

Research Article

Quantitative Similarity Evaluation of Internet Social Network Entities Based on Supernetwork

Xiuzhen Chen,¹ Shenghong Li,^{1,2} Jianhua Li,^{1,2} and Zhiyuan Zhang¹

¹ School of Information Security Engineering, Shanghai Jiaotong University, Shanghai, China

² School of Electronic, Information, and Electrical Engineering, Shanghai Jiaotong University, Shanghai, China

Correspondence should be addressed to Xiuzhen Chen, chenxz@sjtu.edu.cn

Received 17 March 2011; Revised 10 May 2011; Accepted 11 May 2011

Academic Editor: Christian Callegari

Copyright © 2011 Xiuzhen Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to accurately characterize similarities of entities is the basis of detecting virtual community structure of an Internet social network. This paper proposes a supernetwork based approach of quantitative similarity evaluation among entities with two indices of friend relation and interest similarity. The supernetwork theory is firstly introduced to model the complex relationship of online social network entities by integrating three basic networks: entity, action, and interest and establishing three kinds of mappings: from entity to action, from action to interest, and from entity to interest, that is, one hidden relation mined through the transfer characteristic of visible mappings. And further similarity degree between two entities is calculated by weighting the values of two indices: friend relation and interest similarity. Experiments show that this model not only can provide a more realistic relation of individual users within an Internet social network, but also, build a weighted social network, that is, a graph in which user entities are vertices and similarities are edges, on which the values record their similarity strength relative to one another.

1. Introduction

Many systems can be represented as complex networks or graphs, that is, collections of vertices joined in pairs by edges. Examples include Internet and World Wide Web, citation networks, social networks, and biological networks. Especially, with the rapid development of the Internet, more and more people treat it as a place to express their voice, vent their feelings, and communicate with each other. Great changes of interpersonal communication have taken place, moving from the real social network to virtual network communities, which provide open platforms for expression and communication so that strangers from real world could share their ideas, form one group with strong influence, and even lead network events. Thus, complex networks have attracted considerable attention in many fields for representation of a variety of complex systems. Detecting community structure, which looks on communities as groups of nodes within which there are higher density of edges and between which the edges are sparser, has become one of the hot research topics in the field of complex networks. Its main task is to divide a whole network into subgroups with strong

similarity for analyzing the dynamic evolution regularity of a virtual community and identifying functional units such as collections of pages about a single topic on the Web networks or cycles in metabolic networks. Community structure detection has been widely used in many fields, such as crime investigation and web search.

In recent years, many kinds of algorithms are put forward to uncover the features of a virtual community [1–5]. All existing mining methods are mainly based on three types of information: topology structure of a network (such as clustering coefficient, betweenness), dynamic properties of the whole network, and graph partition using spectral analysis theory of matrix in mathematics. Great achievements have been made in the field of community structure detection of complex networks. Moreover, most complex networks studied have been binary in nature, that is, the edges between vertices are either present or not. Such networks are represented by (0, 1) or binary matrices. They ignore that there may be stronger or weaker social ties between individual nodes [6]. Most methods of analyzing nodes' connections only take into account a single factor, such as friend relation of individuals or hyperlink of web pages. This

has had effect on mining effective and realistic community structure. Two strangers may belong to the same group if they have strong interests on some topic. In fact, the similarity degree of two entities in a community depends on multifactors/attributes, such as friend relation, hobby, and social status. Similarity analysis of vertices is the basis of mining communities of a complex network. Thus, it is significant for detecting community structure to design a novel model exploring the similarities of nodes based on multifactors.

This paper builds an Internet social network modeled as a graph in which user IDs or entities are vertices and their similarities are edges. Aiming at this network, we put forward a novel supernetwork-based model of characterizing network social groups by multi factors. This model consists of three basic networks, two visible mappings and one hidden network mapping. The similarities of individuals in defined social networks are calculated by weighting the values of two indices: friend relation and interest similarity. We will solve the problem of getting intrinsic weight values, that is, connection strength values between any two user IDs in online social networks [3]. A weighted social network is constructed so as to lay the foundation for community structure detection, that is, dividing a whole network into subgroups with strong similarity.

The remainder of this paper is organized as follows. In Section 2, we review related work about community structure detection of complex networks. Section 3 presents a new supernetwork based relation model of characterizing network social groups by multi factors. Section 4 gives its concrete algorithm implementation. Experimental analysis and discussion are described in Section 5. Finally, Section 6 outlines conclusion and future work.

2. Related Work

Researchers have uncovered that many complex systems in various fields share some significant topological properties such as small-world property, scale-free property, and community structure, which have already attracted much attention. In particular, the community structure is indeed of considerable importance because it is helpful to understand both structure and function of networks. So far, most works of detecting community structure focus on two kinds of networks: computer-generated networks and real-world networks, including Zachary club network, jazz musician network, *C.elegans* metabolic network, a university e-mail network, football match network [3, 7, 8], and Web network [9, 10]. This paper treated Internet social networks as the analysis target and further built the weighted social networks. Because the online social networks share many similarities with Web networks, such as running in Internet environment and having virtual community structure, the following section shows related work of detecting community structure of Web networks and online social networks, and methods of analyzing the weighted networks.

