A lot of improved algorithm do the deep discussion and research on it, such as the improved algorithm of based on item recursive relation [4] and combining the feature of controversy [5], and the improved algorithm of based on project keywords [7], which achieve good effect of recommended.

With the development of Web2.0 technology, the Tag is widely used in the fields of social network and e-commerce. Tag is that users do personalized Tag initiatively for the item content. To a great extent, it summarizes the subjective impression and the general description of resources for users. Not only can it be more real and objective to reflect preference choice to items of users, but also it can reflect the content characteristics and classification information of the item in a certain extent.

The starting point of the improved algorithm IBCF which is based on the Tag is: if the Tag set that user Tags item A and other Tag set Tags item B is very similar, we can think that there is similarity between these two items. When the new users select the item A, recommend system will recommend item B to new users, which is very similar with item A.

This paper will bring in Tag matrix on the base of principles of the traditional Item- Based collaborative filtering algorithm. Calculate comprehensive similarity degree combined with the original score matrix, as basis of producing recommended set. The last experimental results show that the improved algorithm can effectively improve the accuracy of recommendation algorithm.

2. Improved IBCF algorithm based on Tag

2.1 TF-IDF weightiness

TF - IDF (Term Frequency - Inverse Document Frequency) [8][9]weight is often used in information retrieval and text mining, which is used to evaluate important degree of a word to a certain document. Term Frequency (TF) refers to the proportion of the number of times of that the specific words appear in a document. The importance of words t_i to a specific document can be expressed as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (2.1)$$

Among them, the numerator $n_{i,j}$ expresses the appearance number of words t_i in the document, and denominator expresses the total number of all words appear in document.

Inverse Document Frequency(IDF) is used to measure the general importance of a word in a document set. The total number of document in the document set divided the Document number which contains the words, and quotient takes the logarithm. Such as 3.2 formula shows:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \qquad (2.2)$$

Among them, $|D|$ expresses amount of document library files. $|\{j : t_i \in d_j\}|$ expresses number of file which contains the words (namely file number which $n_{i,j} \neq 0$). The purpose of denominator that uses 1 +$\left|\{j : t_i \in d_j\}\right|$, is to prevent that dividend is zero if the word is not in document set.

Finally, the weight of importance can be expressed as:

$$\text{TF}\times\text{IDF}= tf_{i,j} \times idf_i \qquad (2.3)$$

If regard all Tags marking a item as document, we get a weight TFxIIF (Tag Frequency - Inverse Item Frequency). The weight describes the degrees of the importance of the Tag in the Tag set of item. Get user - item Tag matrix filled with TFxIIF which is similar to score matrix, Column vector can be used to calculate the similarity between the items.

In the algorithm design, we also regard Tag as an independent entity, and the relevant frequency can be used to calculate the similarity. Suppose users set is $U = \{u_1, u_2 ... u_m\}$, item set is $I = \{i_1, i_2 ... i_n\}$, Tag set is $T = \{t_1, t_2 ... t_l\}$,and the score matrix that already exists is $R = (r_{ij})_{m\times n}$, among which r_{ij} is the score that user u_i is to the item i_j . Algorithm is also divided into three steps: the user information acquisition, the production of neighbor and the production of recommended set.

2.2 the user information acquisition

Calculate Item-Tag matrix

Suppose:The Tag set of the item I_j is T_j^I .Amount of mark of the item is $C_j^I(A)$, Amount of that marks Tag t_k of the item. Then calculate frequency that Tag t_k appears in the item I_j is $f_j^I(k)$. Formula shows:

$$f_j^I(k) = \frac{C_j^I(k)}{C_j^U(A)} \qquad (2.4)$$

Suppose: $C^I(A) = |I|$ expresses the total number of items. $C^I(k) = \left|\{I_j \mid t_k \in T_j^I\}\right|$ expresses the number of Tags that contains Tag t_k . Then we define iif_k as inverse item frequency of Tag t_k . It can be calculated by formula (2.5).

$$iif_k = \log\frac{C^I(A)}{C^I(k)} \qquad (2.5)$$

Define TFxIDF frequency of Tag t_k in the item I_j as $ti(k, j)$ = $f_j^I(k)$. iif_k .Get matrix M^I , expressed by formula (2.6).

