Multimed Tools Appl (2017) 76:17577–17594
DOI 10.1007/s11042-017-4408-4

A novel relationship strength model for online social networks

Chunhua Ju[1,2,3] **· Wanqiong Tao**[3]

Received: 20 May 2016 / Revised: 20 December 2016 / Accepted: 18 January 2017 /
Published online: 2 February 2017

Abstract One of the key foundations of personalized recommendation in a social network is the relationship strength between social network users. The improvement for recommendation accuracy is mostly tied to the precise evaluation of the relationship strengths. With most of the selected factors affecting the relationship strength between users are too simple, the existed researches show low accuracy in calculating the strength, especially those factors related to topic and indirect links. We propose an online social networks users relationship strength estimation model which incorporates topic classification and indirect relationship. We adopt K-means clustering method using ABC algorithm to cluster all the interactive activity documents and calculate the correlation between clusters and activity topic name. After that, we compute the relationship strength between users which belong to the same topic on top of the user profile and interaction data. To accomplish this we employ a language model based on sentiment classification approach and take similarity, timeliness, and interactivity into account. We conduct experiments on two microblog datasets and the results show that the proposed model is promising and can be used to improve the performances of various applications.

Keywords Online social networks · Topic classification · Indirect relationship · K-means clustering method · ABC algorithm · Sentiment classification

✉ Wanqiong Tao
wwwwq721com@126.com

[1] Contemporary Business and Trade Research Center, Zhejiang Gongshang University, Hangzhou 310018, China

[2] Contemporary Business and Collaborative Innovation Research Center, Zhejiang Gongshang University, Hangzhou 310018, China

[3] College of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

1 Introduction

With the rapid development of Internet technology and wireless communication technology, Online Social Networks (OSNs), such as Facebook, Microblog, Myspace and Twitter are playing an increasingly important role in our everyday life [21]. We can use social networks to connect with our family, friends or colleagues, to share the content such as photos with others, to gossip, and to obtain real- time up-to-date information of the news and events that are most important to us [20]. So far, the number of online social networks users has already reached 1400 million in 2012. In China, the most popular OSN is Sina microblog. It was initiated by SINA Corporation on 14 August 2009, and has 503 million registered users up to December 2012. About 100 million messages are posted each day on Sina microblog. Online social networks make the transmission and sharing of information resources more convenient [18]. And spreading to social networking information service gradually, which provide users with more abundant services than the traditional. But, persons should spend a lot of time and energy on searching information they needed from databases due to the explosion of information. Obviously, this condition decreases the efficiency largely. In real life, acquaintances' and friends' recommendations are very significant means that contribute to the user consumption behavior [11]. It is not the same as the degree of closeness of relationship between different users, the closer to recommend, the greater the chance of success. Granovetter once called the degree of intimacy as relationship strength in his iconic paper *The Strength of Weak Ties* [6]. So associating individual social service with user relationship strength in the social network, we can achieve accurately providing users with some social service information and content which they really interested [19].Therefore, in recent years, relationship strength between online social network users has become a hot topic of research. The accurate calculation of online social networks user relationship strength is one of the important premises of accurately realizing the personalized service, such as friends recommending.

Comparing with general friends, people are more inclined to contact with their relatives and close friends, the relationship between them is called strong ties which are belonging to the direct relation. On the contrary, the relationship of two users is weak ties when they are casual acquaintances. However, the indirect relationship may also have played a great role. Even when there is no direct link between two users, indirect relationship is the only sign of their relationship strength [12]. What is more, the relationship strengths between different users is different, and strength will be affected by many factors and changes quickly.

However, in the past, a lot of researches calculate relationship strength based on the user”s personal information and interaction information between diverse users [5, 15, 20]. User's personal information includes both the static features (e.g., age, gender, education) and the dynamical features (e.g., interest, location), which is an effective way to help providing the personalized information services [22]. As a general rule, users may have more similar interests and hobbies with similar personal information, and their relationship strength will be stronger. Interaction information includes commenting on friends' posts, pointing praise for friends' posts, sending messages to friends and so on. The more frequent the interaction the more closely relationship the two users are. Currently, more approaches are put forward to calculate the user relationship strength in online social networks. Nonetheless, these researches are segmentary. These methods are just a general way to calculate the relationship between the users. It only considers direct relation and ignores the importance of indirect relation. Which lead to inaccuracy of relationship strength estimation as a result. Also, relationship strength between the same two users may also be different in different topics. The relationship between the same two users in different topics may not have the same strength. If two users are colleagues, they will have stronger relationship strength in the job topic. If two users are tour pal, they will have stronger relationship strength in travelling topic.

Therefore, in view of the shortage of the present study, we propose a user relationship strength estimation model in online social networks based on fusion of topic classification and indirect relationship. The main contributions of this paper are summarized as follows:

(1) In the process of the calculation of the relationship strength, the similarity, timeless and interactivity are fused in, and all kinds of influencing factors are considered more comprehensively.

(2) We assign each interactive activity document to an activity topic. And not only the influence of the direct relationship on the relationship between friends is considered, but also the influence of the indirect relationship on the relationship between friends (see formula (13)) is considered in each activity topic. Both the methods can improve the accuracy of the calculation of users' relationship strength, which is demonstrated in Fig. 4 and Fig. 5.

The remainder of this paper is structured as follows. We review the related works in Section 2. In Section 3, we briefly introduce the framework of measuring the relationship strength between different users on each topic in online social networks. The details of the proposed approach are elaborated in Section 4. In Section 5, we show the initial experimental results of our approach on the Sina microblog and Tencent microblog dataset. Finally, we conclude the paper and discuss the directions of the future works in Section 6.

