Journal of Web Engineering, Vol. 15, No. 3&4 (2016) 249–276

SEARCHING FOR RELEVANT TWEETS BASED ON TOPIC-RELATED USER ACTIVITIES

TOMOYA NORO, TAKEHIRO TOKUDA
Department of Computer Science, Tokyo Institute of Technology
Meguro, Tokyo, 152-8552, Japan
noro.t.tech@gmail.com, tokuda@cs.titech.ac.jp

Received March 12, 2015
Revised February 6, 2016

Twitter is one of the largest social media. Although it can be used to get information on a topic of interest, it is not easy for us to find tweets relevant to the topic due to a massive amount of tweets and the small size of each tweet. Some relevant tweets may not include any terms explicitly related to the topic, and general content-based keyword search techniques and query expansion techniques are not effective for finding such relevant tweets. To solve this problem, we present a method for finding tweets on a topic of interest based on the Twitter user activities related to the topic such as tweet, retweet, and reply. The method consists of two phases: the preparation phase and the main phase. In the preparation phase, we create a user-tweet reference graph representing the relation between users and tweets based on the past user activities related to the topic, calculate the influence of each user and tweet in the topic, then define two types of each user's power, called "Voice" and "Impact", indicating "how much voice the user has on the topic" and "how much impact the user has on the other users' tweets on the topic". In the main phase, we calculate the relevance of newly-arrived tweets to the topic according to the Voice and the Impact score of the users who posted, retweeted, or replied to each of the tweets, then rank the tweets by the relevance score. The two phases are processed independently. Once the preparation phase is completed, the main phase can return the final result any time. Experimental results show that "who retweeted or replied to the tweet" is more effective for judging the relevance of each tweet to the topic than "who posted the tweet", and our method can find relevant tweets which do not include any terms explicitly related to the topic. We compare our method with an indegree-based method and a PageRank-based method, and show that our method outperforms the methods compared.

Keywords: Social media, Twitter, social network analysis, search, graph-based approach
Communicated by: M. Gaedke & Q. Li

1. Introduction

Social media play an important role of platforms for collecting, providing, and sharing information among users [1, 2, 3]. One of the popular services is Twitter, which has 300 million monthly active users posting 500 million tweets per day in 2015 [4].

In order to get information about a topic of interest efficiently on a daily basis, we find and follow some Twitter users who usually post valuable tweets on the topic. Many researches on finding topic-related users have been done recently [5, 6, 7, 8, 9, 10]. However, not many

users limit their tweets to a particular topic and, as a result, we are forced to see some tweets of no interest in exchange for receiving some tweets on the topic of interest.

Another way to get information on a topic of interest in Twitter is keyword search. However, it is not easy for us to find relevant tweets since each tweet text is short (less than or equal to 140 characters). Some relevant tweets may not match the keywords (low recall) and some tweets matching the query may not be relevant to the topic of interest (low precision). Unlike general document search, we think the query expansion techniques are not effective for the tweet search.

Suppose, for example, we would like to search for tweets on whaling. If we give "whale" and "whaling" to the keyword search (OR search), we cannot find any tweets including "dolphin". Some tweets including "whale" is not related to whaling but related to watching whales in the ocean or aquariums. Although some tweets on whaling include "Sea Shepherd" (an anti-whaling group), "Sea Shepherd" is often abbreviated as "SS" and there are many other terms which can be abbreviated as "SS" in Twitter. Some of these cases could be solved by using some language resources such as DBpedia [11] and YAGO [12, 13]. However, some tweets may not include any terms explicitly related to whaling. The following tweets are real examples (URLs are omitted).

1. One Pod Escapes! http://...
2. Great article about recent London protest!! http://...

One user was watching whaling in Japan and posted the first tweet with a photo. "Recent London protest" in the second tweet indicates the protest against whaling in Japan. In both cases, it is difficult to know if they are related to whaling from the tweet texts. However, if we know that the posters of these tweets and the users related to the tweets (the users who retweeted or replied to the tweets) are interested in whaling and they usually post, retweet, and reply to tweets on whaling, we guess that these tweets are related to whaling, i.e. we can judge whether a tweet is related to the topic of interest by "who posted, retweeted, or replied to the tweet".

The purpose of our research is to find tweets relevant to a topic of interest. We refer to the topic of interest as "target topic" or "topic" in this paper. We focus on the persistent topics discussed in Twitter on a daily basis, and we do not consider temporary topics and unexpected events such as natural disaster, terrorism, and so on. Our approach is based on the user activities, i.e. who posted, retweeted, or replied to each tweet. We assume that the relevant tweets are posted, retweeted, or replied to by users who have "power" in the topic. In this paper, we consider two types of the power, called "Voice" and "Impact", indicating respectively how much voice a user has on the topic and how much impact a user has on the others' tweets on the topic. Some users post and retweet many valuable tweets on the topic. This type of users have the power of Voice. Tweets retweeted or replied to by some users with the power of Impact will draw more attention than tweets retweeted or replied to by some ordinary users. Some users have the both types of the power, and some other users have either of them. For example, some famous specialists who usually deliver some information or their opinions on the topic will have both powers. Their tweets will be valuable and reliable since they are specialists, and, if they retweet or reply to some others' tweets, the tweets will draw more attention for the same reason (i.e. the tweets will get an endorsement from the

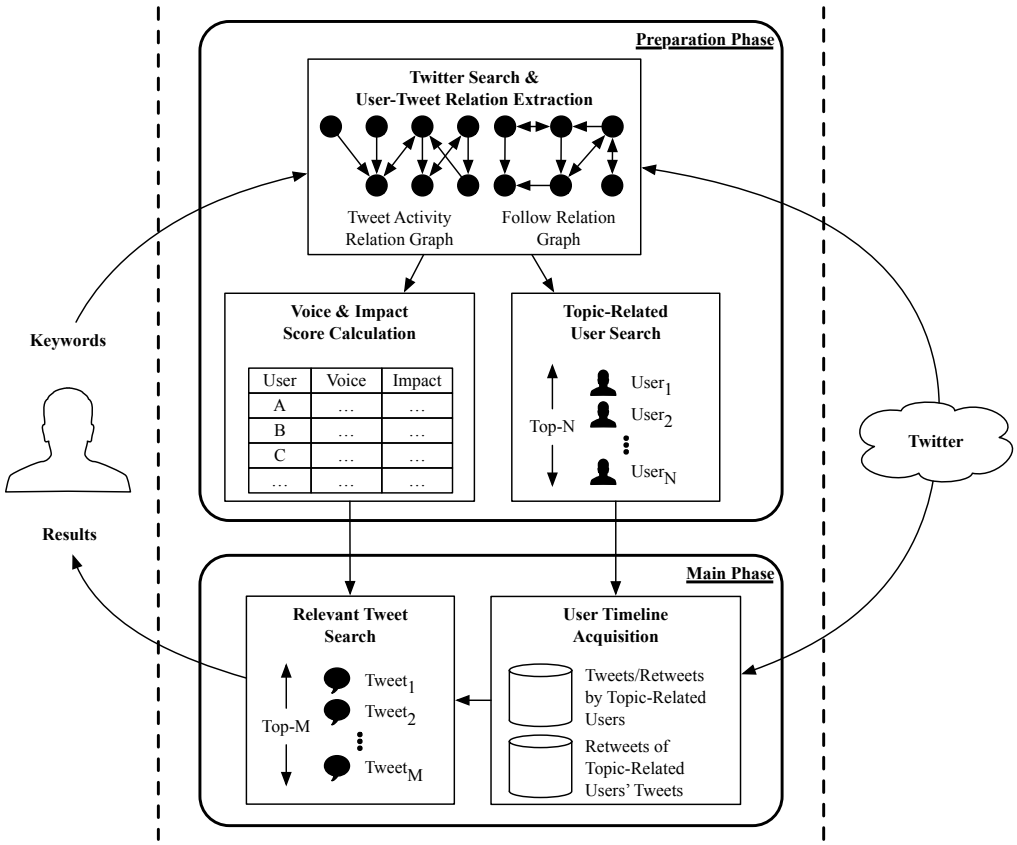

Fig. 1. System structure of our method

specialists).[a] On the other hand, some ordinary users who often retweet valuable tweets on the topic will have only the power of Voice, since they are not so popular as to add much value to the tweets. In our method, we calculate the Voice and the Impact scores from the past activities of each user.

The system structure of our method is shown in Figure 1. It consists of two phases: the preparation phase and the main phase. Given an input query (keywords) representing the target topic, the preparation phase firstly searches for tweets matching the query by the Twitter Search, and creates two types of reference graphs consisting of user nodes and tweet nodes, called "tweet activity relation graph" and "follow relation graph". The tweet activity relation graph represents who (user) posted, retweeted, or replied to what (tweet), and the influence of each user and tweet based on the user activities is calculated from the graph. The follow relation graph represents who (user) follows whom (user), and the influence of each user based on the follow relation is calculated from the graph. Then, users related to the topic (topic-related users) are determined according to the influence of each user based on the user

[a]They sometimes retweet or reply to some tweets to express disagreement on them. Even in this case, the tweets will draw more attention.

activities and the follow relation. The Voice score and the Impact score of each user (not only the topic-related users but also all other users appearing in the searched tweets) are calculated from the influence of each user and tweet based on the user activities. After the preparation phase has been completed, the main phase collects original tweets and retweets of the topic-related users, ranks the tweets using the Voice score and the Impact score, then returns the ranking result. Note that the two phases work independently, i.e. once the preparation phase has been completed, the main phase can be processed any time.

Our method has a feature that it can find relevant tweets which do not include any terms explicitly related to the topic. The general content-based keyword search or the query expansion techniques rely on the input queries and they are not effective for finding the example tweets described previously. Our method considers who posted, retweeted, or replied to each tweet in the main phase and can solve the problem. Users who are interested in a particular topic usually post, retweet, and reply to tweets relevant to the topic, and valuable tweets relevant to the topic are often retweeted and replied to by such users. The example tweets described previously were posted by the users who are interested in whaling, and most of the users who retweeted and replied to the tweets are also interested in whaling. Our method is based on this idea.

