International Journal of Hybrid Information Technology
Vol.8, No.2 (2015), pp.149-158
http://dx.doi.org/10.14257/ijhit.2015.8.2.13

Community-based Collaborative Filtering Recommendation Algorithm

Xiaofang Ding, Zhixiao Wang*, Shaoda Chen and Ying Huang

College of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
**softstone416@163.com*

Abstract

Collaborative filtering recommendation technology is by far the most widely used and successful personalized recommendation technology. However, the method currently faced with some problems such as sparse matrix, affecting the accuracy of the predicted results. This paper puts forward a new community detection algorithm based on topological potential theory, and combines it with collaborative filtering recommendation algorithm. The users with similar interests are put into the same community. When searching for the user's nearest neighbor, it target to the users in a specific community or several communities instead of all users, which narrows the search and improves the prediction accuracy. Experimental results suggest that this approach effectively reduces the impact on the prediction accuracy of the sparse matrix, and significantly improves the prediction ability and recommendation quality.

Keywords: *community divide; collaborative filtering; recommendation; topology potential*

1. Introduction

Along with the rapid development of the Internet, the amount of the information on networks shows an exponential growth. Collaborative filtering recommendation algorithm analyzes the information of the users and items collected from the networks and create a user-rating matrix. Doing a comparative analysis with the information from different users can help discover the patterns and knowledge with potential values from these massive information resources efficiently, and give a prediction of the potential interests for users, which provides an efficiency solution for the "information overload" problem [1].

There are three general classes of collaborative filtering algorithm: user-based collaborative filtering, item-based collaborative filtering and model-based collaborative filtering [2-4].

Being the most classical and most widely used personalized recommendation algorithm, collaborative filtering shows obvious advantages, but it also faces some challenges. Since collaborative filtering recommendation algorithm is based on the search of user's nearest neighbor set [5], and with a significant increase in the number of users and items, the size of user set is growing, it's certain to cost more time to search and compute and hence affect the efficiency and quality of recommendation. In addition, the growth of user and item base makes the user-rating matrix sparsity becomes a more serious problem [6]. A large number of items can't get enough ratings from users, makes it difficult for the system to give an accurate prediction, sometimes it even can't deliver a result, affecting the ability and accuracy of the algorithm.

In response to these problems, we propose a collaborative filtering algorithm based on community division. It based on the idea that "people of a mind fall into the same group",

ISSN: 1738-9968 IJHIT

dividing users with similar preference into the same community [7]. Then combine it with collaborative filtering algorithm, and search for the nearest neighbor from the same community with active user instead of the whole network [8], which can narrow the search several times and improve the prediction accuracy.

2. Community Detection Algorithm based on Topology Potential

2.1. Traditional Community Detection Algorithm based on Topology Potential

Topology potential approach is a new community detection method proposed in recent years, it attracts many researchers' attention [9-13] for its' lower time complexity and needing no field or expert knowledge. Complex network can be modeled as a graph, which defined as $G=(V,E)$, such that V is the set of nodes as users of the networks and E is the set of edges as the connection between nodes. Nodes in complex networks are not isolate but connected by edges. We introduce the topological potential theory in physics to describe the relationship and interactions among these nodes. Each node in the network can be abstracted as a field source with a certain mass, and can affect other nodes in the same potential field, all these nodes interconnected and working together will form a field, called topological potential field [14].

There are some deficiencies of current topology-based community detection approaches, when distribute nodes to communities, they need to use additional means such as benefit function, the ratio of the indegree and outdegree of the node, and regulating parameter ξ, which increase the difficulty and complexity of the detection. For instance, in literature [13] the size of the community and ownership of the nodes are totally depend on the preset regulation parameter ξ. For real complex network, it's difficult to predict the size of each community in advance, so it's hard to find a reasonable adjustment parameter.

In response to these problems, we put forward a new community detection algorithm, which utilizes the inherent peak-valley structure in topology potential field to carry out community detection and determines the community attachment of nodes based on the position of them in topology potential field.

