The Design and Implementation of Composite Collaborative Filtering Algorithm for Personalized Recommendation

Liang Hu, Wenbo Wang, Feng Wang, Xiaolu Zhang, Kuo Zhao*
Department of Computer Science and Technology, Jilin University, Changchun 130012, China
Email: {zhaokuo, hul}@jlu.edu.cn

***Abstract*—A composite collaborative filtering algorithm for personalized recommend will be presented to solve the original Collaborative Filtering algorithm problem including "None of User Starting " and "Data Sparsity", and the Spearman rank correlation coefficient will be used as a main correlation coefficient. Top-M commended is going to be used to get the final results in this paper. At last, we will validate that this algorithm is superior to the algorithm of collaborative filtering based on user and the algorithm of collaborative filtering based on item.**

***Index Terms*—composite collaborative filtering algorithm; none of user starting; data sparsity; top-M.**

I. INTRODUCTION

With the rapid development and application of today's science and technology, Modern society has been improved significantly. Those traditional recommended technologies have been explored increasingly more disadvantages. Therefore, a special complex recommended technology [1] need to be presented to satisfy the E-commence needs from public.

Recently, a majority of "recommended technology" have been presented, including [2]: Bayesian network association rules, Clustering C-means algorithm, Hording map and Collaborative Filtering. However, "Collaborative Filtering" is one of the most successful.

The Collaborative Filtering technology was first presented by D.Doldberg etc. With the increasingly develop of science and technology, the application of the recommended technology has been used widely. Such as research group of CroupLens, its main products was called Net pererptions, which is using the technology of "time-line" and could recommend the similar interests to other on-line readers(community),it is also needs the assessments of items from readers; Phoaks[3]was developed by Terveenet etc. It will recommend the websites which will be most possible to be interested in to users, and it mainly analyzes the bulletin wrote by users in the Usernet, then finding the recommended URL in the article, counting the number of URL of people recommending and then getting the final recommending results.

The Collaborative Filtering technology [4,5] is the key point for an efficient recommending system. This kind of system can recommend the efficient data, including: the items people might be interested in, the users who have the similar interests. Based on the previous assessments and the similarity assessment[6,7] about other users, the collaborative filtering algorithm will calculate the value of interests first, and input them to get final results.

The related work is introduced in section 2. Section 3 will further discuss "A special complex recommended technology". The experiment results will be analyzed in Section4 and Conclusion is shown in section 5.

II. RELATED WORK

Collaborative Filtering technology has been classified in two categories: A Collaborative Filtering based on user and a Collaborative Filtering based on item.

The Collaborative Filtering based on user [8]: The highlight of the collaborative filtering based on user is the interests' similarity between each user. We can utilize the interests' similarity to find the nearest neighbors whose interests are familiar with the target one. In addition, every neighbor has its own interests. When using this kind of system, every user needs to give assessments about each item after their applying. Finally, the target user will get the recommends from the collection of these interests based on those assessments.

The main fomular[6,9] :

$$\mathrm{sim}(i,j)=\frac{\sum_{k=1}^{n} R_{i,k} * R_{j,k}}{\sqrt{\sum_{k=1}^{n} R_{i,k}^{2} * \sum_{k=1}^{n} R_{j,k}^{2}}} \quad (1)$$

$$P_{u,i} = \bar{R}u + \frac{\sum_{m=1}^{n} \left(R_{m,i} - \bar{R}_{m}\right) * sim(u,m)}{\sum_{m=1}^{n} sim(u,m)} \quad (2)$$

(R_{ik}, R_{jk}:the assessments from user i and user j)
Sim(i,j) represents the similarity of users or items.

Manuscript received August, 2011; revised September 6, 2011; accepted September, 2011.

