Travellers' Behaviour Analysis Based on Automatically Identified Attributes from Travel Blog Entries

Kazuki Fujii[1], Hidetsugu Nanba[1], Toshiyuki Takezawa[1], Aya Ishino[2], Manabu Okumura[3] and Yohei Kurata[4]

[1] Hiroshima City University, 3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
{fujii, nanba Travellers' Behaviour Analysis Based on Automatically Identified Attributes from Travel Blog Entries, takezawa}@ls.info.hiroshima-cu.ac.jp
[2] Hiroshima University of Economics, 5-37-1 Gion, Asaminami-ku, Hiroshima 731-0192, Japan
ay-ishino@hue.ac.jp
[3] Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan
oku@pi.titech.ac.jp
[4] Tokyo Metropolitan University, 1-1 Minamiosawa, Hachioji, Tokyo, 192-0397, Japan
ykurata@tmu.ac.jp

Abstract. We propose a method to analyse travellers' behaviour using automatically identified travellers' attributes, such as gender and language, from travel blog entries. We consider that travel blog entries are a useful information source for obtaining travel information, because many bloggers' describe their travel experiences in them. Several studies have analysed travellers' behaviour using travel blog entries. However, they used a small number of manually identified travellers' (bloggers') attributes. In our work, we identify travellers' attributes automatically using natural language processing techniques, and conduct a large-scale travellers' behaviour analysis.

Keywords: travel blog, behaviour analysis, travellers' attributes,

1 Introduction

Being aware of travellers' needs is crucial for tourism planning. Traditionally, such analyses were conducted using questionnaires, but these are costly and time-consuming. Recently, travel blog entries have been used instead of questionnaires. In travel blog entries, various travellers' experiences and opinions are described, and they can help identify travellers' needs. For example, Wenger [1] analysed travel blog entries written by travellers visiting Austria, and found that many females visited Austria to enjoy dining experiences. However, few blog entries were used in this analysis, because the analysis was conducted manually. In this paper, we propose a method for analysing travellers' behaviour from vast numbers of blog entries using natural language techniques. In our approach, we identify travellers' (bloggers') at-

tributes automatically using method based on machine learning. Then, we use the attributes to characterize travellers' behaviour.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our method of identifying travellers' (bloggers') attributes in travel blog entries. To investigate the effectiveness of our method, we conducted some experiments, and Section 4 reports the experimental results. Using our method, we conducted travellers' behaviour analysis, and Section 5 reports the results. We present some conclusions in Section 6.

2 Related Work

Questionnaires have been frequently used in marketing surveys to determine tourist policies. Xia *et al.* [2] asked 464 visitors to Phillip Island about their gender, tourist spots they planned to visit, and their residence, and conducted behaviour analysis using decision-tree methods. Similarly, Jonsson *et al.* [3] also conducted a questionnaire investigation of 163 visitors to a Caribbean island, and analysed travellers' behaviour. However, it is costly and quite time-consuming to conduct questionnaire investigations. Therefore, we aim to analyse travellers' behaviour automatically from travel blog entries, because there are many descriptions of travellers' experiences and many opinions in travel blog entries.

The number of studies using travel blog entries for travellers' behaviour analysis has been increasing. Mack *et al.* [4] and Akehurst [5] reported that travel blog entries are useful for tourism marketing, although they are not as reliable as traditional word of mouth. Li and Wang [6] analysed impressions of China from Taiwanese travellers using travel blog entries in a Chinese travel portal site. They manually classified blog entries to several categories, such as "scenery", "purchase", and "stay", and analysed entries in each category. They reported that the impressions of hot springs were good, and suggested that these hot springs should be promoted as a major tourist attraction in China. Classifying travel blog entries for travellers' behaviour analysis is a common point with our work. However, our work classifies entries automatically, which enables us to conduct a larger-scale analysis.

