On-site likelihood identification of tweets for tourism information analysis

Kazutaka Shimada, Shunsuke Inoue and Tsutomu Endo
Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502 Japan
{*shimada. endo*}*@pluto.ai.kyutech.ac.jp*

***Abstract*—Tourism is one of the most important key industries. The Web contains much information for the tourism, such as impressions and sentiments about sightseeing areas. Analyzing the information is a significant task for tourism informatics. One approach to extract tourism information is to extract sentences with keywords related to target facilities and events. However, all sentences with keywords might be not tourism information. In this paper, we propose a method for measuring tourism information likelihood. The target resource for the analysis is information on Twitter. The task is to identify whether each tweet has high on-site likelihood. We introduce a filtering process and a machine learning technique for the task. Our method obtained 80.5% on the precision rate.**

***Keywords*-Tourism information on the Web, Twitter, On-site likelihood**

I. Introduction

Tourism for many local cities is one of the most important key industries. The activation of tourism leads to the activation of the local industries and communities. In this situation, the World Wide Web plays a large role [6]. Although a huge number of online documents are easily accessible on the Web, the quality of the information is a mixture of the good and bad. Finding important information relevant to the target needs has become increasingly significant. We develop a tourism information analysis system which extracts information about tourism from the Web, analyzes the extracted information in various perspectives, and visualizes the output of the analysis [8]. Figure 1 shows the outline of the system. By using this system, people involved in the tourism can easily understand and organize significant information of the target city. The target resource for the system is information on Twitter[1]. It is one of the most famous microblogging services and text-based posts of up to 140 characters. The posted sentences are described as "tweets". In microblogging services such as Twitter, users tend to post tweets in real time. It denotes that tweets often contain significant information of events for tourism as lifelog data.

One approach to extract tourism information is to extract tweets with keywords related to target facilities and events. However, all tweets with keywords might be not tourism information. For example, the tweet "I'm on the way to work now. near `LOCATION`" is not suitable as an input for an opinion analysis system because it does not include sentiments about sightseeing as experience. Therefore, we need to judge the adequateness of each tweet.

In this paper, we propose a method for evaluating the adequateness for the tourism information analysis system. We focus on on-site likelihood of tweets. The on-site likelihood estimation is to identify whether a tweet is posted at the target facility. High on-site likelihood denotes the adequateness as the input for the analysis system. In addition, tweets with high on-site likelihood are useful to analyze the behavior of tourists because their location is identified. The on-site likelihood identification contains two processes; a filtering process and a machine learning process.

In this paper, we explain the outline of the basic information extraction process in our system in Section II. Then, we describe the on-site likelihood identification method in Section III. The method consists of a filtering process and a classification task with a machine learning technique. In Section IV, we evaluate the performance of our method, and conclude the paper in Section V.

II. Tourism information extraction

In this section, we describe a basic method to extract tourism information.

The extraction process is basically as follows:

Step1: Acquisition of basic queries,
Step2: Selection of related words,
Step3: Query generation and retrieval.

The basic information for this process is extracted from portal sites for tourism which the city and tourist association established. Here "basic query" denotes tourist facilities, restaurants and events, such as festival, which are written in the portal sites. Figure 2 shows an example of the tourism portal site about Iizuka city[2]. It consists of (1) facility or event names, (2) a link to detailed information of each entry and (3) basic information of each entry. We define the facility or event names as the basic queries. For example, "Kaho performing theater" and "Ito Den-emon residence" are basic queries. The number of basic queries is approximately 200 words.

We need to consider a problem of basic queries. Sightseers do not always mention the basic queries, i.e., facility or event

[1]http://twitter.com

[2]http://portal.kankou-iizuka.jp/ It is one of our target cities.

