An Item Based Collaborative Filtering Recommendation Algorithm Using Rough Set Prediction

Ping SU
Zhejiang Business Technology Institute,
Ningbo 315012, P. R. China
e-mail: supingzjbti@163.com

HongWu YE
Zhejiang Textile & Fashion College,
Ningbo 315211, P. R. China
e-mail: yehongwuzjbti@163.com

***Abstract*—Recommender systems represent personalized services that aim at predicting users' interest on information items available in the application domain. Collaborative filtering technique has been proved to be one of the most successful techniques in recommendation systems in recent years. Poor quality is one major challenge in collaborative filtering recommender systems. Sparsity of users' ratings is the major reason causing the poor quality. To solve this problem, this paper proposed an item based collaborative filtering recommendation algorithm using the rough set theory prediction. This method employs rough set theory to fill the vacant ratings of the user-item matrix where necessary. Then it utilizes the item based collaborative filtering to produce the recommendation. The experiments were made on a common data set using different filtering algorithms. The results show that the proposed recommender algorithm combining rough set theory and item based collaborative filtering can improve the accuracy of the collaborative filtering recommendation system.**

Keywords-recommender system; item based collaborative filtering; rough set; sparsity

I. INTRODUCTION

While the rapid growth and wide application of the Internet and information technology has provided an unprecedented abundance of information resources, it has also led to the problem of information overload. Thus, methods to help find resources of interest have attracted much attention from both researchers and vendors. To deal with the problem, the personalized recommendation systems play a more important role [1,2].

Recommender system plays an important role particularly in an electronic commerce environment as a new marketing strategy. Although a multifarious of recommendation techniques has been developed recently, collaborative filtering (CF) has been known to be the most successful recommendation techniques and has been used in a number of different applications such as recommending web pages, movies, tapes and products. The CF assumes that a good way to find a certain user's interesting content is to find other people who have similar interests with him. CF methods operate upon user ratings on observed items making predictions concerning users' interest on unobserved items. However, in most cases in real-world applications, the ratio of rated items to the total of available items is very low. The absence of a sufficient amount of available ratings significantly affects CF methods reducing the accuracy of prediction. The sparsity of ratings problem is particularly important in domains with large or continuously updated list of items as well as a large number of users. The sparsity problem may occur when either none or few ratings are available for the target user, or for the target item that prediction refers to, or for the entire database in average. Different treatments are required and different prediction techniques must be employed depending on the sparsity conditions, making the selection of an appropriate approach a cumbersome task [3,4]. Current CF approaches are limited in the sense that they address specific aspects of the above problem.

To solve this problem, in this paper, we proposed an item based collaborative filtering recommendation algorithm using the rough set theory prediction. This method employs rough set theory to fill the vacant ratings of the user-item matrix where necessary. Then it utilizes the item based collaborative filtering to produce the recommendation. The experiments were made on a common data set using different filtering algorithms. The results show that the proposed recommender algorithm combining rough set theory and item based collaborative filtering can improve the accuracy of the collaborative filtering recommendation system.

II. USING ROUGH SET THEORY TO PREDICTION WHERE NECESSARY

In our proposed approach, we employ the rough set theory [5,6,7] to predict the vacant ratings where necessary.

A. Basic definition

Definition1 S = (U, A, {Va}, a) is the information system, where U is a no empty finite set, named discussed field. A is a no empty finite set too, named property set. Va is the value field of the property of a A. U → Va is a mapping relation, which makes any element of discussed field U have the exclusive value when getting property a from Va. If A is composed by condition attribute set C and conclusion attribute set D, meanwhile C and D satisfy $C \cup D = A$, $C \cap D = \Phi$, then S is a decision making system. To show

978-0-7695-3615-6/09 $25.00 © 2009 IEEE
DOI 10.1109/JCAI.2009.155

simply , (U , C ∪{ d}) can be used to express decision making system.

Definition2 For knowledge denotation system S = (U , A ,{ Va} , a) , suppose R ∈A , X ∈U ,POSR (X) = R- X ,NEGR (X) = U - R- X and BNR (X) = R- X - R- X are respectively called positive fields , negative fields and border of R below X.

Definition3 UB (x , X) = card([x]B ∩X) / card ([x]B) is the reliance degree of element x to set X. where , card denotes the base of gather.

Definition4 Given R is an equivalent relation family and r ∈ R , when ind(R) = ind (R - { r}) , r is omissible for R , or else r is not omissible. All sets of no omissible relation in R are called the core of R, noted as CORE(R) . Therefore, CORE(R) = ∩RED(P), RED(P) is all the reduced family of P.

