Effectiveness of Incorporating Follow Relation into Searching for Twitter Users to Follow

Tomoya Noro and Takehiro Tokuda

Department of Computer Science, Tokyo Institute of Technology
Meguro, Tokyo 152-8552, Japan
{noro,tokuda}@tt.cs.titech.ac.jp

Abstract. Twitter is one of the most popular microblogging services that facilitate real-time information collection, provision, and sharing. Following influential Twitter users is one way to get valuable information related to a topic of interest efficiently. Recently many researches on this issue have been done and, in general, it is said that the follow relation is not useful for measuring user influence. In this paper, we study effectiveness of incorporating not only the tweet activity (retweet and mention) but also the follow relation into searching for good Twitter users to follow for getting information on a topic of interest. We present a method for finding Twitter users based on both the follow relation and the tweet activity, and show the follow relation could improve the performance as compared with methods based on only the tweet activity.

Keywords: Social network analysis, microblog, Twitter, search, influential users, power iteration algorithm.

1 Introduction

Currently Twitter has become a more and more important platform of information collection, provision, and sharing. If we would like to get information of a topic of interest and discuss the topic with others on a daily basis, we usually follow some users who provide valuable information on the topic [1]. For example, if we look for information about dementia, we could find that some doctors and care staff members deliver information about the topic, and some people who have family members with dementia post tweets about their daily care. We can get various information on dementia by following such users. However, it is not easy for us to find good users to follow in a massive number of users.

Many researches on this issue have been done recently. Measuring user influence on a particular topic will be one solution. They measure each user's influence based on the tweet content, the follow relation, the tweet activity such as retweet and mention, and so on, and some of them pointed out that the follow relation is not useful for measuring user influence. Cha et al. investigated

[1] In this paper, we do not consider temporary topics such as incidents and events (natural disaster, terrorism, FIFA World Cup, etc). If we would like to get information on such topics, we would take different actions such as keyword search.

S. Casteleyn, G. Rossi, and M. Winckler (Eds.): ICWE 2014, LNCS 8541, pp. 420–429, 2014.

characteristics of Twitter users, then concluded users who have many followers are popular but not necessarily influential, while users who are retweeted or mentioned many times have ability to post valuable tweets or ability to engage others in conversation [2]. We have also been working on this issue and showed a user search method based on the tweet activity outperforms a method based on the follow relation [6].

Here is one question. Is the follow relation really useless for finding good users to follow? Actually some users follow almost all of their followers. We can easily get many followers if we search for such users, follow them, wait for them to follow us, and remove them if they do not follow us. The follow relation built in this way is meaningless since most of the followers may not be interested in our tweets and we have little influence on them. However, in general, influential users have many followers. Users who have a small number of followers may not be good users to follow since they should have more followers if they provide a lot of valuable information. The search accuracy could be improved by dealing with both the tweet activity and the follow relation.

In this paper, we study effectiveness of incorporating not only the tweet activity but also the follow relation into searching for good Twitter users to follow for getting information on a topic of interest. We present a method for finding good users to follow based on both the follow relation among users and the tweet activity of each user. In evaluation, we compare the method with other methods without taking the follow relation into account and show incorporating both the tweet activity and the follow relation could improve the search performance.

This paper is organized as follows. In section 2, we discuss some researches on searching Twitter users to follow on a topic of interest. We present our method based on both the follow relation and the tweet activity in section 3, then show some evaluation results in section 4. Lastly we conclude this paper in section 5.

2 Related Work

Twitter provides its own services for user search and recommendation. Given some keywords, Twitter mainly shows some users whose screen names or profiles match the keywords and does not care whether they actually post tweets related to the keywords. The Twitter recommendation service shows users based on the follow relation (users who have mutual followers and/or friends), and does not consider their activity and the tweet content either.

Twittomender [3] finds users related to a particular user or query by using lists of followers, friends and terms in the user's tweets. TwitterRank [8] considers the follow relation and topical similarity to find influential users. Both of the methods are based on the follow relation and they do not consider the tweet activity such as retweet and mention.