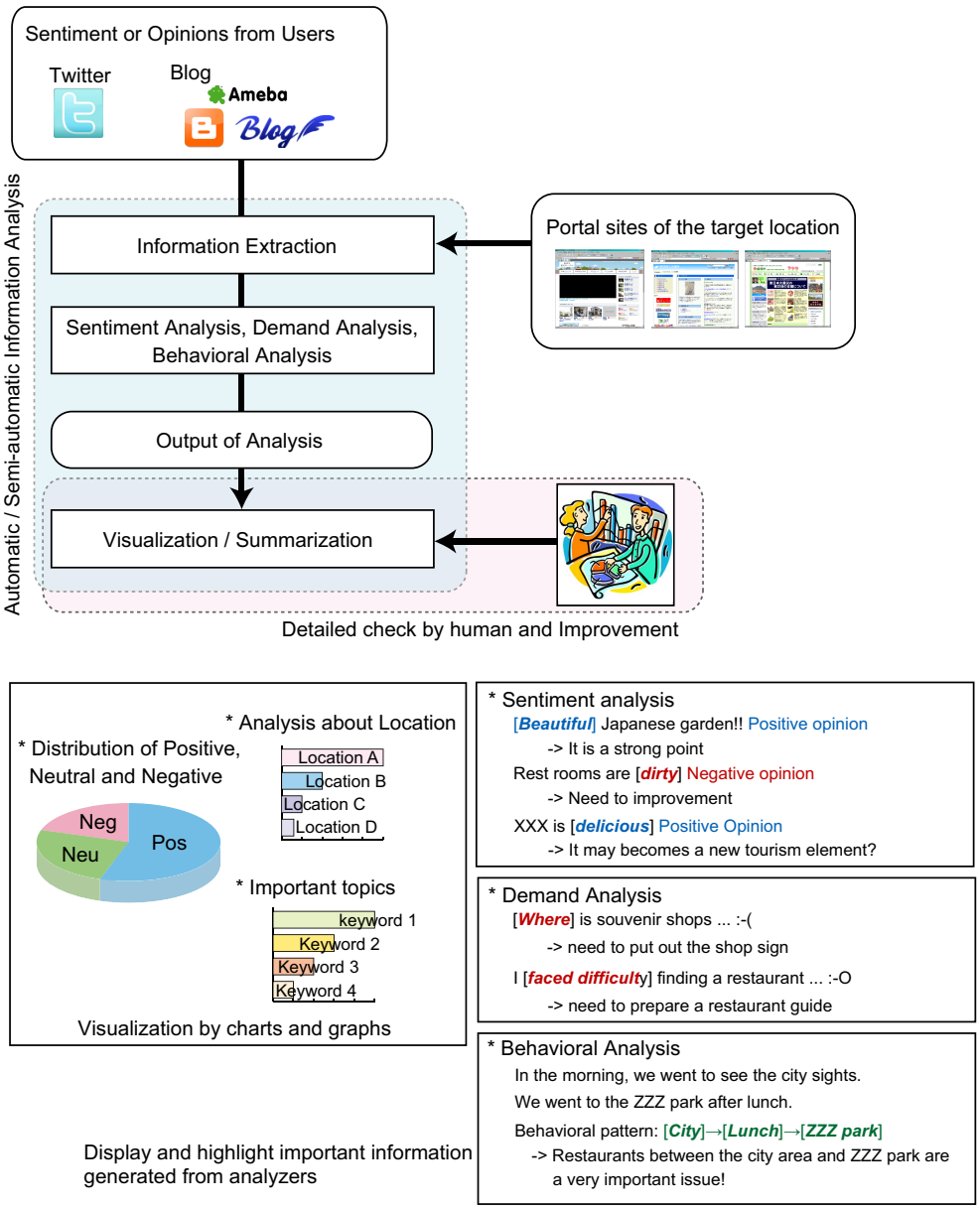

Figure 1. The outline of our prototype system.

names, in tweets. Moreover, they might mention information which is related to the location or event names and does not appear in the portal site. Therefore, we need to acquire related words of the basic query, i.e., query expansion. First, we need to divide each sentence into words. For the process we use MeCab[3], which is one of the most famous Japanese language morphological analyzers. For the selection of related words, we introduce a weighting approach, which is well-known as Okapi-BM25 [5]. The importance of a word is computed by

$$score(D, Q) =$$

$$\sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{\mathrm{f}(q_i, D) \cdot (k_1 + 1)}{\mathrm{f}(q_i, D) + k_i \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \quad (1)$$

where

$$\mathrm{IDF}(q_i) = \log \frac{N - \mathrm{n}(q_i) + 0.5}{\mathrm{n}(q_i) + 0.5} \quad (2)$$

$\mathrm{f}(q_i, D)$ is the frequency of a word q_i in a document D. $\mathrm{n}(q_i)$ and $|D|$ are the number of documents containing q_i and

[3]http://mecab.sourceforge.net/

Figure 2. The portal site on the Web.

