Journal of Computers Vol. 30 No. 5, 2019, pp. 205-212
doi:10.3966/199115992019103005015

Two-layer Neighbor Selection Scheme Based on Kendall Correlation Coefficient and Standard Deviation

Xing-Cao[1,2*], Yun-Liu[1,2], Kun-Mi[3], Qun-feng Lu[4]

[1] School of Electronic and Information, Beijing Jiaotong University, Beijing 100044, China
(18111006, liuyun)@bjtu.edu.cn

[2] Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing 100044, China

[3] Beijing Huayu Software Co., Ltd, Beijing 100044, China
mik@thunisoft.com

[4] School of Information and Communication Engineering, University of Electronic Science and Technology Xian 710000, China
luqunfeng@uestc.edu.cn

Received 13 January 2019; Revised 12 March 2019; Accepted 12 March 2019

Abstract. Collaborative filtering is the basis of the recommendation system. Collaborative filtering generally finds users (neighbors) with the same interests as the target users among a large number of users. How to determine whether the users have the same interests as the target users and how to form the neighbors of the target users. Sorting directories has become a major issue. However, there are some shortcomings in the existing neighbor selection scheme. For example, when calculating the deviation of two users from the same group of items ratings, the existing two-layer neighbor selection scheme only considers the sum of the individual items rating differences. It is unfair to users who are not much different from the target users. We propose a trustworthiness calculation scheme based on Kendall correlation coefficient and standard deviation. Specifically, when the sum of the difference between the two neighbors and the target user is the same, the trustworthiness of the user with high Kendall correlation coefficient and small standard deviation is higher.

Keywords: Kendall correlation coefficient, neighbors, recommender system, standard deviation, trust

1 Introduction

With the rapid development of the Internet, the number of users has increased, and tens of thousands of messages have been generated every day. People are gradually entering the era of information overload from the era of information scarcity, so it is a big challenge for both information consumers and information producers. How can information consumers find the information they are interested in in a large amount of information, and how the information producers push the information they produce to consumers, so the recommendation system came into being. The recommendation system conducts personalized calculation by studying the user's interest preference, and the system discovers the user's interest points, thereby guiding the user to discover their own information needs [1].

Collaborative filtering recommendation is currently the most widely used and most successful recommendation system [2]. Collaborative filtering recommendation algorithms fall into two categories, namely user-based collaborating algorithm (user-based collaborating algorithm) and item-based collaborative filtering. The user-based collaborative filtering algorithm is to discover the user's likes of

* Corresponding Author

goods or content (such as product purchase, collection, content comment or sharing) through the user's historical behavior data, and measure and score these preferences [3]. The relationship between users is calculated according to the attitudes and preferences of different users for the same product or content, and product recommendation is performed among users with the same preferences. So how to choose the user with the same interests as the target user is the key to the algorithm.

This paper aims to improve the recommendation accuracy of the user-similarity based CF, The existing two-layer neighbor selection scheme introduces the Good value and the Bad value, and the Bad value represents the difference in rating between the neighboring user and the target user for the same items [4]. However, when calculating the Bad value, only the sum of the differences between the ratings of the two users is considered, and the influence degree of the different rating differences is ignored. We have improved the existing scheme and introduced the Kendall coefficient and standard deviation in calculating the Bad value, which improves the trustworthiness of the neighbors.

2 Related Work

In this section, we have detailed the shortcomings of the existing two-layer neighbor selection scheme towards online recommender systems.

The existing two-layer neighbor selection scheme towards online recommender systems consists of two parts, the capability evaluation module and the trust evaluation module [5].

Ui stands for the user, Ii stands for the items (movie), and Ri stands for the user's rating of the i project as shown in Table 1.

Table 1. An example of a user-item rating metrix

	I_1	I_2	I_3	I_4	I_5
U_1	-	2	-	-	4
U_2	2	2	3	3	5
U_3	5	1	-	3	1
U_4	3	3	4	5	1
U_5	1	3	2	4	5

2.1 Capability Evaluation Module

The existing two-layer neighbor selection scheme use one of the most popular similarity calculation methods, Pearson correlation coefficient [6], Pearson correlation coefficient is calculated as:

$$\mathrm{sim}(u,v)=\frac{\Sigma_{i\in I_{uv}}(R_{ui}-\overline{R_u})(R_{vi}-\overline{R_v})}{\sqrt{\Sigma_{i\in I_{uv}}(R_{ui}-\overline{R_u})^2\Sigma_{i\in I_{uv}}(R_{vi}-\overline{R_v})^2}} \quad \textbf{(1)}$$

where Rui and Rvi are the ratings of user u and user v for item i, respectively; Iuv is the set of commonly rated items by both users; and Ru and Rv are the average rating values of user u and user v, respectively.

