Time-Variable Analysis of Accommodation Reviews Based on a Hierarchical Topic Model

Yujiro Sato*, Ryosuke Yamanishi†, Yoko Nishihara†
*Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan
†College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan
Email: {is0309he@ed, ryama@media, nisihara@fc}.ritsumei.ac.jp

***Abstract*—Accommodation reviews are valuable resources for future guests to know the opinions of users who have already stayed at a particular place before. However, it is difficult for users to extract the information specific to each topic such as facilities, access, and breakfast. We consider that the seasonal features of accommodations are especially important to ensure a comfortable and enjoyable stay. This paper proposes a hierarchical topic analysis with time variation to extract seasonal features for accommodations. The proposed method extracts seasonally important words and shows the similarity of topics that the important words belong to between seasons. In this paper, we discuss the effectiveness of the extracted features as references for guests to choose accommodations.**

Keywords–review analysis; consumer decision support; hLDA.

I. INTRODUCTION

Since the mainstream of accommodation reservations is online, travelers need to decide on accommodation based on the information on the Web. Consumers-stated preferences for decision criteria are various [1]. In particular, the amount of reviews has been found to promote accommodation room occupancy [2]. However, it is hard to check a large number of reviews. We thus consider that the value of reviews would be increased by helping users to easily check the reviews.

According to Dickinger *et al.*, recommendations from friends and online accommodation reviews should be the most important factors that influence online hotel booking [3]. Online accommodation reviews have been widely studied [4], and such research enables us to use the analysis results by text mining to help travelers with their decision making. According to Vermeulen, negative as well as positive reviews increase the consumer awareness of the accommodation [5]. Also, they showed that positive reviews can improve consumer attitudes toward the accommodations.

This paper focuses on seasonal features in accommodation reviews. We believe that the value of each accommodation also depends on seasonal events held in the neighborhood. The presentation of seasonal features might become one of the determinants of accommodations. The tf-idf method is a well-known representative method providing word importance in documents and it is used in several applications such as documents classification [6]. The first step of our proposed method is extracting higher tf-idf words from monthly reviews of an accommodation. The second step is forming hierarchical topics that contain higher tf-idf words. Finally, comparing the analysis results of each month showed the seasonal features of the accommodation. In order to provide accurate information to the consumers, the category should be taken into consideration [7]. Features change according to the season for each category, and it influences the users' decision making. In the proposed method, a topic model (i.e., latent semantic analysis) is used for category acquisition. One of the topic models is an Unsupervised Learning method: Latent Dirichlet Allocation (LDA) [8]. Our purpose is to extract topics from the documents based on the assumption that a document has multiple topics. Han *et al.* analyzed hotel reviews using LDA [9]. This research has succeeded in extracting the relationships between emotions and evaluations by topic analysis of accommodation reviews. Another LDA extension method is hLDA (hierarchical Latent Dirichlet Allocation) [10][11]. The hLDA probabilistically estimates topics, assuming that the topics contained in the document have a hierarchical structure. Regarding the feature of accommodation reviews, Wang *et al.* defined a new problem in opinionated text data analysis called Latent Aspect Rating Analysis (LARA) [12]. This study focuses on the inclusion relation in the category of accommodation reviews. For example, the category for "meal" includes more detailed information such as "meal price" and "meal quality."

In this paper, we hypothesize that it is possible to extract the inclusion relations of topics in accommodation reviews by using hLDA. We incorporate the time change analysis of the accommodation reviews using the results of hLDA and important words extracted by using the tf-idf method. The problem tackled in this paper is when consumers are not able to know the seasonal features from information on the Web. The decision making of the consumers should become easier as this problem is resolved. In Section 1, we have introduced the background and the relevant studies. In Section 2, we will introduce the data to be used, and in Section 3, we will propose an analysis method. In Section 4, we will discuss the evaluation method, and in Section 5, we will discuss the results. Finally, we will discuss the prospects of this research.

II. DATA

In this study, we use 5,082,427 Rakuten Travel reviews (the data was retrieved on March 29, 2020 [13]). The data was collected from 29,400 accommodation reviews for the time period 1996 through 2016. The top 10% of accommodations with the highest number of reviews were used for our analysis.

The importance of preprocessing in natural language processing has been widely known [14]. With preprocessing, the accuracy of the analysis would be improved by narrowing down the parts-of-speech to be analyzed as the stop-words. In this paper, we analyze only nouns to extract the characteristics of accommodations. We exclude the following words from the analysis:

1) Nouns whose meaning can be not understood.

