Attribute Inference by Link Strength Modeling in Online Social Networks with User Tags

Ya Xiao[1], Zhijie Fan[2,3], Chengxiang Tan[1], Qian Xu[1], Wenye Zhu[1]

[1] College of Electronic and Information Engineering, Tongji University, China
[2] Software School, Fudan University, China
[3] The Third Research Institute of the Ministry of Public Security, China
1710053@tongji.edu.cn, aaronzfan@126.com, jerrytan@tongji.edu.cn, 1310512@tongji.edu.cn, 1310513@tongji.edu.cn

Abstract

The link strength between two users in online social networks is generally latent and can not be observed directly. The strength is usually related to the interests, behaviors, posted texts, common friends, and common followings of two users. Most previous works have ignored the distinctions of the link strengths among different pairs of users, and some works simply classify the relationships into strong and weak instead of a particular value. Given the importance of link strength in link prediction or item recommendation system, in this paper, we propose a novel method for modeling the strength of links in social networks by jointly taking the common friends, common followings, user behaviors, and user tags into consideration. A new method to construct the tags for each user based on the semantics of open information is also presented. The attribute inference and tag prediction approach based on link strength is put forward and evaluated by the experiments on a real-world dataset, the inferred results prove the feasibility of the proposed model and demonstrate that the model substantially outperforms the compared methods.

Keywords: Social networking, Link strength, Attribute inference, User behaviour, Tags

1 Introduction

Online social networks (OSNs) are increasingly essential for people to connect with their friends, share opinions and obtain information. Technologies about OSNs have also developed rapidly, researchers focused mainly on two areas for their feasibility in the use of reality: link prediction and item recommendation. The link prediction [1-3] intends to predict the probability of having a link between two users and can be used to recommend new friends or find communities. User profiling and item recommendation methods such as widely used collaborative filtering [4-6] are useful in advertisement injecting and recommendation systems. The profiling items usually include user's attributes, such as gender and location, and tags for users, like topics of interested or event tags. Item recommendation is to predict and recommend the attributes or tags for users.

Link strength is used to measure the closeness of two users, and modeling of link strength is a premise of both link prediction and item recommendation, most of the existing works on these two areas [6-7] merely consider the link weight between two users as a constant, which is inappropriate in real world. In most OSNs, the relationship between two users is observed as two different types, one is that two users follow each other, named friends here, another is one-way following, such as we usually follow a movie star, but he/she does not follow us in turn. However, the potential relationships are much more complicated than these two types, sometimes people follow each other, but they do not have a close relationship and not share the same hobbies or same friends.

Modeling the relationship strength in OSNs is important and can also influence the results of the link or item prediction. Series of works [8-10] have concentrated on the relationship strength estimation, but they only involved the profile or topics of users and interaction activities, and did not take the common friends/followings, and fields of interest into consideration. The most commonly used method for modeling users' interested fields is the topic model, however, in the micro-blog scenario, each posted text is short and lacks sufficient information to extract the topics, so the traditional method is improved in our work by utilizing the semantic meaning and extending the words in each text.

The link strength of two users generally has a close relationship with the users' attributes, areas of interest, friends and followings, so that in this work we consider all these four aspects when modeling the link strength according to the public information and social network

*Corresponding Author: Zhijie Fan; E-mail: aaronzfan@126.com
DOI: 10.3966/160792642020052103007

structure. Since users' attributes are usually not open to the public, we propose a new method to predict the unobserved attributes based on the link strength and attributes obtained.

The contributions of this paper can be formalized as follows: (1) A new directed and weighted social-behavior-tag network model is proposed. (2) This paper presents a novel method for managing interest tags for each user based on the semantics of posted texts, the lexical database is used to help extending the words to their synonyms. (3) A new link strength estimation model is proposed in accordance with the social friends, followed users, topic and interest tags on the open texts. (4) Finally, we put forward a tag prediction and attribute inference method weighted by link strength based on the idea of random walk, the real-world dataset is used to verify the effectiveness of our proposed model.

2 Related Work

The related work of this paper can be grouped into three categories, namely tag-based user profiling, link strength modeling and attribute inference.

The tag-based user profiling is to profile users with predefined tags [11], high weighted words [12] or topic tags [13]. The TF-IDF [14] or BM25 [15] term weighting method and topic model like LDA [16] can be used to extract tags. The sources which can be used to extract tags include self-descriptions, posted texts, registered profile and information from neighbors [17]. However, the predefined tags lack of flexibility and the traditional tag extraction method is not adaptable for micro-blog scenarios, so the tags extracted from sources can be extended. [13] use the structure of ontology category to extend the topic tags by adding the child nodes of the topic into the tags; they assumed that when a user is interested in a subject s, one should acquire the tastes from all child subjects of s. However, this assumption is not pervasive since the child subjects usually have a large amount and a person likes a certain subject may dislike some of the subcategories. Link strength modeling methods like [8] and [9] used the Gaussian distribution to model the conditional probability of the relationship strength given profile similarity and user interactions. [18] considered the uploaded photos, texts, common interest groups and common friends to model the link strength. However, none of these methods takes into account the latent semantic meaning of texts.

The attribute inference problem is usually based on social relationships and topological information [19]. [18] proposed a homophily principle of social influence which states that users who have links are likely to adopt similar attributes. [7] used both the social friend and behavior information and utilized the random walk with restart to predict the private attribute values based on the social-behavior-attribute network (SBA). [4] addressed the attribute prediction as a classification problem and used the semi-supervised learning, local and global consistency method. The known attributes are treated as features to predict hidden attributes. Inferring the unknown attributes may help with the recommendation system, but may also cause the problem of privacy disclosure [20-21]. As long as we study clearly about the technological process of the attribute inference, we can have more countermeasures to protect privacy. Our model considers not only the social relations, user behaviors and known attributes used in previous works, but also the posted texts and tags in predicting the unknown attributes.

3 Preliminaries

3.1 Social-behaviour-tag Network

Gong and Liu [7] have proposed a network framework called social-behavior-attribute (SBA) network considering both the behaviors and attributes of social users for the purpose of predicting latent attributes. However, with the improvement of security awareness, people would hide their attributes information in OSNs, such as location or major; thus we additionally introduce the latent tags in the network topology as an important part of link strength modeling and attribute inference. A new and more general network named social-behavior-tag (SBT) network with directed and weighted links is proposed, in this framework, the tags are derived from the posted texts of users, which are always public and open to access.