2 Related works

The rapid development of online social networks has led to the prosperity of researches focusing on modeling the social network. Kim M et al. [8] established an intelligent movie recommended system with a social trust model, and which was based on a social network for analyzing social relationships between users and generated group affinity values with user profiles. Xiang Lin et al. [10] set a method for relationship intensity in weighted social network graphs, which were based on the trust propagation strategy and the direct relationship intensity. Deng [4] introduced two independent ways: diversity and weighted frequency, and based on text information of subject to infer social strength between users. Xiang et al. [17] developed an unsupervised model to estimate relationship strength from interactive activity (e.g., communication, tagging) and user similarity. Zhao et al. [19] built an algorithm of relationship strength in social network integrating personality traits and interactions computing. Zhao et al. [20] proposed a general framework to measure the relationship strengths between different users, taking consideration not only the user's profile information but also the interactive activities and the activity fields. Pham H et al. [13] created an entropy-based model (EBM) that not only inferred social connections but also estimated the strength of social connections by analyzing people's concurrences in space and time. Li Peng et al. [9] built an improved LDA topics model base on microblogs chain structure, which distributing weight according to microblogs published time and social activities information including publish, comments and retweet activities, and took background knowledge to enrich semantic features of this structure.

While these researchers neglect the assignment of activity topics. The same users will have different relationship strengths in different activity topics, because the users' relationship strength may vary a lot because of plenty of factors. For example, relationship strength affected by the time factor, if users communicate more frequently means the shorter the time interval of their communication, and they will be more intimate. Users who belong to a same activity topic may share the same interests, and they will have more communication, so their relationship strength will be stronger. Meanwhile, combining with the assignment of activity topic when the research of user relationship strength applies to personal recommendation will be more targeted. What's more, most of researches at present ignore the indirect relationship strength. In some case, it is insufficient. For example, user A and user B aren't direct friends, but they have a common friend C. Although A can't influence B directly, A can influence B indirectly through C. If without the calculation of indirect relationship strength, we are unable to measure the effect of A to B.

Our work makes up the shortage of researches at present. We concentrate on modeling the relationship strength rather than the link existence. Meanwhile, we also aim to assign activity topics for interactive activities, and calculating the relationship strengths between different users on the same activity topic based on the similarity, timeliness and interactivity, which utilizing a language modelling method. Using this approach to calculate the users' relationship strength can be applied in personalized recommendation service more accurately. Locating the corresponding interest topic for target user quickly and improve the targeted recommendations, achieve precise recommendation finally. We estimate relationship strengths between different users in the same activity topic. For instance, when make discussion about the topic of travel, A is the publisher and B is the reviewers. We assume that B always agrees with A when A recommended the relevant content of tourism. So, when we want to make a recommendation for B, we can take A as a start point, then the probability that B can accept is higher, that is, the higher the probability of successfully recommend. The strength of the relationship is called the comprehensive relationship strength, which including direct relation and indirect relation based on the users' personal information and interactive activities. Our approach overcomes the limitation that can only calculate the relationship strength between users who has direct association.

3 The SNS user relationship strength model

In this section, we propose a model to measure the relationship strengths between different users in each topic, which is shown as Fig. 1. In this model, we fuse users' similarity, timeliness and interactivity. The users' similarity involves user's personal information and Official Accounts concerned by user. The user's personal information includes user ID, nickname, gender, birthday, job, address, education and label. Timeliness indicates the frequency of the interaction and the time long elapsed since the last time interaction. Interactivity is evaluated through some interactive activities, such as praises, comments as well as forwards of friends' posts. In this paper, we only estimate the relationship strengths between mutually concerned users, not consider the relationship strengths between unilaterally concerned users. Because of the raw data is crawled directly from microblog, so before using these data, we need to preprocess them. Data preprocessing mainly refer to remove the stop words. Stop words include some common words such as pronouns and modal particles. The occurrence

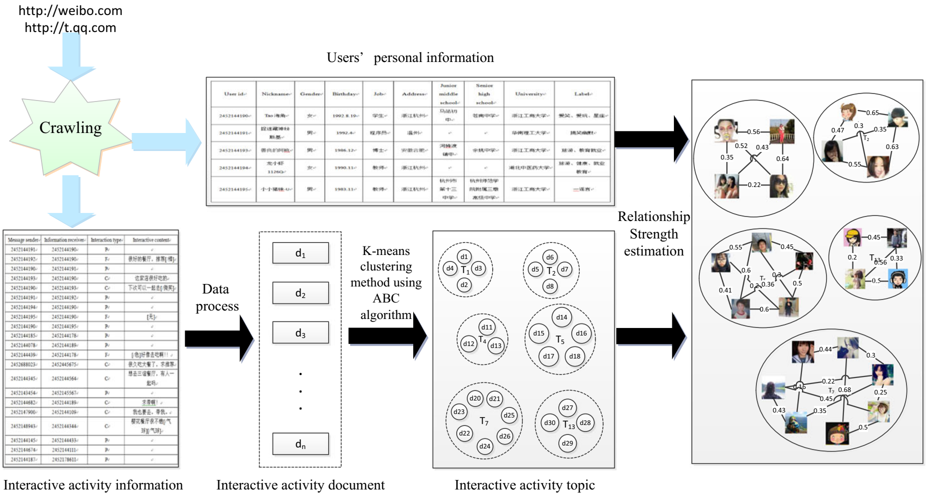
http://weibo.com
http://t.qq.com
Crawling
Users' personal information
Relationship Strength estimation
Data process
K-means clustering method using ABC algorithm
d1
d2
d3
dn
T1
T2
T4
T5
T7
T13
Interactive activity information
Interactive activity document
Interactive activity topic

Fig. 1 The model of relationship strength estimation between online social networks users

frequency of these words is very high but for their subject didn"t help. In this paper, we utilize the stop words dictionary to remove stop words, and then transfer the original data into corresponding documents.

The calculation of users' relationship strength in each activity topic is divided into two steps. Each interactive activity document is subdivided into a certain activity topic. Then we estimate the relationship strengths between different users in each activity topic.

In the first step, interactive activity document is divided into a certain activity topic. We utilize K-means clustering method based on the ABC algorithm [3] and calculate the correlation between each cluster and all activity topic names. In this paper, we consider 13 activity topics, including traffic, sports, military, medicine, politics, education, environment, science and technology, economy, art, law, agriculture and space technology.

In the second step, we estimate users' relationship strength in online social networks according to users' similarity, timeliness and interactivity. This relationship strength is comprehensive relationship strength, which covers direct contact and indirect contact. In view of the indirect contact, we only consider the case that contains only one intermediate node user. When estimate the relationship strength refers to interactivity, we refer to a language modeling based sentiment classification of text, and divide the sentiment of each interactive activities document into "agree" or "disagree" [16].

