Recommending Funding Collaborators with Scholar Social Networks

Juan Zhao, Kejun Dong, Jianjun Yu
Computer Network Information Center
Chinese Academy of Sciences
Beijing, China
{zhaojuan, kevin, yujj}@cnic.cn

Abstract— **Applying for research funding projects is becoming one of the most important ways for scientists to carry on the research. How to find an appropriate collaborator/applicant is a major concern for scientists. Social networks provide one means of visualizing existing and potential collaborations. In this paper, we study the funding collaborators recommendation problems. We solve the problem by starting with analyzing the researchers' motivations for finding collaboration, which are (i) to form a competitive team (ii) to expand cooperation circle, which little work noticed. We model the funding relation as a complex network called co-applicant network. Based on that, we propose a utility function to take all the aspects of recommendation into account. And we propose a novel recommendation algorithm by modeling the utility function based on the group relations in the co-applicant network. We experiment our approaches on National Science Foundation of China (NSFC) funding projects and achieve effective results.**

Keywords—funding collaboration recommendation; social network; recommender system

I. INTRODUCTION

Applying for research funding projects is becoming one of the most important ways for researchers to carry on the research. Partially driven by the stricter conditions and more competitiveness on the application, finding and choosing competitive collaborators become more and more important for researchers to obtain the funding. Due to the limit energies on meeting at conferences and casual conversations that bring to little knowledge about a potential collaborator, researchers need more effective ways to find appropriate collaborators. Despite a great deal of theoretical development in information retrieval and data mining techniques, advanced search tools for finding collaborators on funding project application are still in their infancy. Currently, recommending collaboration with social networks has been well studied and could be effectively helpful [1][2], but most are concentrated in the co-authorship, which is different from funding collaboration.

Seeking collaborations could be viewed as a team formulation problem. Studies on team formulation propose two important factors in forming an effective team, which are the degree of competence of the team members and the degree of cohesiveness among them. The former evaluates whether the team formed by members can cover all the required skills. The cohesiveness is the evaluation on how the team can function together well [1]. Studies [2][3] [4] propose methods based on graph theory to measure cohesiveness, such as the communication cost, the density of the subgraph and so on, based on the assumption that it is easier for people to work with the acquainted people.

However, a team formed by the acquainted people would work effectively, and may not help bring in new resources and opportunities, which could be more important in obtaining national funding. Therefore, researchers usually would like to find new researchers or groups to expand the cooperation circle, which will yield different strategies for choosing collaborators. In order to achieve cohesiveness, a well connected sub-graph would be the best choices in the graph. On the contrary, to enlarge cooperation circle, how to expand the egocentric network in the graph should be considered.

This paper considers that forming an effective team and finding new collaborators to enlarge the cooperation circle are two major important motivations for researchers in finding and choosing collaborations. The team formulation problem mainly focused on the first motivations, but little work noticed the second one. And how to combine both motivations would be a big challenge. We solve the problem by evaluating the utility /goodness which the collaborators could bring in. There are four factors which influence the utility, 1) the competence of the collaborators, 2) the communication cost, both of which need to be considered in team formulation, 3) the degree of the egocentric network expansion, which is about the second motivation, 4) the possibility that the collaborators would collaborate with, which is not related about motivation but about the recommendation effectiveness. Thus, the problem becomes recommending collaborators, which would bring more utilities to the target user.

How to design a novel utility function to make balanced among all these factors is still challenged, as reducing the communication cost and expanding degree of the egocentric network expansion are conflicted. We are inspired by the deep observations that the collaborations between researchers are closely influenced by the group relations in which the researchers are. A group is several persons who collaborated more frequently. If a person in the group has collaborated with the target user, then other persons in the same group would be easy to get to know and build trust.

Based on these observations, we model the funding relation as a complex network called co-applicant network. We propose a novel recommendation approach to finding new trusted collaborators, and model the utility function based on the group relations in the co-applicant network that addresses all three challenges including balancing different factors influencing the utility function. We collect the NSFC funding dataset and experiment our approaches on it. The experimental results show that our methods perform very well and satisfy the target users' needs.

