Collaborative Filtering Recommendation Algorithm Based on Improved Similarity Computing

Aili LIU[1, a], Baoan LI[1, b*]

[1]Computer School, Beijing Information Science and Technology University, Beijing, China

[a]bjliuaili@163.com, [b]liba2010@139.com

*Corresponding author

Keywords: Personalized Recommendation; Collaborative Filtering; MAE (Mean Absolute Error); User Characteristic; Item Attribute

Abstract. At presently the most widely used algorithm is collaborative filtering in the Personalized Recommendation Systems for E-Commerce. Aiming at the problem that the recommendation is not accurate due to the data sparsity, a collaborative filtering algorithm based on user characteristics and item attributes preference was proposed in this study. It obtained the nearest neighbor users and similar items by analyzing the user characteristics, item attributes and the data of user's historical scores, and then computing the similarity between the two users based on the user-based collaborative filtering algorithm. It gave an algorithm which has the lower value of MAE and may improve the accuracy of the recommendation services.

Introduction

With the rapid development of Internet technology, people's lives have becoming more convenient. While at the same time, the problem of information overload is becoming serious. For e-commerce systems, the traditional search service model has been unable to meet the needs of accurate and efficient, so personalized recommendation systems for E-Commerce arising at the historic moment. Personalized recommendation services for E-Commerce can provide information and advice to customers so that to help customers deciding what products should to be bought, and then imitating sales people to help customers completing the purchase process [1]. The key task of the recommendation system is to recommend the related items according to the user interests analyzed from the relevant historical behavior of visitors. There are more current mainstream recommended methods such as Content-based recommendation, Collaborative Filtering recommendation, recommendation based on Association Rules and Hybrid recommendation. There are many famous recommendation examples such as *Amazon*, *Jingdong*, *Taobao* and *movie* which has used intelligent recommendation system to provide personalized recommendation service for users. Although the existing personalized recommendation technologies have achieved great success in various fields, but they also need to be improved for the accuracy of recommendation [2]. Now the Collaborative Filtering technology is the widely used technology, while it is also facing many questions, such as cold start, data sparsity and scalability. If these questions can be improved effectively, especially the recommended inaccuracy caused by the sparse data, it will have an important effort on personalized recommendation.

Some researchers have been carried out on personalized recommendation related research, mainly includes the following research. Among the Content-based recommendation, the TF-IDF [3] (Term Frequency-Inverse Document Frequency, abbreviated TF-IDF) method proposed by Salton is commonly used. There are many famous systems based on content filtering system, such as the Massachusetts institute of technology Letizia [4] and newsgroups filtration system NewsWeeder [5] and others. In the 1992, the Collaboration Filtering was proposed by Goleberg et al. [6] and applied in the email recommendation system in the Tapestry. Since then, as a kind of filtering technology, based on reducing information overload, is widely used in the Internet [7]. The representative of the application of collaborative filtering recommendation technology is the GroupLens [8] system developed by the Minnesota state university GroupLens team. Collaboration Filtering is divided

into user-based and item-based two kinds. Its main idea is to analyze the user's interest, find the specified users similar (interest) users in the user group, and get the target users neighboring collection. Then combine these similar user evaluations for information, and forming system for the specified user for this information to predict the degree of preference. Thereby the recommended [9] is generated. For the data sparsity problem in the Collaboration Filtering system, Lee using a pseudo scoring [10], Ahn and others used heuristics algorithms measure the similarity between users [11], Sarwar used of the singular value decomposition (SVD) clustering method to reduce the user-item rating matrix of dimension resulting in relatively dense data to solve the data sparsity problem [12]. Yuan proposed the collaborative filtering algorithm based on clustering [13], Liu proposed the collaborative filtering algorithm based on clustering of the attributes to predicted the score [14]. Agrawal put forward the Association Rules technology at first [15]. It is a recommendation algorithm based on data mining technology. In the study the improved collaborative filtering algorithm was approved and used. Aiming at the problem that the recommendation is not accurate due to the data sparsity, a collaborative filtering algorithm based on user characteristics and item attributes preference was studied.

Improved Collaborative Filtering Algorithm

Similarity computing based on user characteristics

Web site generally in the registration requirements of the user was filled in gender, age, occupation, etc. the impacts of these factors on the recommendation in the calculation of similarity ware considered according to the use of the data set.