A user v's capability to be used for predicting target user u's preferences (i.e. marked as ava(u, v)), is calculated as in below equation [4].

$$ava(u,v)=\begin{cases}0, I_vCI_u \ or \ \mathrm{sim}(u,v)<0\\ \frac{|I_{uv}|}{|I_u|}\times\mathrm{sim}(u,v), \quad \text{other}\end{cases} \quad \textbf{(2)}$$

In ava(u, v), ava(u, v) is set to 0 when the set of items rated by the neighboring user is a true subset of the item set of the target user(For example, the item rating set of the U_1 user is the true subset of the U_3 user in Table 1). This solves the problem that neighbors cannot recommend each other. The Pearson correlation coefficient calculates the similarity of a pair of users, mainly based on their rating values for rating items, while ignoring the total number of these items. So the introduction of |Iuv|/|Iu| in ava(u, v) [7] solves this problem.

2.2 Trust Evaluation Module

A user v's behavior contains both a good portion (i.e. consistent preference) and a bad portion (i.e. inconsistent preference), marked as gi(u, v) or bi(u, v), each of which can be quantified as a continuous value in the range of (0, 1) [8]. Specifically, gi(u, v) and bi(u, v) are computed as follows.

$$g_i(u,v) = 1 - \frac{|R_{vi} - R_{ui}|}{R_{\max} - R_{\min}} \tag{3}$$

$$b_i(u,v) = \frac{|R_{vi} - R_{ui}|}{R_{\max} - R_{\min}} \tag{4}$$

Where Rmax and Rmin represent the maximum and minimum values in the recommended system, respectively, where Rmax is 5 and Rmin is 1. Furthermore, the total number of good/bad behaviors conducted by user v is calculated as the sum of his/her good/bad behavior values on all the commonly rated items [4].

$$G(u,v) = \Sigma_{i \in I_{uv}} g_i(u,v) \tag{5}$$

$$B(u,v) = \Sigma_{i \in I_{uv}} b_i(u,v) \tag{6}$$

At the end, the trustworthiness of user v from target user u's perspective (i.e. marked as tru(u, v)) is calculated as in (7). The more good behaviors a user v conducts, the higher trust value he/she will obtain [4, 9].

$$tru(u,v) = \frac{G(u,v) + 1}{G(u,v) + B(u,v) + 2} \tag{7}$$

3 Proposed Scheme

In this section, we have detailed the shortcomings of the existing two-layer neighbor selection scheme, and discuss the proposed scheme in details.

3.1 Inadequacies of the Trust Evaluation Module

The difference in preference between the neighbor user and the target user is mainly measured by the bad behavior value of the neighbor user, that is, B(u, v). However, even when B(u1, v) = B(u2, v), the difference between User 1 and User 2 and the target user is different.

Table 2.

	I_1	I_2	I_3	I_4	I_5
U_1	1	1	1	1	1
U_2	5	1	1	1	1
U_3	2	2	2	2	1
U_4	4	2	1	1	1
U_5	3	2	2	1	1

Assume that user U1 is the target user, and calculate the B (Ui, U1) values of users U2, U3, U4, U5 and user U1 respectively. It can be seen that the B (Ui, U1) values of each neighbor are equal. Intuitively, User 3 should be most similar to target User 1 in comparison with other users, because User 3's preference for each project is similar to Target User 1, which is relatively stable. However, User 2's rating for item 1 is completely opposite to that of the target user, and interest preference fluctuations are relatively large.

3.2 Improved Trust Evaluation Module

We propose a trust evaluation module that uses Kendall correlation coefficient and standard deviations. At the end, the proposal of a two-layer neighbor selection strategy is presented.

