Hot Topic Detection in Local Areas Using Twitter and Wikipedia

Shota Ishikawa*, Yutaka Arakawa*, Shigeaki Tagashira*, Akira Fukuda*
*Graduate School / Faculty of Information Science and Electrical Engineering
Kyushu University
744 Motooka, Nishi-ku, Fukuoka, Japan
Email: {ishikawa, arakawa, shigeaki, fukuda}@f.ait.kyushu-u.ac.jp

***Abstract*—As microblog services become increasingly popular, spatial-temporal text data has increased explosively. Many studies have proposed methods to spatially and temporally analyze an event, indicated by the text data. These studies have aimed a extracting the period and the location in which a specified topic frequently occurs. In this paper, we focus on a system that detects hot topic in a local area and during a particular period. There can be a variation in the words used even though the posts are essentially about the same hot topic. We propose a classification method that mitigates the variation of posted words related to the same topic.**

I. Introduction

Over the past several years, participation in social media, e.g., posting and/or reading, has gradually become a routine part of many peoples' lives. The posts cover a wide range of topics, including daily activities, events, opinions, comments, photographs, and links to web pages. The popularity of this form of communication has been driven by advances in mobile phone technology. Smartphone, which enable access to internet services, are becoming increasingly popular at an unprecedented rate. Social media applications for smartphones have also been developed and popularized. These client applications have features that exploit smartphone's ancillary functions such as a global posting system (GPS) and a camera, which, for example, enable users to post still or video images, and determines their current location. These associated features greatly improve the usefulness and will further spur the growth of social media services. The number of people who use a variety of microblog service is astonishing. In particular, $Twitter$ has attracts an estimated 200 million participants globally since 2006 when the service started (Twitter, Inc. estimated that the number of active users is 100 million) [1]. In Japan, the growth rate of Twitter users is remarkable. In [9], a good correlation was reported between smartphone owners and Twitter users. Since smartphones are spreading increasingly rapidly, we can predict that the number of users participating in social media will continue to rise.

Due to the convergence of technology and high participation rates, it is now possible to readily collect more spatial-temporal text-based Twitter posts, known as "tweets," than before, since a tweet can be associated with not only the posting time but also the posting position. By analyzing the contents of tweets, it is possible to forecast a market [4], sense a circumstance (weather and noise level) in a specific area [7], and identify hot topics that coincide at a particular time and location; it would lead several studies on complex networks.

Examples of hot topics are predictable events such as a soccer game or a festival, and unpredictable events such as a natural disaster or traffic accident. These events would probably be common hot topics for users in corresponding location. Additionally, in [3], it was reported that word (hot topic) is often used at a specific location in the analysis of the relationship between the input words typed by users and the users' positions. It is expected that this word would also be of interest to other users in the same area. Therefore, it would help a user to detect the hot topics associated with their location. This detection can also contribute to artificial intelligence services such as suggesting keywords for web search system, which would save time while inputing the words into a search engine. Furthermore, it is possible to evaluate the efficacy of a specific advertising campaign and comprehend a disseminating flow of the detected hot topics by observing the spatial-temporal changes in hot topics. It would be helpful to produce a marketing strategy. Similarly, this analysis would be applied to the spreading model of an infection disease and the network of people's relationship, which has been focused in complex network science.

We propose a novel detection scheme for hot topics on Twitter. The basic approach is to classify tweets into topics according to their content and select the top topics, ranked according to the number of topics' tweets, as hot topics. In this classification, the tweets, including words referring to the same event must be exactly associated with the corresponding topic. However, due to semantic fluctuations, this classification does not work particularly well. For example, tweets can use different words to refer to the same event, and consequently, they will be classified as different topics. In this paper, we identify semantic fluctuation as spelling, spatial, and temporal fluctuations. For example, the word "stadium" included in a tweet could refer to a baseball game in a particular region, but it could also refer to a soccer match in a different location This is an example of a spatial fluctuation. Similarly, the word "festival" has seasonal or temporal fluctuation. It could imply a music festival held in May or a food festival held in July, although both events take place in the same region.