II. Related work

A. Expert Recommender Systems

Collaborator recommendation is closely related to expert recommender (ER) [5], which intend to recommend qualified experts to a user who has a need for a particular expertise. An early ER typically builds profiles on experts by extracting expertise related data, such as publication information from websites, and then selects the experts whose profile match with the target user. Then the ER attempts to use social networks as a mechanism for recommending the experts who may be more available and willing to interact. ReferralWeb [6] was a typical ER, where co-authoring and co-citation relationships are mined to create a social network and the resulting visualization is used to find a possible expert. The social network in ReferralWeb can also compute the social distance between two researchers in order to answer queries about how far one researcher is from another and who is between.

The main difference of collaborator recommendation is that when seeking a collaborator, factors additional to expertise and connectivity need to be included, such as the mutual benefit, collaboration style, or experience in the field, which play a part in determining the collaboration.

B. Team formation.

Lappas et al. [1] proposed the problem of team formation: finding a group of individuals who can function as a team to accomplish a specific task in a given social network. The formulation of the problem is that given a social network where nodes are labeled with a set of skills, and given a task that requires a certain set of skills to be satisfied, the goal is to find a connected subgraph in which all skills are present and the cohesiveness is good. [2], [3], [4] use different methods based on graph theory to measure cohesiveness, such as the communication cost, the density of the subgraph and so on.

The difference in our work with the above papers is that strategy of finding a collaborator is different from finding a team member. As collaboration is a process where people or organizations work together toward an intersection of common goal, such as sharing knowledge in co-publishing papers or complementing strength in the funding application, thus the cohesiveness is less important than the mutual benefit between each other.

C. Link prediction

The problem is that we are given a graph and a node *s* for which we would like to predict new links, which could be used in recommending relations in social networks. The link prediction problem was initially introduced in [7]. The study examined the effectiveness of several simple heuristics such as the number of common friends, node proximity and PageRank [8] in the prediction.

Our study is different from these research works in goals and solutions. These studies focus on predicting or recommending the future relations based on the current interaction behaviors. However, in some cases, researchers may always choose previously collaborated ones due to the limitations on space and information systems. Our goal is to provide more new options for researchers to broaden their collaborations.

III. Modeling Funding Network

Formally, a collaboration network can be defined as a graph, which is composed of nodes, where each node is an individual. Edge represents a binary collaboration relationship among nodes. Examples of the graph from the introduction are a co-author network, where nodes are authors and edges are formed by their co-author relationship. We use a co-applicant network modeling the network of applying funding projects among scientists. In the **co-applicant network**, nodes are applicants who have applied for the national funding programs, and edges are built if two applicants have been co-applicants in a program.

The simplest graph is the undirected graph without assuming weights, however, many of the concepts we consider in this paper could be generalized to the case of weights and directions. The weight of the edge is mainly determined by their collaborated frequency. We take the Newman formula1.It can be applied to different types of collaboration networks with different means of the parameters.

$$w_{ij} = \sum_{p} \frac{\delta_i^p \delta_j^p}{n_p - 1} \quad (i \neq j)^1 \qquad (1)$$

In the co-applicant network, *p* is the project number, n_p is the total applicants in the project. If δ_i^p is equal to 1, *i* will be the one of the applicants in project *p*, otherwise δ_i^p is equal to 0, *i* will not be the applicant in project *p*. We claim that edge weight is an important indicator of the strength of the relationship between the individuals.

A. Recommending model based on clusters

Definition 1. **Group (cluster)**. In the co-applicant network, a group is a set of individual nodes who collaborated with each other more often. Groups are not completely isolated or not connected. Groups do not necessarily correspond to a real organization in the real world.

If several persons in a group have collaborated with others in another group, we think that:

1) The two groups have some relations. Thus the communication cost between these two groups would be lower compared with other groups which have no direct relation.

