UR Rank: Micro-blog User Influence Ranking Algorithm Based on User Relationship

Wenbin Yao[1,2], Yiwei Yang[1,2(✉)], and Dongbin Wang[3]

[1] Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China
yiwei_yang@163.com
[2] National Engineering Laboratory for Mobile Network Security, Beijing University of Posts and Telecommunications, Beijing, China
[3] Key Laboratory of Ministry of Education for Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. In this paper, a novel UR Rank (User Relationships based Ranking) algorithm is proposed for ranking the influence of the user. We first explore five factors that affect user relationship. They are following rate (*FR*) factor, activity (*ACT*) factor, authority (*ATR*) factor, interaction (*ITA*) factor and similarity (*SML*) factor. Then those factors are used in Support Vector Regression (SVR) model to predict the relationship between users. We assimilate such predicted relationship into a PageRank based transition probability to identify influential users. The experiments on a real micro-blog data set demonstrate that UR Rank algorithm has better performance and is more persuasive than the existing algorithms.

Keywords: Micro-blog · User influence · Support vector regression PageRank

1 Introduction

User influence is defined as the interaction of an individual in ideas, feelings, attitudes, or behavior with other individuals or groups [1]. It is of great significance to identify influential users in the aspect of situation awareness, product promotion [2], expert recommendation [3] and public opinion guidance.

Most current user influence researches are based on the PageRank [4] algorithm. Some studies concentrate on incorporate temporal factors into PageRank algorithm [6, 7]. Bartoletti et al. [6] compute user reputation by using an arbitrary ranking algorithm in the most recent time window, and combine it with a summary of historical data. Hu et al. [7] studied temporal dimension in assessing the authority of nodes by adopting three temporal factors. However, they did not consider the difference between users, which has led to the weakness in practical application.

Recently, several researches such as [8–10] have applied interaction relationship of users to the ranking model. Ma et al. [8] focus on user behavioral characteristics and

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 394–404, 2018.
https://doi.org/10.1007/978-3-030-00916-8_37

predict the probability that user will respond using logistic regression (*LR*). However, when the data dimension is high, the algorithm of *LR* is not very applicable. Weng et al. [9] measure the influence taking both the topical similarity between users and the link structure into account and proposed Twitterrank algorithm. However, they only consider the topic similarity between users, but ignore other factors that influence the user relationship. Ding et al. [10], consider the four relationships between users: repost, reply, copy and read between users and measure the influence of users by random walks of multi-relation data in micro-blog. However, the user relationship changes over time, so the pure formula definition of the user's four relations in this paper is unreasonable.

In this paper, we focus on the prediction of user relationships, which can be regarded as a preprocessing step for mining influential users. Then we apply the predicted relationship to the weight division of the PageRank algorithm to avoid the interference of inactive users to the ranking.

On analyzing the user influence, we organized our studies in tree questions: what factors are closely related to the relationship between users, how to predict the relationship between users and how to apply the user relationship to the influential user mining. The highlights of our work can be summarized as follows:

- We take the probability of comment and forward behavior as the index to measure the relationship between users and explore five factors. Besides, we describe the relationship between each factor and index by real data.
- Using the mining factors, we fit a regression model to predict the intensity of user relationship. We use a machine learning technique called support vector regression. Support vector machine are complex machine learning models which are better suited for data which has more complex patterns than linear regression can handle.
- By using the predicted user relationship as the basis of the weight distribution of the influence value, we improve the PageRank to avoid the interference of inactive user to the ranking.

The rest of this paper is organized as follows. In Sect. 2, we show the UR Rank algorithm. The experiment result is shown in Sect. 3. Finally, we make a conclusion and present some future researches in Sect. 4.

2 Influence Ranking Based on User Relationship

2.1 Factors Mining

The selection of factors is the basis of the prediction between users. The user relationship is affected by the superior user's attributes and the interactive behavior between users. In this section, we explore the impact of the following rate (*FR*) factor, activity (*ACT*) factor, authority (*ATR*) factor, interaction (*ITA*) factor, similarity (*SML*) factor on the user relationship. The user relationship (*UR*) is expressed as the average of the comment and forward probability. For users (v, u), user v followed u and the factors that affect their relationship are described below.

