

Overlapping Community Detection in Social Network Based on Microblog User Model

Yajun Gu¹, Bofeng Zhang^{1,2*}, Guobing Zou¹, Mingqing Huang¹, Keyuan Jiang²

¹School of Computer Engineering & Science
Shanghai University
Shanghai, China
e-mail:bfzhang@shu.edu.cn

²Department of Computer Information Technology & Graphics
Purdue University Calumet
Indiana, USA
e-mail:jiang@purduecal.edu

Abstract—Online social networks have found a significant increase in their popularity in recent years. All the networks have community structure, and one of the research problems mostly frequently tackled is the discovery of communities. An overlapping community is a network structure that allows one node to be a member of multiple communities. The method presented in this paper aims at detecting overlapping communities in social networks, and its novelty lies in that it combines with the Microblog User Model (MUM) which can reflect the interest of the user accurately. First, the MUM network, which is an undirected and weighted network, is constructed by computing the similarity among MUMs. Afterwards, Overlapping Community Detection based on MUM (OCD-MUM) is performed to partition the network. A community stops expanding when the fitness function reaches a local maximum. The communities detected are locally optimized. A user's interest is not only decided by the MUM, but it is also affected by the communities the user belongs to. The community model can reflect the interest of the community. The MUM is updated with community model of its communities, and therefore the interest of the user can be predicted by these communities. Our experiment result shows that OCD-MUM has a higher modularity Q value than traditional methods and the predicted interest is more close to the real world situations.

Keywords—microblog user model; social network; overlapping communities detection; interest prediction

I. INTRODUCTION

With the rapid development of Web 2.0 in the last few years, online social networks, such as Facebook, Twitter, Sina Weibo and Tencent microblog, have become important media platforms for users to know each other and communicate with their friends all over the world.

Within social networks there exist communities which may consist of groups of related individuals. A community can be represented as a network within which there are nodes representing individual community members. Nodes within a community are much more connected to each other than to the nodes in other communities. It has been well understood that people in social network are naturally characterized to belonging to multiple communities. For example, an individual usually may connect to several groups such as

family, friends and colleagues. A researcher may be active in several areas including science, mathematics and geography. How to find communities effectively has been one of the important research areas in complex network.

Efforts have been done in detecting separate communities. Han proposed a method (CSD-IS) for detecting communities in weighted network based on the interest similarity [1]. The weight of the edge represents the interest similarity among users. Xing proposed a new microblog community discovery algorithm based on user interest [2]. Yan proposed a microblog community discovery algorithm based on both links and topics [3]. Although these three methods can discover microblog communities effectively, their algorithms are limited to the cases where each user belongs to a single community. Their methods do not work for the real-world situations.

For the recent years, researchers have focused on detecting the overlapping communities. The work on detecting overlapping communities was previously proposed by Palla with the clique percolation algorithm (CPM) [4] which is suitable for networks with dense connected parts. Baumes proposed the iterative scan algorithm (IS) [5] that expands seeded small cluster cores by adding or removing nodes until the local density function cannot be improved. The quality of discovered communities depends on the quality of seeds. Lancichinetti proposed the local fitness maximum algorithm (LFM) [6] that expands a community from a random seed node until the fitness function is locally maximal. LFM depends significantly on a parameter of the fitness function that controls the size of the communities. Steve proposed Cluster-Overlap Newman Girvan Algorithm (CONGA) [7] that extends Girvan and Newman's divisive clustering algorithm by allowing a node to split into multiple copies. Copra is an extension of the label propagation algorithm for overlapping community detection [8]. Each node updates its belonging coefficients by averaging the coefficients over all its neighbors. Copra produces a number of small size communities in some networks.

The afore-mentioned methods can detect the overlapping communities effectively, but they are not suit for weighted networks. There exist many weighted networks [9] in the real world. The weight of a network edge is of practical significance. For example, the weight may represent the

*Corresponding author.

This work was supported by the National Natural Science Foundation of China (61303096), and Shanghai Natural Science Foundation (13ZR1454600).

similarity or close strength among users in a social network. In this paper, we propose a method of Overlapping Community Detection based on MUM (OCD-M) in order to partition the weighted network into overlapping communities.

