Research on the cold start of collaborative filtering

Zhonghua Lu

Linyi University, Shandong, China

sduman@126.com

Key Words: Collaborative filtering; Cold start; Recommendation system; Electronic commerce system

ABSTRACT. The recommendation system is the information service tool that can satisfy the user's personalized demand. It can effectively solve the problem of information overload. Collaborative filtering recommendation technology is widely used, but the technology is difficult to recommend new users and new projects, which is called the cold start problem. This paper introduces the collaborative filtering technology and the cold start problem, then introduces a new model of collaborative filtering to solve the problem of cold start.

Introduction

The appearance and application of Web2.0 greatly changes the online behavior of Internet users, which from the search and browse to the mutual influence and share. User selection of network presents exponential growth, it is more difficult for users to find useful information, that is, we often say that the problem of information overload. To solve this problem, the recommendation system came into being, and has a large number of applications in the electronic commerce system through which the user is recommended to meet the interests of books, movies and music. Collaborative filtering technology is one of the most famous and commonly used recommendation technologies. Collaborative filtering technology is based on the user's score, which can deal with the complex objects, and has no special requirement to the recommended object. Therefore, it is widely used in all kinds of recommender systems. Collaborative filtering recommendation technology is faced with a series of problems and the cold start problem is one of the major impacts. Cold start problem is a classical problem in collaborative filtering recommendation algorithm, which has been affecting the recommendation quality of the traditional collaborative filtering recommendation system. For the electronic commerce recommendation system, every day a large number of new users access the system and a large number of new items were added. In order to better retain customers and tap potential customers, we need effective recommendation for new users and new items. The cold start problem is caused by the lack of scoring information, and the method that not considering the content alleviates the cold start problem. Here, we proposed a method to improve the accuracy of collaborative filtering algorithm.

Collaborative filtering technology

Collaborative filtering technology mainly includes user-based collaborative filtering and item-based collaborative filtering based on project.

According to the user - item score data set, the method uses statistical techniques to search for a group of users with similar historical preferences as the target users, called "the neighbors". User-based collaborative filtering technology can be divided into three steps:

1) Nearest neighbor selection. Using the similarity measure method, a group of K users with a high degree of history score similarity is generated for the target user u. Common similarity measure method includes cosine similarity, Pearson correlation coefficient and adjusted cosine similarity. The specific calculation formula is as follows.

Cosine similarity：sim(u_1,u_2) = $\cos(\overrightarrow{u1}, \overrightarrow{u2}) = \frac{\overrightarrow{u1}\cdot\overrightarrow{u2}}{\|\overrightarrow{u1}\|\times\|\overrightarrow{u2}\|}$

Pearson correlation coefficient：sim(u1, u2) = $\frac{\sum_{i\in I}(R_{u1,i}-\overline{R_{u1}})\times(R_{u2,i}-\overline{R_{u2}})}{\sqrt{\sum_{i\in I}(R_{u1,i}-\overline{R_{u1}})^2}\sqrt{\sum_{i\in I}(R_{u2,i}-\overline{R_{u2}})^2}}$, Where I refers to the score set of u1 and u2, while $\overline{R_{u1}}$, $\overline{R_{u2}}$ refers to the average score of u1 and u2 respectively.

Adjusted cosine similarity, sim(u1, u2) = $\frac{\sum_{i\in I}(R_{u1,i}-\overline{R_i})\times(R_{u2,i}-\overline{R_i})}{\sqrt{\sum_{i\in I}(R_{u1,i}-\overline{R_i})^2}\sqrt{\sum_{i\in I}(R_{u2,i}-\overline{R_i})^2}}$, where $\overline{R_i}$ represents the average score of the item I.

2) Score forecast. After the k nearest neighbor selection, the target user is calculated by using the weighted average method. UN represents k nearest neighbor set.

$$R_{u,i} = \frac{\sum_{u'\in UN}(sim(u,u')R_{u',i})}{\sum_{u'\in UN}(sim(u,u'))}$$

3) Project recommendation. The items of N highest score was recommended to the user.

Experiment

Movielens is the data set, and the evaluation index is the absolute mean error MAE and the average error square root RMSE. The smaller the MAE and RMSE, the higher the accuracy of the prediction. Pu,I refers to the predictive score of the item I, and ru,I refers to the actual score of the item i.

$$\text{MAE} = \frac{1}{K}\sum_{u,i}|P_{u,i}-r_{u,i}|$$

$$\text{RMSE} = \sqrt{\frac{\sum_{u,i}(P_{u,i}-r_{u,i})^2}{K}}$$

User attributes set D = {d1, d2, d3}={age, occupation, gender}, and Item category set C ={c1,c2,c3,c4}= { fun, intellectual, adventurous, romantic}. At the same time, we set up four kinds of scenarios according to the different weight w of the user attribute elements.