(1) *Following Rate (FR): FR* is defined as the popularity of user u, that is the ratio of the followers number ($FoNum_u$) to the friends number ($FrNum_u$), as shown in formula (1). The higher of the following rate, the higher the probability of other users to comment and forward the user's information.

$$FR = \frac{FoNum_u}{FrNum_u} \tag{1}$$

(2) *Activity (ACT): ACT* is defined as the tweet frequency of the u, that is the ratio of tweet number ($TNum_u$) to the time (ΔT), as shown in formula (2). The higher of the *ACT*, the greater the probability that the information will be read by his followers. Therefore, *ACT* gives us an idea of the likelihood of user's tweet being spread.

$$ACT = \frac{TNum_u}{\Delta T} \tag{2}$$

(3) *Authority (ATR): ART* is represented by user's level. Micro-blog user's level is divided into 48 levels, the higher the user's level, the greater of his authority. The authority degree of (v, u) is divided into 48 intervals from 1 to 48, and the average *UR* value is taken as the dependent variable.

(4) *Interaction (ITA): ITA* is defined as the ratio of number of v forwarded u (Fw_{uv}) to the total number of v forwarded (Fw_v), as shown in formula (3). The interaction intensity of (v, u) is divided into 20 intervals from 0 to 1, and the average *UR* value is taken as the dependent variable.

$$ITA = \frac{Fw_{uv}}{Fw_v} \tag{3}$$

(5) *Similarity (SML): SML* is defined as the *Jaccard* of the two user's tag, as shown in formula (4). $note_u$ is the tags set of u and $note_v$ is the tags set of v. The similarity degree of (v, u) is divided into 20 intervals from 0 to 1, and the average *UR* value is taken as the dependent variable.

$$SML = \frac{note_u \cap note_v}{note_u \cup note_v} \tag{4}$$

2.2 User Relationship Prediction Model

To evaluate the usefulness of the five mining factors, we first propose the support vector regression model. SVR (support vector regression) is a regression model based on SVM (support vector machine). The goal of SVR is to minimize the prediction error. It makes as many samples as possible on the optimal regression hyperplane, but it is not required to be absolute in the hyperplane but the distance from the hyperplane is small enough. Therefore, the loss function is introduced and the definition of the loss function is shown in formula (5).

$$Err\left(UR_{vu}, UR'_{vu}\right) = \begin{cases} 0, & \left|UR'_{vu} - w\phi(UR_{vu}) + b\right| \leq \varepsilon \\ \left|UR'_{vu} - w\phi(UR_{vu}) + b\right| - \varepsilon, & \left|UR'_{vu} - w\phi(UR_{vu}) + b\right| > \varepsilon \end{cases} \tag{5}$$

where UR_{vu} is the value of the optimal hyperplane, and the value UR'_{vu} is the predicted user relationship of (v, u). The constant ε represents a sufficiently small loss value. It controls the fitting degree of the function and the training sample. If all the samples are in the pipeline with a diameter of 2ε, the error of the algorithm is 0.

In order to increase the fault tolerance of the algorithm, we need to set a soft edge. Therefore, we added penalty factor C and slack variable ξ. The final SVR model is shown in formula (6).

$$\min \frac{1}{2}\|w\|^2 + C\sum\nolimits_{i=1}^{M} \xi_{vu} + \xi^*_{vu} \quad s.t. \begin{cases} UR'_{vu} - w \cdot UR_{vu} - b \leq \varepsilon + \xi_{vu} \\ w \cdot UR_{vu} - UR'_{vu} + b \leq \varepsilon + \xi^*_{vu} \\ \xi_{vu}, \xi^*_{vu} \geq 0 \end{cases} \tag{6}$$

2.3 UR Rank

Given a directed graph $G = (V, E)$, V is the set of users and E is the set of links. For each directed edge $e = (v, u) \in E$, the direction of the arrow indicates the direction of the following. Let PR_v be the PageRank value of v and $w(v, u)$ be the proportion of importance propagated from v to u. In PageRank algorithm, $w(v, u)$ is normally set to $1/d^o(v)$, where $d^o(v)$ is the friends number of v in G. If u has not follow any others, he will assimilate all the PageRank value of others. Therefore, to avoid the trap problem, it introduces the jump factor $1 - \alpha$. Besides, $d^i(u)$ is the follower number of u, and N is the number of users in G. In the $t + 1$ interaction, the PageRank value of u is shown in formula (7).

