International Journal of Hybrid Information Technology
Vol.9, No.3 (2016), pp. 1-10
http://dx.doi.org/10.14257/ijhit.2016.9.3.01

Research on Information Entropy Measure based on Collaborative Filtering Algorithm

Jingxia Guo[1] and Jinggang Guo[2]

[1]Bao Tou Medical College, BaoTou 014060, china
[2]Inner Mongolia press and Publication Bureau, Hohhot 010050, china
[1]Guojing7223113@163.com and [2]112680605@qq.com

Abstract

Most existing calculations of similarities suffer from data sparsity and poor prediction quality problems. For this issue, we proposed a similarity measurement algorithm based on entropy. The entropy is computed by the difference of two users' ratings, and we also consider the size of their common rated items, the size is bigger, the weight of their similarity is higher. Experiments show that the algorithm effectively solves the problem of the inaccuracy of similarities in data sparsity or small size neighborhood environments, and outperforms other state-of-the-art CF algorithms and it is more robust against data sparsity.

***Keywords**: Data mining, Recommendation Systems, Collaborative Filtering, Similarity Measure*

1. Introduction

In the collaborative filtering algorithm based on memory, the most popular is collaborative filtering algorithm based on a nearest neighbor [1-5]. In the use of such method to make recommendations to user, it involves these steps:

(1) Collect information which represents user's interest, such as purchase records, user rating of item etc [6-8].;

(2) Look for similar users, calculate user similarity through their common evaluation data and find the nearest neighbors having similar interest as target users' [9-10];

(3) Generate recommendations to target users in the nearest neighbor set. Apparently, similarity calculation is foundation and core to the whole collaborative filtering algorithm. The selection of appropriate similarity measure method is significant to the entire collaborative filtering approach [11-15].

In the algorithm, what's used mostly so far is correlation coefficient similarity (Pearson correlation, Spearman correlation) [16-18] and cosine similarity. Although they consider diversity of user assessment standards, such similarity measuring methods still have shortcomings in the application of the collaborative filtering:

(1) In the case of improving high-dimensional sparseness, the scale K of the intersection of concern circle among users (jointly rated item) is generally smaller and not identical; traditional similarity measuring methods can't accommodate to the situation, easily overestimating or underestimating the real similarity among users;

(2) Lower recommendation precision due to data sparseness.

Table 1 lists out user's scorings of items. The scoring scale of every user is [1-5]. For user u1 and u2, it needs to find out firstly the rating u1: (1,2,1,2,1)and u2: (4,5,4,5,4)of their commonly appraised items; Then utilize Pearson correlation to compute their similarity Sim(u1,u2)=1. Based on correlation coefficient, they are completely alike. However, user u1 marked lowly those items while user u2 rated highly, therefore, their similarity is not so high. For user u2: (4,5,4,5,4) and u3: (5,4,5,4,5), employ Pearson correlation to reckon their similarity Sim(u2,u3)=-1, totally negative correlated, but

ISSN: 1738-9968 IJHIT

actually they have very strong similarity because they both rated highly those items. For instance, we determine u1and u5 which is closer to u4. First of all, for user u1 and u4, they rated commonly only one item and the score is 5; at this moment, their Pearson correlation is Sim(u1,u4)=1; for user u5 and u4, their commonly evaluated item is u5(2,3,4,3) and u4(2,3,4,5); use Pearson correlation to compute their similarity Sim(u5,u4)=0.67. From correlation coefficient, we find u1 is much closer to u4; but in fact, u1 and u4 assessed commonly only one item; u5 and u4 did four items, of which three ratings are identical. Hence, u5 and u4 should be more similar. Besides, if one user's ratings of all items keep the same, such as u:(2,2,2,2,2) and v:(4,4,4,4,4), with traditional cosine vector similarity measuring method, it's impossible to get accurate similarity between them. Those weaknesses cause lower precision of recommendation in light of data sparsity.