Leavitt et al. [5] measured user influence by using ratio of being retweeted and mentioned to the number of tweets the user posted. They do not consider the follow relation. Anger et al. [1] defined user influence based on ratio of retweeted tweets and ratio of the user's followers who retweeted the user's tweets or mentioned the user. Although they consider both the tweet activity and the follow

relation, they use the follow relation to observe how many followers retweeted the user's tweets or mentioned the user and do not consider who follows whom.

We presented a method for finding good users to follow for getting information about a topic of interest by using the tweet activity [6]. Although we showed that a search method based on the tweet activity outperforms a method based on the follow relation, we study effectiveness of considering both the tweet activity and the follow relation in this paper.

3 Method for Finding Good Twitter Users to Follow

3.1 Overview

Our process of finding good users to follow on a topic of interest goes as follows.

1. Given some keywords representing the topic of interest, collect tweet data and user data by using the Twitter APIs.
2. Create a user reference graph based on the follow relation and a user-tweet reference graph based on both the tweet activity and the follow relation.
3. Calculate score of each user from the two graphs and rank the users.

The data collection process in the first step goes as follows.

1. Given some keywords representing a topic of interest, get tweets matching the keywords posted in the last N days. Duplicate tweets (exactly the same tweet text posted by the same user) are removed to exclude spammers who post the same tweets repeatedly. Let this tweet set be T_0.
2. For each tweet in T_0, get ID and poster's name of the tweet and user names in the tweet text (user mention). If the tweet is a reply tweet/retweet, get ID and poster's name of the replied/retweeted tweet. Let the set of tweets and the set of users be T_{all} and U_{all} respectively.
3. Get the follow relation among users in U_{all} ($F \subseteq U_{all} \times U_{all}$).

3.2 Score Calculation

In order to define the score of each user, we assume the followings.

1. Users who post many valuable tweets about the topic are worth following.
2. Valuable tweets attract attention from many users.
3. Each user pay attention to tweets the user retweets or replies to.
4. Each user also pay attention to tweets posted by the user's friends.

The first assumption means that users who post many tweets related to the topic should be ranked higher. However, some users who post many valueless tweets such as spam tweets will also be ranked higher if we consider only this assumption. To exclude such users, we take the other assumptions into account. A user's retweeting or replying to a tweet means that the user is interested in the tweet. Each user may pay attention to a tweet posted by the user's friends to some extent even if the user did not retweet or reply to the tweet.

Based on these assumptions, we define the score of each user u as follows.

$$\text{Score}(u) = \text{TC}(u)^{w_c} \times \text{UI}(u)^{w_i} \times \text{FR}(u)^{w_f}$$
$$\text{such that} \quad w_c + w_i + w_f = 1 \wedge w_c \geq 0 \wedge w_i \geq 0 \wedge w_f \geq 0 \quad (1)$$

$\text{TC}(u)$, $\text{UI}(u)$, and $\text{FR}(u)$ are respectively "tweet count (TC) score", "user influence (UI) score", and "follow relation (FR) score" of user u ranging between 0 and 1. The TC score is based on the number of tweets each user posted, and reflects the first assumption. The FR score is based on the follow relation among users, and reflects the second and forth assumptions. The UI score is based on both the tweet activity and the follow relation, and reflects the second, third and forth assumptions.

3.3 Tweet Count Score (TC Score)

The TC score is calculated by counting not only each user's original tweets but also retweets in T_0. We count retweets as each user's own tweets [2]. The score is normalized so that the largest value should be 1.

$$\text{TC}(u) = \frac{\log(1 + |\{t | t \in T_0 \wedge t.user.id = u.id\}|)}{\max_{u' \in U_{all}} \log(1 + |\{t | t \in T_0 \wedge t.user.id = u'.id\}|)} \quad (2)$$

$t.user.id$ indicates poster's ID of tweet t and $u.id$ indicates ID of user u.

3.4 User Influence Score (UI Score)

The basic idea is as follows.

1. If user u_i retweets or replies to user u_j's tweet, u_j has an influence on u_i.
2. Users who post many tweets paid attention to by many users are influential, especially if their tweets are often paid attention to by influential users.

