Neurocomputing 210 (2016) 107–115

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Identifying users across social networks based on dynamic core interests

Yuanping Nie*, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, Bin Zhou

College of Computer, National University of Defense Technology, China

ARTICLE INFO

Article history:
Received 8 May 2015
Received in revised form
25 August 2015
Accepted 19 October 2015
Available online 11 June 2016

Keywords:
Identification
Dynamic core interests
Topic analysis
Cross media analysis

ABSTRACT

With the development of social networks, most of users hold serval accounts in different social network platforms. It is a very important task to match users' varying identities in the internet. Plenty of existing approaches attempt to link users via comparing social structures, mapping users' profiles and analyzing users' authority. Those existing approaches fail to consider the dynamic changes of users. In the paper, we introduce human behavioral limitations in social networks. And then based on the limitations, we propose a dynamic core interests mapping (DCIM) algorithm, which jointly consider the users' social network structures and users' article content to identify users over platforms. The algorithm firstly models user's core interests and then calculates the similarity of two target users using DCIM. Our experiments use real world datasets from Twitter and BlogCatalog. The results of experiments show that our method is effective on mapping users across social networks. And the algorithm is significantly more effective than baseline methods such as FNN and MAG.

1. Introduction

With the development of Web2.0, social networks such as Twitter, Facebook and Sina Weibo have revolutionized our social life by allowing everyone to share their knowledge and interests with ease and fun like never before. Those social media sites have owned billions of users. The recent investigation shows that Facebook, the most popular social media site, has more than 1 billion users. And, those users have produced huge information: over 500 million tweets are posted on Twitter [1]. How to manage this huge information will be a key for better business intelligence [2]. Mapping users from different platforms is an important way to manage internet information. Identification of users across social media sites will offer the following benefits:

Recommendation: The huge redundant information produced by Internet costs people much more time to find the useful information. It will be more precise to give users recommendation by analyzing the users cross-platform identities information.

Trust mechanism: For the social networks rarely have identity authentication systems, the information provided by users could be false, conflicting, missing or deceptive. Identifying users across social networks can help researchers to confirm users' identities and improve the consistency of user information.

Web development analysis: With the rapid development of internet, the websites come out and die in a very short period time. The movement of users responds the websites' development. The linkage between users in different platforms will provide a great chance to analyze the users' movement through the websites.

However, it is difficult to automatically identify the different social web identities of the users, for there are many reasons. Firstly most social networks allow users completing their user profiles freely, which may be totally unconnected with their real identity. Secondly some users might enter varying and unrelated profiles deliberately in different platforms. The other users who are willing to choose the same username might not be allowed because the username may have been used by another user. Thirdly people use different social networks with different purposes. Heterogeneity in the network structure and profile fields between the services becomes a complicated factor in the task of linking online accounts [3]. The primary obstacle is that the connection relationships between user identities across social media sites are usually unavailable [4].

User's behavior analysis in social networks becomes popular recently. Social networks allow people to follow and communicate with users they are interested. It is said that "Old habits, die hard", which means that users will present their behavioral habits and characteristics in the internet no matter what identities they choose. Due to different personality and human limitations, there exist some behavioral limitations of users. Those unique behavior can create useful redundant information. The analysis of users' behaviors will help to find the relation of two users from different social networks.

In this paper, we focus on identifying users across social media sites based on their behavioral limitations. These works are based

* Corresponding author.
E-mail address: yuanpingnie@nudt.edu.cn (Y. Nie).

http://dx.doi.org/10.1016/j.neucom.2015.10.147
0925-2312/

on our precious work [4]. The users' behavioral limitations are:

(1) Because of the limitation of human energy, a user has limited interests. Among those interests, there are long-term and core interests. And there must be temporary and marginal interests. People's core interests are stable. It means that a persons core interests will not change in a short period.
(2) When natural people's core interests change, their virtual identity's core interests will transfer too and the transference will be synchronized.
(3) People are more willing to make friend with the one who has the same opinion or interesting than the person who holds different attitudes. At the same time, a user may more like to receive information from the side he supports.
(4) Active social network users will show parts of the same core interests and opinions in their real life during the action in social networks.

According to behavioral limitation (4), we know that there will be common virtual circle when one natural person holds two accounts in different social networks. The common virtual circle contains parts of user's core interests and those core interests will have synchronous transference based on behavioral limitations (1) and (2). The user's core interests can be described by their structure in social media sites. Based on those behavioral limitations, we propose a user's dynamic core interests matching algorithm (DCIM). Our algorithm provides an score *S im* to list the similarity between two users.

The main contributions of this paper are shown as follows:

1. We propose an assumption that users' interests in social networks can be divided into two groups: temporary interests and core interests. And we model the users' core interests by their structure and authority in social networks.
2. We jointly model users' social network structure and their content based topic analysis to connect users from different social network platforms.
3. We consider the users' core topic dynamic changes in web actions. The dynamic characteristics analyzing will improve accuracy of the identification.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 describes definition of our method. Section 4 explains the proposed approach and Section 5 presents experimental results. Section 6 concludes the paper with remarks while proposing future work.

2. Related work

Identifying or mapping users across websites becomes a popular research topic. And some previous work has been done recently, such as [5–9]. The research has been focused on three different ways shown as follows:

