A comparative study of potential-of-interest days on a sightseeing spot recommender

Kazutaka Shimada, Hisashi Uehara and Tsutomu Endo
Department of Artificial Intelligence
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka 820-8502 Japan
shimada@pluto.aikyutech.ac.jp

***Abstract*—We have already proposed a sightseeing spot recommendation system based on information on the Web. An input for the prototype system was a user's favorite location or facility. Our system computed a similarity measure between a target location that a user selects and each sightseeing spot in our database. One interesting feature for the similarity calculation in our system is a time sequence of each sightseeing spot. The prototype system used the number of hits in Yahoo Chiebukuro for the feature. We regard the time sequence as the potential-of-interest days. In this paper, we focus another information resource for the time sequence feature; Panoramio. We compare the two information resources, and analyze the difference. We show the potential merits of combination of Yahoo Chiebukuro and panoramio for sightseeing spot recommendation.**

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 industries and communities. In this situation, the World Wide Web plays a large important role [2], [9]. Many researchers have proposed recommendation systems for sightseeing. General approaches usually handle text-information. Kurashima et al. have proposed methods for mining and visualizing local experiences from blog entries [3], [4]. However, the coverage of only text-based systems is not always enough. Kanazawa et al. [1] have proposed an association retrieval system based on analysis of impression words to express a destination image. Kurata has proposed a system for assisting the user's tour planning in a collaborative manner [5], [7]. Okuyama and Yanai[8] have proposed a travel planning system based on geotagged photos on the Web.

We have already proposed a sightseeing spot recommendation system based on information on the Web [12]. An input for the prototype system was a user's favorite location or facility. Our system computed a similarity measure between a target location that a user selected and each sightseeing spot in our database. The resource for the similarity calculation is one of the most important points. We focused on several information resources on the Web as the collective intelligence. In the previous work, we used five features: (1) keywords, (2) time sequence, (3) category information on Yahoo Chiebukuro, (4) surrounding area information and (5) map images. One interesting feature for the similarity calculation in our system was the time sequence of each sightseeing spot. The prototype system used the number of hits in Yahoo Chiebukuro.

In this paper, we focus another information resource; Panoramio. We compare the two information resources, and analyze the difference. We show the potential merits of combination of Yahoo Chiebukuro and Panoramio for sightseeing spot recommendation.

The remainder of this paper is structured as follows: In Section II, we explain our prototype system with five features. Next, we introduce "potential-of-interest days" based on Panoramio in Section III. Then, we discuss the common and different points between Chiebukuro, namely the time sequence feature in the prototype system, and Panoramio in Section IV. Finally, we conclude our work in Section V.

II. Prototype system

We have developed a prototype system for sightseeing spot recommendation [12]. In this section, we explain the basic ideas of the prototype system. Figure 1 shows the outline of the system. The purpose of our system is to extract multiple features about sightseeing spots from various information resources on the web, and to visualize them for the sightseeing spot recommendation. The system uses Japanese Wikipedia[1], blogs, map information and Yahoo Chiebukuro (Yahoo Answers)[2] as resources of information for the recommendation. Our system extracts five features from these resources.

A. Word importance

The first feature is a word importance measure. For the feature, we use Yahoo Chiebukuro and blog entries. Yahoo Chiebukuro is one of the most famous community-driven Q&A site in Japan. Users can submit questions and answer questions that are submitted by other users. It contains hundreds of millions of a pair of a question and answers. It also contains several categories for each pair of a question and answers. We use entries in the category "Travel". Each entry includes the best answer that a questioner selected. We utilize the pair of a question and the best answer for the computation of an importance measure of each word. For blogs, we use words in the snippet from a blog search API.

First, our system determines candidate words for the calculation. The target is the title of each entry in Wikipedia. If the title words appear in the category "Travel" of Yahoo Chiebukuro, we regard them as the candidate words. The number of candidate words is 67644.