4 Methodology of the relationship intensity measurement

4.1 Activity topic determination

The dataset is downloaded from microblog including user personal information (user ID, nickname, gender, birthday, job, address, education and label), commonly concerned information (Official Account which is concerned commonly by friends), interactive activity information (praises, comments as well as forwards of friends' posts). In order to utilize the dataset, we need to preprocess the raw dataset, based on the method of stop word dictionary to remove the stop words. After that, we transfer each information data into the corresponding document. Given the interactive activity document set $D = \{d_1, d_2, ..., d_N\}$, where N is the number of the documents. Given the user set $U = \{u_1, u_2, ..., u_s\}$, where s is the number of the users. And for each interactive activity document, the related users refer to post senders, post commenters, post praises or post forwarders. Then, each document will be linked to one or more users. We use a matrix $UD = \{ud_{ij}\}$ to record the relationship between a user and a document, if the indicator ud_{ij} equals to 1, it means the document j is related to user i, otherwise, 0 means not related.

In our paper, we utilize cluster-level based activity topic assignment method to assign an activity topic to each document d_i. It is divided into two steps: firstly, we utilize K-means clustering method based on ABC algorithm [3] to cluster all of the documents. Secondly, we calculate the correlation between clusters and all activity topic names, and assign topic names to each cluster.

In the process of interactive activity document clustering, we use ABC algorithm to overcome the local optimal problem of K-means. As we know, in traditional K-means algorithm, the k value and centre points which will often have a big influence on clustering results. In our clustering method, we use ABC algorithm to determine the optimal value of centre points. According to many literatures in setting the value of k, consider that the optimal value k satisfies the case of $k_{\text{opt}} \le k_{\max}$ and $k_{\max} \le \sqrt{n}$, n indicates the sum number of data. In this clustering algorithm, solutions equate to the cluster centers. The steps of K-means method using ABC algorithm are described as follows.

Step 1. Generate the initial population of solutions (the number of solutions is less than the predicted value k mentioned above) and the maximal search times *limit*. Let the K-means' cost function Eqs. (1) as the objective function.

$$f_i = \frac{1}{k}\sum_{j=1}^{k}\sum_{x_i \in C_j} d(x_i, c_j) \tag{1}$$

Step 2. Produce new solutions for the employed bees, evaluate them and apply the greedy selection process.

Step 3. Calculate the probabilities of the current sources with which they are preferred by the onlookers.

Step 4. Assign onlooker bees to employ bees according to probabilities, produce new solutions and apply the greedy selection process. That is making a clustering iteration of K-means.

Step 5. If the search times *Bas* of an employed bee is more than the threshold *limit*, stop the exploitation process of the sources abandoned by bees and send the scouts in the search area for discovering new food sources, randomly.

Step 6. Memorize the best food source found so far.

Step 7. If the termination condition is not satisfied, go to step 2, otherwise stop the algorithm.

Step 8. Determine the optimal centre points. Then assemble the dataset by these cluster centers and get the final results.

Through K-means clustering method based on ABC algorithm, we will get a set of clusters $C = \{c_1, c_2, ..., c_M\}$. Specifically, each cluster is consisted of one or more documents, and we record its normalized word frequency as an R-dimensional vector TF^m, where element tf_r^m is the normalized frequency of the word w_r in W. In this paper we set 13 activity topic including traffic, sports, military, medicine, politics, education, environment, science and technology, economy, art, law, agriculture and space technology.

We let $Sem(c_m,A_l)$ to denote the correlation between cluster c_m and activity topic name A_l, which is shown as below.

$$Sem(c_m,A_l) = \sum_{r=1}^{R} tf_r^m \times Google\_distance(w_r,A_l) \tag{2}$$

In eq. (2), $Google\_distance(w_r,A_l)$ is the standard Google distance between word w_r and activity topic name A_l. And the definition of standard Google distance between the search word x and search word y is shown as below.

$$NGD(x,y) = \frac{maxlogf(x), logf(x) - logf(x,y)}{logM - minlogf(x), logf(x)} \tag{3}$$

M indicates the total number of web pages searched by Google. $f(x)$ and $f(y)$ are the click number of search word x and search word y. $f(x,y)$ represents the web page number that has x and y at the same time.

We set a threshold value in advance, and then calculate the relevance between cluster c_m and activity topic A_l. We take the most relevant activity topic name as the name of this cluster, whose relevancy should exceed the threshold value which is set with us in advance. Otherwise, if all the values of relevancy are less than the threshold value, then the activity field will be named the field of "others".

4.2 Measurement of user relationship strength

The relationship strength is touching on two users. One user is set as the start point while the other as a destination point. The start point user is called source user, and the destination point is called target user. Relationship strength denotes the intimating degree between source user and target user. Not only the direct relationship strength is considered, but also the indirect relationship strength is taken into account, for both of them may exist between the same users at the same time. For example, user A, B and C they all know each other. If we want to make a recommendation from B for A, then A and C will both gain recommendation information from B. It means A can get this recommended information from B through two routes. One is from B to A directly. The other one is from B to C, and then from C to A, it is an indirect route. This can be explained with that may be A trust C more than B, so A can be persuaded by C more easily. In addition to this, the other situation is A only have direct contact with C, and have no contact with B. So the recommended information can only be passed on from B to C, then to A. There is no direct relationship strength between A and B, only indirect relationship strength, their intimating degree only can be estimated by indirect relationship strength.

Therefore, in this paper, we consider the comprehensive relationship, which includes the direct relationship and the indirect relationship. According to the Six Degree of Separation, a user can make contact with any other user through less than 6 people. However, relationship strength between source user and target user will be weaker with the increase of the number of intermediate node users. When the intermediate nodes are more than two, the relationship strength between source and target user becomes so weak, in this condition, we only consider the case that existence of one intermediate node user.

4.2.1 Direct relationship strength estimation

The direct relationship strength is obtained from user's personal information and user's interactive activity information, which simply describes the intimacy between two directly linked users. This method incorporates the similarity, timeliness, and interactivity. Similarity is measured by the degree of the similarity of the user"s personal information and the similarity of the number of Official Accounts which they commonly concerned. Timeliness is calculated by the frequency of the interaction and the number of days elapsed since the last time interaction. Interactivity is measured by the semantic division of "agree" or "disagree" within praises, comments and forwards of friends' posts of interactive activity [1].