The organization of the rest of this paper is as follows. In section 2, we discuss some related works. We describe a method for finding users related to the target topic in section 3, then define the Voice/Impact score and present a method for finding relevant tweets in section 4. We show some evaluation results and have some discussions in section 5. Lastly we conclude this paper in section 6.

2. Related Work

Twitter officially provides the tweet search service. Given an input query, it shows a list of tweets matching the query. The list is sorted in reverse chronological order or by popularity-based ranking. In the popularity-based ranking, it seems that tweets retweeted many times and tweets posted by famous users (official accounts managed by famous organizations) tend to be ranked higher (the detail algorithm is not open to public). If new tweets matching the query are generated, the tweets are automatically added at the top of the list. However, we need to give all keywords representing the target topic, and it cannot search for tweets which do not include any terms explicitly related to the topic.

Twinder [14] is a search engine for Twitter streams to find tweets relevant to a given topic. It uses topic-sensitive features and topic-insensitive features to estimate the relevance of tweets. It considers word occurrence (keyword-based relevance) and named entities (semantic-based relevance) for the topic-sensitive features. We think that it is difficult for Twinder to find relevant tweets which do not include any terms explicitly related to the topic.

Duan et al. proposed a method for ranking tweets relevant to an input query by using content relevance features, Twitter specific features, and account authority features [15]. Their idea of the content relevance features is based on term frequency and inverse document frequency (TF-IDF). However, TF-IDF is not appropriate for estimating the relevance of tweets to an input query, since the tweet length is limited [16]. As is the case in Twinder, their method will also fail to find relevant tweets which does not include any terms explicitly related to the query. Although they consider retweet relations among users by computing the

PageRank score (one of the account authority features), it is not specific to the query.

Some researches consider the task of finding or ranking tweets as a retweet prediction problem. Although most of the researches deal with the problem from a global perspective, Uysal et al. proposed a user-centric approach, which predicts whether a specified user will retweet a given tweet or not by taking author-based, tweet-based, content-based, and user-based features into account [17]. However, the purpose of our research is not exactly the same as theirs. Their approach focuses on only one specified user. The specified user may be interested in more than one topic, and tweets obtained by their method are not necessarily limited to one topic (they do not aim to limit to one particular topic). The purpose of our research is to find tweets relevant to one particular topic, and our method finds such tweets based on the activities of multiple topic-related users.

3. Searching for Topic-Related Users

In this section, we describe a method for finding users related to the target topic (topic-related users) in the preparation phase. The method is inspired by our previous work [8].

In our research, we have the following assumptions about the topic-related users.

1. Good topic-related users usually post valuable tweets on the target topic.
2. The valuable tweets on the topic draw the attention of many users.
3. Each user pays attention to the tweets the user retweets or replies to.
4. Each user also pays attention to the tweets posted by the user's friends (followees).

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. We take the other assumptions into account to exclude such users. The tweet activity such as retweet and reply is a better indicator for measuring how much attention each tweet is paid compared with the follow relation. Cha et al. investigated characteristics of Twitter users, and concluded that 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 [18]. Our idea basically follows their observation and we take the third assumption. However, we do not think the follow relation is not entirely meaningless. Although users with many followers are not necessarily post many valuable tweets, users with a small number of followers may not have the ability to post valuable tweets since, if they post many valuable tweets, they should have more followers. We showed that considering the follow relation as well as the tweet activity improves the performance of the topic-related user search in our previous work [8]. The fourth assumption is based on this observation.

Based on these assumptions, we define the user relevance score of each user u as follows, and select the top-N users ranked by the user relevance score as the topic-related users.

$$\mathrm{UserRel}(u) = \mathrm{TR}(u)^{w_r} \times \mathrm{UI}(u)^{w_i} \times \mathrm{FR}(u)^{w_f} \tag{1}$$

$\mathrm{TR}(u)$, $\mathrm{UI}(u)$, and $\mathrm{FR}(u)$ are respectively "tweet rate (TR) score", "user influence (UI) score", and "follow relation (FR) score" of user u ranging between 0 and 1. w_r, w_i, and w_f are non-negative values where the sum of the values is equal to 1. The TR score is based on the tweet

frequency, and reflects the first assumption. The UI score is based on the tweet, retweet, and reply activities and the follow relation, and reflects the second, third, and fourth assumptions. The FR score is based on the follow relation, and reflects the second and fourth assumptions.

3.1. *Tweet Rate (TR) Score*

The topic-related users post many tweets relevant to the target topic. However, there are some users who post many relevant tweets and much more irrelevant tweets (some users post hundreds of tweets on various topics every day). In order to exclude such users, we consider the tweet rate instead of the number of each user's tweets searched in the preparation phase.

In calculation of the TR score, we count not only each user's original tweets but also retweets as the user's own tweets. Some of the topic-related users usually retweet tweets relevant to the topic originally posted by others, which means they play a role of "filter" searching for valuable relevant tweets and sharing them with their followers.

$$\mathrm{TR}(u) = \frac{|\{t | t \in T \wedge t.user.id = u.id\}|}{|\mathrm{Total}(u)|} \tag{2}$$

T indicates a set of the tweets searched in the preparation phase, and $\mathrm{Total}(u)$ indicates a set of all the tweets posted or retweeted by user u during the same time period as the tweet search. $t.user.id$ and $u.id$ indicate the poster's ID of tweet t and the ID of user u respectively. The score is normalized so that the largest value should be 1.

3.2. *User Influence (UI) Score*

The basic idea is as follows.

1. Users who post many tweets on the target topic paid attention to by many users are good topic-related users. This is based on the first assumption.

2. Tweets of the good topic-related users are often paid attention to by other good topic-related users. This is based on the second, third, and fourth assumptions.

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

$$\mathbf{u} = B_t^T \mathbf{t} \qquad\qquad \mathbf{t} = B_a^T \mathbf{u} \tag{3}$$

$\mathbf{u}$ and $\mathbf{t}$ indicate a column vector of the UI score and a column vector of the TI score respectively. B_t is the tweet-to-user relation matrix based on what (tweet) was posted or retweeted by whom (user), and B_a is the user-to-tweet relation matrix based on who paid attention to what.

To derive the two relation matrices B_t and B_a, we create a reference graph consisting of user nodes, tweet nodes, and directed edges each of which connects a user node and a tweet

node, called "tweet activity relation graph". The tweet activity relation graph is represented as combination of three adjacency matrices A_t, A_r, and A_s.

$$A_t(t_i, u_j) = \begin{cases} 1 & \text{if } t_i \text{ is posted/retweeted by } u_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$A_r(u_j, t_i) = \begin{cases} 1 & \text{if } u_j \text{ retweets/replies to } t_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$A_s(u_j, t_i) = \begin{cases} 1 & \text{if } u_j \text{ follows at least 1 user who posts/retweets } t_i \\ s & \text{otherwise } (0 < s \leq 1) \end{cases} \tag{6}$$

t_i and u_j indicates the i-th tweet and the j-th user respectively. A_t represents what is posted or retweeted by whom, and A_r and A_s represent who retweets or replies to what and who sees what respectively. It is natural for users to reply if they are mentioned by others, and some users often retweet tweets mentioning themselves. These activities are ignored in creation of the tweet activity relation graph since they do not always depend on topics. Activities replying to themselves are also ignored.

The adjacency matrices are transformed into the two relation matrices B_t and B_a as follows.

$$B_t(t_i, u_j) = \frac{A_t(t_i, u_j)}{\sum_k A_t(t_i, u_k)} \tag{7}$$

$$B_a(u_j, t_i) = \begin{cases} \frac{A_r(u_j,t_i)}{\sum_k A_r(u_j,t_k)}(1-d) + \frac{A_s(u_j,t_i)}{\sum_k A_s(u_j,t_k)}d & \text{if } \sum_k A_r(u_j, t_k) \neq 0 \\ \frac{A_s(u_j,t_i)}{\sum_k A_s(u_j,t_k)} & \text{otherwise} \end{cases} \tag{8}$$

d is a damping factor of $0 < d < 1$. The matrix B_t means that the TI score of each tweet is given to the users who posted or retweeted the tweet, and it reflects the first assumption. The matrix A_r in B_a means that the UI score of each user is given to the tweets the user retweeted or replied to, and it reflects the second and third assumptions. Incorporating the matrix A_s into B_a means the UI score of each user is given to all tweets regardless of the user's activities at a certain rate of d. This is based on the idea that the user may see not only the tweets the user retweeted and replied to but also the other tweets, and the tweets posted or retweeted by the user's friends are more likely to be seen than the other tweets since the tweets appear in the user's timeline. It reflects the second and fourth assumptions. If the parameter s in A_s is less than 1, each user gives a larger value to the tweets which appeared in the user's timeline than the tweets which did not appear and, as a result, the tweets seen by many users will get larger value than the tweets seen by a small number of users. If the parameter s is equal to 1, the fourth assumption is ignored.

The UI score and the TI score are calculated using the power iteration method as shown in Figure 2. $\mathbf{u_k}$ and $\mathbf{t_k}$ indicate the UI score and the TI score at the k-th iteration respectively. U and T are a set of user nodes and a set of tweet nodes in the tweet activity relation graph. ε_u and ε_t are error tolerance. Lastly the UI score is normalized so that the largest value should be 1.

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

$\mathbf{u}_0 = (\frac{1}{|U|}, \frac{1}{|U|}, \ldots, \frac{1}{|U|})$;
$\mathbf{t}_0 = (\frac{1}{|T|}, \frac{1}{|T|}, \ldots, \frac{1}{|T|})$;
$k = 1$;
Repeat
$\mathbf{t}_k = B_a^T \mathbf{u}_{k-1}$;
$\mathbf{u}_k = B_t^T \mathbf{t}_k$;
$k = k + 1$;
until $|\mathbf{u}_k - \mathbf{u}_{k-1}| < \varepsilon_u$ and $|\mathbf{t}_k - \mathbf{t}_{k-1}| < \varepsilon_t$;
return $\mathbf{u}_k$ and $\mathbf{t}_k$;

Fig. 2. Calculation of the UI score and the TI score

3.3. *Follow Relation (FR) Score*

The FR score is calculated based on the follow relation using PageRank [19]. A reference graph consisting of user nodes and directed edges each of which connects two of the user nodes, called "follow relation graph", is created from the follow relation. The graph is represented as the following adjacency matrix.