[1] http://blog.twitter.com/2011/09/one-hundred-million-voices.html

Here, we focus on spatial and temporal fluctuations and propose a clustering method to cope with these fluctuations. Basically, we interrelate words based on their interpreted meanings graphically using Twitter and Wikipedia. By analyzing the tweets, we can understand the local meaning of time- and position- dependent words. In Wikipedia, the structure of global meanings for common words in every region has been built. Next, we associate these graphs with topics, on the basis of their similar occurrences on the graph. Finally, we detect hot topics by calculating the frequency of each topic in a specific period.

The paper is organized as follows; In section II, we introduce the related work analyzing spatial-temporal text data and make a sharp distinction between these studies and our approach; In section III, we describe our approach in detail; we then present the result and discussions of a preliminary experiment in section IV; in section V, we propose a technique to speed-up the process; finally, we present our conclusions and suggest possible future work in section VI.

II. Related Work

Several recent research have analyzed the spatial-temporal text data of social media. The main goal of our research is to detect hot topics (including semantic summarizing); other researches have pursued similar goals. For example, Fujisaka et al. [8] collected tweets using Twitter's application programming interface (API), and analyzed the movement histories of several users. They discovered characteristic mobility patterns in an urban area.

Yamanaka et al. [20] has proposed an extraction method that detects events in a given observation area. Initially this method categorizes messages with attached GPS information by using a support vector machine (SVM) model [10]. Secondly, the messages are clustered based on messages' category and the position. Finally, a burst is detected for each cluster. The burst-detection method, which has been proposed by Kleinberg [11], detects whether the interval between messages is more dense than that in a normal condition. However, this method needs to predefine a query set for each area and condition.

In addition, Sakaki et al. [18] has proposed an event-detection method using Twitter. For example, the method focuses on an earthquake as the event. It uses the SVM as an event classifier and then detects the event by calculating the occurrence probability. However, this method only handles specific words relating to the particular event. It lacks general versatility for event detection. Similar to the above method, this method also requires the configuration of a query set for each situation and event in advance. Thus, to apply this method to a wide range of situations, a considerable amount of time would be required to prepare the queries.

In addition to the above studies, a few event detection approaches that do not need predefined query sets have been proposed. Mathioudakis et al. [13] has developed a technique for detecting a trend based on the co-occurrence probability between events. However, in this method, the resultant trend sometimes includes phrases commonly used in Twitter conversations. Becker et al. [5] [6] has detected an event by analyzing tweets and identified whether the event had actually happened. Here, tweets are classified by calculating characteristics from temporal, social, topical, and Twitter-centric features, and then separating events from non-events. However, the process involves using capitalizations rules to split multiple word hashtags into single words. This could not be applied to Japanese as the language does not use capital letters.

Another set of related studies examined semantic summarizations of topics. This approach often exploits hashtags. A hashtag, which is a Twitter specific annotation format, is used to designate or assign a topic in a tweet. For example, in Canada a tweet about professional hockey matches are often tagged with the hashtag, #NHL(National Hockey League). Using hashtags, enable us to summarize tweet topics effectively. Rosa et al. [17] categorized tweets into six predefined topics using hashtags. Long et al. [12] has attempted to summarize tweets using hashtags in $SinaMicroblog$. Moreover, Motooka et al. [14] has researched tweets about a similar event using hashtags. This approach collects a set of users using a specified hashtag. It displays top events ranked according to the similarity between the specified hashtag and all events associated with the tweets. Here, it is important to note that a tweet can have multiple hashtags. However, the number of tweets marked by hashtags is still small; the percentage of hash-tagged tweets in geotagged tweets is only 0.4% [2]. Therefore, in order to ensure a broader event detection the proposed method did not employ a hashtag-based detection.

It is possible to use information resources other than Twitter, used in above approaches, for detecting events or topics. WordNet [1] has published a meaning dictionary that groups words into sets of concepts and links conceptually similar sets to indicate relationship. However, WordNet's primary classification is based on parts of speech, i.e., nouns , verbs, adjectives, and adverbs. Therefore, for example, it is impossible to find a relationship between the noun "earthquake" and the verb "to shake." Consequently, several ontology construction methods using Wikipedia and Folksonomy have been proposed in [16] [19]. Wikipedia has already covered a wide range of vocabulary related to global areas or regions. Moreover, Wikipedia has been built semi-structured data (redirect link, category tree, and infobox), which makes it possible to construct the relationship between concepts. The goal of the ontology construction approach is to clearly define the relationships among concepts (is-a relationship, and a-part-of relationship). Our purpose is also to build the relationships and a path from upper concepts to lower ones. In our approach, we also use Wikipedia to arrange words in terms of semantics.

