Received September 20, 2013；Accepted December 27, 2013

International Journal of Asia Digital Art&Design

# A Method for Composing Ad-hoc Following Networks on Twitter for Sharing Information among Event Participants

Shingo, Tajima
Graduate School of Design, Kyushu University
2ds13084s@s.kyushu-u.ac.jp

Taketoshi, Ushiama.
Facility of Design, Kyushu University
ushiama@design.kyushu-u.ac.jp

**Abstract**
Twitter is one of the most popular micro blog services in the world. It has a large number of users and posts. One of the reasons why so many people uses Twitter is that they can use it easily and can obtain many kinds of information in real time. On Twitter, information is delivered via following networks. The following networks on Twitter are static and have to be specified manually by each user, so a user cannot share information that he/she wants if that was out of his/her following networks. On the other hand, it often happens that users want information about an event such as festival or sport game when they are attending it. In this paper, we propose a method for composing ad-hoc following networks automatically among users who are attending the same event for sharing information about it.

**Keywords:** Twitter, Ad-hoc following network, information sharing

## 1 Introduction

Micro blog systems, which are one of SNS (Social Networking Service), are becoming popular in these days. In Twitter, users can send messages which are called “tweets” within 140 characters and the messages are sent to their followers. For example, if a user (“A”) follows another user (“B”), tweets of “B” are sent to “A” in real time. Moreover, each user can select the followees (twitter users the user is following) without approval of them. Therefore a user can obtain information about topics that the user is interested in by following other users who are also interested in the same topics.

The following network of a user is generally composed of family, friends, and/or acquaintances and it is static. Therefore it does not change dynamically depending on purposes of the user. In other words, Twitter can only support to share information within a closed community. Of course, it is very useful as a daily communication tool, but it would be sometimes inconvenience when following situations:

・ when user want information about an event such as concert or festival which he/she is attending,

・ when user want information about a sport game which he/she is watching, and

・ when user want information about an accident or disaster which he/she is involved.

In these situations (we call it “event” in this paper), a user can’t obtain information about the event that the user is attending because the followees of the user would not attend the same event. When user tries to obtain information about the event in real time, popular solutions are to use BBS or to use other SNS that provide community functions. However, Twitter has a larger number of posts than other BBSs and SNSs. For example, about 2,500 messages (tweets) were posted on Twitter related to “JOIN ALIVE 2012”, which is a summer music festival in Japan. On the other hand, only about 200 messages were posted on 2ch, which is the most popular BBS in Japan. Therefore, we think that it is effective to use Twitter for obtaining information about an event from its participants.

In order to obtain information about a topic, Twitter provides a search function. Using the search function, users can find tweets that contain the words that they send as a query. However, it is very difficult and ineffective for users to check all of the search results because they would contain a lot of tweets, so effective solutions for selecting important tweets are required.

In order to gather tweets that concern a specific topic, the hash tag system is widely used on Twitter. Hash tags are words or phrases prefixed with the symbol #. They could be included in tweets and works as tags on the tweets. However, all of tweets do not contain the hash tags because they have to be attached to tweets explicitly by users. Also, the tweets that concern to the same topic do not always have the same hash tag. For example, different hash tags “#JOIN_ALIVE” and “#JOIN_ALIVE_2012” were used for sharing information about the JOIN ALIVE 2012.

For those reasons, we propose a system that can construct following networks (we call it ad-hoc twitter following network) with users who are attending the same event (we call it a *reporter*). Using our system, you can obtain information about an event on twitter easily and efficiently by following the reporters. Figure 1 shows an example of a conventional follow network on Twitter. On the other hand, Figure 2 shows an example of an ad-hoc following network that is provided by our system.