Kendall correlation coefficient. The Kendall correlation coefficient is a statistical value used to measure the correlation between two random variables [10]. The calculated correlation coefficient is used to test the statistical dependence of two random variables. Suppose that the score sets of two neighbor users for the item are X and Y respectively, and the number of their elements is N. The i-th (1<=i<=N) values in the two sets are represented by Xi and Yi respectively.

Let (X_1, Y_1), (X_2, Y_2), …, (Xn, Yn) be a set of observations of the joint random variables X and Y respectively, such that all the values of (Xi) and (Yi) are unique. Any pair of observations (X_i, Y_i) and (X_i, Y_i), where i<j, are said to be concordant if the ranks for both elements (more precisely, the sort order by X and by Y) agree: that is, if both $(X_i>X_j)$ and $(Y_i>Y_j)$; or if both $(X_i<X_j)$ and $(Y_i<Y_j)$. They are said to be discordant, if $(X_i>X_j)$ and $(Y_i<Y_j)$; or if $(X_i<X_j)$ and$(Y_i>Y_j)$. If $(X_i=X_j)$ or $(Y_i>Y_j)$, the pair is neither concordant nor discordant [11].

There are formulas for calculating the value of the Kendall correlation coefficient [12].

$$Tau = \frac{C - D}{\sqrt{(N_3 - N_1)(N_3 - N_2)}} \tag{8}$$

$$N_3 = \frac{N(N-1)}{2} \tag{9}$$

$$N_1 = \sum_{i=1}^{s} \frac{U_i(U_i - 1)}{2} \tag{10}$$

$$N_2 = \sum_{i=1}^{s} \frac{V_i(V_i - 1)}{2} \tag{11}$$

Where C is number of concordant pairs, D is number of discordant pairs. N1 and N2 are calculated for the sets X and Y, respectively. Now take the calculation of N1 as an example:

Combine the same elements in X into small sets, s represents the number of small sets in set X (for example, X contains elements: 1 2 3 4 3 3 2, then the s obtained here is 2, because only 2, 3 There are the same elements), Ui represents the number of elements contained in the i-th small set.

Table 3.

	I_1	I_2	I_3	I_4	I_5	I_6
U_1	1	1	2	2	3	3
U_2	1	2	2	3	3	2

Assume U1 is the target user, X=[1, 1, 2, 2, 3, 3], Y=[1, 2, 2, 3, 3, 2]. Calculated C=7, D=1, $S_{X=}3$, $S_Y=2$

Standard deviation. According to the deficiencies in Table 2, We introduce standard deviation factors into the scheme, and the standard deviation [13] reflects the degree of dispersion of a data set. The standard deviation of user U3 (σ_3)was found to be the smallest, which is consistent with the subjective situation.

In summary, the improved trust evaluation module is:

$$tru(u, v) = \frac{G(u, v) + Tau/2 + 1}{G(u, v) + B(u, v) + \sigma + 2} \tag{12}$$

3.3 Neighbor Selection Strategy

Among the traditional collaborative filtering algorithms, there are two most popular neighbor selection methods. One is to select a fixed number of neighbors with the highest similarity score, but when a target user has fewer neighbors with higher similarity scores, selecting a fixed number of neighbors will result in lower trust in the set of neighbor users. The neighbors will downgrade the accuracy of the recommendation system. The other is to select the part of the neighbor whose similarity score is above a certain threshold, but you need to set different thresholds again in different environments [14].

Assuming that K users are selected as the neighbors of the target user, the existing two-layer neighbor selection scheme uses the capability evaluation module to select the first-level neighbors. Set the parameter K', calculate ava(u, v) according to formula (2), sort in descending order by size, and select the first K' users as the first layer neighbor. K' is calculated as follows:

$$K' = [e \times K] \tag{13}$$

When the number of users whose ava(u, v) is not 0 is greater than K', the former K' users are selected as the first-level neighbors, which are recorded as N'(U); when ava(u, v) is not When the number of users with 0 is less than K', the user who is not 0 is selected as the first-level neighbor and is recorded as N'(U).

The improved trust evaluation module is used to select the second layer of neighbors. For the user in N'(U), calculate tru(u, v) according to formula (12), sort by descending order, and select the first K users as the second layer neighbor.