2) The two groups which have some cooperation are not possible to be competitors.

3) The cooperation of these two groups is not enough to be combined into one group, so the cooperation between the two groups can be further promoted.

We give another two definitions:

Definition 2. **Group distance**. The group distance between two groups X and Y is represented by the links between nodes in group X with nodes in group Y.

$$DC(X,Y)=\sum_{l=1}\text{paths}^{(l)}\,DC(x_i,y_j) \qquad (2)$$

Definition 3. **Group closeness**. If group distance between two groups X and Y is beyond a threshold *n*, *DC(X,Y)*>=n, X and Y are close.

Our recommending strategy is as follows:

1) For a node *v* in a group A, the candidates should be in other groups which have the closeness with A. In Fig. 1, group A (blue) and group B (orange) are much closer to each other. Group A has some collaboration with B and C (green). For a node *v* in A, the candidate collaborators are in group B and C, marked with a tick.

2) The nodes which have direct links with node *v* could not be recommended. They have cooperated with each other which mean that they have collaborated and do not need to be recommended. It's just like that a book which you already read doesn't need to be recommended again. In Fig. 1, those nodes which have direct edges with *v* are marked with a tick.

3) The nodes in the groups which have no relations with group A should not be recommended. Like those nodes in group D (yellow), marked with a cross, as group A and group D have no collaboration (Fig.1).

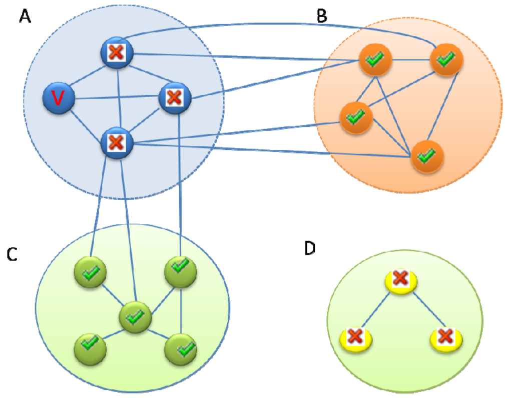


Fig. 1. Co-applicant network model

IV. Problem Statement

Given an undirected graph $G=(V,E)$, $V=\{v_i\}$, $E=\{e(v_i,v_j)|\ v_i,v_j \in V\}$, we have a target node *v*, which we should recommend collaborators for. Suppose we recommend a different node *u* to *v*. Once *v* accepts *u*, there will be edge *v* and u_i in the future, denoted by $e'(v,u_i)$. We define the utility function, denoted by $f(e'(v,u_i))$, to evaluate the utility of the edge formed by *v* and u_i. Our main task is to find those nodes which have better utility function values by constructing edge with the target node. In this section, we first describe the utility function in detail and then present the calculation of the elements in the function.

A. Problem definition

Given a graph *G*=(*V*, *E*), the target node *v* and the requirements $Q=\{q_1,q_2,\ldots,q_t\}$ for the collaborators, the objective is to find nodes $U=\{u_i\}\subseteq V$, so that node u_i could satisfy the requirements *Q* and the utility function $f(e'(v,u_i)) > \varphi$, φ is a threshold.

B. Utility function

The utility function evaluates the utility that the target node builds a relationship with a candidate node. The utility function mainly contains two parts: (1) advantage function denoted as $advantage(e(u,v))$ to evaluate the advantages that the target node could bring; (2) possibility function, to evaluate the possibilities that the two nodes would collaborate. Because recommending a person who is unlikely to collaborate with has no means to the target user.

$$f(e(u,v)) = advantage(e(u,v)) * possibility(e(u,v)) \quad (3)$$

C. The advantage function

The advantage function contains two parts: 1) the competence of the node *v*. *2) the* edge *e (u,v)* which could bring to the egocentric network of the target node.