Generally, the Web network is modeled as a graph in which Web pages are vertices and hyperlinks are edges. Gary William Flake et al. defined a Web community as

a collection of Web pages in which each member page has more hyperlinks within the community than outside the community [9]. And identification of highly related Web pages was put forward based solely on connectivity, without the inherent bias of text-based approaches. The result could be applied to improved search engines, content filtering, and objective analysis of relationships within and between communities on the web. Lefteris Moussiades et al. proposed a novel definition of refined community for the community structure of a Web site. It requires that the number of links connecting a vertex to its community is higher than the number of links connecting the vertex to any other community, but not necessarily higher than the number of links connecting the vertex outside its community. And a novel graph clustering algorithm was put forward to mine refined communities from web sites [10]. It can be seen that much attention was paid to grouping highly related web pages which are identified through hyperlinks of Web sites in Web networks.

In the field of extracting online social community structure, Alan made the first study to examine multiple online social networking sites at scale, including Orkut, YouTube, and Flickr. The publicly accessible user links on each site were crawled for obtaining a large portion of each social network's graph. The authors analyzed the linkages of more than 11 million users from virtual online communities and further illustrated that online social networks have three characteristics: power law, small world, and scale-free [11, 12]. Feng Fu et al. analyzed the structure of one blog community Sina, a kind of social networks, by calculating degree distribution, clustering coefficient, and average shortest path length. They also demonstrated that the blogging network has small-world property and the in and out degree distributions have power-law forms [13]. Makoto Uchida et al analyzed a weblog network in the Japanese Blogosphere in which weblogs are written by bloggers and reflect people's up-to-date interests. And a community extracting method based on modularity is applied on the weblog network where individual posts are vertices and hyperlinks to/from other weblog posts are edges [14].

We found that most Web and online social networks studied in the current literature were unweighted networks. Ju Xiang et al pointed that the strength of connections is neglected and only the network topology is retained in unweighted networks [3]. Two new concepts: intrinsic weight and structural weight of edges are put forward in this paper. The former represents the strength of connections among various elements of networks. The later that is extracted from the topological structure of network is also related to network topology and helpful to identify inter- or intra- community edges. But how to get the weight values of each edge in weighted network is not mentioned. Newman analyzed weighted networks, in which connections have associated weights that record their strength relative to another [6]. He illustrated that weighted networks can be analyzed using a simple mapping from a weighted network to an unweighted multigraph. Like work in [3], the method of determining intrinsic weight values of edges is not presented.

This paper aims at proposing a new way to evaluate the similarity level of online social entities and constructs weighted social networks in which user IDs are nodes and their similarities are edges. In fact, many factors, such as social relation and interest, have an effect on the similarity between two social entities. Thus, it is very important to propose a formal model that combines multi attributes to determine connection strength of nodes in online social networks. We should consider the effect of multi factors on the connection strength of vertices in order to evaluate the similarities of user entities comprehensively.

3. Supernetwork Based Relation Model within Internet Social Groups

With the limit of space, this paper chose two factors: friend relation and interest to measure the similarities of user entities within Internet social groups. This section firstly introduces the basic knowledge of supernetwork. Then the model of entities' relation of an online social group based on supernetwork is presented.

3.1. Introduction of Supernetwork. A supernetwork is a compound network that is "above and beyond" existing networks, which consist of nodes and edges [15]. It can link multire networks according to many kinds of criteria and provide us with tools to study interrelated networks [16]. It also allows for the application of efficient algorithms for computation. Supernetworks are conceptual in scope, graphical in perspective, and with the accompanying theory, predictive in nature. The supernetwork model can be represented by hypergraph. Its formalized definition is shown as follows.

Given a finite set $V = \{x_1, x_2, \dots, x_n\}$ and a set $E = \{e_1, e_2, \dots, e_m\}$ which is a set cluster of V , that is to say $E \subseteq 2^V$, $H = (V, E)$ is a hyper graph of V if $e_i \neq \emptyset$ ($i = 1, 2, \dots, m$) and $\bigcup_{i=1}^m e_i = V$. And all elements of V and E are nodes and edges of the hyper graph, respectively.

Supernetworks have become a powerful tool of analyzing complex networked systems, such as knowledge network, research network, transportation network, and social network. Therefore, it is feasible to model the complicated relationships and further evaluate similarities of entities in Internet social groups based on supernetwork theory.

3.2. Model of Entities' Relationship of an Online Social Group. This paper combines two attributes of virtual individuals: friend relation and interest similarity so as to measure the similarities of different entities. We define three kinds of basic networks: entity network, action network and interest network, and build two visible mappings: from entity to action and from action to interest, and one hidden mapping: from entity to interest. A novel model based on supernetwork for modeling the relationships among user entities put forward by this paper is shown in Figure 1.