[1] http://ja.wikipedia.org/wiki
[2] http://chiebukuro.yahoo.co.jp

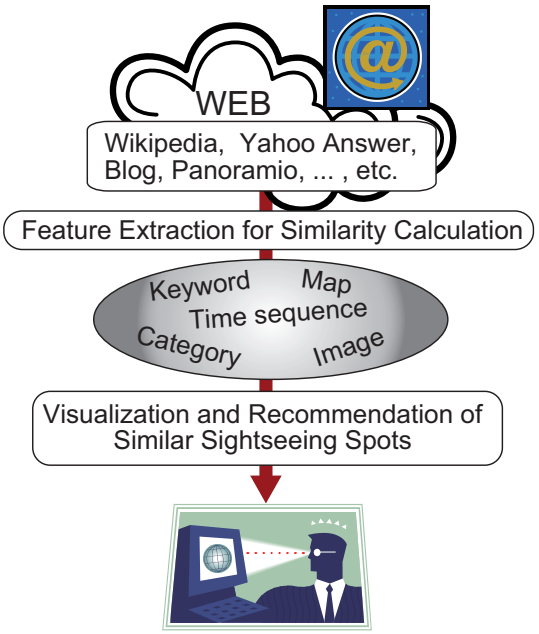

Fig. 1. Outline of our system.

TABLE I. EXAMPLES OF L_k, A_k AND S_k.

k	L_k	A_k	S_k
Fukuoka (City name)	27647	164219	0.168
Fukuoka tower (Spot)	212	325	0.652
Tokyo (City name)	129547	886994	0.146
Tokyo Skytree (Spot)	2171	3984	0.545

Next, we compute the weight of a word k as follows:

$$S_k = \frac{L_k}{A_k} \tag{1}$$

where L_k is the frequency of k in the category "Travel" and A_k is the frequency of k in all categories of Yahoo Chiebukuro. Table I shows examples of L_k, A_k and S_k. The values of sightseeing spots, such as Tokyo Skytree, become larger and those of non-sightseeing spots, such as Tokyo, become smaller.

Finally, we compute the importance of k as follows:

$$I_k = C_k^{\lambda} \times S_k \tag{2}$$

where C_k for Yahoo Chiebukuro is the same as L_k and C_k for blogs is the frequency of k in search results from the blog search API. λ is a decay factor for the frequency and the range is 0 to 1. We determined the value experimentally. In the current system, the value is 0.125.

B. Time sequence

The second feature is the time sequence of each sightseeing spot. Each sightseeing spot usually has strong concentration periods of tourists (hot times). We assume that sightseeing spots with the similar hot time are similar. Therefore, we count the posted dates of Yahoo Chiebukuro entries which contain each sightseeing spot.

In the prototype system, we use two types of time period about time sequence information that is captured from Yahoo Chiebukuro and blogs. The first one is a half-yearly period, namely the January-June period and July-December period. The second one is a twice-a-month period, namely the 1st-15th period and 16th-31th period in a month. Our system counts the appearance frequency of each sightseeing spot from Yahoo Chiebukuro and blogs.

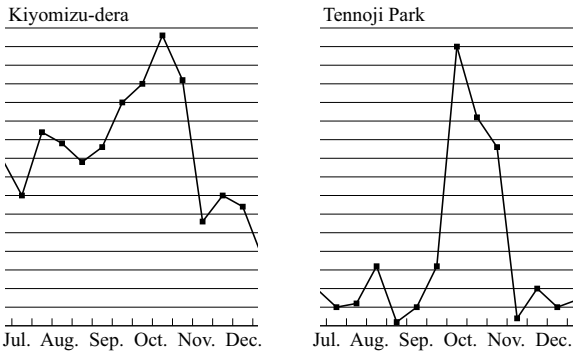

Fig. 2. Time sequence of Kiyomizu-dera and Tennoji Park.

TABLE II. PART OF CATEGORY INFORMATION ABOUT TOKYO TOWER.

Category	Frequency
Academics & Science	508
Sports, Outdoor, & Cars	42
Entertainment & Hobby	424
Health & Beauty	17
Business & Money	37
Computer technology	3

Figure 2 shows an example. The figure is time sequence of Kiyomizu-dera, a famous temple in Japan, and Tennoji Park, a botanical garden. The number of hits for these sightseeing spots increases in autumn, namely October and November. This is the foliage season. This result implies that the two spots are famous for the beautiful colored leaves. We think that this feature is suitable to compute a similarity between two sightseeing spots.

C. Category information

The third feature is a relation between a candidate word and each category of Yahoo Chiebukuro. In Section II-A, we focus on the category "Travel" to detect candidate words related to sightseeing spots. We compute the frequency of entry that includes each candidate word in other categories, such as "Academics" and "Entertainment". It implies latent topics of each sightseeing spot.