There are many traditional ways to compute microblog user's interest [10, 11]. In this paper, we first discuss the construction of the Microblog User Model (MUM) to express users' interest. Afterwards, we then describe the construction of the MUM network by computing the similarity among the MUMs. Next, we detect overlapping communities using the OCD-MUM method. Finally, we construct the community model to reflect the interest of the community. Interest of the user can be predicted by the MUM updated with community models of communities this user belongs to. We conclude with our experiment results which show the interest prediction method achieves more reasonable partition result than the CSD-IS method proposed in [1].

II. CONSTRUCTION OF MICROBLOG USER MODEL NETWORK

A. The MUM Construction

Microblog User Model (MUM) is a model that reflects the microblog user's basic attributions, interested topics and social relationship according to the user's behaviors and intentions in microblog social media. It can be applied to interest prediction, friend recommendation and online advertisement. In a word, MUM is user's virtual image from the real world to microblog social platform.

MUM contains three parts: base information, semantic interest and social relationship. Base information includes name, nick name, gender, location and tag. Semantic interest is mined from their microblog content. Social relationship is derived from the user's followings and followers. As shown in Fig. 1, (a) is the framework of the MUM and (b) is an instance of MUM which represents an actual user.

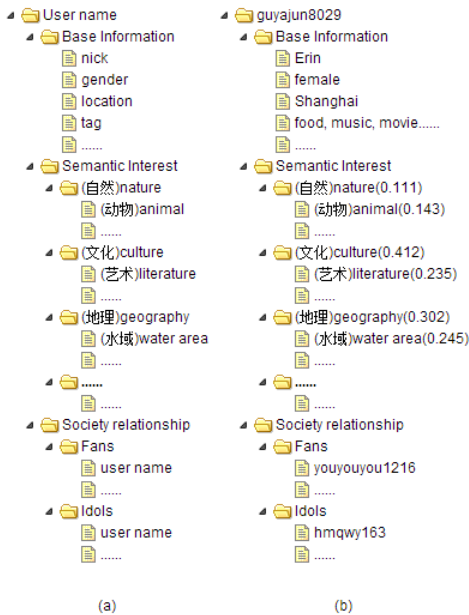


Fig. 1. Construction of MUM

The construction of MUM consists of the following steps:

1) Computing the initial interest to keywords

To create a Chinese comprehensive classification platform, Baidu developed the Baidu Encyclopedia Ontology Base (BEOB) which is organized in 4 levels and with a total of 380,627 keywords (<http://baike.baidu.com>). We can see the large 11 categories at the bottom of the web page. A small part of BEOB is shown in Fig. 2. Each microblog user corresponds to a MUM which in turn corresponds to a MUM Ontology Base (MUMOB) that is similar to BEOB. Both MUMOB and BEOB have the same hierarchical structure. However, MUMOB not only contain the keywords but also contain the calculated interest value to the corresponding keyword.

The initial interest to keywords is calculated with TF-IDF [12] given by (1).

$$TF-IDF(P_i, t_j, P) = \frac{tf(t_j, P_i)}{\sum_{t_k \in P} tf(t_k, P_i)} \times \log\left(\frac{|P|}{|\{P_k \in P: t_j \in P_k\}|}\right) \quad (1)$$

where P_i is a document which is the entire collection of i^{th} user's blogs. Each user corresponds to a document; P is all the documents (blogs); $|P|$ is the number of documents, which is equal to the number of users; t_j is a keyword in MUMOB; $tf(t_j, P_i)$ is the number of times of t_j appears in P_i . TF-IDF value is considered as the user's initial interest to a particular keyword.

While traversing the MUMOB, if a keyword in the MUMOB is found in a user's blogs, the TF-IDF value of the keyword will be computed. If the keyword is not contained in the user's blogs, its TF-IDF value equals zero.

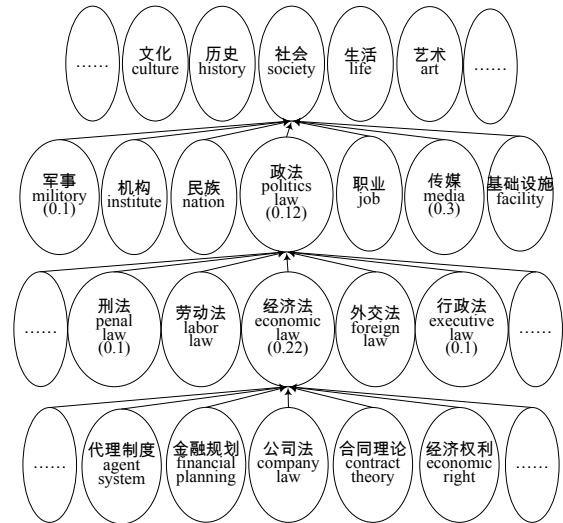


Fig. 2. A part of Baidu Encyclopedia Ontology Base.