Corresponding author: Kuo Zhao: phone:86-431-85168716; fax: 86-431-85166494; email: zhaokuo@jlu.edu.cn

doi:10.4304/jsw.7.9.2040-2045

However, the visible disadvantages [7,10]should not be neglect. Firstly, it is impossible that user have a set of neighbor, and then we can recommend through it. Secondly, Data Sparsity[11]. Therefore, the Collaborative Filtering based on data had been produced.

A Collaborative Filtering based on data [12]: the highlight is "data similarity". The first step, Find the set of neighborhood of the target data, and then predict the users' assessment according to all neighbor's assessments. Lastly, the data which belong to the top-M assessment (has been defined earlier) are the final recommends.

Main fomular[10,13]:

$$\mathrm{sim}(\overline{u},\overline{v})=\cos(\overline{u},\overline{v}) \tag{3}$$

$$C=U^{i}_{j=1}C_{j=1}-U^{i}_{j=1}C_{j}\cap U \tag{4}$$

$$\mathrm{sim}(\overline{n}_i U)=\sum_{s=1}^{k} sim(\overline{n}_i,\overline{i}_s) \tag{5}$$

(U:a set of the useful items ;U={ $i_1,i_2,\ldots i_k$ })

(C:the most similar set of items; C = {nl,n2,...n_t})

As we can see form the above, this algorithm cannot find the potential interests [14]. Users can only get the items which are in the same field without any creativity, this greatly decline the flexibility of this system. A new algorithm was presented in this paper in order to deal with this problem of diversity, called "A Personalized Recommendation Technology "[15].

III . A COMPOSITE COLLEBORATIVE FILTERING ALGORITHM

This algorithm mainly combine the collaborative filtering based on user algorithm and the collaborative filtering based on item algorithm together, and using a spearman rank correlation coefficient[16,17,18] as the correlation coefficient not by Pearson to ensure the equal space in logic area of the data, which do not need to be received in pairs from the normal distribution. In this algorithm, the collaborative filtering algorithm predicts a similar item sets by the data had been given, and then use another algorithm to get the final assessment and solve the problem called singular data.

As we can see form the Figure1:The phone is an interest of the target user, the system compute the algorithm using the assessment of this interest to find his or her other interests in the same field, such as carema, cell-phone, computer etc. Then find the potential users who have these interests, these new users must have other interests; all of these new interests are the final information used to recommend to the target user.

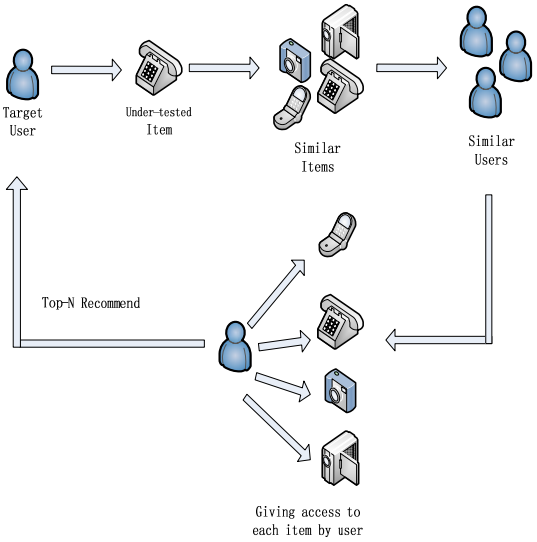

Figure 1. The theory of the composite collaborative filtering algorithm.

The Operating process will be shown below:

Get a milar data set

(a) Calculate the similarity SD_j[19]:

$$\mathrm{Rs}_{j,q}=1-\frac{6\sum d_i^2}{N(N^2-1)}\ (d_i=R_{j,c}-R_{q,c}) \tag{6}$$

(N:number of users;$R_{j,c}$,$R_{q,c}$: the assessment of user j,q)

(b) To compare with all the results above, we use the top-M as the neighborhood SD_j of data-j.

Predict the assessments from user to any items using the similar data.

