International Journal of Grid and Distributed Computing
Vol. 9, No. 10 (2016), pp.291-300
http://dx.doi.org/10.14257/ijgdc.2016.9.10.26

Collaborative Filtering Recommendation Model Based on Normalization Method

Gao Yan

Information Engineering School, Yulin University, Yulin 719000, Shaanxi, China
yulingaoyan@126.com

Abstract

In view of traditional collaborative filtering recommendation does not take into account differences of dimension of user vector and value of evaluation, this paper proposed a collaborative filtering recommendation model based on normalization method. Before calculating the user's or item's similarity, the value of evaluation will be normalized to a range of specifications. Then the similarity of user vector will be calculated, and predictions and recommend will be made. The experimental results show that this model could accurately find similar neighbor users or items, and performances of prediction and recommendation have been largely improved.

Keywords*: normalization, collaborative filtering, similarity measurement*

1. Introduction

With the rapid development of computer network technology, human's activities are increasingly dependent on Internet. People can buy all kinds of goods, obtain information about life, and seek cooperation and assistance through network. But in the face of huge amounts of Internet information, ordinary users find that it becomes more and more difficult to get timely and effective information. Under this circumstance, how to provide efficient personalized information recommendation is particularly important [1].

Collaborative filtering algorithm is one of the most successful recommendation algorithms, and has been widely used in many fields. According to user's interest in hobbies and historical data, collaborative filtering could find neighbors that have some similarities in these aspects, and make recommendation according to neighbor's action and evaluation data. Different with content based recommendation systems, collaborative filtering algorithm is only looking for neighbors that have the similarity evaluation based on user evaluation of a project, ignores data details, and does not extract the project's text feature vectors [2-4]. Collaborative filtering recommendation algorithm is generally divided into two categories: memory based collaborative filtering and model based collaborative filtering. Memory based collaborative filtering algorithm is the most widely used recommendation method, and can be further divided into user based collaborative filtering recommendation and item based collaborative filtering recommendation. User based collaborative filtering recommendation predicts current user's evaluation of a project according to the evaluation of neighbor users on the same project. Item based collaborative filtering recommendation predicts the current item's evaluation according to the evaluation of neighbor item [5-7].

For collaborative filtering recommendation algorithm, the core is how to compute similarity between users or items. However, in the traditional similarity calculation process, differences in the length of vector space between user ratings vectors or project score vectors cannot be handled properly, and numerical differences between common scoring items of neighbor user or multiple users' score on neighbor item are neglected or weakened. In order to solve this problem, this paper proposes a collaborative filtering algorithm based on normalized data processing (NDCF). Before calculation of users'

ISSN: 2005-4262 IJGDC

similarity, user evaluation of item is normalized to a unified range. Then, by using the user based collaborative filtering recommendation algorithm, normalized user vector in different vector space is measured by user's similarity. Therefore, the nearest neighbor is formed, and normalized item rating prediction is made. Finally, according to current user's evaluation on other items, the item's final predict score is given.

2. Principle of Traditional Collaborative Filtering Recommendation

The most popular and most widely used collaborative filtering recommendation method is the recommendation algorithm based on memory. Taking classical user based collaborative filtering recommendation algorithm as an example, its basic idea is to find k neighbors who have highest similarity with current user, then according to these neighbors' evaluation on an item to predict current user's evaluation on the same item [8-10]. This recommendation algorithm can be divided into three steps: data definition, user similarity calculation, item rating prediction and recommendation.