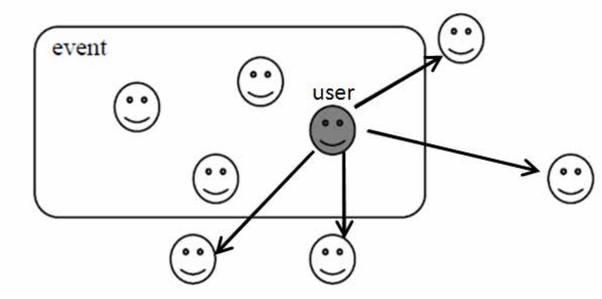


Figure 1: Conventional following network on Twitter

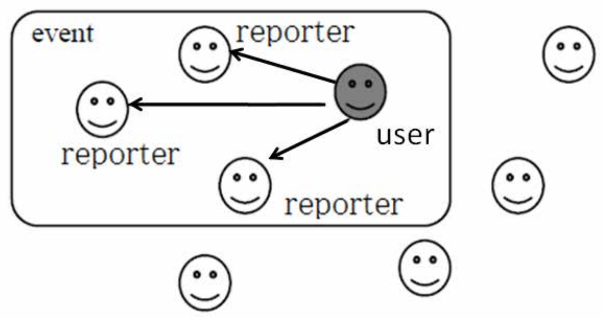


Figure 2: Ad-hoc follow network on Twitter

Figure 3 shows an overview of our system. The system works as follows.

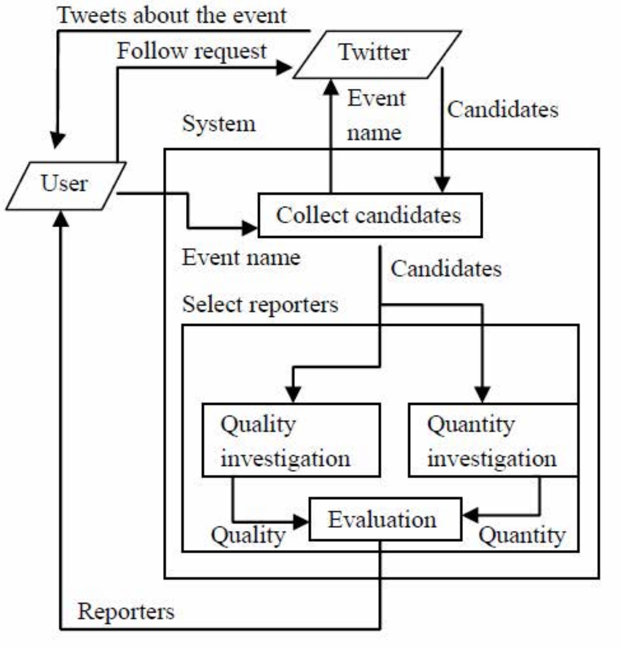


Figure 3: Overview of our system

(1) A user inputs the name of an event (e.g. the name of a festival, a sport game, or a TV program) as a query.

(2) The system finds candidates of reporters from Twitter by using the query.

(3) The system selects conclusive reporters, who post valuable tweets about the event frequently. In order to select the reporters from the candidates, the system evaluates quality and quantity of each candidate's posts. We will explain it in more detail in section 3.

(4) The system recommends the reporters to the user. By following one or more of the recommended reporters during the event, they can compose a follow network for the event temporally and obtain information about the event in real time.

## 2 Related Works

In the last few years, many researchers analyzed Twitter and tried to discover some unknown features in human society such as trend or public opinion, etc. The reason why they use Twitter is that it deals with overwhelming information. Twitter is the most appropriate social media to statistically analyze trend or opinion of the public. Bernardo et al. [6] researched about the correlation between the number of follow/follower on Twitter and the number of real friends. Also, Danah et al. [7] classify the usage of retweets by its purpose. These researches are types of studying about the characteristic of Twitter itself.

On the other hand, there are also many researches that detecting a phenomenon or tendency of society by analyzing Twitter. Wakamiya et al. [1] proposed a Twitter-based estimation of behavior of TV audience for better TV ratings. They analyzed tweets for specifying the users who were watching a certain TV program. Then they estimated the TV ratings from the tweets of the users. They said that the advantages of using Twitter are followings:

(1) people can use Twitter with almost no cost, and

(2) people can obtain not only the TV ratings but also opinions or sentiments of TV audience simultaneously.

Similarly, Jansen et al. [2] proposed an approach to gather consumer opinions concerning a brand, and Akcora et al. [9] utilized Twitter to identify breakpoints and capture trends in public opinion.

Moreover, Twitter is also used to detect events promptly. For example, Sakaki et al. [3] introduced a system that can detect earthquakes by monitoring tweets, and Aramaki et al. [4] proposed an approach to detect Influenza Epidemics using natural language processing techniques. Most of these researches utilized SVM [8] in order to classify the tweets.

