2012 IEEE fifth International Conference on Advanced Computational Intelligence(ICACI)
October 18-20, 2012 Nanjing, Jiangsu, China

Collaborative Filtering Recommendation Algorithm based on Semantic Similarity of Item

Bai Juan

***Abstract*—The accuracy and quality is the best evaluation of recommend system . This paper proposes a collaborative filtering remmendation algorithms based on computing the sematic similarity of items in order to improve the accuracy of items' similarity.The experimental results shows that the optimized algorithm can give a better prediction ,by way of increasing accuracy and reducing cold-start problem of item .**

I. INTRODUCTION

COLLABOTRTIVE filtering is the most popular personalized recommendation technology, and it predicts recommended list of new target users based on the existing user perspective. Its basis is: users are classified based on the users' interests; if users give similar score for some items, it is similar to other items. This similarity can be measured from two aspects: the one is the user 's similarity, called collaborative filtering based on user（User-based）[1], the other is the item similarity, called collaborative filtering based on item（item-based）[2]. Therefore, the calculation of similarity is the key of collaborative filtering algorithm.

At present, there are three kinds of methods commonly used for calculating similarity: 1) the cosine similarity 2) adjusted cosine similarity 3) the Pearson correlation coefficient.Based on these three methods, considering the effects of several factors, the quality of the final recommendation can be effectively improved. The paper[3] classifies the items using clustering algorithm in data mining firstly, and then finds the nearest neighbor in the item space through comparing the similarity. Although you can effectively reduce the original data scale by clustering algorithm, the impact of similarity on the final recommendation is still simply considered in the calculation of similarity. So recommendation elements are relatively simple. The paper[4] adds item attribute similarity to item similarity computation, which is an influential factor on the final similarity computing. This method considers two kinds of elements including the attribute similarity and the score similarity, and to a certain extent solves the cold start problem. But the method is only based on the measurement standard of having attributes or not when considering the item's attribute information. It neglects the correlation of existing attributes. In the calculation of similarity of item, the paper[5] considers the program category information, comprehensively utilizes this information and score information to predict the score value, and has received a certain effect. But in the calculation of category information similarity, program semantic similarity is calculated only from the classification tree hierarchy. It however, ignores the other multiple relationships among semantic concepts. This thereby leads to that the notion of semantic cannot be completely reflected, and affects the calculation of categories information similarity between items. In the calculation of user similarity, this paper[6] takes the similarity into account, while adding the user time function. This has a positive impact on the final recommendation precision. The method provides a better idea and breakthrough point for the following research.

Based on the above analysis, this paper introduces a new semantic similarity calculation method to the item attribute similarity calculation. The cold start problem is effectively avoided by calculating the similarity of semantic concept. At the same time, the similarity accuracy is improved, which has a positive effect on the final recommendation.

Manuscript received September 9, 2012.

J Bai is with Department of Information Engineering, North China University of Water Conservancy and Electronic Power, Zhengzhou, Henan , 450011 (phone:0371-69127299;e-mail:baijuan@ncwu.edu.cn)

II. COLLABORTIVE FILTERING RECOMMENDATION ALGORITHM BASED ON ITEM SEMANTIC SIMILARITY

A. Item Semantic Similarity

In the actual e-commerce, all items provided generally have description related to classification. According to the different categories of items, all items are classified in a tree structure[5], as shown in Figure 1. But in the tree structure, we tend to express the relationship between the upper and lower nodes, while ignoring the correlation among nodes on the same layer. For example, in spite of belonging to the same layer, fantasy and science fiction have higher similarity than fantasy and children. Therefore when drawing lines between the similar attributes of two parties, figure 1 becomes a figure 2. The semantic similarity between any two items can be measured by the semantic distance. And in general, the smaller the distance between two items, the higher their semantic similarity, and vice versa[7].

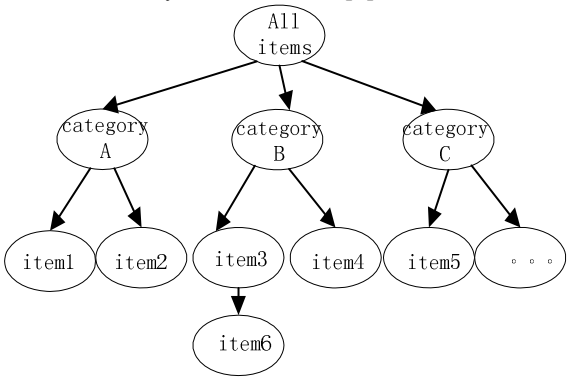

Fig. 1. The classified-tree of item

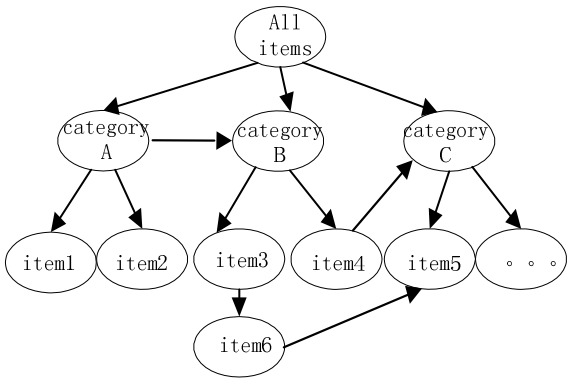

Fig. 2. The classified-tree introducing multiple relation

multiple paths between any two items, and each path includes N edges. We first introduce the concept of a side weights (weight), and the value of this weight and the representative relationship type are related to the hierarchy depth of edges in the network:

$$weight(edge_i) = type(edge_i) \times \frac{1}{2^{H(edge_i)}}$$

The $edge_i$ is the i edge a certain path contains, and type ($edge_i$) is the edge relation type (here, we value it 1, indicating the relationship among items is not equal, subclass or partial), and H ($edge_i$) is the hierarchy depth of edges in the network. We will take the hierarchy depth of a shallower node in the network between the two nodes connected by edges as the hierarchy depth of edges in the network[7].

Definition 1: the semantic distance of any two items is the sum of weights of all edge on a certain path connecting them.

$$dis\tan ce(item_1, item_2) = \sum_{i=1}^{n} weight(edge_i)$$

As shown in Figure 2, there are multiple paths to arrive at a node, so we take the minimum value of semantic distance among all paths. And semantic similarity is calculated according to semantic distance between two items:

$$sim_d(item_1, item_2) = 1 - \sigma\sqrt{\frac{1}{2} \times dis\tan ce(item_1, item_2)}$$

In this formula, σ >1, is an adjustable parameter, at the following realization of the algorithm we take 2 as the calculation of threshold.

B. The Rating Similarity Of Items

Calculating methods of rating similarity of the items mainly have two kinds: cosine similarity and correlation similarity. Cosine similarity calculation method is relatively simple, and it is faster to achieve. But this method is the calculation based on the whole score vector, and in the whole score vector space, zero value is filled for the item which has not been evaluated. This will objectively cause a certain error, thereby leading to inaccuracy in search of the nearest neighbor. The correlation similarity method is based on the intersection operation of two item score vector. Considering the item's average score relatively improves the accuracy in search of the nearest neighbor. In consideration of the above factors, we use correlation similarity to compute the item score similarity in the late stage of the experiment. If the user set of common score of $item_1$ and $item_2$ is expressed with U, the formula is as follows:

$$sim_p(item_1, item_2) = \frac{\sum_{u \in U}(R_{u,item_1} - \bar{R}_{item_1})(R_{u,item_2} - \bar{R}_{item_2})}{\sqrt{\sum_{u \in U}(R_{u,item_1} - \bar{R}_{item_1})^2}\sqrt{\sum_{u \in U}(R_{u,item_2} - \bar{R}_{item_2})^2}}$$

C. Improved Item Similarity

We carry out weighted summation between item's semantic similarity and item's score similarity, and the formula is as follows[4]:

$$sim(item_1, item_2) = \lambda \times sim_d(item_1, item_2) + (1-\lambda) sim_p(item_1, item_2)$$

$$\lambda = \frac{\left|U_{item1item2}\right|}{\left|U_{item1}\right| + \left|U_{item2}\right|}$$

Included among these, ,and it is called dynamic balance factor, representing the weight of users who have score of both item1 and item2 against the sum of users who have score either for tem1 or item2. When more users have score for both two items, its ratio is bigger, indicating that the score has contributed more to item similarity calculation; conversely, the semantic has contributed more to item similarity calculation.

D. Prediction Score

Take several items with maximum similarity as neighbor set of target $item_i$ $S=\{item_1, item_2, \ldots\ldots item_n\}$. Among them, $item_i \notin S$ and the items in set S are arranged from high to low order in accordance with the $item_i$ similarity, which may predict user score of $item_i$ that has not been scored according to a similar neighbor. The formula is as follows:

$$p_{u,item_1} = \bar{R}_{itemi} + \frac{\sum_{item \in S} sim(item_i, item) \times (R_{u,item} - \bar{R}_{item})}{\sum_{item \in S} \left|sim(itemi, item)\right|}$$

III. Experiment and Analysis

A. Experimental Environment

In the experiments, we use the data set downloaded from MovieLens site as experimental samples. The data set includes 100,000 records of evaluation of 943 users about 1682 films. Among them, the score range of users is 1-5, and a larger value indicates that the user has higher preference for the film. In this experiment, we select nearly 30,000 records of 400 users from the data set, and divide them into training set and test set in accordance with the ratio of 80% and 20%. The training set is randomly selected, and the value of MAE is calculated by predicting the user's score for evaluated films. Corresponding parameters are adjusted according to the MAE value, and finally the final performance of algorithm is inspected in test set.

B. Evaluation Standards

The metrical standards of recommendation quality of recommendation system have two kinds. One is a metrical

method of decision support precision, and the other is a metrical method of statistical accuracy. The latter mainly refers to the metrical method of average absolute error of MAE (Mean Absolute Error), which is easy to understand and easy to calculate, and can directly measure the recommendation quality. Therefore, in the experiment we use MAE as a metrical standard, and measure prediction accuracy by calculating the deviation between predicted user score and actual user score. The smaller the MAE value, the higher the quality of recommendation.