 ISBN: 978-1-61208-761-0

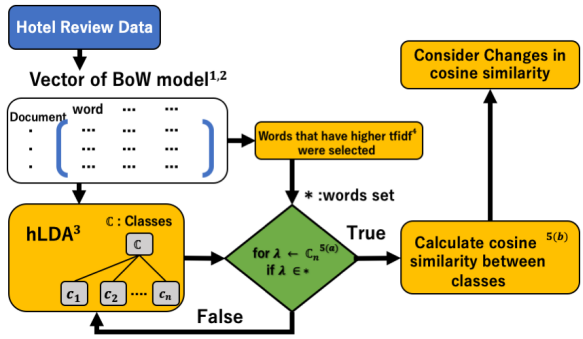

Figure 1. The framework of analysis procedure. In the figure, the index of procedure is shown as a superscript which is detailed in section III.

2) Nouns whose frequency is lower than three.
3) Days and symbols.

According to these procedures, we removed many noise nouns from the analysis: 8,817 of 10,979 nouns and 7,825 out of 9,674 nouns were each removed from the reviews of the accommodation #1 and #2, respectively.

III. The Proposed Method

In the proposed method, hLDA, is usedfor a latent topic analysis, and the tf-idf method is used to extract seasonal features. Using both methods, we extract the latent topics depending on season in the reviews in a hierarchical structure. Extraction and consideration are performed according to the following procedure (see the subscript number in Figure 1).

1) The morphological analyzer extracts the nouns that appear in the document. For the Japanese documents, we use MeCab and NEologd as the morphological analyzer and the dictionary, respectively.
2) The extracted nouns are vectorized based on the Bag-of-Words model.
3) Using the vectors, the latent topics are hierarchically clustered.
4) The reviews are divided for each month in the calendar and 12 documents are generated. The tf-idf value is given to each noun that appears in each document.
5) The following processing (a) and (b) are executed for the top 10% nouns with tf-idf values excluding stop words.
   a) The tendency of nouns extracted in the same cluster is analyzed.
   b) Focusing on the clusters containing the arbitrary noun among plural months, the similarity among the clusters is calculated.

A. Hierarchical topic model

Accommodation reviews have two category types: large and small. Assuming that the structure of the categories can be extracted as topics, the hierarchical relationship of topics is constructed. Therefore, we focused on hLDA, which is an extension model of the LDA. The hLDA analyzes the hierarchical relationship of topics.

1) The nested Chinese Restaurant Process: The nested Chinese Restaurant Process (nCRP) is a stochastic process on a tree structure. This stochastic process is used for hLDA and is represented by the following metaphor using the Chinese Restaurant Process (CRP). The CRP is a distribution obtained by imagining a process by which N customers sit down in a Chinese restaurant with an infinite number of tables [11]. Let the customers be labeled as $1, 2, \ldots, N$ in the order they entered the restaurant. The first customer sits at the first table. The nth customer sits. The probability of sitting at the ith-table is determined by the following distribution (1);

$$p(c_n = i|c_{n-1}) = \begin{cases} \frac{n_i}{\gamma+n-1} \text{ (occupied table } i \text{)}, \\ \frac{\gamma}{\gamma+n-1} \text{ (next unoccupied table)}, \end{cases} \quad (1)$$

where, n_i is the number of customers currently sitting at ith-table, and γ is a meta-parameter that controls how often a customer chooses a new table versus sitting with others, which is relative to the number of customers in the restaurant.

In nCRP, a tree structure is formed based on CRP. nCRP is explained with the following metaphor. Suppose there are an infinite number of restaurants in the city, and each restaurant has an infinite number of tables. There is the restaurant at the root of the hierarchy, and each table at the restaurant specifies other restaurants. The first customer enters the root restaurant and selects the table according to the CRP. Then, the route to the next restaurant would be provided to the customer. The customer selects the table again according to the CRP. This procedure is infinitely repeated, and the path in the tree structure is constructed. All customers select a table, and a subtree consists of an infinitely deep tree branched infinitely. In this study, customers, restaurants and table seats each represents words, hierarchies, and topics.

