Available online at: http://tsdr.psdku.unpad.ac.id
Tourism and Sustainable Development Review Journal (TSDR)
Volume 1 Number 1 (2020): 57-64

Tourism Analysis Using User-Generated Content: A Case Study of Foreign Tourists Visiting Japan on TripAdvisor

Suguru Tsujioka[1], Kojiro Watanabe[2], Akihiro Tsukamoto[3]

[1] Department of Management and Information Science, Shikoku University, Japan
[2,3] Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Japan

Abstract

In recent years, online travel service platforms such as TripAdvisor have been actively used by tourists. These services include user-generated content, which is vast and difficult to interpret manually. Several previous studies used user-generated content (e.g., social networking services and TripAdvisor) for tourism analysis. Most of these studies did not perform a systematic text analysis. In this study, we propose a method of analyzing this content to understand the characteristics of sightseeing attractions. Specifically, we analyzed the reviews of foreign tourists who visited Japanese sightseeing attractions. The review data were collected from TripAdvisor. First, a correspondence analysis was conducted to understand the similarities between sightseeing attractions. Next, a co-occurrence network analysis was conducted to derive the theme clusters for understanding the characteristics of sightseeing attractions based on the words in the review. Finally, individual analyses were conducted based on the description of the derived themes at each sightseeing attraction. The results of the analyses demonstrate that the proposed method is effective for comprehending the characteristics of each sightseeing attraction. The proposed method is useful when using user-generated content for tourism analysis.

Keywords: *user-generated content, TripAdvisor, correspondence analysis, co-occurrence network*

INTRODUCTION

Tourism has been transformed recently by web services. For travel preparation, the ratio of travelers buying travel package products at travel agencies is decreasing, and the ratio of those booking airline tickets and hotels using web services is increasing. When collecting information on travel destinations, the ratio of travelers who bring a paper guidebook to the local area is decreasing, and the ratio of those searching for information online on-demand is increasing. Using online services, travelers can arrange transportation and accommodation inexpensively and access destination information quickly. Under these circumstances, online travel service platforms such as Booking.com (Booking.com, 1996) and TripAdvisor (TripAdvisor LCC, 2017) are prospering.

These online travel service platforms include a massive amount of user-generated content—such as reviews and ratings—from people who have visited each sightseeing attraction. For example, in the section on TripAdvisor about the Fushimi Inari Shrine in Japan, 20,664 reviews (of which 11,233 are written in English) from foreign travelers are posted (as of January 24, 2020). These visitor opinions are often browsed by prospective tourists pondering the destination because they consider the information neutral, unlike their perceptions of marketing by stakeholders. This content represents one type of big data in the tourism industry.

Corresponding author
tsujioka@keiei.shikoku-u.ac.jp
DOI: https://doi.org/10.31098/tsdr.v1i1.9

In this study, we analyze the user-generated content on TripAdvisor, an online travel service platform, to understand the characteristics of each sightseeing attraction in Japan as perceived by foreigners. Understanding these characteristics is useful in fields such as inbound tourism marketing, city planning, and civic activities. Previously, questionnaires have been used for obtaining data regarding the characteristics. However, administering questionnaires and performing analyses are expensive. Therefore, we use user-generated content to obtain factors regarding sightseeing attraction characteristics.

LITERATURE REVIEW

We analyzed Twitter, a source of user-generated content, to help with city planning and tourism. We estimated the characteristics of towns based on tweets (posts on Twitter) (Tsujioka, 2016). In the study, the relationship between tweets and posted locations was verified using correspondence analysis and self-organizing maps. Seven towns in Tokyo were selected as verification targets, and the characteristics of each town were obtained from tweets. In another study, we estimated the posting user's place of residence based on tweets (Tsujioka, Kondo, & Watanabe, 2016). The estimation method included latent semantic analysis and a decision tree with machine learning, enabling the residence of each posting user to be classified with at least 60% accuracy. Analyzing user-generated content can yield abundant information about the subject in question.

