Identifying Obscure Venues Using Classification of User Reviews

Masaharu Hirota
Department of Information Science
Faculty of Informatics
Okayama University of Science
Okayama-shi, Okayama
Email: hirota@mis.ous.ac.jp

Masaki Endo
Division of Core Manufacturing
Polytechnic University
Kodaira-shi, Tokyo
Email: endou@uitec.ac.jp

Hiroshi Ishikawa
Graduate School of Systems Design
Faculty of System Design
Tokyo Metropolitan University
Hino-shi, Tokyo
Email: ishikawa-hiroshi@tmu.ac.jp

***Abstract*—Today, tourism occupies an essential position in many countries as a critical industry. When sightseeing, many people visit different places such as restaurants, hotels, and tourist spots. Some of these venues, while worthwhile, are considered obscure, secret, not well-known, or having little popularity. Their extraction and recommendation are vital to improving the satisfaction of tourists. Although some studies have been proposed on extracting obscure venues based on their degree of popularity, the interest in such venues varies from person to person. In addition, these studies have defined what constitutes an obscure venue and use such criteria for venue extraction. This study proposes a method for discovering obscure venues using classifiers for identifying reviews, including obscure impressions. To achieve this goal, in this study, a model was developed to classify venues as obscure or not obscure using reviews with language indicating their obscurity. This study also analyzes the differences among venues perceived by reviewers as being obscure. We demonstrate the performance of the proposed approach by indicating that the posting destination of obscure reviews differs for each user.**

Keywords–Tourism information; Text classification; Support Vector Machine;Review Analysis.

I. Introduction

In recent years, it has become commonplace for many people to give their opinions and impressions regarding several types of venues, such as tourist spots, hotels, and restaurants, on review websites such as Yelp [1], Expedia [2], and TripAdvisor [3]. In this paper, we call such spots venues. Reviews written about venues describe information regarding the venues themselves and the impressions and behaviors of the users. Such reviews are useful for travel planning, obtaining information on travel destinations, tourist behavior, and visitor impressions of popular tourist spots. Therefore, some studies have extracted tourism information from user-provided reviews [4][5].

Some venues are obscure, secret, or not well-known. Despite not being popular, such venues may be well-regarded by visitors. Because some obscure venues can lead to improved tourist satisfaction and the acquisition of repeat visitors, some methods for describing obscure venues and recommending them to tourists have been proposed [6][7]. Definitions regarding obscure venues have been proposed in such studies. Studies on this subject commonly define an obscure venue as one in which the visibility for tourists is low but the value is high. For example, the authors in [6] defined obscure spots as less known, but still worth visiting, and extracted such spots. Also, [8] extracted hidden tourist spots with low popularity but a high level of satisfaction. However, precisely identifying obscure venues is difficult because the places that people feel are obscure depends on their own personality.

In this research, we identify obscure venues from review sites, and the proposed approach focuses on words in the text of the venue reviews. This study then extracts obscure reviews without directly giving a definition of obscure to accommodate the fact that the impression of a venue differs among different people. For this study, we regard a venue with many reviews written about the impression of its obscurity as an obscure venue (hereinafter referred to as "obscure review").

This study extracted such reviews from all reviews on a particular venue. In this paper, a review is defined as an obscure review if its text contains terms related to "obscure" (hereinafter referred to as "obscure words"). If the ratio of reviews of a venue that includes obscure words accounts for the majority, the venue is defined as obscure.

Although the aim of this research is the identification of obscure venues using user-provided reviews that include obscure words, in most cases the number of reviews on a venue is small, and customers might frequently visit there. Because an obscure venue might be less well-known by people even if worthwhile, there will be few reviews for such venues. In addition, few reviews obtain obscure words. As a result, the number of reviews to be classified as obscure is insufficient for identification of obscure venues. Moreover, it is unrealistic to define all expressions related to the word obscure. Therefore, to extract obscure reviews that do not include obscure words but rather the description of an obscure venue, this study applies the development of a classification model of the representation of contents of a review as obscure or not, regardless of whether a review contains an obscure word. Reviews that do not contain obscure words were classified using the model, and the classifier was evaluated using a dataset of reviews submitted by users.