Sugitani et al. [5] presented a technique for detecting local events by analyzing tweets regionally. They implemented a system that can detect not only big events but also small local events by observing tweets in real time. The advantages of using Twitter, they said, there is time lag between the event occurring time and the time when some articles or blogs about it are uploaded in most cases. It means that some articles and blogs are uploaded after the event. On the other hand, most of tweets about the event are posted in real time. They focused on the fact that most of tweets about an event are posted intensively at the time and place, so they executed clustering for the tweets about time and place. If some peculiar words would be detected in the cluster, they supposed it as a related word about the event and judged that there is an event at the place.

## 3 Method

In this paper, we focus on recommending twitter users (reporters) rather than tweets. The reasons are followings.

- If the system detects important tweets about an event, the system has to search tweets that include characteristic features of the event and select some important tweets from them. In such case, if a tweet does not have any related words, it could not be recommended even if it has important information about the event. In fact, there are a lot of such cases in actual events and twitter timelines. Thus, the above approach would tend to miss important tweets.
- It requires much time to estimate the importance of each tweet every time when user requests. On the other hand, if the user follows the recommended reporters once, they can obtain tweets that have information about an event without consulting the system.

### 3.1 Collecting reporter candidates from Twitter

First step is collecting Twitter users who are supposed to be participating in the event and post tweets about the event. The most typical criterion for evaluation is whether the user posts tweet that includes the name of the event. Hence, the system searches tweets that include the name of event with Twitter API. For example, when user posts query “oman” during the soccer game “Japan vs Oman” which was played on November, 4th 2012, the search results would be like “Oman is leading by one point!” or “Japan vs Oman, so exciting!” However, there may be many tweets that are related to the same soccer game even though they don’t include the word “oman” itself. For example, the tweet “I’m watching the world cup elimination round, Japan is leading by one point!” doesn’t contain the word “oman”, but it can be considered to be related to the same game. Like this, there is a possibility that some users will be candidates even though they don’t post tweets that contain the query word itself. In order to solve this problem, the system uses related words. For example, in this case, if the system specifies related words such as “World Cup” or “elimination round”, the users who posted tweets that contain the related words can be added to the candidates.

### 3.2 Specifying related words

At first, the system collects co-occurring words with the query word in each searched tweet. Fig 4 shows the co-occurring words with the word “oman” and their ranking by the number of occurrences. However, it is not appropriate to use these words as related words. This is because that for example the word “man” and the word “elimination round” shows the same number in the fig 4 but the word “man” is considered to be used in many topics besides the soccer game. On the other hand, the word “elimination round” is considered to be used only in the soccer game topic. In this case, the word “man” has to be considered as less important related word, and the word “elimination round” has to be considered as more important related word. To compute such importances, our system employies cosine distance. The cosine distance is defined as follows.

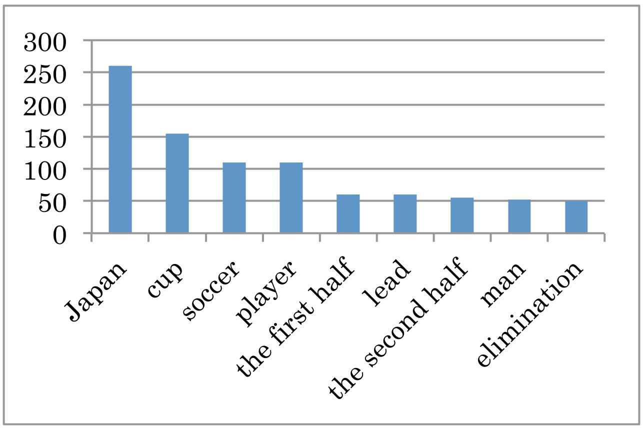


Figure 4: Ranking of co-occurrence words and its number of occurrences

In the formula (1), X refers a set of tweets which contain the query word and Y refers a set of tweets which contain a co-occurrence word (fig 5 shows “the second half” as a example of co-occurrence word).

$$\cos(X,Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}} \qquad \ldots(1)$$

However, it’s almost impossible to count the accurate number of each set of tweets in Figure 5, because twitter has enormous number of tweets. Therefore, we compute the ratio of each number of tweets as the following two steps.

