Real-Time Event Search Corresponding to Place and Time using Social Stream

Ruriko Kudo
Department of Information Sciences
Ochanomizu University
Bunkyo, Tokyo, Japan
ruriko_k@ogl.is.ocha.ac.jp

Miki Enoki
IBM Research - Tokyo
Chuo, Tokyo, Japan
enomiki@jp.ibm.com

Akihiro Nakao
University of Tokyo
Bunkyo, Tokyo, Japan
nakao@iii.u-tokyo.ac.jp

Shu Yamamoto
University of Tokyo
Bunkyo, Tokyo, Japan
shu@iii.u-tokyo.ac.jp

Saneyasu Yamaguchi
Department of Information and Communications Engineering
Kogakuin University
Shinjuku, Tokyo, Japan
sane@cc.kogakuin.ac.jp

Masato Oguchi
Department of Information Sciences
Ochanomizu University
Bunkyo, Tokyo, Japan
oguchi@is.ocha.ac.jp

Abstract—Since the decision was made to hold the Olympic Games in 2020 in Tokyo, the number of foreign tourists visiting the city has been increasing rapidly. Accordingly, tourists have been seeking more sightseeing information. Their interests and visit durations are varied, so they should be provided information that is interesting and timely to each of them. While guidebooks are good for pointing out popular tourist attractions, it is more difficult for tourists to get information on local events and spots that are just becoming popular. We developed a tourist information distribution system that sends information corresponding to a place and time. The system extracts recent major and minor event information from social media streams in a per place and time manner and provides it to tourists. We evaluated the accuracy of its event classification of actual Twitter data with two representative methods: support vector machine (SVM) and random forests. Furthermore, we found that supplementary information from the web can be used to provide more accurate event information.

Keywords-Twitter; tweet; local event; information extract;

I. INTRODUCTION

Since the International Olympic Committee decided to hold the 2020 Games in Tokyo, the number of foreign tourists visiting the city has been increasing rapidly. Approximately 24 million foreign tourists visited Japan in 2016. The year-on-year increase was 21.8 percent, the highest ever [1]. It is important for Japan’s economy to keep attracting visitors. Along with attention to the tourism industry [2], various IoT devices have been developed to enrich the means of tourist information distribution.

In 2015, the University of Tokyo carried out demonstration experiments on various information distribution services using access points compatible with network virtualization [3]. In these experiments, researchers set up access points that implemented “BeaconCast” technology to distribute information on limousine buses running between the airport and the city from wireless beacons to users’ smartphones. This technology enables message delivery when the Internet can not be accessed. These days, by installing the special “LimoCast” application, passengers on the limousine bus can connect to the Internet via normal wireless LAN access by using the access point on the bus. Moreover, they can receive information about the destinations of the bus that is updated in real time. For example, users can receive information on events and even shopping coupons for stores at their destination. “LimoCast” is also useful as a means of delivering information to tourists who do not have a communication contract with local service providers.

Regarding the delivery of information, we can easily get information on popular tourist attractions from guidebooks, but it is not so easy to get information on local events, events currently in progress, or tourist spots that have only recently become popular. On the other hand, in various social networking services (SNSs) such as Twitter [4], many users post their impressions of tourist spots they have visited, as well as movies and stage shows they have watched. In events like festivals, users may tweet their current situation (for example, the degree of traffic congestion) and post photos. Sometimes, users tweet the schedule of the event that they are going to, not just impressions of it. In addition, it is possible to spread tweets using the “retweet” function so that the information is available to more people, which is also used for event announcements. In this way, SNSs contain useful information for people in specific places.

In this paper, we propose a tourist information distribution system that provides information corresponding to a place and time. The system extracts information that is useful to a person (e.g., a tourist) at the time from an SNS and suggests appropriate information in a timely manner. We evaluated the accuracy of the system’s event classification of actual Twitter data with two representative methods: support vector machine (SVM) and random forests. Furthermore, we found that supplementary information from the web provides more accurate event information.
The contributions of this study are summarized as follows.

1) We correctly extracted information from SNSs and provided it to a user despite that messages posted to SNSs have many informal sentences and spelling and grammatical errors.

2) We carried out experiments and evaluations with actual Twitter data and obtained accurate tweet classifications for each event category.