Table 1. Sparse Evaluation Table of User-Item

User	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	…
U1	1		2	1		2	1	5			
U2	4	3	5	4	2	5	4		1		
U3	5		4	5		4	5	1		2	
U4		2			3			5	4	5	
U5		2			3				4	3	
…	…	…	…	…	…	…	…	…	…	…	

2. Existing Solutions

To address those problems, researchers stated to improve those similarity measuring methods, enhancing calculation accuracy while the sparsity keeps the same. J. A. Konstan, *et al.*, [19-20] suggested MAX and MIN [21] improved approaches which are both based on Pearson correlation coefficient, considering the effect of the number of jointly rated items by users on similarity calculation results. It is shown in formula 1 and formula 2.

$$corr_{u,v}^{'} = \frac{\max(|K_u \cap K_v|, \gamma)}{\gamma} \times corr_{u,v} \tag{1}$$

$$corr_{u,v}^{'} = \frac{\min(|K_u \cap K_v|, \gamma)}{\gamma} \times corr_{u,v} \tag{2}$$

The paper [22] presented a new similarity measurement method. It discussed from the aspect of Proximity, influence and popularity (PIP in short) the impact of user rating on similarity among users, alleviating data sparseness in traditional similarity measuring methods and upgrading the recommendation precision. PIP's calculation method is shown as:

$$PIP(r1, r2) = \Pr oximity(r1, r2) \times \operatorname{Im} pact(r1, r2) \times Popularity(r1, r2)$$

The final two user similarity:

$$Sim(u_i, u_j) = \sum_{k \in Ci, j} PIP(r_{ik}, r_{jk}) \tag{3}$$

Moreover, some scholars reduced the coefficient degree of matrix by matrix filling or adding user or item attribute, for the purpose of higher recommendation accuracy. Zhang Guangwei, *et al.*, used similarity calculation method based on cloud model. With its bridge role in qualitative knowledge representation, qualitative and quantitative knowledge transformation, they designed a method to compare user similarity in knowledge level, overcoming deficiencies of traditional methods based on vector. Literature [23] showed a new similarity measuring method based on fuzzy similarity priority comparison. It used similarity priority ratio to search similar users and predicted their marks of unrated items according to item similarity, for enhanced recommendation precision. Inspired from collaborative filtering algorithms based on user and item, [24] considered both user and item similar information while predicting ratings of missing items, finding the similar user set and similar item set of those missing items. Peng Yu, *et al.*, proposed a collaborative filtering recommendation algorithm based on item. By calculating item's rating similarity and also attribute similarity, it made use of two-way implicative predicates to compute item similarity. However, those aforesaid approaches require often the recommendation system to provide some additional information like user personal information, item attribute etc. None of them is the optimal solution because, on one hand, those data are hardly acquired; on the other hand, it increases computational complexity and workload of the system.

The paper proposed the similarity measuring method based on user rating weighed difference entropy, *i.e.*, NWDE. By calculating differences of user ratings and considering the size of user's concern circle intersection, it uses weighted information entropy to measure similar degree of user scoring. This method doesn't require user or commodity's other attribute information. With unchanged data sparsity, it mitigates overestimation or underestimation of user similarity which is found in traditional similarity measuring method, improving the precision of recommendation.

3. Similarity Measuring Technique based on Entropy

3.1 Motivation of this Proposal

Information entropy is used to measure the randomness or dispersion degree of distribution. The distribution is more scattered, *i.e.*, more even, and that the entropy is bigger; the distribution is more orderly, *i.e.*, more concentrated, and that the entropy is smaller. For the given sample set X, its information entropy is acquired by this formula:

$$H(X) = \sum_{i=1}^{n} p(a_i) \log_2 \frac{1}{p(a_i)} \tag{4}$$

Hence, by introducing entropy to collaborative filtering similarity measuring field, we can weigh the dispersion degree of rating differences among diverse users or items. If the entropy of two user's scoring difference is smaller, implying that their differential degree is lower and similar degree is higher. In extreme cases, if all data of two user's grade difference are made the same value, the entropy is 0; conversely, bigger entropy indicates higher difference degree and lower similarity.

In the recommendation process, one user is more likely to accept opinions of others in the same circle of concern. Suppose user u's concern circle (i.e. rating item set) is I_u, user v's is I_v, and the crossed set of the two is $I_u \cap I_v$. If the intersection set is bigger, meaning the two users are more unanimous in the circle and thus they're more likely to accept mutual opinions and more interdependent. In consideration of two user similarity, it's necessary to consider the scale of their intersected circle of concern.