Moreover, different reviewers have posted various reviews on different venues, and the criteria by which a venue is considered obscure differs according to the reviewer. Therefore, this research revealed that the reviewer who posts an obscure review for each venue is different. As a result, this study examined the efficiency of the proposed approach in identifying obscure venues using the obscure-word based classifier without a direct definition of the term obscure.

A summary of contributions from this study is as follows.

- We design a new approach for identifying obscure venues using user-provided venues.
- We propose a classifier for identifying obscure reviews without the word review or obscure words.
- We analyze the posting destination of obscure review differently for each user.

The remainder of this paper is organized as follows. Section

II presents previous studies related to this topic. Section III describes our proposed method for the development of a classifier for discovering obscure reviews by using obscure words and the identification of obscure venues. Section IV describes the experiments evaluating our proposed method using the Yelp dataset and an analysis of the hypothesis that an obscure venue is perceived differently for each user. Section V provides some concluding remarks along with a discussion of results and areas of future work.

II. Related Works

The main aim of our research was to find obscure venues for tourism analysis using user-provided reviews posted to social media sites. This section introduces the related studies published in the area of analysis of tourism information using reviews and extracting obscure venues.

A. Analysis for tourism using reviews

Research has been conducted on the extraction of tourism information through user-generated content on social media sites. In addition, extracting helpful or useful information from text data like reviews and blogs is one of the research tools used to analyze reviews. Our proposed research on extracting obscure venues from reviews is related to the analysis of reviews for recommendation and the analysis of tourism information.

[9] analyzed factors affecting the perceived usefulness of reviews to findings contributing to tourism marketers. [10] predicted where memorable is the travel destination using the user-generated photographs in blogs. [11] proposed a method for identifying dimensions of satisfaction using an unsupervised learning algorithm with numerical and textual information from user-generated online reviews, and analyzed the multiple factors contributing to consumer satisfaction. [12] predicted how helpful a review is and presented a list of ranked reviews based on an evaluation. [13] proposed a method for detecting reviews that reliably predict foodborne illnesses using review classification. [14] proposed a method for detecting the topic of phrases in helpful recommending reviews. [15] proposed a method for aspect-based opinion mining of tourism reviews to classify them into negative or positive aspects. [16] proposed an approach for sentiment classification of online hotel booking opinions using a dependency tree structure.

These studies analyzed user-provided reviews on social media sites for improving sightseeing satisfaction. This paper tackles the analysis of user perception of obscure venues based on reviews.

B. Extracting obscure venues from social media sites

Studies have been conducted on extracting obscure venues and tourist spots from social media sites. Because obscure spots are expected to spread tourists to other tourist spots and improve the satisfaction of the tourism experience, some studies extracting posts on such spots have been conducted.

[6] proposed a method for evaluating sightseeing spots that are less well-known but are worth visiting. [7] defined the term obscure to indicate spots that are not famous but have high evaluations, and extracted such spots based on name recognition and user evaluations. [8] proposed a method for providing tourism information of hidden spots for increasing tourism satisfaction. [17] extracted hot and cold spots based on a spatial analysis of user-generated content to extract knowledge of tourist behaviors.

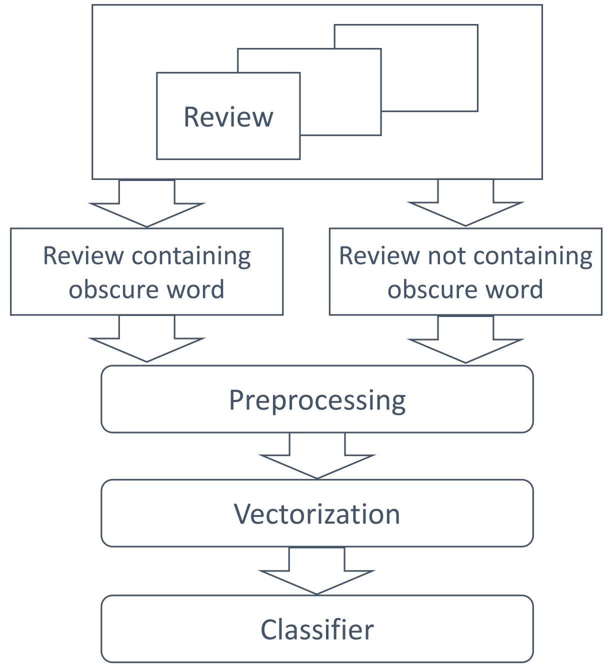

Figure 1. Overview of classifier for extracting obscure reviews using obscure words.