2) Updating the interest value based on hierarchical structure of MUMOB

Considering the semantic influence of hierarchical structure, each keyword interest value at the parent level is updated by its child level, as shown in (2).

$$I = I_f + \frac{\sum_{i=1}^m I_{c_i}}{m} \quad (2)$$

where I_f is the interest value of the parent keyword; I_{c_i} is the interest value of the child keyword; m is the number of child keywords.

For example, in Fig. 2, assuming that the keyword "politics law" has 10 child keywords at the third level, the interest of keyword "politics law" at the second level will be updated to $0.12 + (0.1 + 0.1 + 0.22) / 10 = 0.162$.

3) Normalizing each level of MUMOB

In the MUMOB, there are 11 keywords at the top level, 89 terms at the second level, about 1,068 keywords at the third level, and about 379,559 ones at the bottom level. In order to create the standard for comparisons, each level should be normalized. It means that the sum of 11 interest values at the top level will be equal to 1.0, and so is the sum at other levels. For example, user A and user B are both interested in basketball and swim. A's interest degree in basketball and swim are separately 0.6 and 1.4. While B's interest degree in basketball and swim are separately 0.4 and 0.6. We can know that B is more interested in basketball only after normalized.

4) Updating the interest value with social relationship

The interest of a user is not only related to the user's own microblog content but also related to the interest of other users the user follows. Therefore, the interest value of each keyword should be updated with (3).

$$I'(t_i) = \mu I(t_i) + (1 - \mu) \frac{I_1(t_i) + I_2(t_i) + \dots + I_n(t_i)}{n} \quad (3)$$

where $I(t_i)$ is the user's final interest value to keyword t_i ; μ is an adjustable value representing the influence weight of the user's own interest to keyword t_i ; n is the number of the user's followings; $I_1(t_i)$ is the first following's interest value to keyword t_i . This completes our construction of the MUM.

MUM can be presented as a vector. Each element of the vector represents the user's interest value to the corresponding keyword.

B. The MUM Network Construction

An MUM network is an undirected and weighted network. In an MUM network, each node represents a MUM, each edge indicates that there is common interest between two MUMs, and the weight of the edge represents the similarity between two MUMs. MUM network is constructed in the following steps:

1) Computing each level's similarity

The similarity of each level is given by (4) as follows.

$$S(u_p, u_q, h) = \frac{u_p(h) \bullet u_q(h)}{\|u_p(h)\| \|u_q(h)\|} = \frac{\sum_{i=1}^{t(h)} u_p(h, i) \times u_q(h, i)}{\sqrt{\sum_{i=1}^{t(h)} u_p^2(h, i)} \sqrt{\sum_{i=1}^{t(h)} u_q^2(h, i)}} \quad (4)$$

where $S(u_p, u_q, h)$ is the similarity between MUM u_p and MUM u_q at level h ; $u_p(h, i)$ is u_p 's interest value to the i th keyword at level h ; $t(h)$ indicates the number of keywords at level h .

2) Mixing the levels' similarity

The four levels in MUMOB have different number of keywords. The top level has 11 categories that are very abstract, so the difference between the weight of the first level and that of the second level is relatively large. The lower levels have a more and more detailed classification, so the weights of lower levels vary little. The hyperbolic tangent function is selected to reflect the unbalanced weights. The weight of an edge is calculated by (5) as follows.

$$S(u_p, u_q) = \frac{\sum_{h=1}^{\Phi} W(h) S(u_p, u_q, h)}{\sum_{h=1}^{\Phi} W(h)} \quad (5)$$

where $W(h)$ is the weight of level h , and

$$W(h) = \frac{e^{h\gamma^{[h-M(\Phi)]}} - e^{-h\gamma^{[h-M(\Phi)]}}}{e^{h\gamma^{[h-M(\Phi)]}} + e^{-h\gamma^{[h-M(\Phi)]}}};$$

ϕ is the levels of the MUM. $M(\phi)$ is the hierarchy median, and

$$M(\Phi) = \frac{\sum_{h=1}^{\Phi} h}{\Phi};$$

γ is an adjustable value that adjusts the influence weight of different levels; $S(u_p, u_q, h)$ is the similarity between u_p and u_q at level h ; $S(u_p, u_q)$ is the final similarity between u_p and u_q .