3.2 Algorithm Design

The new algorithm has three steps:

Step one: Evaluate rating difference between two users

Assume jointly rated item set I of user i and j; their commonly rated data is respectively $U_i = \{R_{Ui,I1}, R_{Ui,I2}, R_{Ui,I3}, ..., R_{Ui,In}\}$ and $U_j = \{R_{Uj,I1}, R_{Uj,I2}, R_{Uj,I3}, ..., R_{Uj,In}\}$. The difference $Diff(U_i, U_j)$ between their rating data can be defined as:

$$\begin{aligned} Diff(U_i,U_j) &= \{R_{Ui,I1}-R_{Uj,I1}, R_{Ui,I2}-R_{Uj,I2}, ..., R_{Ui,In}-R_{Uj,In}\} \\ &= \{d_1, d_2, d_3, ...d_n\} \end{aligned} \tag{5}$$

Step two: Calculate weighted information entropy

Firstly, utilize information entropy equation to calculate the entropy value of $Diff(U_i,U_j)$:

$$H(Diff(U_i,U_j)) = \sum_{i=1}^{n} p(d_i)\log_2\left[\frac{1}{p(d_i)}\right] = -\sum_{i=1}^{n} p(d_i)\log_2 p(d_i) \tag{6}$$

In calculating information entropy $H(Diff)$ of $Diff(U_i,U_j)$, the magnitude of rating difference value d_i id reflects varied user similarity; bigger d_i id suggests higher user difference. Thus, we need to modify equation 6 and exert one weight $|d_i|$ as to compute information entropy. Additionally, regarding the size of user's intersected concern circle, we should add to equation 6 the weight 1/n which represents the size of such intersection. For now, the weighted difference entropy $(WDE(U_i,U_j))$ between user i and j is:

$$WDE(U_i,U_j) = -\frac{1}{n}\sum_{i=1}^{n} p(d_i)\log_2 p(d_i)\times|d_i| \tag{7}$$

Step three: WDE normalized to [0,1]

From equation 7 we learn that WDE_{Ua} element value range is from zero to positive infinity. It's required to standardize WDE_{Ua} . Meanwhile, bigger $WDE(U_i,U_j)$ indicates bigger user difference and smaller similarity. So we have to use the following extreme linear model to normalize elements of WDE_{Ua} .

$$NWDE_{Ua}[i] = \frac{Max(WDE_{Ua}) - WDE_{Ua}[i]}{Max(WDE_{Ua}) - Min(WDE_{Ua})} \tag{8}$$

Pseudo code of algorithm is described as follows:

Algorithm 1 NWDE similarity measure

Input: the original user rating data $R_{m\times n}$

Output: the similarity matrix $R_{m\times m}$

```
BEGIN:
    Constant length;
    Preference array x[length];
    Preference array y[length];
    Difference set D [length];
    For(int i=1;i<=sizeof(x);i++)
        D [i] = abs(x[i]-y[i]);
    For(int i=1;i<=sizeof(D);i++)
        WDE+= - p(d[i])*log2(p(d[i]))*|d[i]|;
    WDE = WDE/ sizeof(D);
    Return normalization WDE  (NWDE);
    Repeat until all pairs of users is calculated;
END.
```

4. Experiment Design and Discussion

4.1 Experimental Data

For the experiment, we got data from MoveiLens dataset collected by GroupLens Research group in Minnesota University [25]. The dataset have been so far used mostly for studies on the collaborative filtering recommendation system. MovieLens site is used to receive user's marks of movies and provide accordingly movie recommendation list. Its rating scale is integer from 1 to 5. Higher rating value means user's stronger interest in the movie; otherwise, less/no interest.

MovieLens provides downloading of dataset of three magnitudes: 100,000 comments by 943 users on 1682 movies; 1,000,000 comments by 6064 users on 3900 movies and 10,000,000 comments by 71567 users on 10618 movies. In the experiment, we chose the first data as experimental dataset, of which each user evaluated at least 20 movies. The sparsity of the dataset [26] is 1-100000/(943*1682)=0.93, very discrete. We divided randomly the data to training set and testing set, of which training set is 80% and testing set 20%.

4.2Experimental Evaluation Strategy

The paper firstly confirmed whether NWDE method can effectively eliminate user similarity distortion when data is sparse; then, examined if it can reach better recommendation precision. We used mean absolute error (MAE) and three evaluation criteria commonly used in information retrieval and personalized recommendation field: precision rate, recall ratio and F –measure [27-29].