When the number of users that are not 0 in tru(u, v) is greater than K, the first K users are selected as the first layer neighbors, which are recorded as N(U) [15]; when tru(u, v) is not 0, the user is not 0. When the number is less than K, the user who is not 0 is selected as the first layer neighbor and is recorded as N(U).

4 Experimental Results

In this section, to validate the effectiveness of the proposed improvement, we conducted experiments based on real user data sets. These experiments were performed on an Intel i7 2.5 GHz, 8 GB RAM, Windows 10 computer using the Python programming language. We conduct experiments using the MovieLens-100k dataset collected by the GroupLens Research Project at the University of Minnesota.

4.1 Performance Evaluation Metric

First, we predict the target user's rating for each movie based on the selected N neighbor users, which is calculated according to Equation (14).

$$P_{ui} = \overline{R_u} + \frac{\Sigma_{v \in N(u)} (R_{vi} - \overline{R_v}) \times ava(u, v) \times tru(u, v)}{\Sigma_{v \in N(u)} | ava(u, v) \times tru(u, v) |} \tag{14}$$

To measure the effectiveness of the proposed method, we use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are widely accepted by the research community. MAE and RMSE are used to compute the deviation between the predicted ratings and the actual ratings in all experiments. Specifically, the MAE and RMSE are calculated as [4].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} | r_i - P_i | \tag{15}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r_i - P_i)^2} \tag{16}$$

In the above formula, r_i represents the actual score value of the target user for item i, and p_i is the corresponding predicted score value. N is the total number of items scored by the target user in the test set. The smaller the MAE or RMSE, the better the performance.

4.2 Module Testing

Firstly, the optimal N values of the two schemes are determined. We separately select the N values for the existing two-layer neighbor selection scheme and the improved two-layer neighbor selection scheme.

It can be seen from Fig. 1 that when N=20 and e=2.1, the MAE has a minimum value. As can be seen from Fig. 2, for the improved scheme, when N=40 and e=2.43, the MAE has a minimum value. Because when the number of selected neighbors is small, more users with similar interests to the target users cannot be obtained; when too many neighbors are selected, users with lower scores are included.