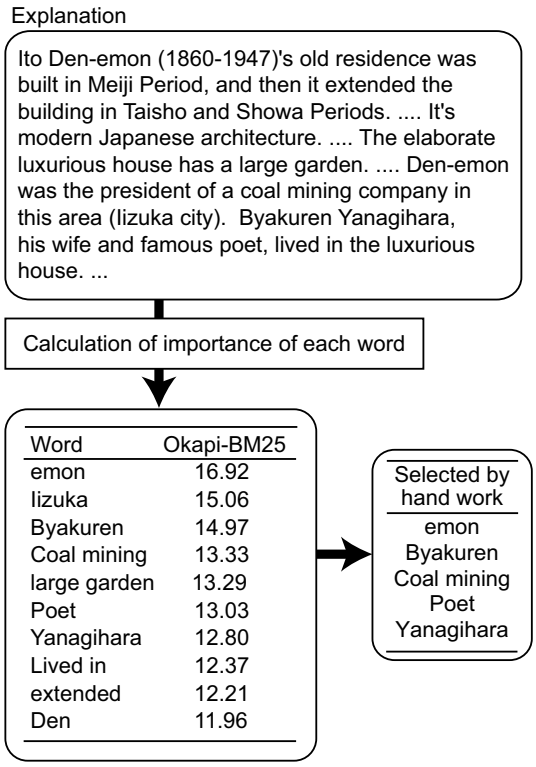

Figure 3. An example of explanation and the related words.

the length of D, respectively. $avgdl$ and n are the average length of documents and the number of documents. b and k are constant factors for weighting.

Figure 3 shows an example of explanation about "Old Ito Den-emon's residence (the 2nd facility in Fig 2)". We obtain a weighted word list from this explanation. In this example, appropriate words, such as "Byakuren[4]" and "Coal mining[5]", are located in top position in the ranking. Finally we select suitable words from the weighted word list, as related words by hand work. This selection process depends on subjective heuristics. In this process, we tend to select proper names (e.g., Byakuren) , the head of a noun (e.g., emon) and attributes of the target (e.g., Coal mining).

We retrieve tweets with Twitter API by using the manually-produced query list, namely the union of basic queries and related words. Final queries for the retrieval are combination of words in the list. However, there is a problem of the final queries. Tweets do not always contain the official name of facilities or events. For example, "Old Ito Den-emon residence", which is one of the most famous facilities in Iizuka, is exceedingly-long words. Therefore, it's unlikely that users input the official name itself. To solve this problem, we manually generate abbreviations of queries. For "Old Ito Den-emon residence", we add some abbreviations such as "Ito residence" and "emon residence" to the query list.

[4]She is a wife of Mr. Ito and famous poet.
[5]Mr. Ito was a rich coal mine owner and is called Coal Mine King

III. On-site likelihood identification

Not all tweets extracted in the previous section might be tourism information. For example, the tweet "I'm going to visit LOCATION" is not always suitable as an input for an opinion analysis system because it does not include sentiments about sightseeing as experience. Therefore we need to identify the on-site likelihood of each tweet. We introduce a filtering process and a machine learning technique for the task. Figure 4 shows the outline of the proposed method. First, we delete obvious noise tweets by using the rule-based filtering approach. Then, we classify tweets into tweets with on-site likelihood and without on-site likelihood.

A. Filtering

Many tweets on Twitter do not contain information with the on-site likelihood. Although one solution to identify the on-site likelihood of each tweet is to classify them by using a machine learning based classifier, biased data usually leads to the development of a unsuitable classifier. Therefore, we need to delete the tweets not containing the on-site likelihood in advance. We apply a rule-based filtering approach to the process.