2.1. Data Definition

Data definition means definition of user rating matrix R. R is an n*m matrix, where n and m respectively represent the number of users and number of items in this matrix. The values of matrix elements represent the users' evaluation values of items, including explicit and implicit rating score. Explicit rating generally refers to user's directly evaluation value on items, implicit rating is predicted value based on user's browsing behavior, such as browsing frequency, time, purchase and other indirect values [11-13]. Matrix user vector (horizontal vector) represents a specific user evaluation of all items; item vector (vertical vector) represents a specific item's evaluation value of all users. If there is evaluation r_{xy} which represents user u_x's evaluation on item i_y, then there will be R[x,y]=r_{xy} in the rating matrix. User rating matrix R is shown as Formula (1).

$$R = \begin{Bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & \cdots & \cdots & r_{nm} \end{Bmatrix} \tag{1}$$

2.2. Similarity Calculation

The essence of collaborative filtering recommendation is search method based on nearest neighbor, and the core is to measure similarity between users or items [14]. Commonly used similarity methods of calculation are vector space similarity measure method(VSS) and Pearson correlation coefficient (PCC). Vector space similarity measure method is to make the user rating of item as an n-dimensional vector, and the similarity between users is determined by the cosine between two user vectors. The similarity value SIM(u,v) between user u and v is calculated by Formula (2). Where ,I_{uv} represents a collection of common rating items between user u and user v.

$$Sim(u,v) = \frac{\vec{R}_u \cdot \vec{R}_v}{\left\|\vec{R}_u\right\|_2 \times \left\|\vec{R}_v\right\|_2} = \frac{\sum_{i \in I_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{i \in I_{uv}} r_{ui}^2} \sqrt{\sum_{i \in I_{uv}} r_{ui}^2}} \tag{2}$$

Pearson correlation coefficient is used to measure linear relationship between variables, considering the difference of user ratings based on vector space similarity measure method, and using user ratings deviation as the score in the calculation. The similarity

value Sim(u,v) between user u and v is calculated by Formula (3). Where, $\overline{u_r}$, $\overline{v_r}$ respectively represent average score user u and user v rated items. I_{uv} represents a collection of common rating items between user u and user v.

$$Sim(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})^2 \quad (r_{vi} - \overline{r_v})^2}} \tag{3}$$

2.3. Item Rating Prediction and Recommendation

After calculating the similarity value between any two users using Formula (2) or Formula (3), N neighbor users with the most similarity value could be got for target user. For an item which has not been scored by target user and needed to be predicted, evaluation value could be got by applying a calculation on neighbors' evaluation of thisitem[15]. In order to improve prediction accuracy, and considering different users' different rating scale, user's average score on all items is used in prediction. And evaluation value could be calculated by Formula (4).

$$\hat{r}_{ui} = \overline{r}_u + \frac{\sum_{v \in N(u)} SIM(u,v).(r_{vi} - \overline{r}_v)}{\sum_{v \in N(u)} |SIM(u,v)|} \tag{4}$$

Where, $\hat{r}_{ui}$ is rating of item i from user u, and $\overline{r_u}$, $\overline{r_v}$ are average ratings of user u and user v for rating items respectively. N(u) is a collection of neighbors of user u. After prediction score of unevaluated items had been calculated, these items could be sorted according to their score, and several highest scoring items are selected to recommend.

3. Collaborative Filtering Recommendation Based on Normalized Method

3.1. Deficiency of Traditional Collaborative Filtering Algorithm

The traditional memory based collaborative filtering recommendation has been applied in many fields because of its simple and practical algorithm, but there are still many problems. Especially in the similarity calculation process, VSS algorithm and PCC algorithm both are using original evaluation value to calculate similarity, without considering differences of user vector and numerical size of score. For example in Figure 1, this is a simple user evaluation matrix, which contains 5 users and 5 items. User's rating value for each item is between 1 and 6.

	i_1	i_2	i_3	i_4	i_5
u_1	2	1	2	3	6
u_2	1	1	2	4	
u_3	2	1		3	3
u_4	2	2	2	2	3
u_5	4	4	4	4	6