III. OVERLAPPING COMMUNITY DETECTION BASED ON MUM

A. Overlapping Community Detection Based on MUM

Fitness function is given by (6) as follows.

$$f_G = \frac{W_{in}^G}{W_{in}^G + W_{out}^G} \quad (6)$$

where G is a community; W_{in}^G is the sum of internal edges' weights; W_{out}^G is the sum of external edges' weights. Here, an external edge is the edge that joins the node in G with the rest of the graph.

The fitness of a node A with respect to subgraph G is defined as the variation of the fitness of subgraph G with and without node A. It is given by (7).

$$f_G^A = f_{G+\{A\}} - f_G \quad (7)$$

Equation (7) represents the increment of fitness function when a node is added to G , and here the subscript $G+\{A\}$ means that node A is added to G .

Assuming that we have initialized a subgraph G including only one node ($W_{in}=0$). A community is formed with the following procedures.

- a) A loop is performed over all neighbor nodes of G not included in G .
- b) The neighbor with the largest fitness is added to G , yielding a larger community G_1 .
- c) The fitness of each node of G_1 is recalculated.
- d) If a node turns out to have negative fitness, it is removed from G_1 , yielding a new subgraph G_2 .
- e) If step d) occurs, repeat from step c), otherwise repeat from step a) for subgraph G_2 .

The process stops when the nodes examined in step a) all have negative fitness.

The overlapping community detection method proposed consists of the following steps:

- a) Compute each node's vertex strength which is the sum of the weights of the edges attached to the node.
- b) Sort the nodes in the descending order of their vertex strengths. While traversing all the sorted nodes.
- c) If a node is not assigned to any community, initialize a community G including only one node. If assigned, skip the node.
- d) Perform the above community forming steps. Now a community is generated.
- e) Stop if all the nodes have been assigned to at least one community. Otherwise, go to step c).

B. Computing Membership Degree of Overlapping Nodes

If a node is assigned to more than one community, it becomes an overlapping node. Assuming that node p is assigned to n communities. Membership degree is given by (8) as follows.

$$a_i = \sum_{q \in C_i} \frac{w_{pq} \in C_i}{m} \quad (8)$$

where a_i is the membership degree of node p belonging to community C_i , m is the number of communities containing the node q , w_{pq} is the weight of the edge connecting p and q . The output is an membership array $[a_1, a_2, \dots, a_n]$. The array should be normalized to create the standard for comparison, meaning that the sum of the elements in the array will equal to 1.0.

C. Predicting the User's Interests by Its Communities

The interest of a user is not only decided by the user's MUM, interest of the communities the user belongs to is also playing an important role. The user's MUM updated with the community model can predict the interest of the user.

1) The Community Model

A community model reflects the interest of the community. The interest of the community is decided by its

MUMs. Each MUM has a different influence weight. If a node's vertex strength is very large, it will have a larger influence weight. The interest of the community is derived by (9) as follows.

$$C^{int} = \sum_{i=1}^{|C|} u_i \times \ln(W_i + 1) \quad (9)$$

where C^{int} is the interest of the community C , $|C|$ indicates the number of MUMs in C , u_i is one MUM, and W_i is the sum of the weights of internal edges attached to the node in community C .

C^{int} is a vector that is similar to MUM and it should be normalized to create the standard for comparison. Every element in C^{int} will be divided by the sum of elements in C^{int} . At last, the sum of the elements in C^{int} equals to 1.0.

2) Interest prediction

The predicted interest is calculated by (10) as follows.

$$u_i^{int} = \alpha u_i + (1 - \alpha) \sum_{j=1}^n a_j C_j^{int} \quad (10)$$

where u_i^{int} is the predicted interest. u_i is a MUM, C_i is a community, n is the number of communities u_i belongs to, a_j is the membership degree of u_i belonging to C_j , C_j^{int} is the community model of the community u_i belongs to, and α is an adjustable value which represents the influence weight of the user's own interest. When $\alpha=1$, the predicted interest is not influenced by the community model.