We use $S(u_i,u_j)$ to indicate the calculation formula of similarity between user u_i and user u_j, and use vectors to express the personal information of u_i and u_j. The personal information includes sex, constellation, job, address and education. And the value of constellation is calculated by the value of birthday. Assuming that a piece attribute of personal information is p, if the attribute of user u_i and u_j are the same, the $p(u_i) = 1$, $p(u_j) = 1$, otherwise, $\mathrm{p}(u_i) = 1$, $\mathrm{p}(u_j) = 0$, and the default is $\mathrm{p}(u_i) = 1$, $\mathrm{p}(u_j) = 0$ without this attribute. The similarity of user"s personal information is built on the cosine similarity [23]. And the similarity of the Official Accounts concerned by users is determined based on Jaccard coefficient, the more common Official Accounts are concerned, the more similar they are. It is shown as follows.

$$S\left(u_i,u_j\right)=\frac{\overrightarrow{P_i}\cdot\overrightarrow{P_j}}{\left\|\overrightarrow{P_i}\right\|\left\|\overrightarrow{P_j}\right\|}+\frac{c_i\cap c_j}{c_i\cup c_j}=\frac{\sum\limits_{k=1}^{n}p_{ik}\times p_{jk}}{\sqrt{\sum\limits_{k=1}^{n}p_{ik}^2}\times\sqrt{\sum\limits_{k=1}^{n}p_{jk}^2}}+\frac{c_i\cap c_j}{c_i\cup c_j}\tag{4}$$

$\overrightarrow{P_i}$ indicates the vector composed of personal information of user u_i, $\overrightarrow{P_j}$ indicates the vector composed of personal information of user u_j. If all the personal information attributes of user u_i and u_j are the same, then $\overrightarrow{P_i}$=(1,1,1,1,1), $\overrightarrow{P_j}$ =(1,1,1,1,1). If all the profile information attributes of user u_i and u_j are not the same. $\overrightarrow{P_i}$=(1,1,1,1,1), $\overrightarrow{P_j}$ =(0,0,0,0,0). c_i indicates the set of the Official Accounts concerned by user u_i, c_j indicates the set of the Official Accounts concerned by user u_j. n represents the total number of users' personal information attributes. P_{ik} represents the value of the kth personal information attributes of the user u_i , p_{jk} represents the value of the kth personal information attributes of the user u_j .

We use $T(u_i,u_j)$ to indicate the calculation formula of timeliness between user u_i and user u_j, which is calculated based on the frequency of the interaction and the number of days elapsed since the last time of their interaction [14]. The higher the interaction frequency is, the smaller the number of the days that elapsed from the last time of their interaction is, and the higher the timeliness is. Its function is shown as follows.

$$T\left(u_i,u_j\right)=\frac{I_{ij}}{2I_i}+\frac{I_{ij}}{2I_j}+rT\tag{5}$$

I_{ij} indicates the number of interactions in an activity topic between user u_i and user u_j, I_i indicates the total number of interactions between u_i and other users in an activity topic. I_j indicates the total number of interactions between u_j and other users in an activity topic. Using *Day* represents the number of days elapsed since the last time interaction between user u_i and user u_j. rT is considered to be the weight of the number of days elapsed since the last time interaction between user u_i and user u_j, and its value is determined by experiment. We randomly select 50 experimental users, and provide them with a list of 50 random friends, and then we require the 50 experimental users to divide their friends into three categories including "close friend", "friend", "general acquaintance". According to the experimental results, in "close friend" category, there are 98% users who interact with experimental users, and the number of days elapsed since the last time interaction between them is less than 7 days. In "general acquaintance" category, there are 89% users who interact with experimental users, and the number of days elapsed since the last time interaction between them is less than 14 days. It is shown in Fig. 2. So, according to this, we identify two time limits, such as 7 days and 14 days. According to the difference of intimacy, if $Day \leq 7$, we make $rT = 3a$. If $7 \leq Day \leq 14$, we make $rT = 2a$. If $7 \leq Day \leq 14$, we make $rT = a$. If $14 \leq Day$, we make $rT = 0$, (we made a = 0.05 in the experiment).

We let $Int(u_i,u_j)$ denote the calculation formula of interactivity between user u_i and user u_j, which is measured by the semantic division of "agree" or "disagree" within praise for friends' posts, commenting on friends' posts and forward friends' microblogs of interactive activity.

In order to calculate $Int(u_i,u_j)$, referring to the paper by Hu et al. [7], we propose a language modeling method to detect text emotional tendencies based sentiment classification of text. A very different is that we assume that the corresponding language model of "agree" may be different with the corresponding language model of "disagree", because "agree" and "disagree" might be inclined to different language habits. So, we can divide the text which is the same represented based on a language model into "agree" and "disagree" through exploring the differences between the language models.

We estimate the language model of the two kinds of emotional tendency from training data at first. Then we use a distance function to compare the distance between the language model of test text and the language model of these two emotions. We define classification function φ is shown as below.

$$\varphi\left(d;\theta_p;\theta_N\right) = Dis\left(\theta_d;\theta_p\right) - Dis\left(\theta_d;\theta_N\right) : \begin{cases} < 0 & \text{“}agree\text{”} \\ > 0 & \text{“}disagree\text{”} \end{cases} \tag{6}$$

θ_p indicates the language model of "agree" emotional tendency, it is probability distribution of n-gram. θ_N indicates the language model of "disagree" emotional tendency, a test text generates a language model θ_d. And $Dis(\theta_d;\theta_p)$ is distance between distribution θ_d and distribution θ_p, $Dis(\theta_d;\theta_N)$ is distance between distribution θ_d and distribution θ_N. If $Dis(\theta_d;\theta_p) < Dis(\theta_d;\theta_N)$, it means that test text d is closer to the "agree" emotional tendency. And if $Dis(\theta_d;\theta_p) > Dis(\theta_d;\theta_N)$, it refers to that test text d is closer to the "disagree" emotional tendency. If $\varphi(d;\theta_p;\ \theta_N) = 0$, its emotional tendency will be deemed to be "neutral", but we don't consider this case in this paper. We use Kullback-Leibler Divergence as distance measure between language models.