How much each tweet is paid attention to by others is measured according to the tweet activity (retweet and reply) and the follow relation. Based on this idea, we define not only the UI score of each user but also tweet influence score (TI score) of each tweet. The UI score is calculated using the TI score of tweets and retweets posted by the user, and the TI score is calculated using the UI score of users who pay attention to the tweet.

We create a user-tweet reference graph consisting of user nodes (U_{all}), tweet nodes (T_{all}), and directed edges each of which connects a user node and a tweet

[2] The retweet activity is incorporated into both the TC score and the UI score. The number of times each user retweeted is considered in the TC score while the user-tweet relation (who retweeted what) is considered in the UI score.

node. The reference graph is represented as combination of three adjacency matrices A_t, A_r, and A_s.

$$A_t(t_i, u_j) = \begin{cases} 1 & \text{if } t_i \text{ is tweeted/retweeted by } u_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$A_r(u_j, t_i) = \begin{cases} 1 & \text{if } u_j \text{ retweets/replies to } t_i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$A_s(u_j, t_i) = \begin{cases} 1 & \text{if } u_j \text{ follows at least 1 user who tweets/retweets } t_i \\ \alpha & \text{otherwise } (0 < \alpha \leq 1) \end{cases} \tag{5}$$

t_i and u_j indicates the i-th tweet and the j-th user respectively ($1 \leq i \leq |T_{all}|$ and $1 \leq j \leq |U_{all}|$). A_t and A_r are derived from the tweet activity of each user, and A_s is derived from the follow relation among users. A_t represents what (tweet) is tweeted or retweeted by whom (user), and A_r and A_s respectively represent who retweets or replies to what and who sees what. The follow relation will be ignored if α is equal to 1.

These adjacency matrices are transformed into the following two matrices.

$$B_t(t_i, u_j) = \frac{A_t(t_i, u_j)}{\sum_k A_t(t_i, u_k)} \tag{6}$$

$$B_a(u_j, t_i) = \begin{cases} \frac{A_r(u_j,t_i)}{\sum_k A_r(u_j,t_k)}(1-d) + \frac{A_s(u_j,t_i)}{\sum_k A_s(u_j,t_k)}d & \text{if } \sum_k A_r(u_j, t_k) \neq 0 \\ \frac{A_s(u_j,t_i)}{\sum_k A_s(u_j,t_k)} & \text{otherwise} \end{cases} \tag{7}$$

d is a damping factor of $0 < d < 1$. Transformation of A_r and A_s into B_a reflects the third and forth assumptions of good users to follow described in section 3.1. Each user pay attention to tweets the user retweets or replies to, and the user also watches all tweets at a certain rate of d regardless of the user's activity of retweet and reply. Tweets posted or retweeted by the user's friends are more likely to be seen than the other tweets, and the idea is also included.

The UI score and the TI score are calculated as follows.

$$\mathbf{u} = B_t^T \mathbf{t} \qquad\qquad \mathbf{t} = B_a^T \mathbf{u} \tag{8}$$

$\mathbf{u}$ and $\mathbf{t}$ indicate a column vector of the UI score of all users and a column vector of the TI score of all tweets respectively. We can calculate the UI score and the TI score using the power iteration method. Lastly the UI score of each user is normalized so that the largest value should be 1.

$$\mathrm{UI}(u_j) = \frac{\mathbf{u}(j)}{\max_k \mathbf{u}(k)} \tag{9}$$

3.5 Follow Relation Score (FR Score)

The FR score is calculated based on the follow relation using PageRank [7]. A user reference graph is created from the follow relation F. Adjacency matrix of the graph is represented as follows.

$$A_f(u_i, u_j) = \begin{cases} 1 & \text{if } u_i \text{ follows } u_j \text{ i.e. } (u_i, u_j) \in F \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$B_f(u_i, u_j) = \begin{cases} \frac{A_f(u_i,u_j)}{\sum_k A_f(u_i,u_k)}(1-d) + \frac{d}{|U_{all}|} & \text{if } \sum_k A_f(u_i, u_k) \neq 0 \\ \frac{1}{|U_{all}|} & \text{otherwise} \end{cases} \tag{11}$$

$$\mathbf{f} = B_f^T \mathbf{f} \tag{12}$$

u_i and u_j indicates the i-th user and the j-th user respectively, and d is a damping factor. $\mathbf{f}$ indicates the column vector of the FR score of all users.

