Measuring Social Relations in the Web based on Search Engine and Text Analysis: A Model and Implementation

Meijuan Yin, Xiaonan Liu, Junyong Luo, Yan Liu, Ziqi Tang

State Key Laboratory of Mathematics Engineering and Advanced Computing, China
raindot_ymj@163.com, nine_day@163.com, luojunyong@vip.371.net, ms_liuyan@aliyun.com, 690245411@qq.com

Abstract

To improve the accuracy and stability of existing methods to measure social relations in the Web, a novel model of relation measuring methods based on both a search engine and text analysis is proposed. The model measures the strength of social relations according to both the co-occurrence of two persons' names in web pages obtained by a web search engine and the co-occurrence of two person names in sentences of the text of web pages as found by text analysis. The formalized description of the model is then presented. To evaluate the effectiveness of the proposed model, the specific implementation of the model is presented in detail. Experimental results show that compared with existing methods only based on a search engine or text analysis, the relation weights obtained by the specific methods based on the proposed model are more accurate and stable.

Keywords: Web social networks, Social network extraction, Relation measuring, Search engine, Text analysis

1 Introduction

Extracting and analyzing social relations based on web information is one of the popular research topics in data mining and social network analysis [1-3]. One method to extract social relations is to automatically extract relations from various sources of information on the web such as web pages [5-6], web snippets [7], e-mail archives [8-9], paper citation information [10], schedule data [11] and relational databases [12]. These data sources are classified into two sorts by the range of people whose social relations can be extracted, (1) the sort covering information of people only in a specific field, such as paper citation information, schedule data, relational databases and so on, which can be used to measure social relations of a very narrow crowd with one or several types of specific relations; (2) the other covering a variety of people in different fields, such as web pages (including web snippets), e-mail archives, instant communication information, social graphs in Social Network Service (SNS) etc., by which social relations of persons in more fields can be measured. In the second sort, except web pages, these sources are not open to the public. It is very difficult to get enough data for social relation measurement from these data sources. Whereas web pages are open to the public. As long as persons' names occur in the web, we can measure their social relations, no matter they are famous or not.

This paper focuses on extracting social networks from web pages, which is named "web social network". Most of existing related methods measure the weight of relations among persons by the co-occurrence coefficient metrics [4, 13-14]. The basic assumption is that the co-occurrence of name pairs in web pages indicates the strength of their relation. As this method does not take the relative position of two co-occurring names in web pages into account, redundant noise relations are often extracted [15]. Only few studies proposed methods to weight relations in web social networks based on the co-occurrence of person names in sentences or text sliding windows of web pages [15-16]. The relation weights measured by this method are more accurate, however, only relations with strong weights can be extracted [15].

To improve the performance of relation measuring methods for extracting web social networks, we propose a novel sort of methods based on both web search engines and text analysis of web pages, and present a model to represent this sort of methods. To evaluate the effectiveness of the model, its specific implementations are presented in detail, including two kinds of relation measuring functions designing and two method instances on the function construction. The experimental results on names of stars and academic researchers show that compared with the typical methods only based on a search engine or text analysis, the designed methods based on the proposed model can measure the strength of relations in web social networks with better accuracy and stability.

The remainder of this paper is presented as follows. Section 2 presents the related work. A novel relation measuring model and its formalization description are presented in Section 3. Section 4 presents the

*Corresponding Author: Meijuan Yin; E-mail: raindot_ymj@163.com
DOI: 10.3966/160792642018031902014

implementation of the model in detail. In Section 5, we evaluate the established method instances on real names of individuals. We conclude the paper in the last section.

2 Related Work

Most existing methods to weight relations in web social networks are based on the co-occurrence coefficient metrics [4], which are computed by the count of co-occurring web pages for two names usually obtained by using a web search engine [13, 17]. Co-occurrence coefficient metrics [18] include Matching coefficient, Mutual Information, Dice coefficient, Jaccard coefficient, Overlap coefficient, and Cosine coefficient.

Many related works have used co-occurrence coefficient metrics to measure weights of relations among individuals [4, 13]. The systems Referral-Web [13] and FLINK [14] used Jaccard coefficient, the recall of relations extracted by which is not very high. Lee et al. [19] used Matching coefficient, which cannot accurately describe the relative closeness of relations. POLYPHONET [20-21] used Overlap coefficient, which compensates for the limits of Jaccard coefficient, but cannot precisely reflect the relative closeness of relations for the individual with the bigger occurrence frequency [22]. Reference [4] compared these metrics and the results show Overlap coefficient performs best. He et al. [22] and Canaleta et al. [23] compared all the metrics except Matching coefficient, and the results indicate that Cosine coefficient performs best and Overlap coefficient is better than Jaccard coefficient.

For individuals in different fields or with greatly dissimilar occurrence frequency in web pages, the results of methods by using only one kind of metrics are usually relatively instable [24]. To compensate the deficiency of instable results of these methods, Jin et al. [24-25] and Matsuo et al. [26] used several metrics concurrently to reduce the difference of the relation number between different individuals. However, by these methods, a relation can only be extracted, yet, its strength can not be measured.

Besides, the methods based on co-occurrence coefficient metrics neglect that contents in web pages are semi-structured, which may result in the extraction of redundant noise relations [15]. To improve this problem, Li et al. [16] measured the relation strength by the frequency of names co-occurring in sentences in web pages and blogs. Di et al. [15] used the co-occurrence of names in sliding windows and beside coordinative conjunction words to measure relations. Relation weights obtained by these two methods are accurate. However, some relations with strong strength cannot be extracted, when not satisfying the strong co-occurrence condition. Besides, only the retrieved top-ranking web pages with names co-occurring are usually analyzed [16], which may lose some important co-occurrence information and reducing the recall of the extracting results.

So existing methods can be grouped into two classes, one is based on a search engine, and the other is based on text analysis. The former can get a high recall of extracted relations, but may result in the extraction of redundant noise relations and the relation weights obtained by different metrics are not stable. The latter methods can get a relatively high accuracy, but may lose some important relations and most of the relations with an inadequately high weight.

3 Relation Measuring Model: SETARM

We draw on the advantages of two sorts of methods and propose a novel sort of methods, which measure relations in web social networks based on both search engines and text analysis. To describe methods of this sort in a unified form, we present a model, called Search Engine and Text Analysis based Relation Measuring model, abbreviated as SETARM model. The model can represent all the specific methods of the proposed sort measuring social relations on web based on both search engines and text analysis in a unified expression.

Methods based on SETARM model first get the co-occurrence of two names in web pages by using a search engine (referred to as web page co-occurrence), and then analyze the texts derived from the retrieved top-ranking web pages got in the first step to obtain the co-occurrence of names in sentences of web pages (referred to as sentence co-occurrence). The relation strength for the corresponding individuals is then calculated according to both the web page co-occurrence and the sentence co-occurrence. Since the model takes into account both the name co-occurrence in the whole web pages and the name co-occurrence in intra-page close positions in some web pages, the relation strength for individuals measured by the model will be more accurate than that measured by only one of the two kinds of co-occurrence mentioned above.