Precision rate is defined as: in Top-N recommendation results, the percentage of correctly recommended items, in the expression:

$$precison = \frac{\text{The correct number of recommended}}{\text{The total recommended}} = \frac{|test \cap TopN|}{N} \tag{9}$$

Recall rate is the percentage of correctly recommended items in the whole testing dataset, acquired by:

$$\mathrm{Re}\,call = \frac{\text{The total number of correct recommendation}}{\text{The total number of test set}} = \frac{|test \cap TopN|}{|test|} \tag{10}$$

In some cases, the recommendation system's precision and recall ratio are both high or low, indicative of different perspectives. F-measure combines both precision and recall rate. It can reflect them equally. It can be obtained by the equation:

$$F-measure = \frac{2 \times Precison \times \mathrm{Re}\,call}{Precison + \mathrm{Re}\,call} \tag{11}$$

Apparently, the smaller MAE is, the better quality the recommendation reaches; higher precision and recall ratio and bigger F-measure suggests better quality of recommendation; otherwise, it's worse.

4.3 Experimental Results and Discussion

First of all, in MovieLens dataset, we used respectively Pearson correlation, Spearman correlation, cosine similarity and the proposed NWDE technique to calculate the similarity between the first and other users. Results are shown in Figure 1.

In Figure 1, axis X refers to user ID and axis Y means similarity score. We find from it cosine similarity has concentrated scores, mostly in the range [0.8, 1]. But, user similarity difference value is not big, which makes it difficult to discern clearly neighboring users with uniform preference as target users. Pearson correlation and Spearman correlation have very close results, in the range [-1, 1]. The new method here can distinguish better user similarity.

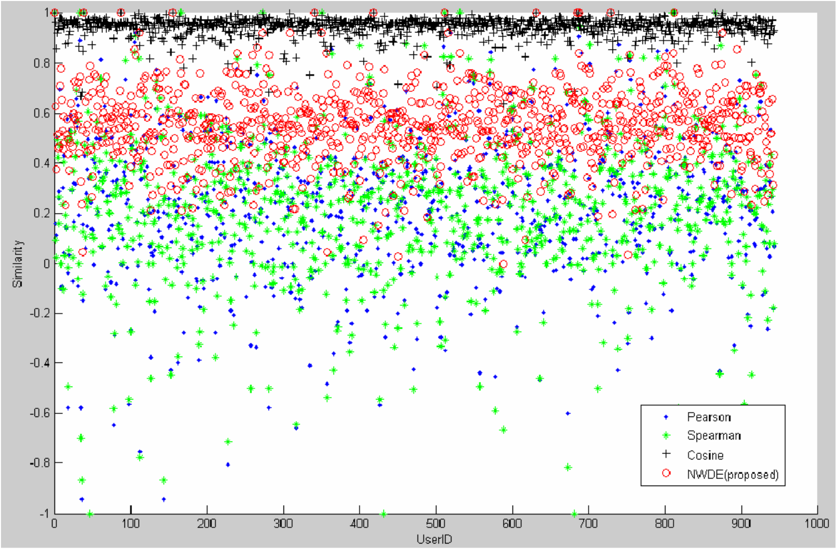


Figure 1. Different Similarity Measure Score

Besides, in the case of sparse dataset, the proposed algorithm alleviates the influence of overestimation or underestimation of user similarity. We take for instance, the common rating data of the first and the 866 user is u1: (5,5,3) and u866: (3,3,2), between them the similarity is 1 with traditional Pearson or Spearman correlation. Obviously their real similarity is not so high as 1. With NWDE, the similarity is 0.7899, to a certain degree, eliminating overestimation of similarity. Moreover, the common rating data of the first and the 47 user is u1:(5, 3, 5, 5) and u47:(4,5,4,4), between them the similarity is -1 with Pearson or Spearman correlation. Clearly, they rated too high and the similarity is not so low. With NWDE, the similarity is 0.7894, more accordant to the real similar degree.

In the following, we discussed the recommendation precision of the proposed NWDE technique. In testing dataset, we predicted user rating by choosing different neighbor numbers to get respectively MAE value of each measuring method.