The main formula is:

$$P_{aj}=\frac{\sum_{m\in Sij} Rs_{j,n}\times r_{a,n}}{\sum_{m\in SIj}|Rs_{j,n}|} \tag{7}$$

Predict the assessments from the target user to the new data.

(a) Get the similarity V(a,i) on SD_j between the target user and any other users.

The formula:

$$V(a,i)=1-\frac{6\sum d_i^2}{N(N^2-1)},d_i=R_{a,k}-R_{i,k} \tag{8}$$

(b) To compare with all the result above, we use the users set of $V_{(a,i)}$ in top-M as the neighborhood of user-a on data.

Get relevated recommends[20].

The main formula is:

$$P_{a,j} = \bar{R}_a + \sum_{k \in NBR_{a,j}} V(a,j)(R_{i,j} - \bar{R}_i) \quad (9)$$

($\bar{R}_i$: average assessment form user i)

The Operating Process[21] is shown below (Figure2):

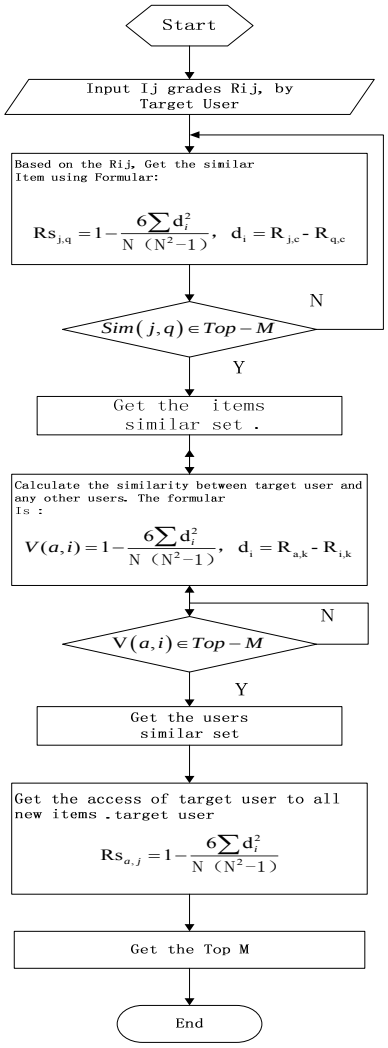

Figure 2. The operating process

IV. EXPRIMENT AND ANALYZING

Expriment environment

Hardware environment: Dual-core CPU,1G,80G

Software environment: windows XP, Java.

Experiment data

The data set is "BookLens data" [22,23,24,25]which is from Minnesota University (http: //grplens.org), it contains 945 Users, 683 Books and a figure of 132046 for assessment.

Access criteria

In this experiment, we use MAE[2,26,27,28] as an access criterion.

$$MAE = \frac{\sum_{i=1}^{N} |Pi - Qi|}{N} \quad (10)$$

P_i: The predicted value of this algorithm;

Q_i: The true value;

N: The number of data

Analyzing data result

The meaning of capital letter:

A: represents a "none of user starting" in composite collaborative filtering algorithm.

B: represents"data sparsity" in composite collaborative filtering algorithm.

C: represents a "Collaborative Filtering based on user"

D: represents a "Collaborative Filtering based on data"

(a) Exq1: User: 356; Book: 178.

The different MAE when the assessment of item is varied. As is shown in the below table1.

TABLE1
U356B178 (MAE)

	0	0.1	0.2	0.3	0.4
None user	0.3587	0.3587	0.3307	0.3080	0.3011
Data sparity	0.2552	0.2552	0.2511	0.2866	0.3224
None user/data sparity	0.3934	0.3934	0.3748	0.3842	0.3520
Item	0.4137	0.4929	0.4962	0.4908	0.4908

Analyze table1 and translate it into a special figure,as we can see from Fig.3.

Generally, The MAE of Collaborative algorithm based on item is superior to the traditional collaborative algorithm when solving the problem of "none user" and the problem of "data sparity". And it is more superior to the algorithm which can only deal with the problem of "data sparity" and the collaborative algorithm which can only deal with "none user".