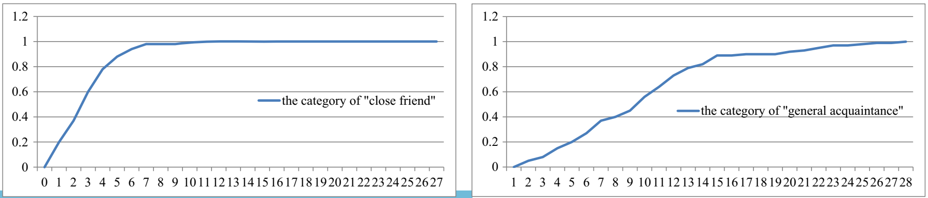

Fig. 2 The user percentage that in the case of the number of the days that elapsed from the last time of interaction less than a certain number in the categories of "close friend" and "general acquaintance"

We use $D(p||q)$ to present the Kullback-Leibler Divergence between two probability distribution $p(x)$ and $q(x)$. The formula is shown below.

$$D(p||q) = \sum_{x} Pr(x) log\left(\frac{p(x)}{q(x)}\right) \tag{7}$$

The KL-Divergence between θ_d and $\theta_p(\theta_N)$ can be calculated by Eqs. (8).

$$\begin{cases} D\left(\hat{\theta}_d||\hat{\theta}_P\right) = \sum\limits_{n-gram} Pr\left(n-gram|\hat{\theta}_d\right) \times log\left(\dfrac{Pr\left(n-gram|\hat{\theta}_d\right)}{Pr\left(n-gram|\hat{\theta}_P\right)}\right) \\ D\left(\hat{\theta}_d||\hat{\theta}_N\right) = \sum\limits_{n-gram} Pr\left(n-gram|\hat{\theta}_d\right) \times log\left(\dfrac{Pr\left(n-gram|\hat{\theta}_d\right)}{Pr\left(n-gram|\hat{\theta}_N\right)}\right) \end{cases} \tag{8}$$

Where $\hat{\theta}$ denotes the estimation model of real model θ. $Pr\left(n\text{–}gram|\hat{\theta}\right)$ denotes the probability of *n-gram* when given the estimated model. So, we can get an emotional classification function as follows shown.

$$\begin{aligned} \varphi\left(d;\hat{\theta}_p;\hat{\theta}_N\right) &= Dis\left(\hat{\theta}_d||\hat{\theta}_p\right) - Dis\left(\hat{\theta}_d||\hat{\theta}_N\right) \\ &= \sum_{n\text{-gram}} Pr\left(n\text{-}gram\middle|\hat{\theta}_d\right) \times log\left(\frac{Pr\left(n-gram\middle|\hat{\theta}_N\right)}{Pr\left(n-gram\middle|\hat{\theta}_P\right)}\right) \end{aligned} \tag{9}$$

In the emotional lexicon of the sentiment classification method, we added some popular phiz of online social networks (such as 💬 [awesome] 、👍 [mighty]) and some popular vocabulary (such as hehe.)

Readers can read the paper wrote by Hu Yi et al. [7] to understand the specific statement of this method.

Interactive activities include praises, comments as well as forwards of friends' posts. When it is praise, it is considered to belong to the "agree" emotional class. When it is a comment, the language model is used to identify the sentiment classification. When it is forwarding behavior, it is divided into two cases including with comments and without comments. It is thought to belong to the "agree" emotional class if without comments. Other, the language model is used to determine the sentiment classification if with comments. According to the emotional classification result of the "agree" and the "disagree", the more interactive activities between the two users belong to "agree", the value of interactivity between them is higher.

We use l_c to indicate the number of instances of interaction between u_i and u_j, and l_a to indicate the number of instances of interaction between u_i and u_j, which is belong to "agree" emotion category. Therefore, fusing the similarity, timeliness and interactivity, we use $Rsd(u_i,u_j)$ indicates the direct relationship strength between users in the same activity topic, which can be expressed as follows.

$$Rsd\left(u_i, u_j\right) = \frac{S \times T \times l_a}{1 + ln(1 + l_c)} \tag{10}$$

4.2.2 Indirect relationship strength estimation

We can calculate the indirect relationship strength, based on the number of relationship paths, the length of the relationship path and the weights of the edges between different users. We let *Rsid*(u_i,u_j) denotes the indirect relationship strength, which describes the closeness of two indirectly linked users. In this paper, we only consider the case that existence of one intermediate node user. Users' relationship strength is greater than 0.5 belongs to strong relationship, otherwise weak relationship. The value of generally strong relationship strength ranges from 0.8 to 0.9 in previous studies. So, according to the real estimation, even for the strong relationship, relationship strength value between the source user and target user is around 0.5 to 0.8 with one intermediate node user. The relationship strength value between the source user and target user is around 0.3 to 0.5 with two intermediate node users. The relationship strength value between the source user and target user is around 0.01 to 0.1 with three intermediate node users. When the intermediate node is more than two, relationship strength between the source and target user becomes weak, so in this paper, we only consider the case that existence of one intermediate node user.

We use *Rsid*(u_i,u_j) indicates the indirect relationship strength between users in the same activity topic.The formula of indirect relationship strength between users is demonstrated below.

$$Rsid\left(u_i,u_j\right)=e^{-2\lambda}\cdot w_1\cdot w_2 \tag{11}$$

Where λ is the attenuation coefficient of the length of each relationship path, $e^{-2\lambda}$ indicates an attenuation function, which ranges from 0 to 1. And w_1, w_2 represent the weights of the first and second edges in the relationship path respectively. Moreover, we let $P_{ij} = \{P_1, P_2, \dots, P_n\}$ denote the relationship paths of u_i and u_j. So, indirect relationship strength between user u_i and user u_j can be shown as follows.

$$Rsid\left(u_i,u_j\right)=\frac{\sum\limits_{i=1}^{n}\left[P_ie^{-2\lambda}\right]}{ne^{-2\lambda}}=\frac{\sum\limits_{i=1}^{n}\left[e^{-2\lambda}w_{1i}w_{2i}\right]}{ne^{-2\lambda}} \tag{12}$$