In related research using travel blog entries, Wenger [1] conducted travellers' behaviour analysis based on travellers' attributes, such as gender and age. They used the dataset of TravelBlog[1], in which some bloggers reveal their attributes in their profiles. However, few bloggers explicitly describe these attributes. We therefore identify bloggers' attributes automatically, and use them for travellers' behaviour analysis.

Recently, several methods for identifying bloggers' attributes, such as gender, age [7, 9], and residential area [8] have been proposed. We identify travellers' gender based on Ikeda's method. With respect to travel-related documents, Saeki *et al.* [10] identified travellers' languages in each geotagged tweet using a Java language-detection[2] program (langdetect), and conducted foreign travellers' behaviour analysis

[1] https://www.travelblog.org/

[2] https://code.google.com/p/language-detection/

in Japan. In the same way, we also apply the program to travel blog entries, and use them for travellers' behaviour analysis.

3 Automatic Identification of Travellers' Attributes

We aim to analyse travellers' behaviour based on the travellers' gender, language, and behaviour using the TravelBlog dataset. Unfortunately, this information is not explicitly written in the dataset. Therefore, we identify attributes automatically using natural language processing techniques. We explain how to identify a blogger's (a traveller's) gender, language, and behaviour (content type of each blog entry) in Sections 3.1, 3.2, and 3.3, respectively.

3.1 Identification of Blogger's Gender

To identify bloggers' genders, we propose three methods: (1) semi-supervised learning approach based on Ikeda's method [7] (SSL), (2) cue-word-based method (CUE), and (3) a combination of the CUE and SSL methods.

Ikeda *et al.* (2008) assumed that each blogger has a writing style based on his/her attributes, such as gender and generation. They identified these attributes from two kinds of data: (i) a small number of blog entries in which bloggers explicitly describe their attributes; and (ii) a large number of entries with no explicit attributes using a semi-supervised technique. They experimentally confirmed that their approach obtained an accuracy score of 0.890 for the identification of bloggers' genders, while a supervised approach using the data (i) obtained 0.760. Therefore, we examine the semi-supervised approach based on the Ikeda's method (SSL).

In addition to the SSL method, we also examine a cue-word-based method. This method focuses on the differences in words that male and female bloggers use. We created two cue word lists by collecting words that are frequently used in blog entries written by male or female bloggers, and used them for the identification of bloggers' genders (CUE).

As our third approach, we examine the combination of the SSL and CUE methods. The basic procedure is the same as that for the SSL method, except that the SSL+CUE method does not use all words in blog entries, only those that appear in two cue-word lists. We believed that using the cue-word lists could identify the features of blog entries more clearly, and would improve the SSL method. In the experiment in Section 4, we will report that our SSL+CUE method did actually improve the SSL method.

3.2 Identification of Travellers' Languages

Many entries in the TravelBlog database were written in English, but other languages, such as French and German, are also used. We focus on the languages that bloggers used, and analyse the travellers' behaviours in each city they visited in terms of their languages. This analysis is particularly useful for tourism policy. For example, if we

find that there are many visitors who use French in a particular city, then it is effective for the city authorities to provide signs in French.

To identify which language a blogger uses in a blog entry, we use the langdetect program. This library automatically estimates probabilities of each language for a given text. We attempt to identify travellers' languages in two different ways: (1) identify the language that has the highest probability score (Top method); and (2) identify all languages having probabilities higher than a threshold value (Threshold method). This threshold value was determined in a pilot study.

3.3 Identification of Content Type

Even when travellers visit the same destination, the purpose of the visit is not always the same. Some travellers might visit tourist spots as their primary purpose, while others might visit to enjoy local dining. We aim to make the visitors' purposes clear by identifying the content type of each blog entry, as shown in Table 1. These types were originally proposed by Fujii *et al.* [11], and they devised a system that identifies one or more content types relevant to a given blog entry written in Japanese. In their method, they first collected cue words that are useful for identifying content types using the information gain (IG) method. Second, they applied Support Vector Machine (SVM) to identify content types using cue words as features for the SVM. In our work, we developed a system that can identify content type for a given blog entry written in English using two different methods. In the first method, we collect 100 cue words for each content type using IG and apply the SVM, using the same procedure as that used by Ishino *et al.* for Japanese blog entries. In the second method, we employ a machine translation method (MT) by combining the first method with a Japanese identifier devised by Ishino *et al.* In the MT method, we translate an English blog entry into Japanese using the Microsoft Translator API[3], and then identify its content type using the Japanese identifier (MT). Then, we use the result as one of the features of SVM, and identify the content type of the given English blog entry (IG+MT).