The fluctuation of MAE: the collaborative algorithm which could solve the problem of both "data sparity" and "none user " fluctuate slightly; while the collaborative algorithm based on user fluctuate significantly when solving either the problem of "data sparity" or the problem of "none user".

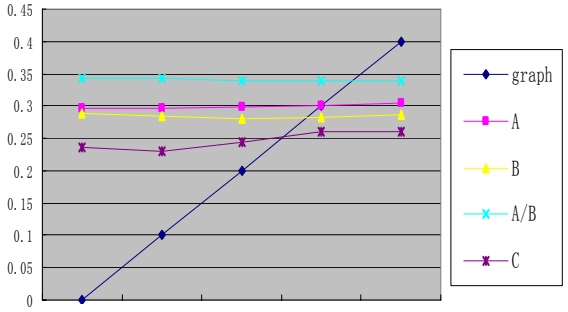

Figure 3. User 356 Book:178

There will be a different MAE when the number of user for assessment is changed. As is shown in the below table2.

TABLE 2
U356B178 (MAE)

	0	0.1	0.2	0.3	0.4
None user	0.2965	0.2968	0.2979	0.3003	0.3049
Data Sparity	0.2883	0.2849	0.2810	0.2828	0.2859
None user/ Data sparity	0.3425	0.3423	0.3392	0.3392	0.3399
Based on User	0.2361	0.2309	0.2442	0.2600	0.2600

Analyze table2 and get the following conclusion:

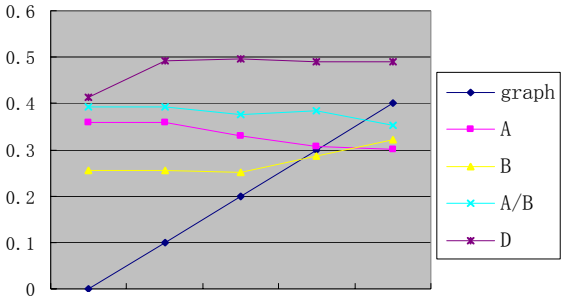

Figure 4. User 356 Book:178

As we can see form the fiture4:

Generally, the algorithm based on item is better than the collaborative algorithm of solving the problem of both “data sparity” and ”none user”, and the collaborative algorithm of solving the problem of “data sparity” and the problem of “none user” respectively.

The fluctuation of MAE: the collaborative algorithm which could solve the problem of both “data sparity” and ”none user“ fluctuate slightly; while the other three algorithms fluctuate significantly.

(b) Exq2: User: 648 Book: 324,.as is shown in Firure5 and Figure6

The different MAE when the assessment of item is varied. As is shown in the below table3

TABLE 3
U648B324 (MAE)

	0	0.1	0.2	0.3	0.4
None user	0.3033	0.3032	0.2812	0.2706	0.2637
Data sparity	0.1903	0.1901	0.1867	0.1886	0.2046
None user/data sparity	0.4112	0.4114	0.4182	0.4345	0.4361
Based on Item	0.4937	0.5755	0.5808	0.5805	0.5794

Analyze table3 and get the following conclusion:

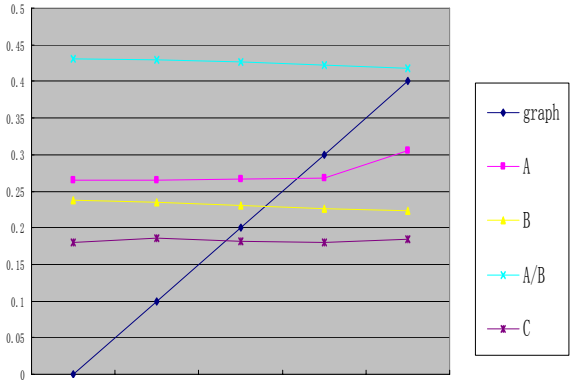

Figure 5. User648 ,Book324

As we can see from Figure5:

Generally, the MAE of collaborative algorithm which can solve”data sparity” and ”none user” are superior to the collaborative algorithm which can solve either “none user” or ”data sparity” and more superior to the algorithm based on item.