There are strong links among these three basic networks contained in the relationship model of online social users. The fact whether two user entities are friend or not can be

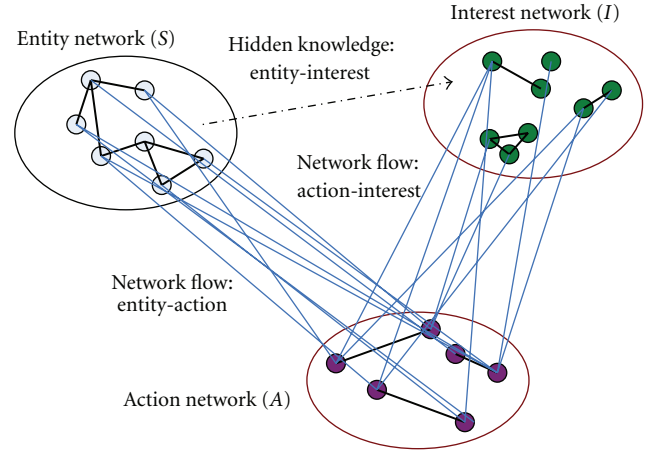


FIGURE 1: Supernetwork based relationship model among entities in social networks.

directly identified by capturing the value of friend attribute of an entity or judged through the number of communications between them. But it is very difficult to obviously determine whether two individual entities share the common interest or not. We only can observe all actions done by entities, such as publishing posts, expressing opinions, concerning homepages, to obtain their interest similarity degree indirectly. The hidden mapping, from entity to interest, can be obtained by computing above supernetwork based relation model.

Given m individual user nodes within one Internet social network, the formal description of supernetwork based relation model is shown in the following two sections.

3.2.1. Three Kinds of Basic Networks

Network 1. Entity. Represented by $G_S = (S, E_{S-S})$, where finite entities' set $S = \{s_1, s_2, \dots, s_m\}$ and edges' set between two nodes $E_{S-S} = \{(s_p, s_q) \mid p, q = 1, 2, \dots, m\}$. Each component of S may be one registered account ID or user name. The variable m is used to represent the number of user entities within a social network. The component (s_p, s_q) of E_{S-S} represents the friend relation of two entities: s_p and s_q . If its value is 1, it means these two entities are friends and there is one edge between them in the constructed hypergraph. Otherwise, no friend relation exists between s_p and s_q .

Network 2. Action. Denoted by $G_A = (A, E_{A-A})$, where finite actions' set $A = \{a_1, a_2, \dots, a_n\}$ and edges' set between two action nodes $E_{A-A} = \{(a_p, a_q) \mid p, q = 1, 2, \dots, n\}$. The variable n represents the number of all actions launched by entities. Each component of A is one action done by entities, such as expressing his opinion about some topic in a forum or participating one network application provided by different communities. Moreover, each network activity can reflect one kind of interest of corresponding ID entity. The component (a_p, a_q) of E_{A-A} represents common interest that two activities: a_p and a_q reflect. If there exists one edge between two action nodes: a_p and a_q , its value is set to 1.

Otherwise, its value is set to 0, which means that they do not have common interest according to prior knowledge of actions and their corresponding interests.

Network 3. Interest. Expressed by $G_I = (I, E_{I-I})$, where finite interests' set $I = \{i_1, i_2, \dots, i_l\}$ and edges' set between two interests $E_{I-I} = \{(i_p, i_q) \mid p, q = 1, 2, \dots, l\}$. The variable l is the number of interests corresponding to all actions. Each component of I denotes one kind of predefined interest. Each boolean component (i_p, i_q) of E_{I-I} means there maybe exists a common interest core between i_p and i_q , which can reflect main characteristics of interests. If its value is 1, there is one common interest core between two kinds of interests. On the contrary, no common interest core exists between them and its value is set to 0. Moreover, some interests with same interest core can be clustered into one hobby. This paper clusters all interests into a set of hobbies so as to lower the dimension of relation matrix between entity and interest, and further to simplify the correlation model proposed in this paper.

3.2.2. Three Kinds of Network Mappings

Mapping 1. From Entity to Action. Conveying all network activities done by entities within a period of time. The mapping from entity s_i to its activities set $A(s_i)$ is shown as $A(s_i) = \{a_j \mid a_j \in A, s_i \in S, \theta(s_i, a_j) = 1\}$, where the fact that $\theta(s_i, a_j)$ is set to 1 signifies entity s_i performed action a_j .

Mapping 2. From Action to Interest. An obvious mapping that represents all interests corresponding to any entity action. The mapping from network action a_i to its interests set $I(a_i)$ is shown as $I(a_i) = \{i_j \mid i_j \in I, a_i \in A, \theta(a_i, i_j) = 1\}$, where $\theta(a_i, i_j) = 1$ denotes action a_i performed by an entity is corresponding to interest i_j .

Mapping 3. From Entity to Interest. A hidden mapping that provides all interests owned by any network user node. The mapping from entity s_i to his interests $I(s_i)$ is represented by $I(s_i) = \{i_j \mid i_j \in I, s_i \in S, \theta(s_i, i_j) = 1\}$, where the fact that value of $\theta(s_i, i_j)$ is equal to 1 means entity s_i has bias on interest i_j . This mapping can not be obtained from raw Web dataset directly. It is gotten through transfer characteristic of two visible network mappings: Mapping 1 and Mapping 2.