Figure 1. Scoring Matrix of Users and Items

If you want to predict user u_2's rating for item i_5, $Sim(u_1,u_2)=0.9548$ and

$Sim(u_2,u_3)$=0.8653 by Formula (2). While, $Sim(u_1,u_2)$=0.6603 and $Sim(u_2,u_3)$=0.4926 byFormula (3). Obviously, $Sim(u_1,u_2)>Sim(u_2,u_3)$.According to this result, similarity between u_1 and u_2 is bigger than u_2 and u_3. However, as can be seen from Figure 1 that score interval between u_1 and u_2 is [1-6], and score interval between u_2 and u_3 is[1-4]. Therefore, similarity between u_1 and $u_2$is smaller than u_2 and u_3 actually. In the calculation of u_2's rating for item i_5, score of neighbor u_3 on item i_5 should be considered instead of u_1. It is not difficult to find that Pearson algorithm cannot effectively deal with different score style of user vectors in different vector space. And vector space algorithm is only concerned with space angle between users' vector, ignoring different length between users' vector. Therefore, it is necessary to find a more reasonable, effective and appropriate calculation method of similarity, and to improve the recommendation performance.

3.2. Similarity Measurement Method Based on Normalization Method

In order to overcome disadvantage of traditional collaborative filtering algorithm, this paper attempts to introduce normalized data processing in traditional algorithm process. Firstly, all of user's score is normalized to uniform range space after formation of user ratings matrix, assuming that any user vector for all item score are not equal. Secondly, in the process of implementation, for each row vector in scoring matrix R, user's highest score values and lowest score are used to normalize each item score. Finally, final score will be located in section [0-1].After normalization, score on item i from user u is calculated by Formula (5).

$$r_{ui}^{'} = \frac{r_{ui} - r_{u\min}}{r_{u\max} - r_{u\min}} \tag{5}$$

Assuming that num indicates number of items that have been evaluated by user u and v. $\vec{u}$ and $\vec{v}$ respectively denote normalized user u and v's score vector, and the score values of vector in each dimension are between 0 and 1. Combined with correction vector space similarity measurement method presented in literature [8],the similarity of user u and v is calculated by Formula (6).

$$Sim(u,v) = 1 - \frac{dist(\vec{u},\vec{v})}{dist_{\max}} = 1 - \frac{\sqrt{\sum_{i\in I_{uv}} (r_{ui}^{'} - r_{vi}^{'})^2}}{\sqrt{\sum_{k=1}^{|I_{uv}|} (1-0)^2}} = 1 - \frac{\sqrt{\sum_{i\in I_{uv}} (\frac{r_{ui} - r_{u\min}}{r_{u\max} - r_{u\min}} - \frac{r_{vi} - r_{v\min}}{r_{v\max} - r_{v\min}})^2}}{\sqrt{\sum_{k=1}^{|I_{uv}|} (1-0)^2}} \tag{6}$$

Where, dist($\vec{u}$, $\vec{v}$) indicates Euclidean distance between vectors in num dimensional vector space. $dist_{max}$ indicates the max Euclidean distance between vectors in num dimensional vector space. $r_{ui}^{'}$ and $r_{vi}^{'}$ respectively indicates score on item i from user u an v after normalization. r_{umax} and r_{umin} indicate the minimum and maximum score of user u in original score matrix R.I_{uv}indicatesa collection of items that have been evaluated by user u and v. $|I_{uv}|$ indicates the number of item in the collection.

$$Sim(u,v) = 1 - \frac{\sqrt{\sum_{i\in I_{uv}} (\frac{r_{ui} - r_{u\min}}{r_{u\max} - r_{u\min}} - \frac{r_{vi} - r_{v\min}}{r_{v\max} - r_{v\min}})^2}}{\sqrt{|I_{uv}|}} \tag{7}$$

According to Formula (6), Sim(u,v)∈[0-1], the greater the value is, more similar user

u and user v are. If similarity value equals to 0, two users are completely different; while 1 means two users are same. According to data in Figure 1, values of $Sim(u_1,u_2)$ and $Sim(u_2,u_3)$ are 0.7915 and 0.8557, the similarity results are consistent with facts. When a user's scores are exactly same for all items, divisor will be zero. But because all items will be involved in various commodities and services of recommendation system, valid user ratings are not all the same, this situation can be ignored.