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

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

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

u_i and u_j indicates the i-th user and the j-th user respectively, and d is a damping factor. U is a set of user nodes in the follow relation graph, and $\mathbf{f}$ is a column vector of the FR score. The FR score is normalized so that the largest value should be 1.

$$\mathrm{FR}(u_i) = \frac{\mathbf{f}(i)}{\max_k \mathbf{f}(k)} \tag{13}$$

4. Searching for Relevant Tweets

In this section, we describe a method for finding tweets relevant to the target topic. The method consists of two parts: calculation of the Voice score and the Impact score of each user based on the past user activities in the preparation phase, and ranking tweets based on who posted, retweeted, or replied to each of the tweets in the main phase.

4.1. *Voice And Impact Score Calculation*

In order to judge the relevance of each tweet to the target topic, we have the following assumptions about tweets relevant to the topic (relevant tweets).

1. The relevant tweets are posted or retweeted by the topic-related users.
2. The relevant tweets are paid attention to (retweeted or replied to) by many topic-related users.

3. Tweets posted, retweeted, or replied to by good topic-related users are more relevant to the topic.

These assumptions are close to the idea of the TI score defined in section 3.2. Tweets with high TI score are paid attention to by users with high UI score, and users with high UI score post and retweet tweets with high TI score. However, we need tweets posted in a certain time period and the follow relation among the users appearing in the tweets to calculate the TI score. Instead, we estimate the TI score of newly-arrived tweets based on the TI score and the UI score derived from the past tweet data.

We have two ideas for estimating the score of a newly-arrived tweet. According to the definition of the TI score, the score of the newly-arrived tweet can be estimated from the UI score of the users who retweeted or replied to the tweet.[b] The other idea of the score estimation is considering the TI score assigned to the past tweets and retweets of the users who posted or retweeted the newly-arrived tweet. For the score estimation, we define two kinds of score of each user, called "Impact" and "Voice". The Impact score is used for the estimation based on the first idea, and the Voice score is used for the estimation based on the second idea.

The Impact score of user u (appearing in the tweets collected in the preparation phase) is defined as follows.

$$\mathrm{Impact}(u) = \begin{cases} \frac{\mathrm{UI}(u)}{|\mathrm{Relate}(u)|+\sigma_i} \times (1-d) + \frac{\mathrm{UI}(u)}{|T|} \times d & \text{if } |\mathrm{Relate}(u)| > 0 \\ \frac{\mathrm{UI}(u)}{|T|} & \text{otherwise} \end{cases} \tag{14}$$

$\mathrm{Relate}(u)$ indicates a set of tweets the user u retweeted or replied to, and T indicates a set of all tweets obtained in the preparation phase (i.e. a set of tweet nodes in the tweet activity relation graph). σ_i is a smoothing parameter ($\sigma_i \geq 0$) and d is the damping factor used in Eq. (8). Some users frequently retweet other users' tweets without taking a moment to read them, and such users have little impact on other users' tweets. The definition of the Impact score (dividing the UI score by the number of tweets the user retweeted or replied to) reflects this idea. The Impact score is used for estimating the TI score from the UI score of the users who retweeted or replied to the tweet based on the definition (Eq. (3)), and we use the unnormalized UI score (each value of the column vector $\mathbf{u}$) for the Impact score calculation.

In the definition of the Voice score of user u (appearing in the tweets collected in the preparation phase), we consider the score for the user's original tweets (Voice_t) and the score for the user's retweets (Voice_r) separately. This is because some users post valuable original tweets related to the target topic and some other users search for and retweet valuable tweets posted by other users (some users do both). Also, we define two versions of the Voice score for each case, called "As-is" version and "Split" version.

$$\mathrm{Voice}_t(u) = \begin{cases} \frac{1}{|\mathrm{Tweet}(u)|+\sigma_v} \sum_{t \in \mathrm{Tweet}(u)} \mathrm{TI}(t) & \text{(As-is)} \\ \frac{1}{|\mathrm{Tweet}(u)|+\sigma_v} \sum_{t \in \mathrm{Tweet}(u)} \frac{\mathrm{TI}(t)}{|\mathrm{Poster}(t)|} & \text{(Split)} \end{cases} \tag{15}$$

$$\mathrm{Voice}_r(u) = \begin{cases} \frac{1}{|\mathrm{Retweet}(u)|+\sigma_v} \sum_{t \in \mathrm{Retweet}(u)} \mathrm{TI}(t) & \text{(As-is)} \\ \frac{1}{|\mathrm{Retweet}(u)|+\sigma_v} \sum_{t \in \mathrm{Retweet}(u)} \frac{\mathrm{TI}(t)}{|\mathrm{Poster}(t)|} & \text{(Split)} \end{cases} \tag{16}$$

[b] To be exact, the TI score of each tweet is calculated not only from the UI score of the users who retweeted or replied to the tweet but also from the UI score of the other users (Eq. (8)).

Tweet(u) and Retweet(u) indicate a set of the user u's original tweets and a set of the user's retweets respectively. Poster(t) is a set of users who posted or retweeted the tweet t. σ_v is a smoothing parameter ($\sigma_v \geq 0$). The Voice_t score is not defined here if the user's original tweet set is empty, and the Voice_r score is not defined if the user's retweet set is empty. We will describe how to deal with the users with the undefined Voice score in the next subsection. The As-is version of the Voice score is the average TI score for the user's original tweets or retweets. The TI score of tweets retweeted or replied to by many users is likely to be high, which means the Voice_r score of users who just retweet only a few tweets retweeted or replied to by many users will be higher than expected. The Split version of the Voice score is presented to avoid the situation. In the Split version, the TI score of each tweet is divided among users who posted or retweeted the tweet.

4.2. *Tweet Ranking*

In the main phase, we collect tweets posted or retweeted by the topic-related users and retweets of the topic-related users' tweets, calculate the tweet relevance score of each tweet from the Voice score and the Impact score, then select the top-M tweets ranked by the tweet relevance score as the tweets relevant to the target topic. The tweet relevance score is defined separately according to the version of the Voice score (the As-is version and the Split version). In the case of using the As-is version of the Voice score, we present three versions of the tweet relevance score. One version considers only the Voice score of the original poster of the tweet (As-is::Original), another version takes the maximum Voice score among the users who posted or retweeted the tweet (As-is::Max), and the last version calculates the average Voice score among the users (As-is::Average).

$$\text{TweetRel}(t) = \alpha \times \text{VR}(t) + (1-\alpha) \times \text{IR}(t) \tag{17}$$

$$\text{VR}(t) = \begin{cases} \text{Voice}(t.user) & \text{(As-is::Original)} \\ \max_{u \in \text{Poster}(t)} \text{Voice}(u) & \text{(As-is::Max)} \\ \frac{1}{|\text{Poster}(t)|} \sum_{u \in \text{Poster}(t)} \text{Voice}(u) & \text{(As-is::Average)} \\ \sum_{u \in \text{Poster}(t)} \text{Voice}(u) & \text{(Split)} \end{cases} \tag{18}$$

$$\text{IR}(t) = \sum_{u \in \text{Related}(t)} \text{Impact}(u) \tag{19}$$

VR(t) and IR(t) are respectively "Voice-based Relevance (VR) score" and "Impact-based Relevance (IR) score". α ranges between 0 and 1. Voice(u) indicates the Voice score of the user u (Voice_t if u is the original poster of the tweet t, and Voice_r if u retweeted t). Poster(t) and Related(t) are a set of users who posted or retweeted the tweet t and a set of users who retweeted or replied to the tweet t respectively. $t.user$ is the original poster of the tweet t. The VR score and the IR score can be considered as the estimated TI score calculated from the Voice score and the estimated TI score calculated from the Impact score respectively.

The Voice score and the Impact score of unknown users (users who did not appear in the tweets collected in the preparation phase) are not defined although they may appear in the newly-arrived tweets. Also, some of the known users have undefined Voice score in the case that Tweet(u) $= \emptyset$ or Retweet(u) $= \emptyset$ in Eqs. (15) and (16). In these cases, the Voice score

and the Impact score are set by using a parameter p ($p \leq 1$) as follows.

$$\mathrm{Voice}_t(u) = p \times \min_{\mathrm{Voice}_t(u') \text{ is defined}} \mathrm{Voice}_t(u') \tag{20}$$

$$\mathrm{Voice}_r(u) = p \times \min_{\mathrm{Voice}_r(u') \text{ is defined}} \mathrm{Voice}_r(u') \tag{21}$$

$$\mathrm{Impact}(u) = p \times \min_{\mathrm{Impact}(u') \text{ is defined}} \mathrm{Impact}(u') \tag{22}$$

Tweets posted, retweeted, or replied to by many topic-related users would be relevant to the topic, while tweets posted, retweeted, or replied to by many unknown users would be irrelevant. Assigning a negative value to the parameter p means penalizing tweets posted, retweeted, or replied to by many unknown users. If p is equal to 0, the unknown users are ignored.

5. Evaluation

In order to evaluate the effectiveness of our method, we carried out two experiments: one experiment is for observing whether tweets relevant to the target topic are ranked higher, and the other experiment is for observing how many relevant tweets which do not include any of the input keywords representing the target topic, called "relevant no-keyword tweets", are found.[c]

5.1. *Experimental Setup*

5.1.1. *Keywords And Tweet Collection*

We selected Japanese keywords (in Japanese characters) representing the following 8 topics as the input query: "nuclear power", "digital book (ebook)", "whaling", "animal test", "dementia", "big data", "Tokyo Olympics", and "euthanasia". We chose these topics since we expect that tweets related to the topics are posted on a daily basis (i.e. they are the persistent topics discussed in Twitter).

We collected the tweet data for the tweet activity relation graph creation, the follow relation graph creation, the topic-related user search, and the calculation of the Voice score and the Impact score in the preparation phrase as follows.