Referring to the architecture toward general social networks extraction from the Web using a search engine [24], the formalization of the SETARM model is described as follows.

The weight of the relation between two individuals with names X and Y in a web social network is denoted as $w(X,Y)$, which represents the strength of social relationship between them. f denotes the relation measuring function used to compute the weight $w(X,Y)$ by the co-occurrence derived from web pages. We consider that the function f can be generally denoted as:

$$w(X,Y) = f(\mathbb{S}(X,Y),\Theta) \rightarrow [0,1] \quad \textbf{(1)}$$

where $\mathbb{S}(X,Y)$ is a 2-dimensional vector space $(S_1(X,Y),S_2(X,Y))$ to represent two sorts of methods to

measure the weight of the relation between X and Y, $S_1(X,Y)$ is one of the methods based on a search engine, and $S_2(X,Y)$ is one of the methods based on text analysis. For example, $S_1(X,Y)$ can be either Jaccard coefficient ($n_{X\wedge Y}/n_{X\vee Y}$), Overlap coefficient ($n_{X\wedge Y}/\min(n_X,n_Y)$), or Cosine coefficient ($n_{X\wedge Y}/\sqrt{n_X n_Y}$) and so on, where n_X and n_Y respectively represent the number of hit pages yielded by a search engine for the query "X" and the query "Y", and $n_{X\wedge Y}$ and $n_{X\vee Y}$ respectively represent the number of hit pages yielded by a search engine for the query "X AND Y" and the query "X OR Y". $S_2(X,Y)$ can be the frequency of the sentences in which X and Y co-occur, or the number of the pages in which X and Y co-occur in one or more sentences. Θ is a *K*-dimensional vector space $(\theta_1,\theta_2,...,\theta_K)$ to represent parameters in the function f. The function f calculates the weight $w(X,Y)$ with the range [0, 1] according to the weight evaluated by the method $S_1(X,Y)$, the weight obtained by the method $S_2(X,Y)$ and the parameter vector Θ. Besides, in the model, if we use $w_1(X,Y)\in[0,1]$ and $w_2(X,Y)\in[0,1]$ to respectively represent the weight evaluated by the method $S_1(X,Y)$ and the weight obtained by the method $S_2(X,Y)$, then the higher the value of $w_1(X,Y)$ or $w_2(X,Y)$ is, the greater the weight $w(X,Y)$ calculated by the function f is, that is to say, the relationship between the function f and the weight $w_1(X,Y)$ or $w_2(X,Y)$ is a positive correlation.

The model seems to be similar to the model of general social networks extraction from the web using a search engine [24], but the meanings of two models are completely different. The SETARM model are based on two sorts of relation measuring methods, the one based on a search engine and that based on text analysis. While the latter model based on different correlation coefficient metrics, which are all based on a search engine. And the measuring results on SETARM are from 0 to 1, which indicates the weight of a relation between two people, while the results on the latter model are 0 or 1, only indicating whether the relation is exist or not. The advantages of the SETARM model can be concluded as: (1) the measuring results can insure both a good recall and a lower influence by noisy pages of names co-occurring. (2) the results describe the relations more distinctly, which can differentiate the relation intensities for different person pairs.

4 Implementation of Methods on SETARM

To evaluate the effectiveness of the SETARM model, the implementation of specific methods based on the model is described in detail as follows, including the design of the function f and the establishment of the methods $S_1(X,Y)$ and $S_2(X,Y)$.

4.1 Relation Measuring Function Designing

The design of the function f is referring to multiple regression analysis in Probability and Statistics and goal programming in Operational Research.

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables, especially a dependent variable and one or more independent variables. Multiple regression analysis is the type with two or more independent variables. There are linear and nonlinear relationships in regression analysis. When the independent variables are given, the general multiple linear regression model is:

$$y=\beta_0+\beta_1 x_1+...+\beta_k x_k,$$

where, β_0 is a constant term, $x_1,...,x_k$ are k independent variables and $\beta_1,...,\beta_k$ are their regression coefficients. Nonlinear regression models include exponential functions, logarithmic functions, trigonometric functions, power functions, Gaussian function, and Lorenz curves. Among them, the power function is the best model to design the function of SETARM, as the relationship between the dependent variable (f) and each independent variable (S_1 , S_2) is positive correlation in our model SETARM and dimensions of three variables are the same. The general power function model is:

$$y=\beta_0\prod_{i=1}^{k}x_i^{\beta_i},$$

where, β_0 is a constant parameter, $x_1,...,x_k$ are k independent variables and $\beta_1,...,\beta_k$ are their regression coefficients. Both the linear regression model and the nonlinear regression model should be concerned when we design the function f of SETARM.

Goal programming is a branch of multi-objective optimization, which in turn is a branch of multi-criteria decision analysis. It is an optimization programming. Goal programming is usually used to perform this type of analysis that is providing the best satisfying solution under a varying amount of resources and priorities of the goals. In this paper, we refer this thought of constraint solving to narrow the solution space of the goal function by constraint conditions step by step, and get the final solution of the function at last.

We denote the weight $w_1(X,Y)$ calculated by $S_1(X,Y)$ as S_1, and the weight $w_2(X,Y)$ calculated by $S_2(X,Y)$ as S_2, then the steps to design the function f based on both the linear regression model and the

nonlinear regression model are as follows.

Step 1: Design the basic expression of function f. There is a basic constraint in the SETARM model that the relationship between the function f and S_1 or S_2 is a positive correlation, that is f should satisfy the constraints of $f(S_1,S_2)>f(S_1^{'},S_2)$, $where\ S_1>S_1^{'}$ and $f(S_1,S_2)>f(S_1,S_2^{'})$, $where\ S_2>S_2^{'}$. We merge the similar terms in the linear model and different forms of the nonlinear model to one term and get the basic expression of function f f as formula (2):

$$\begin{aligned} & f(S_1,S_2)=\theta_0+\sum\theta_k S_1^{\ i}S_2^{\ j} \\ & s.t.\quad i,j\in R,(i,j)\geq 0,i+j\neq 0 \\ & \qquad k\in K,k\geq 1 \\ & \qquad (\theta_0,\theta_k)\geq 0 \end{aligned} \tag{2}$$

Step 2: Narrow the solution space of the expression. As the weights $w(X,Y)$, $w_1(X,Y)$ and $w_2(X,Y)$, respectively calculated by f, $S_1(X,Y)$ and $S_2(X,Y)$, are all the weights to measure the same relation strength, the dimensions of $w(X,Y)$, $w_1(X,Y)$ and $w_2(X,Y)$ should be the same to each other, that is the value of f, S_1 and S_2 are in the same dimension. That is to say the dimension of each item in formula (2) including θ_0 and every $\theta_k S_1^{\ i}S_2^{\ j}$ should be the same with that of the value of f, S_1 and S_2. So $\theta_0=0$ as it is an item without a dimension and $i+j=1$ to make the dimension of the item $\theta_k S_1^{\ i}S_2^{\ j}$ is the same with that of S_1 and S_2. So the solution space of the expression of f is narrowed to the following form:

$$\begin{aligned} & f(S_1,S_2)=\sum\theta_k S_1^{\ i}S_2^{\ j} \\ & s.t.\quad i,j\in R,(i,j)\geq 0,i+j=1 \\ & \qquad k\in K,k\geq 1 \\ & \qquad \theta_k\geq 0 \end{aligned} \tag{3}$$

if we separate the items with $i=0$ or $j=0$ from $\sum\theta_k S_1^{\ i}S_2^{\ j}$, then formula (3) can be equivalent to the following form:

$$\begin{aligned} & f(S_1,S_2)=\theta_1S_1+\theta_2S_2+\sum\theta_k S_1^{\ i}S_2^{\ j} \\ & s.t.\quad i,j\in R,(i,j)>0,i+j=1 \\ & \qquad k\in K,k\geq 3 \\ & \qquad (\theta_1,\theta_2,\theta_k)\geq 0 \end{aligned} \tag{4}$$