In order to calculate similarity between items that have been valued, Formula (8) gives calculation method for Sim(i,j).

$$Sim(i,j)=1-\frac{\sqrt{\sum_{u\in U_{ij}}\left(\frac{r_{ui}-r_{i\min}}{r_{i\max}-r_{i\min}}-\frac{r_{uj}-r_{j\min}}{r_{j\max}-r_{j\min}}\right)^2}}{\sqrt{|U_{ij}|}} \tag{8}$$

Where, U_{ij} indicates collection of users that have evaluated item i and item j. $|U_{ij}|$ indicates number of users in the set. r_{imin} and r_{imax} indicate the min and max score on item i form users. r_{jmin} and r_{jmax} indicate the min and max score on item j from users. Under special cases, when all users give exactly the same score for an item, divisor will be zero. However, because a single item score will be involved in all kinds of user in recommendation system, scores are not all the same in reality, this situation can be ignored.

3.3. Item Rating Prediction and Recommendation Based on Normalization Method

Based on similarity measure algorithm proposed in Section 3.2, a new memory based collaborative filtering method is given in this section. When it is needed to predict unknown score on item i from user u, this method will predict score according to user's original score scale. In the user based item score prediction, $\hat{r}_{ui}$, prediction score of item i form user u, could be calculated by Formula (9).

$$\hat{r}_{ui}=r_{u\min}+(r_{u\max}-r_{u\min})\frac{\sum_{u'\in U}(Sim(u,u')\times\frac{r_{u'i}-r_{u'\min}}{r_{u'\max}-r_{u'\min}})}{\sum_{u'\in U}Sim(u,u')} \tag{9}$$

Where, U indicates collection of k users who had evaluated on item i and have the highest similarity with user u.

In the item based collaborative filtering prediction, $\hat{r}_{ui}$, prediction score of item i form user u, could be calculated by Formula (10).

$$\hat{r}_{ui}=r_{i\min}+(r_{i\max}-r_{i\min})\frac{\sum_{i'\in I}(Sim(i,i')\times\frac{r_{ui'}-r_{i\min}}{r_{i'\max}-r_{i'\min}})}{\sum_{i'\in I}Sim(i,i')} \tag{10}$$

Where, Iindicates collection of k items that had been evaluated by user u and have the highest similarity with item i.

4. Experiment and Analysis

4.1. Experimental Data Set

In order to verify effectiveness of this algorithm, Movie lens movie ratings public data set (http://www.grouplens.org) provided by Group Lens research group is used to verify

this algorithm. The data set is collected from a movie recommendation site, containing 100000 scores on 1682 films from 943 users. Sparsity degree of this set is 93.7%. Part of items of original score matrix are random modified in this experiments to form 10 data sets whose data density respectively are 2%, 4%, 6%, 8%, 10%, 12%, 14%, 16%, 18% and 20%. The reason of using high sparsity degree is that actual evaluation data set is usually very sparse.

4.2. Evaluation Criteria

In order to test accuracy of item rating prediction, the widely adopted Mean Absolute Error(MAE) is used to measure effectiveness of prediction. MAE is average value of absolute deviation between forecast value and real value. Formula (11) gives calculation method of MAE.

$$MAE = \frac{\sum_{ui} | r_{ui} - \hat{r}_{ui} |}{N} \tag{11}$$

Where, r_{ui} indicates actual evaluation of user u on item i. $\hat{r}_{ui}$ indicates predicated evaluation of user u on item i. N indicates the number of predicated values. The lower MAE value is, the higher prediction accuracy is.

4.3. Experimental Scheme

In order to compare the efficiency of normalized collaborative filtering algorithm, five kind of traditional collaborative filtering methods and user based normalization collaborative filtering algorithm (NDCF) are studied for comparison. The first traditional collaborative filtering method is user-based mean (UMEAN), which uses mean score of all items from a user to calculate unknown score of an item. The second traditional collaborative filtering method is item-based mean (IMEAN), which uses mean score of an item from all users to calculate unknown score of an item. The third, fourth and fifth traditional collaborative filtering method are user-based collaborative filtering using PCC (UPCC), item-based collaborative filtering using PCC (IPCC), and collaborative filtering method using UPCC and IPCC (WSRec).

