Feature Representation Extraction Method of Hotel Reviews using Co-occurrence Restriction and Dependency Graph

Koji Tanaka
Hitachi Government & Public Sector Systems, Ltd
Tokyo, Japan
koji.tanaka.st@hitachi.com

Takashi Ikoma
University of Tsukuba
Tokyo, Japan
ikoma.takashi.xj@alumni.tsukuba.ac.jp

Kazuhiko Tsuda
University of Tsukuba
Tokyo, Japan
tsuda@gssm.otsuka.tsukuba.ac.jp

Koichi Tsujii
Nippon Travel Agency Co. Ltd.
Tokyo, Japan
tsujii@gmail.com

Akiyuki Sekiguchi
Meiji University
Tokyo, Japan
sekichan2008@gmail.com

***Abstract*—Hotel reviews posted on accommodation reservation websites are thought to be valuable information for selecting hotel accommodations and also expected to be used for marketing. Since hotel reviews are various in their expressions, it was necessary to make a thesaurus to obtain useful feature representations. Preparing a thesaurus, however, has problems such that it is laborious and requires occasional revisions. In addition, it is necessary to determine subjects of evaluation in advance and set up synonyms for them. Thus, the analysis of subjects not under consideration becomes difficult. In the present study, we first graphed impression comments using co-occurrence restrictions and dependency structures and then extracted feature representations by clustering the graphs. This enabled us to extract feature representations on evaluation from the impression comments in hotel reviews without setting up subjects of evaluation in advance and a thesaurus.**

Keywords—Japanese text analysis; Thesauruses; Graph; Hotel reviews

I. INTRODUCTION

The advent of accommodation reservation websites has made it possible for people searching for hotel accommodations to make reservations anytime for 24 hours, and the number of their users has increased more than 10 times in the past 10 years [1]. Accommodation seekers collect necessary information using various information offered by accommodation reservation websites to find the most desirable accommodations. Today's accommodation seekers do not simply accept only advertisements and official information offered by hotels but also read hotel reviews posted by those who already used accommodation services, and consider their evaluations as important references for selecting suitable accommodations [2]. Furthermore, hotel reviews are also expected by hotels to be used for marketing, for example, for grasping things to improve and selling points [3, 4].

Each hotel review consists of 3 elements: the guest's profile, numerical evaluation and impression comment, which reflects the guest's frank impression of his/her actual stay. Although hotel reviews are useful information, they are simply presented successively. Thus, it is difficult to read through a huge number of hotel reviews. In addition, numerical evaluations are subjective evaluations by individuals, and many guests generally tend to give high scores. Furthermore, there are also many guests who just give same scores, for example, full marks to all evaluation items without individually evaluating them. Therefore, it is difficult to know concrete characteristics of each hotel only by processing its numerical evaluations.

It is thus desired to extract feature representations of evaluation from impression comments in hotel reviews and to use them for selecting hotels and for marketing. Impression comments, however, are freely written by guests and have a wide range of expressions. Thus, it is necessary to prepare a thesaurus suitable for the subjects of evaluation and to uniform expressions [4, 5]. The thesaurus is prepared by extracting by eye feature terms from sampled impression comments [4] or by extracting feature terms through term frequency analysis [6]. Therefore, preparing a thesaurus has various problems such as, in addition to being laborious, that the thesaurus needs to be revised when the number of posted hotel reviews has increased and that the evaluation of subjects not under consideration is difficult because the prepared thesaurus is not intended for them.

In this report, we present the result of studying a method of extracting feature representations on evaluation from the impression comments in hotel reviews without setting up subjects of evaluation or preparing a thesaurus.

0730-3157/17 $31.00 © 2017 IEEE
DOI 10.1109/COMPSAC.2017.126

II. RELATED WORKS

A. Evaluation Analysis Dictionary in Hotel Review Analysis

In the study of hotel review analysis by Tsujii et al. [6], they have applied the result of the study on reputation analysis by Kobayashi et al. [7] to hotel reviews and, by further expanding it, prepared "an evaluation attribute dictionary."