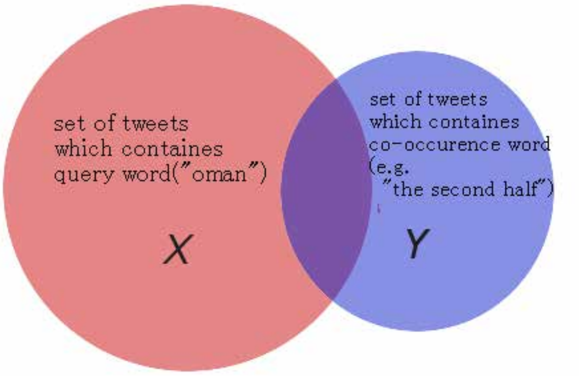


Figure 5: Conceptual diagram of cosine distance

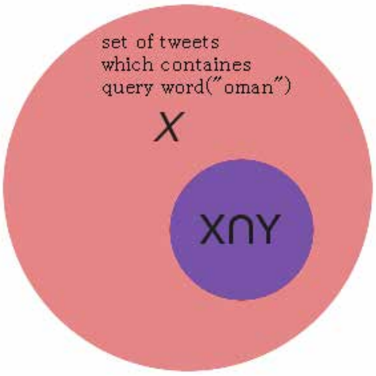


Figure 6: Computation of the ratio (1)

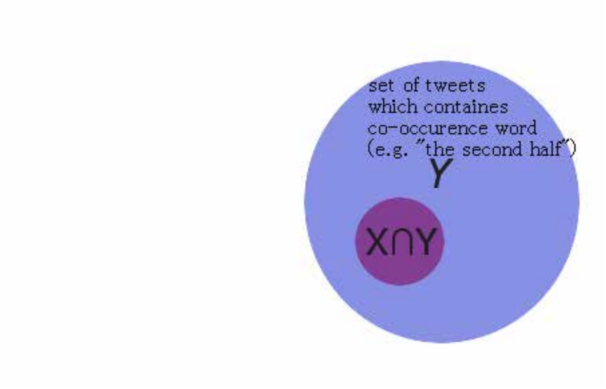


Figure 7: Computation of the ratio (2)

1. In the set of tweets that contain the query word, compute the ratio of the number of tweets that also contains the co-occurrence word at the same time.
2. In the set of tweets that contains the co-occurrence word, compute the ratio of the number of tweets that also contains the query word at the same time.

For instance, when the results of steps 1 and 2 are 1/10 and 1/8 respectively, the formula of the ratio is shown as follows.

$$X : X \cap Y = 10 : 1$$
$$Y : X \cap Y = 8 : 1$$
$$\therefore X : Y = 10 : 8$$

As a result, the formula of cosine distance is shown as follows.

$$\cos(X,Y) = \frac{1}{\sqrt{10 \cdot 8}} = \frac{1}{4\sqrt{5}}$$

Figure 8 shows the ranking of the words with the cosine distance in this case. A vertical axis shows the value of cosine distance between the word and the query word. As this figure shows, the value of general words like "man" became low, and the value of niche words like "elimination round" became high. In this paper, we use the cosine distance as the value of relevancy between the related word and the query word.

As we described in section 3.1, the system adds the users who posted a tweet that contain the related word to the candidates of reporter.

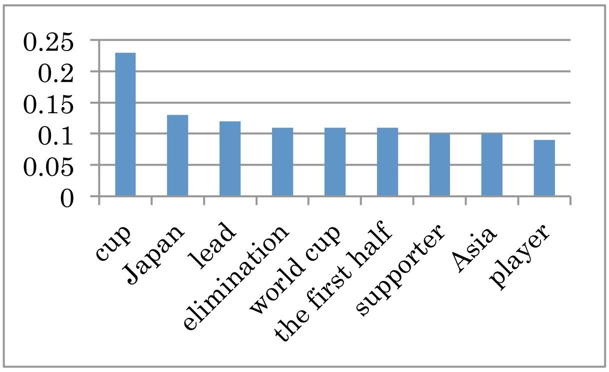


Figure 8: Ranking of cosine distance

## 3.3 Selecting reporters

Next, the proposed system selects conclusive reporters from the candidates. The requirements of reporters are shown as follows:

1. the users who will post tweets that are related to the event constantly, and
2. the contents of tweets are informative.

Therefore, the requirement 1 means the quantity factor of the user's tweets, and the requirement 2 means the quality factor of the user's tweets. We describe how to evaluate the quantity and quality factor in the next section.

### 3.3.1 Importance of quantity factor

The proposed system counts the number of user's tweets that is related to the event in order to evaluate the quantity factor of the user's tweets. In this paper, the system counts only tweets that are posted within the latest 5 hours. Also, if a tweet has at least one related word, the system regards it as related tweet to the event.