$$M^I = \begin{bmatrix} ti(1,1) & ... & ti(1,l) \\ ... & & ... \\ ti(n,1) & ... & ti(n,l) \end{bmatrix} \qquad (2.6)$$

The representation of the matrix is the content shown by table I:

TABLE I Item-Tag matrix

	Tag_1	…	Tag_k	..	Tag_l
$Item_1$	$ti_{1,1}$	…		..	
…		…		..	
$Item_i$	$ti_{i,1}$	…	$ti_{i,k}$	..	$ti_{i,l}$
…		…		..	
$Item_n$		…	$ti_{n,k}$	..	$ti_{n,l}$

Among it:

$$ti_{i,k} = \begin{cases} f_i^I(k)\cdot iif_k & itemI_i \quad marked \quad Tag_k \\ 0 & \text{otherwise} \end{cases}$$

While filling the matrix M^I , if the corresponding item element is zero, the user has not Taged this item. So it also faces the problem of data sparse solution.

2.3 Produce the nearest neighbor set

Item based on the Tag similarity also has a variety of calculation method, which is the same as the traditional collaborative filtering algorithm,such as the cosine similarity, modified cosine similarity, Pearson correlation coefficient and similarity based on the conditional probability, etc. In order to keep consistent with the improved algorithm and the experiment of previous chapter, we introduce the calculation method of similarity here which is used in Pearson correlation coefficient.

2.3.1 Item-Tag similarity degree

Here we use Pearson correlation coefficient to calculate Tag similarity of item I_x and item I_y which all based on the Tag weight matrix. Formula (2.7) shows:

$$sim_I^T(x,y) = \frac{\sum_{k\in T_{xy}} (ti(k,x) - \overline{T}_i(x)) \cdot (ti(k,y) - \overline{T}_i(y))}{\sqrt{\sum_{k\in T_{xy}} (ti(k,x) - \overline{T}_i(x))^2} \sqrt{\sum_{k\in T_{xy}} (ti(k,y) - \overline{T}_i(y))^2}} \qquad (2.7)$$

T_{xy} in the formula expresses two items use Tag whose frequency is not zero. TFxIDF frequency means of item I_x and item I_y are expressed respectively by $\overline{T}_i(x)$ and $\overline{T}_i(y)$.

2.3.2 Item-Score similarity degree

The similarity of score between items can be written directly by pearson correlation coefficient measurement.

Formula (2.8) shows:

$$sim_I^R(x,y)=\frac{\sum\limits_{c\in U_{xy}}(r_{cx}-\overline{R}_x)\cdot(r_{cy}-\overline{R}_y)}{\sqrt{\sum\limits_{c\in U_{xy}}(r_{cx}-\overline{R}_x)^2}\sqrt{\sum\limits_{c\in U_{xy}}(r_{cy}-\overline{R}_y)^2}} \quad (2.8)$$

$\overline{R}_x$ and $\overline{R}_y$ express respectively the mean of item I_x and item I_y that all users rate. $U_{x,y}$ expresses the user set which have rated the item I_x and item I_y.

2.3.3 Comprehensive similarity

Considering the influence of the Tag similarity and rating similarity to recommend set, they are weighted and get comprehensive similarity, using the regulatory factor β to adjust rating similarity and Tag similarity weight value, whose value is between 0 and 1. Recommended system under the different application background use score and Tag differently. For example general e-commerce sites use rating data more, such as dangdang, etc., and the requirement of the DouBan reading Tags is used popularly. β is used to weigh the proportion of the score matrix and Tag matrix according to different recommend system. If ratings data is sparse, and Tag weight matrix data is dense in the recommendation system, it can take the value of the small; If the score matrix data is dense, and Tag matrix data is sparse, then take big. So the Tag based on item comprehensive similarity calculation formula can be expressed as:

$$sim_I(x,y)=\beta\cdot sim_I^R(x,y)+(1-\beta)sim_I^T(x,y) \quad (2.9)$$

Geting method of the item Top - n neighbor set is similar with the traditional coordination filtering algorithm. It can sort according to the degree of correlation from big to small,and select the front of the K neighbor. It also can set a similarity threshold. Greater than the valve is worth considering in the Top - n set. Or integrate two, which the nearest neighbor is only greater than similarity threshold and order in the Top K.