Profile mapping. Some work has been done by comparing users' profiles (especially usernames). According to an investigation, over 50% of users choose the same username in different platforms [10]. A unique username may correspond to different individuals. Common usernames are often owned by different natural persons [9,11]. Zafarani and Liu [7] introduced a methodology (MOBIUS) for mapping individuals across social media sites. The MOBIUS employed the minimal information (usernames) in social network to drive features which can be used by supervised learning to connect users. Bekkerman and McCallum [11] presented two framework to solve users' linkage problem. One framework was based on web link structure and the other one used agglomerative/conglomerative double clustering (A/CDC) method. Zafarani and Liu [5] firstly formalized the problem and provided a method to connect those websites. There were two assumptions in the paper, the URL of a user profile page contained the corresponding username and a natural person's profile usually contained another community's username. Iofciu et al. [6] proposed an approach aiming at specific types of social networks, tagging systems. The proposed approach was jointly considered edit distances of usernames and the similarities of two-platform users tags. Nunes et al. [12], Malhotra et al. [3] and Zhang et al. [13] all identified the linkage of over-platform users based on users profiles (e.g. username, description, profile image, location, etc.). Vosecky et al. [14] proposed a vector-based comparison algorithm. The algorithm collected each user's profiles and proposed representing profiles as a vector. The profile vector represented profile data fields (e.g. phone number, date of birth, etc.). Nunes et al. [12] reduced the user identity resolution problem into a binary classification task. The feature vectors were grouped in three classes for user similarity. The photos of users can be an important features and some researches [15] have been done on web image search reranking and had good performances. Malhotra et al. [3] proposed users' digital footprints. The authors collected username, display name, description, location, profile image, and number of connections to calculate the similarity of user profiles across social media sites. A probabilistic approach [13] was proposed by Zhang, which used a domain-specific prior knowledge to make profile linkage in different social networks. Carmagnola and Cena [16] introduced an approach that bases heuristics on profile attributes such as username, name, location or email address of a user. Goga et al. [17] constructed a similarity matrix, which contained: name similarity, photo similarity, face similarity, and location similarity. And the author did serval pairs of platforms experiments to prove that the method was effective. The work [18] aimed at video retrieval, which may allow user's sharing video analysis for matching users.

Structure analysis. There are some researches that focus on users' social structure. The user's social structure can be defined in different ways, such as structure of users' friendship (neighbors) in social network. Some researches have proved that it is available to map users by analyzing users' structure from different social networks. Narayanan and Shmatikov [19] believed that a natural person usually owned the similar social graph in virtual world. Tan et al. [10] proposed a novel subspace algorithm called manifold alignment on hypergraph (MAH) to utilize social structures to improve the mapping performance. The semi-supervised manifold alignment [20] based on traditional graphs was an intuitive choice. Liu et al. [9] studied on human behaviors with regard to the usages of online usernames. The author focused on the alias-disambiguation step as a pairwise classification problem and proposed a novel unsupervised approach to link users across social networks. Liu et al. [8] proposed a framework called HYDRA to address users linkage. The authors firstly considered the heterogeneous behavior that was decided by long-term behavior distribution analysis and multi-resolution temporal information matching. And the authors constructed structural consistency graph to model core structure of users in social media sites.

Authority identification. Authority identification is the task on identifying authors according to their writing styles by analyzing the corresponding article content. Novak et al. [21] proposed a language model based approach for authorship classification in online forums. Qian et al. [22] proposed a labeling oriented author disambiguation (LODA) approach, to disambiguate authority by combining machine learning and human judgment together. Amitay et al. [23] studied on one individual content generation behavior in several collections of documents. They proposed a method for detecting pages created by one user from different collections of documents based on the overlap between contributions. They used a method based on normalized compression distance (NCD) [24]. Some work has been done based on content [25] to identify content features across a large number of

documents. Rao and Rohatgi [26] found that function words were useful on identifying author in mailing lists. Zheng et al. [27] provided a framework on capturing the writing style features. There are lots of widely used models on detecting users' content topic such as LDA [28], LSI [29], pLSI [30] and DTM [31]. LDA is a hierarchical nonparametric Bayesian approach to discover topic in text corpora. DTM used state space models on the natural parameters of the multinomial distributions to express the topics.

Most of previous works do not consider the dynamic features. In this paper, we jointly consider the users' structures and their topic labels based on the article content. We propose a dynamic core interests mapping algorithm to identify users across social media sites.

3. Problem statement

When we want to map users across social media sites, there are three different kinds of problems need to be solved. The first problem is: Given two users u and v and how to identify whether they belong to the same individual. The second problem is: Given one user u and a group users V, which contains target matching user v. How do we find the user v that is match to the user u. The third problem is: There are two groups of users U and V, which we have no idea of matching users existence. Can we find the two users u and v that belong to the same natural person. In this paper, we focus on solving the first and the second problems. The third problems will be considered in future.

A social network user will have relation with other users and the structure has some properties, which can be summarized as: [identity/labels, following/follower, mention]. Identity and labels: Social network users can be grouped as: public figures and regular users. There are some public figures that hold accounts authenticated by social media sites. Those famous figures include actors, athletes, politicians and entrepreneurs. Their identities can be sure. The remarks from famous accounts are much more widely spread than the regular users. Obviously most of the famous users have certain labels. There is an example Kobe Bryant (e.g. his label is basketball player, Los Angeles Lakers). However, most of the users are regular users. Their identities cannot be identified easily and labels are much more difficult to be confirmed. Follower/following: Twitter and Sina Weibo both have those functions. If user A follows user B, the user A is user B's follower and the user A can read user B's public timeline. We believe that if two users follow each other, then it will be more possible that they are friends in social networks. Usually, public figures have owned much more followers than regular users. In fact, a user's following can represent the user's labels more effectively than followers. Mention: Mention is an action that users can communicate with each other. In Twitter and Sina Weibo, users can mention other users no matter whether they follow or not. We believe that more mention times means more closer relation of two users. According to the limitation 3, mentioned in Section 1, for people are more willing to make friend with the one who has the same interesting, the closer relation means two users more like share the same interests.