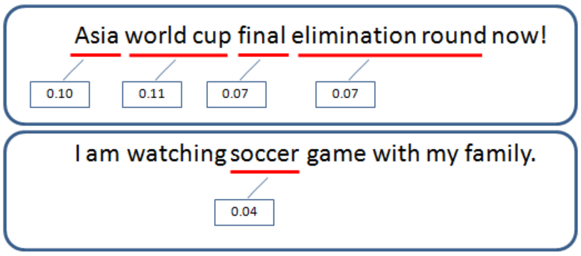
Formula 2 shows the definition of the value of user's quantity evaluation, where *av* refers "amount value", *u* refers "user", *t* refers "tweet", and *T(u)* refers the set of tweets that a user *u* has posted within the latest 5 hours.

$$av(u) = \sum_{t \in T(u)} \phi(t) \quad \ldots(2)$$

$$\phi(t) = \begin{cases} 1(\text{ if the tweet has at least one related word}) \\ 0(\text{ if the tweet doesn't has any related word}) \end{cases} \quad \ldots(3)$$

### 3.3.2 Importance of quality factor

Here, we would explain how to evaluate the value quality factor of the user's tweets. In this paper, the proposed system counts the related words that are included in the user's tweets in order to evaluate the quality factor of the user's tweets. For instance, Figure 9 shows example tweets of a candidate. In this case, the first tweet has four kinds related words and the total number of cosine value is 0.35. Therefore, the quality value of this tweet is 0.35. On the other hand, the second tweet has only one related word and its cosine value is 0.04, so the quality value of this tweet is 0.04. Like this, the system evaluates the quality factor of each tweet and as a result, it can evaluate the value of quality factor of each candidate by computing the average value per tweet.


Asia world cup final elimination round now!

I am watching soccer game with my family.

Figure 9: Example tweets

Because of this evaluation, a candidate who posts so many tweets that are not related to the event is rated too low, even though he/she may be rated high in the evaluation of quantity factor.

Formula (4) shows the definition of the value of user's quality evaluation, where *qv* refers "quality value," *u* refers "user," *t* refers "tweet," *T(u)* refers the number of tweets that the user(u) has posted within the latest 5 hours, and *w* refers the set of related words.

$$qv(u) = \frac{\sum_{t \in T} \sum_{w \in W} \varphi(t,w)\gamma(w)}{|T|} \quad \dots(4)$$

$$\varphi(t,w) = \begin{cases} 1(\text{if the tweet has the related word } w) \\ 0(\text{if the tweet doesn't have the related word } w) \end{cases}$$

$$\gamma(w): \text{ value of cosine distance of the word } w \quad \dots(5)$$

### 3.3.3 Conclusive importance of the candidate

The proposed system computes the value of conclusive importance by calculating the product of quantity value by quality value. We define the *tv*, which refers "total value." Because of the criteria of the quantity and quality value, as we described in 3.3.1 and 3.3.2, the candidates who often posts informative tweets that are related to the event are selected as conclusive reporters. The formula of the conclusive evaluation is shown as follows.

$$tv(u) = av(u) \cdot qv(u) \quad \dots(6)$$

## 4. Evaluation

### 4.1 Procedure

In order to evaluate the reporters who are recommended by this system, we did an experiment. The subject of the event is "えべっさん", which was held in Osaka, Japan on January, 11th 2013.

We did this experiment relatively by comparing the tweets that were posted by the reporters and the tweets that were posted with the hash tag "#えべっさん." Moreover, we also evaluated the effectiveness of the related words by comparing the results of tweets that were acquired using the related words and not using it. The concrete procedure is shown as follows.

1. We send a query "えべっさん" to the system at 8 o'clock on the event day, and follow five reporters. Also, we followed more five reporters that were recommended by not using related words(using only query word).
2. One hour later after following, we collected the latest 20 tweets that were posted by the reporters.
3. At the same time, we collected the latest 20 tweets that were acquired by researching with hash tag "#えべっさん".
4. We sent out questionnaires about the sets of tweets that were collected in the above procedure. The contents of questionnaires are, "If you are attending the event and you want information about the event, do you think the tweet is informative or not? Please evaluate each tweet from one to five." The criteria of the evaluation are that one is "Not informative, it's not needed to get information", and three is "Neither", and five is "So informative." The subjects are not informed whether each tweet is posted by reporters or got by the hash tag.

The table 1 shows the results of this experiment. Each column shows the average value of the evaluation to the tweets by the subject. As this table shows, the average value of the tweets posted by reporters is higher than the average value of the tweets that were got by researching hash tag. Also, the value of t-test is p=0.02<0.05, which verify a significant difference of this results.