Table II shows the frequency of each category about "Tokyo Tower". In the table, the frequency of "Academics & Science" is large. The reason is that some people get interested in the architecture and construction of Tokyo Tower, namely the point of view of architectonics. This is a latent topic of Tokyo Tower.

D. Surrounding environment

The fourth feature is based on the surrounding environment of each sightseeing spot. Yahoo local search API[3] returns the

[3]http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/localsearch.html

TABLE III. SURROUNDING ENVIRONMENT OF TOKYO SKYTREE (WITHIN 5KM)

Surrounding environment	Frequency
Japanese food	4356
Western food	1356
Chinese food	849
Drugstore	839
Electrical store	976
Convenience store	1293

Standard map Night-view map

Fig. 3. Standard map and night-view map.

surrounding environment such as the number of restaurants and convenience stores located near an input query, namely a sightseeing spot. It implies a kind of local information of each spot.

Table III shows a part of the frequency list of the surrounding environment about Tokyo Skytree. There are many facilities near Tokyo Skytree. This indicates that Tokyo Skytree is located in an urban area.

E. Map information

The fifth feature is captured from map images. We collect map images of sightseeing spots by using Yahoo map API[4]. Each map image consists of a standard map and a night-view map. Figure 3 show an example of a pair of map images. We generate two color histograms, namely the standard histogram and the night histogram, from each image. The histogram contains the geographical information of each spot.

For example, the image of a standard map in Figure 3 contains blue regions. Therefore our system can estimate that the spot is located near the sea. The image of a night-view map also contains blue regions. This indicates that the spot is located in an inner urban area.

F. Similarity calculation

Our prototype system computes a similarity measure between a user input and each sightseeing spot on the basis of the five features described in the previous sections.

The word importance feature Key_s is as follows:

$$Key_s = \{x_1, x_2, \ldots, x_{n_k}\} \tag{3}$$

where n_k is the number of candidate words (67644 words). The element x in Key_s is the importance value computed by Eq 2.

The time sequence feature $Time_s$ is

$$Time_s = \{x_1, x_2, \ldots, x_{n_t}\} \tag{4}$$

where n_t is the number of time sequence patterns; 18 patterns which are the combination of the early part and latter of each year from 2004 to 2012 and 24 patterns which are twice-a-month periods in a year. The element x in $Time_s$ is the frequency in each patterns.

The category feature $Cate_s$ is

$$Cate_s = \{x_1, x_2, \ldots, x_{n_c}\} \tag{5}$$

where n_c is the number of categories without "Travel" in Yahoo Chiebukuro and 14 categories. The element x in $Cate_s$ is the frequency in each category.

The surrounding environment feature $Surr_s$ is

$$Surr_s = \{x_1, x_2, \ldots, x_{n_e}\} \tag{6}$$

where n_e is 58 types in the prototype system. The element x in $Surr_s$ is the frequency of each type, e.g., restaurant and convenience stores located within 5km about a sightseeing spot.

The map feature Map_s is expressed as follows:

$$Map_s = \{x_1, x_2, \ldots, x_{n_m}\} \tag{7}$$

where n_m is the number of color patterns in the histogram. The element x in Map_s is the frequency of each color in an image.

Finally, we integrate these five features for each sightseeing spot s as follows:

$$Vec_s = \{Key_s, Time_s, Cate_s, Surr_s, Map_s\} \tag{8}$$

Our system computes a similarity between $s1$ and $s2$ by using the COS measure.

$$Cos(s1, s2) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} y_i^2}} \tag{9}$$

where x_i and y_i are vectors in Vec_{s1} and Vec_{s2} for $s1$ and $s2$ respectively.

On the basis of this similarity calculation, we developed a prototype system[5]. Figure 4 shows the input interface of our system. In our previous work [12], we evaluated the system, and showed the effectiveness of the combination of the five features. The paper showed that the approach with only text information, namely only word importance feature, was not sufficient. By using the five features, our system could recommend intriguing, interesting and diverse sightseeing spots for inputs. A detailed discussion of the effectiveness can be found in [12].

[4]http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/static.html

[5]http://tlr.pluto.ai.kyutech.ac.jp/

Fig. 4. The prototype system.