The formula (4) shows that the expression of f includes both the expression of the linear coupled combination of S_1 and S_2, $\theta_1S_1+\theta_2S_2$, and an infinite number of the expression of the nonlinear coupled combination of them, $\theta_k S_1^{\ i}S_2^{\ j}$.

To study the relationship between the function f and the single type of coupled combination of S_1 and S_2, we choose the single expression of the nonlinear coupled combination of S_1 and S_2, $\theta_k S_1^{\ i}S_2^{\ j}$, and the expression of the linear coupled combination of them, $\theta_1S_1+\theta_2S_2$, as two typical solutions of f on the conditions of different type of coupled combination of S_1 and S_2.

(1) The function of the nonlinear coupled combination of S_1 and S_2:

$$\begin{aligned} & f_{NLC}(S_1,S_2)=\theta_3\cdot S_1^{\ i}\cdot S_2^{\ j} \\ & s.t.\quad i,j\in R,i,j>0,i+j=1 \\ & \qquad \theta_3\geq 0 \end{aligned} \tag{5}$$

if we let $\theta'=j$, then $i=1-\theta'$ and the above expression can be equivalent to the formula (6).

$$\begin{aligned} & f_{NLC}(S_1,S_2)=\theta_3\cdot S_1^{\ (1-\theta')}\cdot S_2^{\ \theta'} \\ & s.t.\quad 0<\theta'<1,\theta_3\geq 0 \end{aligned} \tag{6}$$

(2) The function of the linear coupled combination of S_1 and S_2:

$$\begin{aligned} & f_{LC}(S_1,S_2)=\theta_1\cdot S_1+\theta_2\cdot S_2 \\ & s.t.\quad (\theta_1,\theta_2)\geq 0. \end{aligned} \tag{7}$$

Step 3: Deduce the range of the parameter value in the expression. There is another basic constraint in the model that the range of the value for f is [0, 1], so the specific solutions $f_{NLC}(S_1,S_2)$ and $f_{LC}(S_1,S_2)$ should also satisfy the constraint, that is the inequalities $0\leq\theta_3\cdot S_1^{\ (1-\theta')}\cdot\ S_2^{\ \theta'}\leq 1$ and $0\leq\theta_1\cdot S_1+\theta_2\cdot S_2\leq 1$ should be satisfied. Thus we can determined the domains of parameters θ_1, θ_2 and θ_3.

$if(0\leq\theta_3<1)\Rightarrow 0\leq\theta_3\cdot S_1^{\ (1-\theta')}\cdot S_2^{\ \theta'}<S_1^{\ (1-\theta')}\cdot S_2^{\ \theta'}\leq 1$, that is $0\leq f_{NLC}(S_1,S_2)\leq 1$, which dissatisfies the value constraint of f. Also $if(\theta_3>1)\Rightarrow\exists(S_1=S_2=1)$, $\theta_3\cdot S_1^{\ (1-\theta')}\cdot S_2^{\ \theta'}=\theta_3>1$, that is $f_{NLC}(S_1,S_2)>1$, which also dissatisfies the constraint of the value of f. So $\theta_3=1$.

Similarly, $if\ (0\leq\theta_1+\theta_2<1)\Rightarrow 0\leq\theta_1\cdot S_1+\theta_2\cdot S_2\leq\theta_1+\theta_2<1$, that is $0\leq f_{LC}(S_1,S_2)\leq 1$, which dissatisfies the value constraint of f. Also $if(\theta_1+\theta_2>1)\Rightarrow\exists(S_1=S_2=1)$, $\theta_1\cdot S_1+\theta_2\cdot S_2=\theta_1+\theta_2>1$, that is $f_{LC}(S_1,S_2)>1$, which dissatisfies the value constraint of f. So $\theta_1+\theta_2=1$. Besides, the constraint of $(\theta_1,\theta_2)\geq 0$ should also be satisfied, so $0\leq(\theta_1,\theta_2)\leq 1$.

In summary, the domains of the parameters θ_1, θ_2 and θ_3 in (6) and (7) are $\theta_1+\theta_2=1$, $0\leq(\theta_1,\theta_2)\leq 1$ and $\theta_3=1$.

If we let $\theta=\theta_2$, then $\theta_1=1-\theta$ and the final expressions of two specific functions of f are as

follows:

$$f_{NLC}(S_1,S_2) = S_1^{(1-\theta')} \cdot S_2^{\theta'} \quad s.t. \quad 0 < \theta' < 1 \tag{8}$$

$$f_{LC}(S_1,S_2) = (1-\theta)\cdot S_1 + \theta \cdot S_2 \quad s.t. \quad 0 \le \theta \le 1 \tag{9}$$

4.2 Specific Methods Establishing on SETARM

To construct the method of measuring relations based on the function f, the methods $S_1(X,Y)$ and $S_2(X,Y)$ in the model must be determined after the design of the function f, as both of the methods $S_1(X,Y)$ and $S_2(X,Y)$ have several ways to measure relations.

Among the basic methods based on a search engine, Cosine coefficient shows the best performance, and it can characterize the co-occurrence of names in web pages in a better way [22-23]. So we take Cosine coefficient as the method $S_1(X,Y)$. There are mainly three methods based on text analysis of web pages, including the frequency of the sentences with names co-occurring (referred to as "sentence co-occurrence frequency"), the number of the pages with names co-occurring in one or more sentences (referred to as "sentence co-occurrence page number"), and the rate of the pages with names co-occurring in sentences in all the web pages with names co-occurring (referred to as "sentence co-occurrence page rate"). The first two methods ignore the number of the web pages with names co-occurring, which may result in that the weight cannot reflect the relative strength of relations objectively. So we choose the last one as the method $S_2(X,Y)$.