The SBT network is shown in Figure 1 and denoted as $G=(V,E,T^{v},T^{\varepsilon})$. V is the set of all nodes; each node is expressed as an id number and belongs to the following four types: social nodes, i.e., users, behavior nodes, attribute nodes and tag nodes. $n=|V|$ is the number of all nodes. E is the set of links, $m=|E|$ is the total number of links. T^{v} is a mapping set which describes the type of each node. T^{ε} is the set representing the type of each link, for example, social type $T_{ij}^{\varepsilon}=FRIENDS$ indicates that the user i and j are friends, on the other hand, $T_{ij}^{\varepsilon}=FOLLOWS$ means user i follows user j. We consider the posted texts as behavior nodes in this paper, the behavior type $T_{ij}^{\varepsilon}=POST$ is that user i is the original author of behavior node j, and $T_{ij}^{\varepsilon}=REPOST$ shows the text j is reposted by user i. The type of a link also indicates the link weight, type *FREIENDS* has a greater weight than type *FOLLOWS*, while the weight of type *REPOST* is smaller than *POST*. The types of a link can be distinct in different OSNs. Attribute links connect users and attribute values, each has a weight that represents the degree of affinity between a user and the value. Links among the tag nodes are called inter-tag link, indicating the relationship among different tags. The SBT is adaptable to different OSNs since SBT can generally cover all the items in OSNs, the corresponding relationships of entities in SBT and real word OSNs are described in Table 1.

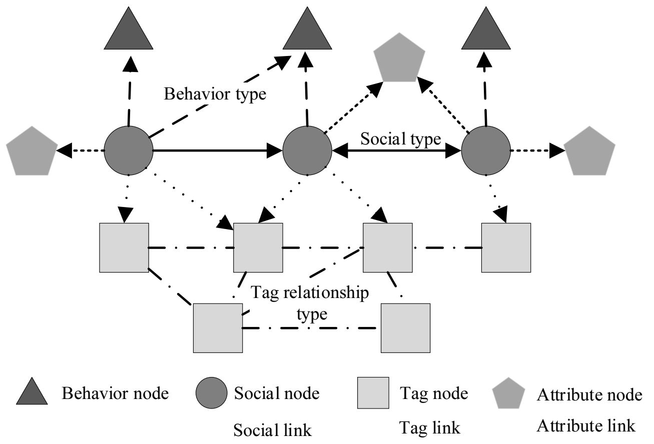


Figure 1. Social-behaviour-tag network

Table 1. Corresponding relationships of entities in SBT and real word OSNs

SBT	OSNs
Social nodes	Users
Social links	Follow, friends
Behavior nodes	Items like pictures or texts which are liked, retweeted or tweeted by users
Behavior links	Like, dislike, scan, comment, tweet, retweet, et al.
Attribute nodes	Attributes like gender, location, major, employer, et al.
Attribute links	The relationships between user and attributes, like “is a”, “live in”, et al.
Tag nodes	Tags extracted from the texts in behavior nodes like comments, posts, or answers to a question
Tag links	Links between users and tags, can be weighted by the links between user and behavior nodes which extract the tags

3.2 Problem Statement

Assuming that we are given an SBT network G with social nodes and links, behavior nodes and links, attribute nodes and links, and a set of behavior item, i.e., post or repost texts set C, the first work is to obtain the interest tags for each user by analyzing C to establish the tag nodes and tag links. A graph of tags will be built, denoted as A. The tag graph is intended to illustrate the relationships of words like synonymous or agent-object relationship, as the inter-tag links shown in Figure 1. The second work is using the $\{G, C, A\}$ to model the strength function S of each social link, the function S_{ij} measures the strength of social relationship between user u_i and u_j. Finally, predicting the unknown attributes can prove the effectiveness of the modeled link strength. The attribute inference problem is to predict the most probably appears attribute value for a specific user based on the set $\{G, C, A, S\}$ and the existing attribute values.

4 Proposed Model

4.1 Framework of Attribute Inference Model

The framework of inferring users’ attributes is shown in Figure 2. The texts, i.e., description, post or repost contents in online social networks, connected with users are used to extract tags. Tags extracted from texts can demonstrate the interests of users, thus can build the user-tag links. The user-tag and user-user relationships are used to model the similarity based social links. Based on the social links and existing user-attribute links, an attribute inference method is proposed to predict the unknown attribute links.

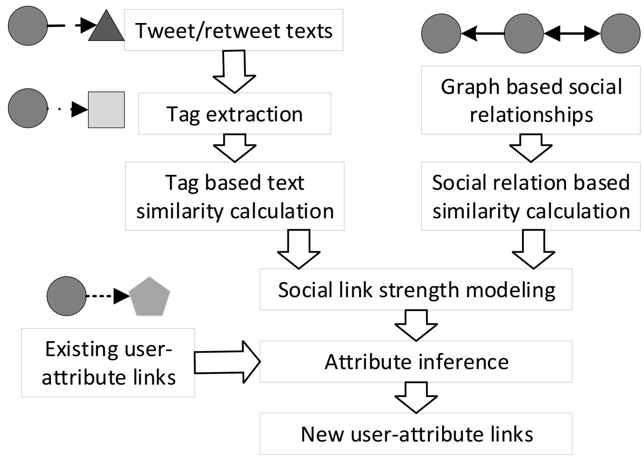


Figure 2. Framework of the proposed attribute inference model

4.2 Tag Extraction

The tags for each user consist of two parts, topic tags and interest tags. Topic tags can be extracted with the widely used topic model, such as LDA [16]. However, most online social networks have a limitation on the number of words user posts, short and abundant texts can cause a dispersive topic distribution. Since the topic tags can not precisely represent the latent information of users’ behaviors, we present additional interest tags based on the semantics of the posted short texts.