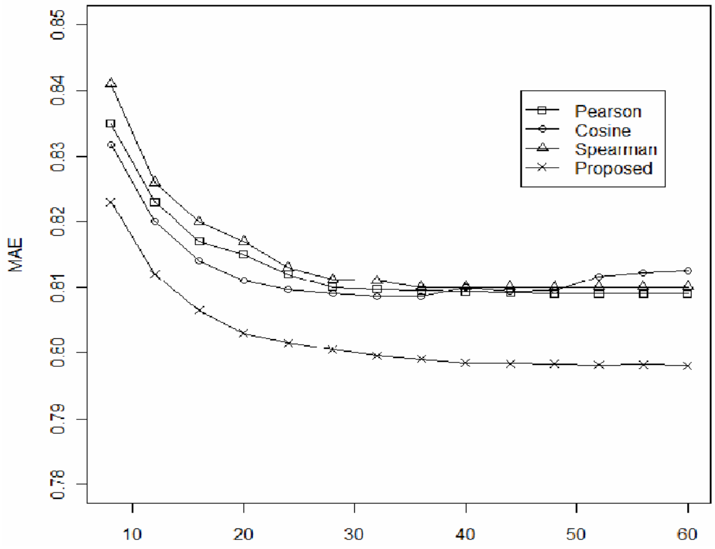


Figure 2. The Similarity Measure Method for the Accuracy of Recommendation

Figure 2 shows the size of MAE values for different neighbor numbers with Pearson correlation, Spearman correlation, cosine similarity and NWDE measuring methods. We chose 8,12,16,20,24,28,32,36,40,44,48,52,56,60 different numbers of neighbors. From the picture, we know that with more neighbors, Pearson correlation, Spearman correlation and cosine similarity reached approximate MAE value; while NWDE here got significantly low MAE value, with enhanced recommendation precision.

In addition, to validate the quality of recommendation results, we evaluated those methods in terms of accuracy rate, recall ratio and F-measure. Results are put in Table 2.

Table 2. The Similarity Measure Method for the Quality of Recommendation

similarity measure method	Precision	Recall	F-measure
Pearson	0.3237	0.1686	0.2288
Spearman	0.3156	0.1766	0.2314
Cosine	0.3015	0.1614	0.2166
NWDE	0.3224	0.1795	0.2365

From the Table 2, except Precision, NWDE got the second best marks; its Recall and F-measure got the best scores. On the whole, NWDE realized the best quality of recommendations. Simultaneously, compared to precision rate, four methods had lower recall ratio, perhaps being affected by the magnitude of both N value and testing dataset.

5. Conclusion

In collaborative filtering algorithm based on memory, the most widely used similarity measurement method for similarity based on correlation coefficient (Pearson and Spearman) and similarity based on cosine vector. However, there are still some drawbacks of these traditional similarity measure methods:

In the case of high dimensional sparse data, users' the circle of concern the intersection of scale K is mostly small and inconsistent. The traditional similarity measure methods cannot adapt to this situation, easy to exaggerate or reduce the similarity of users. For this issue, the proposed a similarity measurement algorithm based on entropy. The algorithm is not the help other information users or items, the difference between users by calculating the score, used the information entropy of an adjustment to measure different user rating similarity. At the same time, in the calculation of user similarity considered users the size of common concern circle. The size is bigger, the weight of their similarity is higher. Experiments show that the algorithm effectively solves the problem of the inaccuracy of similarities in data sparsity or small size neighborhood environments, and outperforms other state-of-the-art CF algorithms and it is more robust against data sparsity.

References

[1] X. Peiyong, "Collaborative filtering algorithm based on personalized recommendation technology", Ocean University of China, **(2011)**.

[2] S. Shan, "The global and local similarity research on Collaborative Filtering Recommendation Based on the measure of Chongqing University", **(2014)**.

[3] Z. Xiaohong, "Research on collaborative similarity measurement method in filtration", The radio communication technology, vol. 01, **(2013)**, pp. 94-96.

[4] D. Ailin, Z. Yangyong and S. bole, "A collaborative filtering recommendation algorithm based on item rating prediction", Journal of software, vol. 09,, **(2003)**, pp. 1621-1628.

[5] Z. Cuicui and L. Lin, "A collaborative filtering algorithm in similarity measure method research", Computer engineering and applications, vol. 08, **(2014)**, pp. 147-149.