Many studies have analyzed TripAdvisor. O'Connor (2010) coded the word data used in reviews to specific themes to obtain the context of the word or comment. These studies have found that the posting users on TripAdvisor prioritize hotel location, room size, and staff (service) in hotel ratings. Miguéns, Baggio, and Costa (2008) analyzed TripAdvisor using the component-based structural equation modeling approach of partial least squares. The study examined travel consumers' perceptions regarding the credibility of user-generated content sources and how such perceptions influence attitudes and intentions toward the use of user-generated content for travel planning. The results of the examination demonstrated that perceived trustworthiness positively influences attitudes toward using user-generated content for travel planning.

User-generated content can provide extensive information about posting users' perceptions. In particular, the analysis of travel platforms such as TripAdvisor can reveal the evaluation axis of visitors to hotels and sightseeing attractions. Accordingly, the user-generated content on TripAdvisor (instead of questionnaires) can be used as base data for tourism analysis.

RESEARCH METHOD

In this study, we analyzed the 15 sightseeing attractions in Japan identified as most popular with foreign tourists on TripAdvisor (Tsunagu Japan, 2019). Table 1 presents the URL of each sightseeing attraction's introduction on TripAdvisor. We collected all reviews written in English as user-generated content from the webpages at these URLs. Some of the reviews written in English were written by Japanese residents, so these were excluded from the analysis database in this study. We analyzed 37,488 reviews, as presented in Table 2.

Several stop words were defined and excluded before conducting our analysis because these words frequently appear in reviews about specific sightseeing attractions and create excessive co-occurrence structures. The following stop words were established:

(1) Functional words unrelated to sentence meaning (e.g., articles and particles)
(2) Words included in the name of the sightseeing attractions (e.g., Fushimi and shrine)
(3) Words indicating the location of the sightseeing attractions (e.g., Kyoto and Kamakura)

By excluding these words before conducting the analysis, the characteristics of each sightseeing attraction could be characterized more accurately.

FINDINGS AND DISCUSSION

In this study, correspondence, co-occurrence network, and cross-tabulation analysis were performed. KH Coder (Higuchi, 2001) was used for much of the analyses. It is a free software for quantitative content analysis or text mining. The outline and results of the above three analyzes are described below.

Correspondence Analysis

The collected reviews were analyzed using constituent word-based correspondence analysis to obtain an approximate overview of the target sightseeing attractions. The analysis considered words that occurred more than 1,000 times; consequently, 108 words were identified as factors in the analysis. The results of the correspondence analysis are depicted in Figure 1, in which squares represent sightseeing attractions, and circles represent words identified as factors of the reviews. The size of each square denotes the number of reviews for the sightseeing attractions, and the size of each circle denotes the frequency of the word.

Ranking	Sightseeing attraction (Loation)	URL (after https://www.tripadvisor.com/)
1	Fushimi Inari-taisha Shrine (Kyoto, Kyoto Prefecture)	Attraction_Review-g14124535-d321456-Reviews-Fushimi_Inari_taisha_Shrine-
2	Hiroshima Peace Memorial Museum (Hiroshima, Hiroshima Prefecture)	Attraction_Review-g298561-d320360-Reviews-Hiroshima_Peace_Memorial_Museum-
3	Miyajima and Itsukushima Shrine (Hatsukaichi, Hiroshima	Attraction_Review-g1022438-d1161271-Reviews-Miyajima-Hatsukaichi_Hiroshima_Prefecture_Chugoku.html
4	Todai-ji Temple (Nara, Nara Prefecture)	Attraction_Review-g298198-d319876-Reviews-Todai_ji_Temple-Nara_Nara_Prefecture_Kinki.html
5	The Hakone Open-Air Museum (Hakone, Kanagawa Prefecture)	Attraction_Review-g298171-d320696-Reviews-The_Hakone_Open_Air_Museum-
6	Shinjuku Gyoen National Garden (Shinjuku, Tokyo Prefecture)	Attraction_Review-g1066457-d479258-Reviews-Shinjuku_Gyoen_National_Garden-
7	Sanjusangendo Temple (Kyoto, Kyoto Prefecture)	Attraction_Review-g14124527-d321411-Reviews-Sanjusangendo_Temple-
8	Mt.Koyasan Okunoin (Koya, Wakayama Prefecture)	Attraction_Review-g1121341-d324938-Reviews-Koyasan_Okunoin-Koya_cho_Ito_gun_Wakayama_Prefecture_Kinki.html
9	Himeji Castle (Himeji, Hyogo Prefecture)	Attraction_Review-g298191-d320231-Reviews-Himeji_Castle-Himeji_Hyogo_Prefecture_Kinki.html
10	Kinkakuji Temple (Kyoto, Kyoto Prefecture)	Attraction_Review-g298564-d321400-Reviews-Kinkakuji_Temple-Kyoto_Kyoto_Prefecture_Kinki.html
11	Kenrokuen Garden (Kanazawa, Ishikawa Prefecture)	Attraction_Review-g298115-d321201-Reviews-Kenrokuen_Garden-Kanazawa_Ishikawa_Prefecture_Hokuriku_Chubu.html
12	Naritasan Shinsho-ji Temple (Narita, Chiba Prefecture)	Attraction_Review-g298161-d1314231-Reviews-Naritasan_Shinsho_ji_Temple-Narita_Chiba_Prefecture_Kanto.html
13	Hase-dera Temple (Kamakura, Kanagawa Prefecture)	Attraction_Review-g303156-d319981-Reviews-Hase_dera_Temple-Kamakura_Kanagawa_Prefecture_Kanto.html
14	Nara Park (Nara, Nara Prefecture)	Attraction_Review-g298198-d319880-Reviews-Nara_Park-Nara_Nara_Prefecture_Kinki.html
15	Nikko Tosho-gu Shrine (Nikko, Tochigi Prefecture)	Attraction_Review-g298182-d1311878-Reviews-Nikko_Tosho_gu-Nikko_Tochigi_Prefecture_Kanto.html