4. Algorithm of Similarity Evaluation among Social Entities

The supernetwork based entities relation model proposed in this paper calculates the similarity degree of two entities by weighting the value of two indices: friend relation and interest similarity. In our model, similarity matrix X of user entities is defined by (1), where F and H are the index matrices of friend relation and interest similarity, respectively, W_f and W_h are weightings whose values are assigned by system

administrators according to the importance assigned to each index,

$$\begin{aligned} X &= W_f \times F + W_h \times H, \\ W_f, W_h &\in [0, 1] \quad W_f + W_h = 1. \end{aligned} \quad (1)$$

If there are m entities within an Internet social network, X , F , and H are three matrices with the dimension $m \times m$. The following two sections describe how to determine the values of F and H .

4.1. Approach to Quantify Friend Relation. For any element f_{pq} ($p, q = 1, 2, \dots, m$) of friend relation matrix F , its value can be two possibilities: 1 and 0, which represent two entities: s_p and s_q are friends or strangers, respectively. Friend relationship F is clarified in the following two situations.

- (a) When information about friend list of entity s_p or entity s_q is provided in the social networks, f_{pq} can be obtained directly by extracting their friend list. It is shown as follows:

$$f_{pq} = f(s_p, s_q) = \begin{cases} 1, & s_p \in G(s_q) \\ 0, & s_p \notin G(s_q), \end{cases} \quad (2)$$

where $G(s_q)$ is the set of s_q 's friends. If s_p belongs to $G(s_q)$, f_{pq} is set to 1. On the contrary, it is set to 0.

- (b) When no information about friend list of entity s_p or s_q can be obtained directly from Internet social networks, the value of friend relation is obtained by the number of interactive activities $O_{s_p-s_q}$ between entity s_p and entity s_q , which can be counted from posts published by entity s_p and replied by entity s_q or published by entity s_q and replied by entity s_p . If $O_{s_p-s_q}$ is larger than a predefined threshold T_o , these two entities are treated as friends and the value of f_{pq} is set to 1. Otherwise, it is set to 0. It is given as follows:

$$f_{pq} = f(s_p, s_q) = \begin{cases} 1, & O_{s_p-s_q} \geq T_o \\ 0, & O_{s_p-s_q} < T_o. \end{cases} \quad (3)$$

4.2. Approach to Quantify Interest Similarity Degree. It is well known that cosine similarity is widely used to measure similarity between two vectors by calculating the cosine of angle between them. It is often used to compare documents in text mining, including clustering and classifying Web texts [17]. The result of the cosine function is equal to 1 when the angle is 0, and it is less than 1 when the angle is of any other value. Based on this fact, this paper proposes one method of calculating interest similarity matrix H of virtual individuals based on cosine similarity of two vectors in hidden relation matrix T of entities and interests. Firstly, the operation of matrix iteration is done on three obvious mapping relations so as to get the hidden mapping relation T : from entity to interest according to the proposed supernetwork model

(see Figure 1). And further interest similarity matrix H among user ID nodes is acquired by calculating the cosine between two paired vectors of T . The details of the algorithm are shown as follows.

Input. Network user entities set $S = (s)_{1 \times m} = [s_1, s_1, \dots, s_m]$, network actions set $A = (a)_{1 \times n} = (a_{11}, a_{12}, \dots, a_{1n})$, interests set $I = (i)_{1 \times l} = [i_1, i_2, \dots, i_l]$, hobbies set $B = (b)_{1 \times w} = [b_1, b_2, \dots, b_w]$.

Output. Interest similarity matrix H .

Step 1. Constructing the activities set $A(s_i)$ of s_i ($i = 1, 2, \dots, m$) according to network actions done by entity s_i .

Step 2. Establishing the interest set $I(a_j)$ of a_j ($j = 1, 2, \dots, n$) according to prior relation knowledge of network actions and their corresponding interests.

Step 3. Building the hobby set $B(i_t)$ of interest i_t ($t = 1, 2, \dots, l$) based on prior knowledge of interests and their corresponding hobbies.

Step 4. Generating three relation matrices: M_1 , M_2 , and M_3 based on vectors $A(s_i)$, $I(a_j)$, and $B(i_t)$ by supernetwork mapping rules, respectively. They represent three kinds of relations between entity and action, between action and interest, and between interest and hobby, respectively. These three matrices are shown as follows:

$$\begin{aligned}
 M_1 =_{\text{entity}} & \left\{ \begin{array}{c} \text{action} \\ \left[\begin{array}{cc} y_{11}, \dots, y_{1n} \\ \vdots \quad \quad \quad \vdots \\ y_{m1}, \dots, y_{mn} \end{array} \right] \\ \end{array} \right\}, \\
 M_2 =_{\text{action}} & \left\{ \begin{array}{c} \text{interest} \\ \left[\begin{array}{cc} z_{11}, \dots, z_{1l} \\ \vdots \quad \quad \quad \vdots \\ z_{n1}, \dots, z_{nl} \end{array} \right] \\ \end{array} \right\}, \\
 M_3 =_{\text{interest}} & \left\{ \begin{array}{c} \text{hobby} \\ \left[\begin{array}{cc} v_{11}, \dots, v_{1w} \\ \vdots \quad \quad \quad \vdots \\ v_{l1}, \dots, v_{lw} \end{array} \right] \\ \end{array} \right\}.
 \end{aligned} \tag{4}$$

The components of the above three matrices are binary, that is, the value of each component is 0 or 1. If entity s_i performed action a_j , the value of component y_{ij} of matrix M_1 is set to 1. Otherwise, it is set to 0. Similarly, the component z_{ij} of M_2 represents whether action a_j is

corresponding to interest i_j . The component v_{ij} in M_3 signifies whether interest i_j belongs to the hobby b_j or not.