(1) Sex characteristics

Different gender of the users have different needs for the commodity, the similarity measure formula based on sex is:

Sex(u,v)=1, u.sex=v.sex; Sex(u,v)=0, u.sex!=v.sex;

(2) Age characteristics

Users of different age groups have different needs, the similarity measure formula based on age is shown as:

Age(u,v)=1, |u.age-v.age|<=10; Age(u,v)=10/|u.age-v.age|, |u.age-v.age|>10.

(3) Job characteristics

the similarity measure formula based on job is listed as:

Job(u,v)=1,u.job=v.job; Job(u,v)=0, u.job!=v.job.

So the formula of the similarity based on user's characteristic was showed as Eq. 1.

$$SimA(u,v)=a*Sex(u,v)+b*Age(u,v)+c*Job(u,v). \quad (1)$$

Among them, a, b, c is the percent of sex characteristics, age characteristics and job characteristics, and a+b+c=1.

Similarity calculation based on the preference of item attributes

The similarity calculation based on the preference of item attributes can help the recommendation system to find the implicit relationship among the sparse user evaluation data. The basic idea is that the degree of similarity between users is influenced by item's score and the degree of preference for a certain type of item. When two user evaluation the similar item, we considered that have high similarity between them. With I={I1,I2,...,In} and P={P1,P2,...,Pk}respectively present the item and attribute set, then the item-attribute matrix can be shown in Table 1.

Table 1. The item-attribute matrix

	P_1	P_2	...	P_k
I_1	h_{11}	h_{12}	...	h_{1n}
I_2	h_{21}	h_{22}	...	h_{2n}
...	...	...	...	...
I_n	h_{n1}	hn_2	...	h_{nk}

Among them, hij indicates whether the item has the property p, if there was 1 and 0 otherwise. The preference that user u to item i can be expressed by Iu,i=Scorei/Score. Scorei represents the score that u to class i, and Score represents the score that u to all items. Assume that the project has a total of k attributes, through the above formula to accumulate the item attribute preference vector in different categories Iu=(Iu,1,Iu,2,Iu,3,...,Iu,k), then similarity calculation based on the preference of item attributes is listed as Eq. 2.

$$\mathrm{Si}\, m_s(u,v) = \frac{\sum_{i=1}^{K} I_{u,i} I_{v,i}}{\sqrt{\sum_{i=1}^{k} I^2_{u,i}} \sqrt{\sum_{i=1}^{k} I^2_{v,i}}} \tag{2}$$

Combination recommendation algorithm

According to the Eq.1 and Eq.2 calculating the similarity respectively, then the similarity of the final between two users is calculated by the weighted method, and can be expressed by Eq. 3.

$$Sim(u,v) = m \times Sim_R(u,v) + n \times Sim_S(u,v) + (1-m-n) \times Sim_A(u,v) \tag{3}$$

Among them, m and n presents the percent of self similarity, the specific values can be obtained through experiments by combining the conventional filtering algorithm. SimR(u,v) is the similarity based on the matrix of user-item score, SimS(u,v) is the similarity based on the preference of item attributes, SimA(u,v) is the similarity based on user characteristics.

Results and discussion

Experiment using Movie Lens data set (http://www.grouplens.org/), which is by the university of Minnesota GroupLens Research project team to provide a research recommendation system based on Web, used to receive the user's score and provides the recommendation list. In the experiment, the training data set (u1. Base) was used to implement the target users to score of films, and then compare forecast results with the test data set (u1. Test) scores. Data set includes 943 users to 1682 films of 100000 rating score interval of 1 to 5. The experiment of the density of user ratings for: 100000 / (943 * 1682) * 100% = 6.30%. So the degree of user ratings sparse is 93.70%, thus the score data is sparse.

Evaluation index

Personalized recommendation system is widely applied in a kind of evaluation standard is MAE (Mean Absolute Error). MAE is through the target user prediction score compared with target user actual score, represented by the deviation degree between the two recommended the accuracy of the results. MAE value is smaller, says the higher prediction accuracy and the better the quality of the recommendation. MAE calculation formula can be represented as Eq. 4.

$$MAE = |E| = \frac{\sum_{i=1}^{n} |p_i - q_i|}{N} \tag{4}$$

Among them, pi for predicting the user's score, qi is the user's actual score; N is the number of users.

Experimental design and results analysis

The purpose of this experiment is to test the effectiveness of the proposed algorithm in this paper.

Firstly, based on the traditional collaborative filtering algorithm, is determined in the paper that a=0.5, b=0.2, c=0.3, m=0.6, n=0.3. Then the improved collaborative filtering algorithm is compared with the traditional user-based and item-based collaborative filtering method, and the recommendation effect is judged. The number of neighbors is increased from 10 to100, and the test results are shown in Fig. 1.