IV. EXPERIMENTS

A. Experiment Data

The data set used in this paper was collected from Tencent microblog API (<http://dev.t.qq.com/>). Once every 15 days, 12,761 users' base information, followings, followers and microblog content about 70 latest blogs were collected. A total of 14 batches of data were collected and stored. In this paper, the second and the sixth batches of data were used. The first batch of data is not very perfect, it is the first collection after all. So the second batch of data is selected as the training set. The sixth batch of data is collected two months later after the second collection and it is considered as the test set.

Because there are so many keywords in MUMOB that requires intensive computation time when constructing the MUM network, in this paper only the top two levels are considered after MUMOB is constructed. It does not mean that we never use the third level and the fourth level. All the four levels are used in the process of constructing the MUMOB. After we construct MUMOB, we only select the top two levels of MUMOB to use. Thus, a MUM can be presented as a 100-dimension vector with 11 keywords at the top level and 89 keywords at the second level. Each element of the vector represents the user's interest value to the corresponding keyword.

In (3), we set $\mu=0.8$, because the user's own interest to a keyword plays dominant role and the followings' interest values to the keyword have had a small influence. For more accurate value of μ , please wait for follow-up researchers' further study.

MUM has two levels, so we set $\phi=2$ in (5). And the second level has about 8 times as many keywords as the top level. Thus the second level has larger influence weights, which is about 8 times as large as the weight of the first level. Therefore, we set $\gamma=64$ in $W(h)$ of Equation (5). If we do not take the computation quantity into consideration, we can set $\phi=3$ in (5). In this case, the third level has about 97.1 times as many keywords as the top level. Then γ is calculated and it equals to 97. Generally, MUM has no more than three levels.

Initially, there is an edge connecting each pair of nodes. There are so many edges that it is difficult to partition the network. So we select a threshold M to remove some edges. Only the edges whose weight greater than the threshold are retained. In this paper, $M = 0.497$. The MUM network created from the collected data contains 12,761 nodes and 250,113 edges. It was processed by our OCD-MUM method and the resultant partitions are shown in Fig. 3. It was drawn using the Gephi 0.8.2 software.

We pick out a part of Fig. 3. As shown in Fig. 4, there are 3 communities: 211, 223 and 248. The number on the node represents its community id. Three nodes are shared between 211 and 223. There are no overlapping nodes between 223 and 248.

The comparison of partition result is shown in TABLE I. N is the number of the edges in the graph. Our OCD-MUM algorithm partitioned a total of 252 communities, a number which is better than the partition result of 87 communities by the CSD-IS algorithm because of the bigger modularity Q . The modularity Q is given by (11) which was proposed in [9] by Newman. It is suitable for weighted networks. In practice this is an excellent guide to whether a particular division of a network into communities is a good one. Nonzero values around 0.3 or more usually indicate

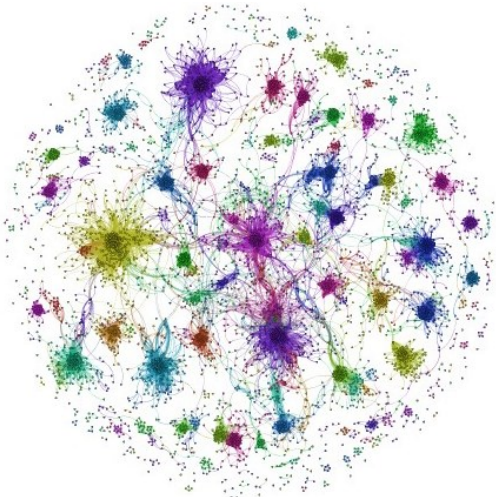


Fig. 3. Overlapping communities detected in the MUM network

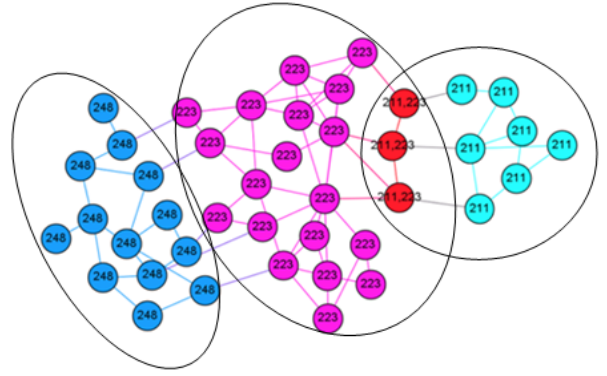


Fig. 4. Three communities and three overlapping nodes.

good divisions. The maximum possible value of Q is 1 [9].

$$Q = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \quad (11)$$

where A_{ij} represents the weight of the edge between i and j , k_i is the vertex strength of a vertex i , and $m = \frac{1}{2} \sum_{i,j} A_{ij}$. c_i is the community to which vertex i is assigned, $\delta(c_i, c_j)$ is 1 if $c_i=c_j$ and 0 otherwise.