Step 5. Calculating relation matrix T of entities and interests through transfer characteristic of network mapping relations. It can be obtained by the following equation:

$$T = M_1 \times M_2 \times M_3 =_{\text{entity}} \left\{ \begin{array}{c} \text{interest} \\ \left[\begin{array}{cc} t_{11}, \dots, t_{1w} \\ \vdots \quad \quad \quad \vdots \\ t_{m1}, \dots, t_{mw} \end{array} \right] \\ \end{array} \right\}. \tag{5}$$

In matrix T , the larger the value of t_{ij} ($i = 1, \dots, m$, $j = 1, \dots, w$) is, the higher is the degree of interest i_j owned by the node entity s_i .

Step 6. Computing the cosine of two vectors: \vec{T}_i and \vec{T}_j of T for getting similarity H_{ij} which is one element of interest similarity matrix H . The similarity degree H_{ij} ($i, j = 1, 2, \dots, m$) of entity i and entity j is given as follows:

$$\begin{aligned}
 H_{ij} &= \cos(\vec{T}_i, \vec{T}_j) = \frac{\vec{T}_i \cdot \vec{T}_j}{\|\vec{T}_i\| \times \|\vec{T}_j\|} \\
 &= \frac{\sum_{k=1}^w t_{ik} \times t_{jk}}{\sqrt{(\sum_{k=1}^w t_{ik}^2) (\sum_{k=1}^w t_{jk}^2)}}.
 \end{aligned} \tag{6}$$

The larger the value of H_{ij} ($i, j = 1, \dots, m$) is, the higher is the interest similarity degree of two entities: s_i and s_j in online social networks.

5. Experiment and Analysis

5.1. Experiment Environment. Because of the limit of Internet access, this paper chose one real social network: RenRen as a platform for verifying the validity of proposed entity similarity relationship model [18, 19]. RenRen, which currently has about 338 million registered users, provides the largest social network service in China and adopts the real name system for user registration. Its service is basically the same as Facebook's in terms of functions and features. It can be regarded as China's version of Facebook. A registered user may easily find his friends according to the educational information in the profile and add them to the friend list. He can also update personal state, publish blogs, upload/share photos, leave a message, or comment on a photo, blog, state, and so on. Especially, there is an area of common web-pages which provides user entities with an interactive platform through logs, music, and videos. According to the linkage structure, the area of common web pages is divided into three levels: domain, topic, and common web page. It involves seven domains: Company, Sports & Events, Famous Icon, Relaxation, Education & Science, Movie & Cartoon, and Media & Organization. And each domain of the common web pages consists of several kinds of topics. For example,

the domain Sports & Events is further divided into World Cup, Asian Games, World Exposition, and so on. There is one common web page corresponding to each topic, where its creator shares photos or videos, and publishes logs. If other registered users have interest on some topic, they usually leave their comments on the photos, videos, or logs in the corresponding common web page. Thus, this page is called focused common web page by some users who express their ideas about it.

In the social network RenRen, the common web pages focused by users can well mirror their interests. In order to identify the actions, interests, and hobbies of an individual social user, this paper looks on his comments on the common webpages as their actions set. And further, the topics corresponding to the focused pages and domains which the focused web pages belong to are regarded as users' interests and hobbies, respectively.

5.2. Experiment Results and Analysis. Virtual network community RenRen provides people with a communication platform. Individual posts and comments posted by registered users can well reflect their concerns. Strangers from the real world may form one group with strong influence, and even lead to network collective events as long as they have strong interest in one topic. On the contrary, friends who don't have common concerns may not form one influential group. Thus, this paper assigned the values 0.3 and 0.7 to W_f and W_h , respectively, to simulate different levels of importance assigned to each index.

We chose 12 registered network users as analysis objects because of the space limit. In real society, the selected 12 network users who are from different cities are master students of our university. Z. Y. Zhang is the student leader in our school. Thus, almost everyone knows him. He likes to collect common homepages related to the interests of celebrity entrepreneur, celebrity model, astrology, video games, and Athletic club. F Liu, X. Y. Wang, S. X. Du, Y. Ding, and C. G. Liu work at the same laboratory: network attack and defense. Both F. Liu and S. X. Du pay close attention to the homepages about Famous Icon, Relaxation, Movie & Cartoon, Sports & Events, and Education & Science. T Qiao, X. X. Wei and Y. Wang engage in analysis of information content security and work at the lab of content security. H. J. Jin and H. Li work at another lab: security management. Coincidentally, both H. Jin and X. X. Wei like to concern some web pages with the subject of Famous Icon, Relaxation, and Sports & Events, such as World Cup, AC Milan, YuShu earthquake. Only J. Zha does research work about encryption and decryption in information security lab. Moreover, there is a tendency among young students that more and more people add their friends into the friend list of RenRen social network, express their ideas about the events they focus on and know about latest activities by accessing the posts written by their friends.