III. Proposed Detection Method

In this section, we present a detailed description of our proposed hot-topics detection method.

[2]We surveyed tweets posted in Tokyo area from Jun to Sep. in 2011. As a result, among 277,249 geotagged tweets, there are only 1,088 hashtagged tweets.

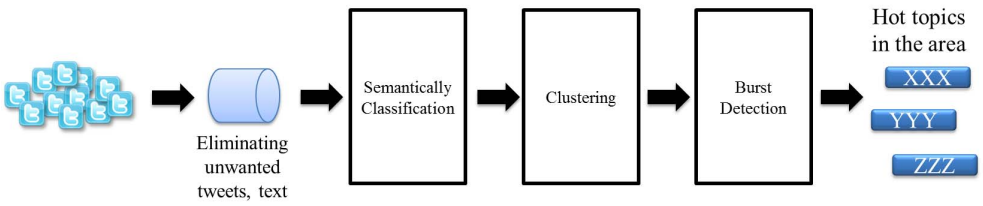

Fig. 1. Flow of proposed method.

We collect tweets that are associated with geotag using Twitter's streaming API [3]. Firstly, we eliminate tweets posted from $foursquare$ [4]. Because most of these tweets only contain the location information and a URL. Thus, they are not useful samples for our hot topic detection method. Similarly, we also eliminate URLs from the text in all the tweets [17].

Next, we perform a morphological analysis using MeCab (Japanese morphological analysis engine) [5] to decompose the text into parts of speech. Since it is difficult to accurately associate the qualified word with an adjective or adverb, we only focus on nouns and verbs extracted by the morphological analysis. In this paper, we do not focus on the morphological analysis's methodology; however in future, we intend to improve the accuracy of the extraction process.

After building a set of semantically related words contained in the obtained tweets, we detect hot topics from the set. The detection method consists of the following procedures.

- Building relationship among a set of words (section III-A)
- Classify the words into topics (section III-B)
- Detect hot topics using a burst detection method (section III-C)

A. Building relationship among a set of words

It is possible that different words indicate the same topic, and the converse being true. This possibility is termed as "semantic fluctuation." As described previously, we consider that the semantic fluctuation in Twitter consists of spelling, spatial, and temporal fluctuations. For example of spatial fluctuation, although a word "stadium" included in a tweet is concerned with the topic of baseball in a region, the word might express about that of soccer elsewhere. As for temporal fluctuation, a word "festival" implies different topics in each season, e.g., the word indicates not only a music festival held on May but also a food festival held on July even in the same region.

While building relationship among a set of words, we have assumed that each tweet refers to a particular topic. This is a reasonable assumption considering the limit on the number of input characters (a tweet must not exceed 140 characters).

To determine relationships between words, firstly we build relationships between pairs of words contained in each tweet,

[3]http://dev.twitter.com/pages/streaming_api
[4]Foursquare : www.foursquare.com
[5]MeCab : http://mecab.sourceforge.net/

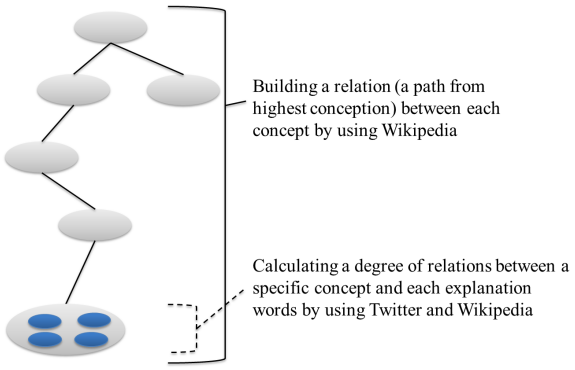

Fig. 2. Building conceptual path and relationship between concepts and words.

i.e., the link of weight (w_i) is established between any combination of words. Secondly, if any two tweets contain the same pair of words, the link of weight (w_r) is established between the pairs, where w_r is proportional to the number of word combinations. Additionally, if a word exists in a hash-tagged tweet, the weight of the link related with the word is w_h ($w_i < w_h$). If a word is included in Wikipedia, the rank of the word is added by one. Finally, we obtain links from upper concepts defined in Wikipedia, by the matching category name or Infobox template, and scraping, as proposed in [19].