Table 1. Content types and their descriptions

Content type	Criterion
Watch	Sightseeing for watching enjoyment
Experience	Experience (scuba diving, dance)
Buy	Shopping or souvenir stores
Dine	Drinking and dining
Stay	Accommodation

[3] https://datamarket.azure.com/dataset/bing/microsofttranslator

4 Experiments

To confirm the effectiveness of our method, we conducted three experiments: (1) identification of blogger's gender; (2) identification of the language used; and (3) identification of the content type of each blog entry. We describe them in Sections 4.1, 4.2, and 4.3, respectively. In the experiments, we used blog entries from the TravelBlog dataset, which we mentioned in Section 3.

4.1 Identification of Traveller's Gender

4.1.1 Experimental Settings

Data

We used 228 bloggers for this experiment, 77 males and 151 females. We obtained the gender information from each blogger's profile. In this experiment, we used blog entries written in English.

Machine Learning and Evaluation Measure

We employed SVM with a linear kernel for machine learning, and conducted two-fold cross-validation. We used accuracy determined by the following equation as an evaluation measure.

$$Accuracy = \frac{\begin{array}{c}Number\ of\ travellers\\ whose\ genders\ are\ correctly\ identified\end{array}}{Number\ of\ travellers}$$

Alternatives

We examined the following four methods.

- **Baseline:** Identify all bloggers as female.
- **SSL:** Semi-supervised learning based on Ikeda's method [7].
- **CUE:** Use two different cue word lists as features for machine learning. These lists were created by collecting approximately the top 100 words that appeared in blogs written by males or females.
- **SSL+CUE:** A combination of the SSL method and the CUE method. When we conduct the SSL method, we use words in the above two lists as features for machine learning.

Results

We show the experimental results in Table 2. Of the four methods, our method SSL+CUE obtained the best accuracy score.

We expected Ikeda's method to obtain a high accuracy score, because they also identified bloggers' genders, and obtained a score of 0.890. Unfortunately, their method obtained 0.667 with our data, which is much smaller than their experimental result. We consider that this must be because of the different characteristics of the two datasets. Ikeda's experiment used blog entries about various topics, such as sport, politics, automobiles, beauty salons, and sweets, while we only used blog entries

about travel. As a result, Ikeda's method could not capture the gender-related features of blog entries, and did not work well in our dataset. On the other hand, the two cue word lists assisted to capture the gender-related features of blog entries, and as a result, SSL+CUE outperformed SSL.

Table 2. Evaluation results for the identification of bloggers' genders

Method	Accuracy
Baseline	0.662 (151/228)
SSL	0.667 (152/228)
CUE	0.776 (177/228)
SSL+CUE	0.877 (195/228)

4.2 Identification of Traveller's Language

4.2.1 Experimental Settings

Data

We manually identified the language that each of 109 bloggers used in their blog entries, and used these for our experiment. The statistics are shown in Table 3. Note that some bloggers used more than one language.

Table 3. The number of bloggers for each language

Language	Bloggers	Language	Bloggers
English	83	Portuguese	1
German	10	Swedish	1
Spanish	9	Afrikaans	1
Dutch	9	Hungarian	1
French	6	Finnish	1
Danish	5	Slovene	1
Italian	2	Romanian	1
Japanese	2		

Evaluation Measure

We used recall and precision as evaluation measures. Of these, we consider that precision is more important, because a low precision score causes inaccurate analysis, as we will describe in Section 5. In addition, as the number of blog entries has been increasing, the low recall can be improved.

