# Research on personalized recommendation system on Item-based collaborative filtering algorithm

Ji-chun ZHAO [1, 2, a], Shi-hong LIU[2], Junfeng ZHANG[1, 3]

[1] Beijing Academy of Agriculture and Forestry Sciences

[2] Institute of Agricultural Information, Chinese Academy of Agricultural Sciences

[3] The Research Center of Beijing Engineering technology for Rural Remote Information Services, Beijing, China

[a] email: zhaojichun_0@163.com

**Keywords:** Distance Learning; Item based collaborative filtering algorithm; Personalized Learning

**Abstract.** A lot of learning resources and information occupy rural distance education platform, and farmers don't know how to find useful information in the learning platform. Personalized distance learning system can provide farmers with required learning resources. The paper analyzes classical collaborative filtering algorithms, and Item based collaborative filtering algorithm is used in distance education platform, which is experimented with the distance learning platform data. The test result show that item based collaborative filtering algorithm in prediction accuracy and coverage is better with the growing of the sparsity of the data set, and the average accuracy is82.99%, the average coverage is 99.06%.

## Introduction

Beijing Academy of Agriculture and Forestry Sciences has built a distance education platform for farmers, which has run about six years, and has accumulated a lot of data. The platform realize the functions of VOD(video on demand), video living, learning question answering, and learning data management. Now the registration users is more than 40 million, and the videos teaching resources has reached more than 9,000 items. Usually the farms are perplexed in the face of great information, so it is necessary to research personalized recommendation system in distance education platform. which can analysis the user's behavior of individual to provide them with useful information. The paper analyzes classical collaborative filtering algorithms, and Item based collaborative filtering algorithm is used in distance education platform, which is experimented with the distance learning platform data.

Collaborative filtering is a technique used by some recommender systems. Collaborative filtering has two senses, a narrow one and a more general one. In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets. Collaborative filtering systems have many forms, but many common systems can be reduced to two steps. Look for users who share the same rating patterns with the active user (the user whom the prediction is for). Use the ratings from those like-minded users found to calculate a prediction for the active user. Collaborative Filtering in Recommender System is shown in Figure 1.

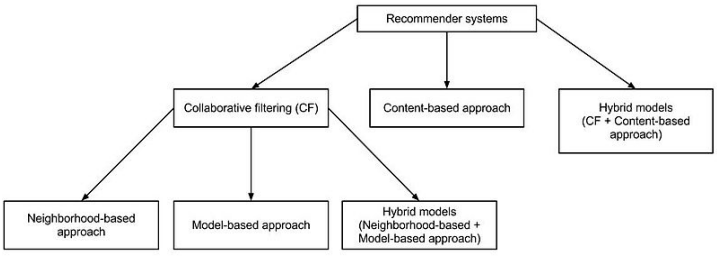


Figure 1 Collaborative Filtering in Recommender Systems

 

**Item base collaborative filtering algorithm**

The basic idea of Item base collaborative filtering algorithm is to calculate the similarity between items according to the historical data of all users preferences, and then the articles of similar users like are recommended to the users. For example, the items of A and C are very similar, if a user like A, and then C is recommended to the user. The process of Item base collaborative filtering algorithms is shown in Figure 2.