In a funding application, people with a good authority and experience in funding program would be preferred. We observe that the more projects the applicant applied for, the more likely to be successful in applying next program. So the relationship in the co-applicant network is used to find a collaborator who has good authority. So we use authority value to denote the competence of a node and use the random walk algorithm [8] to compute the authority value.

$$p_{i+1} = (1-\beta)E^T p_i + \frac{\beta}{N}(l_N)p_i = \left((1-\beta)E^T + \frac{\beta}{N}l_N\right)p_i \qquad (4)$$

For evaluating the advantages of building the relationship, we use the extended degree of the edge, which demonstrates to what large extent, the egocentric network of the target node could extend with establishing the new edge.

For example, in Fig. 2 (a), Node A has no relation with node C, D, E, F before connecting with node B. Therefore, the distance between A and C, D, E, F is infinite. When A connects to B, A could arrive at node C, D, E and F in two walks.

In another scenario, as shown in Fig. 2 (b), node A is directly connected with node C, D, E and F, thus, node A could arrive at them in one walk. After connecting with B, the node A could arrive at the node C, D, E in two walks. The shortest distance between them is not reduced but the accessible paths increased. Edge (A, B) would bring a larger extended degree to A compared with the previous scenario in Fig. 2 (a).

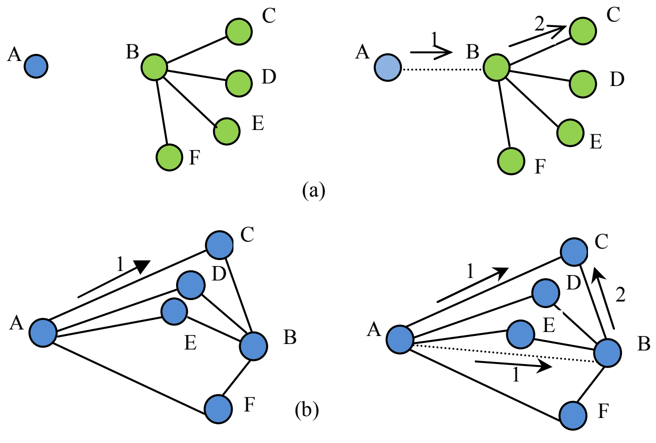


Fig. 2. The demonstration of extended degree

Definition. Extended degree. Given a graph *G*=(*V, E*), a node *v* connects with the node *u*, the increased nodes that *v* could arrive at in a limited walks.

$$extend(e(u,v)) = \sum_{l=1} (A^{l}{}_{G+e(u,v)}(v) - A^{l}{}_{G}(v)) \qquad (5)$$

$A^{l}{}_{G}(v)$ denotes the total number of the nodes that *v* could arrive at in one walk in the graph G.

Extended degree between different groups

Our recommending strategy is to encourage more collaboration between groups. For a given node *v*, connecting with nodes in different groups could bring more extend degree goodness comparing with connecting with those in the same group. Therefore, we define the extended degree between different groups.

Given a graph *G*=(*V, E*), a node *v* in a group A connects with the node *u* in another group B, the increased nodes in the group B that *v* could arrive at in a limited walks.

$$extend_{u\in C_i, v\in C_j}(e(u,v)) = \sum_{l=1} (A^{l}{}_{G+e(u,v),C_i}(v) - A^{l}{}_{G,C_i}(v)), i \neq j \qquad (6)$$

$A^{l}{}_{G,C}(v)$ denotes the total number of the nodes in the group C that could arrive at in one walks in the graph G.

For example, in Fig. 3, given the node A in the cluster C2, compare the extend degree of connecting to the node B and node C in cluster C1.The degree of the node C is three, connected with two nodes in C1 and one node in C2. If node A connects with node C, node A would relate with the two nodes, which node C connects with in C1.The degree of the node B is five, which is connected with nodes in the same cluster. If node A connects to node B, node A would relate with the five nodes in the C1. Therefore, the extend degree of edge (A,B) would be better than the extend degree of edge (A,C).