Definition 1. *User's graph in social networks*: Let u and v be two users, which are the pair of users we try to map from different social network platforms. For each social network platform, we build a graph X and Y for user x and y. The vertices donate the users and the edges represent the relations of users. We define the graph X and Y as $G_X(V_X, E_X)$ and $G_Y(V_Y, E_Y)$, where V_X is the set of vertices corresponding to users and E_X is the set of edges corresponding to social relations. Users follow friends with the same interests. So users' friends can represent their interests. The followed users' labels can be described by the authors article content. The vertices can be labeled by their content based topic $p_i(\theta)$. Users' topic labels can be defined as $V(p_i(\theta))$, where θ is topic distribution of the ith user. And the edges can be indicated by mention times m. The bigger m is, the closer relation two users have:

$$G_X(V_X, E_X) = G_X(V_X(p_i(\theta)), E_X(m)) \tag{1}$$

Definition 2. *User's core interests*: According to behavioral limitation 1, although different social network platform has different purposes, the active user will hold parts of core interests in each platform. It means if given two users belong to the same individual, there should be some core interests in both platforms. So the user's interests can be grouped as *Core interests* and *Marginal interests*. We define user X's core interests, which can be written as $G_{XC}(V_{XC}, E_{XC})$ and marginal interests can be $G_{XM}(V_{XM}, E_{XM})$. It can be concluded as:

$$G_X(V_X, E_X) = G_{XC}(V_{XC}, E_{XC}) + G_{XM}(V_{XM}, E_{XM}) \tag{2}$$

Definition 3. *Dynamic core interests matching*: As we discuss in Section 1, user's core interests will transfer synchronized. It means that the distribution of two mapping users' core topics are similar. We provide a score of similarity $S\ im$ to measure dynamic core interests' similarity between user X and Y. $S\ im$ is decided by function f about X and Y.

$$S\,im = f(G_{XC}, G_{YC}) \tag{3}$$

We define that for problem 1, the $S\ im$ can provide researchers the similarity of two users. And for the problem 2, we define the most similar user Y is user X's matching user.

4. User's dynamic core interests matching algorithm

4.1. User's core interests

Social networks contain a lot of information and functions. The social network users' activities have become more complicated and unpredicted. How to find user's core interests from redundant data is a task of our work. However, it is effective to find target user's core circle of friends first. The structure of two target users X and Y is shown in Fig 1. Two users have friends in each social networks. If two users belong to the same individual, there should be similar circle of friends, which may not have equal size. It is a challenge to find the similar circle, since we cannot know all those users' identities. However, there are two assumptions that will help us to find user's core interests. The first one is the more times that two users communicate (mention actions) with each other, the closer relation that two users have. The second one is that the core circle of friends will exist in different social networks. In the paper, we define the user X's core circle of friends as $G_{XC}(V_{XC}, E_{XC})$. According to the first assumption, let user X has p friends and user Y has q friends. The pair of users' communication similarity can be derived using the following equation:

$$S_m = \left| \frac{m_{xi}}{\sum_1^p m_{xi}} - \frac{m_{yj}}{\sum_1^q m_{yj}} \right| \tag{4}$$

where m_{xi} is the mention times of user X and ith friend, while m_{yj} is the mention times of user Y and jth friend.

In the second assumption, there are some famous users who make their identities public in social media sites. Those users' labels can be found easily. However, for most users are anonymous, we cannot determine their labels by their identities. In order to capture the users' intersection core circle of friends, we employ latent Dirichlet allocation (LDA) model [28], which is one of the most widely applied model for mining the topics of documents. Different from the normal article, social networks such as Twitter and Weibo have the short content length (usually the length limitation is 140 words). The situation will lead to poor performance of LDA model. To address this issue, we gather all of user

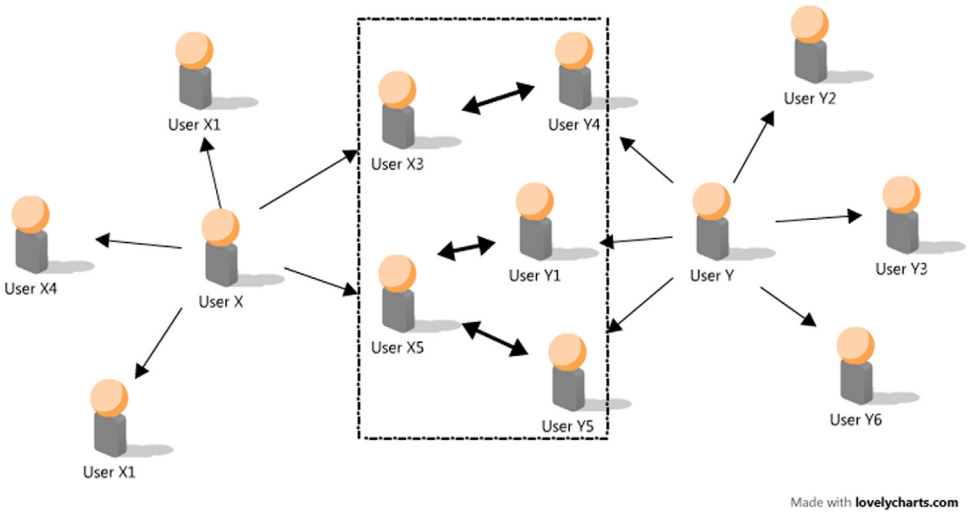

Fig. 1. The structure of matching users.

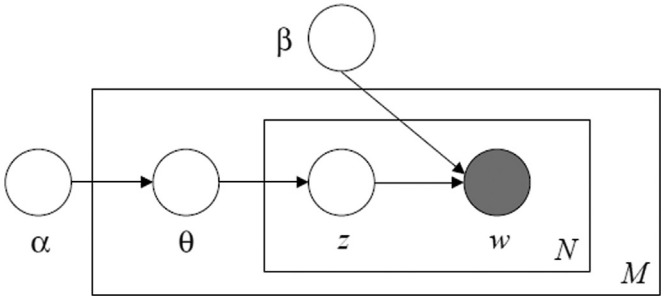

Fig. 2. The generation process of LDA model.