The fluctuation of MAE: all of these algorithms are fluctuated slightly

Generate different MAE when assessments of user are varied. As is shown in the below table4.

TABLE4
U 648B324 (MAE)

	0	0.1	0.2	0.3	0.4
None user	0.2652	0.2654	0.2666	0.2676	0.2692
Data spary	0.2379	0.2354	0.2305	0.2264	0.2235
None user/ data spariy	0.4311	0.4298	0.4269	0.4226	0.4182
Based On User	0.1808	0.1862	0.1813	0.1807	0.1841

Analyze table.4 and get the following conclusion:

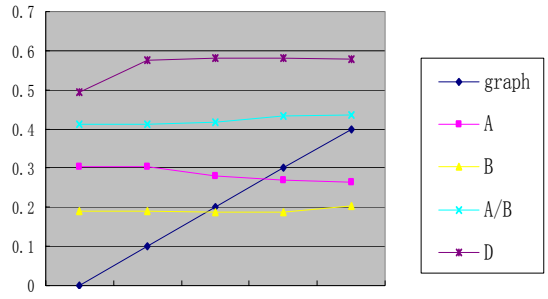

Figure 6. User648, Book324

As we can see from Figure5:

Generally, The MAE of algorithm of based on user is superior to the collaborative algorithm when solving both "none user" and "data sparity" and more superior to the algorithm which can only deal with the "data sparity" and the algorithm which can only deal with "none user".

The fluctuation of MAE: all of these algorithms are fluctuated slightly

Conclusion: with the data being increased, we get the same results.

(c) Exp3: A deeply Comparison between Exp1 and Exp2.

We can get two graphs from the above two experiments. As are shown in Figure7 and Figure 8:

From these two charts, we can get: MAE is declined in both A and B when the data is increasing. A more intuitive chart is shown in Figure 9:

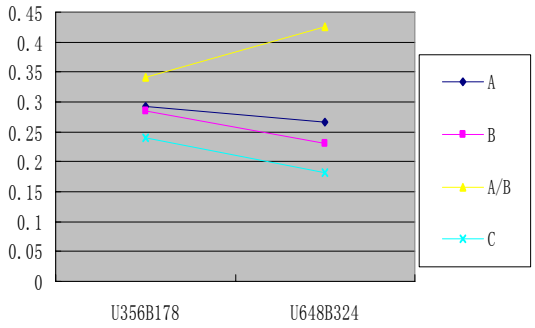

Figure 7. Comparison 1

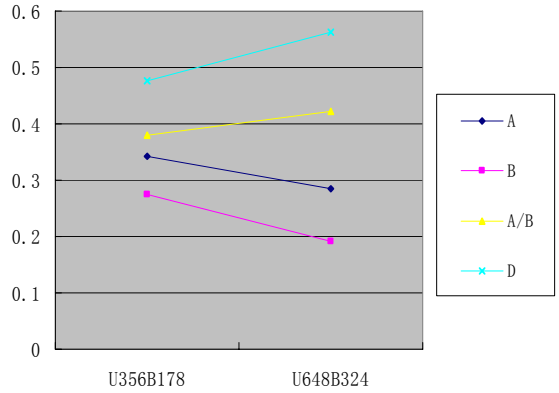

Figure 8. Comparison 2

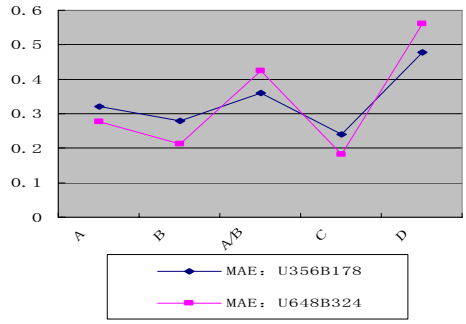

Figure 9. Comparison3

Obviously the largest error-rate of the Collaborative Filtering based on data. Although MAE of Collaborative Filtering based on users is lower than that of our new algorithm, it cannot solve the problem of "none of user starting" and" data sparsity" as we presented in section 2. So the composite collaborative filtering algorithm is valuable to be presented.