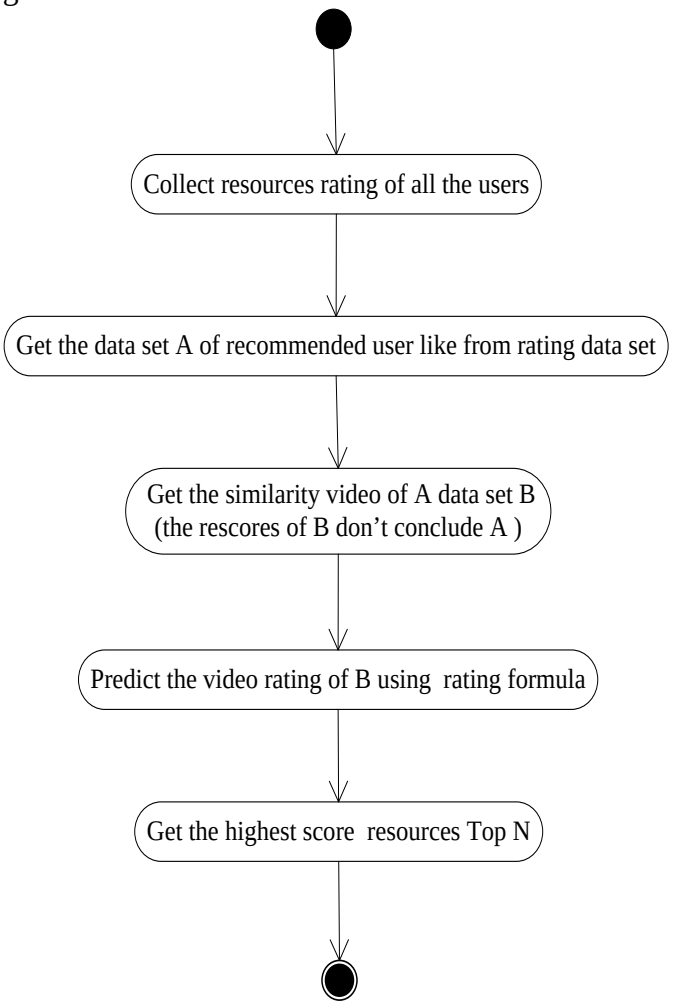


Figure 2 The process of Item base collaborative filtering algorithms

Assuming the users of the system is {U1, U2,... }. The resources in the system of {V1, V2,... }. The process of recommendation resources for user $U_k$ is shown follows:

1 Collect user video score, the score contained two part.

a) Direct rating, the video resources are marked directly by user from the score function interface.

b) implicit rating, the video resources are marked by the system according to user's behaviors of browsing video, adding video to learning plan and so on. The rating data set is gotten in the set form of $\{(V_1,P_1),\ldots,(Vm,Pm)\}$, and V is the video, P is the rating data. All the video resources rating data can get. For example , The video of $V_K$ rating is $R_k$, $R_k = \{(U_1, P_1),\ldots,(U_n, P_n)\}$.

2 $\delta = \{V_x,\ldots,V_y\}$, $\delta$ is the video set that the user of $U_K$ likes, which is obtained from rating data.

3 According to the score data of each video, the similar resources of $\delta$ are searched and resources that UK has rated are removed, and then alternative recommended video set $\varepsilon = \{V_i,\ldots,V_j\}$ is obtained.

4 The video rating of user $U_k$ for $\varepsilon = \{V_i,\ldots,V_j\}$ is forecasted.

a) Get the video set $\epsilon = \{V_m,\ldots,V_n\}$ which Uk likes.

b) The predicting rating formula of UK for $V_h$ ($V_h \in$ ∂ is $(\sum_{i=m}^{n}(S_{hi} \times V_{ki}))/\sum_{i=m}^{n} S_{hi}$, $S_{hi}$ is the similarity between $V_h$ and $V_i(V_i \in \epsilon)$, $V_{ki}$ is the video rating of $U_k$ for $V_i$.

5 read the score the highest Top N resources are recommended that the rating data are highest.

## Material and method

### Test Environment

The test client environment is shown in Table 1

Table 1 the test client environment

| The hardware environment | | |
|---|---|---|
| Name | Hardware device | Configuration |
| Client | Notebook | Intel i3-2350M cpu @ 2.30GHz 2.30GHz；4G 500G Hard disk；100M Ethernet card； |
| **The software environment** | | |
| Name | Soft type | Vision |
| IE | Browner | IE9 vision：9.0.8112.16421 |
| Win7 | Operating system | Win7 sp1 |
| Oracle | Database | Oracle 11g |
| Eclipse | Development tool | My Eclipse 8.5 |
| PLSQL Developer | Database link tool | PLSQL Developer 8.0 |

### Data set

Movielens100k is as the basis for data set, which has 100,000 video score record, 943 users and 1682 different video resource ID, and each user at least has 20 movies evaluations, the value is from 1 to 5, the higher value indicates a higher degree of user preference for the movie.