$$PR_u^{(t+1)} = \alpha \sum_{v \in d^i(u)} \frac{PR_v^t}{d^o(v)} + \frac{1-\alpha}{N} \tag{7}$$

In this section, we describe a new ranking function that incorporates the five factors into PageRank. In the PageRank $w(v, u)$ is weight for a transition form v to u, and is set to $1/|d^o(v)|$. In our UR Rank, the main purpose is to predict the user relationship by SVR model and investigate the contribution of user relationship to node ranking by weighting in node transitions. The UR Rank can be rewritten as follows, where R_v^t is the influence value of v in t interaction and UR_{vu} is the predicted relationship of (v, u).

$$R_u^{(t+1)} = \alpha \sum_{v \in d^i(u)} \frac{UR_{vu}}{\sum_{u' \in d^o(v)} UR_{vu'}} R_v^t + \frac{1-\alpha}{N} \tag{8}$$

According to the formula (8), the UR_{vu} is not only relevant to the degree of v but also related to the relationship degree of u and its directed neighbors. Given two users u and v, with u having posted a tweet, the task of UR_{vu} prediction is to predict the probability that v will comment and forward this tweet. Using the mining five factors we fit a SVR model, it is better suited for data which has more complex patterns. UR Rank is summarized in Algorithm 1, where F_{vu} is the factor set of v and u, ε is a small constant.

Algorithm 1. Influence Rank Based User Relationship (UR Rank)

Input: Social graph $G(V, E)$, F_{vu}, $(v, u) \in E$
Output: The influential user list $L = (R, V)$
1: init each $R_v = 1$, t=1, ε
2: **for** every user $u \in G$ **do**
3: **for** every edge $(v, u) \in d^i(u)$ **do**
4: $UR_{vu} = SVR(F_{vu})$
5: **end for**
6: **end for**
7: **while**($\exists u$, $|R_u{}^t\text{-}R_u{}^{t-1}| > \varepsilon$)
8: **for** every user $u \in G$ **do**
9: $R^{(t+1)}{}_u = \alpha \sum_{v \in d^i(u)} \frac{UR_{vu}}{\sum_{u' \in d^o(v)} UR_{vu'}} R^t{}_v + \frac{1-\alpha}{N}$
10: **end for**
11: $t = t + 1$
12: **end while**
13: $L = rank(R)$
13: **return** L

3 Experiments and Discussion

In the section, we conduct plentiful experimental work to compare our proposed algorithm UR Rank with existing ranking algorithms. We first describe the data set and data characteristics. Then we predict the *UR* (user relationships) and evaluate the performance of support vector regression. Finally, we propose the evaluation index and prove the effectiveness of our proposed algorithm.

3.1 Dataset

To evaluate the proposed approach, we collected micro-blog data from Sina Weibo [13]. It contains five parts: user information, following relation, tweet information, forward relation and comment relation. In order to verify the two aspects of user relationship prediction and influence ranking, we need to divide the whole data set.

The main principle of segmentation is to make the relationship between users to focus as much as possible. So we divide the data according to the time sequence of user registration. The amount of data is shown in Table 1, Set1 is the training set, Set2 is the test set, and the Set3 is the prediction and user ranking set.

Table 1. Data set list.

Datasets	#User	#Follow	#Tweet	#Forward	#comment
Set1	14605	320095	19358	86384	90157
Set2	7513	178571	9978	45465	48532
Set3	41524	893053	57234	230286	241578

3.2 User Relationship Prediction

Based on Set1, we apply scatter diagram and bar graph to show the influence of single factor on user relationship. The *UR* is defined as formula (9), where $P_{c,vu}$ is the probability of v comments u and $P_{f,vu}$ is the probability of v forwards u. The details are shown in Fig. 1.

$$UR = \frac{P_{c,vu} + P_{f,vu}}{2} \tag{9}$$

As shown in Fig. 1, the five factor are positively related to *UR*. Although *FR*, *ACT*, and *ATR* are the attributes of u, they are also related to the relationship of (v, u). From Fig. 1(a), it is found that when $Log(FR) > 0$, the trend of scatter is quite consistent. In Fig. 1(b), the points are relatively scattered, and when $Log(ACT) > 5$ the correlation is obvious. Besides, the interaction factors *ITA* and *SML* are divided into 20 intervals from 0 to 1, they will directly affect the user relationship (UR). As shown in Fig. 1(d) and (e) they are also positively related to *UR*. However, we can see from (e), when the *SML* is close to zero, the *UR* is higher. The reason is that only a small number of labels appear on the platform several times, and the number of them is shown in the (f). In the process of computing the similarity, the labels of some users are not common, so the tendency of interest is not obtained. However, this unusual label may have some relevance. In other words, the computation of interest similarity is 0, and there may be some similarity between them.