The term weighting can be performed by an improved term frequency and inverse document frequency (TF-IDF) algorithm [14]. Each text consists of various terms, i.e., words, after filtering the stop words, a text may remain only several words and most words do not appear repeatedly, resulting in a less precision with traditional TF-IDF algorithm. Therefore one efficient way to extract useful features from short texts is to enrich the information of texts. We consider the semantic meaning of words to extend the words with other similar words, in this way, the short texts will contain much more serviceable information to help with the tag extraction. Since some different words may have a similar semantic meaning, it is useful to extend the words to their synonyms or words with close relationships. The open knowledge base such as DBpedia [22] and lexical database like WordNet [23] can express the relationship of words in English. In WordNet, the nouns, verbs, adjectives and adverbs are all grouped into sets of cognitive synonyms, relationships like synonymy, hyponymy and meronymy are used to connect different words. Another lexical database in Chinese is HowNet [24], which expresses the inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. Many works have proven the effectiveness of introducing lexical resources into various natural language processing tasks [25-26]. In this work, we use these two lexical databases to extend the texts. We use the synonyms and hypernyms of each word to extend the word. To avoid the case that one word has many various synonyms and hypernyms, we only utilize the nearest two synonyms and hypernyms at most. For example, the word ‘chef’ is a kind of ‘occupation’, and the agent part is ‘cook’, so the word ‘chef’ can be extended to (‘chef’, ‘occupation’, ‘cook’).

Suppose each text in the set *C* is denoted as *c*, the original vector feature is indicated as $F_c =$ $(d_1, d_2, ..., d_n)$, where $d_i = (1, ..., n)$ is the *i*-th word in text *c* after filtering the stop words. By extending the synonyms and words have close relation with each word d_i, the new vector feature is expressed as $F'_c = \{d_1, (d_{21}, d_{22}, d_{23})d_1, (d_{31}, d_{32}), ..., d_n\}$, where $d_{ij} (i = 1, ..., n; j = 1, ..., m)$ are the *m* synonyms extended by word d_i. Term weighting for each word d_i in F'_c is performed by:

$$w_i = tf(\frac{f_i}{n}) \times idf(\frac{N}{df(t_i)}) = \frac{d_i}{n} \times \frac{1}{\log(\frac{df(t_i)}{N} + 1)}, \quad \textbf{(1)}$$

where f_i is the frequency of word *i* in text *c*, *n* is the frequency of word *i* in all the texts *C*, *N* is the number of texts in *C*, $df(t_i)$ is the number of texts which contain word *i*.

Sort the words by term weighting values and choose the top -α words as tags for each text, as well as the tags for the user who tweeted the text. The inter-tag links represent the relationship types of tags, like synonymy, hyponymy and meronymy.

4.3 Link Strength Modeling

The social link strength between a pair of users (u_i, u_j) is modeled based on social links, behavior links and tag links and is then be used to predict the unknown attributes. Considering two types of social links, i.e., friends and follows, and two kinds of behavior links post and repost, the social link strength is calculated on behalf of the similarity of post or repost texts, also the similarity on common friends and common followings.

The similarity of two texts is denoted as $sim(c_i, c_j)$ and involves both document-level and word-level similarity. Results of LDA topic model gives the probability matrix of each text being a specific topic, denoted as $p = (pt_1, pt_2, ..., pt_n)$, where t_n is the number of topics, and pt_i is the probability that this text belongs to topic t_i. The tags are generated by the tag extraction process. The topic similarity is used to measure the document-level similarity since each document is formulated into a probability matrix, and the tag similarity is used to measure the word-level similarity.

We employ a commonly used method Jensen-Shannon divergence (JSD) [27] to calculate the topic similarity between two different texts. JSD is a method based on the Kullback-Leibler divergence (KLD) [28] and is suitable to measure the similarity between wo probability distributions, as shown in Equation (2), where *p*, *q* represent the topic probability distribution vector of texts c_i and c_j respectively, and the KLD is calculated by Equation (3).

$$D_{Js}(p, q) = \frac{1}{n}\left[D_{KL}(p, \frac{p+q}{2}) + D_{KL}(q, \frac{p+q}{2})\right]. \quad \textbf{(2)}$$

$$D_{KL}(p, q) = \Sigma_i P(i)\frac{P(i)}{Q(i)}. \quad \textbf{(3)}$$

Depending on the processes in tags extraction, the distribution result of term weighting in each tweet *c* is denoted as $\{w, w_2, ..., w_n\}$, where *n* is the number of words in *c* and w_i is calculated by Equation (1). The tag similarity between two tweets c_i, c_j is calculated by combining simhash and Hamming distance [29]. Since the lengths of two tweets are different, simhash can transfer the tweets into two symbols with the same length, which is convenient to measure the similarity between two sets of tags. The simhash value for each tweet *c* is computed by the following procedures.

- After word segmentation and stop word filtering, add weight, i.e., w_i, for each remaining word.
- Hash each word into a hash value.
- Weight the hash value of each word with w_i, if the *j*-th bit of the hash value is 1, the *j*-th component is incremented by the weight; if the *j*-th bit of the hash value is 0, the *j*-th component is decremented by the

weight.

- Add each bit of the weighted value sequences in words and turn into one sequence.
- Turn the sequence into 0 and 1. If the *i*-th bit is bigger than 0, the i-th component is denoted as 1, if the *i*-th bit is less than 0, the correspondence component is denoted as 0. The sequence is the final simhash fingerprint.

After calculating the simhash value for c_i and c_j, the Hamming distance is used to count the number of positions at which the corresponding symbols are different. The proportion of common symbols is treated as the tag similarity between c_i and c_j, denoted as $S_{hamming}(c_i, c_j)$.

Based on the topic and tag similarity, the similarity of two tweets sim (c_i, c_j) is calculated with Equation (4), where (c_i^v, c_j^v) are topic probability distribution vector of c_i and c_j, α is an adjustable parameter to adjust the importance of topic and tag similarity.

$$sim(c_i, c_j) = a \times D_{JS}(c_i^v, c_j^v) + (1-a) \times S_{hamming}(c_i, c_j) \quad \textbf{(4)}$$

Link strength based texts similarity, denoted as S_{ij}^t, is computed with Equation (5), where C_{u_i} and C_{u_j} are the tweet set in which all the texts are tweeted by user u_i and u_j respectively.

$$S_{ij}^t = \frac{\Sigma_{c_i \in c_{u_i}} sim(c_i, c_j)}{|c_{u_i}||c_{u_j}|} \quad \textbf{(5)}$$

The link strength based on repost similarity S_{ij}^γ is calculated same as S_{ij}^t in Equation (5), while C_{u_i} and C_{u_j} are the repost set in which all the texts are repost by user u_i and u_j respectively. When the repost texts by u_i and u_j are directed to the same behavior node, which means they reposted the same text, then the similarity $sim(c_i, c_j)$ is set to 1. Link strength based on common friends, denoted as S_{ij}^f, and strength based on common follows S_{ij}^c are calculated as Equation (6).

$$S_{ij}^{f/c} = \beta f(u_i, u_j) + (1-\beta)\frac{|F_{u_i} \cap F_{u_j}|}{F_{u_i}}, \quad \textbf{(6)}$$

where β is set to a constant 0.5. Function $f(u_i, u_j) = 1$ when u_i and u_j are friends, $f(u_i, u_j) = 0.5$ when u_j is followed by u_i, and $f(u_i, u_j) = 0$ when u_i has no relationship with u_i. F_{u_i} is the friends set of user u_i when calculating S_{ij}^f, and when computing S_{ij}^c, F_{u_i} is the set of users followed by u_i.