2.2. New Community Detection Algorithm based on Topology Potential

Linkages and connections among nodes in complex network contribute to the topology potential field. In the field, nodes get together due to their mutual effect. Each community corresponds to a local higher potential region in the topology potential field, where the representative node of the community values the maximum potential [15]. The topology potential distribution of a social network always shows an inherent peak-valley structure, in which some nodes with higher potential value are located in higher position, some smaller lie in the lower position.

Utilizing the inherent peak-valley structure in topology potential field, we divide the node into following four kinds according to their positions in it: peak, valley, slope position and edge.

Peak represents a maximum potential value of local higher potential region which correspond to a community, and node in peak position is the center of the region and is the representative of the community [9]. If the distance of two peak nodes is closer than $\lfloor 3\sigma/\sqrt{2} \rfloor$ [13], selecting a higher potential one as the representative node for the community and the smaller one as the merge node of peak.

Nodes in the valley of the topology potential field value smaller potential, they usually located in the junction of several local high potential areas and they are the overlap nodes of the community.

Nodes in the slope of the topology potential field are located between the peak one and valley one, they are internal nodes of the local high potential area.

Nodes in the edge of the topology potential field keep very weak links with other nodes, some even have no link, they are isolated nodes and don't belong to any community.

2.3. Algorithm Description

Algorithm 1: Community detection based on topology potential

Input: a complex network $G=(V,E),|V|=n\ |E|=m$

Output: peak nodes, internal nodes of each community and isolate nodes

Step 1: compute potential value of all nodes and establish the topology potential field for G

Step 2: seek all peak nodes in the topology potential field with hill-climbing method

Step 3: choose representative node for each community from peak nodes

Step 4: choose a representative node at random from nodes searched out in step 3, and expand them in breath-first way

Step 5: figure out the community identification of current node according to its location. If it's in the valley, mark it as an overlap node and stop expansion, if it's in a slope location, identify it as an internal node of the community and continue to expand until get to valley

Step 6: repeat step 4 and 5 until all representative nodes have been expanded

Step 7: after expanding for all representative nodes, identify the left nodes which haven't been visited as isolated nodes, they were in the merge of topology potential field

Step 8: output the peak and internal nodes of each community and isolated nodes.

3. Community-based Collaborative Filtering

Community-based collaborative filtering consists of three parts: (1) build social networks for users, (2) community detection based on topology potential, (3) predict and recommend items for users on the basis of detection results in part (2).

3.1. Build User Social Network

User social network is based on user-rating matrix and similarity matrix.

At first, assume that there are n users, and m items in the network, with the information of users and items we can get a user-rating matrix R with n rows and m columns as follow:

$$\mathrm{R}=\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1j} & \cdots & R_{1m} \\ R_{21} & R_{22} & \cdots & R_{2j} & \cdots & R_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{i1} & R_{i2} & \cdots & R_{ij} & \cdots & R_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{nj} & \cdots & R_{nm} \end{bmatrix} \tag{3-1}$$

R_{ij} is the rating of user i on item j, $R_{ij}=\{0,1,2,3,4,5\}$, if user i doesn't rate on item j, then R_{ij} values 0, and a higher value of R_{ij} implies more preference the user to item.

Next, compute the user similarity based on user-rating matrix and relevant formula and build similarity matrix as follow [16]:

$$S=\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1j} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2j} & \cdots & S_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{i1} & S_{i2} & \cdots & S_{ij} & \cdots & S_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{n1} & S_{n2} & \cdots & S_{nj} & \cdots & S_{nn} \end{bmatrix} \quad (3\text{-}2)$$

S_{ij} is the similarity between user i and j, it's obvious that S is a symmetric matrix since the similarity of user i and j values the same as it of user j and i.

