Item-based Collaborative Filtering Recommendation Algorithm Combining Item Category with Interestingness Measure

Suyun Wei, Ning Ye, Shuo Zhang , Xia Huang, Jian Zhu
College of Information Science and Technology
Nanjing Forestry University
Nanjing, 210037, China
{weisuyun, yening, zhangshuo, huangxia,zhujian}@njfu.edu.cn

***Abstract*—In order to overcome the limitations of data sparsity and inaccurate similarity in personalized recommendation systems, a new collaborative filtering recommendation algorithm by using items categories similarity and interestingness measure is proposed. In this algorithm, first the items categories similarity matrix is constructed by calculating the item-item category distance, and then analyzes the correlation degree of different items by using interestingness measure, last an improved collaborative filtering algorithm is proposed by combining the information of items categories with item-item interestingness and utilizing improved conditional probability method as the standard item-item similarity measure. Experimental results show this algorithm can effectively alleviate the dataset sparsity problem and achieve better prediction accuracy compared to other well-performing collaborative filtering algorithms.**

Keywords- recommendation systems; collaborative filtering; item similarity; item category; interesingnesst measure

I. INTRODUCTION

Collaborative filtering (CF) is one of the most significant techniques applied to recommender system, which is the pivotal method to enhance Electronic Commerce Systems performance. Its basic idea is to provide goods and score prediction to the user based on the ideas of other users of high similarity in preference between themselves and the target user.

Traditional collaborative filtering recommender system has exposed some of their bottlenecks such as data sparsity, cold-start and scalability widely and so on. In order to produce accuracy prediction and ensure the request of the real time in recommendation system, large quantities of scholars came up with new ideas to make up the weakness. Sarwar [1] employed the singular value decomposition(SVD) to tackle $m \times n$ rectangular diagonal rating matrix with decomposition and then gain a Reconstruction Matrix closest to itself with the rank of $k(k<min(m,n))$, so it can lower the data sparsity and boost the scalability to conduct collaborative filtering recommendation based on the lower order approximate matrix. Goldberg [2] advanced collaborative filtering algorithm by applying Principal Component Analysis (PCA) to the joke recommending system developed by University of California, Berkeley earliest. Horting [3] is a graph-based technique in which edges indicate the similarity between two users. Predictions are produced by traversing the graph to nearby nodes and combining the opinions of the nearby users. Papagelis [4] proposed to create a social network derived from the rating activities and trust towards items of users, so that the items without co-rated scores can engender transitive connection of user similarity. Clustering technique [5] improved collaborative filter expansibility by reducing the scope of searching the nearest neighborhood.

In the collaborative filtering algorithm, the accuracy of the similarity measure between items is vital to the quality of the recommender systems. Traditional similarity measure is adverse in the situation where the matrix data is extreme sparse, thereby recommender system predicts in poor precision. Hence, how to calculate the similarity effectively is the key to improve the recommender system. To address the issue, a new algorithm based on item category and interestingness for collaborative filtering (CICF) is proposed, putting forward the novel calculation methodology of item-item similarity based on conditional probability measure. This method builds the item category similarity matrix to enhance the similarity rate between items. Item interestingness is adopted to calculate the degree of item-item correlation and adjust the item similarity measure formula to guarantee the users becoming into neighbors only if they co-rated enough items and the similarity of their rate scores is high and amend the traditional shortage in calculating similarity of users. The experimental result shows an improvement in accuracy in contrast with classic collaborative filtering algorithm.

II. CONVENTIONAL ITEM-BASED COLLABORATIVE FILTERING RECOMMENDER ALGORITHM

Item-based collaborative filtering algorithm [6] held the viewpoint that users usually prefer to purchase the items similar or relevant to the things they have bought based on, so prediction rating on the target item was given based on the rating of the item in the nearest neighbor set by the user. Due to the steady similarities between items, it is quicker to compute off-line than on-line by shorten the time of computation.

Item-based collaborative filtering algorithm is processed in item-user rating matrix. User-item matrix usually is described as a $m \times n$ ratings matrix R_{mn}, where row represents m users and column represents n items. The element of matrix r_{ij} means the score rated to the user i on the item j, which commonly is acquired with the rate of user' interest.