3) We implemented a multilingual translation system from Japanese tweets and a ranking mechanism that decides the priority of information delivery. In addition, we showed that supplementary data from the web can be used to provide more accurate event information.

The remainder of this paper is organized as follows. We introduce related research in Section II and give an overview of our system in Section III. In Section IV, V, and VI, we discuss the details of our system. We conclude the paper in Section VII with a brief summary and discuss future research directions.

II. RELATED WORK

There are a number of methods for extracting sightseeing information and regional features by analyzing data from SNSs [5], [6], [7]. Kurashima et al. [6] and Yizhu et al. [7] attempted to extract regional features from the geo-topic field. They applied topic modeling technology to SNSs with geotags to understand the features of each region. Shen et al. [5] proposed new methods of sightseeing value assessment by analyzing geo-social images. In these studies, the data to be analyzed was SNS data with geotags. However, on Twitter, only 2 percent of tweets include geotagged SNS data [9]. In this research, data to which position information is not added is also analyzed by analyzing the text. In addition, these methods perform static information analysis, which analyzes accumulated information. This research differs from them in that it aims to collect and deliver information dynamically and in real time, that is, by using a social media stream.

Chao et al. [8] developed a method for event detection using Twitter. Many event-detection methods target events that are just happening. These methods are based on the idea that when a lot of posts with similar content are posted locally, there is a strong possibility that some kind of event is occurring in that area. In addition, in these methods, local events such as earthquakes and demos with geographical localities are often detected. In this research, since the information is supposed to be presented to tourists, the extracted information is not just on an event that is just happening right now, but also about events a little in the future. The detection targets also include events that do not have geographical locality such as certain festivals and live events.

III. OVERVIEW OF PROPOSED SYSTEM

To present information useful for tourists in a timely manner, we propose the following system (see Fig. 1).

The procedure is as follows. Note that processes I-III are shown in Fig. 1.

I. Extract tweets

(1) By using the keyword search function of Twitter’s API [11], get tweets that include place names. For example, when applying the proposed system to a limousine bus connecting Narita Airport with the Tokyo Metropolitan Government Building, a place name within one of the 23 wards of Tokyo (Minato, Shinagawa, etc.) and a major station name (Shinjuku, Ikebukuro, etc.) are set as keywords. In this way, tweets containing place names in Tokyo are gathered.

(2) Analyze the tweet content and organize information.

(3) Distribute tweets to files corresponding to each day in which the event is held.

II. Classify events mentioned in tweets

(1) Extract tweets containing the content to be presented.

(2) Classify the extracted tweets by event category.

(3) Summarize tweets whose contents are similar.

III. Prepare information for delivery

(1) Collect and analyze the location information of the limousine bus and event venue.

(2) Prioritize the information to be provided.

(3) Translate information to be provided into multiple languages.
IV. Analyze Tweets Content

Here, we discuss the text analysis unit (Fig. 1(a)), which is part of the tweet extraction function of the proposed system. 

In text analysis, we extract the text, account name, ID, RT number, whether it is an official account, and the number of followers from the collected search JSON, and summarize it into a JSON file. In addition, we ascertain the date and time of the event and the venue by analyzing the content of the tweets.

A. Collecting dates and times of events

We use regular expressions to determine whether the date and time are included in the text of the tweet collected by the keyword search. We extract tweets that include dates and times and put them together in a file for each day of the event.

B. Collecting event places

It is not preferable to display the place name used in the keyword search as the event venue because the granularity of the place name is very large. In this system, we get the building name and address using an event spot dictionary (Fig. 1(b)) and regular expressions.

The event spot dictionary holds information related to amusement facilities, museums, shopping facilities, entertainment facilities, hot springs, theaters, and halls. If there is an element in the event spot dictionary in the content of the tweet, that element is stored in the JSON file as the event venue. Also, when the address of the event location is described in the text, the regular expressions are used to get and store the address. If neither of the above details is mentioned, the place name set by the keyword search is also stored in the JSON file.

V. Tweet Event Classification Evaluation

In this section, we evaluate the event classification of tweets (process no. II in Fig. 1).

Tweets that contain places and times often are not so useful to tourists. So, we classify tweets that are considered useful and exclude those that are not. In addition, we classify tweets into categories in order to provide information according to the user’s preferences.

The details of using machine learning to judge which tweets are the useful and an experiment evaluating the accuracy of the category classification are described below.