Then we can get the adjacency matrix A with the similarity matrix, it's as follow:

$$A=\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1j} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2j} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{i1} & A_{i2} & \cdots & A_{ij} & \cdots & A_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nj} & \cdots & A_{nn} \end{bmatrix} \quad (3\text{-}3)$$

A is established on the basis of the similarity threshold value set in advance [17], when S_{ij} is larger than the threshold, set $A_{ij}=1$, otherwise $A_{ij}=0$. $A_{ij}=1$ means there is a line connecting between nodes i and j in users social network. When all connections are identified, the user social network is set up. It's the final preparation before community division, the element value of matrix A is limited to taking 1 or 0, $A_{ij}=1$ means there is connection between user i and j (or a closely link), $A_{ij}=0$ means there is no association (or a weak link) between user i and j.

After the work for user social network, we can apply algorithm 1 as a basis for the community division.

3.2. Recommend based on Community

Now we can predict ratings and recommend items for user after community detection.

Since the community members are more likely have similar preference [18], when giving recommendation for user u we refer the preferences of users in the same community with u instead of all users in the network to short the running time of the method [8], which is the main difference with normal collaborative filtering algorithm.

3.3. Community-based Collaborative Filtering Algorithm Description

Algorithm 2: Community-based collaborative filtering
Input: users' ratings on items, records to be predicted
Output: results of the prediction
Step 1: build user-rating matrix on the basis of users' ratings on items

Step 2: compute user similarity, and build the similarity matrix

Step 3: set a similarity threshold and establish a user adjacency matrix

Step 4: confirm if there are connections between any two users with the adjacency matrix in step 3, and set up the user social network

Step 5: use the social network in step 4 as the input of algorithm 1, and figure out all peak nodes, internal nodes and isolated nodes

Step 6: detect the community for the target user from step 5, and search the nearest neighbor set in it

Step 7: compute the predicted rating for user on the basis of neighbor set in step 6.

Community-based collaborative filtering algorithm is means to predict ratings for users on the basis of combination between community detection and collaborative filtering. Putting users with similar preference in the same community, community detection algorithm improves the efficiency on the search of neighbor set while keep the prediction accuracy, and reduces the effects of the sparse matrix and brings a certain improvement in prediction ability.

4. Simulation

4.1. Experimental Dataset

We conduct experiments on MovieLens dataset to test the rationality and feasibility of the algorithm proposed in this paper. The dataset consists of over 100,000 ratings (1-5) from 943 users on 1682 movies, and each user has rated at least 20 movies and each movie is rated by at least one user. Ratings are on a scale of 1 to 5, and a higher scale means more preference of user to the film. The information we mostly used in our experiments of the dataset includes user id, item id and rating. We pre-processed this data into the user-rating matrix R before use it.

MovieLens dataset is divided into training set and testing set. Training set is processed as the source of data sample to build user-rating matrix and user similarity matrix, while testing set is for testing and evaluating the result. Since MovieLens contains more than 100,000 records, too much for us to process and operate, we select randomly 100 users' ratings on over 900 movies as the training set and other 100 records of the same users and movies as testing set.

4.2. Evaluation

We select Mean Absolute Error (MAE), one of the most widely recognized evaluations as a measurement for our experiments [19]. MAE measures the prediction accuracy by calculating the deviation between the experiment results and the actual ratings. A smaller MAE value means a smaller deviation between predicted rating and the real data, which suggest higher prediction accuracy and quality. MAE is formulated as:

$$MAE = \frac{1}{|P|} \sum_{(u,i)\in p} \left| r_{(u,i)} - r_{(u,i)}' \right| \tag{4-4}$$

Where $|p|$ is the size of the testing set, (u,i) is the record of user u and item i, $r_{(u,i)}$ is the actual rating of user u on item i, while $r_{(u,i)}'$ is the predict result of the system.

4.3. Experimental Results

We conduct 3 groups of experiments to examine the improvement of our algorithm.