1. Given some keywords representing the target topic, get tweets matching the keywords posted in the last 5 days. Duplicate tweets are removed to exclude spammers who post the same tweets repeatedly.[d] Tweets of less than 10 characters (after removal of the tweet entities) are also removed since short tweets have little information.[e]

2. For each tweet in the tweet set, get the tweet poster's name and user names mentioned

[c] The basic idea of the topic-related user search comes from our previous work and the evaluation is described in [7, 8].

[d] Given two tweets, they are duplicate tweets if the same user posted the both tweets and the texts of the both tweets (after removal of the tweet entities such as user mentions, URLs, and hashtags) are exactly the same.

[e] This threshold is designed for Japanese tweets, and it will be different for different languages. The threshold should be larger in the case of English since average length of English tweets is about 30 characters longer than that of Japanese tweets although the difference in entropy per tweet is small [20].

Table 1. Tweet collection period and keywords used for collecting tweets

Topic	Tweet collection period		Keywords	
nuclear power	Prep. Phase:	Dec. 15-20, 2014	Any:	原子力
	Main Phase:	Dec. 20-22, 2014	None:	–
digital book	Prep. Phase:	Dec. 19-24, 2014	Any:	電子書籍
	Main Phase:	Dec. 24-26, 2014	None:	–
whaling	Prep. Phase:	Dec. 12-17, 2014	Any:	捕鯨
	Main Phase:	Dec. 17-19, 2014	None:	表捕鯨, 裏捕鯨, 回目
animal test	Prep. Phase:	Dec. 12-17, 2014	Any:	動物実験
	Main Phase:	Dec. 17-19, 2014	None:	–
dementia	Prep. Phase:	Dec. 16-21, 2014	Any:	認知症
	Main Phase:	Dec. 21-23, 2014	None:	–
big data	Prep. Phase:	Dec. 15-20, 2014	Any:	ビッグデータ
	Main Phase:	Dec. 20-22, 2014	None:	–
Tokyo Olympics	Prep. Phase:	Dec. 20-25, 2014	Any:	東京五輪, 東京オリンピック
	Main Phase:	Dec. 25-27, 2014	None:	–
euthanasia	Prep. Phase:	Dec. 19-24, 2014	Any:	安楽死, 尊厳死
	Main Phase:	Dec. 24-26, 2014	None:	里親, 殺処分, 動物, 犬, 猫

in the tweet. If the tweet is a retweet or reply tweet, get the poster's name of the tweet retweeted or replied to.

3. Get the follow relation among users in the user set.

The top 50 users ranked by the user relevance score (UserRel) were selected as the topic-related users. The Voice score and the Impact score of all of the users collected in the process were calculated regardless of the topic-related user selection.

When we carried out the tweet collection and the keywords we used for it are shown in Table 1. All of the tweet collection started at 12:00pm JST (at noon) of the first day and finished at 12:00pm JST of the last day. We collected tweets including any of the words in "Any" (i.e. OR search), but tweets including any of the words in "None" were excluded. Two words in "Any" for Tokyo Olympics are synonyms. The first word in "Any" for euthanasia is a translation of euthanasia, and the second word means death of dignity. Although these two words are not exactly synonyms, the words often appear in the discussion on euthanasia. In the case of whaling, the keyword for the topic is also used in an online game, and tweets posted by users who play the game are also collected. The three words in "None" are prepared to exclude such tweets, since these words often appear in the tweets related to the game. The similar situation occurred in collecting tweets on the topic of euthanasia. The first word in "Any" is often used for killing abandoned animals as well as persons suffering from serious illness and so on. The words in "None" are prepared to exclude such tweets related to killing animals.[f]

In the main phase, we collected original tweets and retweets posted by the 50 topic-related users in 2 days following the time period of the tweet data collection in the preparation phase (Table 1). We also collected retweets of the topic-related users' tweets posted in the same time period. As with the tweet data collection in the preparation phase, the duplicate tweets and the short tweets were removed. Also, the tweets retweeting or replying to the

[f]The words in "None" means adopters, killing abandoned animals legally, animals, dogs, and cats respectively.

Table 2. The number of tweets and users (the preparation phase)

	Tweet collection				Reference graph	
Topic	Total	RT	Reply	Mention	Tweet	User
nuclear power	26,420	17,440	799	636	9,374	13,043
digital book	25,689	6,499	1,181	461	19,459	13,118
whaling	4,511	3,187	244	82	1,382	4,176
animal test	11,304	10,317	153	32	1,141	9,985
dementia	9,449	3,609	898	113	5,991	7,650
big data	3,914	2,514	71	88	1,442	3,550
Tokyo Olympics	10,438	4,229	476	305	6,529	7,676
euthanasia	2,046	569	351	23	1,518	1,676

Table 3. The number of tweets (the main phase)

Topic	Total	RT	Reply	Mention
nuclear power	21,677	17,339	1,097	306
digital book	4,268	878	103	8
whaling	3,171	1,965	390	51
animal test	8,198	6,815	311	52
dementia	3,759	2,447	236	61
big data	1,180	488	23	37
Tokyo Olympics	17,082	10,192	441	176
euthanasia	2,318	723	848	36

tweets mentioning the posters themselves and the reply tweets to the posters themselves were removed, as with the tweet activity relation graph creation (section 3.2).

The number of tweets and users collected in the preparation phase, and the number of tweets collected in the main phase are shown in Table 2 and 3. "Reply" means tweets replying to the tweets specified in the "in_reply_to" attribute, and "mention" means tweets including user mentions but not specifying their target tweets.

We carried out a preliminary experiment like the experiment in [8] using tweet data collected in different time period to determine the parameters for calculating the user relevance score (UserRel). w_r, w_i, and w_f in Eq. (1) ranged from 0.20 to 0.50, from 0.30 to 0.60, and from 0.10 to 0.20 respectively so that the sum of the parameters is equal to 1, and s in Eq. (6) ranged from 0.01 to 0.20. d in Eqs. (8) and (11) was fixed to 0.15. Both σ_i and σ_v in Eqs. (14), (15), and (16) were fixed to 1. Although the best parameter setting depended on the topics, we determined the value of each parameter as shown in Table 4.

5.1.2. *Estimation of the total number of tweets*

In calculation of the TR score (Eq. (2)), we need the total number of tweets posted or retweeted by each user during the same time period as the tweet search (i.e. the last 5 days). However, it is not practical to get the value of each user due to the Twitter API rate limit (we need to call the API for getting each user's timeline as many times as the number of the users appearing in the searched tweets). Instead, we estimate the value of each user from the date and time when the user account was created (the "created_at" attribute) and the total number of tweets posted or retweeted by the user during the whole time period (the "statuses_count" attribute), which can be obtained from the user information data.

Table 4. Value of each parameter

w_r, w_i, and w_f in Eq. (1)	0.4, 0.4, and 0.2
s in Eq. (6)	0.1
d in Eqs. (8) and (11)	0.15
σ_i and σ_v in Eqs. (14), (15), and (16)	1 and 1

As we described in section 3.1, the purpose of using the tweet rate instead of the tweet count is to exclude users who post and retweet many tweets on the target topic and much more irrelevant tweets (e.g. general news accounts, bots, etc.). Such users post and retweet tweets regularly, and we can estimate the total number of tweets posted or retweeted by the users. If these users post and retweet not only relevant tweets but also many irrelevant tweets, they will be properly excluded (their TR score will be low). On the other hand, general users (not bots but humans) do not post and retweet tweets regularly, and the value estimated by the method may be different from the actual total number of tweets. In the case of overestimation (the estimated value is larger than the actual total number of tweets), the user's TR score will be lower than expected. In the case of underestimation, the user's TR score will be higher.

One problem is that, in some cases, the estimated value will be less than the number of the user's tweets collected in the preparation phase, and the TR score of the user will be larger than 1. Actually, the estimated value of 989 users (about 1.6% of all users collected) was less than the number of their tweets collected in the preparation phase (the estimated value of 886 users (about 1.5%) was less than 1). In order to avoid the situation, we added 5 to the estimated value of all users collected, which means all users post or retweet at least one tweet in a day. The TR score of 18 users (0.03%) were still larger than 1, and we set the TR score of these users to 1.

5.1.3. *Recalculating the Voice/Impact Score*

In some cases, only a few users have extremely large value of the Voice score and/or the Impact score compared with the other users, and the tweet ranking depends almost exclusively on the users' activities. Table 5 shows the 0%, 10%, 25%, 50%, 75%, 90%, and 100%-quantile of the scores (the 0%, 50%, and 100%-quantile are respectively the minimum, median, and maximum score). For comparison, the scores are normalized so that the median is 1 (all scores are divided by the median, and they indicate the ratio to the median). The maximum scores of Voice_t (As-is version), Voice_r (Split version), and Impact are 1-2 orders of magnitude higher than the 90%-quantile, and we can see that users with such extremely high score will dominate the final ranking result easily. In order to avoid such situation, we dampened the Impact score of the known users (the users who appeared in the tweets collected in the preparation phase) as follows.

$$\text{Impact}(u) \leftarrow -\frac{1}{\log(\text{Impact}(u)) - 1} \tag{23}$$

The Impact score of the unknown users were recalculated by using Eq. (22). The same process is also applied to the Voice score ($\text{Voice}_t(u)$ and $\text{Voice}_r(u)$). The distribution of the new score is shown in Table 6.