Table 1. Target sightseeing attractions and URLs of introductions on TripAdvisor

Ranking	Sightseeing attraction	The number of the review				The number of the review for our analysis
		English	Japanese	Other language	Total	
1	Fushimi Inari-taisha Shrine	13,232	3,528	7,444	24,204	12,624
2	Hiroshima Peace Memorial Museum	3,935	1,200	1,775	6,910	3,703
3	Miyajima and Itsukushima Shrine	2,099	1,387	1,284	4,770	1,949
4	Todai-ji Temple	2,397	1,365	1,823	5,585	2,172
5	The Hakone Open-Air Museum	1,468	833	396	2,697	1,418
6	Shinjuku Gyoen National Garden	3,545	1,234	1,332	6,111	3,403
7	Sanjusangendo Temple	1,615	1,232	1,059	3,906	1,493
8	Mt.Koyasan Okunoin	425	526	286	1,237	394
9	Himeji Castle	1,910	2,030	1,316	5,256	1,783
10	Kinkakuji Temple	8,731	1,955	6,196	16,882	8,442
11	Kenrokuen Garden	1,985	2,566	1,130	5,681	1,933
12	Naritasan Shinsho-ji Temple	811	1,030	258	2,099	751
13	Hase-dera Temple	478	813	398	1,689	454
14	Nara Park	2,762	1,126	1,714	5,602	2,533
15	Nikko Tosho-gu Shrine	870	1,669	717	3,256	804