The web crawler technology was adopted to obtain the source code of each user ID's homepage. And further we got its friend list by extracting information between tag "id" and tag "vip" and its focused common web pages. Their friend relations in virtual network world are shown in Figure 2.

It can be seen from Figure 2 that entity Z. Y. Zhang is friend of all other entities except S. X. Du, which is consistent with his wide social relation in real society. There are many lines between F. Liu, X. Y. Wang, S. X. Du, Y. Ding, and C. G. Liu. Three users T. Qiao, X. X. Wei, and Y. Wang are friends with each other. H. Li and H. J. Jin are well known to each other. Figure 2 well reflects the real work relationship of 12 registered users.

These 12 social network users focused on 36 common web pages, that is, 36 kinds of interests involved in this experiment, such as TOEFL and World Cup. There are seven hobbies: Company, Sports & Events, Famous Icon, Relaxation, Education & Science, Movie & Cartoon, and Media & Organization in the area of common web pages of RenRen community. Each kind of interest can be corresponding to one hobby. According to prior knowledge of actions, interests and hobbies, we got the supernetwork relation model of a group of 12 social entities from RenRen social network. It is given in Figure 3 where triangle, square, small dot, and large dot represent entity, action, interest, and hobby, respectively.

According to the algorithm of similarity evaluation among user entities introduced in Section 4, we further got relation mappings from entities to hobbies and their degrees, which are shown in Table 1 where the first column and first row represent node entities and user hobbies, respectively.

It can be seen from Table 1 that T Qiao and X.Y Wang have the highest concerning degree on hobby Relaxation. The concerning degree of Y Wang on hobby Famous Icon is 5.5. By analyzing his focused common web pages from sampled dataset, we found that he collected 8 web pages: AC Milan, Eason Chen, World Cup, Liu Yan, Lady GaGa, ZL Zhang, redemption, and YUSHU earthquake. And half of them, including Eason Chen, Lady GaGa, ZL Zhang, and redemption, show that he has interest in musician, celebrity host, celebrity model, and celebrity actress, which belong to one hobby: Famous Icon. Table 1 also shows that 12 network users have no concern on two hobbies: Media & Organization and Company. Meanwhile, we analyzed the sampled web pages and found that 12 registered users don't concern web pages related to these two hobbies.

Finally, the values of similarity degree of 12 user entities are shown in Table 2. Figure 4 shows the weighted network of social entities' similarities based on the values of friend relation and interest similarity.

The result of entities' similarities, shown in Figure 4, effectively illustrates the effect of two factors: friend relation and interest similarity on the comprehensive similarities of entities. The obtained results are explained as follows.

- (1) We found that the value of similarity degree between F. Liu and S. X. Du is 0.99. It is the highest among all values of similarity degrees because they are friends, share five kinds of hobbies: Famous Icon, Relaxation, Movie & Cartoon, Sports & Events, and Education & Science and have strong bias towards the former three hobbies. The most correlation degrees between entity Z. Y. Zhang and other entities are above 0.7 because of his wide friendships. It can be seen from Figure 2 that entity Z. Y. Zhang is friends with all other entities

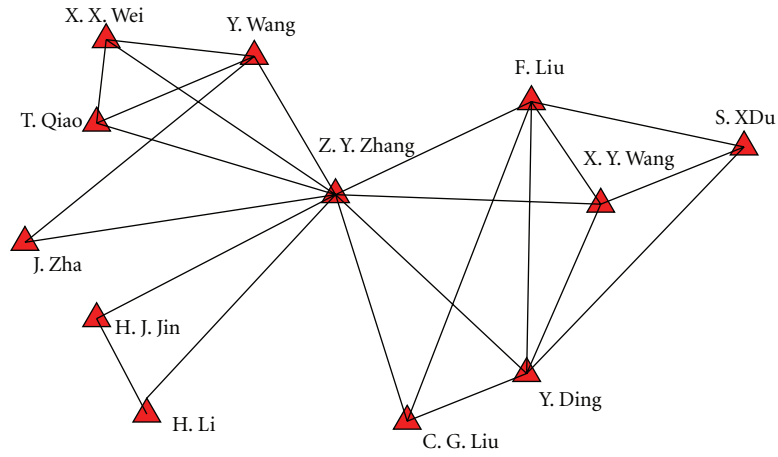


FIGURE 2: Friend relations of entities.