B. Clustering

We utilize a clustering method to classify the words into topics. This clustering method introduces a similarity degree of association between words and topics. We adopt an incremental clustering method such as that proposed in [2], rather than recreating the sequential clustering. The proposed method calculates a degree of similarity $sim(m_i, c_j)$ between words $M = \{m_1, .., m_i, .., m_n\}$ and existing clusters $C = \{c_1, .., c_j, .., c_k\}$. If $sim(m_i, c_j)$ exceeds τ, the word is classified into the maximized cluster c_j. On the other hand, when $sim(m_i, c_j)$ is less than τ, we classify the word into a new cluster c_{k+1}. In addition, a threshold τ is determined empirically.

C. Detecting Burst Topics

Next, we use a burst-detection method to determine the frequency of a topic in a given period. The method has been proposed in [11]. This method detects whether the interval of

the arriving messages is denser than that in a normal condition through comparison with other document streams such as bulletin board threads and current news articles.

The burst-detection method defines a probabilistic automaton (A) consisting of two states: (1) When A is in state q_0, messages arrive at a slower rate. (2) When A is in state q_1, messages arrive at a faster rate. The period (T) is defined as the interval between the arrival of the first message and that of the last message, $n+1$. If the arrival time is random, a gap time x between messages i and $i+1$ follows an exponential distribution. According to Poisson distribution, in state q_0, the probability of arrival of the next message at interval x is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$, where $\alpha_0 = n/T$. In state q_1, the gap time is shorter than in state q_0. Consequently, the probability of interval x between any two consecutive message is $f_1(x) = \alpha_1 e^{-\alpha_1 x}$, where $\alpha_1 > \alpha_0$.

In addition, we determine a given set of n messages with a specified arrival time as inner-arrival gaps $\mathbf{x} = (x_1, x_2, ..., x_n)$, where $x_i > 0$. Similarly, we set the conditional probability of a state sequence $\mathbf{q} = (q_{i_1}, q_{i_2}, ..., q_{i_n})$. Each state sequence q derives a density function f over sequences of gaps, which is represented by the following formula.

$$f_q(x_1, ..., x_n) = \prod_{t=1}^{n} f_{i_t}(x_t) \tag{1}$$

Hence, when the inner-arrival gaps is equal to $\mathbf{x}$, a conditional probability that the state sequence is formed by $\mathbf{q}$ exists and, is given by the following formula.

$$Pr[\mathbf{q}|\mathbf{x}] = \prod_{t=1}^{n} f_{q_t}(x_t) \tag{2}$$

We have assumed that a maximum likelihood (burst) state is equal to $\mathbf{q}$ when it takes the highest value among probability $Pr[\mathbf{q}|\mathbf{x}]$. i.e., it is equivalent to minimum of the following values.

$$-\ln Pr[\mathbf{q}|\mathbf{x}] = \sum_{t=1}^{n} -\ln f_{i_t}(x_t) \tag{3}$$

We can detect a cluster burst by finding the state $\mathbf{q}$ that has the lowest value among those described.

IV. Experiments

We conducted a preliminary experiment to examine the semantic fluctuation of words included in tweets. In this experiment, we collected tweets that were associated with geotags in a 5 × 5 km area around a baseball stadium. Fig. 3 shows the possibility of words and Fig.4 shows the log distribution of words rank and frequency of words.

From the result, we can observe that the spatial words "Tokyo" and "Shinjyuku," which are related with the analyzed area, frequently appear in the whole range of date. However, the temporary words "strike" and "preemptive point," expressly related to baseball, only appear on a specific date, i.e., a baseball game is held on that date.