TABLE I. OBSCURE WORDS.

secret grate spot	secret grate place
kept secret place	kept secret spot
little known hot	spot secret spot
little known hot place	best kept secret
secret place	

This research used a classifier to extract obscure venues using reviews that include the word obscure to comprehensively deal with familiarity and user interest. The main characteristic of this research is the extraction of sightseeing spots recognized by reviewers as obscure venues.

III. Proposed Method

In this section, we describe our proposed method for discovering obscure venues based on user reviews.

This study extracted reviews including obscure words from the Yelp website and generated a classifier for both obscure and non-obscure reviews. We demonstrate an overview of our proposed classifier in Figure 1. First, we extract obscure and non-obscure reviews from the training dataset. Next, we apply preprocessing and a vectorization method. Finally, we create a model of the classifier using a vector to classify a review as obscure or not.

After this process, the classifier is applied to all reviews on a venue, and the venue is classified as obscure or non-obscure based on the reviews classified as obscure.

A. Obscure words

This section explains obscure words for extracting obscure reviews. In this research, obscure words are used to identify obscure venues from all reviews in a venue. This study defined nine obscure words, as shown in Table I. The criterion for selecting obscure words is to select an English phrase manually that seems to represent a word indicating obscurity, and not an expression that has no meaning other than obscurity. Because these words do not cover all words expressing user perceptions of obscurity, we conduct supervised learning using reviews including these words.

B. Preprocessing

This section describes the preprocessing applied to vectorize the reviews for machine learning. First, reviews written in English were extracted from all reviews. The texts from the extracted reviews were converted into lower-case texts. Next, we apply stop-word elimination and stemming to each word. This study defined 319 stop words, such as "the" and "and," which are commonly used in sentences.

C. Vectorization

Next, the preprocessed reviews were vectorized. First, Term Frequency (TF) and Inverse Document Frequency (IDF) were applied to the texts for determining what words in reviews might be more efficient for extracting obscure reviews. In this paper, we calculated the TFIDF of each word t in review r. The term frequency $tf(t,d)$ and inverse document frequency $idf(t,D)$ are calculated using the follow equations:

$$tf(t,r) = \frac{f_{t,r}}{\sum_{t \in r} f_{t,r}} \tag{1}$$

$$idf(t,R) = \log \frac{|R|}{|\{r \in R : t \in r\}|} \tag{2}$$

where the number of reviews is $|R|$, and $f_{t,r}$ is the number of occurrences of word t in review r.

Then, the TFIDF of each word t in review r in reviews R is calculated through the following equation:

$$tfidf(t,r,R) = tf(t,r) \times idf(t,r) \tag{3}$$

Next, to decrease the number of dimensions, a Principal Component Analysis (PCA) was conducted [18]. This process resulted in a feature vector of each review.

D. Classification of obscure reviews

In this section, we describe the procedure for generating a classification model of reviews regardless of whether they are obscure reviews. Our method proposed in this study identifies obscure venues using obscure reviews even if the review does not include obscure words. Therefore, our proposed method creates a classifier for identifying such reviews that do not include obscure words but when their content represents an obscure venue.

A method is proposed to classify the reviews into obscure or non-obscure reviews. In this research, we apply a binary classification method using vectors generated as described in Section III-C. The first class is thus obscure reviews, which consists of reviews that contain an obscure word. The other class is non-obscure reviews, which consists of reviews that do not contain an obscure word. This study used a binary classification Support Vector Machine (SVM) [19] to classify reviews as obscure or not obscure.

E. Identification of obscure venue

Herein, we describe how to find obscure venues using a classifier. Figure 2 shows an overview of the procedure for identification of an obscure venue. We collect all review texts of a venue and apply the classifier described in Section III-D to the reviews. Finally, we count the reviews classified as obscure or non-obscure reviews of a venue. As a result, this study regards an obscure venue as one in which the percentage of obscure venues is greater than the threshold. In this paper, when the ratio of reviews classified as obscure among all reviews on a venue is larger than half, the venue is considered obscure, otherwise it is non-obscure.