Unlike normalization of the TC score and the UI score, the FR score of each user is not divided by the maximum value. As described in section 1, users who have many followers are not necessarily influential, and it is said that the follow relation is not useful for measuring user influence. However, we think the follow relation could be used for excluding uninfluential users who have few (influential) followers. Instead of dividing the FR score of each user by the maximum value, we set upper limit of the FR score to the minimum score of the top-P% users and divide the score of each user by the limit.

$$\mathrm{FR}(u_i) = \frac{\min(\mathbf{f}(i), \mathit{limit})}{\mathit{limit}} \tag{13}$$

limit indicates the minimum FR score of the top-P% users. This normalization can weaken influence of the users who have high FR score since the score of all of the top-P% users will be set to 1.

Some alternative ways for normalization may be considered. For example, some may think of normalization by setting the upper limit to a predetermined proportion to the maximum value (e.g. $\mathit{limit} = 0.1 \times \max_k \mathbf{f}(k)$) or determining the number of users to be capped (e.g. Top-50 users). However, the number of users to be ranked depends on topics of interest, and distribution of the FR score also varies by the topics. We think that determining the percentage of users to be capped is better from our observation.

4 Evaluation

4.1 Experimental Setup

We selected the following 7 Japanese keywords (in Japanese characters) as input query representing topics of interest: “nuclear power”, “animal test”, “whaling”, “dementia”, “digital book”, “basic income” and “fair trade”. We chose these topics since we expect that tweets related to the topics are posted on a daily basis (independent of season). Tweets and other data were collected 6 times on different days. For each time, we get tweets posted in the last 5 days. The average number of tweets and users we collected is shown in Table 1. “Reply” means tweets replying to tweets specified in “reply-to” attribute, and “Mention” means tweets including user names but not specifying their target tweets.

Table 1. The average number of tweets and users

Keyword	$\|T_0\|$				$\|T_{all}\|$	$\|U_{all}\|$
	Total	Retweet	Reply	Mention		
nuclear power	26,937.7	14,878.0	1,008.7	1,336.0	28,124.0	13,435.3
animal test	2,591.7	1,539.8	185.7	126.5	2,818.3	1,349.3
whaling	4,057.7	1,045.7	254.0	321.0	4,249.7	3,112.7
dementia	5,497.5	1,670.7	832.3	163.5	6,255.0	4,959.3
digital book	19,208.0	4,408.2	1,307.8	1,273.3	20,333.5	12,976.8
basic income	400.7	148.8	68.3	19.8	449.0	251.5
fair trade	779.2	364.8	70.7	14.7	849.8	662.2

Table 2. Value of each parameter

d in Eqs. (7) and (11)	0.15
P for FR(+lim)	5%
α in Eq. (5) for UI(+fol)	0.1
w_c and w_i in Eq. (1) for TC+UI	0.6 and 0.4
w_c, w_i and w_f in Eq. (1) for TC+UI+FR	0.6, 0.2, and 0.2

We set up the following methods for comparison.

TC+UI(+fol)+FR(+lim): Rank users based on the TC score, the UI score with the follow relation, and the FR score with upper limit.

TC+UI(-fol)+FR(+lim): Rank users based on the TC score, the UI score without the follow relation, and the FR score with upper limit (α in Eq. (5) is set to 1).

TC+UI(+fol)+FR(-lim): Rank users based on the TC score, the UI score with the follow relation, and the FR score without upper limit (normalization of the FR score is done by dividing each score by the maximum value).

TC+UI(-fol)+FR(-lim): Rank users based on the TC score, the UI score without the follow relation, and the FR score without upper limit.

TC+UI(+fol): Rank users based on the TC score and the UI score with the follow relation (w_f in Eq. (1) is set to 0).

TC+UI(-fol): Rank users based on the TC score and the UI score without the follow relation

TC: Rank users based on only the TC score.