[6] W. Junho and S. Shan, "An improved collaborative filtering recommendation algorithm of similarity measure", Computer science, vol. 05, **(2014)**, pp. 68-71.

[7] X. Chao, "The sparsity of data collaborative filtering algorithm research and implementation oriented", Beijing University of Posts and Telecommunications, **(2013)**.

[8] X. Ji and L. Yanbing, "A collaborative filtering method of similarity measure based on user interest degree", Computer application, vol. 10, **(2010)**, pp. 2618-2620.

[9] Z. Z. Lan, "Study on Personalized Recommendation Algorithm Based on collaborative filtering", Huazhong Normal University, **(2009)**.

[10] L. Chenghua, "Research on collaborative filtering model and user characteristics based on interest degree", Tianjin University of Finance Economics, **(2012)**.

[11] P. Yu, "Collaborative filtering recommendation multi project content based on personal characteristics of users", Southwestern University, **(2007)**.

[12] Y. Son, "The research and design of personalized recommendation system based on user interest", Beijing Jiaotong University, **(2008)**.

[13] W. Huiping, "The algorithm of recommendation similarity and user interest based personalized project categories", Taiyuan University of Technology, **(2008)**.

[14] Z. Yonglong, "Algorithm of collaborative filtering recommendation project based on feature model", Nanjing University of Science and Technology, **(2008)**.

[15] Z. Guangwei, L. Deyi, L. Peng, K. Jianchu and C. Guisheng, "A collaborative filtering algorithm based on cloud model", Journal of software, vol. 10, **(2007)**, pp. 2403-2411.

[16] K. Goldberg, T. Roeder, D. Gupta and C. Perkins, "Eigentaste: a constant time collaborative filtering algorithm", Information Retrieval, vol. 4, no. 2, **(2011)**, pp. 133–151.

[17] J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, "Evaluating collaborative filtering recommender systems", ACM Transactions on Information Systems, vol. 22, no. 1, **(2004)**, pp. 5–53.

[18] J. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, "Evaluating collaborative filtering recommender systems", ACM Transactions on Information Systems (TOIS), vol. 22, no. 1, **(2004)** January, pp. 5-53.

[19] M. Hao, K. Irwin and R. L. Michael, "Effective missing data prediction for collaborative filtering", Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, **(2007)** July 23-27.

[20] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction, MIT Press, Cambridge, Mass, USA, **(1998)**.

[21] "A. H. N. H. J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem", Information Sciences: an International Journal, vol. 178, no. 1, **(2008)**, pp. 37-51.

[22] Z. Guangwei, L. Deyi, L. Peng, K. Jianchu and L. Hesong, "Collaborative filtering recommendation algorithm based on cloud model", Journal of software, vol. 18, no. 10, **(2007)**, pp. 2403-2411.

[23] G. Songjie, "Calculation of personalized recommendation and a new similarity method", Application of computer systems, vol. 17, no. 07, **(2008)**, pp. 87-89.

[24] H. Ma, I. King and M. R. Lyu, "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.

[25] "MovieLens Dataset", http://MovieLens.umn.edu/.

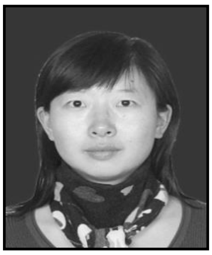
[26] D. Ailin, Z. Y. Yong and S. bole, "Collaborative filtering algorithm", Journal of software, based on item rating prediction, vol. 14, no. 9, **(2003)**, pp. 1621-1628.

[27] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce", ACM Conference on Electronic Commerce, **(2000)**, pp. 158-167.

[28] Y. Shih and R. Liu, "Hybird recommendation approaches: collaborative filtering via valuable content information", In proceedings of 38th Hawaii International Conference on System Sciences, **(2005)**, pp. 217b.

[29] Z. Huang, H. Chen and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering", ACM Transactions on Information Systems, vol. 22, no. 1, **(2004)**, pp. 116-142.

Authors



Jingxia Guo, She received her M.S degree from Inner Mongolia Normal University. She is a lecturer in Bao Tou Medical College. Her research interests include Data mining.

Jinggang Guo, He received his B.S degree from Xi'an University of Posts and Telecommunications. He is a engineer in Inner Mongolia press and Publication Bureau. His research interests include Data mining.