Let $n(X)$ and $n(Y)$ respectively be the number of hit pages yielded by a search engine for the query "X" and "Y". Let $n(X,Y)$ be the number of hit pages for the query "X AND Y", and let $n_t(X,Y)$ be the number of web pages with names co-occurring in one or more sentences, which is calculated according to text analysis of the whole hit pages for the query "X AND Y". Then by the basic methods $S_1(X,Y)$ and $S_2(X,Y)$ we have chosen, the formulas to calculate the relation strength between X and Y are as follows:

$$w_1(X,Y) = \frac{n(X,Y)}{\sqrt{n(X)}\cdot\sqrt{n(Y)}} \tag{10}$$

$$w_2(X,Y) = \frac{n_t(X,Y)}{n(X,Y)} \tag{11}$$

in formula (11), if the value of $n(X,Y)$ is great, it is so time-consuming to extract and analyze texts from all of the retrieved web pages that it is infeasible to do text analysis and get the value of $n_t(X,Y)$. However, as it is well-known that the top M search results yielded by a search engine are usually more accurate than the search results followed by the top M results, and for example the value of M is 100 for Google [27], it is reasonable to use the top search results to measure the relation weight between two persons, and many researchers have done text analysis on only the top M search results rather than on the whole search results in web social network extraction [3-4, 28]. So we use the rate of pages with names co-occurring in sentences in the top M web pages yielded by a search engine for the query "X AND Y" instead of the formula (11) to calculate $w_2(X,Y)$, and the corresponding formula is as follows:

$$w_2(X,Y) = \frac{n_t'(X,Y)}{\min\{M, n(X,Y)\}} \tag{12}$$

in formula (12), M denotes the number of the top search results ranked by a search engine, and the top M search results are relatively accurate in the hit pages, and for different search engines, the values of M vary. $n_t'(X,Y)$ denotes the number of pages with names co-occurring in one or more sentences in the top $\min\{M, n(X,Y)\}$ hit pages.

So according to formula (8) and formula (9) of function f and formula (10) and formula (12) of the basic methods $S_1(X,Y)$ and $S_2(X,Y)$, we can finally construct two specific methods to measure relation in web social networks based on the SETARM model. The expressions are as follows:

$$w(X,Y) = \begin{cases} f_{NLC}(X,Y) = \left(\frac{n(X,Y)}{\sqrt{n(X)}\cdot\sqrt{n(Y)}}\right)^{1-\theta'} \cdot \left(\frac{n_t'(X,Y)}{\min\{M,n(X,Y)\}}\right)^{\theta'}, 0 < \theta' < 1 \\ f_{LC}(X,Y) = (1-\theta)\cdot\frac{n(X,Y)}{\sqrt{n(X)}\cdot\sqrt{n(Y)}} + \theta\cdot\frac{n_t'(X,Y)}{\min\{M,n(X,Y)\}}, 0 \le \theta \le 1 \end{cases} \tag{13}$$

The relation measuring methods based on $f_{NLC}(X,Y)$ and $f_{LC}(X,Y)$ are respectively referred to as the NLC method and the LC method. Both of them calculate the weight $w(X,Y)$ based on the two kinds of co-occurrence computed by Cosine coefficient based on the hits of a search engine and by the sentence co-occurrence page rate derived from text analysis on the co-occurrence pages. The NLC method is the nonlinear coupled combination of two basic methods of Cosine coefficient and the sentence co-occurrence page rate, while the LC method is the linear coupled combination of them. Also the contributions of the two basic methods to the weight $w(X,Y)$ are determined by the parameters θ' and θ (referred to as contribution factors). The values of two contribution factors will be appropriately set by the experiments.

5 Experimental Results and Analysis

In related works, there are still no standard data sets for the experiments to evaluate the results of the relation measuring method in web social networks. Researchers usually build their own data sets for their experiments and label the data sets manually, and then evaluate the effectiveness of their methods by comparing with the labeled results [3, 17].

5.1 Data Sets

We deliberately choose names of two class as the test data. One class is names of Chinese stars with very high occurrence frequency. Google hits of these name pairs are normally about one hundred thousand. The other is names of Chinese academic researchers with relatively low occurrence frequency in web pages. Google hits of these name pairs are normally several tens.

The data sets we built include four groups of names in Chinese, as shown in Table 1. The first two groups include ten entertainment stars and ten sports stars derived from the 2011 Chinese celebrities in the Forbes china site [29]. The third and fourth groups respectively contain ten academic researchers from seven Chinese universities and research institutions, whose names are derived from the program committee of two academic conferences [30-31].

Then we constructed four test groups, as shown in Table 2. Group A and Group B, are respectively derived randomly from the first two lines and the last two lines in Table 1, while for Group C and Group D, half of the names in four groups are selected randomly from the first two lines of Table 1, and other names are derived randomly from the last two lines of Table 1. Thus, names of Group A and B are respectively in only one field, while names of Group C and D are in two fields.

Table 1. Four Groups of candidate data for the experiments

Group	Name list
1. entertainment stars	Jackie Chan, Andy Lau, Sally Wu, Lee Bing Bing, Zhang Ziyi, Zhao Benshan, Yimou Zhang, Jay Chou, Louis Liu, Jacky Wu
2. sports stars	Yao Ming, Roger Yi, Liu Xiang, Guo Jingjing, Li Na, Zheng Jie, Lin Dan, Ding Junhui, Zhang Yining, Liu Guoliang
3. Chinese academic researchers in the fields of Process and Information Retrieve	Yu Shiwen, Wan Xiaojun, Yan Hongfei, Bai Shou, Wang Bing, Shi Zhongzhi, Liu Ting, Che Wanxiang, Sun Maosong, Liang Xun
4. Chinese academic researchers in the field of Database	Tang Shiwei, Zhou Longxiang, Luo Xiaopei, Gao wen, Li Jianzhong, Zhou Lizhu, Du Xiaoyong, Meng Xiaofeng, Tang Changjie, Fan Ming

Table 2. Randomly produced test data for the experiments

Test data	Name list	Total number
Group A	Jackie Chan, Zhang Ziyi, Zhao Benshan, Yao Ming, Guo Jingjing, Zheng Jie, Ding Junhui	7
Group B	Yu Shiwen, Yan Hongfei, Bai Shou, Sun Maosong, Tang Shiwei, Luo Xiaopei, Meng Xiaofeng, Tang Changjie	8
Group C	Jackie Chan, Yimou Zhang, Yao Ming, Guo Jingjing, Zheng Jie, Lin Dan, Zhang Yining, Yu Shiwen, Wan Xiaojun, Sun Maosong, Meng Xiaofeng, Tang Changjie, Du Xiaoyong, Fan Ming	14
Group D	Zhang Ziyi, Zhao Benshan, Roger Yi, Ding Junhui, Yan Hongfei, Bai Shou, Shi Zhongzhi, Che Wanxiang, Liang Xun, Tang Shiwei, Luo Xiaopei, Gao Wen	12

5.2 Evaluation Methods

To evaluate the effectiveness of our approach, we have asked five experts, who are famous academic researchers in the field of web social network in China, to manually label the strength of each relation in each test group. Each expert assigned a score of 0-4 to each relation according to the same criteria listed in Table 3. As we believe that it should be determined by all five experts and a relatively larger score labeled means the expert knows the relation well, the sum of five scores was used as the golden standard weight for a relation.

Table 3. The labeled score and the corresponding closeness of relations

Score	Closeness of relations
0	There is no relationship between two individuals.
1	The strength of the relationship between two individuals is very weak.
2	The strength of the relationship between two individuals is middle.
3	The strength of the relationship between two individuals is strong.
4	The strength of the relationship between two individuals is very strong.