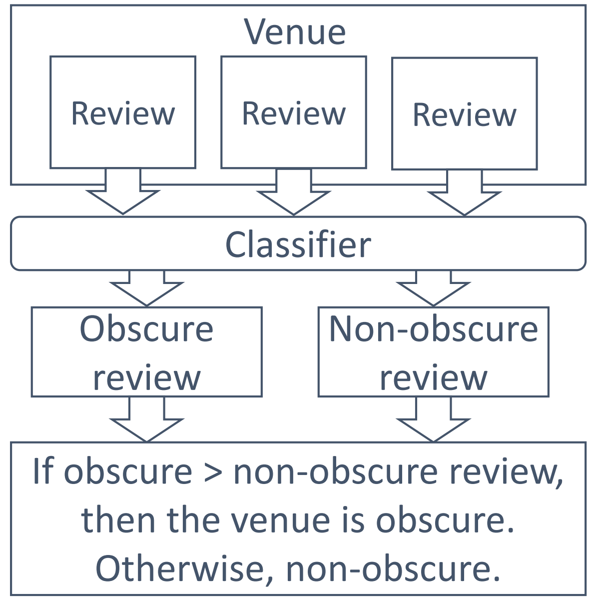

Figure 2. Overview of procedure for identification of obscure venues using obscure and non-obscure reviews.

IV. Experiments

In this paper, we evaluate the performance of our proposed method through an evaluation experiment based on classification. First, we describe the experimental conditions of the dataset and the evaluation criteria. Next, we describe our experiments conducted for an evaluation of obscure review discovery. Finally, we evaluate and discuss the differences in which each reviewer evaluates a venue as obscure or not. In addition, we used the Python software scikit-learn [20] for implementation of the SVM, PCA, TFIDF, and evaluation criteria in the following experiments.

A. Dataset

Herein, we describe the dataset used for this experiment, namely, the Yelp Dataset Challenge (round 9) [21], which includes 144,072 venues and 4,153,150 reviews. This study comprises 1,978 reviews that mention an obscure word at least once.

B. Experimental conditions

This section describes the procedure used for the creation of classifiers for obscure reviews. The training data for the SVM includes 140 reviews that present an obscure word and are proven to be about an obscure venue, and 1,000 reviews that do not include an obscure word.

This experiment used a Gaussian kernel for the SVM kernel function. In addition, the hyperparameters of the SVM were searched through a grid search with five cross-validations, using parameters with the highest F-values measured through this experiment. The number of dimensions found through the PCA was 100.

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-697-2

TABLE II. CLASSIFICATION RESULTS OF OBSCURE REVIEWS.

	Precision	Recall	F-value	Accuracy
Obscure review	0.92	0.73	0.81	0.98
Non-obscure review	0.96	0.99	0.98	
Average	0.95	0.96	0.95	

TABLE III. TOP-10 OF VENUE WITH A HIGH PERCENTAGE OF OBSCURE REVIEWS.

Venue	Obscure reviews	All reviews	Percentage
Fashion 1	4	5	0.80
Restaurants 1	4	5	0.80
Fitness & Instruction1	4	5	0.80
Health & Medical 1	4	5	0.80
Shopping 1	4	5	0.80
Restaurants 2	4	5	0.80
Home Services 1	3	4	0.75
Shopping 2	3	4	0.75
Beauty & Spas 1	3	4	0.75
Restaurants 3	3	4	0.75

In this paper, four evaluation criteria were used for the classification performance: accuracy, precision, recall, and F-value.

C. Classification result of obscure reviews

In this section, we describe and discuss the evaluation results of classifying reviews into obscure or non-obscure reviews. Table II shows the evaluation results of the classification of obscure reviews through the procedure described above. In Table II, “Obscure review” shows the reviews that include an obscure word, whereas “Non-obscure review” shows reviews that do not include an obscure word. Comparing the results shown in Table II for obscure and non-obscure reviews, the evaluation scores of the non-obscure reviews are lower than those of the obscure reviews. In particular, there is a vast difference between both scores regarding the recall rate. The evaluation score is achieved because reviews with an obscure word are misclassified as non-obscure in certain cases because the number of reviews in the training dataset is unbalanced. However, the purpose of this research is to identify obscure venues using extracted obscure reviews. As shown in Table II, the precision of the obscure reviews was 0.92, which shows that it is rare for a classifier to misclassify the content of reviews unrelated to obscurity. With the following, we worked on finding obscure venues through this classifier.