Table 5. Quantiles of the Voice score and the Impact score

Voice_t (As-is)	Min	10%	25%	Med	75%	90%	Max
nuclear power	9.594e-01	9.594e-01	9.616e-01	1.000	1.781	4.256	6.902e+02
digital book	9.968e-01	9.968e-01	9.971e-01	1.000	1.330	1.852	1.066e+02
whaling	9.966e-01	9.966e-01	9.966e-01	1.000	1.084	1.871	3.150e+02
animal test	9.923e-01	9.923e-01	9.923e-01	1.000	1.747	4.359	1.152e+02
dementia	9.970e-01	9.970e-01	9.970e-01	1.000	1.329	1.748	4.944e+01
big data	9.757e-01	9.757e-01	9.757e-01	1.000	1.301	1.819	7.464e+01
Tokyo Olympics	9.915e-01	9.915e-01	9.915e-01	1.000	1.382	2.171	8.024e+01
euthanasia	1.000	1.000	1.000	1.000	1.015	1.865	1.752e+01
Voice_t (Split)	Min	10%	25%	Med	75%	90%	Max
nuclear power	3.828e-01	9.841e-01	9.850e-01	1.000	1.314	1.687	1.522e+01
digital book	4.677e-01	9.982e-01	9.982e-01	1.000	1.183	1.500	1.396e+01
whaling	1.263e-01	9.372e-01	9.999e-01	1.000	1.011	1.333	5.639
animal test	6.990e-03	9.997e-01	9.997e-01	1.000	1.261	1.922	8.035
dementia	7.548e-02	9.985e-01	9.985e-01	1.000	1.025	1.337	4.325
big data	1.235e-02	9.877e-01	9.877e-01	1.000	1.078	1.355	3.171
Tokyo Olympics	4.129e-01	9.482e-01	9.972e-01	1.000	1.048	1.418	2.911
euthanasia	3.576e-01	1.000	1.000	1.000	1.006	1.235	5.783
Voice_r (As-is)	Min	10%	25%	Med	75%	90%	Max
nuclear power	1.333e-02	4.474e-01	1.605e-01	1.000	2.614	5.511	8.539
digital book	1.081e-01	1.482e-01	3.256e-01	1.000	3.075	4.992	1.057e+01
whaling	5.496e-03	1.682e-02	1.000	1.000	1.000	1.000	1.000
animal test	3.554e-02	1.000	1.000	1.000	1.000	1.993	2.663
dementia	7.934e-02	1.100e-01	2.732e-01	1.000	2.266	3.162	3.475
big data	9.028e-02	3.774e-01	1.000	1.000	1.000	3.974	3.974
Tokyo Olympics	1.533e-01	2.122e-01	4.517e-01	1.000	3.515	4.818	8.845
euthanasia	2.533e-01	2.594e-01	4.268e-01	1.000	1.813	2.551	2.743
Voice_r (Split)	Min	10%	25%	Med	75%	90%	Max
nuclear power	3.648e-01	6.505e-01	7.198e-01	1.000	1.349	1.753	1.288e+01
digital book	2.367e-01	6.667e-01	7.665e-01	1.000	1.265	1.802	7.135e+01
whaling	1.000	1.000	1.000	1.000	1.000	4.675	6.304e+01
animal test	1.000	1.000	1.000	1.000	1.000	1.016e+01	1.048e+03
dementia	2.571e-01	2.571e-01	4.314e-01	1.000	2.455	3.171	2.799e+01
big data	1.000	1.000	1.000	1.000	1.282e+01	5.500e+01	2.386e+02
Tokyo Olympics	4.594e-01	6.719e-01	7.600e-01	1.000	1.338	1.613	8.099
euthanasia	4.399e-01	4.399e-01	0.746	1.000	1.254	1.814	9.411
Impact	Min	10%	25%	Med	75%	90%	Max
nuclear power	1.122e-04	2.890e-04	6.643e-04	1.000	1.499	2.169	1.111e+02
digital book	3.807e-01	8.126e-01	8.131e-01	1.000	4.489	6.784e+03	1.046e+06
whaling	1.702e-03	1.348e-02	2.713e-02	1.000	1.000	2.993	1.342e+02
animal test	2.062e-03	1.000	1.000	1.000	1.000	1.016e+01	4.105e+03
dementia	3.780e-02	5.000e-01	5.001e-01	1.000	2.872e+02	1.007e+03	2.364e+04
big data	1.631e-03	1.304e-01	1.650e-01	1.000	1.282e+01	3.066e+01	4.679e+02
Tokyo Olympics	1.930e-01	4.662e-01	4.668e-01	1.000	7.977e+02	1.145e+03	2.636e+04
euthanasia	3.576e-01	1.000	1.000	1.000	2.356	5.678e+02	1.074e+04

Table 6. Quantiles of the Voice score and the Impact score (recalculated)

Voice_t (As-is)	Min	10%	25%	Med	75%	90%	Max
nuclear power	9.965e-01	9.965e-01	9.997e-01	1.000	1.051	1.138	2.208
digital book	9.997e-01	9.997e-01	9.998e-01	1.000	1.025	1.055	1.645
whaling	9.996e-01	9.996e-01	9.996e-01	1.000	1.009	1.072	2.600
animal test	9.992e-01	9.992e-01	9.992e-01	1.000	1.058	1.169	1.876
dementia	9.997e-01	9.997e-01	9.997e-01	1.000	1.027	1.055	1.577
big data	9.973e-01	9.973e-01	9.973e-01	1.000	1.029	1.069	1.875
Tokyo Olympics	9.992e-01	9.992e-01	9.992e-01	1.000	1.031	1.077	1.677
euthanasia	1.000	1.000	1.000	1.000	1.002	1.071	1.442
Voice_t (Split)	Min	10%	25%	Med	75%	90%	Max
nuclear power	9.258e-01	9.987e-01	9.987e-01	1.000	1.023	1.046	1.294
digital book	9.400e-01	9.998e-01	9.998e-01	1.000	1.014	1.035	1.284
whaling	8.188e-01	9.931e-01	1.000	1.000	1.001	1.032	1.227
animal test	6.721e-01	1.000	1.000	1.000	1.023	1.069	1.258
dementia	8.049e-01	9.999e-01	9.999e-01	1.000	1.002	1.028	1.159
big data	6.781e-01	9.987e-01	9.987e-01	1.000	1.008	1.034	1.142
Tokyo Olympics	9.247e-01	9.951e-01	9.997e-01	1.000	1.004	1.033	1.109
euthanasia	9.009e-01	1.000	1.000	1.000	1.001	1.023	1.231
Voice_r (As-is)	Min	10%	25%	Med	75%	90%	Max
nuclear power	6.224e-01	6.961e-01	7.955e-01	1.000	1.156	1.315	1.431
digital book	8.042e-01	8.272e-01	8.906e-01	1.000	1.140	1.214	1.348
whaling	4.086e-01	4.681e-01	1.000	1.000	1.000	1.000	1.000
animal test	6.530e-01	1.000	1.000	1.000	1.000	1.123	1.185
dementia	7.494e-01	7.744e-01	8.538e-01	1.000	1.121	1.179	1.197
big data	7.240e-01	8.662e-01	1.000	1.000	1.000	1.280	1.280
Tokyo Olympics	8.183e-01	8.449e-01	9.140e-01	1.000	1.175	1.229	1.348
euthanasia	8.438e-01	8.461e-01	8.970e-01	1.000	1.087	1.144	1.157
Voice_r (Split)	Min	10%	25%	Med	75%	90%	Max
nuclear power	9.220e-01	9.652e-01	9.732e-01	1.000	1.026	1.049	1.273
digital book	8.936e-01	9.676e-01	9.785e-01	1.000	1.020	1.051	1.544
whaling	1.000	1.000	1.000	1.000	1.000	1.156	1.569
animal test	1.000	1.000	1.000	1.000	1.000	1.181	1.850
dementia	8.975e-01	8.975e-01	9.340e-01	1.000	1.082	1.107	1.389
big data	1.000	1.000	1.000	1.000	1.230	1.416	1.670
Tokyo Olympics	9.356e-01	9.660e-01	9.763e-01	1.000	1.026	1.044	1.227
euthanasia	9.208e-01	9.208e-01	9.703e-01	1.000	1.024	1.066	1.307
Impact	Min	10%	25%	Med	75%	90%	Max
nuclear power	5.747e-01	6.013e-01	6.268e-01	1.000	1.034	1.067	1.621
digital book	9.558e-01	9.902e-01	9.902e-01	1.000	1.674	1.731	2.973
whaling	6.450e-01	7.290e-01	7.625e-01	1.000	1.000	1.105	1.733
animal test	7.121e-01	1.000	1.000	1.000	1.000	1.179	2.193
dementia	8.459e-01	9.629e-01	9.629e-01	1.000	1.460	1.625	2.274
big data	6.827e-01	8.715e-01	8.846e-01	1.000	1.227	1.330	1.802
Tokyo Olympics	9.171e-01	9.597e-01	9.598e-01	1.000	1.580	1.632	2.270
euthanasia	9.395e-01	1.000	1.000	1.000	1.057	1.658	2.386

5.2. *Experiment 1: Tweet Ranking Based on Tweet Relevance Score*

5.2.1. *Evaluation Metric*

In order to observe whether tweets relevant to the target topic are ranked higher, we evaluated the top 50 tweets ranked by the tweet relevance score. In the case that two or more than two tweets tied the tweet relevance score, they were sorted in reverse chronological order (i.e. from the latest one to the earliest one). α in Eq. (17) and p in Eqs. (20), (21), and (22) ranged from 0.0 to 1.0 and from -20 to 1.0 respectively.

The evaluation was done with respect to normalized discounted cumulative gain (nDCG) [21].

$$\mathrm{DCG}_{50} = rel_1 + \sum_{i=2}^{50} \frac{rel_i}{\log_2 i} \tag{24}$$

$$\mathrm{maxDCG}_{50} = 2 + \sum_{i=2}^{50} \frac{2}{\log_2 i} \tag{25}$$

$$\mathrm{nDCG}_{50} = \frac{\mathrm{DCG}_{50}}{\mathrm{maxDCG}_{50}} \tag{26}$$

rel_i indicates the relevance between the tweet ranked the i-th and the target topic, which is judged on a scale of 0 to 2. Tweets relevant to the topic were assigned the score of 2, while irrelevant tweets, tweets with no concrete content, and tweets for affiliate advertising were assigned the score of 0. Tweets indirectly relevant to the topic were assigned the score of 1. For example, in the case of the topic "nuclear power", tweets about renewable energy (in the context of comparison with nuclear power) were assigned the score of 1. In the case of the topic "Tokyo Olympics", tweets calling for a boycott of the Tokyo Olympics over the reason which is not directly related to the event were assigned the score of 1, while tweet calling for a boycott of the event over no clear reason were assigned the score of 0.[g] The judgment was done by one of the authors and one member in our research group. In the judgment, we shuffled the tweets in order to make it difficult for the annotators to guess by which method each of the tweets is ranked high.