TABLE I. THE COMPARISON OF PARTITION RESULT BETWEEN THE TWO METHODS

| Algorithm | Number | Modularity Q | Time Complexity |
|-----------|--------|--------------|-----------------|
| OCD-MUM | 252 | 0.473 | $O(n^2)$ |
| CSD-IS | 87 | 0.375 | $O(N^4/n^2)$ |

B. Evaluation Criteria

The MUM constructed with the sixth batches of the collected data reflects the user's practical interest. The similarity between the predicted interest and the practical interest can be computed by using the cosine function. Now there are two groups of communities: the first group has 252 communities detected by the OCD-MUM and the second group has 87 communities identified by CSD-IS proposed in [1]. Assuming that a MUM is in community A belonging to the first group while in community B belonging to the second group. Community A corresponds to a community model while community B corresponds to a community model too. The MUM will be updated by two community models to get two predicted interest values. Then we will compute which one is more close to the practical interest.

Computing the similarity between predicted interest and practical interest is done with the following steps:

- Constructing MUMs with the second batch of data. The interest of each user is computed and stored in a 100-dimension vector.
- Updating each MUM with the community model of the community the user belongs to. The predicted interest is computed and stored in a 100-dimension vector.

- c) Constructing MUMs with the sixth batch of data. The practical interest is calculated and stored in another 100-dimension vector.
- d) Computing the similarity between the predicted interest and the practical interest using the cosine function.

A total of 1,477 users were selected to compare the two community partition results. Because some users do not often write blogs, we select 1477 users who are active and often update their blogs. Most of them write or forward at least 15 blogs during the two months. When the predicted interest is not influenced by the community model, the 1,477 similarity values were computed, and used as benchmark. The similarity values were calculated with the α values ranging from 0.4 to 1.0. As shown in TABLE II.

TABLE II. SIMILARITY BETWEEN THE PREDICTED INTEREST AND THE PRACTICAL INTEREST

| α | 0.4 | | 0.45 | | ... | 1.0 | |
|------------|---------|--------|---------|--------|-----|---------|--------|
| | OCD-MUM | CSD-IS | OCD-MUM | CSD-IS | | OCD-MUM | CSD-IS |
| U_1 | 0.544 | 0.882 | 0.569 | 0.874 | ... | 0.761 | 0.761 |
| U_2 | 0.276 | 0.269 | 0.280 | 0.273 | ... | 0.302 | 0.302 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| U_{1477} | 0.501 | 0.871 | 0.537 | 0.868 | ... | 0.801 | 0.801 |
| average | 0.544 | 0.534 | 0.549 | 0.538 | ... | 0.553 | 0.553 |

C. Comparing the Prediction Result

TABLE III lists the similarity values that are greater than the benchmark of two methods, where columns “Count1” and “Count2” each represent the number of similarity values greater than the corresponding elements in benchmark, and columns “Percent1” and “Percent2” are the improved percentage over the benchmark. TABLE III shows that “Count1” is always greater than “Count2”. This is a very positive indication that our OCD-MUM method is superior to the CSD-IS algorithm in terms of detecting reasonable communities.

TABLE III. THE NUMBER OF SIMILARITY VALUES GREATER THAN BENCHMARK

| α | OCD-MUM | | CSD-IS | |
|----------|---------|----------|--------|----------|
| | Count1 | Percent1 | Count2 | Percent2 |
| 0.40 | 709 | 48.00% | 667 | 45.16% |
| 0.45 | 729 | 49.36% | 686 | 46.45% |
| 0.50 | 753 | 50.98% | 707 | 47.87% |
| 0.55 | 768 | 52.00% | 726 | 49.15% |
| 0.60 | 799 | 54.10% | 738 | 49.97% |
| 0.65 | 823 | 55.72% | 755 | 51.12% |
| 0.70 | 846 | 57.28% | 773 | 52.34% |
| 0.75 | 868 | 58.77% | 789 | 53.42% |
| 0.80 | 886 | 59.99% | 817 | 55.31% |
| 0.85 | 914 | 61.88% | 837 | 56.67% |
| 0.90 | 954 | 64.59% | 850 | 57.55% |
| 0.95 | 985 | 66.69% | 864 | 58.50% |
| 1.00 | 0 | 0 | 0 | 0 |