UI(+fol): Rank users based on only the UI score with the follow relation.

UI(-fol): Rank users based on only the UI score without the follow relation

FR(-lim): Rank users based on only the FR score without upper limit

We carried out a preliminary experiment using tweet data collected on different days to determine the parameters appeared in section 3. The value of each parameter is shown in Table 2.

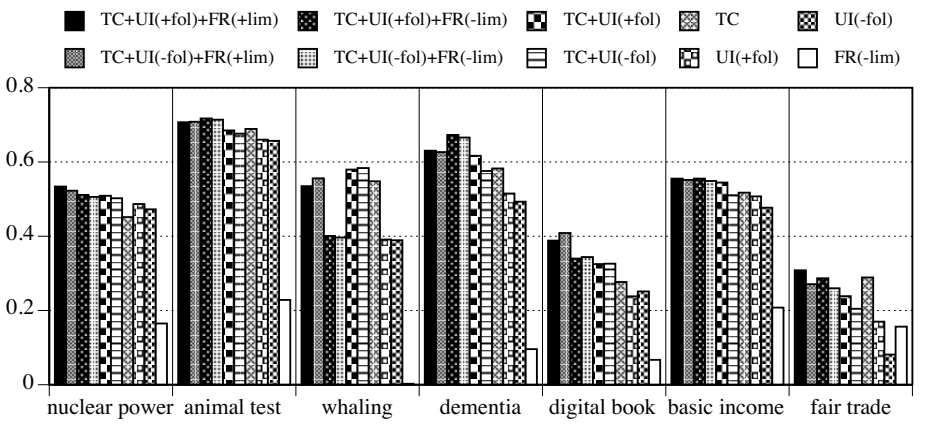

Fig. 1. Average nDCG of each method

Evaluation is done on the top 20 users ranked by each method with respect to normalized discounted cumulative gain (nDCG) [4].

$$\mathrm{DCG}_{20} = rel_1 + \sum_{i=2}^{20} \frac{rel_i}{\log_2 i} \qquad \mathrm{maxDCG}_{20} = 2 + \sum_{i=2}^{20} \frac{2}{\log_2 i} \tag{14}$$

$$\mathrm{nDCG}_{20} = \frac{\mathrm{DCG}_{20}}{\mathrm{maxDCG}_{20}} \tag{15}$$

rel_i indicates relevance score between the user ranked i-th and the input keyword, which is judged on a scale of 0 to 2. Users who often post related tweets are assigned the score of 2, while users who rarely post related tweets are assigned the score of 0. Users who post a lot of unrelated tweets like advertisement and spams are also assigned the score of 0. The judgment was done by watching their tweets posted after the data collection period (for about one month) to check whether they continuously post related tweets.

4.2 Result

The result is shown in Figure 1. We can see that considering the follow relation in calculation of the UI score improves the search result on average (e.g. methods including UI(+fol) vs methods including UI(-fol)).

Except for the case of "whaling", incorporating the FR score is also effective (e.g. TC+UI(+fol)+FR(+lim) vs TC+UI(+fol)). The FR(-lim) method found no relevant user in the case of "whaling". When incidents related to whaling occur, many major news organizations will post tweets related to the topic and will get high ranking in the FR score since they have many followers. However, their interest is not always focused on the topic. On the other hand, users who usually talk about or discuss the topic do not have strong follow relation with others

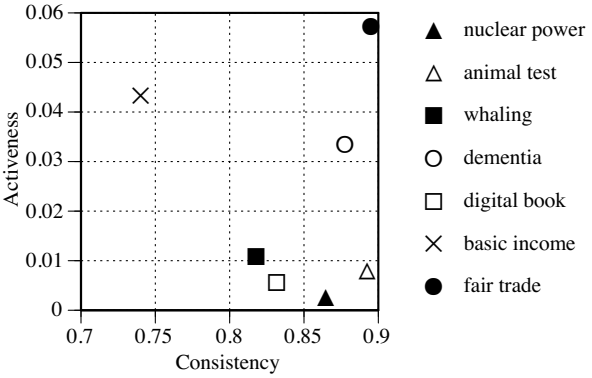

Fig. 2. Consistency and activeness of each keyword

compared to such news organizations. We think this is why the performance were not improved by incorporating the FR score.