Alternatives

For the identification of the languages that each blogger uses, we used the langdtect program, with the following two methods.

- **Top:** languages having the highest probability among automatically identified languages by the langdetect program.
- **Threshold**: languages with probabilities higher than a threshold value, which was determined in a pilot study.

Results

We show the experimental results in Table 4. The Top method obtained the higher precision score of the two methods. In the analysis in Section 5, we used the Top method.

Table 4. Evaluation results for the identification of bloggers' languages

Method	Precision	Recall
Top	**0.972**	0.797
Threshold	0.887	0.887

4.3 Identification of Content Types of Each Blog Entry

Data

We examined using 660 travel blog entries, all written in English. For these entries, we assigned content types. A summary of the data is shown in Table 5. Here, more than one content type was assigned to several entries, so the total number of entries in Table 6 is larger than 660.

Table 5. Summary of the data using in the examination of content type identification

Content type	Buy	Dine	Experience	Stay	Watch	Other
Entries	30	97	143	61	316	155

Machine Learning and Evaluation Measures.

We employed SVM with a linear kernel. We evaluated our methods and a baseline method by recall and precision.

Alternatives

We examined the following methods.

- IG: SVM-based approach with cue words, which were identified using information gain method.
- MT: Machine translation-based approach using Microsoft translator and Japanese content-type identifiers [11].
- IG+MT: SVM-based approach using cue words and the MT result as features.
- Baseline: SVM-based approach using all words as features.

Results
We show the experimental results in Table 6. Our method, IG+MT, outperformed the baseline method by 0.449 points, confirming the effectiveness of our method.

Table 6. Evaluation results for identification of blog entry type

Method	Precision	Recall
IG	0.574	0.296
MT	0.458	0.406
IG+MT	**0.597**	0.336
Baseline	0.148	0.702

5 Analysis of Travel Blogs Based on Bloggers' Attributes

We analysed 7,490 blog entries focusing on bloggers' attributes and the content types of each blog entry. These blog entries were written by 1,302 bloggers who had visited Japan.

5.1 Basic Statistics of Visitors to Japan Based on the Automatically Identified Attributes

First, we show some basic statistics of foreign visitors to Japan based on the attributes of visitors and content types of blog entries.

Gender
We show the number of bloggers for each gender in Table 7.

Table 7. The number of automatically identified bloggers for each gender

Gender	Travellers
Male	513
Female	789

Language
We show the number of bloggers for each language in Table 8. We removed the English-speaking travellers from the results, because, the number of English-speaking travellers in TravelBlog is much larger than others, as we mentioned previously. From the results in Table 8, French and German are most often used after English.

Table 8. The number of automatically identified travellers for each language

Language	Travellers
French	22
German	13
Dutch	10
Spanish	7
Finnish	7

Content Type
We analysed the content types of the 7,490 blog entries automatically. We show the results in Table 9. The primary purpose of visitors to Japan is watching. On the other hand, only one blog entry had the content type are "buy".

Table 9. Automatically identified content types of blog entries

Content type	Entries
Buy	1
Dine	1134
Experience	315
Stay	319
Watch	3213

5.2 Analysis Based on Attributes of Travellers and Content Types

Using automatically identified travellers' attributes, we conducted travellers' behaviour analysis. In this report, we mainly focused on Japan as a test case, but it is possible to conduct the same analysis in any city in any country, because we have already identified all travellers' attributes.

5.2.1 Analysis Based on Travellers' Languages and Content Types of Travel Blog Entries

We analysed travellers' languages in each prefecture. The top three prefectures with the largest numbers of languages used by travellers are shown in Table 10.