4.2.3 Comprehensive relationship strength estimation

Let *Rs*(u_i,u_j) be the comprehensive relationship strength between user u_i and user u_j, which contains the direct relationship strength and the indirect relationship strength. It is illustrated in the equation below.

$$Rs\left(u_i,u_j\right)=\alpha Rsd\left(u_i,u_j\right)+\beta Rsid\left(u_i,u_j\right) \tag{13}$$

In the formula, α denotes the weight coefficient of direct relationship strength, and β denotes the weight coefficient of indirect relationship strength. They satisfy $\alpha + \beta = 1$, α, $\beta > 0$.The values of α and β will change dynamically with factors such as the number of user interaction. If α is larger, β will be smaller, it shows the proportion of the direct relationship strength between the users is larger and larger, and the proportion of the indirect relationship intensity will be smaller and smaller with the increase of the number of direct interaction. The users' relationship strength between 0.5 and 1, they belong to strong relationship. The users' relationship strength between 0 and 0.5, they belong to weak relationship. For two users, the probability of their relationship strength is a strong relationship or a weak relationship which both take half part. So, we introduce the influence function of relationship strength, and its formula is shown as follows.

$$\alpha(k) = 1-\left(\frac{1}{2}\right)^{\frac{k}{n-k}} = \begin{cases} 1-\left(\frac{1}{2}\right)^{\frac{k}{n-k}} & n-k \neq 0 \\ 1 & n-k = 0 \end{cases} \tag{14}$$

$\alpha(k)$ represents the dynamic change function with a variable of the number of interaction k. When n-k = 0, that is $k = n$, it means the comprehensive relationship strength between the two users is all derived from the direct relationship strength, and without indirect relationship strength, at this time, $\alpha = 1$. When $k = 0$, $\alpha(k) = 0$, it means there is no direct link between the two users, and the comprehensive relationship strength all comes from the indirect relationship strength.

5 Experiments

5.1 Experimental settings

The dataset is down from Sina microblog and Tencent microblog, which are popular online social networks [2]. To download data from Sina microblog and Tencent microblog, we first randomly select a certain number of users as seed nodes. After obtaining their consents, we collect all their friends who mutually concerned with them. For each of these users, we download their user's personal information, including user ID, nickname, gender, birthday, job, address, education and label. The detailed example of the user's profile is given in Table 1. And these persons' commonly concerned Official Accounts, as Table 2 shows.

Furthermore, for each of these users, we download the interactive activities (praise, comment, forward) between September 2014 and October 2014.The detailed example of the interactive activities information is given in Table 3.

To evaluate the performance, we adopt a manual labeling procedure to generate the ground truths. We randomly choose 30 friends for each seed user and ask all of them to label the relationship strengths. For each user, we provide a list of his or her friends and an microblog topic, then the user labels the relationship strengths on the specific microblog topic with each of his/her friends on the scale of "strong", "weak". When two users label different relationship strengths between them, we will ask them to re-label the relationship strengths.

Table 1 An example to illustrate the user's profile information

Attribute	Value
User ID	2452144190
Nickname	Tao Haijiao
Gender	Female
Birthday	1992.8.19
Job	Student
Address	Hangzhou
Junior middle school	Mazhan junior high school
Senior high school	Cangnan middle school
University	Zhejiang Gongshang University
Label	Dream、Playing、Persistent、Love to laugh、Constellation

Table 2 Commonly concerned Official Accounts of mutually concerned friends

Friend1	Friend2	Commonly concerned official accounts
2,452,144,190	2,452,144,191	158,730 158,731 158,732
2,452,144,190	2,452,144,192	158,786
2,452,144,190	2,452,144,193	158,780 158,784
2,452,144,190	2,452,144,194	158,786 158,780 158,732
2,452,144,190	2,452,144,195	158,786

In the experiment, we first test the performance change of relationship strength with the number of seed nodes. We select different number of seed nodes. Figure 3 shows this result measured by "precision", which indicates the ratio of the correct number of relationship strengths belonging to the strong relationship or weak relationship according to the calculation results of the relationship between users and the number of the total relationship. The result tells that the overall precision achieves the highest point when the number of seed nodes is 10. When the number of seed nodes is less, the total number of users who participated in the experiment is less, and there are less interactive activity documents, thus the words in documents are not so rich. On the other hand, when we select too many seed nodes, there will be appear a phenomenon that newly added user already exists, so the number of users

Table 3 The example of interactive activities information

Message sender	Information receiver	Interaction type	Interactive content
2,452,144,191	2,452,144,190	P(praise)	
2,452,144,192	2,452,144,190	F(forward)	A good restaurant, recommended. [/good]
2,452,144,190	2,452,144,191	P(praise)	
2,452,144,193	2,452,144,190	*C* (comment)	The food in this restaurant is very delicious.
2,452,144,190	2,452,144,193	*C* (comment)	We can go together next time. [/smiling]
2,452,144,191	2,452,144,192	P(praise)	
2,452,144,194	2,452,144,190	P(praise)	
2,452,144,195	2,452,144,190	F(forward)	[nothing]
2,452,144,190	2,452,144,195	P(praise)	
2,452,144,185	2,452,144,178	P(praise)	
2,452,144,078	2,452,144,189	P(praise)	
2,452,144,439	2,452,144,178	F(forward)	[/Drool] Really want to eat ah!!
2,452,688,023	2,452,445,675	*C* (comment)	Long time didn't go to eat big dinner, please recommend.
2,452,144,345	2,452,144,564	*C* (comment)	I want to go to three restaurants, anybody together
2,452,143,454	2,452,145,567	P(praise)	
2,452,144,682	2,452,144,189	*C* (comment)	O takes!
2,452,147,900	2,452,144,109	*C* (comment)	I also want to go, take me.
2,452,148,943	2,452,144,344	*C* (comment)	Sakura restaurant is very good [/Sun][/Sun]
2,452,144,145	2,452,144,433	P(praise)	
2,452,144,674	2,452,144,111	P(praise)	
2,452,144,187	2,452,178,611	P(praise)	

In the above table, *P* represents praise, *C* represents comment, *F* represents forward