5.2.2. *Result*

Figure 3 shows the average nDCG value of each version of the tweet relevance score calculation in the case of changing the parameter α from 0.0 to 1.0 (the parameter p is fixed to -3.0). Note that the VR score in the tweet relevance score calculation is ignored in the case that α is equal to 0.0, which means any version of the score calculation will get the same result. We can see that the Split version and the As-is::Original version achieved the best performance in the case that α is equal to 0.0, and the As-is::Max version and the As-is::Average version got a slightly better result in the case that α is around 0.4. The result of any version in the case that α is equal to 0.0 is better than the result in the case that α is equal to 1.0, which

[g] This tweet data set with the relevance annotation is available at
`https://dl.dropboxusercontent.com/u/5107943/TwData/index.html`

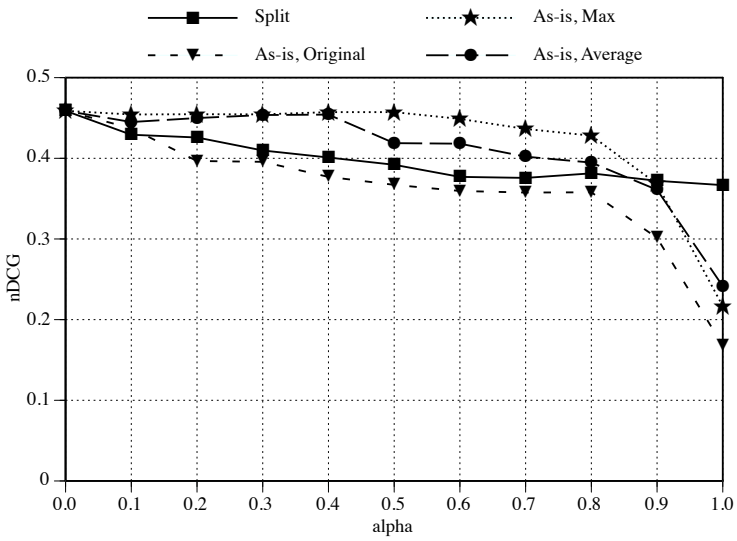

Fig. 3. Average nDCG ($p = -3.0$)

indicates that "who retweeted or replied to the tweet" is more effective than "who posted the tweet" for judging the relevance of tweets to the target topic.[h]

The average nDCG value of each version in the case of changing the parameter p from -20 to 1.0 (the parameter α is fixed to 0.0, 0.4, and 1.0) is shown in Figure 4, 5, and 6. We focus on this range for the parameter p since the result was almost unchanged in the case that p is less than -20. If we set smaller value to p, tweets retweeted and replied to by many unknown users will be ranked lower, and only tweets retweeted and replied to the known users (appearing in the tweets collected in the preparation phase) will remain high in the ranking result.[i] As mentioned previously, any version will get the same result in the case that α is equal to 0.0. From the result, we can see a significant drop in the performance of any version when p is more than 0.0 (except for the case that α is equal to 1.0), which indicates that the unknown users' activities should be considered as negative factors. In the case that α is equal to 1.0, only the VR score is considered in the tweet relevance score calculation. In this case, the As-is::Original version and the As-is::Max version calculate the VR score from the only one user's Voice score. As a result, any tweet will be ranked high if the tweet is posted or retweeted by one user with high Voice score (regardless of the other users who posted or retweeted the tweet), and the performance does not depend on the value of p. Even in the case of the As-is::Average version, since the version takes the average Voice score, tweets posted by the users with high Voice score and retweeted by no user tend to be ranked higher and

[h] Although the VR score and the Voice score consider not only "who originally posted the tweet" but also "who retweeted the tweet", we can say that "who originally posted the tweet" is not so effective compared with "who retweeted the tweet", since the performance of the As-is::Original version was worse than that of any other version.

[i] The upper bound of the range for p is set to 1.0 due to the definition (section 4.2). If p is larger than 1.0, the Voice score and the Impact score of unknown users will be higher than the scores of some of the known users appearing in the tweets collected in the preparation phase.

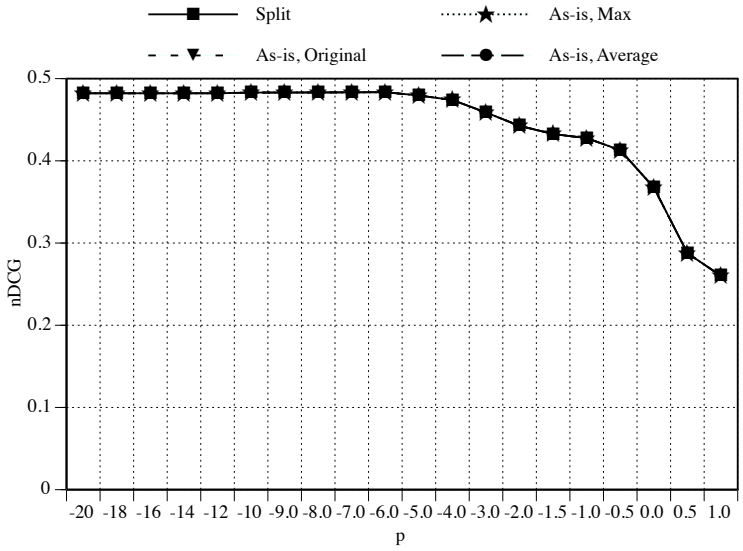

Fig. 4. Average nDCG ($\alpha = 0.0$)

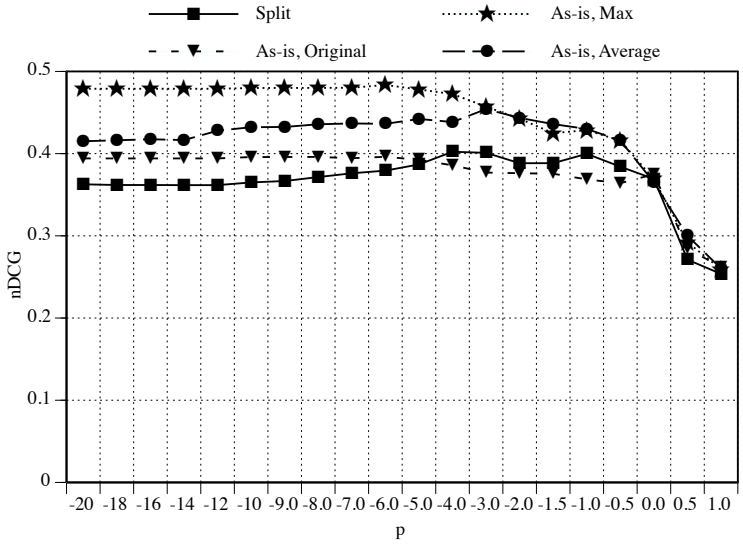

Fig. 5. Average nDCG ($\alpha = 0.4$)

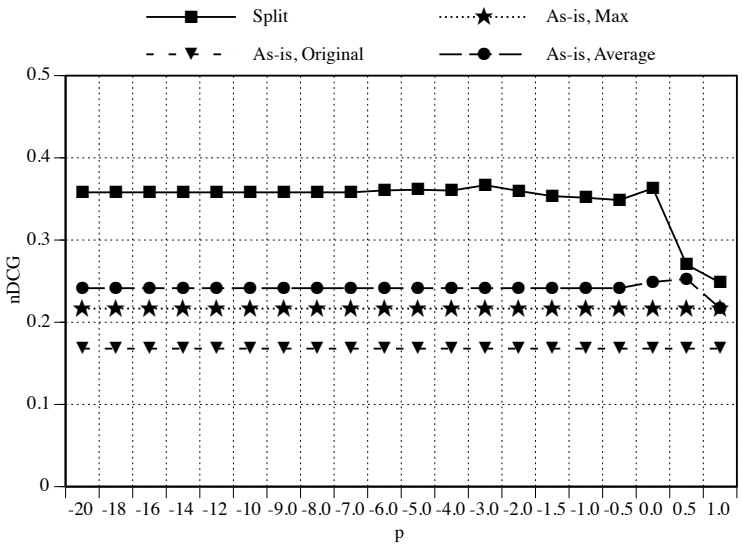

Fig. 6. Average nDCG ($\alpha = 1.0$)

the performance does not depend largely on the value of p (the average Voice score of tweets retweeted by many users will drop due to the users with low Voice score). As we described in section 4.1, the As-is version of the Voice score of a user tends to be higher than expected if the user just posts only a few tweets retweeted or replied to by many users. The performance of the Split version is better than any other version since the Voice score of such users is reduced and the VR score of a tweet is calculated by summing up the Voice score of all users who posted or retweeted the tweet.

5.2.3. *Comparison with Other Methods*

We compared the performance of our method with some other ranking methods.

- **Retweet and reply count (RT)**, which ranks the collected tweets in descending order of the number of times retweeted or replied to.
- **Tweet Influence score (TI)**, which ranks the collected tweets in descending order of the TI score defined in section 3.2.

In both of the methods, if two or more than two tweets have the same score, they are ranked in reverse chronological order. Indegree and PageRank are often used for graph-based ranking approaches. The RT method can be considered as an indegree-based ranking method. Instead of calculating the standard PageRank from a reference graph of tweets, we employ the TI method since the basic idea of the TI score is related to that of the standard PageRank (the tweet activity reference graph consisting of tweets and users is used for simultaneous calculation of the TI score and the UI score).