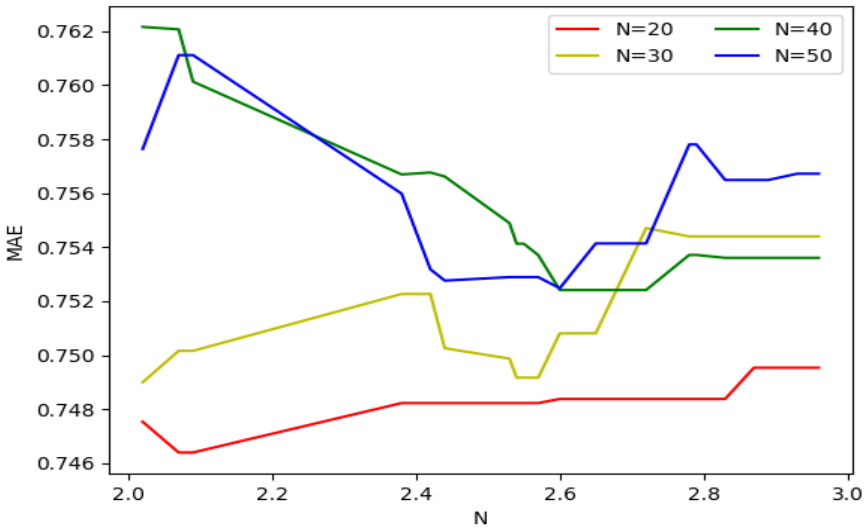

Fig. 1. The existing two-layer neighbor selection scheme

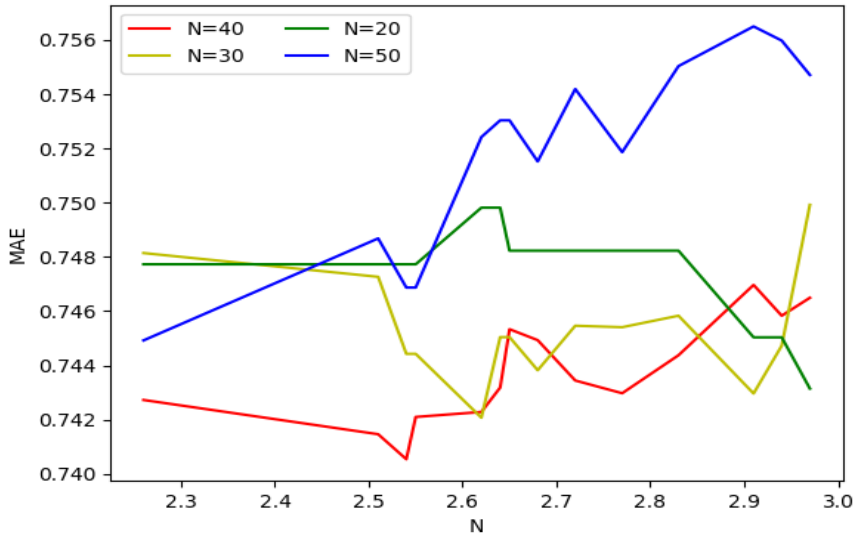

Fig. 2. The improved two-layer neighbor selection scheme

Fig. 3 shows the performances of different schemes with different e values, where the x-axis represent the e value and the y-axis represent the MAE, respectively. In Fig. 3, we observe that the proposed scheme achieves the best MAE for all different e values. Fig. 4 shows the performances of different schemes with different e values, where the x-axis represent the e value and the y-axis represent the RMSE, when the e value is less than 2.93, the RMSE value of our proposed scheme is smaller than the existing neighbor selection scheme. And get the best RMSE when e value is equal to 2.71.