A. Category setting and experimental data

"Useful tweets" in this paper are tweets on events in the following categories.

Music Event” ”Vaudeville/Comedy”
"Theater” ”Movie” ”Art"

These categories were set by referring to the event category of “Tokyo Walker” [13], which is a representative city information magazine.

Learning data was created from tweets collected from September to October 2016. As for the learning data for judging whether a tweet is useful, tweets were classified manually as useful or not. Likewise, the category classification was performed manually. The experimental data consisted of 1,000 tweets on a December 13 event among those collected between December 6 and 13, 2016.

To automatically classify tweets, their text had to be preprocessed. Here, we extracted the necessary parses by using MeCab [15] and created a bag-of-words vector model of the tweets. The parts of speech extracted by MeCab were nouns, verbs, adjectives, and adverbs.

As a verification in the evaluation experiment, we manually divided the tweets into “correct answer” tweets (positive examples) and “incorrect answer” tweets (negative examples) and calculated the precision, recall, and F-measure.

B. Experimental method

Experiments were conducted using the following three methods. Note that support vector machine (SVM) and random forest are supervised learning techniques.

1) SVM+ IPADIC [16]
2) SVM + mecab-ipadic-NEologd [17]
3) Random Forest + mecab-ipadic-NEologd

The first method classifies tweets by SVM using MeCab’s dictionary with the default IPADIC. However, IPADIC has not been updated since March 2007, so words popular since then have not been recorded.

The second method uses mecab-ipadic-NEologd. This is a customized system dictionary for MeCab that includes new words obtained from a large number of linguistic resources on the Web, and it has the following features.

- It accepts proper nouns and new words mistakenly divided into multiple morphemes in IPADIC.
- It is updated at least twice a week.
- It utilizes language resources on the Web to add new unique expressions as needed.

SVM-light [14] is used as the classifier. Tweets are first classified by SVM using one-versus-one classification to determine whether they are useful or not useful. Next, useful tweets are classified by category, and one-versus-the-rest classification is performed.

Since SVM decides one of two classes in the category classification, its discrimination model is constructed for a certain category and all other categories, and it is classified into multiple categories. Since five categories were chosen, SVM discrimination models were created and judged for each of the five categories. Tweets that were not identified in any category were set as "Other.”

The third method uses a random forest as a classifier. A random forest is a group-learning algorithm that uses a
decision tree as a weak learner. An SVM classifying two
classes needs to learn each class. On the other hand, since
the random forest can perform multi-class classification,
it does not judge whether a tweet is useful or not and classifies
into six classes of "Not Useful", "Music Event", "Comedy",
"Theater", "Movie", "Art" all at once. Classification was done
using the Python machine learning library scikit-learn [18].

C. Experimental results

1) Support Vector Machine + IPADIC: Here, we describe
the results of classifying events in tweets by using an SVM,
which is a learning machine capable of classifying two
classes.

Table I lists the results of the one-versus-one classification,
and Table II lists the results of the one-versus-the-rest
classification.

Both classifications I and II had high recall, but the
precision and F-measure in Table II were biased.
The precision of "Theater" was extremely low probably
because the feature in which the weight is large even in
the classifiers other than "Theater" also exists in the feature
in which the weight is large in the classifier of "Theater."
Figure 2 shows a list of some of the features with high
weights in the classifiers of "Theater" and "Music Event."

Among the features listed above, the "opening" feature
is commonly included. When features with large weights
commonly appearing in classifiers of more than one category
exist, a certain tweet is determined to be correct in a
number of categories, and the precision of the classification
decreases.

![Diagram](image-url)

Figure 2. Examples of features with high weights

Therefore, we created one-to-one classifiers between
"Theater" and "Music Event" and excluded tweets that were
not "Theater." Table III shows the results of re-classifying
tweets classified as "Theater" by the category classification
using the newly created classifier.

As shown in Table III, although the recall decreased
slightly, the precision and F-measure improved. Therefore,
many results were mixed in the category "Theater," and we
were able to eliminate "Music Event."

2) Support Vector Machine + mecab-ipadic-NEologd:
The results of the experiments using mecab-ipadic-NEologd
as the morpheme dictionary of MeCab are listed in Tables
IV and V.