Figure 7 shows the average nDCG value of our method (the parameter p is fixed to -3.0, and α is 0.0, 0.4, or 1.0), the RT method, and the TI method. From the result, we can see

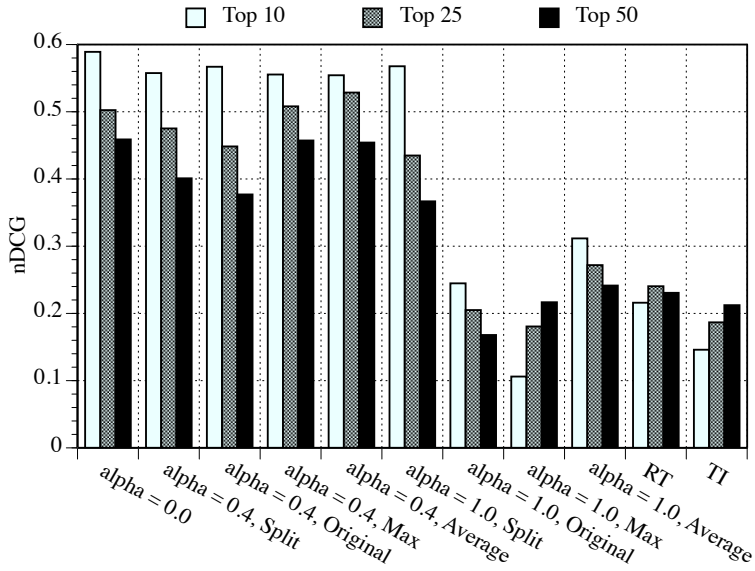

Fig. 7. Average nDCG at the top 10, 25, and 50 tweets ($p = -3.0$)

that our method outperforms the RT method and the TI method in the case that α is equal to 0.0 and 0.4. In the main phase, we collected the tweets posted or retweeted by the topic-related users selected in the preparation phase. The topic-related users do not always post, retweet, or reply to tweets on the target topic. Some irrelevant tweets many users retweeted or replied to are ranked high by the RT method since the method considers only the number of times retweeted or replied to. Although the TI method considers who tweeted, retweeted, or replied to what, the same problem still remains since it does not consider whether they are related to the target topic.[j] Our method can exclude some of the irrelevant tweets since it considers both of them. On the other hand, if we use only the VR score for the tweet relevance score calculation (i.e. when α is equal to 1.0), the difference of the performance between our methods and the methods compared is small (the performance of our method is even worse) except for the case of the Split version.

We also observed correlation among the ranking results by using the Kendall's τ rank correlation coefficient [22] and the Jaccard coefficient [23]. The Kendall's τ rank correlation coefficient measures the association between two ranking results. It takes the whole result (from the top to the bottom) into calculation although, for our task, only high-ranked tweets should be considered and the actual order of low-ranked tweets is not important. The Jaccard coefficient is also employed for this reason. We take the top 10, 25, 50, and 100 tweets of each ranking result for calculation of the Jaccard coefficient.

The Kendall's τ values and the Jaccard coefficient values among our method (the parameter p is fixed to -3.0, and α is 0.0 or 0.4), the RT method, and the TI method are shown in Table 7 and 8. Values greater than 0.5 are indicated in boldface type. In general, both the

[j] In the main phase, we collect not only original tweets and retweets posted by the topic-related users but also retweets of the topic-related users posted by other users, as described in section 5.1.1.

Table 7. Kendall's τ correlation coefficient between ranking results

		$\alpha = 0.4$				RT	TI
		Split	Original	Max	Average		
$\alpha = 0.0$		0.206	0.161	**0.554**	0.220	**0.651**	0.321
$\alpha = 0.4$	Split	–	**0.811**	0.469	**0.840**	-0.086	-0.007
	Original	–	–	**0.536**	**0.887**	-0.085	-0.008
	Max	–	–	–	**0.616**	0.278	0.348
	Average	–	–	–	–	-0.059	-0.028
RT		–	–	–	–	–	**0.505**

Table 8. Jaccard coefficient between ranking results

Top 10 tweets		$\alpha = 0.4$				RT	TI
		Split	Original	Max	Average		
$\alpha = 0.0$		0.451	0.403	**0.852**	**0.673**	0.043	0.041
$\alpha = 0.4$	Split	–	**0.636**	0.490	**0.694**	0.027	0.036
	Original	–	–	0.442	**0.558**	0.027	0.034
	Max	–	–	–	**0.749**	0.043	0.041
	Average	–	–	–	–	0.034	0.049
RT		–	–	–	–	–	0.336

Top 25 tweets		$\alpha = 0.4$				RT	TI
		Split	Original	Max	Average		
$\alpha = 0.0$		0.372	0.341	**0.767**	**0.552**	0.097	0.099
$\alpha = 0.4$	Split	–	**0.543**	0.404	**0.680**	0.066	0.063
	Original	–	–	0.365	**0.550**	0.065	0.073
	Max	–	–	–	**0.583**	0.097	0.104
	Average	–	–	–	–	0.092	0.084
RT		–	–	–	–	–	0.357

Top 50 tweets		$\alpha = 0.4$				RT	TI
		Split	Original	Max	Average		
$\alpha = 0.0$		0.378	0.329	**0.858**	**0.586**	0.110	0.158
$\alpha = 0.4$	Split	–	**0.616**	0.368	**0.587**	0.058	0.060
	Original	–	–	0.319	**0.561**	0.057	0.097
	Max	–	–	–	**0.595**	0.108	0.145
	Average	–	–	–	–	0.098	0.087
RT		–	–	–	–	–	0.357

Top 100 tweets		$\alpha = 0.4$				RT	TI
		Split	Original	Max	Average		
$\alpha = 0.0$		0.445	0.368	**0.759**	**0.511**	0.154	0.180
$\alpha = 0.4$	Split	–	**0.710**	0.452	**0.802**	0.119	0.103
	Original	–	–	0.379	**0.720**	0.122	0.121
	Max	–	–	–	**0.530**	0.151	0.159
	Average	–	–	–	–	0.130	0.104
RT		–	–	–	–	–	0.397

Table 9. No-keyword tweet ratio ($p = -3.0$)

$\alpha = 0.4$				$\alpha = 0.0$
Split	Original	Max	Average	
0.762	0.716	0.737	0.722	0.730
(125/164)	(106/148)	(146/198)	(140/194)	(143/196)

Kendall's τ value and the Jaccard coefficient value between our method and each of the methods compared (RT and TI) are less than the values among the variations of our method. They are also less than the values between the RT method and the TI method. This means that the ranking result of our method has less agreement with the result of the RT method and the TI method, compared with the agreement between the RT method and the TI method, since our method considers whether the users who retweeted or replied to each tweet are related to the target topic while the RT method and the TI method do not. The Kendall's τ value between our method in the case that α is 0.0 and the RT method is high. This is because our method considers only the IR score in the case that α is 0.0, which is based on the retweet and reply activities. However, the Jaccard coefficient value between our method in the case that α is 0.0 and the RT method compared is low. We can see that the high-ranked tweet set generated by our method differs from the set generated by the RT method.

5.3. *Experiment 2: Finding Relevant No-keyword Tweets*

One feature of our method is that it can find relevant tweets which do not include any terms explicitly related to the target topic. In order to evaluate the feature, we measured "no-keyword tweet ratio", how many relevant tweets which do not include any of the input keywords given in the preparation phase (we refer to the tweets as "relevant no-keyword tweets") are found. The no-keyword tweet ratio is defined as the ratio of the number of the detected relevant no-keyword tweets to the number of all of the detected relevant tweets (i.e. true positives).

Table 9 shows the no-keyword tweet ratio in the case that α is equal to 0.0 and 0.4 (p is fixed to -3.0). From the result, we can see that 71%-76% of the detected relevant tweets do not include any input keywords.

We investigated the detected relevant no-keyword tweets for more details. In the case of "whaling", some users frequently post tweets countering the activities by some anti-whaling groups such as the Sea Shepherd. However, they do not include the terms explicitly representing the topic every time they post such tweets since they usually discuss the topic. Also, although they usually post Japanese tweets, they sometimes argue with the foreign activists in English. Some of the English tweets could be found by our method, while it is difficult for the keyword-based search using Japanese keywords to find relevant tweets written in the other languages.

We found the similar situation in the case of "dementia". Some users who have family members with dementia post tweets about their daily care, and they do not always say "demented mother" and so on. Our method can find such a tweet if some users with high Voice score and/or high Impact score posted, retweeted, or replied to the tweet.

In the case of "big data", some relevant no-keyword tweets about prediction of the Japan's

parliamentary election by the big data (social media) analysis were found. The parliamentary election is not always related to the topic of big data. We think it is difficult for the query expansion technique using some language resources such as DBpedia and YAGO to deal with such dynamic (temporal) relation to the target topic.

5.4. *Discussion And Future Direction*

According to the result of the first experiment, the nDCG value of any version in the case that α is equal to 0.0 was higher than the value in the case that α is equal to 1.0. In particular, the value of the As-is::Original version is the lowest in the case that α is equal to 1.0. This means that "who retweeted or replied to the tweet" is more effective for judging the relevance of the tweet to the target topic than "who posted the tweet". Also, considering the activities of the unknown users as a negative factor is effective (the nDCG value dropped when the parameter p is positive). Our method outperforms the RT method and the TI method, since our method considers both who tweeted, retweeted, or replied to each tweet and whether each user is related to the target topic.

5.4.1. *Dealing with the Activities of the Unknown Users*

Why should the activities of the unknown users be considered as a negative factor? Not so many users post and retweet tweets limited to only one topic. Even the topic-related users detected in the preparation phase do not always post and retweet tweets about the target topic. The group of users interested in the target topic and the group of users interested in another topic are different (some users may belong to the both groups). Tweets on the target topic will be retweeted and replied to by the users in the former group, while tweets on the other topic will be retweeted and replied to by the users in the latter group. In order to distinguish the tweets on the target topic from the others, considering the activities of the unknown users as a negative factor is effective, since it is likely that the unknown users do not belong to the group of users interested in the target topic but belong to the other group. From the experimental results, the case where negative value is set to p achieved the better performance than the case where positive value is set to p. It is true that some of the unknown users may be interested in the target topic (but unfortunately they were not found in the preparation phase). Actually, if p is too small, the performance of our method slightly decreased. This is caused by the unknown users who are interested in the target topic, i.e. the penalty given by the unknown users was too strong.