To estimate the accuracy of the relation measuring results, researchers usually used the Pearson's Correlation Coefficient (referred to as the correlation coefficient) to measure the correlation between the test results and the labeled results [3, 17]. The correlation coefficient of two random variables $\mathbb{X}$ and $\mathbb{Y}$ is calculated by the following formula:

$$\rho_{\mathbb{X}\mathbb{Y}} = \frac{Cov(\mathbb{X},\mathbb{Y})}{\sqrt{D(\mathbb{X})}\cdot\sqrt{D(\mathbb{Y})}} \quad \textbf{(14)}$$

where $Cov(\mathbb{X},\mathbb{Y})$ is the covariance of $\mathbb{X}$ and $\mathbb{Y}$, $D(\mathbb{X})$ and $D(\mathbb{Y})$ are respectively the variances of $\mathbb{X}$ and $\mathbb{Y}$. The correlation coefficient between $\mathbb{X}$ and $\mathbb{Y}$ measures their linear correlation, and when the value is larger than 0.5, the correlation of them is strong [32-34].

When we use the correlation coefficient to estimate the accuracy of results for a relation measuring method, $\mathbb{X}$ and $\mathbb{Y}$ are respectively the weights calculated by the method and the labelled one (that is the golden standard weights) of relations in the same web social networks. The stronger the correlation coefficient between them is, the closer two kinds of weights is. So the larger the correlation coefficient is, the more accurate the results of the method are.

We use Average and Mean Squared Error (MSE) to measure the performance of each method. Average reflects the correlation between the results of the measuring method on different test group and the labeled results. MSE is the expectation of the square of the differences between the correlation coefficient on each group and the Average for one method, which reflects the change degree of the correlation coefficient of each method for different test groups.

5.3 Procedures of the Experiments

To evaluate the effectiveness of the SETARM model, we must compare methods NLC and LC with existing two kind of approaches.

For the approaches based on a search engine, He et al. [22] and Canaleta et al. [23] have compared most of these approaches and the results indicate that Cosine coefficient performs best and Overlap coefficient is better than Jaccard coefficient. So we choose the better two approaches Cosine coefficient (referred to as Cosine) and Overlap coefficient (referred to as Overlap) as the compared methods in our experiments.

There are mainly three methods based on text analysis. The sentence co-occurrence page rate method is better than the other two, the sentence co-occurrence frequency method and the sentence co-occurrence page number method, which has been discussed in section 4.2. As the first one are proposed by this paper, we choose the better one of the latter two approaches, the sentence co-occurrence frequency method (referred to as CoocSentCnt), which is proposed in [16] and more relative with the basic text analysis method in NLC and LC, as the third compared method in our experiments.

The results of all five methods are very dependent on the search engine adopted in the experiments. Google is one of the most influential and reliable web search engines [19], and many studies on web social network extraction or weighting relations from web pages select Google as the adopted search engine [4, 13, 19]. So we chose Google as the only search engine used in our experiments.

We carried out experiments on the four name groups in Table 2. For each group, we measured the relation weights of each name pair by all five methods and the steps are as follows.

Step 1: Doing Google search and downloading web pages. For each name pair X and Y in the group, we used Google search API to respectively obtain the number of hit pages yielded by Google for the queries "X", "Y" and "X AND Y", and downloaded at most top 100 pages by the retrieved URLs for the query "X AND Y".

To avoid retrieving the web pages referred to a namesake rather than the test person, when searching for web pages with a name by Google, we added the keywords of his affiliation organization to the query keywords.

Step 2: Extracting texts from downloaded web pages. We extracted texts from each downloaded pages and kept the original structure of the texts at the same time by using the document object model (DOM) tree, such as ' ' and ' ' are the original flags for a block in the DOM tree, and '\r\n' is the end flag of a text paragraph.

Step 3: Deleting reduplicate web pages. To avoid pages with same contents affecting the statistical results of name co-occurrence in sentences, we identified the reduplicate pages by the text similarity of web pages and deleted them. The remaining pages were regarded as the real top ranking M pages.

Step 4: Analysis on texts of web pages. We analyzed the text of each remaining page to count the number of pages with X and Y co-occurring in sentences and the frequency of sentences in which X and Y co-occur.

We only analyze the paragraphs with names X and Y co-occurring in the texts, and detect sentences by the sentence end punctuations, such as periods, question marks and exclamation points in Chinese texts.

Step 5: Measuring the weights of relations. The weights of relations of each name pair in the group were calculated by the above five test methods based on the results of Step 1 and Step 4.

Step 6: Manually labeling the weight score of each relation. We asked five experts in the field to label the weight of each relation in the test group and computed the score sum for each relation, which is the golden standard weight of the relation.

Step 7: Calculating the correlation coefficient for each method. At last, we computed the correlation coefficient between the weights obtained by each

method and the golden standard weights, and evaluated the performance of each method.

5.4 Results and Analysis

5.4.1 Experiments for the Parameters

The methods NLC and LC both include a contribution factor, the parameters θ' and θ respectively, which determines the contributions of two basic methods on a search engine and on text analysis to relation measuring results, the relation weight of two persons. It is necessary to research which one of two basic methods is more influential to the social relation strength of two persons by co-occurrence of their names in web pages, and how to set the value of the factors. So we conducted some experiments on different values of each parameter to find the appropriate domains of them, which can ensure relatively high performance for NLC and LC.

In the experiments, the parameter θ' of the NLC method was set to nine different values in (0, 1) by the formula (13), and the results of the method NLC with different value of θ' on four groups are listed in Table 4.

Table 4. Correlation coefficients between the results of NLC with different parameters and the labeled scores

	$\theta'=0.1$	$\theta'=0.2$	$\theta'=0.3$	$\theta'=0.4$	$\theta'=0.5$	$\theta'=0.6$	$\theta'=0.7$	$\theta'=0.8$	$\theta'=0.9$
Group A	0.407162	0.464163	0.525798	0.590914	0.657212	0.720898	0.776678	0.818501	0.841161
Group B	0.368429	0.417788	0.465984	0.51042	0.548793	0.579549	0.602065	0.616487	0.623437
Group C	0.686544	0.699842	0.713286	0.726856	0.740446	0.753737	0.765836	0.774066	0.769415
Group D	0.352133	0.393511	0.435322	0.475855	0.513628	0.547587	0.577110	0.601903	0.621859
Average	0.453567	0.493826	0.535098	0.576011	0.61502	0.650443	0.680422	0.702739	0.713968
MSE	0.135986	0.121622	0.1079	0.096579	0.089691	0.088373	0.091343	0.094994	0.094779

In Table 4**,** we present the correlation coefficients between the labeled results and the weights obtained by the method NLC with nine different values of θ' on each test group. And the last two rows are respectively the Average and MSE of the correlation coefficients of the method NLC with the same values of θ' on four groups. The average of all the correlation coefficients of the method NLC with various values of θ' on different groups is 0.60.

In the experiments, the parameter θ of the LC method was set to eleven different values in [0, 1] by the formula (13), and the results on four groups are listed in Table 5.