For each of 10 sparse matrixes, by using half off cross validation method, users' record data is divided 5 times into training set and test set according to ratio of 4:1. Test set data and corresponding MAE value are predicted and calculated according to training set data. Average MAE values from 5 experimental are obtained through statistics. In this experiment, neighbor number k is 10,20,30,40 and 50 respectively. Figure 2, to Figure 6, present experimental results of 5 kinds of neighbor scale.

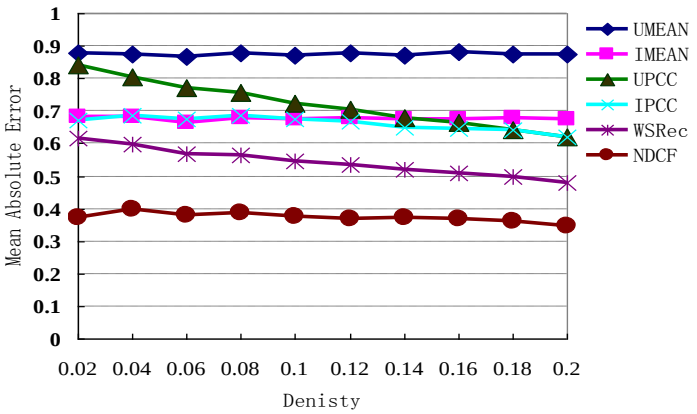


Figure 2. Performance Comparison Chart of 10 Neighbors

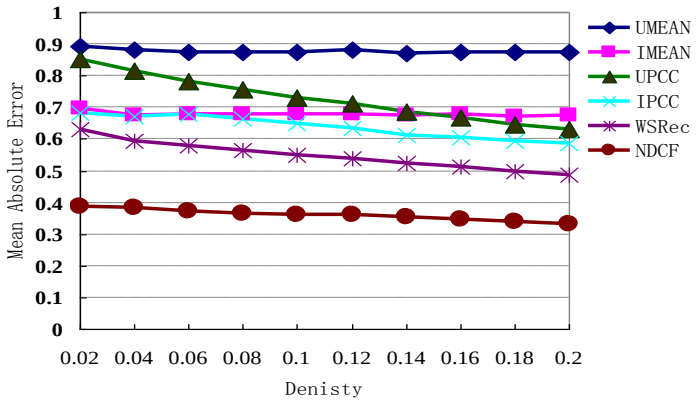


Figure 3. Performance Comparison Chart of 20 Neighbors

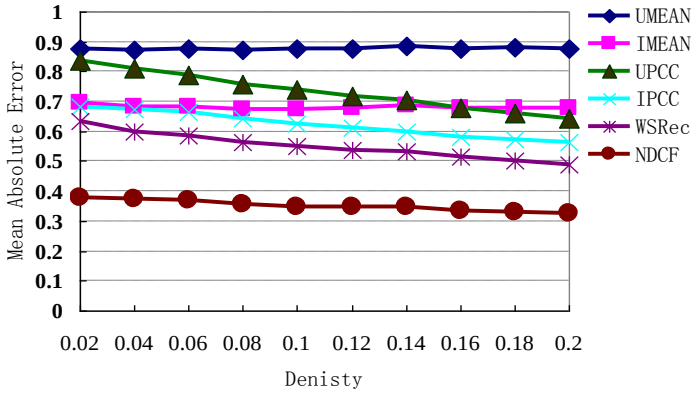


Figure 4. Performance Comparison Chart of 30 Neighbors

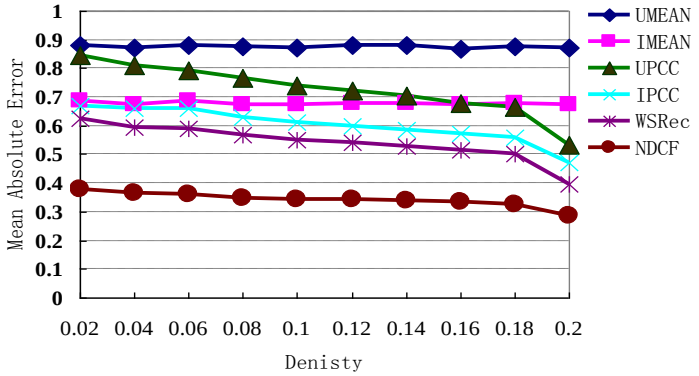


Figure 5. Performance Comparison Chart of 40 Neighbors

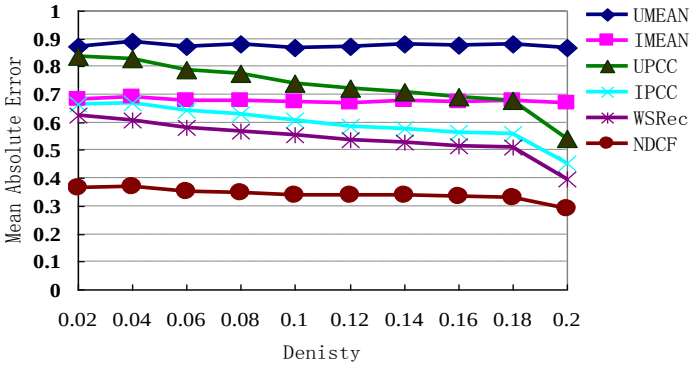


Figure 6. Performance Comparison Chart of 50 Neighbors

According to test results, it can be seen that MAE value from collaborative filtering recommendation based on normalized data processing (NDCF) is obviously better than other traditional recommendation algorithms. But with the increase of density of user ratings matrix, compared to other methods, MAE value of NDCF changed little. The reason is that due to the sparser rating matrix is, the more various user vector dimension is. Therefore, NDCF have better performance than other methods in low density score matrix. Under realistic environment, effective user score matrixes in recommendation system are usually sparse.