When predicting user relationship (*UR*), we select logistic regression and support vector regression models. Figure 2 shows the relationship between the real values and the predicted values of the two models on the same data set (Set2). The (a) is the result of logistic regression and (b) is the result of support vector regression model Besides, the horizontal axis is the true *UR* and the vertical axis is the predict *UR*. The closer the regression forecast result is to the real value, the more accurate the prediction result is. From the comparison results we can see that the horizontal axis in the range of [0, 0.4], map (b) of the prediction points are closer to the real values than map (a). When the horizontal axis in the range of [0.4, 1], the prediction errors of the two models are greater. On the whole, the prediction error of the logistic regression model is 35.1%,

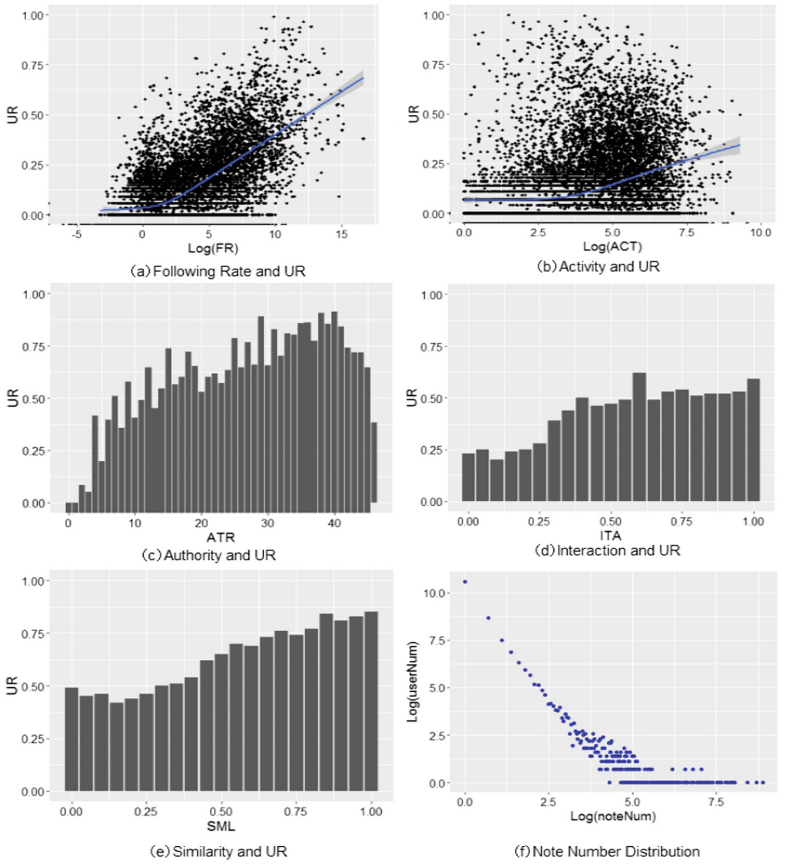

Fig. 1. The influence of five factors on user relationship (*UR*).

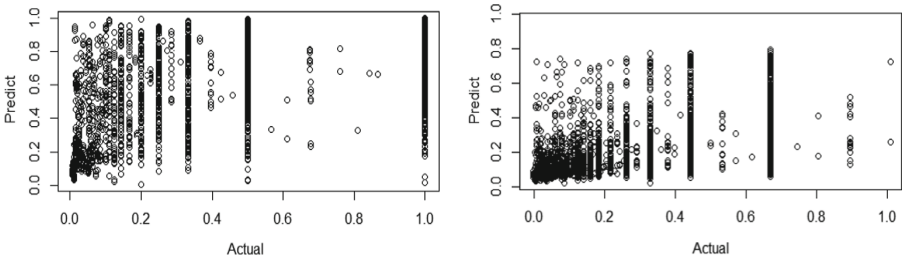

Fig. 2. Comparison of predicted and real values.

and the prediction error of support vector regression is 22.7%. The support vector regression model is more suitable for this problem.