core friends' documents as a new corpus. Then we train LDA model on the new corpus with Gibbs sampling method. The graphic model representation of the LDA model is shown in Fig. 2. We define the set of user's core topics as:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\partial)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{5}$$

where α and β is the given the parameters. θ is the joint distribution of a topic mixture. z is a set of N topics and w is a set of N words. We use Gibbs sampling to obtain samples of the hidden variable assignment and estimate the model parameters of LDA. We can drive the Gibbs update rule as follows:

$$\begin{aligned} p(z_i = k|\vec{z}_{\neg i}, \vec{w}) &= \frac{p(\vec{w}|\vec{z})}{p(\vec{w}, \vec{z}_{\neg i})} = \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \\ &\propto \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})} \cdot \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})} \\ &= \frac{\Gamma(n_k^{(t)} + \vec{\beta}_t)\Gamma(\sum_{t=1}^{V} n_k^{(t)} + \beta_t)}{\Gamma(n_{k,\neg i}^{(t)} + \vec{\beta}_t)\Gamma(\sum_{t=1}^{V} n_k^{(t)} + \beta_t)} \cdot \\ &\quad \frac{\Gamma(n_m^{(k)} + \alpha_k)\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k)}{\Gamma(n_{m,\neg i}^{(k)} + \alpha_k)\Gamma(\sum_{k=1}^{K} n_m^{(k)} + \alpha_k)} \\ &= \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{[\sum_{k=1}^{K} n_m^{(k)} + \alpha_k] - 1} \\ &\propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t}(n_{m,\neg i}^{(k)} + \alpha_k) \end{aligned} \tag{6}$$

According to the properties of the gamma function: $\Gamma(z+1) = z\Gamma(z)$. With Gibbs sampling, we can make the following parameter estimation:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t} \tag{7}$$

$$\theta_{m,t} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \tag{8}$$

Using LDA we can know user's friends' topic distribution $p(\theta)$. Because it is hard to know whether two friends from different social media sites belong to the same individual, we can group the two friends who have the similar topic distribution in to core circle of friends. There are lots of work to measure the similarity of two distributions, such as KL (Kullback Leibler) divergence and JS (Jensen–Shannon) divergence. In this paper, we employ KL divergence to calculate the similarity of two topic distributions. KL is a non-symmetric measure of the difference between two probability distributions. For two vectors p and q, the divergence is shown as follows:

$$D(P\|Q) = \sum_i \ln\left(\frac{p(i)}{q(i)}\right)p(i) \tag{9}$$

when $D(P\|Q) = 0$ means the two distributions are completely the same. The value of D is smaller and the two probability distributions are more similar. The pair of users' topic similarity can be derived as:

$$S_t = \sum_i \ln\left(\frac{p_{xi}(\theta)}{p_{yi}(\theta)}\right)p_{xi}(\theta) \tag{10}$$

where p_{xc} is user X's ith friend topic distribution and p_{yc} is user Y's ith friend topic distribution.

The set of core friends can be measured by:

$$Cor = \alpha \cdot S_c + (1 - \alpha)S_t \tag{11}$$

The Cor keeps the same monotonicity of S_c and S_t. The value of Cor is smaller and more probably the pair of users might be the core friends.

$$f(V_{Xi}, V_{Yi}) = \begin{cases} 1 (Cor \le l) \\ 0 (Cor > l) \end{cases} \tag{12}$$

The core circle of friends can be captured by Eq. (12). l is a

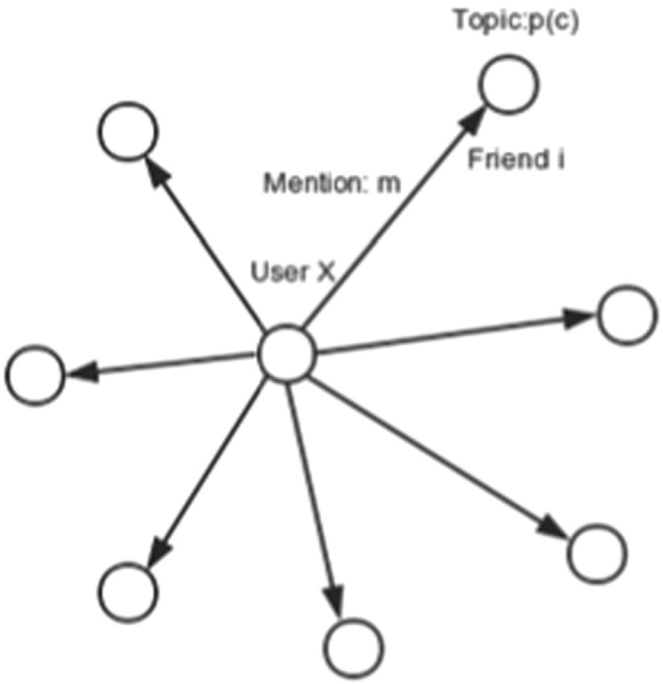

Fig. 3. User's core interests structure.

threshold. The threshold is not the key point in our work. In this paper, we choose the threshold by experience. The structure of user X's core friends' graph is shown in Fig. 3. The user X owns i core friends. The ith core friend of user X is labeled with topics, and the topic distribution is $p(\theta)$, estimated by LDA. The edge length is m, indicated as mention times m.

4.2. Dynamic core topic matching

In this section, we propose a dynamic core topic matching algorithm to identify users across social media sites. After the user's core interests is learnt, which can be described by the core circle of user's friends according to the last section, the core circle can be presented as $G_{XC}(V_{XC}(p(i)), E(m))$. Note that in real social network time t is an important parameter which should not be ignored. Usually, pair of users that belong to the same individual show core interests' similarity across social media sites in a period of time Δt. Further more, we believe that two pairs of users' core interests will show highly consistent relation dynamically. For it is obviously that user's core topics usually change with the time although the core interests are stable, the dynamic core topic will show better performance. In the paper, we use a time sliding window to estimate user's dynamic topic changes. The method choosing time sliding window Δt can be grouped in two categories [32]. The first category is slicing the time t by the natural order such as hours, days, weeks and months. Each time slice has different amount of information. The second category is dividing data into the same size. In this paper, we employ the first category to slice time. We definite the sliding window of time as Δt. In one period of time Δt, the information is exchangeable, while the information is unexchangeable in different Δt. The length of time sliding window Δt's will affect the result of core topic analysis. The Δt should be trained to find a better value. The dynamic core interests matching structure is shown in Fig. 4. It illustrates that in each Δt, the structure of user's core interests may be different, since the edge length will change. And each node may have different topic distribution $p(\theta)$ in different Δt. In the paper, we choose the Δt as 7 days, 14 days and 1 month. If the time is sliced into n pieces, the user's dynamic core interests can be expressed by matrix G_{XC} shown as follows:

$$G_{XC}(t) = \{ p_x(\theta)(t), m_x(t) \} \tag{13}$$

The $G_{XC}(t)$ can be written as discrete representation:

$$G_{XCi} = (p_{xi}(\theta), m_{xi}, \Delta t_i) \tag{14}$$

It can also be written as:

$$G_{XCi} = \begin{Bmatrix} p_{xi}(\theta) & \cdots & p_{xn}(\theta) \\ m_{xi} & \cdots & m_{xn} \\ \Delta t_i & \cdots & \Delta t_n \end{Bmatrix}, \tag{15}$$

where $p_{xi}(\theta)$ is topic distribution of user X, m_{xi} is the mention times between user X and ith friend and Δt_n is the time slice. We use the cosine similarity to measure the similarity of user's dynamic core interests:

$$Sim(G_{XCi}, G_{YCi}) = \frac{\sum_1^n (G_{XCi} \times G_{YCi})}{\sqrt{\sum_1^n G_{XCi}^2} \times \sqrt{\sum_1^n G_{YCi}^2}} \tag{16}$$

We can calculate the unlabeled users X and Y's similarity of core interests across social networks. Then we can rank users across social media sites by ranking the similarity of user's dynamic core interests. In summary, the sketch of the optimization process is described in Algorithm 1.

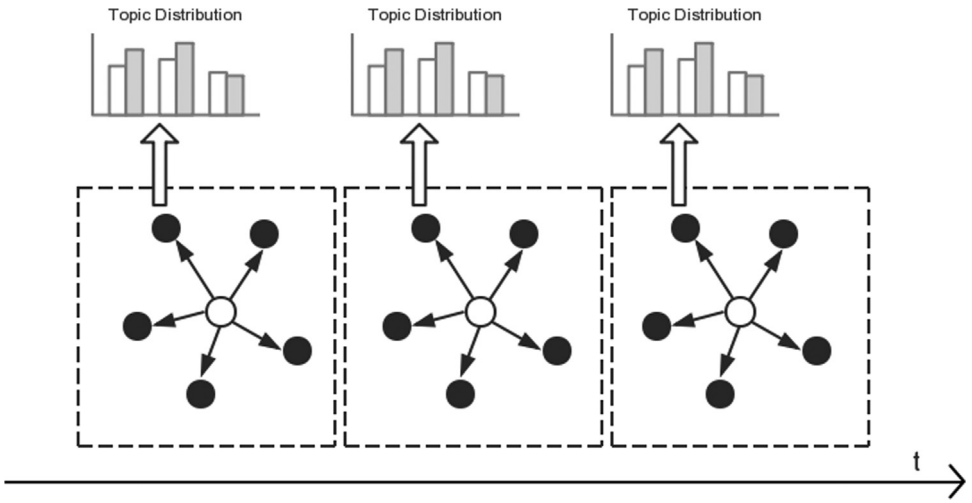

Fig. 4. Dynamic core interests matching structure.

Algorithm 1. The DCIM algorithm.

Input: **Data**: X, Y, **Parameters**: θ, l, n
Output: $S\ im$
1. Calculate the communication similarity of pair of users,
2. Using LDA estimate two users topic distribution $p(\theta)$.
3. Construct the distance S_t of two user's topic using KL divergence.
4. Select the threshold l, judging the user's core topic.
5. Slice the time into Δt_n
6. **While** the stopping criterion is not reached **do**
7. Sliding the time window Δt_n
8. **end while**
9. Obtain the $S\ im$, by Eq. (16).

5. Experiments

In this section, we investigate the use of DCIM algorithm for mapping users across social networks. We present two experiments which are both based on real world dataset. The datasets include Twitter and BlogCatalog. Then we make a discussion on the experiments.

5.1. Experiment 1

We use this experiment to test the existence of users' core interests in social networks.

Dataset: We pick 1000 users from Twitter randomly and crawl their 12 months (from February 5, 2014 to February 5, 2015) information including the users's home timeline and their friends' timeline as datasets. The first 6 months data is for language training and the rests 6 months data is for experiments.

Evaluation metrics: In this experiment, we employ KL divergence *Div* to judge the difference between user topics. It means that when the value of *Div* is smaller, the two factors are more similar.

Firstly we divide the 6 months experiment data into 6 parts and each part has 1 months data. Then we estimate the users' topics by using LDA model for each part. As comparative experiment, we calculate the core topic of users by our algorithm mentioned before. Then we calculate the difference between users' topics *Div* in every neighboring two parts. The 6 months training data is trained using Gibbs sampling method. The parameters include α, β, topic K and the threshold l. The core topic can be captured by S_m according to Eq. (4). We can obtain the 1000 users' mean value of *Div*. The result of *Div* is shown in Fig. 5.

The experiment 1 proves that users' core topics really exist. Fig. 5 illustrates that there are parts of users' topics which are much more stable than the all topics. Fig. 5 also shows that the core topics change with time, although the difference between user core topics is obviously less than the difference between all topics. For the parameters, α and β, are selected by experience. Usually, when $\alpha = 50/K$ and $\beta = 0.01$, the model will have better performance [33]. In the paper, we employ K as 25 in the experiments. The parameter of l performance is shown in Fig. 6.

Fig. 6 shows that the performance with l varies from l=0.1 to l=10. Note that decreasing l will help obtain the performance of the *Div*. The selection of threshold l will make core topics get similar. It illustrates that l decides the range of core interests and people will pay more attention on the topics they are most interested in. When l is set at 0.2, the performance of model can be satisfactory.

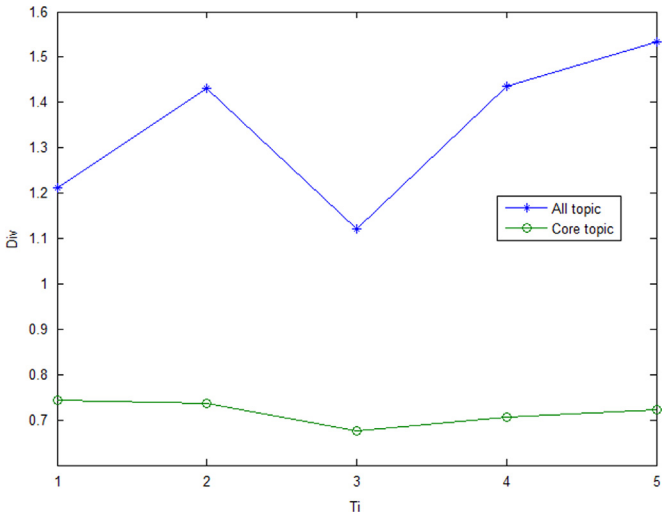

Fig. 5. *Div* on all-topic and core-topic.