2.4 Produce recommend prediction

Based on the 2.6 formula, we can write prediction score calculation methods of target users on the target items directly. Formula (2.10) shows:

$$P_{u,i}=\overline{R}_i+\frac{\sum_{j=N(u)}sim_I(i,j)\cdot(R_{u,i}-\overline{R}_j)}{\sum_{j=N(u)}\left|sim_I(i,j)\right|} \quad (2.10)$$

Among it, $sim_I(i,j)$ expresses similarity degree between the target item i and its nearest neighbors item j, and R_i and R_j express average score respectively that the user score the item i and item j.

2.5 Description of algorithm

The detailed steps description of the TAG - IBCF algorithm based on the Tag described is as follows:

Input: Item - Score matrix $R_{m\times n}$, the Item - Tag matrix $T_{n\times l}$, comprehensive similarity regulatory factor β, comprehensive similarity threshold η, the nearest neighbor set size K.

Output: user top - N recommend set.

Step1. Base on the Item - Score matrix $R_{m\times n}$, and calculate Item - Score similarity ItemSim – R.

Step2. Base on Item-Tag matrix $T_{n\times l}$, and calculate Item - Tag similarity ItemSim – T.

Step3. Calculate item comprehensive similarity ItemSim through the regulatory factor β.

Step4. Determine the nearest neighbors set of the target item based on ItemSim, and predict target user's rating through the nearest neighbor score.

Step5. Take the highest value N item from prediction score, which is the current user's Top - N recommended list.

3. The experiment and analysis

3.1 Data collection and evaluation standard

Experiment this chapter uses MovieLens data set[10], and the data set is divided into two pieces: 80% of the data is as a training set, and the remaining 20% of the data is as a test set. Score matrix and Tag matrix are taken in training set in random, to calculate similarity. And test set data is used for contrast forecast results. The experiment takes mean absolute error (MAE) as evaluation standard. The aim of the experiment is to verify influence of the improved TAG - IBCF algorithm parameters and neighbor set size on the MAE. Among them:

$$MAE_u=\frac{\sum_{i=1}^{T_u}\left|P_i-R_i\right|}{T_u}$$

3.2 The experimental results

3.2.1. Experiment of parameter β

In the formula (2.9), parameter β is used to balance the influence of the similarity calculation between score data and Tag data. If β = 1, the comprehensive similarity degenerate into a simple scoring similarity. If β = 0, the comprehensive similarity degenerate into a simple Tag similarity. In order to evaluate the influence of β on MAE better, take β = {0.1, 0.2, 0.3... 1}, and take the number of nearest neighbor for 10, testing MAE for each β. The experimental result is as shown in figure 1. It expresses that if β takes value range from 0.4 to 0.6, the accuracy of the algorithm is more correct.

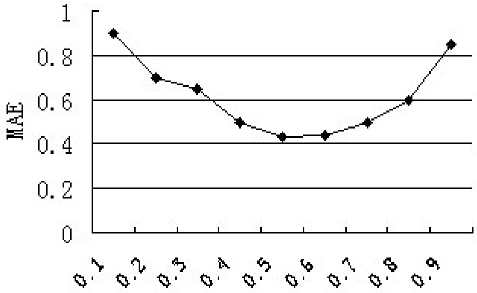

Fig. 1 Relationship between and MAE

3.2.2. Comparing the result of MAE

According to the experimental results, we take = 0.5 as well. With change of the number of nearest neighbors we observe the change of two algorithm's MAE. Here are still using Pearson method to calculate the similarity of the traditional algorithm and improved algorithm. The experimental results show that the improved algorithm TAG-IBCF has a smaller MAE value than traditional collaborative filtering algorithm, so the improved algorithm improves the accuracy of recommendation.