Table 10. Top three prefectures that the number of languages

Prefecture	Visitors	Languages
Tokyo	68	18
Kyoto	31	13
Hiroshima	21	10

Among these three prefectures, in the following, we compare Kyoto and Hiroshima in terms of travellers' languages and content types of travel blog entries. Fig. 1 shows the proportion of blog entries with content type "watch" as a heat map. Red cities indicate that they have more "watch" entries. In this figure, we also show travellers' languages as bar charts[4]. From the results in Fig. 1, we found that the proportion of "watch" entries was higher in Hiroshima and Kyoto than in the other prefectures.

We read the "watch" entries in Kyoto and Hiroshima, and found that many French-speaking travellers visited world heritage locations in both prefectures. From the result, we further investigated the ratio of French-speaking travellers for each prefecture. We show the results in Fig. 2.

[4] When we created this chart, we eliminated English-speaking travellers, because most of the travellers (bloggers) in the TravelBlog dataset used English.

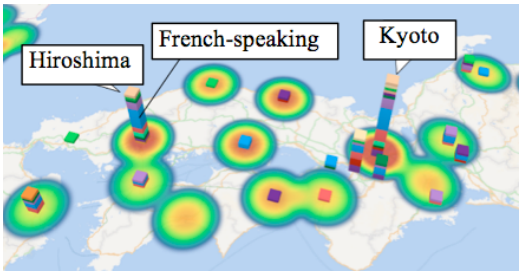


Fig. 1. The proportion of "watch" travel blog entries and travellers' languages for each prefecture

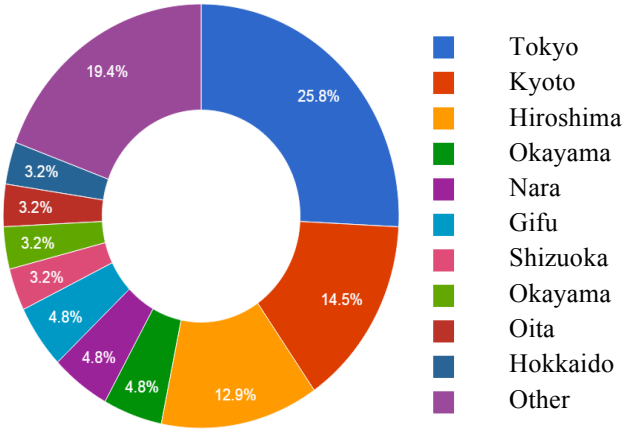


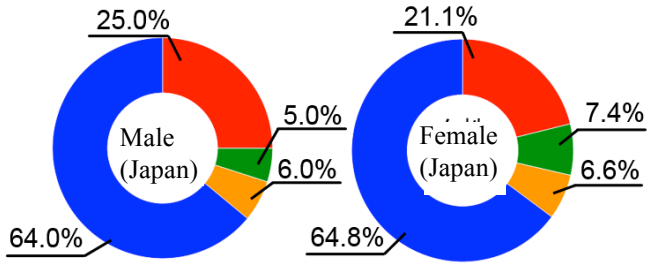
Fig. 2. The proportion of French-speaking travellers among all

From the results in Fig. 2, we found that approximately 30% of French-speaking travellers visited either Kyoto or Hiroshima. We also found that French-speaking travellers tend to visit world heritage locations in other prefectures. From these results, we can conclude that promotion of world heritage locations and providing signs in French are more important in these areas.

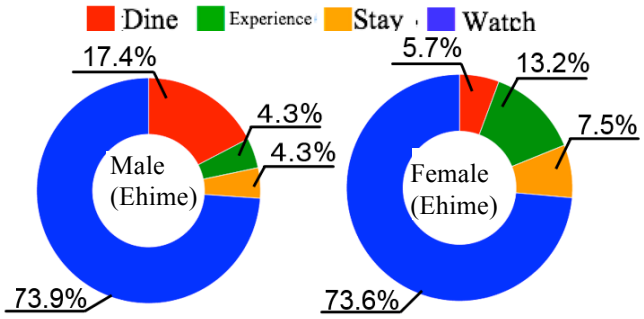
5.2.2 Analysis Based on Travellers' Gender and Content Types of Travel Blog Entries