One critical step in item-based collaborative filtering is to compute the similarity between items and then to select the

978-0-7695-4719-0/12 $26.00 © 2012 IEEE
DOI 10.1109/CSSS.2012.507

most similar items. There are a number of different ways to compute the similarity between items.

Cosine-based similarity: In this case, two items are thought of as two vectors in the *m* dimensional user-space. The similarity between them is measure by computing the cosine of the angle between these two vectors. Formally, in the $m \times n$ ratings matrix, similarity between items *u* and *v*, denoted by *sim*(*u*, *v*) is given by

$$sim(u,v) = \cos(\vec{u},\vec{v}) = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\times\|\vec{v}\|} = \frac{\sum_{i=1}^{m} r_{iu}\times r_{iv}}{\sqrt{\sum_{i=1}^{m} r_{iu}^2}\sqrt{\sum_{i=1}^{m} r_{iv}^2}} \quad (1)$$

where "·" denotes the dot-product of the vectors.

Correlation-based similarity: In this case, similarity between two items *u* and *v* is measured by computing the *Pearson-r* correlation $corr_{u,v}$. To make the correlation computation accurate we must first isolate the co-rated cases(i.e., cases where the users rated both *u* and *v*). Let the set of users who both rated *u* and *v* are denoted by U_{uv} then the correlation similarity is given by

$$sim(u,v) = \frac{\sum_{i\in Uuv}(r_{iu}-\bar{r}_u)(r_{iv}-\bar{r}_v)}{\sqrt{\sum_{i\in Uuv}(r_{iu}-\bar{r}_u)^2}\sqrt{\sum_{i\in Uuv}(r_{iv}-\bar{r}_v)^2}} \quad (2)$$

Here $\bar{r}_u$ is the average rating of the *u*-th item.

Conditional probability-based similarity: An alternate way of computing the similarity between each pair of items *v* and *u* is to use a measure that is based on the conditional probability of purchasing one of the items given that the other items has already been purchased. In particular, the conditional probability of purchasing *u* given that *v* is purchased *P*(*v*|*u*) is nothing more than the number of customers that purchase both items *v* and *u* divided by the total number of customers that purchased *u*, *i.e.*,

$$sim(u,v) = P(v \mid u) = \frac{Freq(uv)}{Freq(u)} \quad (3)$$

where *Freq*(*X*) is the number of customers that have purchased the items in the set *X*. Note that in general *P*(*u*|*v*) ≠*P*(*v*|*u*), *i.e.*, using this as a measure of similarity leads to asymmetric relations.

The most important step in a collaborative filtering system is to generate the output interface in terms of prediction. Once we isolate the set of most similar items base on the similarity measures, the next step is to look into the target users ratings and use a technique to obtain predictions. Here the prediction on an item *i* for a user *u*, denoted by P_{ui} is give by

$$P_{ui} = \bar{r}_i + \frac{\sum_{j\in Inei} sim(i,j)(r_{uj}-\bar{r}_j)}{\sum_{j\in Inei} sim(i,j)} \quad (4)$$

Where I_{nei} is the *top-N* nearest neighbors set of target user *u*.

III. COLLABORATIVE FILTERING ALGORITHM BASED ON ITEM CATEGORY AND INTEREST MEASURE

This paper proposed a collaborative filtering algorithm based on item category and interest measure. Considering the similarity between items within the same category is higher, the distance between items categories is computed through the items categories tree, and then items categories similarity matrix is created. With respect to the similarity deviation resulting from the situation fewer items were co-rated, item interest measure is introduced to amend the item-item similarity. Conventional item-item similarity calculation method based condition probability is ameliorated by adding item category information and the factor of item interest and put forward a new similarity calculation formula furthermore.

A. Item category similarity matrix

In the actual E-commerce websites, all the goods have been consigned to a number of different items categories, and each of the categories had some small classes, thus forming a tree standing on its head, as shown in Fig. 1.

According to items categories $C=\{c_1,c_2,...,c_k\}$, all the goods $I=\{i_1,i_2,...,i_n\}$ are described as a handstand tree, called items categories tree. The distance of categories between two items is measure by computing the items categories tree. Formally, in the categories tree, distance between items *u* and *v*, denoted by *d*(*u*, *v*) is given by

$$d(u,v) = \begin{cases} 0 & u = v \\ l(u,v)/H & u \neq v \end{cases} \quad (5)$$

Here *l*(*u*,*v*) is the longest path of reaching the first common ancestor of two item nodes *u* and *v* in the categories tree, and *H* is the height of the categories tree.