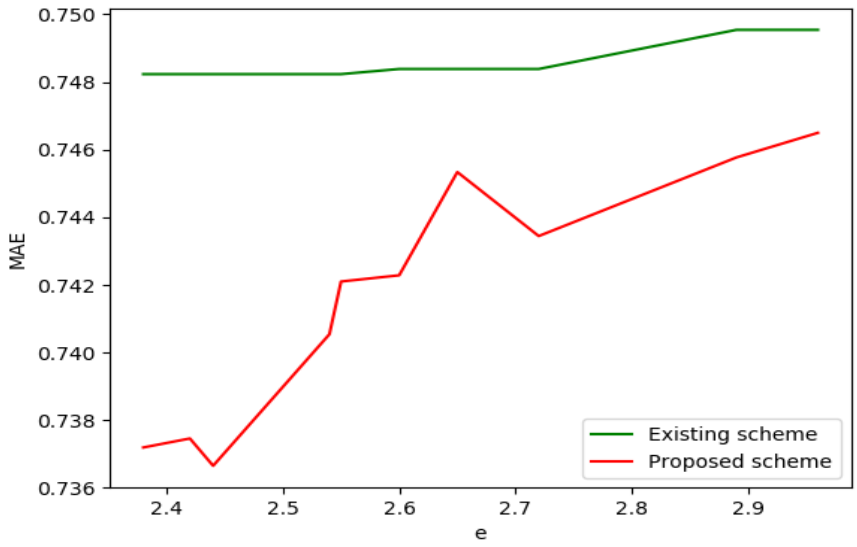

Fig. 3. Impact of parameter e on MAE

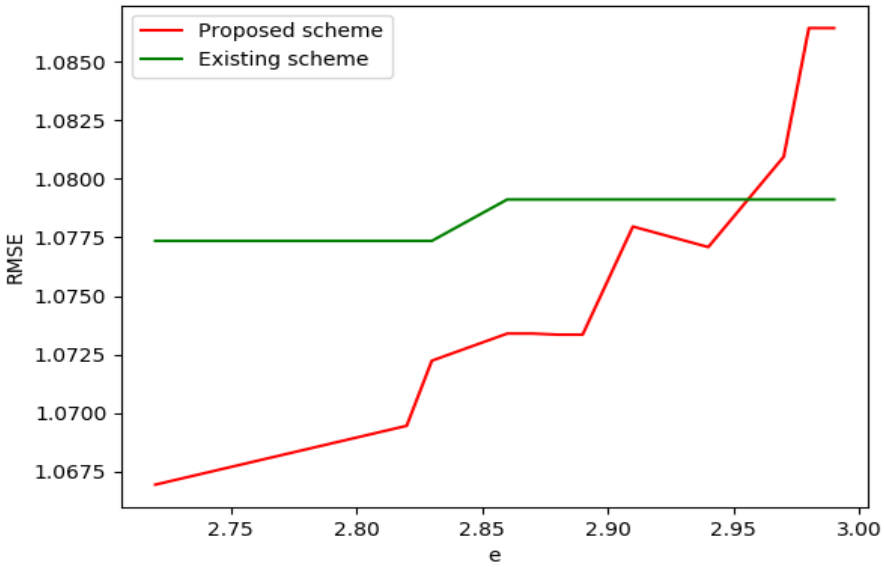

Fig. 4. Impact of parameter e on RMSE

6 Conclusion

In this work, an improved two-layer neighbor selection scheme is proposed for collaborative filtering recommender systems, aiming at improving the recommendation accuracy by introduced Kendall correlation coefficient and standard deviation in the trust evaluation module. Specifically, when calculating the difference between the rating of the neighboring user and the target user for a group of identical items, (1) the rating of the target user is set to set X, the rating of the neighboring user is set to set Y, and then the Kendall correlation coefficient of the set X, Y is calculated to obtain the value t. (2) Calculate the standard deviation of the set X, Y to get the value σ.

To evaluate the performance of the proposed scheme, experiments are conducted on the MovieLens-100K dataset. The experimental results show that the proposed scheme has higher recommendation accuracy and is superior to the existing neighbor selection scheme.

Acknowledgements

This research was funded by the National Key Research and Development Program of China, grant number 2018YFC0831300, and the Fundamental Research Funds for the Central Universities, grant number 2017JBZ107.

References

[1] F. Ricci, L. Rokach, B. Shapira, Recommender systems: introduction and challenges, in F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (Eds.), Recommender Systems Handbook, Springer, 2015, pp. 1-34.

[2] M. Nilashi, O. Ibrahim, K.J.E.S.w. A. Bagherifard, A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques, Expert Systems with Applications: An International Journal archive 92(C)(2018) 507-520.

[3] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proc. the 26th International Conference on World Wide Web, 2017.

[4] Z. Zhang, Y. Liu, Z. Jin, R. J.N. Zhang, A dynamic trust based two-layer neighbor selection scheme towards online recommender systems, Neurocomputing 285(2018) 94-103.

[5] D. Li, C. Chen, L. Qin, S. Li, , Y. Zhao, T. Lu, N. Gu, An algorithm for efficient privacy-preserving item-based collaborative filtering, Future Generation Computer Systems 55(2016) 311-320.

[6] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: I. Cohen, Y. Huang, J. Chen, J. Benesty, Noise Reduction in Speech Processing, Springer, 2009, pp. 1-4.

[7] Y. Xu, R. Hao, W. Yin, Z. Su, Parallel matrix factorization for low-rank tensor completion, Inverse Problems & Imaging 9(2)(2013) 601-624.

[8] S. Deng, L. Huang, G. Xu, X. Wu, Z. Wu, On deep learning for trust-aware recommendations in social networks, IEEE Transactions on Neural Networks and Learning Systems 28(5)(2017) 1164-1177.

[9] X. Wang, X. Zhang, J. Wu, Collaborative filtering recommendation algorithm based on one-jump trust model, Journal on Communications 36(6)(2015) 193-200.

[10] H. Abdi, The Kendall rank correlation coefficient, in N. J. Salkind (Ed.), Encyclopedia of Measurement and Statistics, Sage, Thousand Oaks, 2007, pp. 508-510.

[11] M.-T. Puth, M. Neuhäuser, G.D. Ruxton, Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits, Animal Behaviour 102(2015) 77-84.

[12] K.-S. Wong, M. Seo, M.H. Kim, Secure two-party rank correlation computations for recommender systems, in: Proc. TRUSTCOM '15 Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, 2015.

[13] G. Guo, J. Zhang, N. Yorke-Smith, A novel Bayesian similarity measure for recommender systems, in: Proc. the Twenty-Third International Joint Conference on Artificial Intelligence, 2013.

[14] J. Wang, L. Ke, Feature subspace transfer for collaborative filtering, Neurocomputing 136(2014) 1-6.

[15] S. Gong, A collaborative filtering recommendation algorithm based on user clustering and item clustering, Journal of Software 5(7)(2010) 745-752.