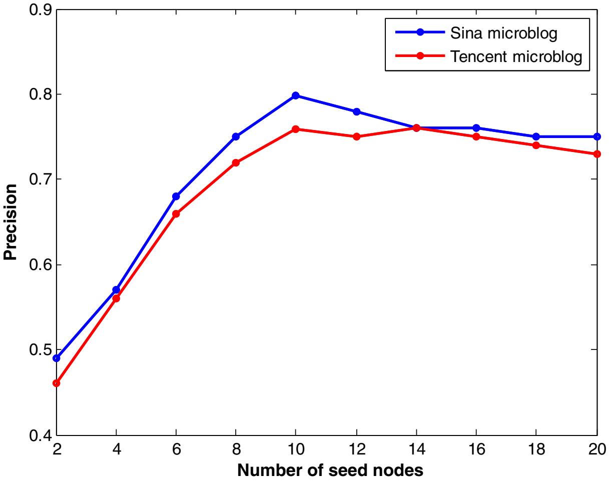

Fig. 3 The performance changes of relationship strength with the number of seed nodes

participated in the experiment will not increase so fast with the increase of seed nodes. Even so, the number of interactive activity documents will increase to a certain extent with the increase on the number of users, though with low discriminative power.

After we obtain the best number of the seed nodes is 10, we respectively and randomly select 10 users in these two datasets as the seed nodes, and collect all their friends who mutually concerned with them, which results in a total of 1581Sina persons and 1235 Tencent persons. For each of these users, we download their personal information and interactive activities between September 2014 and October 2014. It results in a total of 295,300 Sina interactive activity documents and 215,000 Tencent interactive activity documents. After manual labeling procedure, we obtain 17,123 relationship strengths from Sina microblog and 12,566 relationship strengths from Tencent microblog.

5.2 Experimental results

In this part, we exploit Precision (as mentioned above), Recall and the normalized Discounted Cumulative Gain (*nDCG*), to measure the relationship strength estimation result, which is estimated in part 4.2.

Precision and Recall are two measures in the field of information retrieval and statistical classification, which are used to evaluate the quality of the results. For the evaluation of the results of users' relationship strength calculation, Precision indicates the ratio of the correct number of relationship strength belonging to strong relationship or weak relationship according to the calculation results of the relationship between users and the number of total relationship. Recall indicates the ratio of the correct number of relationship strength belonging to strong relationship or weak relationship according to the calculation results of the relationship between users and the number of relationship strength actually belonging to each relationship strength range. For example, there are *A* relationships belonging to strong relationship strength, and *B* relationships belonging to weak relationship strength in fact. According to the results of relational strength calculation, there are *C* strong relationship strengths and *D* weak relationship strengths. Among them, *E* relationships are right in strong relationship strength, and *F* relationships are right in weak relationship strength. Precision is shown as Eqs. (15).

$$P = \frac{E+F}{C+D} \tag{15}$$

Recall is shown as Eqs. (16).

$$R = \frac{1}{2} \cdot \left(\frac{E}{A} + \frac{F}{B} \right) \tag{16}$$

What's more, *nDCG* is an indicator of PageRank that is widely used in the search engine. It considers both the importance of searching results and the relative location of searching results. If the strong correlation takes a higher rank, then the more effective the method is. Otherwise, the method will be punished. The formulation of *nDCG* is shown as follows.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{17}$$

In this equation, *IDCG* is an ideal *CDG*. Then we sort the results manually. In the best order status, we calculate the *DCG* of query, which is called *IDCG*.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{log_2(1+i)} \tag{18}$$

The average *NDCG* is the average *nDCG* of all the users in one certain activity field.

In order to verify our method, we compare it with the following two methods.

1. Linear combination method (here we denote this method as "LCM"): The method calculates the relationship intensity between two users in the same activity field through the linear combination of user profile information and interactive activity information. However, it is limited to the directly linked users.
2. Latent variable model (here we denote this method as "LVM"): This is a latent variable model method, which utilizes the user interactive activity information and the user profile information to estimate the user relationship intensity.

In our experiment, we only consider the top 30 friends of each user. Based on Sina microblog dataset and Tencent microblog dataset, we get the comparison results. The final result of these methods and the result of ours according to Precision and Recall is shown in Table 4.

The table shows that our method is better than LCM and LVM according to Precision and Recall.

Table 4 The final result of these methods and the result of ours

Method	Precision		Recall	
	Sina microblog	Tencent microblog	Sina microblog	Tencent microblog
LCM	0.551	0.520	0.572	0.544
LVM	0.569	0.553	0.585	0.572
Our method	0.798	0.759	0.691	0.671

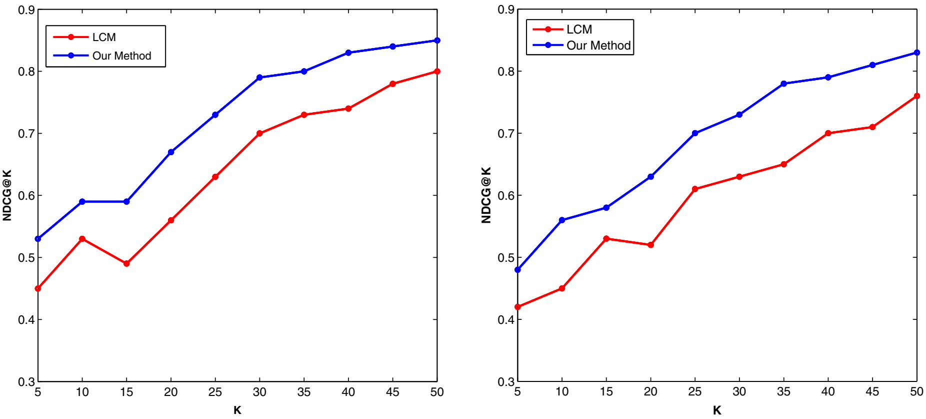

Fig. 4 The average *nDCG* of LCM method and our method based on two microblog dataset

For the linear combination method, we try several combinations of weights. Based on Sina microblog dataset and Tencent microblog dataset, we get the comparison results. The final result of this method and the result of ours are shown in Fig. 4.

In comparison with latent variable model, we compare the average *nDCG* of our method and the latent variable model method in each microblog topic. The result is displayed in Fig. 4.