Both the usefulness judgment and the category classifi-
cation resulted in a higher F-measure as a whole compared
with the experimental results described above. In particular,
the precision and F-measure of the category "Theater" are
lower in Table II. Since a value close to the one of Table III
appears in one classification, it turns out that it is effective
to use mecab-ipadic-NEologd together with SVM.

3) Random Forest + mecab-ipadic-NEologd: In this sec-
tion, we describe the results of the event classification of
tweets using a random forest for the classifier. The results
of the category classification by random forest are shown in
Table VI.

As shown in Table VI, except for "Theater," each cate-
gory's F-measure was around 0.8. In "Theater" as well, we
were able to get the same results as SVM. We confirmed

![Table](image-url)

Table V

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music Event</td>
<td>0.85</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>Vaudeville/Comedy</td>
<td>0.54</td>
<td>0.94</td>
<td>0.68</td>
</tr>
<tr>
<td>Theater</td>
<td>0.35</td>
<td>0.88</td>
<td>0.50</td>
</tr>
<tr>
<td>Movie</td>
<td>0.55</td>
<td>0.86</td>
<td>0.67</td>
</tr>
<tr>
<td>Art</td>
<td>0.37</td>
<td>1.00</td>
<td>0.54</td>
</tr>
</tbody>
</table>

![Table](image-url)

Table VI

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music Event</td>
<td>0.88</td>
<td>0.76</td>
<td>0.81</td>
</tr>
<tr>
<td>Vaudeville/Comedy</td>
<td>0.79</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>Theater</td>
<td>0.50</td>
<td>0.40</td>
<td>0.44</td>
</tr>
<tr>
<td>Movie</td>
<td>1.00</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td>Art</td>
<td>0.87</td>
<td>0.81</td>
<td>0.84</td>
</tr>
</tbody>
</table>
that high values can be obtained with either classifier (SVM
or random forest). However, random forest got a higher F-
measure overall. We consider that the values can be further
increased by increasing the amount of training data and
deleting words that are too frequent and infrequent.

In regard to processing speed, the random forest, which
does not need to perform category classification by brute
force like SVM does, was faster.

In this research, we assumed that the information is to
be processed in real time and presented to users. Thus, we
decided to use random forest for the event classification of
tweets.

VI. PREPARATIONS FOR DELIVERY

Here, we describe the preparations for distributing the
data to users, in particular, the preparation of multilingual
information, prioritized distribution, and collection of event
names.

A. Multilingualization of information

In the previous section, we showed that tweets can be
collected for each category. Since the extracted information
are tweets written in Japanese, it is necessary to translate
them into the languages spoken by the tourists. However, ac-
curate translation is difficult because twitter posts have many
grammatical mistakes, typographical errors, and omissions.
Therefore, we organized the content of the tweets in advance
and provided only necessary information from them. The
presented information is the content extracted in (a) of Fig.
1. Event names are obtained by extracting text enclosed by
symbols such as ””,” and ** using regular expressions.

As an example, the proposed system shows the trans-
lations of tweets posted in Japanese. We used an auto-
matic translation engine ”Everyone’s automatic translation
@ TexTra” [12] developed by the National Institute of In-
formation and Communications Technology. Figure 3 shows
an example of an English translation of information that
was not preprocessed. Words underlined in red are incorrect
translations. Figure 4 shows an example of preprocessing to
organize tweet information. An example of translating Fig.
4 into English is shown in Fig. 5.

From the example, we can see that even if the translation
result is incorrect in several places, by preprocessing the
full text, it is possible to present the necessary information
briefly and correctly.

B. Ranking the information

Information is ranked in order to present information that
meets the conditions of the current location, date, and time
to benefit the user. The order that is beneficial is different
for each user. For example, as a key for ranking, the venue,
popularity degree, holding time, and fee of the event may
be considered. The popularity of the event can be known
by the number of retweet and whether multiple users tweet
about the event.

Let us describe an example of ranking based on an event
venue by using the Distance Matrix API [19] of the Google
Maps API. The Distance Matrix API can get the values
of time and distance based on the recommended route
between the departure and arrival points. One of our use
cases is passengers of a limousine bus travelling from the
airport to the city, so the user’s location information and
destination can be collected. The position of the limousine
bus is set as the departure point. The arrival point is the
event spot described in section IV. Part of the result of
setting "Shinjuku station” as the user’s position and its
ranking is shown in the Fig. 6. Listed is data on Music
Events, whose date is May 21, that were extracted from the
tweet-extraction and category-classification processed data
of tweets collected from April 18th to May 17th, 2017.