The total link strength S_{ij} of (u_i, u_j) is computed as:

$$S_{ij} = w_1 S_{ij}^t + w_2 S_{ij}^\gamma + w_3 S_{ij}^f + w_4 S_{ij}^c, \quad \textbf{(7)}$$

adjustable weighted parameters w_1, w_2, w_3 and w_4 are used to adjust the importance of different link strength. Usually the strength on tweet similarity is more important than retweet, and strength on friends similarity is more important than following.

4.4 Parameter Estimation

The weighted parameters in S_{ij} can be set manually, which however highly relies on the domain knowledge and can not find the optimal combination, also has low feasibility when the strength features change in other conditions. In this section, we use a maximum likelihood estimation method to automatically estimate the weighted parameters.

Suppose the target matrix Y encodes the existing relationships among users, where $y_{ij} = 1$ denotes that user u_i and have a strong link, while $y_{ij} = 0$ represents weak link. Y can be different when facing different target problems. The strength vector and existing link of *i*-th pair of users is denoted as s_i and y_i, the conditional probability of link existence given social link strength can be modeled by

$$P(y_i = 1 | s_i) = \frac{\exp(w \cdot s_i)}{1 + \exp(w \cdot s_i)}, \quad \textbf{(8)}$$

where w is the vector of weight parameters in S_{ij}. The likelihood function is defined as

$$L(w) = \prod_{i=1}^{n} P(y_i = 1 | s_i)^{y_i} (1 - P(y_i = 1 | s_i))^{1-y_i}, \quad \textbf{(9)}$$

where n is the number of pairs of users in training dataset. By taking logarithm, Equation (9) can be derived to

$$L(w) = \sum_{i=1}^{n} [y_i (w \cdot s_i) - \log(1 + \exp(w \cdot s_i))], \quad \textbf{(10)}$$

By maximizing value $L(w)$, the weighted vector w can be estimated. The gradient ascent algorithm [30] is used to maximizing the likelihood function, as shown in Equation (11),

$$w^{t+1} = w^t + \delta \frac{\partial L(w)}{\partial L}, \quad \textbf{(11)}$$

where δ is the learning rate and can be updated during training process.

4.5 Attribute Inference and Tag Prediction Method

After modeling the tags and link strength of users, the classifiers can be used for link prediction to predict strong, weak or no link between two users. However, since the link types, i.e., friend, following or no link, are always observable in most OSNs, it is more serviceable to predict and recommend the items like attributes or user tags. In this section, we propose a new attribute prediction and tag recommendation method to prove the effectiveness of the presented link strength model.

Nodes and links in Figure 1 are used to inference unknown attributes. Given the link strength matrix S calculated in Section 4.3, with each element S_{ij} represents link strength between user u_i and u_j. The relationship between two users is not only related on the link strength, but also the strength proportion among all the neighborhood users. Thus, we further define the effect propagation matrix M which encodes the probability of effect M_{ij} from user u_i to u_j as follows:

$$M_{ij} = \begin{cases} \dfrac{s_{ij}}{\Sigma_v s_{iv}} & if(i,j) \in E, \\ 0 & elsewise \end{cases} \quad \textbf{(12)}$$

The voting value form user u_i to u_j is calculated by

$$V_{ij} = (1-\sigma)M_{ij} + \sigma e_v, \quad \textbf{(13)}$$

where e_v is a vector with the v-th entry equals to the vote value of start node and all other entries equal to 0, σ is the restart probability in random walk, i.e., with probability σ the procedure jumps back to the start node V_s and restart. The higher the vote value from u_i to u_j, the higher probability that u_i and u_j have a close social relationship, and they are more likely to have common tags and attributes. For target user u_j, the attribute value a can be inferred by combining the users who have top-k vote value with u_j. The voting value for user u_j having the attribute value a can be calculated by

$$A_{ja} = \Sigma_{u\in\tau_{a,S}} V_{uj} \cdot \frac{w_{ua}}{\Sigma_{u\in\tau_{u,A}} w_{ua}}. \quad \textbf{(14)}$$

The $\tau_{a,s}$ is a set of candidate social nodes, i.e., users, that have high vote value with user u_j and also have the attribute value a. $\tau_{u,A}$ denotes the attribute values user u has, and w_{ua} is the weight of user u having the attribute value a, i.e., attribute link weight. If the dynamic change or transfer among users in the social network is not taken into consideration, the voting value V_{u_j} can be replaced by link strength value S_{u_j}, which can also indicate the vote capacity from user u to j.

Similarly, the voting value for user u_j having the tag b can be calculated by:

$$A_{jb} = \Sigma_{u\in\tau_{b,S}} V_{uj} \cdot \frac{w_{ub}}{\Sigma_{v\in\tau_{u,B}} w_{ua}}. \quad \textbf{(15)}$$

where $\tau_{b,s}$ is a set of candidate users that have high vote value with user u_j and also have the tag value b. $\tau_{u,B}$ denotes all the tags user u has, and w_{ub} is the weight of user u having the tag b, i.e., attribute link weight. If the dynamic change or transfer among users in the social network is not taken into consideration, the voting value V_{uj} can be replaced by link strength value S_{uj}, which can also indicate the vote capacity from user u to j.

Intuitively, users who have higher link strength with the target user possess higher voting capacities, and an attribute value a or tag b should receive a higher vote if more users with high vote capacities have a or b. An attribute may have several different values, for example, attribute gender has two values male and female. For each value in an attribute, calculate the voting value, and the one whose vote is significantly higher than others can be treated as the predicted attribute value. Similarly, for each user, the tags with high vote value are regarded as the predicted tags. The attribute inference and tag prediction can be further used in item recommendation system.