The filtering process contains two types of rules; deletion rules and non-deletion rules. First, we detect tweets that are matched with the deletion rules. Then we delete the tweets if they do not contain the non-deletion rules. In other words,

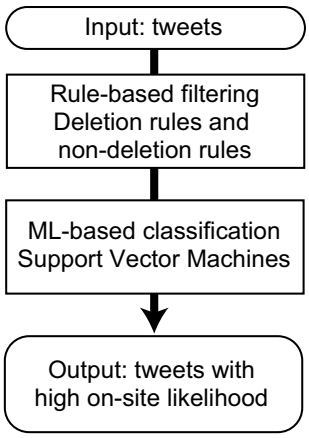

Figure 4. The outline of on-site likelihood identification.

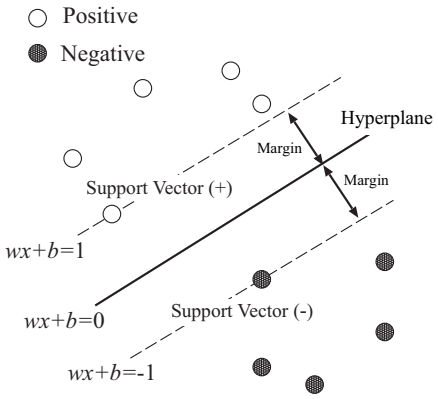

Figure 5. Support vector machines.

we retain the tweet matched with the non-deletion rules even if they correspond to the deletion rules.

The deletion rules are as follows:

- Linguistic rules: we delete tweets including the following words.
  - future: "tomorrow", "next week", "someday". etc.
  - tentative: "may", "guess", etc.
  - indirect: "hear", etc.
  - interaction marks: @ and RT
- Time rule: most tweets on late night do not probably relate to sightseeing information. Therefore we delete tweets that are posted from 11 p.m. to 3 a.m.
- Length rule: In a preliminary experiment, long tweets did not tend to include sightseeing information. Therefore, we delete tweets consisting of more than 100 letters.
- # of nouns: In the preliminary experiment, tweets containing many nouns were often advertising information. We delete tweets containing nouns whose the number is more than a threshold. The threshold in this paper is 36 nouns. This value was determined experimentally.

On the other hand, there are characteristics expressing on-site information with high probability. We introduce non-deletion rules. We do not delete tweets which contain the non-deletion rules. They are as follows:

- Presence of activity: if tweets contain words related to the activity of users, they have a potential value as on-site information. The words are "arrive", "Here we are" and so on. Words related to present and progressive tenses also contain a potential value as on-site information. The words are "looking", "strolling" and so on. Tweets with these words indicate user's action. Therefore we retain tweets with these words.
- Presence of "NAU (now)": NAU is a characteristic keyword on Twitter. Although the presence of NAU does not always indicate on-site information, it seems to suggest high potential for on-site information. For example, "`LOCATION`, NAU" denotes "I'm at `LOCATION`". This tweet includes on-site information. However, "`LOCATION` on TV, NAU" denotes "`LOCATION` is broadcast via television". It is not on-site information. In other words, "NAU" is ambiguous and considerable. Therefore we retain tweets with the word.

By using these rules, we can delete obvious noise tweets from the tweets extracted by Section II and obtain input candidates with high accuracy for the classification process.

B. Classification

The purpose of the filtering process is to delete noise tweets. The filtered tweets consist of tweets with high on-site likelihood and low on-site likelihood. Therefore, we need to classify them into on-site information and non on-site information. We apply Support Vector Machines (SVMs) to the classification task. SVMs are a machine learning algorithm that was introduced by [9]. An SVM is a binary classifier that finds a maximal margin separating hyperplane between two classes. The hyperplane can be written as:

$$y_i = \vec{w} \cdot \vec{x} + b$$

where $\vec{x}$ is an arbitrary data point, i.e., feature vectors, $\vec{w}$ and b are decided by optimization, and $y_i \in \{+1, -1\}$. The instances that lie closest to the hyperplane are called support vectors. Figure 5 shows an example of the hyperplane. In the figure, the solid line shows hyperplane $\vec{w} \cdot \vec{x} + b = 0$.