Table 2 Experimental Scenarios

Scenarios	Weights
Scenario 1	$w_1 = 0.6, w_2 = 0.3, w_3 = 0.1$
Scenario 2	$w_1 = 0.3, w_2 = 0.6, w_3 = 0.1$
Scenario 3	$w_1 = 0.3, w_2 = 0.1, w_3 = 0.6$
Scenario 4	$w_1 = 0.33, w_2 = 0.34, w_3 = 0.33$

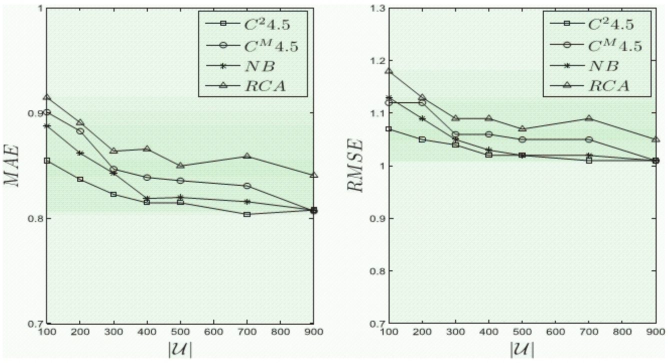

Figure 1 Scenario 1

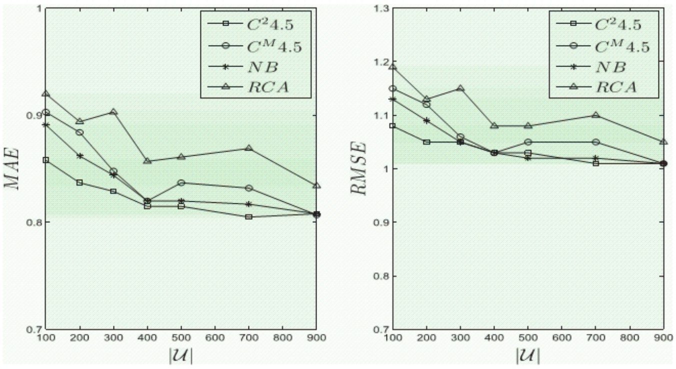

Figure 2 Scenario 2

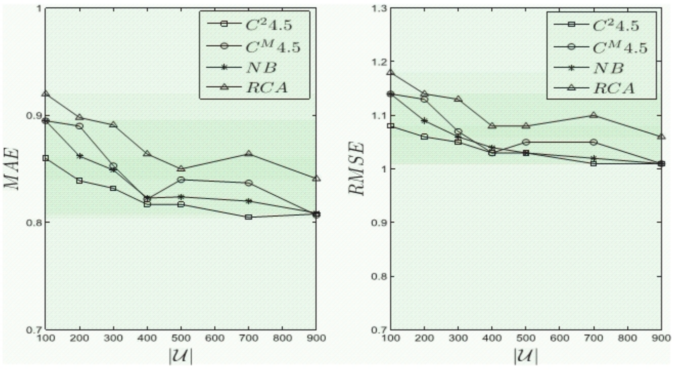

Figure 3 Scenario 3

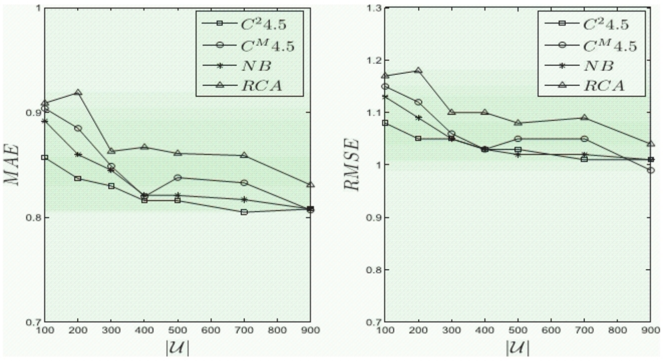

Figure 4 Scenario 4

The results show that when the number is more than 1000, the performance of Naive Bayesian is the best. For the whole model, the smaller the MAE, the better the performance of the model prediction.

References

[1] C. De Rosa, J. Cantrell, A. Havens, J. Hawk, L. Jenkins, B. Gauder, R. Limes, D.

Cellentani, OCLC, Sharing, Privacy and Trust in Our Networked World: A Report

to the OCLC Membership, OCLC, 2007.

[2] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (2005) 734–749.

[3] H.N. Kim, A.T. Ji, I. Ha, G.S. Jo, Collaborative filtering based on collaborative tagging for enhancing the quality of recommendations, Electronic Commerce Research and Applications 9 (1) (2010) 73–83

[4] S. Loh, F. Lorenzi, R. Granada, D. Lichtnow, L.K. Wives, J.P. Oliveira, Identifying similar users by their scientific publications to reduce cold start in recommender systems, in: Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST2009), 2009, pp. 593–600.

[5] Zhang, Z. K., Liu, C., Zhang, Y., & Zhou, Z. (2010). Solving the cold-start problem in recommender systems with social tags, EPL (Vol. 92).

[6] Lam X N, Vu T, Le T D, et al. u1. Addressing cold-start problem in recommendation systems [C] ICMIMC' 08. New York, USA, 2008:208-221

[7] Chu W, Park S T. Personalized recommendation on dynamic contents using predictive bilinear models [C] // Proceedings of the 18th International Conference on World Wide Web.2009: 691-700

[8] Blerina Lika, et al. Facing the cold start problem in recommender systems. Expert Systems with Application 41 (2014) 2065-2073