Group 1: to find out how the similarity threshold affect MAE

When construct user social network, the similarity threshold may affect the quality of community division thus affect the performance of the method. When the similarity threshold

is too large, the adjacency matrix tends to be sparse thus resulted in too many communities and scattered nodes which means the reduction of prediction accuracy. Correspondingly, when similarity threshold is too smaller, the adjacency matrix tends to be dense and can only detect a few communities which will also affect the quality of the recommendation. Figure 1 plots the trend of MAE after community divided when change the similarity threshold from 0.2 to 0.9 (we start from 0.2 because there is no clear community structure when threshold is 0.1).

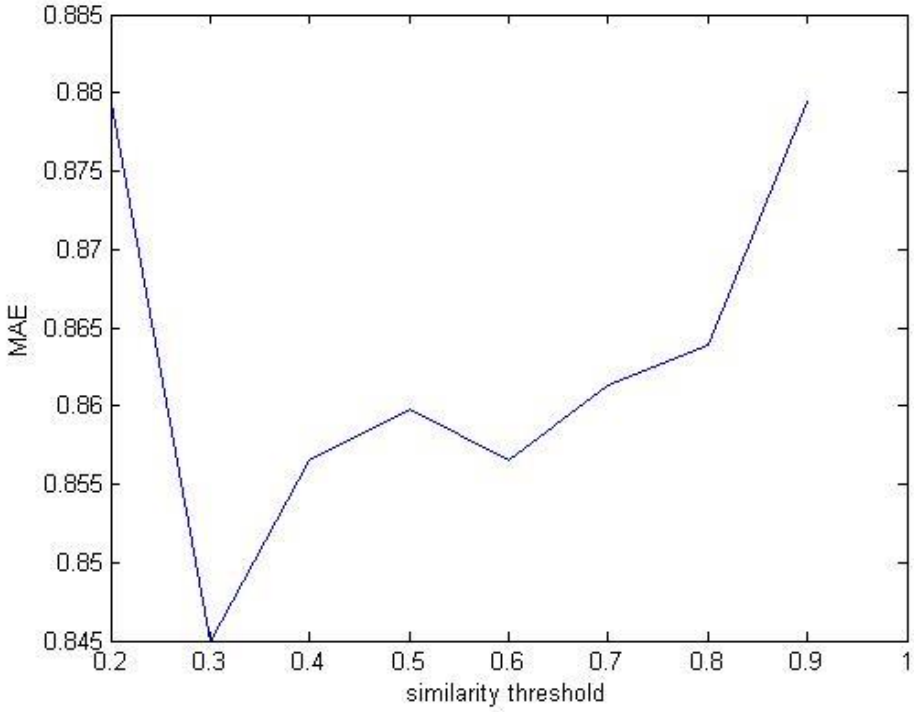

Figure 1. Variety of MAE when Similarity Threshold Change from 0.2 to 0.9

It's obviously that when similarity threshold values 0.3, MAE get the minimum level with the best prediction accuracy.

Group 2: comparison of this algorithm and other two methods in MAE

We compare the performance of three approaches: 1) Collaborative Filtering Based on Community Detecting, referred CFBCD; 2) Collaborative Filtering Based on Top-k nearest neighbor Community Detecting, referred Top-k CFBCD; 3) Collaborative Filtering Based on Pearson Similarity, referred CFBPS. In this group of experiments, we limit number of user nearest neighbor as 5, 10 and 15 in normal community-based collaborative filtering to calculate the MAE and compared the MAE with our new approach. The similarity threshold is set to 0.3 according to the results in experiments of group 1.

Figure 2 shows the experimental results, the abscissa describes the size of neighbor set; the ordinate is the value of MAE. Since the predicted rating of user u on item i is based on all users in the same community with u, the MAE stays in a certain value no matter how the size of neighbor set changes. And Top-k CFBCD based on top k users in the same community with u according to similarity ranking, so MAE values fluctuate along with k changes.