5 Experiments

5.1 Data Sets

We perform the evaluation on a real-world dataset of Sina Weibo, a micro blog in Chinese, which was collected in [31]. Table 2 shows the statistics of the dataset. The social nodes are the users in Sina Weibo website, types of social links contain FRIENDS and FOLLOWS, where FRIENDS is a bi-directional link indicating two users are following to each other and FOLLOWS is a one-way link, meaning the node of start user follows the end user. Behavior nodes are the behavior items, i.e., texts posted by social nodes. Behavior types on links between user and behavior item contain TWEET and RETWEET. Behavior link type TWEET means the user is the original author of the item.

Table 2. Statistics of experimental dataset, where *s* denotes social, *b* denotes behaviour, and *t* denotes tag

#nodes			#links		
s	*b*	*t*	*s*	*b*	*t*
1777k	300k	2k	413504k	33485k	35539k

The Chinese lexical database HowNet is used in tag extraction, and 1,900 high weighting words and 100 different topics are treated as tag nodes.

5.2 Experimental Setups

The HowNet database we use contains 120,496 words with different meaning of both Chinese and English. The example of words and relationships in HowNet is as shown in Figure 3. Words in the post texts in Sina Weibo dataset are expanded according to the relationships in the HowNet.

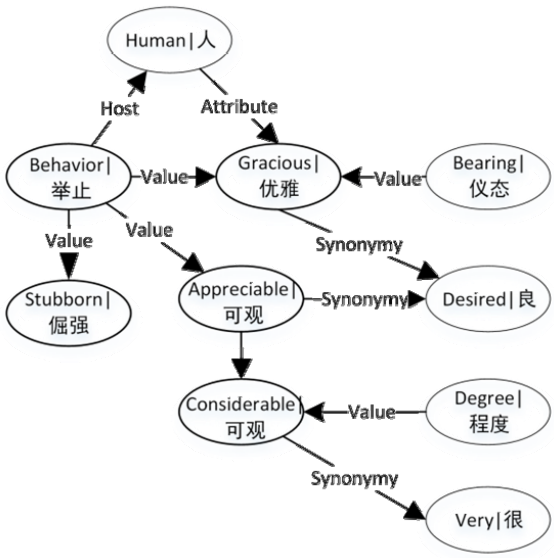


Figure 3. Example of words in HowNet

The link strength for each pair of users is calculated according to the method we proposed in Section 4.3. To better illustrate the result, we selected 1000 users with different social relationships as examples. Figure 4 shows the link strength value of the users who have links, *x*-axis and *y*-axis indicate two user ids respectively, and *z*-axis is the value of link strength. In this figure, the density of existing links and strong or weak link strengths can be clearly discovered. The higher the link strength value, the stronger the relationship is between u_i and u_j. Users of index 0 to 200 have a much denser relationship distribution than users of other ranges, which is accord with the actual distribution. The density and strength can also be applied to community discovery problems.

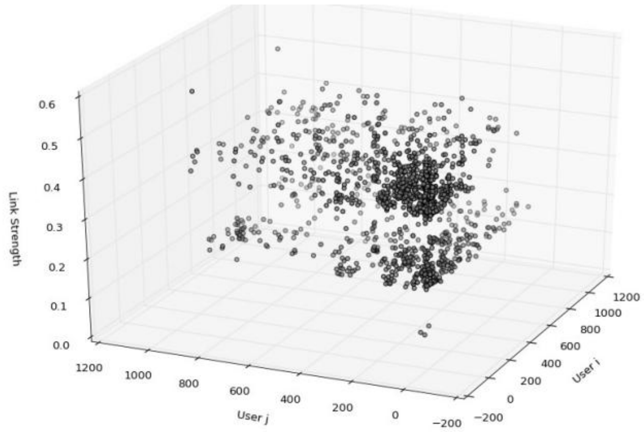


Figure 4. Link strength of each pair of users

To estimate the weighted parameters in Equation (7), we encode the pairs of users from 0 to 1530, i.e., 1530 total existing links among 1000 users. Each pair of users is treated as a training data, the four link strengths are regarded as four features. Figure 5 use three link strengths as examples for visualization, *x*-axis, *y*-axis and *z*-axis are link strength on retweets, friends and follows respectively. The green nodes indicate the pairs of users have weak relationships, while the higher strength values indicate tighter relationships. Figure 5 clearly indicates that the link strengths on follows and friends have better classification capabilities than the link on retweet, thus can better distinguish the weak links from strong ones. The weighted parameters can be estimated properly based on the classification capacities of each link strength.

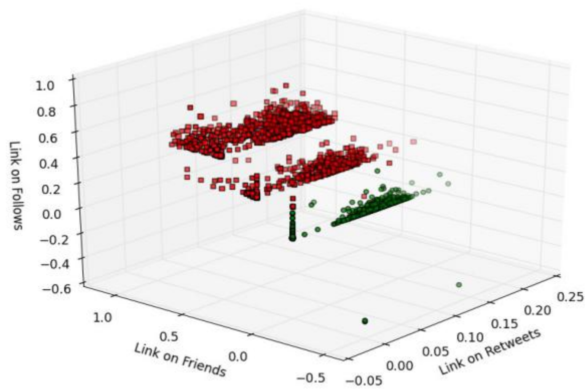


Figure 5. Link strength of each pair of users

For item prediction problem, given a target user, we predict the top-1 candidate attribute value whose attribute vote score is significantly higher than others. we perform evaluations with the attribute location, 34 different provinces in China, one overseas, and one other values are included in the dataset, which means that the attribute location in user profile has 36 different attribute values. We randomly selected three groups of users with each group contains 1000 users as target users to predict the attribute value, the involved number of social nodes is 73,819, the number of social links is 287,108, and the amount of tweet or retweet is 53,631. We use Precision, Recall and F-score to evaluate the predictions, where Precision is the fraction of predicted attributes which belong to the target user. Recall is the fraction of target user's attribute values that are among the predicted *k* attribute values. F-score is the harmonic mean of Precision and 245 Recall, calculated by

$$F - score = \frac{2 \times Precision + Recall}{Precision + Recall}. \quad \textbf{(16)}$$

5.3 Prediction

5.3.1 Parameter Settings

In the process of link strength modeling, $a = 0.4$ for texts content similarity is more important than topic similarity.The weighted parameter of link strength modeling is finally set to $w_1 = 0.2$, $w_2 = 0.1$, $w_3 = 0.4$, $4_1 = 0.3$ according to the importance of different strength and impact on the actual strength of links.