2) Hierarchical Latent Dirichlet Allocation: In the generation process of hLDA, the tree structure is generated by nCRP. The hLDA is conducted according to the previous study [10], [11]. In the implementation of hLDA, it is necessary to set meta-parameters (α, γ, η, number of layers) in advance; affects" should be "affect " appropriateness " of the extraction results. We set $\gamma = 1.0$ and $\eta = 1.0$ referring to the previous study [10] for appropriate extraction for accommodation reviews. The main goal of this paper is to extract the feature of accommodation facilities, so it is desirable to know what is the specific criteria for the topic classification. Therefore, we set the number of layers in a hierarchy to three and the number of sampler iterations to 500. Only converged nouns are used in the analysis.

B. tf-idf

Essentially, tf-idf works by determining the relative frequency of words in an arbitrary document compared to the inverse proportion of the word over the entire document corpus [15]. In this study, since monthly reviews are used as a document set, nouns with high tf-idf values are assumed to be feature nouns representing seasons which rarely appear in other months.

C. Similarity among clusters

In this paper, we extract the nouns featuring the seasons by evaluating the change of the similarity among the clusters for each month constructed by using hLDA. The cosine similarity between the clusters for each month obtained by using hLDA

 ISBN: 978-1-61208-761-0

TABLE I. TOP FIVE NOUNS WITH HIGHER TF-IDF FROM ACCOMMODATION REVIEWS [TRANSLATED INTO ENGLISH BY THE AUTHORS]

	Accommodation #1	Accommodation #2
January	new year, new year's day, new year's end, mochi pounding, superlative degree	new year, new year's, sweets, special, anniversary
May	Golden Week, red snapper, spring, love, holiday	room temperature, Golden Week, weekend, grade, beauty treatment
August	Obon, Gassho, Rokusaburo Michiba, sweetfish, Noryo	pool, summer vacation, beach, sea bathing, barbecue
December	christmas, meal, breakthrough, specialty, superlative degree	Luminarie, christmas, winter, hospitality, special

is calculated. A cluster is formed by a collection of multiple nouns. Let each cluster for the months $m1$ and $m2$ be C_m and C_n. C_m and C_n are represented as (2) and (3), respectively;

$$C_m(1,\ldots,l) = \{W_1, W_2, \ldots, W_l\}, \quad (2)$$
$$C_n(1,\ldots,l) = \{W_1, W_1, \ldots, W_l\}. \quad (3)$$

The cosine similarity between these two sets is calculated by using (4);

$$cos(C_m, C_n) = \frac{\sum_{k=1}^{l} C_m(k) \cdot C_n(k)}{\sqrt{\sum_{k=1}^{l}(C_m(k))^2} \cdot \sqrt{\sum_{k=1}^{l}(C_n(k))^2}}. \quad (4)$$

IV. EXPERIMENT

The target months of analysis were narrowed down to the busy season of accommodations: January, May, August, and December. Table I shows the top five nouns in the two reviews with the higher tf-idf: note, the words are translated from Japanese into English by the authors. Parts of the hLDA analysis results are shown in Figure 2 and Figure 3; note that nouns that did not converge are excluded.

From Table I, other than the nouns that indicate the season itself, items suitable for evaluation such as "meals" and "events" are selected. In this paper, we analyzed the two accommodations randomly selected from the dataset described in Section II as examples for the kick-off of our research project.

We define the class to analyze as follows:

- Small cluster: a single cluster in third layer of hierarchy containing the nouns, e.g., C_{a1} and C_{a2} in both Figure 2 and Figure 3.
- Large cluster: multiple clusters with the same parent as the second layer of hierarchy cluster containing the noun, e.g., all small clusters of C_a an C_b in both Figure 2 and Figure 3.

The next step is to analyze the transition of similarity among months for each category of clusters. This method compares temporal changes in topics on large and small scales. Table II and Table III show the results.

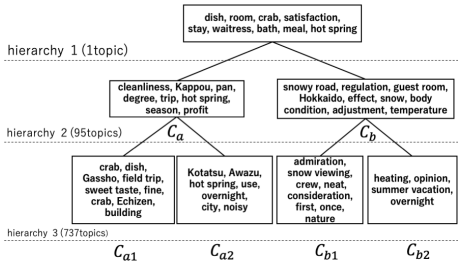

Figure 2. The result of hLDA for accommodation #1 in January.