According to the formula (5), the distance of categories between two items *u* and *v* meet the following properties:

- $0 \le d(u,v) \le 1$
- *d*(*u*,*v*)=0 means two item nodes *u* and *v* belong to the same node, and the distance of categories between two items is minimum.
- *d*(*u*,*v*)=1 means the common ancestor of two item nodes *u* and *v* is the root, and the distance of categories between two items is maximum.
- *d*(*u*,*v*)= *d*(*v*,*u*)(i.e. symmetry).
- *d*(*u*,*v*)>*d*(*i*,*j*) means the distance of categories between two items *u* and *v* is greater than the distance between two items *i* and *j*.

In order to take the category attribute into consideration when calculating item-item similarity ,according to the items categories tree, the distance of categories between items is computed and the category similarity matrix S_{nn} including *n* items is established, whose element values s_{uv}=1- *d*(*u*,*v*) are symmetric around the leading diagonal, i.e. $s_{uv}=s_{vu}$.

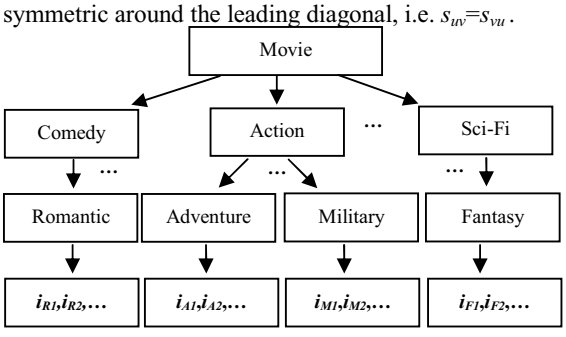

Figure 1. Part of the film categories tree .

B. Item interestingness

Piatetsky-Shapiro proposed three rule interestingness (RI) measures to objectively evaluate the values of patterns [7]. The measures effectively quantify the correlation between items. The measure formula is given by

$$RI(u,v) = P(u,v) - P(u) * P(v) = \frac{Freq(uv)}{m} - \frac{Freq(u) \times Freq(v)}{m^2} \quad (6)$$

Where, in the user-item rating matrix, $P(u,v)$ represents the probability of customers that purchase both items v and u, $P(u)$ represents the probability of customer that purchase u . If $RI(u,v)$ >0 then it means items u and v are in positive correlation, else if $RI(u,v)$ =0 then it means items u and v are independent to each other while if $RI(u,v)$ <0 then it means items u and v are negative correlation.

Theorem 1 if $P(u,v)$=0, $P(u)$=$P(v)$=0.5, then $RI(u,v)$=-0.25, and items u and v reach the maximum in negative correlation.

Theorem 2 if $P(u,v)$= $P(u)$=$P(v)$=0.5, then $RI(u,v)$=0.25, and items u and v reach the maximum in positive correlation.

C. adjusted conditional probability-based similarity

One of the limitations of using conditional probabilities as a measure of similarity is that each item v, will tend to have high conditional probabilities to items that are being purchased frequently. That is, quite often $P(v|u)$ is high, as a result of the fact that v occurs very frequently and not because v and u tend to occur together. This problem has been recognized earlier by researchers in information retrieval as well as recommendation systems [8]. One way of correcting this problem is to divide $P(v|u)$ with a quantity that depends on the occurrence frequency of item v, and use the following formula to compute the similarity between two items.

$$sim(u,v) = \frac{Freq(uv)}{Freq(u) + (Freq(v))^{\alpha}} \quad (7)$$

Where α is a parameter that takes a value between 0 and 1. Note that when α= 0, equation 10 becomes identical to $P(v|u)$, whereas if α= 1, it becomes similar to the formulation in which $P(v|u)$ is divided by $P(v)$.