Table 5. Correlation coefficients between the results of LC with different parameters and the labeled scores

	$\theta=0$	$\theta=0.1$	$\theta=0.2$	$\theta=0.3$	$\theta=0.4$	$\theta=0.5$
Group A	0.355192	0.440723	0.527328	0.610023	0.683509	0.7436
Group B	0.313104	0.465411	0.548714	0.589568	0.609155	0.618473
Group C	0.673296	0.710089	0.726644	0.733068	0.733479	0.734493
Group D	0.311169	0.456564	0.537342	0.58042	0.604011	0.617501
Average	0.41319	0.518197	0.585007	0.62827	0.657539	0.678517
MSE	0.151199	0.111142	0.082123	0.061447	0.053962	0.060616

	$\theta=0.6$	$\theta=0.7$	$\theta=0.8$	$\theta=0.9$	$\theta=1.0$
Group A	0.78833	0.818132	0.835145	0.842223	0.842162
Group B	0.62273	0.624416	0.624764	0.624405	0.623682
Group C	0.73131	0.728644	0.72582	0.723011	0.720303
Group D	0.625543	0.630497	0.633627	0.636927	0.635633
Average	0.691978	0.700422	0.704839	0.706642	0.705445
MSE	0.070781	0.079559	0.085004	0.086996	0.087280

In Table 5, we present the correlation coefficients between the labeled results and the weights obtained by LC with ten different values of θ on each test group. And the last two rows are respectively the Average and MSE of the correlation coefficients of LC with the same values of θ on four groups. The average of all the correlation coefficients of LC with various values of θ on different groups is 0.64.

From Table 4 and Table 5, we can see that both correlation coefficients of NLC and LC on each group are increasing with the parameters' value growing. Also, the MSE are relatively small, when the parameters of NLC or LC are close to 0.5. That is to say, when the contribution factors of two basic methods in the weight of a relation is nearly equal, results of NLC and LC are the most stable.

To get a relatively high performance, we can take the average performance of a method in the correlation coefficient as the benchmark. If the average correlation coefficients of a method with a certain value of the

parameter on each group can simultaneously reach or exceed the average performance of the method with different values of the parameter, we believe that the method with this value of the parameter tends to get relatively high performance in other data sets. By this basic opinion, according to the four correlation coefficients of NLC and LC with a certain value of the parameter on four groups in Table 4 and Table 5, the parameter θ' and θ should be set as $\theta' \geq 0.5$ and $\theta \geq 0.4$.

5.4.2 Performances of Different Methods

To compare the performance of different methods, we first calculate the average correlation coefficients of NLC with various values of parameter $\theta' \geq 0.5$ on each group, and the average correlation coefficients of LC with various values of parameter $\theta \geq 0.4$ on each group. The results are presented on the right two columns in Table 6.

Table 6. Experimental results of different methods

	Overlap	Cosine	CoocSentCnt	NLC($\theta' \geq 0.5$)	LC($\theta \geq 0.4$)
Group A	0.146361	0.355192	0.891162	0.76289	0.7933
Group B	0.32405	0.313104	0.572483	0.594066	0.621089
Group C	0.656903	0.673296	0.671954	0.7607	0.728151
Group D	0.37003	0.311169	0.416131	0.572417	0.626248
Average	0.374336	0.41319	0.637933	0.672518	0.692197
MSE	0.183278	0.151199	0.172309	0.089607	0.072319

In Table 6, in the left three columns are the correlation coefficients of the method on each group. The last two rows are respectively the Average and MSE of the correlation coefficients of each method on four groups. The comparison chart of the Average and MSE of each method on four test groups is shown in Figure 1.

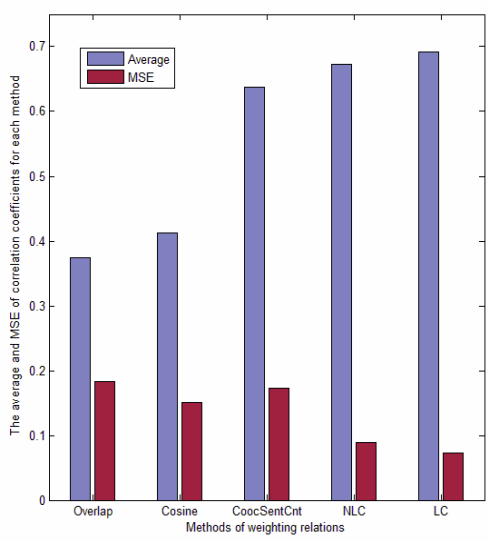


Figure 1. The comparison chart of the performances of five methods

According to Averages in Figure 1, the performances of NLC, LC and CoocSentCnt, the methods based on text analysis, are obviously better than the performances of Overlap and Cosine, the methods based on a search engine. Moreover, the performances of NLC and LC are better than that of CoocSentCnt. That is to say, the results of NLC and LC are more accurate than other methods. According to MSEs in Figure 1, MSEs of NLC and LC are less than 0.1. It is obviously better than those of other methods which are more than 0.15 or even approximately 0.2. That is to say, the stabilities of NLC and LC are better than other methods.

In conclusion, the results of the instance methods on the SETARM model are more accurate and more stable than those of the methods based on a search engine or text analysis.

5.4.3 Performance Analysis

The SETARM model measures the strengths of relations in web social networks according to both the weights calculated by the basic relation measuring method based on a search engine and the weights calculated by the basic relation measuring method based on text analysis. We analyzed that how either of these two basic methods and the way to combine the weights of two methods influence the performance of the SETARM model.

The formula (8) and formula (9) show that both the parameters θ' and θ are the contribution factors of the basic method $S_2(X,Y)$ (referred to as S_2) for calculating the weight $w(X,Y)$ by the function f, and the higher the parameters θ' and θ are, the larger the contribution of the measuring results of S_2 is. The trends of the correlation coefficients of NLC and LC for four test groups changing with the varying of the parameters θ' and θ are respectively drawn in Figure 2 and Figure 3. The trends of the average correlation coefficients of NLC and LC for different test groups changing with the contribution factors of S_2 varying are drawn in Figure 4. To draw the whole changing trend of the correlations coefficients of NLC when the parameter $\theta' \in (0,1)$, we add several special points of the correlation coefficients of NLC at $\theta' = 0$ and $\theta' = 1$ in Figure 2 and Figure 4.