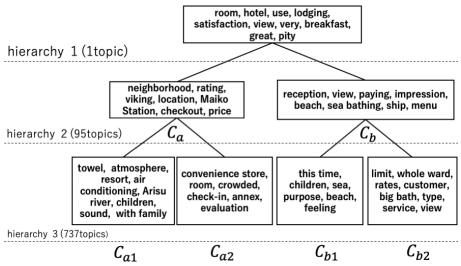

Figure 3. The result of hLDA for accommodation #2 in August.

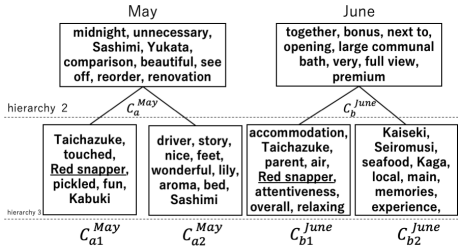

Figure 4. The Hierarchical structure for accommodation #1 in May and June including "red snapper."

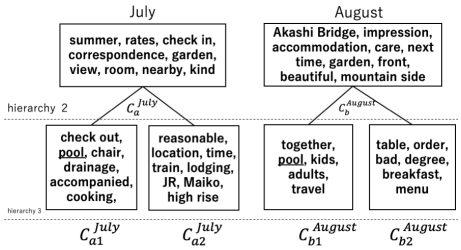

Figure 5. The Hierarchical structure for accommodation #2 in July and August including "pool."

V. DISCUSSION

A. hLDA extraction results

From "cleanliness" and "hot spring" in Figure 2, it can be seen that services and facilities were evaluated as the topics. In contrast, topics related to weather and environment such as "snow" and "temperature" were extracted from C_b. In C_{a1}, evaluations on meals, such as "crab" and "sweet" were extracted. As comparing C_{a2} with C_{b2}, it is clear that food-related topics were classified. C_{b1} had topics for services and environment, while C_{b2} had topics for temperature.

From C_a in Figure 3, it is shown that price and location were evaluated as topics. From "sea bathing" and "view" in C_b, it can be seen that outdoor and facility features were evaluated

TABLE II. THE COSINE SIMILARITY BETWEEN CLUSTERS INCLUDING "RED SNAPPER" WHICH IS EXTRACTED BY TF-IDF METHOD, IN THE REVIEW OF ACCOMMODATION #1 IN MAY.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Large cluster	0.0	0.0	0.0	0.029	1.0	0.046	0.018	0.039	0.021	0.028	0.0	0.0
Small cluster	0.0	0.0	0.0	0.099	1.0	0.199	0.099	0.099	0.099	0.105	0.0	0.0

TABLE III. THE COSINE SIMILARITY BETWEEN CLUSTERS INCLUDING "POOL" WHICH IS EXTRACTED BY TF-IDF METHOD, IN THE REVIEW OF ACCOMMODATION #2 IN AUGUST.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Large cluster	0.0	0.0	0.0	0.0	0.0	0.110	0.042	1.0	0.164	0.0	0.0	0.0
Small cluster	0.0	0.0	0.0	0.0	0.0	0.099	0.099	1.0	0.099	0.0	0.0	0.0

as the topics. "With family" and "beach" were classified into C_{a1} and C_{b1}, it is imagined that we can enjoy a family trip and swimming in the sea; we confirmed that some of the reviews described such experiences. As a result, it can be expected that we should be able to enjoy seasonal foods especially crab and hot pot and snowy weather in accommodation #1 in January. Also, it is expected that we should be able to enjoy swimming with our family in accommodation #2 in August.

B. Evaluation by Cosine similarity

We focus on "Red snapper", which is the specific noun in May in the review of accommodation #1. The similarity between classes including "Red snapper" beyond months is calculated. Table II shows the transition of the similarities between May and each month, respectively. From Table II, it was found that June showed the highest similarity to May, followed by August and April. Though each similarity was low in real values, the similarity can be used in the discussion for relative evaluation. Because it includes various topics in a large cluster, more conceptual changes can be seen. Since the similarity between May and April and the one between May and June are relatively high, it seems that seasonal topics are similar to each other on that combination of months. In the hierarchies of May and June that contain "Red snapper" in Figure 4, it can be seen that "red snapper dishes" were included in both the third layers. In fact, it is known that the season for "Red snapper" should be March through June and September through November. However, we can have the fish during all seasons if we do not mind the freshness, it becomes clear that it is a topic that attracted attention in May and June for this accommodation.