Definition5 Set information system S = (U, A, V, f), define M as :M = (M (i, j)) n×n and M (i, j) = { ak | (ak (xi) ≠ ak (xj)) ∧ (ak (xi) ≠ *) ∧ (ak (xj) ≠ *) } , * express the lack。

Definition6 Set information system S = (U, A, V, f), define miss attribute set of object xi as MAS = { ak | ak (xi) = 3 , k = 1, 2, ..., p}、define nearest set of object xi as NSi = { xj | M (i, j) = <, i ≠ j, j = 1, 2, ..., n}、define miss object set of information system S as MOS = { xi | MASi ≠ <, i = 1, 2, ..., n} define set of object xi as LSi= { xj | P (i, j) = maxx k∈NS i{ P (i, k) } , xj ∈NSi , i ≠ j, j = 1, 2, ..., n} , the similarity of xi and xj on the attribute set A as P (i, j)

B. Prediction where necessary

Based on the basic definition, utilize the maximal similarity object of the object which has the null attribute value having presence. Fill up the object as the one possess most strong null attribute value fill up capable. Thereby, this paper forms the self-contained information system.

Specific algorithm as follows:

```
Input: vacant user rating matrix
Output: complete user rating matrix
Begin
calculate the differentiate M0, MASi0, MOS0
n=0
while(Sn+1≠Sn)
  for each xi in MOSn
    calculate LSin
  for each xi not in MOSn
    for(int k=1;k<p+1;k++)
  ak(xin+1)=ak(xin)
  for each xi in MOSn
  for(int k=1;k<p+1;k++)
      if(LSin==1)
        if(xj in LSin) ak(xin+1)=ak(xin)
      else
        if((xi, xj in LSin)∧( ak ( xin ) ≠ ak ( xjn ) )∧(
                  ak ( xin ) ≠*)∧( ak ( xjn ) ≠*))
          ak(xin+1)=*
        else ak(xin+1)=ak(xin)
  end for
  n=n+1
end while
calculate the empty with the max frequency
End
```

III. USING ITEM BASED COLLABORATIVE FILTERING TO PRODUCE RECOMMENDATIONS

A. Measuring the item rating similarity

There are several similarity algorithms that have been used [2,3,4]: Pearson correlation, cosine vector similarity, adjusted cosine vector similarity, mean-squared difference and Spearman correlation.

Pearson's correlation, as following formula, measures the linear correlation between two vectors of ratings as the target item t and the remaining item r.

$$sim(t,r) = \frac{\sum_{i=1}^{m}(R_{it}-A_t)(R_{ir}-A_r)}{\sqrt{\sum_{i=1}^{m}(R_{it}-A_t)^2\sum_{i=1}^{m}(R_{ir}-A_r)^2}}$$

Where R_{it} is the rating of the target item t by user i, R_{ir} is the rating of the remaining item r by user i, A_t is the average rating of the target item t for all the co-rated users, A_r is the average rating of the remaining item r for all the co-rated users, and m is the number of all rating users to the item t and item r.

The cosine measure, as following formula, looks at the angle between two vectors of ratings as the target item t and the remaining item r.

$$sim(t,r) = \frac{\sum_{i=1}^{m}R_{it}R_{ir}}{\sqrt{\sum_{i=1}^{m}R_{it}^{\ 2}\sum_{i=1}^{m}R_{ir}^{\ 2}}}$$

Where R_{it} is the rating of the target item t by user i, R_{ir} is the rating of the remaining item r by user i, and m is the number of all rating users to the item t and item r.

The adjusted cosine, as following formula, is used for similarity among items where the difference in each user's use of the rating scale is taken into account.

$$sim(t,r) = \frac{\sum_{i=1}^{m}(R_{it}-A_i)(R_{ir}-A_i)}{\sqrt{\sum_{i=1}^{m}(R_{it}-A_i)^2\sum_{i=1}^{m}(R_{ir}-A_i)^2}}$$

Where R_{it} is the rating of the target item t by user i, R_{ir} is the rating of the remaining item r by user i, A_i is the average rating of user i for all the co-rated items, and m is the number of all rating users to the item t and item r.

B. Prediction using item-based CF

Since we have got the membership of item, we can calculate the weighted average of neighbors' ratings, weighted by their similarity to the target item.

The rating of the target user u to the target item t is as following:

$$P_{ut} = \frac{\sum_{i=1}^{c} R_{ui} \times sim(t,i)}{\sum_{i=1}^{c} sim(t,i)}$$

Where Rui is the rating of the target user u to the neighbour item i, sim(t, i) is the similarity of the target item t and the neighbour item i, and c is the number of the neighbours.

IV. EXPERIMENTAL EVALUATION AND RESULTS

In this section, we describe the dataset, metrics and methodology for the comparison between traditional and proposed CF algorithm, and present the results of our experiments.

A. Data set

We use MovieLens collaborative filtering data set to evaluate the performance of proposed algorithm [8]. MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota and MovieLens is a web-based research recommender system that debuted in Fall 1997. Each week hundreds of users visit MovieLens to rate and receive recommendations for movies. The site now has over 45000 users who have expressed opinions on 6600 different movies. We randomly selected enough users to obtain 100, 000 ratings from 1000 users on 1680 movies with every user having at least 20 ratings and simple demographic information for the users is included. The ratings are on a numeric five-point scale with 1 and 2 representing negative ratings, 4 and 5 representing positive ratings, and 3 indicating ambivalence.