5.2. Experiment 2

We use this experiment to detect our algorithm's performance in real world dataset.

Dataset: We build two datasets for experiment 2. In dataset one, we collect 1213 pairs of accounts and their 12 months data from Twitter and BlogCatalog. BlogCatalog is a social network which provides an attribute called my communities for each user. My communities allows user make known to the public their identities in other platforms, which allows us to crawl the linkage of two users from Twitter and BlogCatalog. The users in dataset one have been identified that belong to the same individual. Then we crawl 3000 active users both from two platforms (Twitter and BlogCatalog) and make pairs for those users optional. So we can get 3000 pairs of users for the dataset two. We then collect all the users' articles, all their relations (following/follower) and all comments of relations in social networks. In this experiment, the selected users are English spoken and active users, which have more than 100 relations (following + follower) and have no less than 100 articles in two social network sites.

Evaluation metrics: In the experiments, we use precision and recall to evaluate the effectiveness. Precision is defined as the fraction of the user pairs in the returned result that are correctly linked. Recall is defined as the fraction of the linked pairs which are contained in the returned result.

For the parameters, we firstly set K=25, $\alpha = 50/K$ and $\beta = 0.01$.

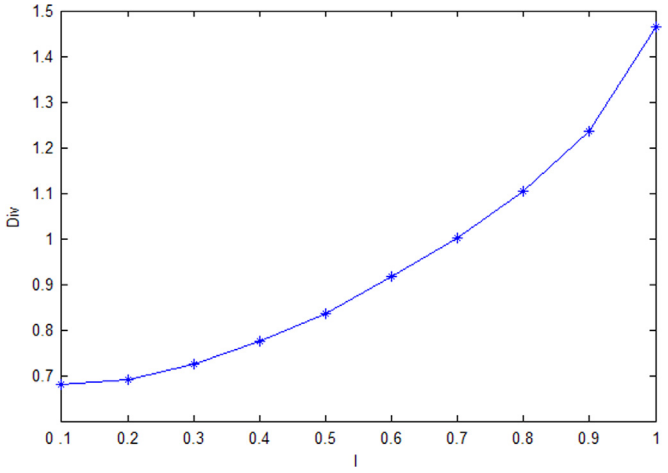

Fig. 6. Performance of l.

We use first 6 months data of dataset one and dataset two for training. And the rest 6 months data is for experiments. We employ three comparative algorithms as baselines in the experiment. The first algorithm is KNN, which is based on k nearest neighbors finding. The second baseline is manifold alignment on traditional graphs (MAG). For the third method, we employ user dynamic interests method (DIM), which does not consider user's core interests. The three baseline approaches have been tuned to a good performance. We set our algorithm's parameter l is 0.3 and set the sliding window as 7 days. Then we get the results of precision and recall.

Experimental results are shown in Figs. 7 and 8. In this test we use 6 months users' data for training. As can be seen, our algorithm has better performances than baselines in all cases of time. That is because our algorithm can make full use of core relations of users. If we change the training data as no training, 2 months data training, 4 months data training and 8 months data training, the results are shown in Table 1.

Table 1 illustrates the performance of our method which depends on the training data. And we find that it can work with acceptable performance when the training data is more than 2 months. The performances of l are reported in Figs. 9 and 10. We set l from 0.1 to 1 and algorithm parameter Δt as 7 days. The threshold l decides the core interests of users. The result shows that when l increases, the performance of precision is worse. When l is at 0.4–0.7, there is an obvious performance drop. Note that imposing larger l leads to better preference of recall, but the improvement is limited. Finally, we vary the proportion of sliding window Δt in this test. The Δt can be selected as 1 day, 3 days, 1 week, 10 days, 2 weeks and 1 month. The result of precision and recall is shown in Fig 11. The result shows that if the sliding window is too small, like 1 day, the performance is much worse. The reason is that the user's core topic cannot be clearly found during such short period of time. For example, a basketball fan may mention no comment on basketball for one day. However, we find that if the sliding window is too large like one month, we cannot obtain ideal results, since it is possible that there may be too many redundant information in such long period.

To summarize, the results show serval observations in those two experiments:

Core topic existence: Fig. 5 illustrates that if we just compare user's all topics in different time, the similarity is unsatisfied. The first reason is that social networks involve almost every corner of our life and there exist a huge redundant information in the users' comments. The second one is social networks' articles are very short (have words limitation) and colloquial (full of abbreviations and spoken-words). Our approach can divide those topic into core

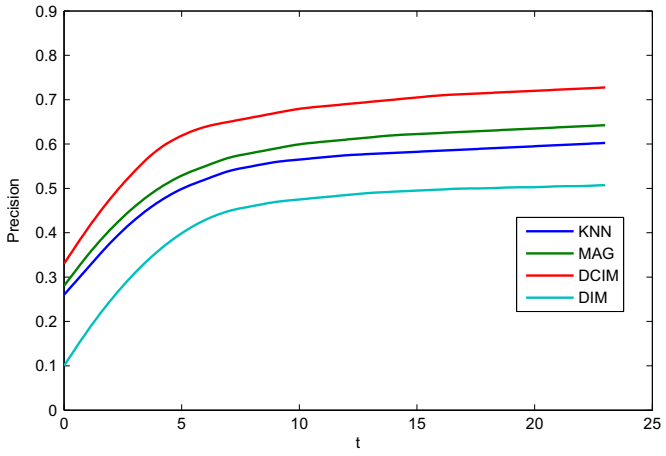

Fig. 7. Precision@t on Twitter–BlogCatalog.