For SVMs, we use 10 features. The features are as follows:

1) BOW: The bag-of-words is a simple and famous feature in natural language processing. In our method, it is a baseline feature set. We use term frequency for the BOW model.
2) SpecificWords: There are some specific words linked to on-site information, such as "arrive", "I'm at" and "NAU (now)". These words are important features. We use the presence of them as the features.

3) PostTime: There are appropriate visiting time patterns for each tourist facility and event. For example, amusement parks are full of customers in the daytime. On the other hand, fireworks shows are events at night. The appropriate posting time of each tweet depends on the target facilities and events. We quantize the posting time by the hour. For example, the feature value is 12 if the posting time is 12:30. This feature is effective by combining the presence of tourist facility and events, namely the BOW feature.
4) PostRange: This is an abstract feature of the PostTime feature. We classify 24 hours into 3 ranges; (1) 4 a.m. - 9 a.m., (2) 10 a.m. - 9 p.m., and (3) 10 p.m. - 3 a.m.
5) Length: Users probably post tweets with mobile devices from facilities and event locations. In general, entering characters by mobile devices is a cumbersome process. In this situation, the length of posted tweets tends to be short. Therefore, we utilize the tweet length as the feature.
6) Tense: The tense is one of the important points for the on-site likelihood. The past and future tenses are not usually related to the on-site information. We focus on the tense linked to adjectives. For example, we distinguish between "tanoshii (enjoyable)" and "tanoshikatta (enjoyed)".
7) NumVerb: tweets with high on-site likelihood tend to use relatively little verb in the preliminary experiment. Therefore, we apply the number of verbs in each tweet to the feature for SVMs.
8) NumNoun: As we mentioned in the filtering process, the number of nouns in each tweet relates to the on-site likelihood. We also use the number of nouns as the feature.
9) RT: RT (ReTweet) is a kind of reference structure on Twitter. In our preliminary experiment, tweets with RT tended to not contain sentiment or opinion information of tourism. It leads to the decrease of on-site likelihood. Our method handles the presence of RT as the feature.
10) LocName: Tweets beginning with a location name often contain high on-site likelihood. We use the presence of a location name at the beginning of a tweet as the feature.

IV. Experiment

We evaluated our method with a test data set. The data set consisted of 2886 tweets. They contained 509 tweets as on-site tweets and 2377 tweets as off-site tweets. There are two points in this experiment; the filtering and the classification.

First, we evaluated the filtering process. The criterion of the filtering was a simple accuracy rate. The method described in Section III-A deleted 1738 tweets from 2886 tweets as noise tweets. In other words, we obtained 1148 tweets as inputs for the next process, namely the classification. 486 tweets of 1148 were on-site tweets (662 tweets were off-site tweets). As a result, our method deleted 23 on-site tweets by mistake. The accuracy of the filtering process was 95.5% (486/509). The mistake of the filtering process leads to the decrease of the accuracy of the classification process because the output of the filtering become the input of the classification directly. Therefore, the mistake is a fatal error of our method even if the error rate is 5%. The improvement of the filtering is the most important future work. On the other hand, our method reduced off-site tweets by approximately 30% (662/2377). This result shows the effectiveness of our filtering method.

Table I
THE EXPERIMENTAL RESULT.

Feature	Precision	Recall
BOW	80.2	62.3
BOW+SpecificWord	79.4	64.6
BOW+PostTime	80.4	63.1
BOW+PostRange	79.9	62.7
BOW+Length	**80.5**	62.5
BOW+Tense	80.3	62.3
BOW+NumVerb	80.3	62.3
BOW+NumNoun	80.2	62.1
BOW+RT	80.3	62.3
BOW+LocName	79.2	62.9
ALL	**80.5**	**65.0**

Next, we evaluated the classification task using SVMs. We used 1148 tweets, which were extracted by the filtering process, as the input of the classification. We evaluated the tweets with 10-fold cross-validation. Evaluation criteria were the precision and recall rates. These criteria are computed by:

$$\text{Precision} = \frac{\text{\# of correct outputs}}{\text{\# of tweets which the method judged as on-site}} \quad (3)$$

$$\text{Recall} = \frac{\text{\# of correct outputs}}{\text{\# of on-site tweets in the input data}} \quad (4)$$