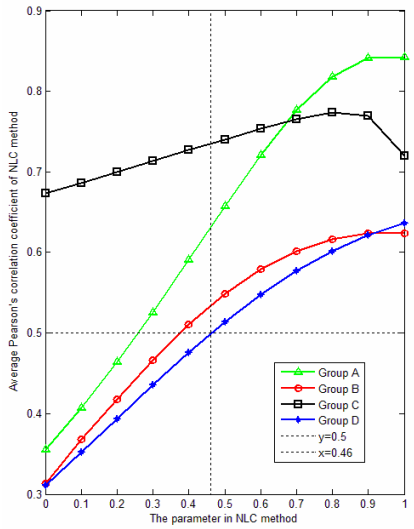


Figure 2. Effect of the parameter θ' on the correlation coefficients of the NLC method

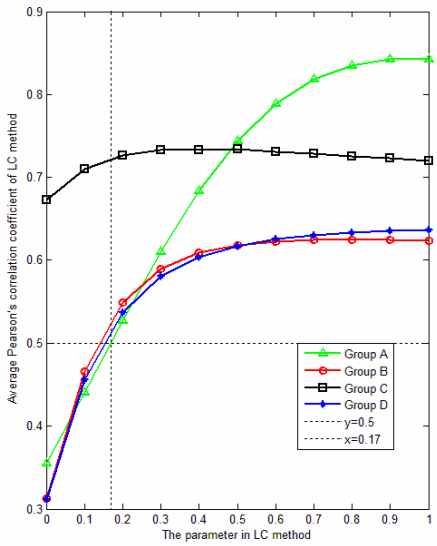


Figure 3. Effect of the parameter θ on the correlation coefficients of the LC method

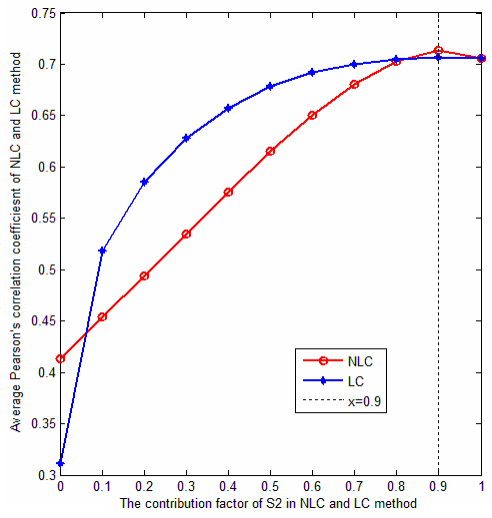


Figure 4. Effect of the contribution factor of S_2 on the average correlation coefficients of NLC and LC

By Figure 2 and Figure 3, the changing the correlation coefficients of NLC and LC are not strictly synchronous for different test groups. However, the overall changing trends of these curves are consistent with each other. That is the correlation coefficients of NLC and LC for each test group are gradually increasing with the values of the parameters θ' and θ growing. Also by the average correlation coefficients of NLC and LC for different test data groups drawn in Figure 4, when the contribution factors of S_2, θ' and θ, are about 0.9, the average correlation coefficients of NLC and LC grows slowly and even become decreasing. These trends show that when constructing relation measuring methods based on the SETARM model, the greater the contribution factor of the weights of the basic method LC is, the more accurate the results of the measuring methods are, and the performance may reach the best when the contribution factor is approximately 0.9.

According to the concept of the correlation coefficient in Section 4.2, when the correlation coefficient is larger than 0.6, the correlation between the results of the method and the labeled results is strong and the results of the method is relatively accurate. By Figure 2 and Figure 3, the correlation coefficients of NLC for all the four test groups achieve 0.5 until the parameter θ' arrives at 0.46, while the correlation coefficients of LC for all the four test groups have reached 0.5 when the parameter θ does not arrive at 0.2. It means that compared to the way of the nonlinear coupled combination of two basic methods, the way of the linear coupled combination of two basic methods can make the integrated method get relatively more accurate results when the contribution factor of the basic method S_2 is smaller for different test groups. That is to say the basic method S_2 can play a better role for the integrated method in the way of the linear coupled combination of two basic methods. Moreover, the Average of the correlation coefficient of LC is larger than that of NLC and the MSE of the correlation coefficient of LC is lower than that of NLC by the average performance shown in Figure 1. It shows that the weights calculated by the integrated method in the way of the linear coupled combination of two basic methods are more accurate and more stable.

In summary, when constructing the instance of the method to measure relations in the Web based on the SETARM model, we should integrate the basic method based on a search engine and the basic method based on text analysis in the way of the linear coupled combination, and the latter one should play a more important role in the measured results.

6 Conclusion

Web pages are one sort of the various data sources on web to measure social relations. This paper presents

a general model of measuring social relations on web pages, named SETARM. From the experiments on the specific methods on the model, we can draw the following conclusions. First, the methods on SETARM can measure the social relation weights between persons, whose names appear in the web at a certain frequency, whether they are famous or not. Second, the relation weights calculated by the methods on SETAMR are more accurate and stable than existing typical methods only based on search engines or text analysis. At last, when measuring social relations by SETARM, the kind of basic methods on search engines and that on text analysis should be integrated in the linear way, and the latter should play a more important role on the relation weights.

The SETARM model can be used to extract web social networks of various individuals and provide more accurate information of relations for a variety of applications on web social networks. Moreover, it can be used to measure the strength of relationships among other types of entities, which is our future work.

Acknowledgement

Supported by the National Natural Science Foundation of China (NSFC) as the project ID of 61309007.

References

[1] H. Kautz, B. Selman, M. Shah, The Hidden Web, *AI Magazine*, Vol. 18, No. 2, pp. 27-36, 1997.

[2] Y. W. Zhao, W.-J. van den Heuvel, X. Ye, A Framework for Multi-Faceted Analytics of User Behaviors in Social Networks, *Journal of Internet Technology*, Vol. 15, No. 6, pp. 985-994, November, 2014.

[3] M. Oka, Y. Matsuo, Measuring the Weight of Relations Between Entities, *Third International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web*, Washington, DC, 2009.

[4] Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, M. Ishizuka, POLYPHONET: An Advanced Social Network Extraction System from the Web, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 5, No. 4, pp. 262-278, December, 2007.

[5] K. Morita, T. Ogawa, H. Kitagawa, M. Fuketa, J.-I. Aoe, A Method of Extraction and Visualisation for Relationships among Objects on Web, *IJISTA*, Vol. 12, No. 3/4, pp. 316-327, September-December, 2013.

[6] N. Takhirov, F. Duchateau, T. Aalberg, I. Sølvberg, KIEV: A Tool for Extracting Semantic Relations from the World Wide Web, *International Conference on Extending Database Technology (EDBT)*, Athens, Grèce, 2014, pp. 632-635.

[7] M. Nasution, S. A. Noah, *A Methodology to Extract Social Network from the Web Snippet*, CoRR abs/1211.5877, 2012.

[8] J. R. Tyler, D. M. Wilkinson, B. A. Huberman, Email as Spectroscopy: Automated Discovery of Community Structure within Organizations, *International Conference on Communities and Technologies*, Amsterdam, Netherlands, 2003, pp. 81-96.

[9] M. Laclavik, S. Dlugolinsky, M. Kvassay, L. Hluchý, Email Social Network Extraction and Search, *Web Intelligence/IAT Workshops*, Lyon, France, 2011, pp. 373-376.

[10] T. Arif, R. Ali, M. Asger, Scientific Co-authorship Social Networks: A Case Study of Computer Science Scenario in India, *International Journal of Computer Applications*, Vol. 52, No. 12, pp. 38-45, August, 2012.