Table 2. Number of reviews analyzed

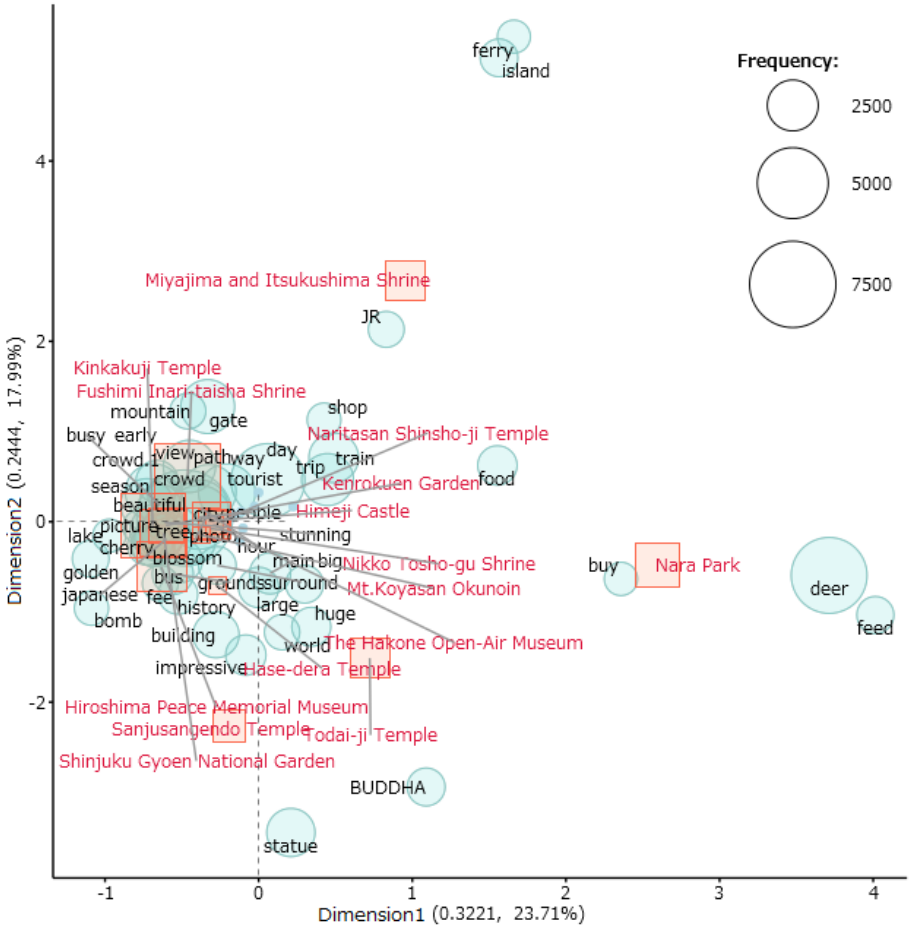

Figure 1. Correspondence analysis results

Figure 1 illustrates that Miyajima and the Itsukushima Shrine, Nara Park, Todaiji Temple, and Sanjusangendo Temple are distinct from the other sightseeing attractions. Miyajima and the Itsukushima Shrine are associated with "ferry" and "island" as characteristic words because the locations are offshore. Similarly, Nara Park is associated with "deer" and "feeding," Todaiji Temple is associated with "BUDDHA," and Sanjusangendo Temple is associated with "statue." These characteristics are not found in other sightseeing attractions.

Clustering with a Co-Occurrence Network of Words

Next, the words were clustered using a co-occurrence network to clarify the factors for sightseeing attraction characteristics. A co-occurrence network was constructed using all of the review texts (Figure 2). Words having a minimum frequency of 700 were used in the network analysis, resulting in 148 words being targeted. The network depicted in Figure 2 was drawn using the 60 words having the highest Jaccard coefficients. Based on this clustering, 15 clusters were obtained.

We focused on 6 (of the 15) clusters that had (a) more than three nodes (otherwise, it would be difficult to define a theme) and (b) a general theme that did not depend on particular sightseeing

attractions. The six themes were "Time," "Fee," "Historical," "Nature," "Transportation," and "Crowd." The constituent words of those clusters are depicted in Figure 2.

Characteristics of Each Sightseeing Attraction

Finally, we performed cross-tabulation based on these six themes to quantitatively clarify the characteristics of each sightseeing attraction (Table 3). Each section in the table lists the number of reviews, including the words representing the theme and the ratio to the total number of reviews. In the cross-table, a p-value of less than 0.01 is considered statistically significant, and the chi-square values demonstrate sufficient degrees of freedom. The sightseeing attractions exhibit statistically significant differences.