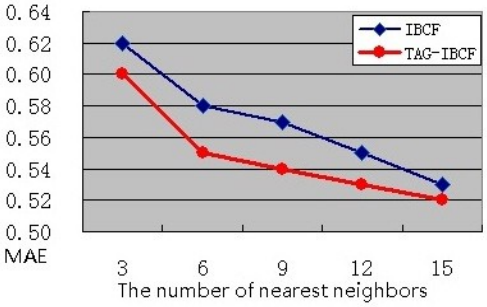

Fig. 2 Comparison of recommendation algorithm MAE

3.3Analysis of Experimental Results

From the figure 1, we also found that value of β that is from 0.4 to 0.6 can have a good result by using the comprehensive similarity to calculate the nearest neighbor after introducing Tag matrix, in the case of the nearest neighbor threshold for 10, comparing the influence on MAE of different β. Because of β as a regulatory factor, it reflects the influence of the balance of score matrix and Tag matrix on recommend result factually. Combined with the description of the 2.9 formula, it is known that it gets better recommend effect if value of β is big when scoring matrix data is more intensive and Tag matrix books more sparse value, and vice versa. The experiment also shows that rating data and Tag data from Movielens 10M data set are in the weight quite reflecting the true state of similarity, so taking β 0.5 will have good recommendation accuracy.

From the figure 2, despite the improved algorithm TAG-IBCF got smaller MAE, the experimental results also show that the algorithm improvement effect is more and more small with the increase of the number of neighbors, and two is almost to the agreement at last. This is due to the Tag brought into the additional information of the description for the item. It can be better mining the true inner correlation contact between items, so as to enhance the performance of recommended. Because the user's knowledge background and usage is different, with the users increasing, the item's Tag information cannot be unified in a clear semantic framework. Tag information itself error will give additional burden to collaborative filtering recommendation system, so improvement effect will be offset part.

4. Conclusion

This paper mainly analyzes that relying on scoring matrix similarity evaluation standard is more onefold, while the current the collaborative filtering recommendation algorithm based on item calculate the nearest neighbor set of item, and it could not really reflect the inherent connection between items. So it introduces the Tag matrix on the basis of the traditional collaborative filtering algorithm based on item, and calculates comprehensive similarity. This method can reflect internal similarity relation between items factually. Finally, the experimental results show that TAG-IBCF recommendation algorithm based on the Tag can achieve better accuracy of recommendation than traditional recommendation algorithm based on items.

5. Acknowledgment

This research is funded by: Provincial natural science research project of Anhui colleges and universities: The Research of Personalized Recommendation Service in Digital Library (Grant No. KJ2012Z267).

6. References

[1] Nicolaus Mote, The New School of Ontologies[J]. CSC I 585. Nov. 30,2004

[2] LIU Jianguo, ZHOU Tao, GUO Qiang,et al. Overview of the Evaluated Algorithms for the Personal Recommendation Systems. Complex Systems and Complex Science. Vol. 6 (2009), p.1–3. (In Chinese)

[3] Li Yucheng,Research on Collaborative Filtering Algorithm[D],Master dissertation,Shanghai,Fudan University,2005

[4] Zhang Liang, Research on Collaborative Filtering Algorithm In Recommendation System[D],Dotoral dissertation,Beijing, Beijing University of Posts and Telecommunications,2009

[5] Guo Hongyan, Collaborative Filtering Algorithm of Recommendation System and the Application Research[D], Dotoral dissertation,Dalian,Dalian University of Technology,2008

[6] Sarwar B M, Karypis, Konstan J A, and Ried J. Item-based collaborative filtering recommendation algorithm[J], Proceedings of the Tenth International World Wide Web Conference, p285-295, 2001.

[7] Li Xuesheng, Research on Collaborative Filtering Algorithm Oriented sparse matrix[D], Master dissertation,Hefei, University of Science and Technology of China, 2011

[8] WU Chun-Xu, LI Jia-Jun, SHI Hui.,A Collaborative Filtering Recommender Algorithm Based on Folksonomy,Computer Systems & Applications,Vol. 19 (2010), p337–347. (In Chinese)

[9] http://zh.wikipedia.org/zh-cn/TF-IDF[EB/OL] [2011-11-30]

[10] MovieLens Dataset[EB/OL], http://www.cs.umn.edu/Research/GroupLens/index.html access on April 15, 2011.