Let us assume that the user score set is $\{p_1, p_2, \ldots, p_n\}$ after prediction, and the corresponding actual score set is $\{t_1, t_2, \ldots, t_n\}$, then mean absolute error MAE is:

$$MAE = \frac{\sum_i^N |p_i - t_i|}{N}$$

C. The Experimental Results And Analysis

In order to test validity and accuracy of collaborative filtering algorithm (ISCF) based on item semantic similarity, we use MAE index to compare it with the existing ICF, collaborative filtering algorithm (PUCF) of improved similarity metrical method in reference[4], and collaborative filtering algorithm integrating item category information (ILCF) in reference[5]. In addition to the improvement of algorithm, the number of the nearest neighbor plays a big part in recommendation performance. In the experiment, we increase the nearest neighbor number from 10 to 50, with an interval of 10, and experimental results are shown in table1:

TABLE I
THE EFFECTS OF DIFFERENT RECOMMENDATION ALGORITHMS ON THE ACCURACY OF RECOMMENDATION

The nearest neighbor count	MAE			
	ISCF	ILCF	PUCF	ICF
5	0.7889	0.8122	0.8273	0.8175
10	0.7726	0.7814	0.7858	0.8021
15	0.7438	0.7630	0.7663	0.7768
20	0.7382	0.7556	0.7585	0.7677
25	0.7220	0.7316	0.7462	0.7563
30	0.7055	0.7268	0.7353	0.7623
35	0.6911	0.7118	0.7335	0.7562
40	0.6857	0.7081	0.7246	0.7501
45	0.6622	0.6864	0.7118	0.7342
50	0.6406	0.6653	0.7009	0.7266

The MAE value comparisons between improved algorithm and other algorithms are shown in figure 1:

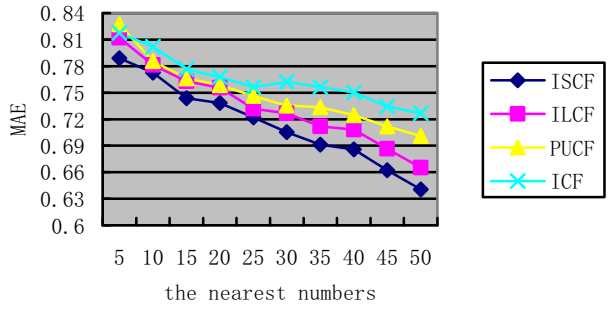

Fig. 3. The MAE curves of ISCF algorithm and other algorithms

Figure 3 shows that, the recommendation accuracy of collaborative filtering algorithm based on the semantic similarity is better than the collaborative filtering algorithm of improved similarity metrical method proposed in reference[4], the collaborative filtering algorithm integrating item category information proposed in reference[5], and the classical collaborative filtering recommendation algorithm based on item. Because the algorithm proposed in this paper has not only considered the item score similarity but also considered the item semantic similarity when calculating item similarity. This has positive effects on the calculation accuracy improvement. After this work, 10 new items are selected, computing the previous 20 neighbors which have the highest similarity with new items in item classification tree. And then the item user score is predicted by the score records of these neighbors. Experimental results show that, this method can get the score rate more than 50%. The reason is that the item semantic similarity in the algorithm is introduced as a dynamic calculation factor. When the users have more score for the item, the item score similarity accounts for a larger weight, but when the users has less score for the item, the item semantic similarity accounts for a larger weight. This can effectively avoid the problem of the cold start of new items by this method.

IV. CONCLUSION

This paper presents a method to improve recommendation quality by calculating item semantic similarity. Experimental results show that this method has a certain application value for increasing the recommendation accuracy and effectively improving the problem of the cold start of new items.

REFERENCES

[1] P Resnick, H Varian."Recommender Systems" in *Communications of the ACM*,1997,40(3):56-58.

[2] B Marko, S.FAB Yoav"Content-based Collaborative Recommendation" in *Communications of the ACM*,1997,40(3):66-72.

[3] X.Xiong, W Aref "G.R-tree with Update Memos Porc" in *the 22nd International Conference on Data Engineering*. Atlanta, Georgia, USA 2006 .

[4] Y.P Wu, J.G Wu. "The Collaborative Filtering Recommendation Algorithm of Improved Similarity Metrical Method". *Computer applications and software*. 10,pp.7-8. 2011

[5] W Shao, F Yuan, Y Zhang. "The Collaborative Filtering Recommendation Algorithm Integrating Item Category Information" *Mathematics in Practice and Theory*, 3,pp.108-112. 2010

[6] Ffiltering Recommendation Algorithm". *Computer applications and software.* 8,pp.72-75. 2011

[7] Y Shu, F Yang. "The Web Service Matching Research Based on Semantic Similarity". *Computer applications and software.* 8,pp.177-180. 2011