B. Performance measurement

The metrics for evaluating the accuracy of a prediction algorithm can be divided into two main categories [8,9]: statistical accuracy metrics and decision-support metrics. Statistical accuracy metrics evaluate the accuracy of a predictor by comparing predicted values with user provided values. Decision-support accuracy measures how well predictions help user select high-quality items. In this paper, we use decision-support accuracy measures.

Decision support accuracy metrics evaluate how effective a prediction engine is at helping a user select high-quality items from the set of all items. The receiver operating characteristic (ROC) sensitivity is an example of the decision support accuracy metric. The metric indicates how effectively the system can steer users towards highly-rated items and away from low-rated ones. We use ROC-4 measure as the evaluation metric. Assume that p1, p2, p3, ..., pn is the prediction of users' ratings, and the corresponding real ratings data set of users is q1, q2, q3, ..., qn. See the ROC-4 definition as following:

$$ROC\text{-}4 = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} v_i}$$

$$u_i = \begin{cases} 1, & p_i \geq 4 \ and \ q_i \geq 4 \\ 0, & otherwise \end{cases} \qquad v_i = \begin{cases} 1, & p_i \geq 4 \\ 0, & otherwise \end{cases}$$

The larger the ROC-4, the more accurate the predictions would be, allowing for better recommendations to be formulated.

C. Comparing the proposed CF with the user based CF

In this paper, we compare the proposed CF that combining the rough set theory and the item based CF with the only utilizing the user based CF. As showing in the Figure 1, it includes the decision support accuracy metrics of ROC-4 for the two comparing methods in relation to the different numbers of recommender items. The obvious conclusion is that the combining method is better than only using the user based CF.

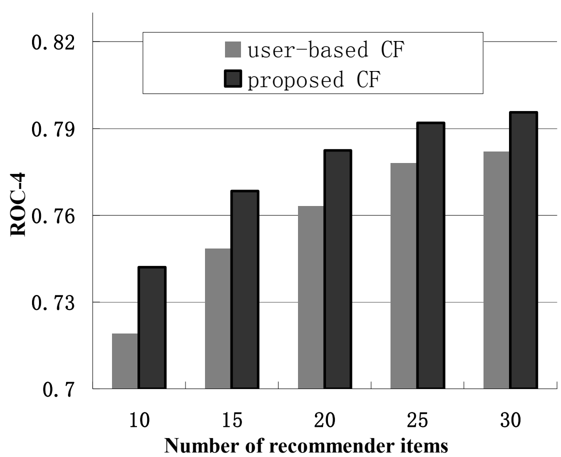

Figure 1. Comparing the proposed CF with the user based CF

V. CONCLUSIONS

Recommender systems represent personalized services that aim at predicting users' interest on information items available in the application domain. Collaborative filtering technique has been proved to be one of the most successful techniques in recommendation systems in recent years. Poor quality is one major challenge in collaborative filtering recommender systems. Sparsity of users' ratings is the major reason causing the poor quality. To solve this problem, in this paper, we proposed an item based collaborative filtering recommendation algorithm using the rough set theory prediction. This method employs rough set theory to fill the vacant ratings of the user-item matrix where necessary. Then it utilizes the item based collaborative filtering to produce the recommendation. The experiments were made on a common data set using different filtering algorithms. The results show that the proposed recommender algorithm combining rough set theory and item based collaborative filtering can improve the accuracy of the collaborative filtering recommendation system.

REFERENCES

[1] SongJie Gong, The Collaborative Filtering Recommendation Based on Similar-Priority and Fuzzy Clustering, In: Proceeding of 2008 Workshop on Power Electronics and Intelligent Transportation System (PEITS2008), IEEE Computer Society Press, 2008, pp. 248-251.

[2] Jong-Seok Lee, Chi-Hyuck Jun, Jaewook Lee, Sooyoung Kim, Classification-based collaborative filtering using market basket data, Expert Systems with Applications 29 (2005) 700–704.

[3] Hyung Jun Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Information Sciences 178 (2008) 37-51.

[4] George Lekakos, George M. Giaglis, Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors, Interacting with Computers 18 (2006) 410–431.

[5] Chong-Ben Huang, Song-Jie Gong, Employing rough set theory to alleviate the sparsity issue in recommender system, In: Proceeding of the Seventh International Conference on Machine Learning and Cybernetics (ICMLC2008), IEEE Press, 2008, pp.1610-1614.

[6] Yee Leung, Wei-Zhi Wu, Wen-Xiu Zhang, Knowledge acquisition in incomplete information systems: A rough set approach, European Journal of Operational Research 168 (2006) 164–180.

[7] ZHANG wei, LIU lu, GE Jian, Collaborative Filtering Algorithm based on Rough Set, MINI- MICRO SYSTEMS,Vol126 No. 11, 2005.

[8] Huang qin-hua, Ouyang wei-min, Fuzzy collaborative filtering with multiple agents, Journal of Shanghai University (English Edition), 2007,11(3):290-295.

[9] Gao Fengrong, Xing Chunxiao, Du Xiaoyong, Wang Shan, Personalized Service System Based on Hybrid Filtering for Digital Library, Tsinghua Science and Technology, Volume 12, Number 1, February 2007,1-8.