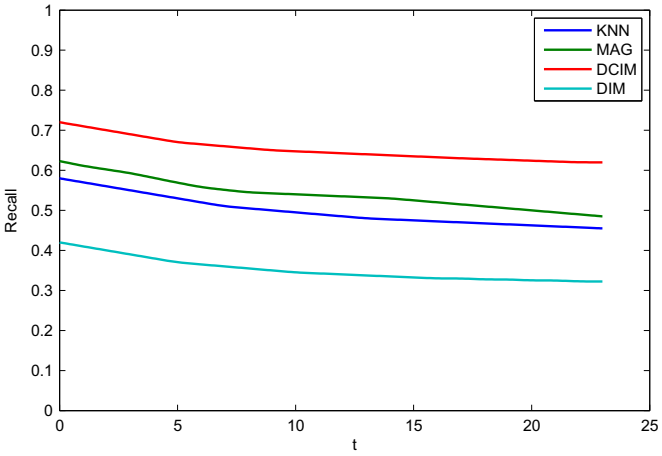

Fig. 8. Recall@t on Twitter–BlogCatalog.

Table 1
Training data performances.

Training	0 month	2 months	4 months	6 months	8 months
Precision	0.243	0.631	0.684	0.727	0.741
Recall	0.215	0.471	0.512	0.682	0.702

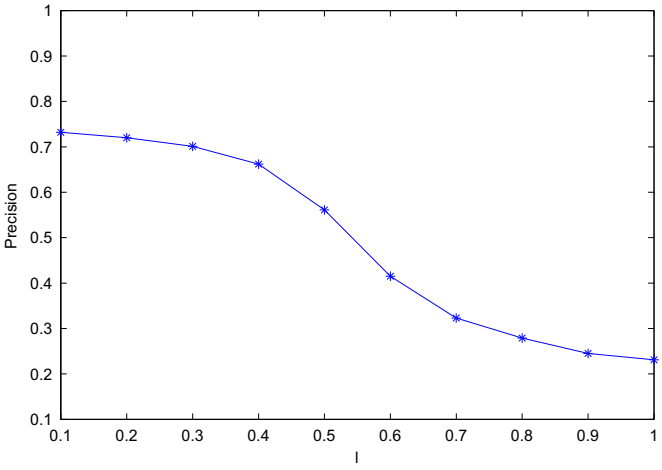

Fig. 9. Precision@l on Twitter–BlogCatalog.

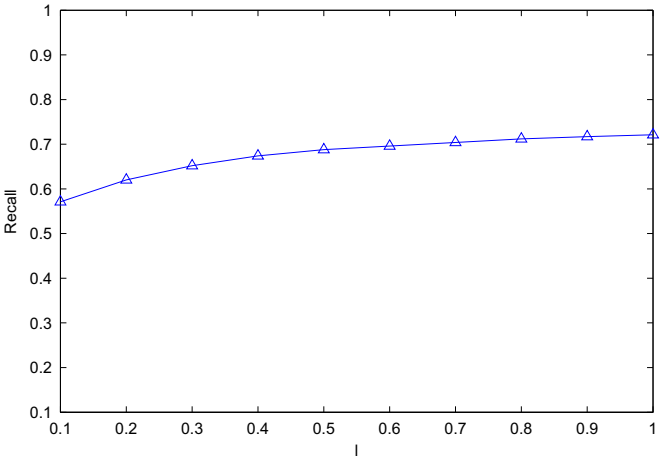

Fig. 10. Recall@l on Twitter–BlogCatalog.

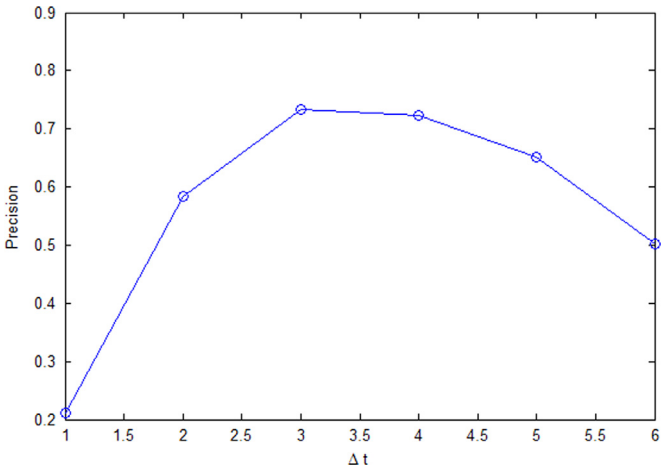

Fig. 11. Performance of Δt.

and marginal topics. The result of experiment 1 shows that the core topics will have better performance than all topics.

Training dependent: Table 1 shows that our algorithm is dependent on training data. Each user has different interests, social relations and writing styles. It is a challenge to estimate user's topic characteristics. Training data can help to improve the model accuracy for models to estimate user's topic based on their writings.

Dynamic characteristics: Fig. 5 shows that users' topic will change with time although their core topic is stable. The result in Figs. 7 and 8 proves that user's two accounts will have the similar core interests dynamic characteristics. The core interests exist in serval social network platforms. And Δt decides the performance of model shown in Fig. 11. Using the DCIM algorithm can match users across social media sites effectively.

6. Conclusions and future work

In this paper, we demonstrate a method for identifying users from different social media sites. We firstly analyze user's topics in social networks and group the users' interests as core interests and marginal interests. Then we propose an algorithm to match users by analyzing their core topic dynamic characteristics. We use the real world datasets for experiments. All the information we crawl from internet is public data on social media sites. The results show that there exist user's core interests in social networks. We then employ our method and three baseline approaches on identifying users from different social network platforms. The results show that our method is effective on mapping users from different social network platforms and our model has better performances than the two baseline methods. In our future work, we aim to deal with the challenges that we faced during the course of this research. For example: we may focus on solving the third problem mentioned in Section 3, and try to improve the accuracy of user's core topic analyses.

Acknowledgments

This work was supported in part by the National Key Fundamental Research and Development Program of China (2013CB329601), National Natural Science Foundation of China (61372191, 61202362, and 61472433), and Project funded by China Postdoctoral Science Foundation (2013M5452560 and 2015T81129).