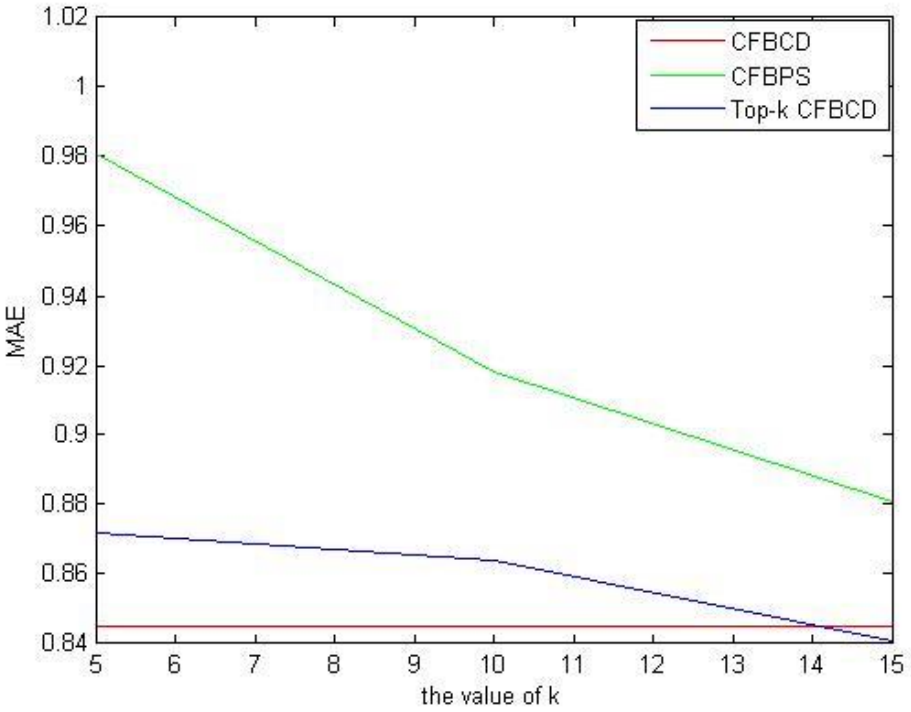

Figure 2. Comparison of Different Methods in MAE

From Figure 2 it's clearly that for normal CFBPS, MAE volatiles more when k get different value and the greater the size of neighbor set the smaller the MAE. As for CFBCD and Top-k CFBCD, the MAE is always smaller than the normal one no matter which value k is set. In addition, when k values smaller (refer to 5 and 10 in this experiment), MAE of CFBCD values is smaller than that of Top-k CFBCD, and when change k to a higher value (refer to 15 in this experiment), MAE of Top-k CFBCD is significantly reduced and is smaller than that of CFBCD.

Group 3: comparison of this algorithm and other two methods in predictive ability

The predictive ability is another standard to measure the performance of a recommendation algorithm except MAE. Since the numerous users and items in real network make sparsity of the user-rating matrix a serious problem, if we search the user neighbor set as the same way of normal collaborative filtering algorithm which choose the top k users ranked by similarity, it's probably to cause the neighbor set too small and even empty to generate a prediction, thereby affecting the prediction ability.

Community-based collaborative filtering algorithm search the users in the same community with target user as the neighbor set, it doesn't cause the situation that a small community just with a few users. Therefore, this method effectively improves the prediction ability of a recommendation system.

Figure 3 compares the numbers of predictions that CFBCD, Top-k CFBCD and CFBPS can give when change k at value 5, 10 and 15. As can be seen, no matter which value k is set, the numbers of predictions with CFBCD is far more than normal collaborative filtering CFBPS, and the number of prediction results will change as k changes in Top-k CFBCD. These happen because numerous users and little rated items cause the sparsity of user-rating matrix, and it can't search enough users for neighbor set thus weaken the prediction ability.