D. Classification results of obscure venue

This section describes and discusses the evaluation results of discovering an obscure venue using a classifier. In this experiment, we apply the classifier to all reviews of a venue and calculate the percentage of reviews classified as obscure.

Table III shows the results of the top-10 venues with a high percentage of reviews classified as obscure. The terms “Obscure reviews” and “All reviews” present the number of obscure reviews and all reviews of a venue. In addition, the name of the venue is anonymous, and is represented by the category name in Yelp and a serial number.

In Table III, we confirm the reviews posted on each venue manually. As a result, those reviews include many phrases of ”I knew for the first time,” ”It was hard to access, but the service was good,” and so. These phrases seem to be related to obscurity. Therefore, we believe that our method discovers venues that people have evaluated as obscure.

TABLE IV. PERCENTAGE OF DIFFERENCES IN REVIEWERS FEELING A VENUE AS BEING OBSCURE.

Pattern ①	50
Pattern ②	883
① / (① + ②)	0.053

E. Analysis of obscurity in each category

In this section, we analyze the obscure venues in each category. We denote the venue where the percentage of obscure reviews is 50% or more, according to the description in Section III-E, and find the proportion of venues classified as obscure within the same category.

We calculate the proportion of venues classified as obscure within a category. Here, we used 27 categories whose number of reviews in a category is 1,000 or more. We show the percentage of obscure venues in each category, as indicated in Figure 3. In this figure, the vertical axis shows the proportion of venues classified as obscure within the same category, and the horizontal axis shows the category names in Yelp. The highest percentage of obscure venues is for “Local Services” at approximately 14%. Subcategories of this category include junk removal & hauling, bike repair / maintenance, and mobile phone repair. In addition, according to Figure 3, the top categories with a high percentage of obscure venues contain many categories used in daily life. In contrast, the categories ”restaurants” and ”nightlife” where many people go to popular venues ranked the lowest. In these categories, popular venues are sometimes a type of sightseeing spot. In addition, it seems that a large number of shops related to food services (such as Mexican restaurants and bars) affects the percentage of obscure venues.

F. Differences between venues evaluated as obscure for each reviewer

This section analyzes the differences among venues considered by reviewers as obscure.

Herein, we show the difficulty of providing a unique definition for obscure venues using our proposed method for obscure venue extraction. Using the classifier described in Section III-D, we classify whether a user review on a venue is obscure or not. Then, if the types of reviews on the venue are different, the venue that the user feels is obscure is different.

This research focused on cases in which two different reviewers posted similar reviews on two venue pairs. Two patterns of venues whose reviews refer to obscurity were considered, as shown in Figure 4. Pattern ① is a case in which two reviewers posted an obscure review and a non-obscure review to different venues. This pattern represents a case in which the reviewer felt that the referred venue was different. Pattern ② is a case in which the reviews posted by two different reviewers are the same for the referred venues. This pattern is one in which the venues the reviewers felt as obscure are the same. Therefore, if there is a certain number of reviews considered as pattern ①, it can be said that the venue perceived as obscure is different for each reviewer; the classification of obscure reviews reveals the contribution of the identification of obscure venues.

The procedure of this experiment is as follows. First, obscure venues to which two users posted similar reviews were extracted. During this experiment, 1,278 obscure venues that had obscure reviews were extracted, comprising more than