On the other hand, there is almost no difference of the results between the reporters that were selected using related words and the reporters that were selected not using related words. The reason of this is that the values of cosine distance of the related words are too low compared with 1, which is the value of cosine distance of the query word. As a result, when the system computes the quality evaluation of each tweet, the values of related words are almost meaningless. To solve this problem, we should compute the value of related words and the query word by respective measures.

Table 1: Results of the experiment

| | Reportes | Reporters(without related words) | Hashtag |
|---|---|---|---|
| Subject A | 2.85 | 2.79 | 1.95 |
| Subject B | 2.65 | 2.58 | 2.55 |
| Subject C | 2.75 | 2.89 | 1.95 |
| Average | 2.75 | 2.75 | 2.15 |

### 4.2. Discussion

We conducted an experiment for an event which is called "えべっさん", and got the good result. However, it doesn't mean that our proposed method can be applied effectively to all events. For instance, if the event is too small and the participants are very few, the results would be not so good. Because if the number of participants is small, there would be also few twitter users who posts about an event. In future, we are planning to do experiment for a small event and consider new approaches for it. Also, though the system currently uses related words that are included in tweets of each candidate in order to compute the quality factor of the candidate, it is not perfect. For example, the recommended reporters would post tweets that include similar words. To solve this problem, we have a plan to use the count of retweets or replies to the tweet to evaluate the quality factor of the tweet.

## 5. Conclusion

In this paper, we proposed a system that can construct ad-hoc follow networks on twitter for sharing information about an event automatically. In conventional systems, we have to search tweets by one or more hash tags in order to get information about an event. However, there are some problems of such systems. For example, there are many futile tweets that include a hash tag, conversely, there are also many informative tweets that doesn't contain the hash tag. These problems occur because appending hash tag to the tweet is optional function. Therefore, we propose a new approach that recommending "reporters" who will post informative tweets about an event constantly. The important point of this study is that how the system selects the conclusive reporters from the candidates. In this paper, we evaluated each candidate by looking at importance of both quality factor and quantity factor of the candidate's tweets.

To evaluate the reporters who were recommended by our system, we conducted an experiment for an event which is called "えべっさん." We asked three subjects to evaluate each tweet from one to five. The result shows that the average points that are evaluated to the reporter's tweets exceed the average points of hash tag tweets, and the value of t-test verifies a significant difference of this result.

## References

[1] Shoko Wakamiya, Ryong Lee, Kazutoshi Sumiya: Twitter-based TV Audience Behavior Estimation for Better TV Ratings, DEIM Forum (http://db-event.jpn.org/deim2011/), 2011

[2] Bernard J. Jansen, Mimi Zhang, Kate Sobel, Abdur Chowdury: Twitter Power - Tweets as ElectronicWord of Mouth, Journal of the American Society for Information Science and Technology, Vol.60, Pages 2169-2188, 2009

[3] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo: Earthquake Shakes Twitter Users - Real-time Event Detection by Social Sensors, WWW '10 Proc. of the 19th international conference on WWW, Pages 851-860, 2010

[4] Eiji Aramaki, Sachiko Masukawa, Mizuki Morita: Twitter Catches The Flu - Detecting Influenza Epidemics using Twitter, EMNLP '11 Proc. of the Conference on Empirical Methods in Natural Language Processing, Pages 1568-1576, 2011

[5] Takuya Sugiya : Detecting local events by analyzing tweets regionally (in Japanese), DICOMO2012(http://www.dicomo.org/), Pages 1704 – 1711, 2012

[6] Huberman, Bernardo A., Romero, Daniel M. and Wu, Fang, Social Networks that Matter: Twitter Under the Microscope (December 5, 2008). Available at SSRN: http://ssrn.com/abstract=1313405 or http://dx.doi.org/10.2139/ssrn.1313405

[7] danah boyd, Scott Golder, Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. System Sciences (HICSS), 2010 43rd Hawaii International Conference on.

[8] T. Joachims. Text categorization with support vector machines. In Proc. ECML'98, pages 137–142, 1998.

[9] Cuneyt Gurcan Akcora, Murat Ali Bayir,, Murat Demirbas, Hakan Ferhatosmanoglu : Identifying Breakpoints in Public Opinion, SOMA '10 Proc. of the First Workshop on Social Media Analytics, Pages 62-66, 2010