References

[1] A. Ahmed, E.P. Xing, Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 1140–1150.
[2] S. Kumar, R. Zafarani, H. Liu, Understanding user migration patterns in social media, in: AAAI, Citeseer, 2011.
[3] A. Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru, V. Almeida, Studying user footprints in different online social networks, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 1065–1070.
[4] Y. Nie, J. Huang, A. Li, B. Zhou, Identifying users based on behavioral-modeling across social media sites, in: Web Technologies and Applications, Springer, 2014, pp. 48–55.
[5] R. Zafarani, H. Liu, Connecting corresponding identities across communities, in: ICWSM, Citeseer, 2009.
[6] T. Iofciu, P. Fankhauser, F. Abel, K. Bischoff, Identifying users across social tagging systems, in: ICWSM, 2011.
[7] R. Zafarani, H. Liu, Connecting users across social media sites: a behavioral-modeling approach, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 41–49.
[8] S. Liu, S. Wang, F. Zhu, J. Zhang, R. Krishnan, Hydra: large-scale social identity linkage via heterogeneous behavior modeling, in: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, ACM, 2014, pp. 51–62.
[9] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, H.-W. Hon, What's in a name? An unsupervised approach to link users across communities, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 495–504.
[10] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, C. Chen, Mapping users across networks by manifold alignment on hypergraph, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[11] R. Bekkerman, A. McCallum, Disambiguating web appearances of people in a social network, in: Proceedings of the 14th International Conference on World Wide Web, ACM, 2005, pp. 463–470.
[12] A. Nunes, P. Calado, B. Martins, Resolving user identities over social networks through supervised learning and rich similarity features, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM, 2012, pp. 728–729.
[13] H. Zhang, M.-Y. Kan, Y. Liu, S. Ma, Online social network profile linkage, in: Information Retrieval Technology, Springer, 2014, pp. 197–208.
[14] J. Vosecky, D. Hong, V.Y. Shen, User identification across multiple social networks, in: First International Conference on Networked Digital Technologies, 2009, NDT'09, IEEE, 2009, pp. 360–365.
[15] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Trans. Image Process. 21 (11) (2012) 4649–4661.
[16] F. Carmagnola, F. Cena, User identification for cross-system personalisation, Inf. Sci. 179 (1) (2009) 16–32.
[17] O. Goga, D. Perito, H. Lei, R. Teixeira, R. Sommer, Large-Scale Correlation of Accounts Across Social Networks, Technical Report, 2013.
[18] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009) 733–746.
[19] A. Narayanan, V. Shmatikov, Myths and fallacies of personally identifiable information, Commun. ACM 53 (6) (2010) 24–26.
[20] J. Ham, D. Lee, L. Saul, Semisupervised alignment of manifolds, in: Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence, vol. 10, 2005, pp. 120–127.
[21] J. Novak, P. Raghavan, A. Tomkins, Anti-aliasing on the web, in: Proceedings of the 13th International Conference on World Wide Web, ACM, 2004, pp. 30–39.
[22] Y. Qian, Y. Hu, J. Cui, Q. Zheng, Z. Nie, Combining machine learning and human judgment in author disambiguation, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 1241–1246.
[23] E. Amitay, S. Yogev, E. Yom-Tov, Serial sharers: detecting split identities of web authors, System 14 (2005).
[24] R. Cilibrasi, P.M. Vitányi, Clustering by compression, IEEE Trans. Inf. Theory 51 (4) (2005) 1523–1545.
[25] O. De Vel, A. Anderson, M. Corney, G. Mohay, Mining e-mail content for author identification forensics, ACM Sigmod Rec. 30 (4) (2001) 55–64.
[26] J.R. Rao, P. Rohatgi, Can pseudonymity really guarantee privacy?, in: USENIX Security Symposium, 2000.
[27] R. Zheng, J. Li, H. Chen, Z. Huang, A framework for authorship identification of online messages: writing-style features and classification techniques, J. Am. Soc. Inf. Sci. Technol. 57 (3) (2006) 378–393.
[28] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[29] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, JASIS 41 (6) (1990) 391–407.
[30] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 50–57.
[31] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 113–120.

[32] C. Kai, B. Zhou, Y. Jia, Z. Liang, LDA-based model for online topic evolution mining, Comput. Sci. 37 (11) (2010) 156–159.
[33] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (Suppl 1) (2004) 5228–5235.

Yuanping Nie, born in 1989. Ph.D. candidate in the National University of Defense Technology. His main research interests include social network analysis and artificial intelligence.

Yan Jia, born in 1960, is a Ph.D. supervisor in the National University of Defense Technology. Her main research interests include database, social network analysis and data mining.

Shudong Li received his M.S. degree in Applied Mathematics from Tongji University, China, in June 2005 and his Ph.D. degree in Information Security at Beijing University of Posts and Telecommunications, China, in July 2012. His current research interest includes complex networks and its application, wireless sensor network security, and social network analysis.

Xiang Zhu, born in 1988, is a Ph.D. candidate in the National University of Defense Technology. His main research interests include social network analysis and data mining.

Aiping Li received the B.S. and Ph.D. degrees in Computer Science and Technology at National University of Defense Technology (NUDT), China, in 2000 and 2004, respectively. He has been a professor in the Institute of Computer and Technology of National University of Defense Technology, since 2006. He visited the University of New South Wales, Australia. His research interests include uncertain databases, data mining, spatial databases and time series databases. He is a member of the IEEE.

Bin Zhou is a professor of Computer Science at the National University of Defense Technology in Changsha, China. His main research interests include web text mining, online social network (OSN) analysis, big data processing, etc. He has published more than 100 research papers (20+ was SCI indexed and 60+ EI indexed) on these topics. He also received several academic rewards, including 2 national science and technology progress awards (second class), 4 science and technology progress awards of Hunan Province (first class twice and second class twice). Recently, he has been involved in several international conference program/organization committees relating to OSN and bid data processing, such as APWeb 2014, ASONAM2014, CCF Bigdata 2014, etc.