To compare the different travel purposes between males and females, we calculated the proportions of each content type for each gender in Japan. We show the results in Fig. 3 (a). The results indicate that there are small differences between males and females. However, when we focus on particular areas, we found obvious differences between genders. Fig. 3 (b) shows the proportions of each content type for each gender in Ehime prefecture. The figure shows that the secondary purpose of males who visit Ehime is to dine, while that of females is to experience. Ehime prefecture is famous for its hot springs, and we found that many female visitors enjoyed them. From

these results, we can conclude that the purposes of visiting a particular area are not always the same for each gender, and different promotions are required. For example, in Ehime prefecture, providing toiletries and bath towels will please females, which seems likely to increase the number of female visitors to Ehime prefecture.



(a) Japan



(b) Ehime prefecture

Fig. 3. Proportions of content types of travel blog entries for each gender

6. Conclusion

In this paper, we have analysed travellers' behaviour using travel blog entries. For this analysis, we used travellers' attributes identified automatically from travel blog entries. To identify gender, our method, SSL+CUE, obtained an accuracy score of 0.877, which outperformed the baseline method by 0.223. To identify languages, we used the langdetect program, and confirmed that it obtained precision of 0.972 and recall of 0.797. To identify the content type of each blog entry, our method, IG+MT, obtained precision of 0.597 and recall of 0.327. Using 7,490 travel blog entries with these attributes, we conducted behaviour analysis of 1,302 travellers, who visited Japan. We were able to derive useful information for tourist authorities.

7. Acknowledgements

This study was carried out under support of Ministry of Internal Affairs and Communications' SCOPE (Strategic Information and Communications R&D Promotion Programme).

References

1. Wenger, A.: Analysis of Travel Bloggers' Characteristics and their Communication about Austria as a Tourism Destination, *Journal of Vacation Marketing*, 14(2), pp. 169-176 (2008)
2. Xia, J., Ciesielski, V. and Arrowsmith, C.: Data Mining of Tourists' Spatio-temporal Movement Patterns - A Case Study on Phillip Island, *Proceedings of the 8th International Conference on GeoComputation*, pp. 1-5 (2005)
3. Jonnson, C. and Devonish, D.: Dose Nationality, Gender, and Age Affect Travel Motivation? A Case of Visitors to the Caribbean Island of Barbados, *Journal of Travel and Tourism Marketing*, 25(3-4), pp. 398-408 (2008)
4. Mack, R. W., Blose, J. E. and Pan, B.: Believe it or not: Credibility of Blogs in Tourism, *Journal of Vacation Marketing*, 14(2), pp. 133-144 (2008)
5. Akehurst, G.: User Generated Content: the Use of Blogs for Tourism Organizations and Tourism Consumers, *Journal of Service Business*, 3(1), pp. 51-61 (2009)
6. Li, Y.R. and Wang, Y.Y.: Exploring the Destination Image of Chinese Tourists to Taiwan by Word-of-Mouth on Web, *Proceedings of World Academy of Science Engineering and Technology*, 7, pp. 977-981 (2013)
7. Ikeda, D., Takamura. H and Okumura, M.: Semi-Supervised Learning for Blog Classification, *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1156-1161 (2008)
8. Yasuda, N., Hirao, T., Suzuki, J. and Isozaki, H.: Identifying Bloggers' Residential Areas, *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 231-236 (2006)
9. Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.: Effects of Age and Gender on Blogging, *Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs*, pp. 199-205 (2006)
10. Saeki, K., Endo, M., Hirota, M, Kurata, Y., and Ishikawa, H.: Language-Specific Analysis of Domestic Places Visited by Foreign Tourists Using Crawled Twitter Data, *Tourism and Informatics*, 11(1), pp. 45-56 (2015) (in Japanese)
11. Fujii, K., Nanba, H., Takezawa, T., and Ishino, A. Enriching Travel Guidebooks with Travel Blog Entries and Archives of Answered Question, *Proceedings of ENTER 2016*, pp. 157-171 (2016)