[11] T. Miki, S. Nomura, T. Ishida, Semantic Web Link Analysis to Discover Social Relationships in Academic Communities, *Proceedings of the 2005 Symposium on Applications and the Internet (SAINT'05)*, Trento, Italy, 2005, pp. 38-45.

[12] R. Soussi, E. Cuvelier, M.-A. Aufaure, A. Louati, Y. Lechevallier, DB2SNA: An All-in-One Tool for Extraction and Aggregation of Underlying Social Networks from Relational Databases, T. Özyer, J. Rokne, G. Wagner, A. H. P. Reuser (Eds.), *The Influence of Technology on Social Network Analysis and Mining,* Springer, 2013, pp. 521-545.

[13] H. Kautz, B. Selman, M. Shah, Referral Web: Combining Social Networks and Collaborative Filtering, *Communications of the ACM*, Vol. 40, No. 3, pp. 63-65, March, 1997.

[14] P. Mika, Flink: Semantic Web Technology for the Extraction and Analysis of social Networks, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 3, No. 2-3, pp. 211-223, October, 2005.

[15] N. Di, C.-L. Yao, X.-M. Li, Extraction, Measurement and Analysis of Social Network in Chinese Web, *Journal of Guangxi Normal University: Natural Science Edition*, Vol. 25, No. 2, pp. 169-172, 2007.

[16] X. Li, B. Liu, P. S. Yu, Mining Community Structure of Named Entities from Web Pages and Blogs, *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, Stanford, CA, 2006, pp. 108-114.

[17] H.-H. Chen, M.-S. Lin, Y.-C. Wei, Novel Association Measures Using Web Search with Double Checking, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 1009-1016.

[18] C. D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[19] S. H. Lee, P.-J. Kim, Y.-Y. Ahn, H. Jeong, Googling Social Interactions: Web Search Engine Based Social Network Construction, *PLoS One*, Vol. 5, No. 7, p. e11233, July, 2010.

[20] Y. Matsuo, H. Tomobe, K. Hasida, M. Ishizuka, Finding Social Network for Trust Calculation, *Proc. 16th European Conference on Artificial Intelligence*, Valencia, Spain, 2004, pp. 510-514.

[21] Y. Matsuo, H. Tomobe, K. Hasida, H. Nakashima, M. Ishizuka, Social Network Extraction from the Web Information, *Japanese Society for Artificial Intelligence*, Vol. 20, No. 1E, pp. 46-56, January, 2005.

[22] J. He, Y. Liu, Q. Tu, C. Yao, N. Di, Efficient Entity Relation Discovery on Web, *Journal of Computational Information*

Systems, Vol. 3, No. 2, pp. 203-213, January, 2007.

[23] X. Canaleta, P. Ros, A. Vallejo, D. Vernet, A. Zaballos, A System to Extract Social Networks Based on the Processing of Information Obtained from Internet, *Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence*, Sant Martí d'Empúries, Spain, 2008, pp. 283-292.

[24] Y. Z. Jin, Y. Matsuo, M. Ishizuka, Extracting Social Networks among Various Entities on the Web, *Proceeding of the 4th European Semantic Web Conference* (ESWC2007), Innsbruck, Austria, 2007, pp. 251-256.

[25] Y. Z. Jin, Y. Matsuo, M. Ishizuka, Extracting a Social Network among Entities by Web Mining, *ISWC'06 Workshop on Web Content Mining with Human Language Technologies*, Berlin, Germany, 2006, pp. 251-266.

[26] Y. Matsuo, K. Hasida, H. Tomobe, M. Ishizuka, Mining Social Network of Conference Participants from the Web, *Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)*, Halifax, NS, Canada, 2003, pp. 190-193.

[27] Y. Matsuo, H. Tomobe, T. Nishimura, Robust Estimation of Google Counts for Social Network Extraction, *AAAI'07 Proceedings of the 22nd National Conference on Artificial Intelligence*, Vancouver, British Columbia, Canada, 2007, pp. 1395-1401.

[28] The 2011 Chinese Celebrities in the Forbeschina Site, http://www.forbeschina.com/review/201105/ 0009378.shtml.

[29] M. Oka, Y. Matsuo, Weighting Relations in Social Networks Using the Web, *The 23rd Annual Conference of the Japanese Society for Artificial Intelligence*, Kagawa, Japan, 2009, pp. 1-2.

[30] The Program Committee of the 7th National Information Retrieval Conference of China (CCIR2011), http://ir.sdu.edu.cn/ccir2011/organization.htm.

[31] The Program Committee of the 27th National Database Conference of China (NDBC2010), http://ndbc2010.ruc.edu.cn/committee.html.

[32] EXPLORABLE, *Statistical Correlation*, http://www.experiment-resources.com/statistical-correlation.html.

[33] E. Garcia, *A Tutorial on Correlation Coefficients*, http://www.miislita.com/statistics/on-the-non-additivity-correlation-coefficients.pdf.

[34] T. J. Cleophas, A. H. Zwinderman, T. F. Cleophas, E. P. Cleophas, *Statistics Applied to Clinical Studies*, Springer Verlag, 2012.

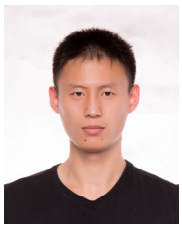
Biographies

Meijuan Yin was born in Anhui Province, China at Nov. 1977. She was conferred a M.Sc. in computer science by Zhengzhou Information Science and Technology Institute at Zhengzhou, China, in 2003. She is working on the Ph.D. in computer software and academic of the same university. After graduating from the university, she became an assistant of State Key Laboratory of Mathematics Engineering and Advanced Computing of the university in 2003 and turned to a lecturer in 2005. Her current research interests include data mining, social network analysis, and information security. Ms. Yin joined China Computer Federation (CCF) as a common member in 2007 and received the IEEE membership in 2010.

Xiaonan Liu was born in Liaoning Province, China. He was conferred a M.Sc. in computer science by Zhengzhou Information Science and Technology Institute at Zhengzhou, China, in 2006. He is working on the Ph.D. in computer software and academic of the same university. He graduated from the university, and became an assistant in 2000 and turned to a lecturer in 2006. Now, he is an associate professor of State Key Laboratory of Mathematics Engineering and Advanced Computing of the university. His research interests include binary translation, compile, and decompile.

Junyong Luo received a M.Sc. in computer science and engineering from Zhengzhou Information Science and Technology Institute at Zhengzhou, China, and became a teacher of the university in 1992. He was developed into a professor and doctoral supervisor of computer science and engineering in 2005. His research has covered many areas, including database, network security, data mining, and information security. His current research projects are on knowledge discovering, social network analysis and parallel computing.

Yan Liu was born in Shandong Province, China. She is working on the Ph.D. in computer software and academic of Zhengzhou Information Science and Technology Institute. After graduating from the university, she became an associate professor of State Key Laboratory of Mathematics Engineering and Advanced Computing of the university in 2012. Her current research interests include data mining and information security.



Ziqi Tang was born in Jiangxi Province, China, in 1994. He received his B.S. degree in network engineering from Zhengzhou Information Science and Technology Institute in 2015. He is a graduate student in the same university now. His research interest includes web mining, data analyzing and information retrieve.