At the same time, it is observed that just like other algorithms, with the increase of K value of neighbor scale, MAE value drops. Figure 7, indicates that when scoring matrix density is equal to 0.1, and with neighbor size K increasing from 10 to 50, change of MAE value of NDCF algorithm indicates that NDCF prediction accuracy becomes better and better. But when K value comes in to interval of 40 and 50, MAE value curve becomes flattened. If K value increases further, noise neighbors will increase, and forecast accuracy will be affected. However, compare to other recommendation, collaborative filtering recommendation method based on normalization method can adapt to more extensive data sparsely, and rating prediction performance is improved obviously.

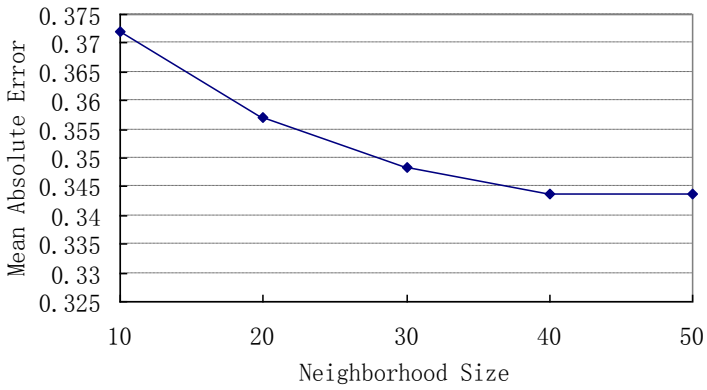


Figure 7. Influence of Neighbor Number on MAE Value

5. Conclusion

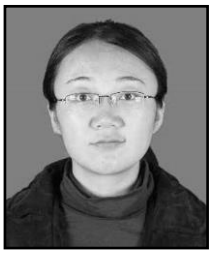
Aiming at shortcomings of traditional collaborative filtering algorithm, this paper proposes a collaborative filtering recommendation algorithm based on normalized data processing, and wants to solve the problem of prediction and recommendation of personalized item rating. NDCF includes a new user or item similarity computing method which could more accurately find similar neighbor users or items, and improve

performance of prediction and recommendation. Experimental results show that compared with traditional recommendation algorithm, NDCF can significantly improve the performance of rating prediction.