5.3.2 Tag Prediction

This section demonstrates the experimental result for tag prediction task. After calculating the link strength between different users, we select the user who have the highest link strength with the target user. The tags and texts related to the selected user are regarded as the predicted tags for target user. The similarity between predicted texts and the actual texts for different target users are show in Figure 6.

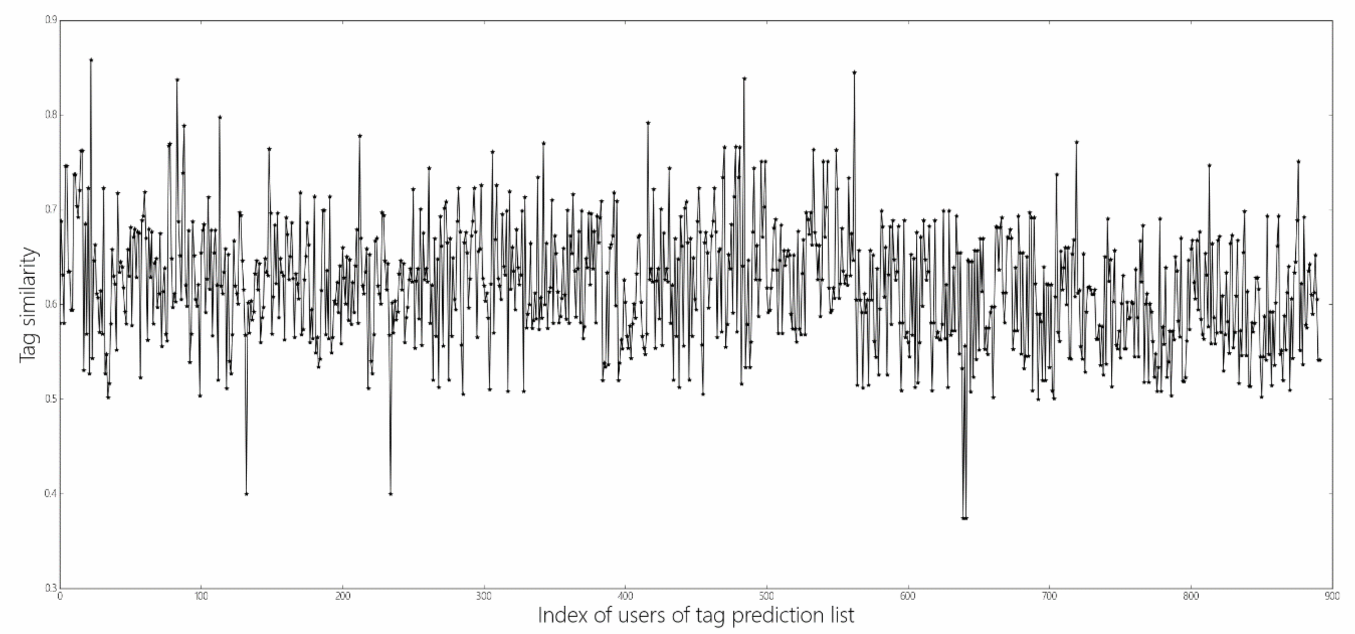


Figure 6. Predicted tag similarity for different users

It can be seen that except for several users, most similarities are higher than 0.5 and average at 0.6195. While the similarities using the traditional LDA method average at about 0.5. This means that tags predicted using our method is convincing and useful.

5.3.3 Compared Methods

We compare our method with the following methods. Each method calculates a score of every location value for the target user. The 36 scores for various location values are compared, and return the *k* attribute values that have the highest scores. The top-*k* scores are more likely to be the predicted location attribute value.

- **Random.** This method is the fraction of a target user *j* who has a particular location value *a*, the fraction is treated as the score of the location for the target user.
- **CSN.** This method calculates the number of common social neighbors of target user *j* and attribute value *a*, i.e., the number of social neighbors of *j* who have attribute *a*.
- **AILS.** Our proposed attribute inference method based on link strength, named AILS, combines social relationships, behaviors and tags together and computes the scores for all test users.
- **AILS-Behav.** A variant of our method that uses only the behavior based link strength.
- **AILS-Social.** A variant of our method that uses only the social neighbor based link strength.
- **VIAL.** This method is proposed in [7]. We compare the top-*k* results for inferring cities in VIAL with our method on predicting locations.

5.3.4 Prediction Results

Figure 7 shows the Precision, Recall and F-Score of all compared methods for the top- $(k=1,2,3)$ prediction on location values. Precision is the fraction of predicted locations which are correct. Recall is the average value of the Recalls for each location. It can be observed that AILS outperforms other methods on Precision, Recall and F-Score, including the state-of-the-art method VIAL. AILS improves VIAL on Recall and F-Score a lot in the top-*k* prediction of locations. The minor surpass of AILS than VIAL on Precision may because VIAL considers all the users who have the same attribute with target attribute, while AILS considers only the users with strong link strength to improve the efficiency and lower down the cost. Leveraging of link strength in AILS as weights of social links instead of the constant 1 in VIAL helps improve the characterization of each user and therefore remarkably improved the Recall. AILS performs better than AILS-Social and AILS-Behav, which means that the combination of both social and behavior links can better characterize a user than using only the behavior or social links. The methods CSN and AILS-Social also perform well because most attributes rely heavily on the social relationships, for example, people usually tend to make friends with those who have the same school or live in the same place with them. Although AILS-Social is less computationally costly than AILS, it also performed quite well as shown in Figure 5. This is because in many cases, users have much more friends who have one attribute value than other values, for example, someone who has 100 friends located in Shanghai, but only 20 friends located in other places. Under these circumstances, the predicted results calculated by AILS-Social or AILS will have few differences, so the outperforms of AILS is not obvious. To better illustrate the surpass of AILS, we typically selected 1000 users who do not satisfy the previously described circumstances. It turns out that AILS improved AILS-Social by 20.6% on precision, indicating that it is much more serviceable using AILS than AILS-Social on predicting the attributes.