We can see in Fig. 4 that our method is superior to the linear combination method with several combinations of weights. And we can also see from Fig. 5 that our method is better than the latent variable model method in each microblog topic, which indicates that our method is feasible and effective. This is because our method not only takes accounting the different relationship strengths in different microblog topics, but also considers the direct relationship strength and the indirect relationship strength.

6 Conclusion and future works

In this paper, we proposed a relationship strength estimation model between different users in online social networks, which fusing of activity topics and indirect relationship. In our experiment, we exploited Sina and Tencent microblog dataset to verify our method with the users' personal information and the microblog interactive activity information. These data were leveraged in the proposed users' relationship strength model to estimate the relationship strengths. And in this model, we considered the direct relationship and indirect relationship on

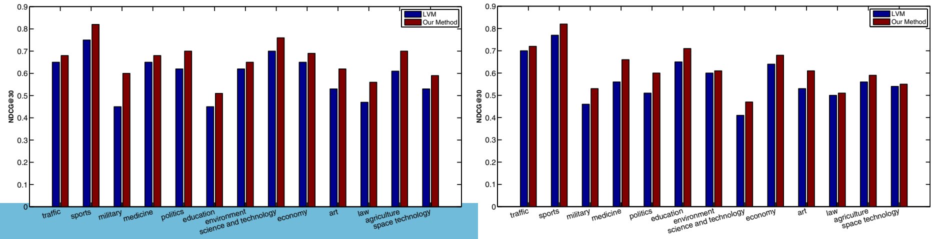

Fig. 5 The average *nDCG* of LVM method and our method based on two microblog dataset

each microblog topic. We conducted experiments on two microblog dataset and the results demonstrated the feasibility and effectiveness of our approach. However, there are still some shortcomings in our paper, like the limited dataset in our experiment, which is resulted from the limited time. In our future research, we will conduct experiments with more users. What's more, users' relationship strength in online social networks can improve the range and performance such as link prediction, news feed, item recommendation, and visualization. And we will consider more interesting applications whose performance can be obviously improved with the estimated relationship strength.

Acknowledgments This research is supported by The National Key Technology R&D Program of China (Grant 2014BAH24F06); Natural Science Foundation of China (No.71571162); Zhejiang Province philosophy social sciences planning project (No.16NDJC188YB); College Students" science and technology innovation activities of Zhejiang Province (2016R408080). This research is supported by the Contemporary Business and Trade Research Center of Zhejiang Gongshang University which is the Key Research Institutes of Social Sciences and Humanities Ministry of Education (14JJD630011). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Baecchi C, Uricchio T, Bertini M et al (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia[J]. Multimedia Tools and Applications 75(5):2507–2525
2. Cheng W, Liu B (2015) Empirical study on the personal information disclosure of micro-blog UsersBased on credibility analysis: taking the Sina micro-blog as example. Journal of Intelligence 08:169–176
3. Chunhua Ju, Chonghuan Xu (2013). A new collaborative recommendation approach based on users clustering using artificial bee colony algorithm. The Scientific World Journal 2013: Article ID 869658
4. Deng Z-S (2015) An entropy model to infer social strength based on texts of subjects. Jisuanji Yu Xiandaihua 02:30–33
5. Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In Proceedings of the SIGCHI conference on human factors in computing systems. ACM 211–220
6. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
7. Hu Y, Lu R, Li X et al (2007) Research on language modeling based sentiment classification of text. Journal of Computer Research and Development 44(9):1469–1475
8. Kim M, Park SO (2013) Group affinity based social trust model for an intelligent movie recommender system[J]. Multimedia tools and applications 64(2):505–516
9. Li P, Yu Y, Li Y et al (2015) Improve LDA microblogs topics model based on weight microblogs chain. Application Research of Computers 07:1–5
10. Lin X, Shang T, Liu J (2014) An estimation method for relationship strength in weighted social network graphs. Journal of Computer and Communications 2(4):82–89
11. Liu F, Lee HJ (2010) Use of social network information to enhance collaborative filtering performance. Expert Syst Appl 37(7):4772–4778
12. Nuñez-Gonzalez JD, Graña M, Apolloni B (2015) Reputation features for trust prediction in social networks. Neurocomputing 166:1–7
13. Pham H, Shahabi C, Liu Yan (2013) Ebm: an entropy based model to infer social strength from spationtemporal data. In: Proceedings of ACM SIGMOD conference 265–276
14. Shen H, Yuan Q (2014) A classification study on the strength of social relationship based on social network. Journal of the China Society for Scientific and Technical Information 8(33):846–859
15. Wilson C, Boe B, Sala A et al. (2009) User interactions in social networks and their implications. In Proceedings of the 4th ACM European conference on computer systems 205–218
16. Wu F, Huang Y, Song Y (2016) Structured microblog sentiment classification via social context regularization. Neurocomputing 175:599–609
17. Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: Proceedings of ACM International Conference on World Wide Web
18. Xu K, Zou K, Huang Y et al (2016) Mining community and inferring friendship in mobile social networks. Neurocomputing 174:605–616

19. Zhao Y, Li Y (2012) Research on forecasting personality traits and relationship strength of social network users. In Proceedings of the seventh MAM conference on business intelligence 10
20. Zhao X, Yuan J, Li G et al (2012) Relationship strength estimation for online social networks with the study on Facebook. Neurocomputing 95:89–97
21. Zhao W, Zhao Y, Zhu Q et al (2013) A simulation study on information diffusion in social network under Web2.0 environment. Journal of the China Society for Scientific 32(5):511–521
22. Zhou X, Wang W, Jin Q (2015) Multi-dimensional attributes and measures for dynamical user profiling in social networking environments[J]. Multimedia Tools and Applications 74(14 k):5015–5028
23. Zhu W (2014) Research on user similarity function of recommender systems. Chongqing: College of Computer Science, Chongqing University, 1–49

Chunhua Ju is a professor, doctoral supervisor and division chief of science and technology department in Zhejiang Gongshang University who focuses on intelligent information processing, data mining and E-commerce. And he won the award for "New Century Excellent Talents in University" of China. In the past several years, he led more than 6 national projects. He has published more than 30 papers which are SCI and EI indexed.

Wanqiong Tao is a postgraduate in Zhejiang Gongshang University. Her research focuses on intelligent information processing and data mining.