For example, one university professor in economics, who are interested in issues on nuclear power, was detected as a topic-related user of the topic "nuclear power". Although he usually post tweets related to the topic and he is certainly a topic-related user, he sometimes post tweets on other topics such as financial issues of the government, corruption scandals, and so on. He is popular, and his tweets are often retweeted and replied to by many users regardless of the topic. However, the users who retweet or reply to tweets on nuclear power and the users who retweet or reply to tweets on another topic such as financial issues are different. The users who are interested in the topic of nuclear power (let the user group be "A") are likely to retweet and reply to the tweets on the topic of nuclear power, and the users who are interested in the topic of financial issues of the government (let the user group be "B") are likely to retweet and reply to the tweets on the topic of financial issues. The users in the group A are

Table 10. The number of duplicate tweets

Topic	duplicate	individual	users
nuclear power	4,114	888	459
digital book	39,204	5,428	1,254
whaling	317	71	59
animal test	1,662	274	123
dementia	4,058	608	382
big data	334	67	54
Tokyo Olympics	7,426	1,392	831
euthanasia	404	97	86

the known users and the users in the group "B" (but not in the group "A") are the unknown users. If a tweet is retweeted or replied to by many users in the group A but not in the group B, the tweet may be related to the topic of nuclear power, while, if a tweet is retweeted or replied to by many users in the group B but not in the group A, the tweet may be related to the topic of financial issues. In order to distinguish this situation, considering the activities of the unknown users as a negative factor is effective. If the parameter p is positive, the score of the tweets on the topic of financial issues will be inflated by the unknown users (the users in the group B but not in the group A). If p is negative, the activities of the unknown users will penalize the tweets on the topic of financial issues. As a result, the case where negative value is set to p outperforms the case where positive value is set to p.

5.4.2. *Outliers*

As we described in section 5.1.3, some users have extremely large value of the Voice and the Impact score (before recalculation) compared with other users. They may be outliers. Some of them are bots or generally popular users such as news accounts, well-known company accounts, celebrities and so on. Bots automatically post and retweet a lot of tweets. Generally popular users are not necessarily the topic-related users, but their tweets (on any topics) are likely to be retweeted and replied to by many users. In order to exclude the influence of bots, we removed duplicate tweets from the collected data as described in section 5.1.1. Bots often post the same tweets repeatedly, and we can weaken their influence. Table 10 shows the number of duplicate tweets removed from the tweets collected in the preparation phase. The second and the third columns indicate the total number of tweets detected as duplicate tweets and the number of individual duplicate tweets. The fourth column indicates the number of users who posted duplicate tweets.

On the other hand, the influence of the generally popular users' tweets irrelevant to the target topic can be weakened by considering the activities of the unknown users as a negative factor. Tweets posted by the generally popular users are likely to be retweeted and replied to many unknown users if the tweets are not related to the topic.

One practical solution to dealing with such outliers would be that we prepare a list of bots and generally popular users then remove their activities from the collected data. However, not only bots and generally popular users are outliers. In some cases, "temporary active users" will be outliers. Although they do not usually post and retweet tweets related to the topic, they actively interacted with some users about the topic in the tweet collection period. The

Voice and the Impact score of such users will be high, but they will not post any tweets related to the topic after finishing the interaction. They are neither bots nor generally popular users, and we cannot prepare a list of such users beforehand. Instead, we use the recalculated Voice and Impact score. The influence of such users will be weakened since not many users with high Voice and/or Impact score will retweet or reply to their tweets irrelevant to the topic.

5.4.3. *Comparison with Related Works*

As we mentioned in section 2, there are several methods for ranking or searching for tweets relevant to a given query. Twinder [14] uses the keyword-based relevance features and the semantic-based relevance features. The keyword-based features are extracted based on word occurrence, and DBpedia Spotlight, an ontology-based named entity recognizer, is used for extracting the semantic-based relevance features. Duan et al. [15] also use the content relevance features based on TF-IDF and Okapi BM25. However, these methods which consider the content-based features have a limitation that they fail to find relevant tweets which do not include any terms explicitly related to the given query. The relevance score of each tweet based on TF-IDF and Okapi BM25 will be low in the case that the tweet does not include any terms related to the query. Although Twinder, which uses the query expansion technique based on DBpedia and other language resources, may cope with some of the situation, it still fails to find relevant tweets in the case that they have photos or URL links to external Web pages relevant to the query but does not include any related terms in the tweet texts (the examples shown in section 1). Also, it is difficult for the ontology-based method to deal with temporal relation such as the relation between "big data" and "Japan's parliamentary election" (section 5.3). Contrary to these methods, our method considers the relation between users and tweets (retweet and reply), and can solve the problem, as shown in the second experiment.

5.4.4. *Selection of the Target Topic And Keywords*

Our method has some limitations. Our method does not work well if the user interaction is not active (i.e. if the number of retweets and reply tweets is too small) since it depends on the tweet, retweet, and reply activities. If there are some active users who are interested in and usually discuss the topic, our method works effectively. As we mentioned in section 1, our research focuses on the persistent topics discussed in Twitter on a daily basis. In the case of such topics, we can expect to find some active topic-related users.

The keyword selection in the preparation phase may influence the final result. If we select keywords used in multiple topics, we could not detect the topic-related users properly. For example, as we described in section 5.1.1, the keyword selected for the topic of whaling was used not only in the topic of whaling but also in the topic of an online game. As a result, tweets related to whaling and the online game will be mixed together. In order to avoid the situation, we prepared "negative keywords", and excluded tweets including any of the negative keywords in the preparation phase.

6. Conclusion

General content-based keyword search techniques and query expansion techniques are not

effective for finding relevant tweets which do not include any terms explicitly related to the topic. To solve this problem, we presented a method for finding tweets on a topic of interest based on the user activities such as tweet, retweet, and reply. Our method consists of the preparation phase and the main phase. In the preparation phase, we first collect tweets matching the query representing the target topic, then calculate the Voice score and the Impact score of each user from the influence of each user and tweet based on the user activities and find the topic-related users by considering the influence of each user based on the user activities and the follow relation. In the main phase, we collect tweets posted or retweeted by the topic-related users, then calculate the tweet relevance score of each tweet by using the Voice score and the Impact score of the users who posted, retweeted, or replied to the tweet. The two phases are processed independently. Once the preparation phase is completed, the main phase can be processed any time. The experimental results showed that "who retweeted or replied to the tweet" is more effective for judging the relevance of the tweet to the topic than "who posted the tweet", and our method can find relevant tweets which do not include terms explicitly related to the topic.

References

1. P.B. Brandtzæg and J. Heim. *Why People Use Social Networking Sites.* In *3rd Int. Conf. on Online Communities and Social Computing*, pages 143–152, 2009.
2. A. Smith. *Why American Use Social Media*, 2011. `http://www.pewinternet.org/2011/11/15/why-americans-use-social-media/`.
3. F. Gruber. *Why User Social Media in the First Place?*, 2014. `http://tech.co/why-use-social-media-2014-06`.
4. Twitter. *About Twitter, Inc.* `https://about.twitter.com/company`.
5. J. Hannon, M. Bennett, and B. Smyth. *Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches.* In *4th ACM Conf. on Recommender Systems*, pages 199–206, 2010.
6. J. Weng, E.-P. Lim, J. Jiang, and Q. He. *TwitterRank: Finding Topic-sensitive Influential Twitterers.* In *3rd ACM Int. Conf. on Web Search and Data Mining*, pages 261–270, 2010.
7. T. Noro, F. Ru, F. Xiao, and T. Tokuda. *Twitter User Rank Using Keyword Search.* In *Information Modelling and Knowledge Bases XXIV*, volume 251 of *Frontiers in Artificial Intelligence and Applications*, pages 31–48. IOS Press, 2013.
8. T. Noro and T. Tokuda. *Effectiveness of Incorporating Follow Relation into Searching for Twitter Users to Follow.* In *14th Int. Conf. on Web Engineering*, pages 420–429, 2014.
9. K. Slabbekoorn, T. Noro, and T. Tokuda. *Towards Twitter User Recommendation Based on User Relations and Taxonomical Analysis.* In *Information Modelling and Kowledge Bases XXV*, volume 260 of *Frontiers in Artificial Intelligence and Applications*, pages 115–132. IOS Press, 2014.
10. F. Xiao, T. Noro, and T. Tokuda. *Finding News-Topic Oriented Influential Twitter Users Based on Topic Related Hashtag Community Detection. Journal of Web Engineering*, 13(5&6):405–429, 2014.
11. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. *DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal*, 6(2):167–195, 2015.
12. F.M. Suchanek, G. Kasneci, and G. Weikum. *YAGO: A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Web Semantics*, 6(3):203–217, 2008.
13. F. Mahdisoltani, J. Biega, and F.M. Suchanek. *YAGO3: A Knowledge Base from Multilingual Wikipedias.* In *7th Biennial Conf. on Innovative Data Systems Research*, 2015.
14. K. Tao, F. Abel, C. Hauff, and G.-J. Houben. *Twinder, A Search Engine for Twitter Streams.* In *12th Int. Conf. on Web Engineering*, pages 153–168, 2012.

15. Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. *An Empirical Study on Learning to Rank of Tweets*. In *23rd Int. Conf. on Computational Linguistics*, pages 295–303, 2010.
16. A. Singhal, C. Buckley, and M. Mitra. *Pivoted Document Length Normalization*. In *19th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 21–29, 1996.
17. I. Uysal and W.B. Croft. *User Oriented Tweet Ranking: A Filtering Approach to Microblogs*. In *20th ACM Int. Conf. on Information and Knowledge Management*, pages 2261–2264, 2011.
18. M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *4th International AAAI Conf. on Weblogs and Social Media*, pages 10–17, 2010.
19. L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical report, Stanford University, 1998.
20. G. Neubig and K. Duh. *How Much Is Said in a Tweet?* In *AAAI 2013 Spring Symposium on Analyzing Microtext*, pages 32–39, 2013.
21. K. Jarvelin and J. Kekalainen. *Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems*, 20(4):422–446, 2002.
22. M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.
23. P. Jaccard. The distribution of the flora in the Alpine zone. *New Phytologist*, 11(2):37–50, 1912.