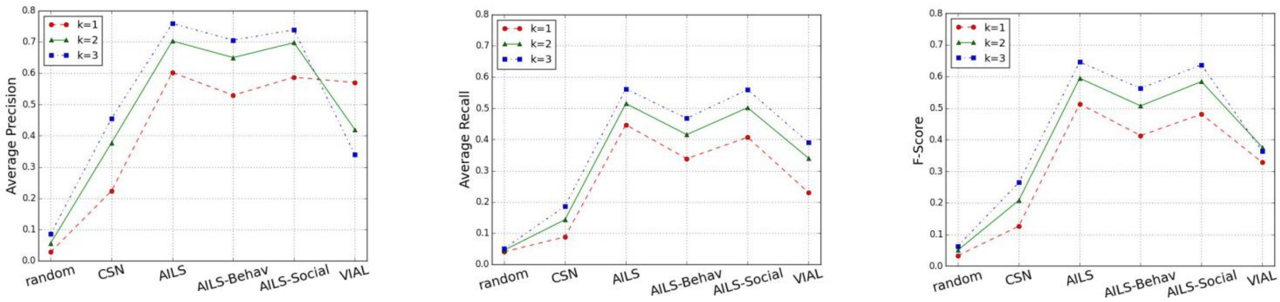


Figure 7. Average precision, recall and F-score of compared methods

When predicting the location value of a target user, after calculating the vote value for each location, the one whose vote is significantly higher than others is considered as the top-1 predicted location, as shown in Figure 8. User101, the user indexed as 101, has the highest vote value at location 27 and the value is much higher than all other locations, so the location 27 is user101's predicted location. For the users have more than one significant high votes, like user216 in Figure 8; both location 2 and 4 have a much higher vote than other locations, these two locations can be treated as top-2 prediction for user216. In a word, the vote values for 36 locations can be grouped into two clusters by a clustering algorithm, the cluster whose votes are significantly higher than the votes in another cluster contains the final predicted location value(s).

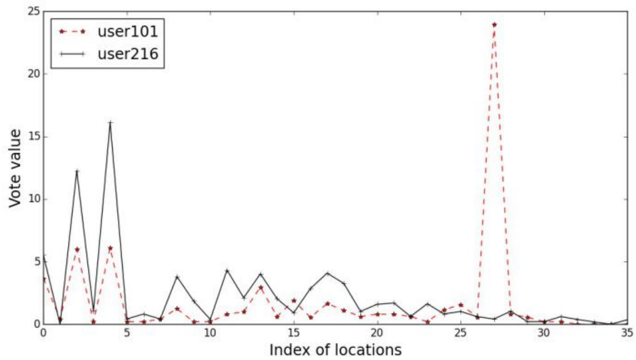


Figure 8. Vote values of each location for user No. 101 and 216

6 Conclusions

In this work, we propose an attribute inference method based on the link strength modeling in online social networks by jointly combine the behaviors, social relationships, and user tags. A novel weighted and directed social-behavior-tag network structure is proposed firstly considering the open access information in most OSNs. Next, we present a new tag extraction method based on a lexical database by extending the semantics of the posted texts, this aims to solve the problem caused by the limits of words in tweets. A method on modeling the link strength of social neighbors is also proposed, the link strength model takes into account the common friends, common followings, texts tweeted or retweeted by the users. Finally, an item prediction method which focused on attribute inference is put forward. The experiments are established to evaluate the performance of attribute inference weighted by link strength, a real-world data set is used and proves the feasibility and effectiveness of the proposed model. The model we propose is also scalable to other online social network scenarios. The item prediction can be used in recommendation system to improve the recommend accuracy, and can also be used in attack models to help attackers get more information of a particular user. As part of our future work, the dynamic change of social networks will be taken into consideration to better illustrate the OSNs in reality and the method can be further optimized to improve the accuracy and efficiency.

Acknowledgments

This work is supported in part by the National Key R & D Program of China under Grant 2017YFB0802300 and 2017YFC0803700.

References

[1] N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. Shin, E. Stefanov, E. R. Shi, D. Song, Joint Link Prediction and Attribute Inference Using a Social-attribute Network, *ACM Transactions on Intelligent Systems and Technology*, Vol. 5, No. 2, Article No. 27, April, 2014.

[2] L. Wu, Y. Ge, Q. Liu, E. H. Chen, B. Long, Z. Y. Huang. Modeling Users' Preferences and Social Links in Social Networking Services: A Joint-evolving Perspective, *Thirtieth AAAI Conference on Artificial Intelligence, Phoenix*, USA, 2016, pp. 279-286.

[3] Z. L. Li, X. Fang, O. R. L. Sheng, A Survey of Link Recommendation for Social Networks: Methods, Theoretical Foundations, and Future Research Directions, *ACM Transactions on Management Information Systems (TMIS)*, Vol. 9, No. 1, Article No. 1, February, 2018

[4] Y. Ding, S. Yan, Y. B. Zhang, W. Dai, L. Dong, Predicting the Attributes of Social Network Users Using a Graph-based Machine Learning Method, *Computer Communications*, Vol. 73, pp. 3-11, January, 2016.

[5] L. Gao, J. Wu, C. Zhou, Y, Hu, Collaborative Dynamic Sparse Topic Regression with User Profile Evolution for Item Recommendation, *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, USA, 2017, pp. 1316-1322.

[6] J. Jia, B. Wang, L. Zhang, N. Z. Gong, AttriInfer: Inferring User Attributes in Online Social Networks Using Markov Random Fields, *Proceedings of the 26th International Conference on World Wide Web, Perth, Australia*, 2017, pp. 1561-1569.

[7] N. Z. Gong, B. Liu, You Are Who You Know and How You Behave: Attribute Inference Attacks via Users' Social Friends and Behaviors, *USENIX Security Symposium, Austin*, USA, 2016, pp. 979-995.

[8] R. Xiang, J. Neville, M. Rogati, Modeling Relationship Strength in Online Social Networks, *Proceedings of the 19th international conference on World wide web*, North Carolina, USA, 2010, pp. 981-990.

[9] X. J. Zhao, J. Yuan, G. D. Li, X. M. Chen, Z. Li, Relationship Strength Estimation for Online Social Networks with the Study on Facebook, *Neurocomputing*, Vol. 95, pp. 89-97 October, 2012.