This paper has further improved conditional probability-based similarity by combining the information of items categories with item-item interestingness, and then extended the similarity measure of equation (8) in the following way.

$$sim(u,v) = \frac{Freq(uv)}{Freq(u) + (Freq(v))^{\alpha} + \sum_{i=1}^{m} |r_{iu} - r_{iv}|} + \beta \times RI(u,v) + \gamma \times s_{uv} \quad (8)$$

Where $\sum_{i=1}^{m} |r_{iu} - r_{iv}|$ is used to weaken the negative effect of the big rating disparity. Interestingness measure $RI(u,v)$ is used to analyze the correlation degree of different items, and β is a parameter that ameliorates the item-item interestingness, when $RI(u,v)$>0 means items u and v are in positive correlation while $RI(u,v)$<0 means items u and v are in negative correlation. Considering the similarity between items within the same category is higher, the element of the items categories similarity matrix is put in the item similarity, and γ is a parameter that adjusts the items categories similarity.

IV. Experimental Results and Analysis

A. Data set

We use data set provided by MovieLens. The website is created by GroupLens Research at University of Minnesota.100K of public data set is selected in this paper, one hundred thousand records, including 1628 movies rated by 943 users. Each user at least rates 20 movies; rate them from one to five stars, five being the best. Users show their interest by number they rated. We divided the database into a training set and a test set which 80% of the data was used as training set and 20% of the data was used as test set.

B. Evaluation Metrics

Mean Absolute Error (MAE) between ratings and predictions is widely used to evaluate the quality of collaborative filtering methods. The MAE is a measure of the deviation of recommendations from their true user-specified values.For each ratings-prediction pair$<p_i, q_i>$ this metric treats the absolute error between them, i.e., $| p_i, q_i |$ equally. The MAE is computed by first summing these absolute errors of the n corresponding ratings-prediction pairs and then computing the average. Formally,

$$MAE = \frac{\sum_{i=1}^{n} |p_i - q_i|}{n} \quad (9)$$

The lower the MAE, the more accurately the recommendation engine predicts user ratings.

C. Adjusting parameters

The value of the parameters has a significant impact on the recommendation quality, as different values lead to substantially different sensitivity. There are three parameters: α which is used to control the extent to which the similarity to frequently purchased/occurring items will be de-emphasized, β which is used to adjust the item-item interestingness and γ which is used to decide the effect of items categories similarity.

To study the sensitivity of the recommendation algorithm on this parameter *α, we* performed a sequence of experiments in which we varied α from 0.1 to 1.0 in increments of 0.1. Fig. 2 shows the MAE achieved on the different values of α. If $0.3 \le \alpha \le 0.6$, then our algorithm achieved good performance.

To experimentally determine the impact of the parameter β on the quality of the prediction, we selectively varied the value of β to be used for similarity computation from 0.1 to 1.0 in increments of 0.1. Fig. 3 shows the value of β has a significant impact on the recommendation quality, as different values of β lead to substantially different MAE. If $0.4 \le \beta \le 0.9$, then our algorithm achieved consistently good performance, and when β=0.9, MAE is least.

The parameter of γ is used to adjust items categories influence on item similarity. We selectively varied the value of γ to be used for similarity computation from 0.1 to 1.0 in increments of 0.1. Fig. 4 shows when $\gamma < 0.8$, MAE is decreasing with increasing of γ. When $\gamma = 0.8$, MAE is least and our algorithm achieved best performance.

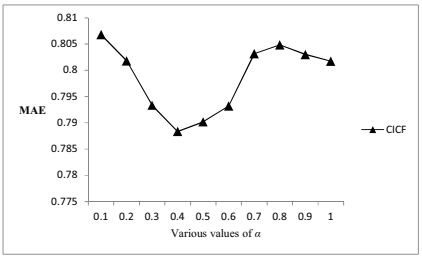

Figure 2. Sensitivity of the parameter α on CICF algorithm.

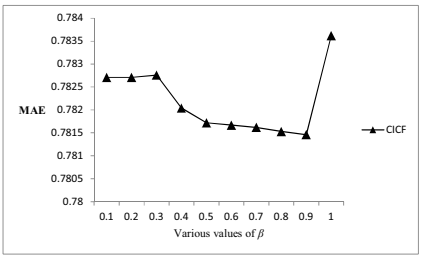

Figure 3. Sensitivity of the parameter β on CICF algorithm.