3.3 Ranking Result

To better evaluate UR Rank, we conducted two different types of experimented on data Set3 to contrast it with existing PageRank and Twitterrank. First we compare their similarities in global ranking. Then we evaluate their performance in information dissemination.

(1) *Global Ranking:* The influence ranking list obtained from UR Rank is a relationship based influence. To compare the similarity of our algorithm with existing ranking algorithms, we calculate the commonly used Kendall τ [15] rank correlation coefficient. It determines the rank consistency of two lists containing the same users.
Our algorithm (UR Rank) was evaluated against PageRank and Twitterrank. As shown in Table 2, the UR Rank is more similar to Twitterank than to PageRank, because Twitterrank take topical similarities into consideration. This result shows that UR Rank has a certain fluctuation compared with the other two ranking algorithms, but the overall ranking list is credible.

Table 2. Kendall τ rank coefficient.

Pairs	τ
UR Rank and PageRank	0.52
UR Rank and Twitterrank	0.67

(2) *Information Dissemination:* We further investigate the performance of UR Rank, PageRank and Twitterrank with respect to information dissemination. In this section, we propose four indexes: forward number, comment number, approver number and topic number. For the top 10 users, the mean values of the 4 indexes involved in the tree algorithms are listed in Table 3.

From the statistical results of the four indexes in Table 3, it can be seen that the average numbers of the Top-10 users in UR Rank are higher than those of other two Top-10 users. More specifically, we compare the cumulative amount of the four indexes in three algorithms from Top-1 to Top-100. The results are shown in Fig. 3.

Table 3. Top-10 mean values of the four indexes.

Algorithm	Forward mean	Comment mean	Approver mean	Topic mean
UR Rank	77619.4	25713.4	31026	4.2
PageRank	43376.4	9057	17174.8	3
Twitterrank	45018.3	13224.1	17474.2	3.3

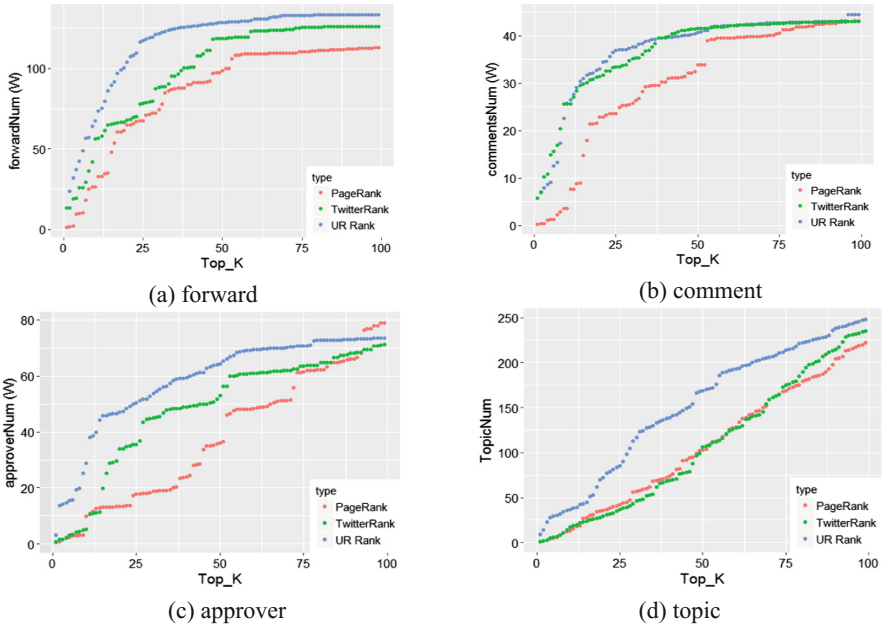

Fig. 3. The cumulative amount comparison of the four indexes in three algorithms from Top-1 to Top-100, ((a) forward cumulative number; (b) comment cumulative number; (c) approver cumulative number; (d) topic cumulative number).