III. Potential-of-interest days on Panoramio

One interesting feature in the prototype system is the time sequence feature described in Section II-B. It is hard to capture from text-based approaches. On the other hand, we can easily apply this approach, namely time sequence, to another information resource because it is only to extract the posted date information of the targets. Kurata [6] has proposed a concept which called potential-of-interest maps. The key idea is to visualize the sightseeing potential (or potential-of-interest) of places in a tourist area, on the basis of data about locations where previous visitors have found something impressive.

We introduce potential-of-interest days for the sightseeing spot recommendation. The time sequence is a kind of potential-of-interest days. We focused on the frequency in Yahoo Chiebukuro in the prototype system. In this paper, we focus on Panoramio[6]. Panoramio is a geolocation-oriented photo sharing website. If a user posts an photo image to it, the image is displayed on Google map on the basis of geolocation of the image. Only the images that are taken at that place are uploaded in Panoramio. Therefore it provides useful and practical information for sightseeing.

We count the appearance frequency of each sightseeing spot from Panoramio during the same periods of the Yahoo Cheibukuro, and visualize the data[7].

IV. Discussion

In this section, we compare the data from Yahoo Chiebukuro and Panoramio in terms of potential-of-interest days, and discuss the common and different points. Here we

[6]http://www.panoramio.com/

[7]http://tlr.pluto.ai.kyutech.ac.jp/panoramio/

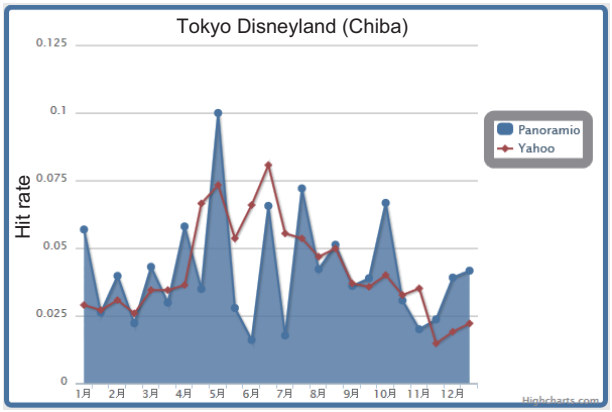

Fig. 5. Similar time sequence.

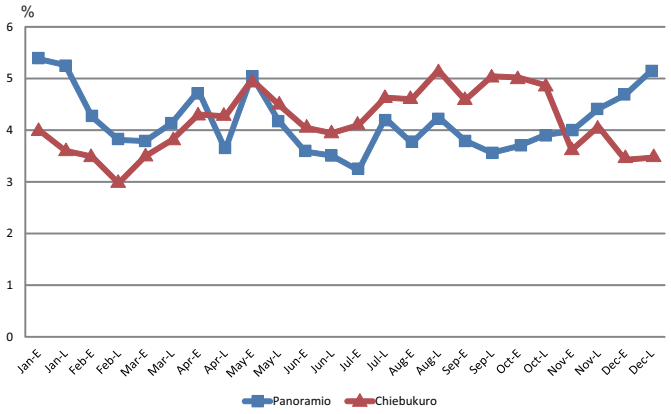

Fig. 6. Average ratio of posting in each period.

consider twice-a-month periods, namely the 1st-15th period and 16th-31th period in a month[8].

The values on the time sequence in our system are the ratio of each period to the total frequency of each spot. For Yahoo Chiebukuro, they were the number of posted Q&A pairs that related to shightseeing spots in each period. For Panoramio, they were the number of posted images on shightseeing spots in each period. The difference percentage which was less than 0.1 between Chiebukuro and Panoramio in a period was a mere 6% in the data. Even less than 0.05 was 20% of all. Many points between Yahoo Chiebukuro and Panoramio contained similar values in the time sequence. Figure 5 shows an example that contained the similar time sequence; Tokyo Disneyland. The tendency of the progress is similar although some different points exist.

On the other hand, some different points between Chiebukuro and Panoramio exist in each period as a whole. Figure 6 shows the average ratio of posting between them in each period. In the figure, the suffix "-E" and "-L" denote the early part and the latter part of each month, respectively. For Panoramio, the year change period, namely December and January, produced high posting rates. As compared with

[8]We can not obtain enough posts in the early and mid-2000s because Panoramio is a relatively-new web service. Therefore, we did not compare half-yearly periods of them.

TABLE IV. NUMBER OF DIFFERENT POINTS BETWEEN CHIEBUKURO AND PANORAMIO.