### The division of training and test sets

Due to different degrees of sparse data, the recommendation system can simulate real world working situations more effectively, an effective verification system information in different conditions. The data sets were randomly taken out a fixed percentage of the record as the training set, the remaining data is as a test set, and then the sparsity of the training set is calculated.

### Video Rating prediction with test set

According to the algorithm, the similarity is calculated in accordance with the training set of data, the test set is calculated for each record score prediction based on the training results.

### Calculating MAE

Predictive accuracy of a classical approach is the actual scoring average absolute error of the measurement system and the user's predicted score, which is MAE（Mean Absolute Error). Prediction scores set is {p1, p2, ...,pn}, and the set of the actual score is {q1, q2, ... qn}, the MAE is calculated as follows:

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

### The accuracy and coverage

Movielens100k uses 5-point test data set, we generally think that users don't like video with 1-2 score, and users will like the video with 3-5 score. So this test cut off the point for predicting scores less than 2 points, and then we calculate the remaining video prediction score for accuracy (MAE) and the coverage.

## Result and conclusion

Through the algorithm , The MAE and coverage is shown in the following Table 2. The change curve of MAE and coverage is shown in Figure 3 and Figure 4.

Table 2 The MAE and coverage

| Data volume | the proportion of the training set | Training set sparsity | MAE | Selection of prediction score | Coverage | Selection of Coverage |
|---|---|---|---|---|---|---|
| 100000 | 20% | 1.49% | 0.8511 | 0.8504 | 97.17% | 96.56% |
| 100000 | 40% | 2.73% | 0.8388 | 0.839 | 99.51% | 99.06% |
| 100000 | 60% | 3.94% | 0.8343 | 0.8338 | 99.77% | 99.29% |
| 100000 | 80% | 5.14% | 0.8336 | 0.8326 | 99.80% | 98.98% |