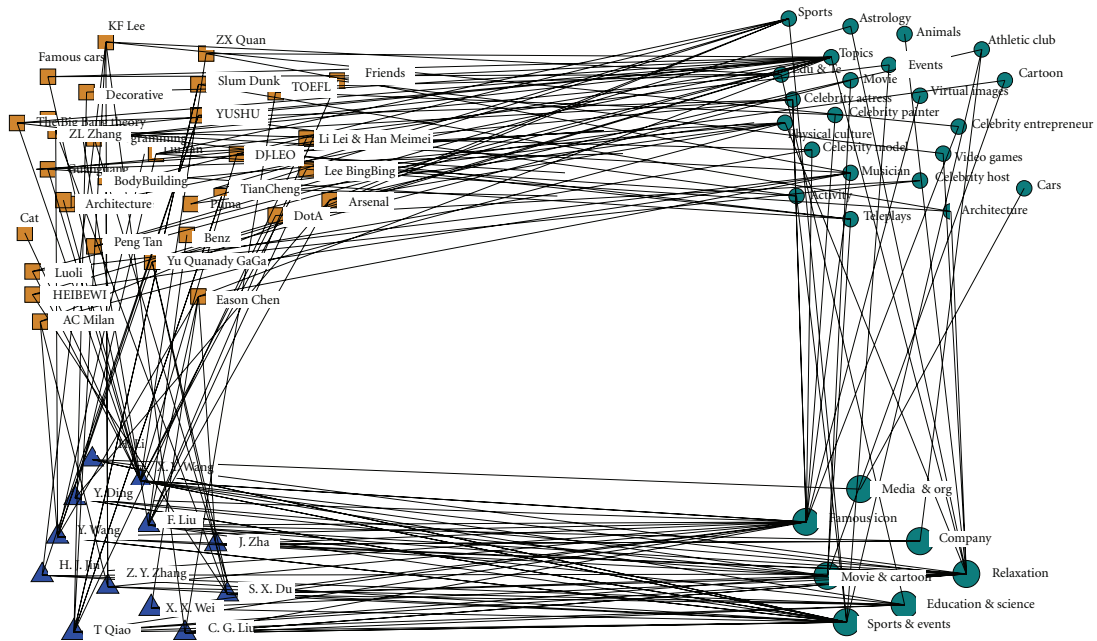


FIGURE 3: Supernetwork mined from RenRen.

TABLE 1: Relation mapping between entities and hobbies.

	Famous Icon	Relaxation	Movie & Cartoon	Media & Organization	Company	Sports & Events	Education & Science
Z. Y. Zhang	1	2	0	0	0	2	0.5
Y. Ding	0.5	2.5	1.5	0	0	4.5	0
F. Liu	2.5	3	2.5	0	0	1.5	1
Y. Wang	5.5	3.5	1	0	0	4.5	0
S. X. Du	3.5	4	2.5	0	0	1	1
C. G. Liu	2.0	1	0	0	0	1	0.5
T. Qiao	2.5	6	2.5	0	0	2	1.5
J. Zha	1	0.5	0	0	0	0	0
H. Li	1	1.5	0	0	0	1	0
X. Y. Wang	2.5	6	2.5	0	0	2	1.5
X. X. Wei	1	1.5	0	0	0	1	0
H. J. Jin	1	0.5	0	0	0	1	0.5

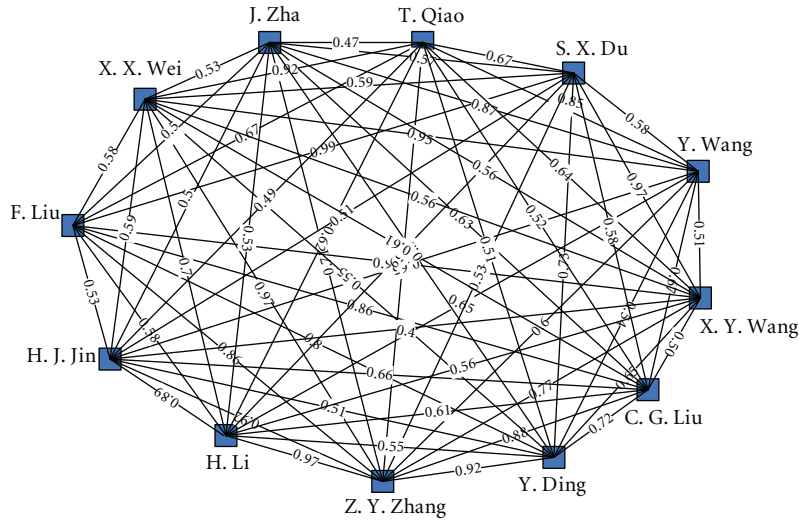


FIGURE 4: Weighted network of entities within an Internet social network.

TABLE 2: Similarity degrees of 12 user entities.

	Z. Y. Zhang	Y. Ding	F. Liu	Y. Wang	S. X. Du	C. G. Liu	T. Qiao	J. Zha	H. Li	X. Y. Wang	X. X. Wei	H. J. Jin
Z. Y. Zhang	1	0.92	0.86	0.92	0.53	0.88	0.90	0.71	0.97	0.77	0.97	0.92
Y. Ding	0.92	1	0.80	0.54	0.73	0.72	0.51	0.20	0.55	0.65	0.55	0.51
F. Liu	0.86	0.80	1	0.59	0.99	0.86	0.67	0.50	0.58	0.93	0.58	0.53
Y. Wang	0.92	0.54	0.59	1	0.58	0.67	0.85	0.87	0.65	0.51	0.95	0.65
S. X. Du	0.53	0.73	0.99	0.58	1	0.58	0.67	0.57	0.59	0.97	0.59	0.51
C. G. Liu	0.88	0.72	0.86	0.67	0.58	1	0.52	0.63	0.61	0.50	0.61	0.66
T. Qiao	0.90	0.51	0.67	0.85	0.67	0.52	1	0.47	0.62	0.64	0.92	0.49
J. Zha	0.71	0.20	0.50	0.87	0.57	0.63	0.47	1	0.53	0.56	0.53	0.50
H. Li	0.97	0.55	0.58	0.65	0.59	0.61	0.62	0.53	1	0.56	0.70	0.89
X. Y. Wang	0.77	0.65	0.93	0.51	0.97	0.50	0.64	0.56	0.56	1	0.56	0.40
X. X. Wei	0.97	0.55	0.58	0.95	0.59	0.61	0.92	0.53	0.70	0.56	1	0.59
H. J. Jin	0.92	0.51	0.53	0.65	0.51	0.66	0.49	0.50	0.89	0.40	0.59	1