Threshold	0.1	0.15	0.2	0.3
Min_{50}	1614	826	485	222
Min_{100}	878	442	239	110
Min_{200}	304	151	98	41

Panoramio, Chiebukuro kept high values from May to October. Table IV shows the number of different points between Chiebukuro and Panoramio in several thresholds and minimum supports. The Min in the table denotes the minimum support value. For example, there were 1614 periods in the case that the different value between Chiebukuro and Panoramio was more than 0.1[9] and the number of posts was more than 50. For approximately 80% of them, the number of Panoramio's posts was larger than Chiebukuro. In addition, some sightseeing spots did not possess any posts in Chiebukuro although it was a several percent. In other words, the data of Panoramio contained posts about minor sightseeing spots as compared with that of Chiebukuro. These results show the potential efficacy of Panoramio data as the potential-of-interest days.

Figure 7 shows two examples of different time sequences between Chiebukuro and Panoramio. Figure 7 (a) is the time sequence of Koshien Stadium, which is located in Kobe, Hyogo Prefecture. In Japan, an annual high school baseball tournament held in this stadium during the summer. It is one of the most popular sport events. Therefore, the number of postings during this period becomes large because entries of Chiebukuro contain many topics[10].

Figure 7 (b) is the time sequence of Niji-no-matsubara, which is an old pine forest in Saga Prefecture. It is a scenic and popular tourist spot. The time sequence of Panoramio contained some salient points in March, April and November although that of Chiebukuro was nearly flat. We verified posted photos in the periods manually. Some photos about cherry blossoms were posted in March and April. Photos in November were related to the autumn color of leaves. These characteristics could not be captured from the time sequence of Chiebukuro. This result shows the effectiveness of information from Panoramio as the potential-of-interest days.

Next, we computed burst points in Chiebukuro and Panoramio. Here the burst is a point where the difference between a current period and the previous period exceeds a certain threshold value. We compared four thresholds; 0.1, 0.2, 0.3 and 0.4. Table V shows the result. For example, the number of burst points of Chiebukuro was 2896 in the case that the threshold was 0.1. This result shows that the time sequence of Panoramio tends to move up and down. In other words, the information on Panoramio tends to yield obvious bursts in time sequence.

Information on Chiebukuro and Panoramio possesses different characteristics. Panoramio data are posted from persons that are there without a doubt. Therefore, it tends to indicate the potential-of-interest days. On the other hand, Chiebukuro data cover a wider field than Panoramio because entries on

[9]For example, it denotes the difference of "Hit rate" between Chiebukuro and Panoramio in Figure 5 and Figure 7.

[10]Note that we did not analyze any text information for the time sequence feature. It is just the number of postings in each period.