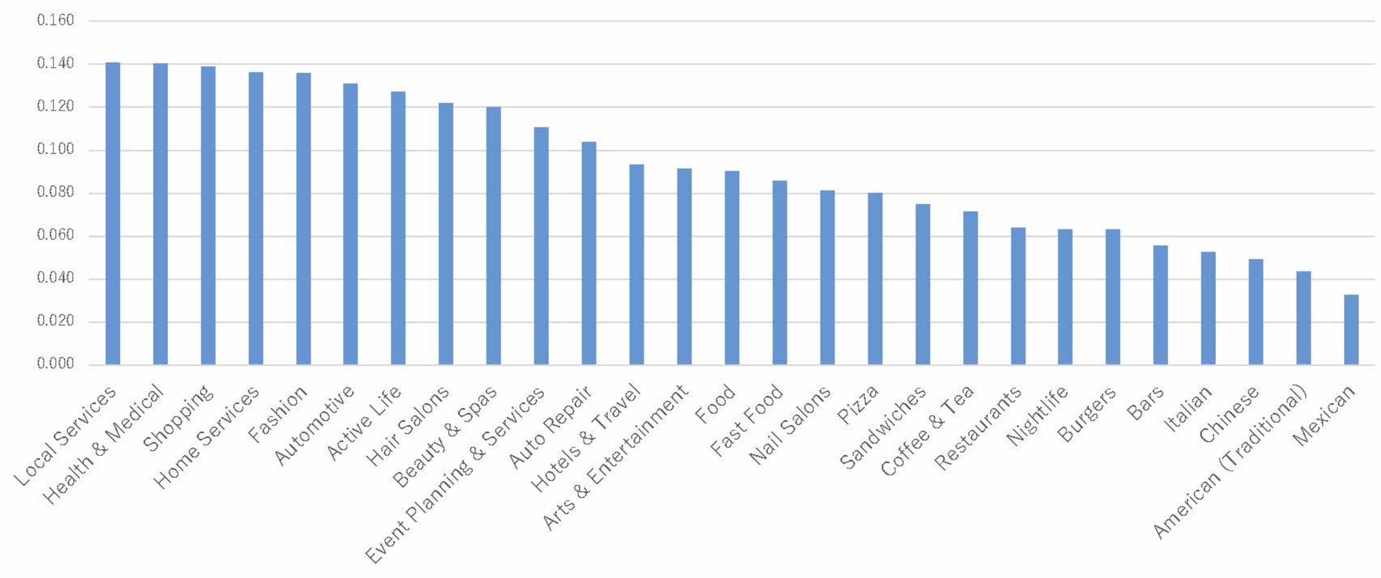

Figure 3. The percentage of obscure venues in each category

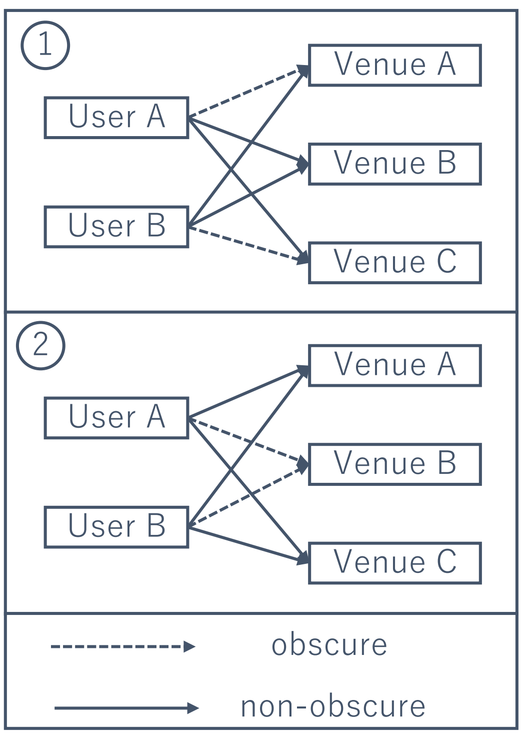

Figure 4. Pattern in which two reviewers evaluate venues as obscure.

50% of all reviews; there were 696 reviewers. The classifier was then applied to the written reviews as described in Section 4.2. The numbers of the two patterns were calculated based on the classification results.

Table IV shows the experimental results. From Table IV, pattern 1 comprised approximately 5.3% of the total. In other words, the combination of 5.3% of reviewers differs from the venue that was perceived as obscure. This result shows that the venues perceived as an obscure venue are not necessarily the same for all reviewers. Therefore, the approach of abstractly treating as obscure a review that includes an obscure word without criteria on the obscure venue used to extract the venue has the potential to be effective.

V. Conclusion

In this research, we proposed a method for identifying obscure venues by extracting reviews that include descriptions regarding obscure posts on Yelp. Through reviews that include obscure words, a classifier was created to differentiate the reviews describing obscurity from those that do not, based on reviews in which the reviewers recognize the venues as being obscure. Experimental results showed that the classifier is useful for extracting obscure reviews. Furthermore, this study formulated and verified the hypothesis that venues perceived as obscure by reviewers are different. As a result, the venues perceived as being obscure are not necessarily the same for all reviewers.