Effectiveness of setting the upper limit for normalization of the FR score differs according to keywords (e.g. TC+UI(+fol)+FR(+lim) vs TC+UI(+fol)+FR (-lim)). In the case of "whaling", a large decline caused by incorporating the FR score is prevented. This is because influence of users who is ranked high on the FR score but do not focus on the topic (e.g. major news organizations) are reduced. Setting the upper limit worsen the search performance in the case of "dementia" since, unlike the case of "whaling", some users who usually focus on the topic also have strong follow relation with others and difference between their FR score and the score of major news organizations is not large.

If we compare methods including UI(+fol) with methods including UI(-fol) in more details, we can see that incorporating the follow relation in calculation of the UI score seems not to improve the search performance in the case of "whaling" and "digital book". To analyze this issue, we calculate two measures. One measure is "consistency", how much their tweet activity (retweet and reply) is consistent with their follow relation, and the other measure is "activeness", how many tweets posted by their friends they retweet or reply to. They are defined as follows.

$$\text{Consistency} = \frac{Consistent}{Consistent + Inconsistent} \tag{16}$$

$$\text{Activeness} = \frac{Consistent}{Consistent + Ignored} \tag{17}$$

"*Consistent*" means the number of times they retweeted or replied to their friends' tweets and retweets, and "*Inconsistent*" means the number of times they retweeted or replied to tweets and retweets posted by none of their friends. "*Ignored*" means the number of times they do not retweeted nor replied to their friends' tweets and retweets. From the result shown in Figure 2, we can see both consistency and activeness of "whaling" and "digital book" are low compared

with other keywords. We guess users who are interested in the topics are likely to see tweets of non-following users and to communicate with them while they do not communicate with their friends so much. This situation would occur if they usually talk about or discuss topics of interest by using hashtags. It can also be seen in the case of "basic income" but, in this case, many of them communicate with their friends as well as other users (activeness is high in the case of "basic income"). On the other hand, both consistency and activeness of "fair trade" and "dementia" are high. In both cases, we can see effectiveness of incorporating the follow relation into calculation of the UI score. The percentage of retweets, reply tweets and mention tweets of "digital book" and "whaling" is low (36.3% and 40.0% respectively) as shown in Table 1, which would also be one factor that worsens the performance.

5 Conclusion

In this paper, we presented a method for finding good Twitter users to follow for getting information about a topic of interest based on both the tweet activity and the follow relation, and showed incorporating both of them improves the search performance on average except for the case that performance of the method based only the follow relation is extremely bad. We also measured "consistency" and "activeness", and found effectiveness of incorporating the follow relation into the UI score is high if the two scores are high.

References

1. Anger, I., Kittl, C.: Measuring influence on Twitter. In: 11th International Conference on Knowledge Management and Knowledge Technologies, vol. 31. ACM Digital Library (2011)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in Twitter: The million follower fallacy. In: 4th International AAAI Conference on Weblogs and Social Media, pp. 10–17 (2010)
3. Hannon, J., Bennett, M., Smyth, B.: Recommending Twitter users to follow using content and collaborative filtering approaches. In: 4th ACM Conference on Recommender Systems, pp. 199–206 (2010)
4. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
5. Leavitt, A., Burchard, E., Fisher, D., Gilbert, S.: The influentials: New approaches for analyzing influence on Twitter. Web Ecology Project (2009), `http://www.webecologyproject.org/2009/09/analyzing-influence-on-twitter/`
6. Noro, T., Ru, F., Xiao, F., Tokuda, T.: Twitter user rank using keyword search. In: Information Modelling and Knowledge Bases XXIV. Frontiers in Artificial Intelligence and Applications, vol. 251, pp. 31–48. IOS Press (2013)
7. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford University (1998)
8. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: Finding topic-sensitive influential Twitterers. In: 3rd ACM International Conference on Web Search and Data Mining. pp. 261–270 (2010)