As shown in the figure, the traveling time is calculated
accurately using fine geographical name granularity, and the
ranking is determined. However, the event names in (i), (ii),
and (iii) were not able to be extracted. In the next section,
we describe a method for improving the collection of event
names.

C. Acquisition of information from the web

Here, we describe a method to improve the event name
collection by using external information.
Twitter postings have no constraints other than the limit of 140 characters; they hence have a high degree of freedom in their format, and the way they are written depends on the individual user. The information, such as date and time, where descriptions are determined to a certain extent can be extracted automatically using regular expressions. However, it is difficult to automatically extract event names because they can be written in various ways.

As a method of extracting event names from a free description such as a tweet, we can fill in missing information by using scraping or machine learning on external information sources. Scraping is a procedure to collect HTML data from a page of a website and to extract and reformat specific data. In this study, we used scraping to get the content of the title tag from HTML data of a website.

Shown below is a comparison of the collection rates of extracting highlighted parts in tweets by using regular expressions such as "", `", and ** and of getting external information by using scraping. We used the tweets of D in Fig. 7 for the data set. The number of tweets of A to D are as follows.

- A. Categorized tweets: 3280 tweets
- B. With URL: 1752 tweets
- C. With URL of image: 986 tweets
- D. With URL of website: 766 tweets

A consists of categorized tweets collected from April 18th to May 17th, 2017 and whose event dates are May 19, 20, and 21st, 2017. The tweet texts of B include the URL. There are two kinds of URL listed in B: the URL of the image attached to the tweet (C in Fig. 7) and the URL of the external website (D). The URL of the website is not listed in tweets other than D, but we can retrieve the website by putting the keyword in the tweet in the search engine. Therefore, we are planning to apply scraping to tweets other than D as well.

The collection rates calculated using 206 randomly chosen tweets from D in Fig. 7 are shown in Table. VII.

<table>
<thead>
<tr>
<th>Method</th>
<th>Collection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) emphasis expression</td>
<td>0.33</td>
</tr>
<tr>
<td>(2) scraping</td>
<td>0.43</td>
</tr>
<tr>
<td>(3) emphasis expression + scraping</td>
<td>0.62</td>
</tr>
</tbody>
</table>

In method (1), the event name was accurately extracted in 67 cases. In method (2), 89 event names were extracted. In 60 cases, method (1) failed, but method (2) succeeded. Not many emphasis expressions could be taken, but in many cases scraping was successful. Using method (3), the collection rate increased to 0.62.

It is difficult to extract necessary items automatically from the tweets, but if we take information from the web and use it as reinforcement, the amount of useful information will increase.

VII. CONCLUSIONS

We proposed a tourist information distribution system that corresponds to a place and time. The system extracts information that is useful for travelers (tourists) from an SNS and suggests appropriate information in a timely manner. By using SNS data, we can get local information that cannot be
easily collected from guidebooks, search engines, etc. The interests and visit durations of travelers are varied, so it is worth providing them with information that is interesting to them.

We classified the events described in tweets from the viewpoint of information presentation to match the user’s preferences. In the event classification, a comparison was made between SVM and random forest, and it was found that random forest is superior at classifying tweets into multiple categories in terms of both classification accuracy and processing speed.

To provide information in multiple languages to tourists, we also organized the information and made high precision translations. To present the beneficial information to users, the information was ranked. As an example, the travel time from the place of the user to the event place was used as the condition of ranking. The result of an analysis of tweet text is used, and the travel time based on more accurate position information can be calculated.

Finally, we collected information that would be difficult to obtain from tweets written in various formats by scraping it from websites. We showed that we can use this supplementary information from the Web to provide more accurate event information.

Future tasks include raising the accuracy of event classification and extraction of event names. Also, in the current system, we only cover tweets that include the place name and date and time in its text, and the types of tweets to be analyzed are somewhat small in number. We need to determine how to process tweets if some of the information is missing. We will also increase the number of tweet types to be analyzed and make it possible to provide more diverse information than now.

VIII. ACKNOWLEDGMENTS

This work is supported by Core Research for Evolutionary Science and Technology (CREST).

REFERENCES

[8] Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, and Jiawei Han, "GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams", Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.