Future studies will include a more detailed experiment and analyze the number of obscure venues and the various categories present in each city. This paper is limited to analyzing obscure venues extracted using our proposed method in a qualitative manner. For a discovered venue, it is necessary to analyze whether it is obscure or not and to evaluate how useful the information is. For this purpose, we will conduct questionnaires by evaluators on the obscure venues by our proposed method. Further studies may apply our classifier to other cities to discover unique, obscure venues.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 16K00157 and 16K16158, and Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas ”Research on social big data.”

References

[1] “Yelp,” URL: https://www.yelp.com/ [accessed: 2019-02-27].

[2] “Expedia,” URL: https://www.expedia.com/ [accessed: 2019-02-27].

[3] "Tripadvisor," URL: https://www.tripadvisor.com/ [accessed: 2019-02-27].

[4] D. Ukpabi, S. Olaleye, E. Mogaji, and H. Karjaluoto, "Insights into online reviews of hotel service attributes: A cross-national study of selected countries in africa," in Information and Communication Technologies in Tourism 2018, B. Stangl and J. Pesonen, Eds. Cham: Springer International Publishing, 2018, pp. 243–256.

[5] V. Browning, K. K. F. So, and B. Sparks, "The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels," Journal of Travel & Tourism Marketing, vol. 30, no. 1-2, 2013, pp. 23–40.

[6] C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa, "Anaba: An obscure sightseeing spots discovering system," in 2014 IEEE International Conference on Multimedia and Expo, vol. 00, 2014, pp. 1–6.

[7] D. Kitayama, "Extraction method for anaba spots based on name recognition and user's evaluation," in Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, ser. iiWAS '16. ACM, 2016, pp. 12–15.

[8] S. Katayama, M. Obuchi, T. Okoshi, and J. Nakazawa, "Providing information of hidden spot for tourists to increase tourism satisfaction," in Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ser. UbiComp '18. ACM, 2018, pp. 377–380.

[9] Z. Liu and S. Park, "What makes a useful online review? implication for travel product websites," Tourism Management, vol. 47, 2015, pp. 140 – 151.

[10] M. Toyoshima, M. Hirota, D. Kato, T. Araki, and H. Ishikawa, "Where is the memorable travel destinations?" in Social Informatics. Cham: Springer International Publishing, 2018, pp. 291–298.

[11] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," Tourism Management, vol. 59, 2017, pp. 467 – 483.

[12] C. Vo, D. Duong, D. Nguyen, and T. Cao, "From helpfulness prediction to helpful review retrieval for online product reviews," in Proceedings of the Ninth International Symposium on Information and Communication Technology, ser. SoICT 2018. ACM, 2018, pp. 38–45.

[13] Z. Wang, B. S. Balasubramani, and I. F. Cruz, "Predictive analytics using text classification for restaurant inspections," in Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, ser. UrbanGIS'17. ACM, 2017, pp. 14:1–14:4.

[14] R. Dong, M. Schaal, M. P. O'Mahony, and B. Smyth, "Topic extraction from online reviews for classification and recommendation," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI '13. AAAI Press, 2013, pp. 1310–1316. [Online]. Available: http://dl.acm.org/citation.cfm?id=2540128.2540317

[15] M. Afzaal, M. Usman, A. C. M. Fong, S. Fong, and Y. Zhuang, "Fuzzy aspect based opinion classification system for mining tourist reviews," Adv. Fuzzy Sys., vol. 2016, Oct. 2016, pp. 2–.

[16] T. S. Bang and V. Sornlertlamvanich, "Sentiment classification for hotel booking review based on sentence dependency structure and sub-opinion analysis," IEICE Transactions on Information and Systems, vol. E101.D, no. 4, 2018, pp. 909–916.

[17] E. van der Zee, D. Bertocchi, and D. Vanneste, "Distribution of tourists within urban heritage destinations: a hot spot/cold spot analysis of tripadvisor data as support for destination management," Current Issues in Tourism, vol. 0, no. 0, 2018, pp. 1–22.

[18] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1, 1987, pp. 37 – 52, proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[19] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, 1995, pp. 273–297.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[21] "Yelp dataset challenge (round 9)," URL: https://www.yelp.com/dataset/challenge [accessed: 2019-02-27].