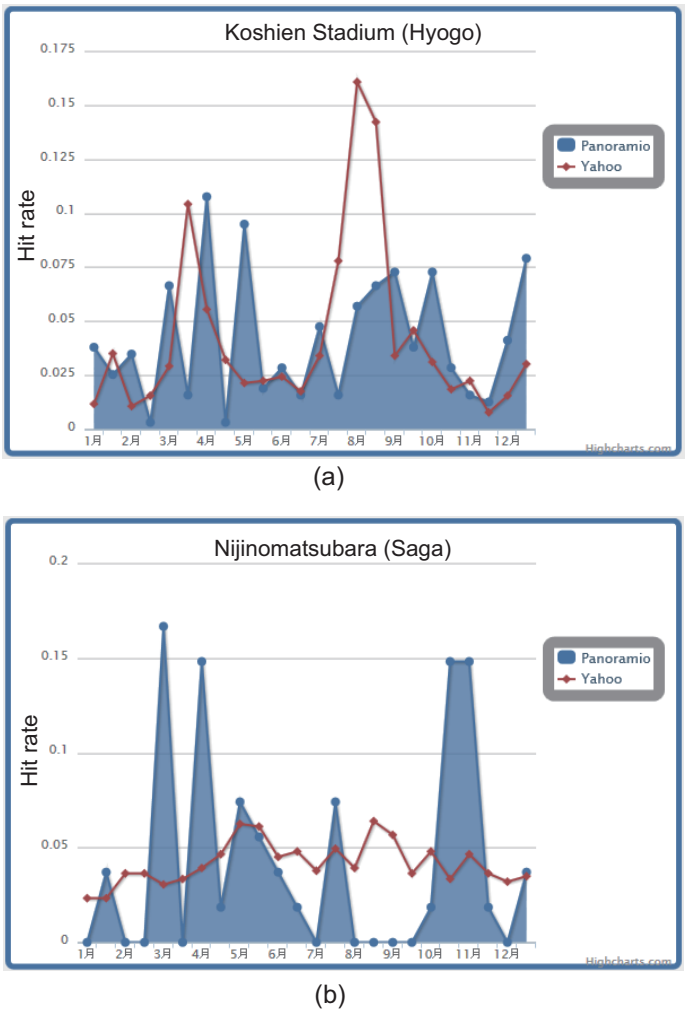

Fig. 7. Different time sequences.

TABLE V. BURST POINTS.

Threshold	0.1	0.2	0.3	0.4
Chiebukuro	2896	554	207	76
Panoramio	11842	4567	2363	1352

Chiebukuro are posted from persons that are not there, have been there and want to go there. The wide-coverage information is important for recommendation, especially local spots, although it also contains much noise information. The suitable combination of these information resources has an important role for the sightseeing spot recommendation system. This is important future work.

V. CONCLUSION

In this paper, we focused Panoramio as a new time sequence feature for our system. We compared two information resources; Yahoo Chiebukuro, which was used in the previous system as the time sequence feature, and Panoramio. We analyzed the common parts and the different parts of them. The data from Panoramio contained information that could not be captured from Yahoo Chiebukuro. The data from Chiebukuto covered a wider field than Panoramio although they contained

noise information. It shows the potential merits of combination of Yahoo Chiebukuro and Panoramio for sightseeing spot recommendation.

We have proposed methods for tourism information analysis using Twitter [11], [10]. Incorporating them to the recommendation system in this paper is important future work. In this paper, we verified the effectiveness of the new feature, namely a time sequence on Panoramio, as the potential-of-interest days. We need to consider other information resources as a new feature for the improvement of our system.

REFERENCES

[1] Yuya Kanazawa, Yosuke Hidaka, and Katsuhiko Ogawa. Destination retrieval system using an association retrieval method. *International Journal of Future Computer and Communication*, 2(3):169–173, 2013.

[2] Hidenori Kawamura, Kenji Suzuki, Masahito Yamamoto, and Hitoshi Matsubara. Tourism informatics (special feature: New informatics). *IPSJ Magazine*, 51(6):642–648, 2010.

[3] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka. Blog map of experiences: Extracting and geographically mapping visitor experiences from city blogs. In *Lecture Notes in Computer Science 3806*, pages 496–503. Springer-Verlag, 2005.

[4] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka. Mining and visualization of visitor experiences from urban blogs. In *Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006)*, 2006.

[5] Yohei Kurata. Interactive assistance for tour planning. In *Spatial Cognition 2010, Lecture Notes in Artificial Intelligence 6222*, pages 289–302, 2010.

[6] Yohei Kurata. Potential-of-interest maps for mobile tourist information services. In *ENTER 2012 (Information and Communication Technologies in Tourism 2012)*, pages 239–248, 2012.

[7] Yohei Kurata and Tatsunori Hara. Ct-planner4: Toward a more user-friendly interactive day-tour planner. In *ENTER 2014 (Information and Communication Technologies in Tourism 2014)*, pages 73–86, 2014.

[8] Kohya Okuyama and Keiji Yanai. A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web. In *Proceedings of Pacific-Rim Conference on Multimedia*, 2011.

[9] Hajime Saito. Analysis of tourism informatics on web (special issue: Tourism informatics and artificial intelligence). *Journal of the Japanese Society for Artificial Intelligence*, 26(3):234–239, 2011.

[10] Kazutaka Shimada, Shunsuke Inoue, and Tsutomu Endo. On-site likelihood identification of tweets for tourism information analysis. In *Proceedings of 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012)*, 2012.

[11] Kazutaka Shimada, Shunsuke Inoue, Hiroshi Maeda, and Tsutomu End. Analyzing tourism information on twitter for a local city. In *Proceedings of SSNE2011, International Workshop on Innovative Tourism*, 2011.

[12] Kazutaka Shimada, Hisashi Uehara, and Tsutomu Endo. Sightseeing location recommendation system based on collective intelligence. *Society for Tourism Informatics*, 10, 2014.