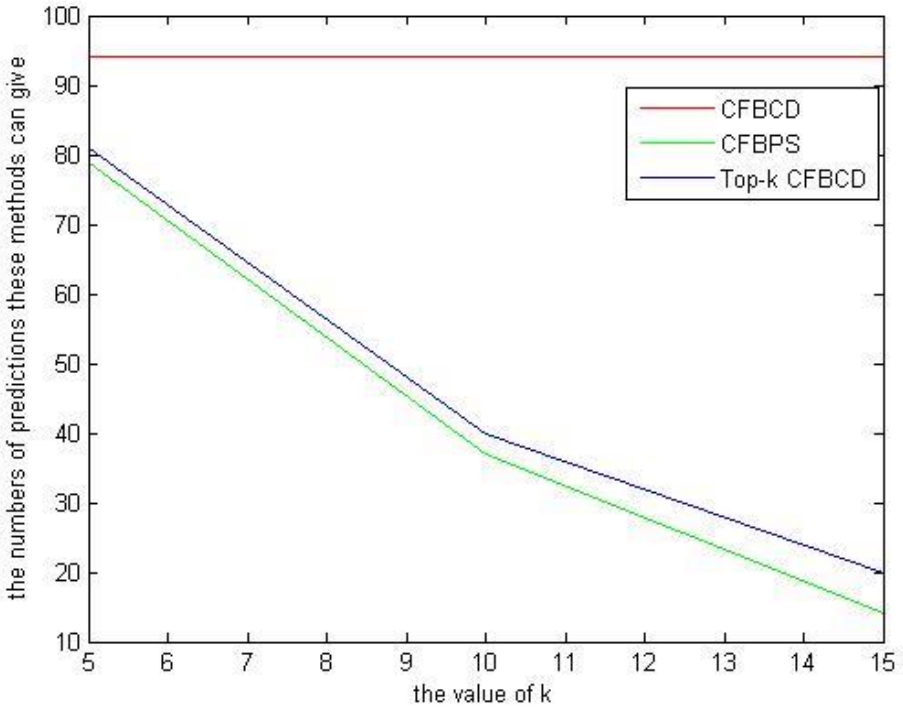

Figure 3. Comparison of Different Methods in Quantity

We can summarize the applications of CFBCD and Top-k CFBCD in two situations: when there are a small number of users or the items are little rated, which means a high sparsity of user-rating matrix, the prediction ability is likely to be improved if CFBCD is chosen. And when deal with densely-populated communities or a small sparse user-rating matrix, the Top-k CFBCD would be better which can short the running time of searching user neighbor set, and improve the algorithm efficiency without reducing the prediction accuracy.

5. Conclusion

Being the most widely used and successful personalized recommendation by far though, collaborative filtering recommendation still faces some challenges. Since collaborative filtering recommendation algorithm is based on the search of user's nearest neighbor set, and with a significant increase in the number of users and items, the user set is growing, it's certain to cost more and more time to generate a recommendation thus reduce the quality of recommendation. In addition, the growth of user and item base makes the user-rating matrix sparsity becomes a more serious problem. A huge number of items can't get enough ratings from users, makes it difficult for the system to give an accurate recommendation; sometimes it can't even deliver a result, affecting the ability and accuracy of the algorithm.

We introduce the topological potential theory, and put forward a new community detection algorithm, which put the users with similar interests into the same community according to the position of the nodes in topology potential field, and then apply the traditional collaborative filtering on the communities. When search for the user's nearest neighbor, it target to the users in a specific community or several communities instead of all users, which narrows the search and improves the prediction accuracy. Experimental results suggest that

this approach effectively reduces the impact on the prediction accuracy of the sparse matrix, and improves the prediction ability and recommendation quality a lot.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.61402482), the Fundamental Research Funds for the Central Universities (No.2014QNB23).