References:

[1] C. Li, C. Y. Liang and L. Ma, "A Collaborative Filtering Recommendation Algorithm Based on Domain Nearest Neighbor", Journal of Computer Research and Development, vol. 45, no. 9, **(2008)**, pp. 1532-1538.
[2] F. Xie, Z. Chen and H. F. Xu, "TST: Threshold Based Similarity Transitivity Method in Collaborative Filtering with Cloud Computing", Tsinghua Science and Technology, vol. 03, **(2013)**, pp. 318-327.
[3] K. Treerattanapitak and C. Jaruskulchai, "Exponential Fuzzy C-Means for Collaborative Filtering", J. Journal of Computer Science & Technology, vol. 03, **(2012)**, pp. 567-576.
[4] C. Federico and B. Latchezar, "Identifying Composite Crosscutting Concerns with Scatter-Based Graph Clustering", Journal Wuhan University Journal of Natural Sciences, vol. 02, **(2012)**, pp. 114-120.
[5] S. Yan, Y. L. Li, Y. X. Song, L. Lin, P. F. Jia and S. Q. Cai, "Identification of Chinese Materia Medicas in Microscopic Powder Images", J. Tsinghua Science and Technology, vol. 02, **(2012)**, pp. 209-217.
[6] W. J. Park, S. M. Lee and S. H. Lee, "Designing similarity measurement with distance measure and application on laterally directional mode flight test", Journal of Central South University, vol. 04, **(2012)**, pp. 1032-1039.
[7] X. P. Yuan, J. Long and Z. P. Zhang, "Near-duplicate document detection with improved similarity measurement", Journal of Central South University, vol. 08, **(2012)**, pp. 2231-2237.
[8] J. Huo, N. Yang and M. L Cao, "A reliable algorithm for image matching based on SIFT", J. Journal of Harbin Institute of Technology, vol. 04, **(2012)**, pp. 90-95.
[9] G. W. Zhang, J. C. Kang and H. S. LI, "Context Based Collaborative Filtering Recommendation Algorithm", J. Journal of System Simulation, vol. 18, no. 22, **(2006)**, pp. 595-601.
[10] H. Sun, Z. Zheng and J. Chen, "NE.CF: A Novel Collaborative Filtering Method for Service Recommendation", 2011 IEEE International Conference on Web Services (ICWS), **(2011)**, pp. 702-703.
[11] Z. D. Zhao and M. S. Shang, "User-based Collaborative-Filtering Recommendation Algorithms on Hadoop", Third International Conference on Knowledge Discovery and Data Mining, **(2010)**, pp. 478-481.
[12] R. Salakhutdinov, A. Mnih and G. Hinton, "Restricted Boltzmann machines for collaborative filtering", Proceedings of the 24th international conference on Machine learning, **(2007)**, pp. 791-798.
[13] S. Banerjee and K. Ramanathan, "Collaborative filtering on skewed datasets", Proceedings of the 17th international conference on World Wide Web, **(2008)**, pp. 1135-1136.
[14] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model", Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2008)**, pp. 426-434.
[15] H. Sun, Y. Peng and J. Chen, "A New Similarity Measure Based on Adjusted Euclidean Distance for Memory-based Collaborative Filtering", J. Journal of Software, vol. 6, no. 6, **(2011)**, pp. 993-1000.

Author



GaoYan. She is a lecture of Information Engineering School of Yulin University. Her current research interests include recommendation algorithm, data processing and computer software.