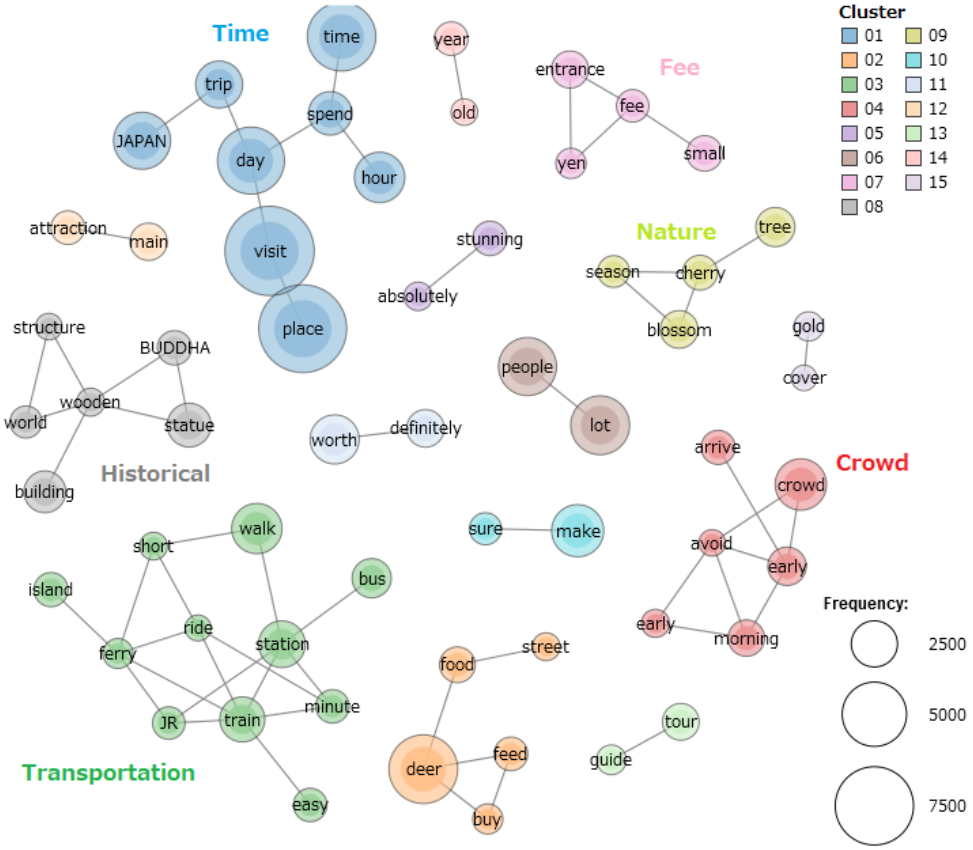

Figure 2. Co-occurrence network for all review texts

	Time		Fee		Historical		Nature		Transportation		Crowd		N of Reviews
Fushimi Inari-taisha Shrine	8,464	(67.05%)	872	(6.91%)	458	(3.63%)	324	(2.57%)	5,838	(46.25%)	5,468	(43.31%)	12,624
Hiroshima Peace Memorial Museum	2,396	(64.70%)	307	(8.29%)	488	(13.18%)	15	(0.41%)	475	(12.83%)	258	(6.97%)	3,703
Miyajima and Itsukushima Shrine	1,618	(83.02%)	141	(7.23%)	84	(4.31%)	40	(2.05%)	1,657	(85.02%)	348	(17.86%)	1,949
Todai-ji Temple	1,472	(67.77%)	321	(14.78%)	1,542	(70.99%)	70	(3.22%)	622	(28.64%)	311	(14.32%)	2,172
The Hakone Open-Air Museum	965	(68.05%)	105	(7.40%)	198	(13.96%)	34	(2.40%)	334	(23.55%)	73	(5.15%)	1,418
Shinjuku Gyoen National Garden	2,394	(70.35%)	774	(22.74%)	134	(3.94%)	1,385	(40.70%)	1,042	(30.62%)	568	(16.69%)	3,403
Sanjusangendo Temple	899	(60.21%)	112	(7.50%)	1,117	(74.82%)	24	(1.61%)	291	(19.49%)	115	(7.70%)	1,493
Mt.Koyasan Okunoin	303	(76.90%)	26	(6.60%)	43	(10.91%)	91	(23.10%)	191	(48.48%)	53	(13.45%)	394
Himeji Castle	1,359	(76.22%)	184	(10.32%)	350	(19.63%)	160	(8.97%)	721	(40.44%)	302	(16.94%)	1,783
Kinkakuji Temple	4,363	(51.68%)	749	(8.87%)	696	(8.24%)	435	(5.15%)	2,005	(23.75%)	2,250	(26.65%)	8,442
Kenrokuen Garden	1,476	(76.36%)	192	(9.93%)	71	(3.67%)	697	(36.06%)	431	(22.30%)	335	(17.33%)	1,933
Naritasan Shinsho-ji Temple	595	(79.23%)	64	(8.52%)	134	(17.84%)	29	(3.86%)	358	(47.67%)	113	(15.05%)	751
Hase-dera Temple	306	(67.40%)	94	(20.70%)	197	(43.39%)	28	(6.17%)	167	(36.78%)	49	(10.79%)	454
Nara Park	1,579	(62.34%)	319	(12.59%)	201	(7.94%)	154	(6.08%)	1,030	(40.66%)	257	(10.15%)	2,533
Nikko Tosho-gu Shrine	639	(79.48%)	101	(12.56%)	230	(28.61%)	69	(8.58%)	270	(33.58%)	147	(18.28%)	804
total	28,828	(65.73%)	4,361	(9.94%)	5,943	(13.55%)	3,555	(8.11%)	15,432	(35.19%)	10,647	(24.28%)	43,856
chi-square	495.599**		768.113**		10204.494**		6856.474**		3804.453**		3844.434**		