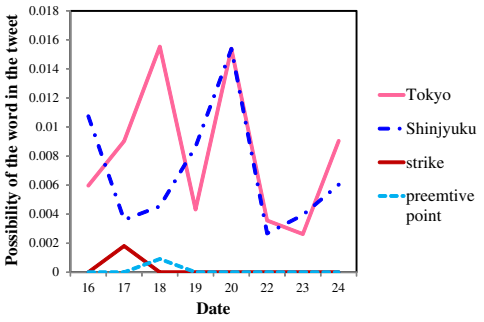

Fig. 3. Possibility of the word.

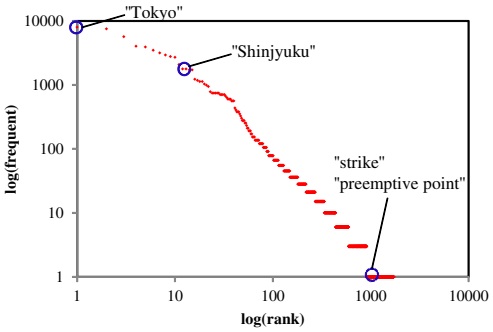

Fig. 4. Log distribution of words rank and frequency

In addition, the words "strike" and "preemptive point" did not appear on every days when a baseball game were held. This is because the percentage of tweets that indicate the event is primitively low. The contents of major tweets are conversations and do not indicate specific topics. The distribution of word in these tweets sampled through the API follws a Zipfian distribution, as shown Fig. 4. On the other hand, the number of topical tweets increases sharply only when the specific event occurs, likely to cause bursts in temporal. Namely, these burst words depend on time. Therefore, we can pick up a word which is related with the specific topic and is not general word in the tail of Zipfian distribution by calculating IDF (Inverse Document Frequency) (we consider tweets per unit time as a single document), and collect topical words efficiently.

V. Technique to speed up the process

In our approach, we analyze tweets in each geographical region, which are positioned in a regularly spaced grid, as shown in Fig. 5. Analyzing tweets in every region requires a considerable amount of computation time. If our target application is not responsive in real time due to the time required to process the calculations, the system would not be helpful for users. Thus, we propose decreasing the computation time by reducing the absolute number of analyzed areas.

Although Fujisaka et al. proposed an area splitting method in [8], their purpose was to avoid the use of the API as much as possible. In our approach, our purpose is to consistently

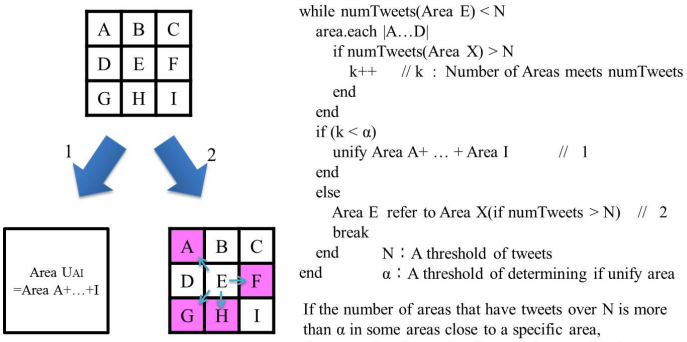

Fig. 5. Saving calculation time by expanding the area and using hot topics of adjacent areas.

reduce the number of analyzed areas. Besides, the number of tweets varies in each area, as shown in [9]. If the difference in the number of tweets in each area is caused by a simple splitting method, the statistical result would be unreliable. In that case we would detect hot topics in a specific region where the number of tweets is a few, a group of few tweets (compared to the average number of tweets in whole regions) indicates an event that would even affect a result. Therefore, to speed up the process, if the number of tweets in a region does not reach a certain number, we could expand the grid area until the number of tweets reaches a specified threshold. As a consequence, the number of tweets in every grid is normalized and the number of grids is reduced (see Fig. 5).

We define the threshold number of tweets as N. If the number of tweets in area E is less than N, the number of adjacent areas are calculated. Then, if the number of adjacent areas is over α, we discontinue detecting hot topics and use the result for hot topics in these areas because we consider that degree of interest in a topic is a spatial continuous value and the hot topics also appear in adjacent areas. On the other hand, if the number of areas is less than α, we integrate these adjacent areas into one area.