V. CONCLUSION

Personalized recommendation technology has been widely applied in various fields, increasingly more efficient recommendation algorithms have been proposed these years. In this paper, the experiment results are based on the restrictive data. In the future we will try to do the experiments on a more optimized data set which will be obtained through cloud computing [30], This will validate the algorithm for further. We will also try to use other related algorithms to optimize this technology, such as quantum behavior of micro-group optimization [29,30], fuzzy C means clustering algorithm [31].

ACKNOWLEDGMENT

This work was supported in part by the National Grand Fundamental Research 973 Program of China under Grant No. 2009CB320706, the National High Technology Research and Development Program of China under Grant No. 2011AA010101, the National Natural Science Foundation of China under Grant No. 61103197 and 61073009, Program of New Century Excellent Talents in University of Ministry of Education of China under Grant No. NCET-06-0300, the Youth Foundation of Jilin Province of China under Grant No. 201101035, and the Fundamental Research Funds for the Central Universities of China under Grant NO.200903179.

REFERENCES

[1] Pretschner, "A. Ontology based personalized search [MS. Thesis. Lawrence", KS: University of Kansas, 1999

[2] DENG Ailin, ZHU Yangyong,SHI Bole, "An algorithm Collaborative Filtering based on item[J]",Journal of Software,2003,14(9),pp.1621-1624

[3] Wang F H, Shaob H M. "Effective personalized recommendation based on time-framed navigation