References

[1] K. Choi and Y. Suh, "A new similarity function for selecting neighbors for each target item in collaborative filtering", Knowledge-Based Systems, vol. 37, (**2013**), pp. 146-153.
[2] P. Sun, Z. Li, Z. Han and F. Wang, "An Overview of Collaborative Filtering Recommendation Algorithm", Advanced Materials Research, (**2013**), pp. 756-759, pp. 3899-3903.
[3] R. Latha and R. Nadarajan, "User relevant for item-based collaborative filtering", Computer Information System and Industrial Management, 12th IFIP TC8 International Conference (CISIM 2013), Proceedings: LNCS, vol. 8104, (**2013**), pp. 337-347.
[4] B. Gilbert, H. Hazem, E.-H. Wassim and N. Lama, "A hybrid approach with collaborative filtering for recommender systems", 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC), (**2013**), pp. 349-355.
[5] Y. Ren, G. Li, J. Zhang and W. Zhou, "The maximum imputation framework for neighborhood-based collaborative filtering", Soc. Netw. Anal. Min., vol. 4, (**2014)**, p. 207.
[6] H. Antonio, B. Jesus and S. Francisco, "Collaborative Filtering: the aim of recommender systems and the significance of user", Social Network Technologies and Applications, vol. 23, no. 1, (**2010**), pp. 317-330.
[7] D. Elnaz, A. Mohsen and K. Keivan, "A Social Network-Based Approach to Expert Recommendation System", 7th International Conference on Hybrid Artificial Intelligent System (HAIS), vol. 7208, (**2012**), pp. 91-102.
[8] A. Dang and E. Viennet, "The Collaborative Filtering in Social Network: A Community-based Approach", 2013 International Conference on Computing, Management and Telecommunications (ComMan Tel), (**2013**), pp. 128-133.
[9] Y. Han and D. Li, "A novel measurement of structure properties in complex networks", Lecture Notes of the Institute for Computer Sciences, Social Information and Telecommunications Engineering, vol. 5, (**2009**), pp. 1292-1297.
[10] W. Y. Gan, N. He, D. Y. Li and J. M. Wang, "Community discovery method in networks based on topology potential", Journal of Software, vol. 20, no. 8, (**2009**), pp. 2241-2254.
[11] Y. Han, D. Li, T. Wang, "Identifying different community members in complex networks based on topology potential", Frontiers of Computer Science in China, vol. 5, no. 1, (**2011**), pp. 87-99.
[12] J.-p. Zhang, H.-b. Li, J. Yang, J.-b. Bai, L.-j. Zhang and Y. Chu, "Variable scale network overlapping community identification based on identity uncertainty", ACTA ELECTRONICA SINICA, vol. 40, no. 12, (**2012**), pp. 2512-2518.
[13] J.-p. Zhang, H.-b. Li, J. Yang, J.-b. Bai, Y. Chu and L.-j. Zhang, "Community discovery method with uncertainty measure of overlapping nodes based on topology potential", Journal of Harbin Institute of Technology (new series), vol. 19, no. 2, (**2012**), pp. 16-22.
[14] Z. Wang, D. Zhang, G. Yu and T. Zhu, "Semantic field model", Journal of Tongji University (Natural Science), vol. 37, no. 11, (**2009**), pp. 1526-1530.
[15] P. De Meo, E. Ferrara, G. Fiumara and A. Provetti, "Enhancing community detection using a network weighting strategy", Inform. Sci., vol. 222, (**2013**), pp. 648–668.
[16] X. Chen, "Design and Implementation of Community Collaborative Filtering Method Based on Topological Potential", Beijing University of Posts and Telecommunications, (**2011**).
[17] G. Wang and H. Liu, "Survey of personalized recommendation system", Computer Engineering and Applications, vol. 48, no. 7, (**2012**), pp. 66-76.
[18] L. Hu, G. Song, Z. Xie and K. Zhao, "Personalized recommendation algorithm based on preference features", Tsinghua Science and Technology, vol. 19, no. 3, (**2014**), pp. 193-299.
[19] MS Shang, CH Jin, T Zhou and YC Zhang, "Collaborative filtering based on multi-channe difushion", Phsica A., vol. 388, no. 23, (**2009**), pp. 4867-4871.