For accommodation #2, we focus on "pool" in August for a discussion. Table II shows the similarity between classes including "pool" beyond months. From the results, it can be seen that "pool" appeared from June to September.

In large clusters, the classes including "pool" between August and September showed the highest similarity. However, the similarity between July and August was relatively low, though "pool" generally shows popularity in the season. From Figure 5, it can be seen that family trips like "adults" and "kids" were in the same cluster as "pool" in August. In the actual reviews, family travel styles appeared more frequently in August than July. The season for family trips was suggested from the results. From the result of analysis, we found that the user group using the pool changed according to the season. The analysis revealed the seasons that attract attention and the seasons for family trips.

C. Weakness of the System

In this analysis, the review data was divided into months. Therefore, it can be said that this analysis method is weak for features that straddle the months and features for each one day. In addition, this analysis method does not completely extract features for irregularly held events and review sentences for questions. The background of this method is that it is assumed that it will help consumers to think about which month they will travel when making travel plans. Also, this method was adopted in consideration of the difference between the date of staying and the date of writing the review.

VI. Conclusion and Future Work

In this paper, we have shown the analysis and discussion of using hLDA to extract seasonal features from accommodation reviews. As a result, we were able to extract the seasonal characteristics of accommodation facilities in a hierarchical structure. However, we need to consider a more practical use. We will go to the big goal of "consumer decision support" as the next step of our research. The feature directions can be as follows: (1) hLDA hyperparameters for accuracy, (2) visualization of the results, (3) regional differences with time variation, and (4) find useful information for accommodation. The task (4) is considered to improve the services of accommodation facilities and dynamic pricing.

Acknowledgment

We show our appreciation to Rakuten Dataset provided by Rakuten with a support of National Institute of Informatics Research Data.

References

[1] K. Kim, O.-J. Park, S. Yun, and H. Yun, "What makes tourists feel negatively about tourism destinations? application of hybrid text mining methodology to smart destination management," Technological Forecasting and Social Change, vol. 123, 2017, pp. 362–369.

[2] P. De Pelsmacker, S. Van Tilburg, and C. Holthof, "Digital marketing strategies, online reviews and hotel performance," International Journal of Hospitality Management, vol. 72, 2018, pp. 47–55.

[3] A. Dickinger and J. Mazanec, "Consumers ' preferred criteria for hotel online booking," Information and communication technologies in tourism 2008, 2008, pp. 244–254.

[4] R. Filieri, "What makes an online consumer review trustworthy?" Annals of Tourism Research, vol. 58, 2016, pp. 46–64.

[5] I. E. Vermeulen and D. Seegers, "Tried and tested: The impact of online hotel reviews on consumer consideration," Tourism management, vol. 30, no. 1, 2009, pp. 123–127.

 ISBN: 978-1-61208-761-0

[6] B. Trstenjak, S. Mikac, and D. Donko, “Knn with tf-idf based framework for text categorization,” Procedia Engineering, vol. 69, 2014, pp. 1356–1364.

[7] K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, “Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews,” Journal of Hospitality Marketing & Management, vol. 25, no. 1, 2016, pp. 1–24.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, 2003, pp. 993–1022.

[9] H. J. Han, S. Mankad, N. Gavirneni, and R. Verma, “What guests really think of your hotel: Text analytics of online customer reviews,” 2016.

[10] D. M. Blei, T. L. Griffiths, and M. I. Jordan, “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies,” Journal of the ACM (JACM), vol. 57, no. 2, 2010, pp. 1–30.

[11] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, “Hierarchical topic models and the nested chinese restaurant process,” in Advances in neural information processing systems, 2004, pp. 17–24.

[12] H. Wang, Y. Lu, and C. Zhai, “Latent aspect rating analysis on review text data: a rating regression approach,” in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 783–792.

[13] “Rakuten dataset,” URL: https://www.nii.ac.jp/dsc/idr/en/rakuten/ [accessed: March 2020].

[14] S. Vijayarani, M. J. Ilamathi, and M. Nithya, “Preprocessing techniques for text mining-an overview,” International Journal of Computer Science & Communication Networks, vol. 5, no. 1, 2015, pp. 7–16.

[15] J. Ramos, “Using tf-idf to determine word relevance in document queries,” in Proceedings of the first instructional conference on machine learning, vol. 242. Piscataway, NJ, 2003, pp. 133–142.

 ISBN: 978-1-61208-761-0