clustering and association mining" , Expert systems with applications, 2004, 27(3),pp.365-377
[4] Ma H, King I, Lyu M R, "Effective missing data prediction for collaborative filtering" , In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York: ACM, 2007, pp.39-46
[5] LI Tao, WANG Jiandong, "Collaborative filtering recommendation algorithm based on clustering basal users", Systems Engineering and Electronics,2007,29(7),pp.1178-1182
[6] Al-Shamri M Y H, Bharadwaj K K, "Fuzzy-genetic approach to recommender systems based on a novel hybrid user model" ,Expert Systems with Applications, 2007, 35(3),pp.1386-1399
[7] LI Cong, LIANG Changyong, DONG Ke, "A collaborative filtering recommendation algorithm based on item category similarity", Journal of Hefei University of technology,2008,31(3),pp.360-363
[8] WENG Xiaolan, ZHUANG Yonglong, "A collaborative filtering algorithm based on clustering of items' character" , Computer Applicationns and Software,2009,26(7),pp.260-262.
[9] D. Quade,I. A. Salama, "Concordance of Complete or Right-Censored Rankings Based on Spearman's Footrule" ,Communications in Statistics - Theory and Methods,2006,35(6),pp.1059-1069.
[10] Conor V. Dolan, "Investigating Spearman's Hypothesis by Means of Multi-Group Confirmatory Factor Analysis", Multivariate Behavioral Research, 2000 25(1),pp 21-50.
[11] Chen Huayue Chen Huayue, "Personal Recommendation Based on the Weighted Association Rules and browser behavior(C) ".
[12] Roh T H, Oh K J, Han I, "The collaborative filtering recommendation.
[13] Based on SOM cluster-indexing CBR". Expert Systems with Applications, 2003, 25(3),pp. 413-423.
[14] LIU Yang, "A Kind of Adaptive Recommend System Based on Collaborative Filtering Technology", Journal of liaoning university of petroleum & chemical technology, 2007,27(3),pp.75-78.
[15] Sun Shouyi,Wang Wei, "A personalized book recommendation system ",modern intelligence,2007,11(11)
[16] Zhang Bingqi, "The Representation, Acquisition and Inference of Personalized Requirements: A Case Study(C)",institute of computing technology chinese academy of science.
[17] Deng Ailin, Zhu Yangyong, Shi Bole, A Collaborative Filtering Recommendation Algorithm Based on Item Rating Prediction,Journal of Software,2003,14(9).pp.1622-1628.
[18] Zhang Zhongping, Guo Xianli, Optimized collaborative filtering recommendation algorithm based on item rating prediction.Applicationresearchofcomputers,2008,25(9),pp. 2659-2683
[19] Breese J,Hecherman D,Kadie C.Empirical analysis of predictive algorithms for collaborative filtering.In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence(UAI'98).1998.pp.43-52
[20] Sarwar B, Karypis G, Konstan J, Ried1 J. Analysis of recommendation algorithms for E-commerce. In:ACM Conference on Electronic Commerce.2000.pp.158-167
[21] Hu huirong, Personalized recommendation system, Science and Technology innovation Herald,2009(08),pp.177-203
[22] Zheng xianrong,Tang Zeying, Cao Xianbin; Non-lineal gradual forgetting collaborative filtering algorithm capable of adapting to users' drifting interest, Computer aded engneerng.2007.16(2),pp.69-73
[23] Huang Liming,Wu Xiaojun, Wang Shitong,A Study on Fuzzy Clustering Based on Multistage FCM Algorithm,Journal of nanjing normal university(natural Science).1999,22(4)
[24] K.Alsabti,S.Ranka and V.Singh. An efficient K-means clustering algorithm.
[25] IPPS/SPDP Workshop on High Performance Data Mining. Orlando,Florida,1998.
[26] J.an, M.Kamber. Data mining: concept and technique. Morgan Kaufmann Publishers, 2000.
[27] M.Ester, J.Sander. A density-based algorithm for discovering clusters in large spatial databases. Proc. 1996 Int. Conf. Knowledge Discovery andData Mining. Portland, USA, Aug 1996:226-331.
[28] S-H.Min, I.Han, Detection of the customer time-variant pattern for improvingrecommender systems.Expert Systems with Applications. 2005, 28(2):189-199.
[29] L.Terveen,W.Hi11:A system for sharing recommendations. Communications of the ACM.1997, 40(3): 59-62.
[30] M.Balabanovic, Y. Shoham. Combining content-based and collaborative recommendation. Communications of the ACM. 1997, 40(3):66-72.
[31] S.W.Changchiena, C-F.Lee,Y.-J.Hsu, On-line personalized sales promotion in electronic commerce, Expert Systems with Application. 2004, 27(1):35-52

Liang Hu was born in Changchun, China, on Feb. 1968. He received his Ph.D. degree in computer software and theory from Jilin University of China in 1999.

He is currently a professor in the Department of Computer Science and Technology, Jilin University, China. As an author or co-author, he has published more than 180 research papers, including 80 journal papers, 8 books, and has 1 National invention patent of China as well as 18 computer software copyright registration certificates of China. His research interest covers network security and distributed computing, including related theories, models, and algorithms of PKI/IBE, IDS/IPS, and grid computing.

Kuo Zhao was born in Tonghua, Jilin Province, China, on Jan. 1997. He received the B.E degree in computer software in 2001 form Jilin University, China, followed by M.S degree in computer architecture in 2004 and Ph.D. in computer software and theory from the same university in 2008.

He is currently associate professor in the Department of Computer Science and Technology, Jilin University, China. As an author or co-author, he has published more than 40 research papers, including 15 journal papers, and has 1 National invention patent of China as well as 10 computer software copyright registration certificates of China. His research interests are in operating systems, computer networks and information security. He is the corresponding author of this paper.