To sum up, we combine the authority function (Pagerank) and extend degree function to denote the advantage function.

$$advantage(e(u,v)) = \lambda \cdot p_u + \gamma \cdot extend_{u\in C_i, v\in C_j}(e(u,v)) \qquad (7)$$

Where, λ and γ are the factors which could influence the portion of the authority and extend degree in the advantage function.

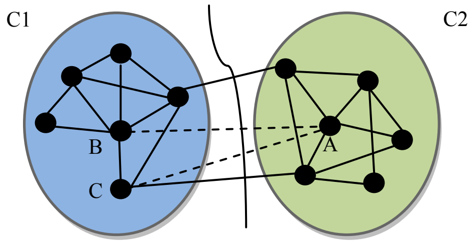


Fig. 3. The demonstration of extended degree between different groups

D. The possibility function

We define the possibility function $possiblity(e(u,v))$ to evaluate the possibility of establishing the edge between node *v* and node *u*. We observed that two acquainted people would easily establish the relation. Thus the possibility could be computed by the distance between the two nodes.

$$possiblity(e(u,v)) = 1 / (distance(u,v)) \qquad (8)$$

When two nodes have no relations, *distance(u,v)* could be infinite. We assume *possibility (e(u,v))*=0 in this situation.

We also give the definition of the possibility in the clusters.

$$possiblity_{u\in C_i, v\in C_j}(e(u,v)) = \frac{1}{distance(u,v) * DC(C_i, C_j)} \qquad (9)$$

Finally, we combine the advantage and possibility function to form the utility function.

$$f(e(u,v)) = (\alpha p_u + (1-\alpha) extend(e(u,v))) * \frac{1}{distance(u,v)} \qquad (10)$$

The utility function based on clusters is:

$$f(e(u,v)) = (\alpha p_u + (1-\alpha) extend(e(u,v))) \times \frac{1}{dist(u,v) * Distance(C(u), C(v))} \qquad (11)$$

V. COLLABORATORS RECOMMENDING ALGORITHMS

In this section, we give our algorithms on recommending collaborators. Assume we get a collection of applicants in a nationally funded program in the past few years, a source person *s* and a query *q* on the applicant's subject. We collect the basic profile information Prf(*v*) of each person by deep web technology from the Web.

1) *Topic matching.*

By computing *rel_l(v, q)*, the relevancy of a person to a topic *q,* based on one's profile, and filtering low *rel_l(v, q)* , we get a candidates collection $V\{v_1, v_2 \ldots v_n\}$.

2) *Group Clustering*

In this step, we use an undirected graph. The relationship between the nodes is co-applicant. We use a cluster to present the intensively collaborated groups and the bottom–to-up hierarchal clustering method Girvan-Newman [9]. Firstly, we pick each node on the *V* as the center node and find all the nodes that have a relationship with it to form the initial cluster $C\{c_1, c_2 \ldots c_m\}$. Then we use a group proximity measure computing algorithm based on link-based proximity [10]. We calculate the links between clusters based on the relationship between persons on these clusters. If the person on the cluster c_i has a relationship with another person on the c_j, there is a link between c_i and c_j. If the number of links is beyond a threshold *N*, c_i is supposed to be close to c_j , and they can be merged as a new cluster. After several iterations, we get the final close clusters $C'\{c_1, c_2 \ldots c_t\}$. The number of links between each pair is less than *N*.

Scan the candidates collection $V\{v_1, v_2 \ldots v_n\}$ to select those nodes that are in the close clusters but not directly connected to the target node to form a new candidates collection $V'\{v_1, v_2 \ldots v_m\}$.

3) *Computing the utility function value of each candidate node*

For each candidate node in $V^{'}$ $\{v_1, v_2 \ldots v_m\}$, compute the utility function *f(e(s,v$_i$)) in* formula 11.

4) *Identifying and ranking the final collaborators.*

Set the threshold φ to pick up the nodes whose utility function value *f(e(s,v$_i$)) is* above φ to form the final recommending collection. We rank the nodes in the descending order in accordance with the function value.