As mentioned above, we decrease the number of areas by expanding areas recursively and referring to adjacent hot topics. Consequently, the computational effort is lowered.

VI. Conclusions

In this paper, we proposed a novel detection scheme for hot-topics on Twitter. The basic approach is to classify tweets into topics according to their content and to select top topics ranked according to the number of the topics' tweets. This detection can contribute to artificial intelligence services, such as suggesting keywords for web search system and thus save time while inputting words into search engines. In addition, it helps to comprehend the trend flow in a marketing analysis by observing the detected hot topics changes in temporal-spatial. In future, we intend to elaborate the detail of the proposed system and implement the same.

References

[1] Japanese WordNet, http://nlpwww.nict.go.jp/wn-ja/index.en.html
[2] J. Allan, R. Papka, and V. Lavrenko, "On-line New Event Detection and Tracking," *Proc. the 21st ACM International Conference on Information Retrieval (SIGIR)*, pp. 37–45, 1998.
[3] Y. Arakawa, S. Tagashira, and A. Fukuda, "Relationship Analysis between User Contexts and Input Word with Twitter," Transactions of Information Processing Society of Japan, Vol. 52, No. 7, pp. 2268–2276, 2011. (in Japanese).
[4] S. Asur, and B. A. Huberman, "Predicting the Future with Social Media," *Proc. the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 492–499, 2010.
[5] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," *Proc. the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.
[6] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," Technical Report cucs-012-11, Columbia University, 2011.
[7] M. Demirbas, C. Akcora, M. Bayir, Y. Yilmaz, and H. Ferhatosmanoglu, "Crowd-Sourced Sensing and Collaboration Using Twitter," *Proc. the 2010 IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–9, 2010.
[8] T. Fujisaka, R. Lee, and K. Sumiya, "Exploring Urban Characteristics Using Movement History of Mass Mobile Microbloggers," *Proc. the 11th Workshop on Mobile Computing Systems & Applications (HotMobile)*, pp. 13–18, 2010.
[9] Hakuhodo DY Media Partners Institute of Media Environment,"2011 Media Teiten chosa," http://www.media-kankyo.jp/upload/files/article_128/teiten2011.pdf, 2011. (in Japanese).
[10] T. Joachims, "Text Categorization with Support Vector Machines," *Proc. the 10th European Conference on Machine Learning (ECML)*, pp. 137–142, 1998.
[11] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Proc. the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 91–101, 2002.
[12] R. Long, H. Wang, Y. Chen, O. Jin, and Y. Yu, "Towards Effective Event Detection, Tracking and Summarization on Microblog Data, " *Proc. the 12th International Conference on Web-age Information Management (WAIM)*, pp. 652–663, 2011.
[13] M. Mathioudakis and N. Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream, " *Proc. the 2010 ACM International Conference on on Management of Data (SIGMOD)*, pp. 1155–1158. 2010.
[14] R. Motooka, T. Yumoto, M. Nii, Y. Takahashi, and K. Sumiya, "A Similar Event Search System Using Hashtag of Twitter," The Database Society of Japan, The 3rd Forum on Data Engineering and Information Management (DEIM), A1-5, 2011. (in Japanese).
[15] S. P. Ponzetto and M. Strube , "Deriving a Large Scale Taxonomy from Wikipedia," *Proc. the 22nd National Conference on Artificial Intelligence (AAAI)*, pp. 1440–1445, 2007.
[16] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical Clustering of Tweets, " The 3rd Workshop on Social Web Search and Mining (SWSM), 2011.
[17] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," *Proc. the 19th International Conference on World Wide Web (WWW)*, pp. 851–860, 2010.
[18] S. Tamagawa, S. Sakurai, T. Tejima, T. Morita, N. Izumi, and T. Yamaguchi , "Learning a Large Scale of Ontology from Japanese Wikipedia," Transactions of the Japanese Society for Artificial Intelligence, Vol. 25, No. 5, pp. 623–636, 2010. (in Japanese).
[19] T. Yamanaka, Y. Tanaka, Y. Hijikata, and S. Nishida, "A Supporting System for Situation Assessment using Text Data with Spatio-temporal Information," Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 22, No. 6. pp. 691–706, 2010. (in Japanese).