[10] L. Weng, M. Karsai, N. Perra, F. Menczer, A. Flammini, Attention on Weak Ties in Social and Communication Networks, *Complex Spreading Phenomena in Social Systems*, Springer, 2018.

[11] C. C. Hung, Y. C. Huang, J. Y. Hsu, D. K. Wu, Tag-based User Profiling for Social Media Recommendation, *Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI*, Chicago, USA, 2008, pp. 49-55.

[12] C. C. Tu, Z. Y. Liu, M. S. Sun, Tag Correspondence Model for User Tag Suggestion, *Journal of Computer Science and Technology*, Vol. 30, No. 5, pp. 1063-1072, September, 2015.

[13] J. X. Zheng, B. F. Zhang, X. D. Yue, G. B. Zou, J. H. Ma, K. Y. Jiang, Neighborhood-user Profiling Based on Perception Relationship in the Micro-blog Scenario, Web Semantics: Science, *Services and Agents on the World Wide Web*, Vol. 34, pp.13-26, October, 2015.

[14] J. E. Ramos, Using tf-idf to Determine Word Relevance in Document Queries, *Proceedings of the First Instructional Conference on Machine Learning*, Washington, USA, 2003, pp. 133-142.

[15] I. Cantador, A. Bellogín, D. Vallet, Content-based Recommendation in Social Tagging Systems, *Proceedings of the fourth ACM conference on Recommender Systems*, Barcelona, Spain, 2010, pp. 237-240.

[16] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving lda Topic Models for Microblogs via Tweet Pooling and Automatic Labeling, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin*, Ireland, 2013, pp. 889-892.

[17] H. F. Zhang, F Xiong, Y Liu, H. C. Chao, Modeling and Analysis of Information Dissemination in Online Social Networks, *Journal of Internet Technology*, Vol. 16, No. 1, pp. 1-10, January, 2015.

[18] J. Zhuang, T. Mei, S. C. H. Hoi, X. S. Hua, S. Li, Modeling Social Strength in Social Media Community via Kernel-based Learning, *Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, USA, 2011, pp. 113-122.

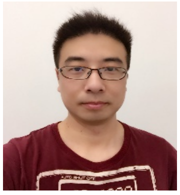
[19] L. Sheng, X. Y. Guang, X. Y. Ma, The Spread of Rumors and Positive Energy in Social Network, *Journal of Internet Technology*, Vol. 19, No. 5, pp. 1515-1524, September, 2018.

[20] C. M. Chen, K. H. Wang, K. H. Yeh, B. Xiang, T. Y. Wu, Attacks and Solutions on a Three-Party Password-based Authenticated Key Exchange Protocol for Wireless Communications, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 10, No. 8, pp. 3133-3142,

August, 2019
[21] C. M. Chen, B. Xiang, Y. N. Liu, K. H. Wang, A Secure Authentication Protocol for Internet of Vehicles, *IEEE Access*, Vol. 7, No. 1, pp. 12047-12057, January, 2019.
[22] J. Lehmann, R. Isele, M Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. V. Kleef, S. Auer, C. Bizer, *DBpedia– A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*, Semantic Web, Vol. 6, No. 2, pp. 167-195, 2015.
[23] C. Fellbaum, *The Encyclopedia of Applied Linguistics*, John Wiley and Sons, 2012.
[24] Z. Dong, Q. Dong, C. Hao, Hownet and Its Computation of Meaning, *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, Association for Computational Linguistics*, Beijing, China, 2010, pp. 53-56.
[25] Y. L. Niu, R. B. Xie, Z. Y. Liu, M. S. Sun, Improved word Representation Learning with Sememes, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver*, Canada, 2017, pp. 2049-2058.
[26] X. K. Zeng, C. Yang, C. C. Tu, Z. Y. Liu, M. S. Sun, Chinese LIWC Lexicon Expansion via Hierarchical Classification of Word Embeddings with Sememe Attention, *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, USA, 2018, pp. 5650-5657.
[27] D. M. Endres, J. E. Schindelin, A New Metric for Probability Distributions, *IEEE Transactions on Information theory*, Vol. 49, No. 7, pp. 1858-1860, July, 2003.
[28] T. V. Erven, P. Harremos, Rényi Divergence and Kullback-Leibler Divergence, *IEEE Transactions on Information Theory*, Vol. 60, No. 7, pp. 3797-3820, July, 2014.
[29] G. S. Manku, A. Jain, A. D. Sarma, Detecting Near-duplicates for Web Crawling, *Proceedings of the 16th International Conference on World Wide Web*, Alberta, Canada, 2007, pp. 141-150.
[30] P. C. Pendharkar, A comparison of Gradient Ascent, Gradient Descent and Genetic-algorithm-based Artificial Neural Networks for the Binary Classification Problem, *Expert Systems*, Vol. 24, No. 2, pp. 65-86, April, 2007.
[31] J. Zhang, J.Tang, Y. Y. Zhong, Y. C. Mo, J. Z. Li, G. J. Song, W. Hall, J. M. Sun, StructInf: Mining Structural Influence from Social Streams, *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, USA, 2017, pp. 73-80.

Biographies

Ya Xiao received the B.S. degree in School of Computer Science and Engineering, Tongji University, Shanghai, China, in 2015. She is currently pursuing her Ph.D. degree in Tongji University of Computer Science and Engineering. Her research interests include social networking and natural language processing.



Zhijie Fan received his Ph.D. degree at Department of Computer Science and Technology, Tongji University, Shanghai, China, 2019. He is the researcher of computer science at The Third Research Institute of Ministry of Public Security. He is currently work as a postdoctor in Fudan University. His research interests include cyber security, privacy preservation, data analyzing and IoT security.

Chengxiang Tan received the Ph.D. degree in engineering from Northwestern Polytechnic University, China, in 1994. He is currently a Professor of computer science with Tongji University. His research interests include cyber security, privacy preservation and data analyzing.

Qian Xu received his M.S. degree from the Department of Computer Science and Technology, Tongji University, China, in 2012. He is now pursuing the Ph.D. degree at the Department of Computer Science and Technology, Tongji University. His research interests include cryptography and cloud security.

Wenye Zhu received his B.S. degree from the Department of Computer Science and Technology, Tongji University, China, in 2013. He is now pursuing the Ph.D. degree at Department of Computer Science and Technology, Tongji University. His research interest is cloud security.