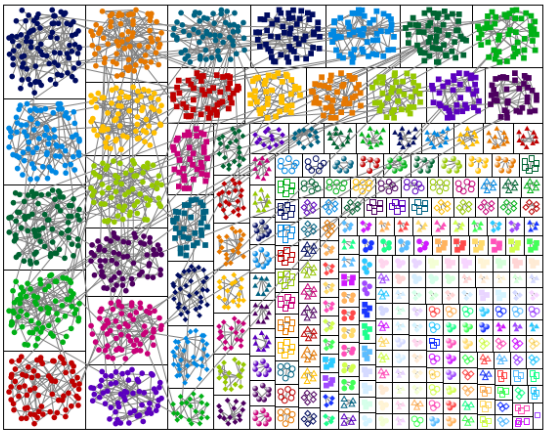
VI. EXPEREIMENTS AND EVALUATION

We collected the 61,483 applicants and 193,850 relations from the National Science Foundation of China (NSFC). Taking into account the actual amount of computation and realities, we selected the funding program applications in the computer field since 2000 year to form the experimental data. The proposed algorithm is based on cluster collaboration. In order to improve computational efficiency, we filter out those isolated edges (the edge of the two nodes of degree 1), which don't contribute to the algorithm.

We finally get an undirected graph composed of 2575 nodes and 2868 edges. We implement the Girvan-Newman algorithm based on the undirected graph to get 266 clusters shown in Fig. 4. The same group of nodes in the cluster is placed in a box, and the same colors, and the type of association between the edge of the group can be displayed.

We present an example to demonstrate our recommendation algorithm. Take the person "Lu Huaxiang" as the target person, whom we intend to recommend collaborators for. His research topic is "Network", so firstly we filter out the applicants with the keyword. We get the initial candidate set which contains 544 candidate nodes. In the Fig. 5 to Fig. 6 as follows, we use node A0 to represent the target person “"Lu Huaxiang".

Then in the global co-applicant network, given the target node in a cluster, we compute the distance between this cluster and other related clusters. We set the average distance as the threshold φ and get three close clusters which distance is above the threshold. Then scanning the candidate set and selecting the nodes in the close cluster but not directly connected to the target node, the candidate sets down to 35 nodes.



Created with NodeXL (http://nodexl.codeplex.com)

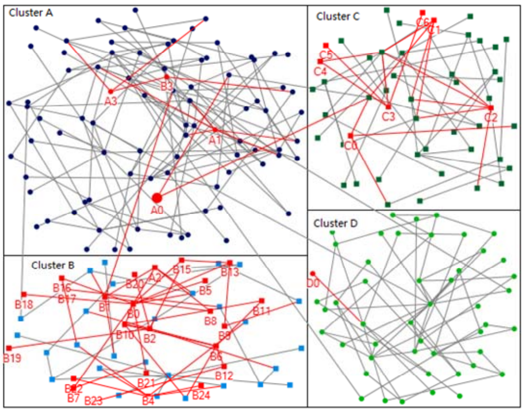
Fig. 4. Clusters by Girvan-Newman

Next we compute the Pagerank value, the extended degree value and the possibility function value, then we get the utility function value of each candidate node and rank the candidate nodes in the descending order of their utility function value. The Table I shows the top 10 results. The same character means the nodes are in the same cluster, for example, C0 and C1 are in the same cluster C. The results show that only consider the Pagerank may recommend a distant candidate which is not the best choice. The utility function considers those factors such as Pagerank, collaboration possibility, extended degree together and makes a good balance. Next we visualize the recommendation results in different views.