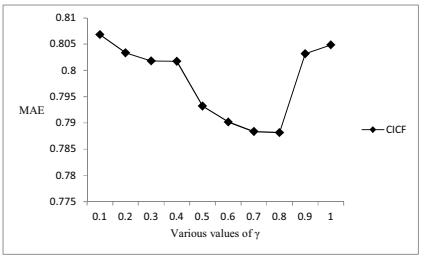

Figure 4. Sensitivity of the parameter γ on CICF algorithm.

D. Comparison with the conditional probability -based Recommendation Algorithm

Finally, to compare the performance of the proposed collaborative filtering algorithm based on item category and interest measure (CICF) with traditional conditional probability-based collaborative filtering algorithms (CPCF),We performed an experiment where we varied the number of neighbors from 10 to 40 in increments of 5 to be used and computed MAE of both recommendation algorithms. These results are shown in Fig. 5. We can observe that our proposed collaborative filtering algorithm provides better quality than the traditional conditional probability-based collaborative filtering algorithms at all different numbers of neighbors.

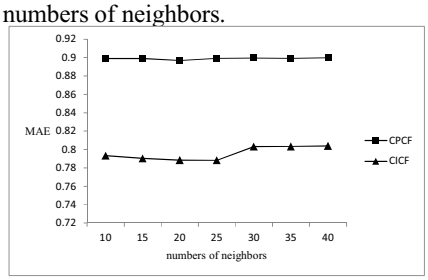

Figure 5. Comparison of prediction quality of CICF and CPCF algorithms.

V. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this paper we presented and experimentally evaluated a *top-N* recommendation algorithm that uses items categories similarity and item-item interestingness to compute the recommendations. Our results showed that our proposed algorithm provides more accurate recommendations than those provided by traditional CF techniques. Our next job is to take the advantage of all kinds of technologies, such as clustering, matrix decomposition and cloud model to improve the algorithm. Furthermore, how to combine different attribute of users and items in the algorithm is a fascinating and challenging research area.

ACKNOWLEDGMENT

The work is partially supported by National Basic Research Program of China (No. 2012CB114505) , China National Funds for Distinguished Young Scientists(No. 31125008) ,Natural Science Foundation of Jiangsu Province(No.BK2009393), Jiangsu Qing Lan Project and Scientific and Technological Innovation Foundation of Jiangsu (No. CXLX11_0525), Technological Innovation Foundation of Nanjing Forestry Scientific (No. 163070079), the Innovation Fund Project for College Student of Jiangsu (No. 164070742).

REFERENCES

[1] B.M. Sarwar, G. Karypis, J.A. Konstan, et al. “Application of dimensionality reduction in recommender system -- A Case Study,” Proc. ACM Web KDD Web Mining for E-Commerce Work shop, Boston, MA, United States , 2000,pp.82-90.

[2] K. Goldberg, T. Roeder, D. Gupta, et al. “Eigentaste: a constant time collaborative filtering algorithm,” Information Retrieval, vol.4(2),pp. 133-151, 2001.

[3] C.C.Aggarwal, J.L.Wolf, K.L.Wu, et al. “Horting hatches an egg: a new graph-theoretic approach tocollaborative filtering,” Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, New York, ACM Press, 1999. pp.201-212.

[4] M. Papagelis, D. Plexousakis and T. Kutsuras, “Alleviating the sparsity problem of collaborative filtering using trust inferences,” Proc. 3rd Int. Conf. on Trust Mnagement, Berlin, Springer-Verlag, 2005, pp.224-239.

[5] L.H Ungar and D.P Foster, “Clustering methods for collaborative filtering,” Proc. Workshop on Recommendation Systems at the 15th National Conf. on Artificial Intelligence, Menlo Park, CA: AAAI Press, 1998, pp.112-125.

[6] B.M. Sarwar, G. Karypis, J.A Konstan, et al. “Item-based collaborative filtering recommendation algorithms,” Proc. 10th Int. Conf. on the World Wide Web, New York, ACM Press, 2001, pp.285-295.

[7] L. Geng , J. H. Hamilton. “Interestingness Measures for Data Mining: A survey,”. ACM Computing Surveys, vol.38(3),2006, pp.11-29.

[8] G. Karypis. “Evaluation of item-based top-n recommendation algorithms,” Proc. 10th Int. Conf. on Information and Knowledge Management, New York, ACM Press, 2001, pp.247-254.