The experimental result is shown in Table I. The mark "+" denotes the combination of each feature; e.g, "BOW+SpecificWord" denotes SVMs with the bag-of-words and specific word features. "ALL" denotes the method with all features mentioned in Section III-B. The method with all features produced the best precision and recall rates. As a feature for the combination with BOW (baseline), the SpecificWord feature was the most suitable in terms of the recall rate. The feature consists of the related word to on-site information, such as "arrive". It was an intuitive result. Although the precision rate was relatively favorable (80.5%), the recall rate was insufficient (65.0%). To improve the recall rate, we need to add suitable keywords to the SpecificWord feature.

Finally, we verified the effectiveness of the filtering process. Table II shows the result of a comparison of the method

Table II
THE EFFECTIVENESS OF THE FILTERING PROCESS.

Feature	Precision	Recall
With Filter	80.5	65.0
Without Filter	75.0	58.2

with filtering process and that without the filtering process. The method without the filtering denotes the result for 2886 tweets with 10-fold cross-validation. We used all features for these methods. Applying the filtering led to the improvement of both the precision rate (5.5 points) and the recall rate (6.8 points). This result shows the effectiveness of our method with the filtering.

V. DISCUSSION AND CONCLUSIONS

In this paper, we focused on on-site likelihood of tweets for evaluating the adequateness for the tourism information analysis system. The on-site likelihood identification was to identify whether a tweet is posted at the target facility. We proposed two processes for the on-site likelihood identification task. The filtering process was based on two types of rules; deletion rules and non-deletion rules. The accuracy was 95.5%. Although the accuracy was relatively favorable,the filtering errors became a critical problem in our method because the output was the input of the next process, namely classification process. The improvement of the filtering method is important future work. We applied a machine learning technique, namely SVMs with 10 features, to the classification process. For the task, we obtained 80.5% as the precision rate. We also verified the effectiveness of the filtering process for the classification task. However, the data set in the experiment was the closed data, namely known data for the method, for the filtering and the classification tasks. Therefore, we need to evaluate our method with the open data set, namely unknown data for the method.

Several researchers have studied similar tasks. Inui et al. [4] have proposed a task which were called experience mining. The task included factuality analysis, which was to judge whether the event indeed took place or not. Aramaki et al. [1] have proposed a machine learning based method to extract influenza tweets from Twitter. This task was also a kind of fact detection. By using this technique, they detected the influenza epidemics. Sakaki et al. [7] have also studied a event detection task from Twitter. Cheng et al. [2] have proposed a method for predicting a user's location based purely on the content of the user's tweets. They identified local words in tweets. Eisenstein et al. [3] have also discussed lexical variations across geographic areas from tweets. Incorporating knowledge of these studies to our method is considerable future work.

We identified the on-site likelihood from a tweet. However, handling only one tweet is not appropriate for the identification. Tweets that are posted within a time sequence often contain the relation between them. For example, the tweet "I enjoyed the place" contains uncertainty as to the on-site likelihood because it might be a tweet about a yesterday's event. However, if the tweet are posted after the tweet " `LOCATION` now!" at short intervals, the tweet has high on-site likelihood even if it includes the past tense. The time relation and the content of the previous tweet are important to improve the identification accuracy, especially the recall rate. We need to handle co-occurrences of tweets for the on-site likelihood identification task.

REFERENCES

[1] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[2] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–769, 2010.

[3] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[4] Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 314–321, 2008.

[5] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *In Proceedings of the Third Text REtrieval Conference (TREC 1994)*, 1994.

[6] Hajime Saito. Analysis of tourism informatics on web. *Journal of the Japanese Society for Artificial Intelligence*, 26(3):234–240, 2011.

[7] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquakeshakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW2010)*, 2010.

[8] Kazutaka Shimada, Shunsuke Inoue, Hiroshi Maeda, and Tsutomu Endo. Analyzing tourism information on twitter for a local city. In *Proceedings of SSNE2011*, pages 61–66, 2011.

[9] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1999.