TABLE I. RECOMMENDATION RESULTS

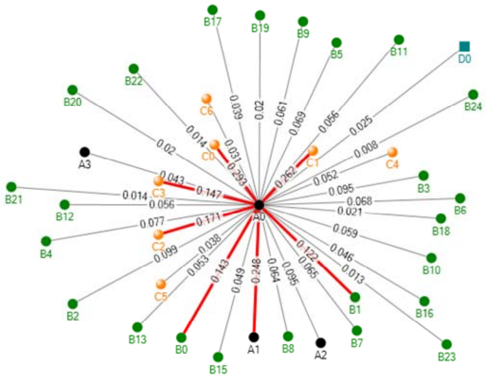
FIGUREID	**POSSIBLITY**	**PAGERANK**	**EXTEND**	***UTILITY FUNCTION VALUE***	**DISTANCE**
C0	0.25	0.973985	0.2	**0.293**	2
C1	0.25	0.748374	0.3	**0.262**	2
A1	0.166667	0.990606	0.5	**0.248**	6
C2	0.125	0.865443	0.5	**0.171**	4
C3	0.125	0.773233	0.4	**0.147**	4
B0	0.071429	1	1	**0.143**	7
B1	0.1	0.818326	0.4	**0.122**	5
B2	0.055556	0.982512	0.8	**0.099**	9
B3	0.083333	0.738257	0.4	**0.095**	6
A2	0.125	0.560486	0.2	**0.095**	8

Fig. 5 visualizes the relationship between the candidate nodes and the target node in a group / cluster view. The nodes in the same group are in the same shape and color, while the recommended nodes are marked in red. The results show that the recommended nodes are distributed throughout the network, not limited in the same cluster with the target node. Fig. 6 shows the relations between the recommended nodes with the target node in the co-applicant network, which the distance between the nodes represents the actual distance in the co-applicant network and the value of the edge is the value of the function value. In Fig.6, we can see that the top recommended nodes are in close distance to the target user.



Created with NodeXL (http://nodexl.codeplex.com)

Fig. 5. Distribution of recommended candidate nodes in the co-applicant network



Created with NodeXL (http://nodexl.codeplex.com)

Fig. 6. Visualization of the recommendation results

The experimental results show that the advantage of our recommendation algorithm is recommending nodes distributed in a different group from the target node, reaching circle expansion purposes, and is conducive to balance looking for a complimented collaborator and expanding their personal relationship circles.

VII. Conclusions

In this paper, we have applied the social network analysis to the study of collaborative recommendation for funding program application. We proposed a funding network model and approaches on recommending the collaborators. We gave the evaluation of the experiment. As the future work, we plan to combine the co-authorship network mining to find more completed research collaboration.

Acknowledgment

The research was supported by the Chinese Academy of Sciences Special Fund under Grant XXH12503 and Around Five Top Priorities of “One-Three-Five” Strategic Planning, CNIC under Grant No. CNIC_PY-14XX.

REFERENCES

[1] Kun Liu,Theodoros Lappas. Finding a Team of Experts in Social Networks [C].KDD'09, Paris,France: 2009.

[2] The community-search problem and how to plan a successful cocktail party, Sozio, Mauro and Gionis, Aristides. KDD’ 2010 ,939—948.

[3] Anagnostopoulos, Aris and Becchetti, Luca and Castillo, Carlos and Gionis, Aristides and Leonardi, Stefano. Online team formation in social networks .WWW’ 2012 ,839—848

[4] Capacitated team formation problem on social networks, Majumder, Anirban and Datta, Samik and Naidu, K.V.M.. KDD’ 2012 ,1005--1013

[5] Perugini, S., Gon, calves, M.A., Fox, E.A.: Recommender systems research: A

[6] Henry Kautz, Bart Selman, Mehul Shah, “Referral Web: combining social networks and collaborative filtering”, Communications of the ACM , Volume 40 Issue 3, March 1997.

[7] Liben-Nowell, D. and J. Kleinberg. The link prediction problem for social networks. in CIKM '03. 2003. New York, NY, USA: ACM.

[8] Arasu, A., Novak, J., Tomkins, A., Tomlin, J. 2002.PageRank computation and the structure of the web:Experiments and algorithms. WWW’2002.

[9] B. Saha and L. Getoor.2008. Group proximity measure for recommending groups in online social networks. In The 2nd SNA-KDD Workshop'08. ACM, August 2008

[10] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "14.3.12 Hierarchical clustering" The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 520–528