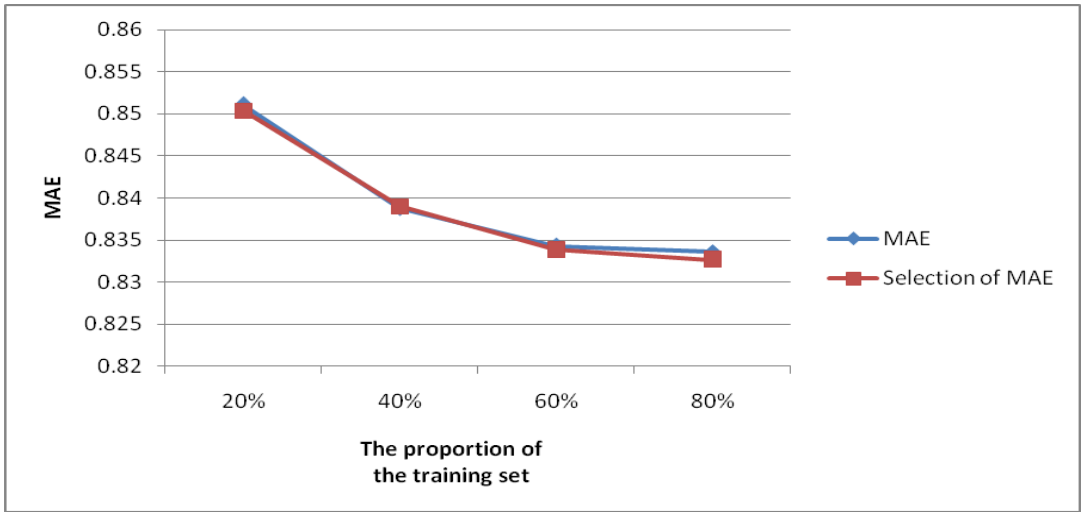


Figure 3 the change curve of MAE

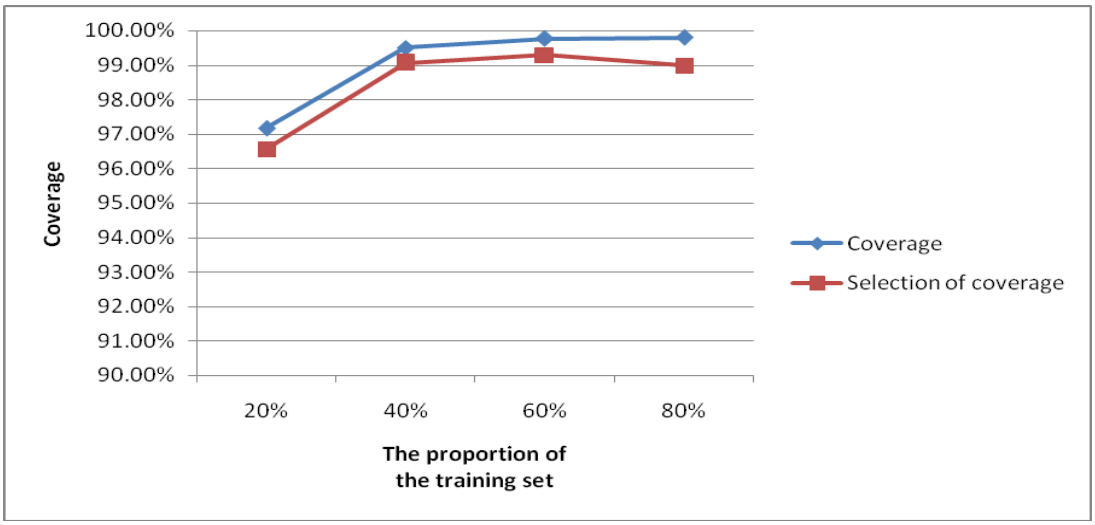


Figure 4. the change curve of coverage

Data sparsity is an important index in the prediction accuracy. The accuracy MAE and coverage of item base collaborative filtering algorithm showed a significance difference with the data set sparsity change. The test result show that item based collaborative filtering algorithm in prediction accuracy and coverage is better with the growing of the sparsity of the data set, and the average accuracy is82.99%, the average coverage is 99.06%.

**Acknowledgement**

In this paper, the research was sponsored by National Science and Technology Support Program(Project No. 2014BAD10B02) , which is construction and application of provincial rural information service platform in developed area, and supported by Distance Education Innovation Team Project of Beijing Academy of Agriculture and Forestry Sciences.

## References

[1] Sarwar B,Karypis G,Konstan J. Item-Based collaborative filtering recommendation algorithms [A].Hong Kong:ACM Press,2001.285-295.

[2] Deshpande,Karypis G. Item-Based Top-N Recommendation Algorithms[J].ACM Transactions on Information Systems,2004,(01):143-177.doi:10.1145/963770.963776.

[3] Breese J,Hecherman D,Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[Technical Report MSR-TR-98-12,][R].San Francisco California,1998.43-52.

[4] Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer, 2011, pp. 1-35.

[5] Terveen, Loren; Hill, Will (2001). "Beyond Recommender Systems: Helping People Help Each Other". Addison-Wesley. p. 6. Retrieved 16 January 2012.

[6] Heckmann D., Schwartz T., Brandherm B. et al. GUMO-the general user model ontology [C]. In: International Conference on User Modeling, Edinburgh, UK, 2005: 28–432.

[7] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Bosagh Zadeh WTF: The who-to-follow system at Twitter, Proceedings of the 22nd international conference on World Wide Web

[8] http://en.wikipedia.org/wiki/Collaborative_filtering

[9] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl, GroupLens: an open architecture for collaborative filtering of netnews, Computer Supported Cooperative Work, pp175-186, Chapel Hill, North Carolina, 1994.