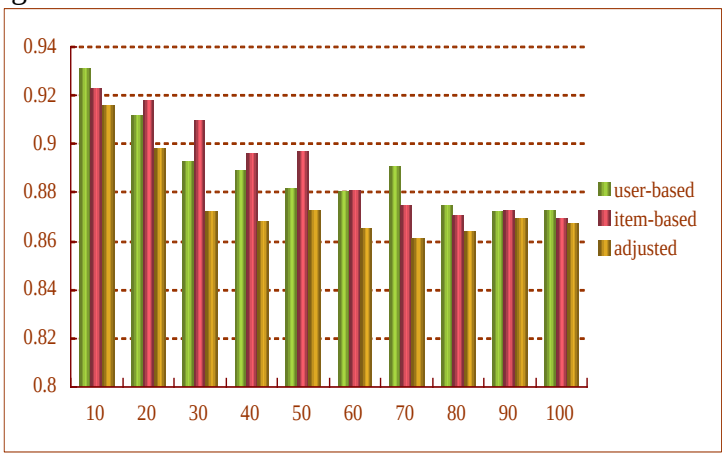

Fig.1. The experimental results

From the experimental results, under different nearest neighbor values, the MAE value of the improved collaborative filtering algorithm is significantly less than the traditional user-based collaborative filtering and item-based collaborative filtering. Therefore, the improved algorithm can reduce the problems caused by the sparse data, and has better recommendation quality.

Conclusion

Aiming at the problem that the recommendation is not accurate due to the data sparsity, in this study, a collaborative filtering algorithm based on user characteristics and item attributes preference is proposed. The improved algorithm calculated the similarity between users by using the weighted method and then by predicting the target user's score to item. It explored an approach which can effectively alleviate the problems caused by the sparse data and improve the accuracy of the recommendation.

Acknowledgement

The work was supported by Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges under Beijing Municipality (Grant No. IDHT20130519), and also Project of Graduate Education Quality Engineering of Beijing Information Science and Technology University (Grant No. YJT201511) and Project of the Specialty Construction and Comprehensive Reform under Beijing Municipality (Grant No. 71M1510818).

References

[1] Resnick, Varian. Recommender Systems. Communications of the ACM, 1997, 40(3):56-78.

[2] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[3] Gerard, Salton. Automatic Text Processing [M]. Addison-Wesley, Dec.1988.

[4] Chun Zeng, Chunxiao Xing, Lizhu Zhou. Overview of personalized service technology. Journal of Software. 2002, 13(10):1952-1961.

[5] Lang,K.Newsweeder. Learning to Filter News. In Proceedings of the 12th International Conference on Machine Learning, Lake Tahoe, CA, 1995, 331-339.

[6] Konstan J A, Miller B N, Maltz D, Herlocker J L, Gordon L R,Riedl J. Grouplens. Applying Collaborative Filtering to Usenet News. Communications of the ACM,1997,40(3):77-87

[7] Qinfang Xin. Research on personalized recommendation system for E-commerce [D]. HuaQiao University, 2010

[8] Resnick P, Iakovou N, Sushak M, et al. GroupLens: An open architecture for collaborative filtering of net news. Proc. 1994 Computer Supported Cooperative Work, Chapel Hill, 1994, 175 -186.

[9] Peiyong Xia. Research on personalized recommendation technology in collaborative filtering algorithm [D]. China Ocean University, 2011.

[10] Lee TQ, Park Y, Park YT. A time-based approach to effective recommender systems using implicit feedback. Expert Systems with Applications, 2008, 34(4): 3055-3062.

[11] Ahn HJ. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences, 2008, 178(1): 37-51.

[12] Sarwar B, Karypis G, Konstan J, et al. Application of Dimensionality Reduction in Recommendation System-A Case Study[C]. ACM Web KDD 2000 Web Mining for E-commerce Workshop, 2000.

[13] Yuan Li. Research on Collaborative Filtering Personalized Recommendation Algorithm Based on Clustering [D].Central China Normal University, 2014.

[14] Xianfeng Liu, Tongcun Liu. Research on the recommendation algorithm of project evaluation based on attribute clustering[J].Statistics and Decision,2012,18:9-11.

[15] Rakesh Agrawal, Tomas Imielinski, Arun Swarni. Mining association rules between sets of items in large databases[C]. Proc. of the ACM Sigmund Conference on Management of Data.1993, 207-216.