Table 3. Cross-table based on a theme

The theme “Time” is mentioned frequently for all sightseeing attractions because of the inclusion of a wide range of words such as “place,” “trip,” and “Japan.” Moreover, the attractions include various time-related phrases, such as length of stay and travel time. Therefore, although clustering with co-occurrence networks is too coarse for this theme, it has the advantage of providing non-arbitrary classification results.

The theme “Fee” has a high ratio for Shinjuku Gyoen National Garden and Hasedera Temple. Many of these reviews express that these attractions are reasonably priced (e.g., “I thought was fair” and “It’s worth the price”). The target words were used in each review using the Key Word In Context (KWIC) method, which provides support with interpreting the meaning of words in context (Garrett, 2006).

The theme “Historical” has a high ratio for three temples: Todaiji, Sanjusangendo, and Hasedera. The reviews of these temples include many mentions of "BUDDHA" and "statue," as depicted in Figure 1. As presented in Table 3, these three sightseeing attractions have no similarities, except for the "Historical" theme. Consequently, the theme "Historical" contributes significantly to the proximity of these three sightseeing attractions, as depicted in Figure 1.

CONCLUSION

In this study, user-generated content on TripAdvisor was analyzed to obtain the characteristics of sightseeing attractions. The analysis method included correspondence and co-occurrence network analyses. Correspondence analysis was useful for understanding similarities between sightseeing attractions, while co-occurrence network analysis identified the themes that define sightseeing attractions and clarified their constituent words. These methods provided an approximate overview of the sightseeing attractions. However, the results obtained by these two methods did not yield clear relationships between the words and the sightseeing attractions.

Next, cross-tabulation was performed using the themes obtained from the co-occurrence network. We presented the sightseeing attractions quantitatively based on those themes. However, the interpretation of the words representing the themes should be judged in context, accomplished using the KWIC method.

User-created content is a big data alternative to questionnaires and is useful for tourism analysis. However, it is difficult to interpret the abundant data manually. Using the method proposed in this study can help researchers to understand user-generated content quantitatively. Today, there are many sources of user-generated content, including social networking services. Tourists often post their impressions on content sites, and our proposed method supports analyzing this content.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP 19K12569.

REFERENCES

Booking.com. (n.d.). Retrieved from https://www.booking.com/index.html

Garrett, J. (2006). KWIC and dirty? Human cognition and the claims of full-text searching. *Journal of Electronic Publishing*, *9*(1).

Higuchi, K. (2001). KHcoder. Retrieved from https://khcoder.net/en/

Miguéns, J., Baggio, R., & Costa, C. (2008). Social media and tourism destinations: TripAdvisor case study. *Advances in Tourism Research*, *26*(28), 1-6.

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management*, *19*(7), 754-772.

TripAdvisor LCC. (2017). TripAdvisor. Retrieved from https://tripadvisor.mediaroom.com/

Tsujioka, S. (2016). Town characteristics estimation using geotagged Twitter data-A case study in the Tokyo area. In *Proceedings of International Conference on Civil, Architectural, and Environmental Engineering* (pp. 143-147).

Tsujioka, S., Kondo, A., & Watanabe, K. (2016). Estimation of residence information of Twitter users based on their posted messages: Data for tourism development. *International Journal of Research in Chemical, Metallurgical, and Civil Engineering*, *3*(1), 180-183.

Tsunekawa, K. (2019, August 5). The top 30 sightseeing attractions in Japan as voted by international travelers. Retrieved from https://www.tsunagujapan.com/top-30-attractions-in-japan-for-international-travelers/