It is obvious in Fig. 3 that (1) the cumulative values of the Top-k in UR Rank are higher than other two algorithms and this is because when calculating user relations, we use the probability of forwarding and comment as the indexes, and (2) when k is small, our proposed method is clearly superior to other two ranking algorithms. This shows that the top ranked users in our ranking list perform better in information dissemination, and (3) The result of our algorithm is much closer to the result of Twitterrank because Twitterrank considers the topic relevance between users. However, in figure (d) we can see that the total number of topics involved by Top-k users in our ranking list is significantly higher than that of the other two algorithms. Therefore, this proves that the influence of Top-k users in our list is not limited by the topic, and they can influence other users from more topics.

In summary, it is effective to consider the user relationship when ranking the influential users and using the support vector regression to predict the user relationship. In general, user relations changing over time, and are not satisfied with pure formula calculation. Therefore, it is reasonable to predict the user's relationship using the recent data of the user.

4 Conclusion and Future Work

This paper introduces user attributes into the influence ranking and proposes the relationship-based influence ranking algorithm. To predict the user relationship, we first mining the five influencing factors and analyze the correlation of them. Then we predict the user relationship using support vector regression model and calculate the weight of PageRank value transmission. Finally, the experiment results on three data set prove that the proposed UR Rank algorithm performs best in the influence spread and they are stable for Top-k users. In the future, we will future explore the factors importance analysis and temporal user relationship prediction.

Acknowledgements. This work was partly supported by the NSFC-Guangdong Joint Found (U1501254) and the Co-construction Program with the Beijing Municipal Commission of Education and the Ministry of Science and Technology of China (2012BAH45B01) and National key research and development program (2016YFB0800302) the Director's Project Fund of Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education (Grant No. 2017ZR01) and the Fundamental Research Funds for the Central Universities (BUPT2011RCZJ16, 2014ZD03-03) and China Information Security Special Fund (NDRC).

References

1. Rashotte, L.: In blackwell encyclopedia of sociology. J. Demogr. **16**(4), 208–210 (2007)
2. Kempe, D., Kleinberg, J.M., Tardos, E.: Maximizing the spread of influence through a social network. In: 9th International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)
3. Song, X., Tseng, B.L., Lin, C.Y., et al.: Personalized recommendation driven by information flow. In: 29th International Conference of the ACM SIGIR on Research and Development in Information Retrieval, pp 509–516. ACM, New York (2006)
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical Report, Stanford InfoLab, November 1999. http://ilpubs.stanford.edu:8090/422
5. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
6. Bartoletti, M., Lande, S., Massa, A.: Faderank: an incremental algorithm for ranking Twitter users. In: Cellary, W., Mokbel, M.F., Wang, J., Wang, H., Zhou, R., Zhang, Y. (eds.) WISE 2016. LNCS, vol. 10042, pp. 55–69. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48743-4_5
7. Hu, W., Zou, H., Gong, Z.: Temporal PageRank on social networks. In: Wang, J., et al. (eds.) WISE 2015. LNCS, vol. 9418, pp. 262–276. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26190-4_18
8. Ma, X., Li, C., Bailey, J., Wijewickrema, S.: Finding influentials in Twitter: a temporal influence ranking model. arXiv preprint arXiv:1703.01468 (2017)
9. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential Twitterers. In: Third International Conference on Web Search and Data Mining, pp. 261–270. ACM (2010)
10. Ding, Z., Zhou, B., Jia, Y., et al.: Topical influence analysis based on the multi-relational network in microblogs. J. Comput. Res. Dev. **50**(10), 2155–2175 (2013)

11. Velissarios, Z., Andreas, K., Christos, M.: Real time analytics for measuring user influence on Twitter. In: 27th International Conference on Tools with Artificial Intelligence, pp. 591–597. IEEE (2015)
12. Zhang, J.X., Zhang, R.I., Sun, J.C., et al.: TrueTop: a sybil-resilient system for user influence measurement on Twitter. IEEE/ACM Trans. Netw. **24**(5), 2834–2846 (2016)
13. Data tang: 63641 sina micro-blog data set. http://www.datatang.com/data/46758. Accessed 20 Nov 2016
14. Ritterman, J., Osborne, M., Klein, E.: Using prediction markets and Twitter to predict a swine flu pandemic. In: First International Workshop on Mining Social Media, pp. 9–17. ACM (2009)
15. Knight, W.R.: A computer method for calculating Kendall's tau with ungrouped data. J. Am. Stat. Assoc. **61**(314), 436–439 (1966)