In Fig. 5, the average of similarity values were calculated with the α values ranging from 0.4 to 1.0. When $\alpha=0.8$, the proposed OCD-MUM method shows the best performance. Hence, we set $\alpha=0.8$.

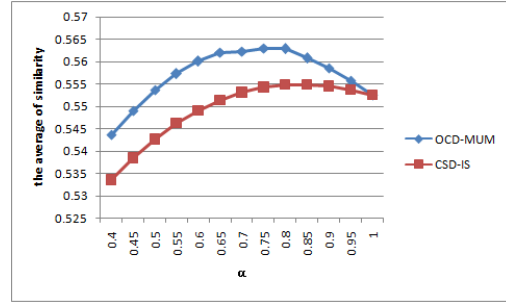


Fig. 5. The average of similarity values.

V. CONCLUSION AND FUTURE WORK

Presented in this work is our study of the problem of detecting overlapping communities in social network based on the MUM. The main contributions of this study are as follows.

First, the MUM is constructed using base information, follower-following relationship and microblog content. The BEOB is crawled from the Baidu Encyclopedia. It has relatively comprehensive classification keywords. The MUM can reflect user’s interest more accurately. Furthermore, the MUM network is constructed by computing the similarity between MUMs. Different levels of MUM have different influence weights. The hyperbolic tangent function is used to reflect the unbalanced weight. Finally, the proposed OCD-MUM method can better detect the overlapping communities of similar interest. Our method shows a higher modularity Q value than the CSD-IS method. The interest prediction result demonstrates that the MUM network is better identified by OCD-MUM.

Although the experiment result shows that the OCD-MUM has great effectiveness, there is still a lot of work to be done. In addition to the users’ microblog content and followers-following relationship, other factors such as the tag, comment, forward and reply interaction relationship can be taken into further consideration to construct the MUM. Another direction is to construct MUMs with more levels to represent the user’s interest in finer-grain.

ACKNOWLEDGMENT

The authors are very grateful to the anonymous reviewers. Their many helpful and constructive comments and suggestions helped us significantly improve this work.

REFERENCES

[1] Han R, Meng S, Lu Y, et al. "Research on algorithm of community structure detection based interest similarity". Railway computer application, 2010.

- [2] Xing D, Wang X. "Community Discovery Technology Research based on MicroblogMedia". Intelligent computer and applications, 2013.
- [3] Yan G, Shu X, Ma Z, Li X. "Community discovery for microblog based on topic and link analysis". Application Research of Computers, 2013.
- [4] Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." Nature 435.7043 (2005): 814-818.
- [5] Baumes, Jeffrey, et al. "Finding communities by clustering a graph into overlapping subgraphs." IADIS AC 5 (2005): 97-104.
- [6] Lancichinetti, Andrea, Santo Fortunato, and János Kertész. "Detecting the overlapping and hierarchical community structure in complex networks." New Journal of Physics 11.3 (2009): 033015.
- [7] Gregory, Steve. "An algorithm to find overlapping community structure in networks." Knowledge discovery in databases: PKDD 2007. Springer Berlin Heidelberg, 2007. 91-102.
- [8] Gregory, Steve. "Finding overlapping communities in networks by label propagation." New Journal of Physics 12.10 (2010): 103018.
- [9] Newman, Mark EJ. "Analysis of weighted networks." Physical Review E 70.5 (2004): 056131.
- [10] Tang, Fan, et al. "Friend Recommendation Based on the Similarity of Micro-blog User Model." Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCoM), IEEE International Conference on and IEEE Cyber, Physical and Social Computing. IEEE, 2013.
- [11] Calegari, Silvia, and Gabriella Pasi. "Personal ontologies: Generation of user profiles based on the YAGO ontology." Information processing & management 49.3 (2013): 640-658.
- [12] M Chowdhury, Gobinda. "Introduction to modern information retrieval." Facet publishing, 2010.