except S. X. Du. On the contrary, the remaining 11 registered users have at least 6 strangers. Moreover, Z. Y. Zhang has at least two kinds of common hobbies: Famous Icon and Relaxation with any other network users. It can be concluded that network users, who are friends and share more hobbies, have larger similarity degree values and may cluster into one powerful virtual community.

- (2) By comparing Figures 2 and 4, we found that entity H. J. Jin only knows two entities: Z. Y. Zhang and H. Li. Although he does not know Y. Wang, C.G Liu and X. X. Wei, they have high similarity degrees with values of above 0.55 because they share many hobbies. The value of similarity degree between H. J. Jin and X. X. Wei is equal to 0.59. The reason is that both H. J. Jin and X. X. Wei share three kinds of hobbies: Famous Icon, Relaxation and Sports & Events. This well illustrates the fact that two strangers with much common hobbies may contribute to the formation of a virtual community.

- (3) Table 2 shows that Y. Ding and J Zha have the smallest similarity degrees 0.2. Figure 1 shows that they are not friends. We also found from Table 1 that they have weak bias towards two common hobbies: Famous Icon and Relaxation. The experiment results further prove that two strangers who share less common hobbies would not form one virtual network group and have less effect on the tendency of popular network opinions.

6. Conclusion

We have proposed an approach of quantitative similarity evaluation of Internet social users based on supernetwork theory utilizing two indices of friend relation and interest similarity. Three kinds of basic networks: entity, action, and interest are integrated for describing the complex correlation relationship of network users. An algorithm of cosine similarity of two vectors is introduced for getting the interest similarity degree between two network entities. Experimental results in real network environment illustrate

that two indices of friend relation and interest similarity are more efficient than previous approaches with only one factor for describing the correlation degree of network nodes. It successfully builds a weighted online social network, that is, a graph in which user entities are vertices and similarities are edges. Moreover, the weight values of each edge record their similarity strength relative to one another.

There is much more to be explored in the future. Our future work will focus on studying the method of detecting community structure of a weighted social network so as to understand both structure and function of online social networks.

Acknowledgments

This paper was supported by State Key Development Program of Basic Research of China (no. 2010CB731403/2010CB731406) and National Natural Science Foundation of China (no. 61071152).

References

- [1] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Article ID 026113, pp. 1–15, 2004.
- [2] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, Article ID 066111, pp. 1–6, 2004.
- [3] J. Xiang, K. Hu, and Y. Tang, "A class of improved algorithms for detecting communities in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 13, pp. 3327–3334, 2008.
- [4] Y. Pan, D. Li, J. Liu, and J. Liang, "Detecting community structure in complex networks via node similarity," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 14, pp. 2849–2857, 2010.
- [5] F. Xie, M. Ji, Y. Zhang, and D. Huang, "The detection of community structure in network via an improved spectral method," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 15-16, pp. 3268–3272, 2009.
- [6] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, Article ID 056131, pp. 1–9, 2004.
- [7] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.
- [8] Network Data, <http://www-personal.umich.edu/~mejn/net-data>.
- [9] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of web communities," *Computer*, vol. 35, no. 3, pp. 66–71, 2002.
- [10] L. Moussiades and A. Vakali, "Mining the community structure of a web site," in *Proceedings of the 4th Balkan Conference in Informatics*, pp. 239–244, Thessaloniki, Greece, September 2009.
- [11] A. Mislove, M. Marcon, P. G. Gummadi et al., "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, (IMC '07)*, pp. 29–42, San Diego, Calif, USA, October 2007.
- [12] A. E. Mislove, *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*, Ph.D. thesis, RICE University, 2009.
- [13] F. Feng, L. Liu, and K. Yang, "The structure of self-organized blogosphere," <http://arxiv.org/abs/math/0607361>.
- [14] M. Uchida, N. Shibata, and Y. Kajikawa, "Identifying the large-scale structure of the blogosphere," *Advances in Complex Systems*, vol. 12, no. 2, pp. 207–219, 2009.
- [15] A. Nagurney, "Supernetworks: an introduction to the concept and its applications with a specific focus on knowledge supernetworks," <http://supernet.som.umass.edu/articles/uksupernetworks.pdf>.
- [16] A. Nagurney, "Supernetworks: on the relationship between supply chain and transportation network equilibria: a supernetwork equivalence with computations," *Transportation Research Part E*, vol. 42, no. 4, pp. 293–316, 2006.
- [17] Cosine Similarity, http://en.wikipedia.org/wiki/Cosine_similarity.
- [18] <http://www.renren.com>.
- [19] Renren, <http://en.wikipedia.org/wiki/Renren>.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