In the study by Kobayashi et al. [7], the positive or negative evaluative expression used in opinion analysis consists of three constituents: the <subject>, the <property>, which is concrete items of the subject, and the <evaluation>. In the study by Tsujii et al. [6], the <subject> is set to be the item for numerical evaluation in hotel reviews, such as the "bath" or the "room," the <property> to be concrete items of evaluation, such as the "large public bath" and the "bed," and the <evaluation> to be the condition of the item of evaluation, such as "large." Then they have prepared an evaluation attribute dictionary, in which the positive or negative polarity is attached to the dependency relationship between the <property> and the <evaluation>.

It can be said that the <subject> and the <property> express semantic similarity between words and that the relationship between the <property> and the <evaluation> is expressed in terms of their dependency relationship.

B. Thesauruses

Thesauruses have been used as dictionaries expressing semantic similarity between words not only in natural language processing but also in a wide range of research areas [8]. A typical one is WordNet [9]. In preparing such a thesaurus, however, an enormous amount of manual labor is required to add or update its entry concepts, and it is thus difficult to include latest concepts and an uncommon vocabulary. As a result, it is needed to develop methods of creating accurate thesauruses (semi-)automatically at low cost. Since the accuracy of a thesaurus strongly depends on the group of documents to be analyzed (corpus) and the method of analysis, a variety of methods have been proposed to construct thesauruses by using various analysis methods for various corpora. For example, there are a method using a word clustering method [10] and one using link structures between Web pages [11, 12]. These methods utilize the fact that words used in documents of similar contents tend to be those having semantic similarity.

There is also a study of constructing a thesaurus based on the co-occurrence relationship of words [8]. It is known that there are semantic restrictions in the dependency relationship of words: for example, an adjective "large" modifies a noun expressing a space or an area, such as a "room" or a "bath." Such restrictions are called co-occurrence restrictions. Among the studies of extracting synonyms using co-occurrence restrictions, there is a study which utilizes the frequencies of dependency relationship of nouns with other words [13] and one that utilizes the connection between nouns within compound nouns [14]. There is also a study of extracting synonymous expressions by using co-occurrence restrictions for all parts of speech [15].

III. CO-OCCURRENCE RESTRICTIONS AND DEPENDENCY STRUCTURE GRAPHS

As described in II. A above, the relationship between the <property> and the <evaluation> in the analysis of hotel reviews is expressed in terms of dependency relationship. In addition, as described in II. B, there are co-occurrence restrictions in the dependency relationship of words, and it is expected that the semantic similarity of words can be extracted by using co-occurrence restrictions. Therefore, in the present study we have expressed dependency structures in terms of graph structures and extracted feature representations on evaluation from impression comments in hotel reviews by clustering the graphs.

Its detailed procedure is presented in the following and in Fig. 1.

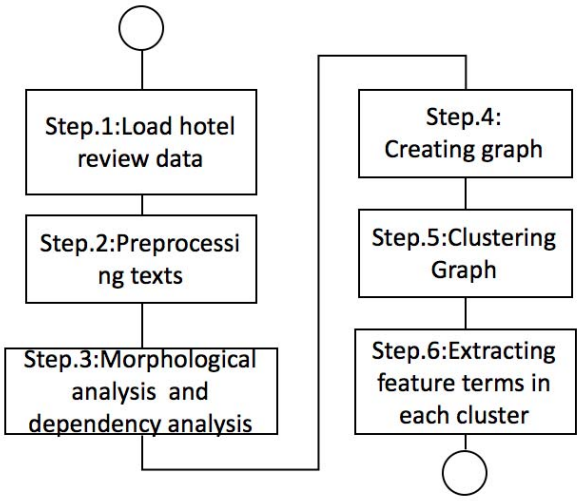

Fig. 1. Example of a figure caption. *(figure caption)*

Step 1. Hotel review data is read. The information to be read in hotel review data is the name of the hotel being reviewed and its impression comments written in Japanese.

Step 2. Text data (impression comments) is preprocessed. Since em (double-byte) characters and en (single-byte) characters can mixedly exist as the characters of the same meaning in Japanese, these writings need to be unified to obtain a correct result in morphological analysis in the next step. In addition, some characters look similar and are often misused interchangeably, such as the prolonged sound mark "ー" and the em minus sign "−," resulting in a wrong result in morphological analysis. Correction for such misuses is also done in the preprocessing.

Step 3. Morphological analysis and dependency analysis are conducted. We have used the morphological analyzer MeCab [16] and the morphological dictionary mecab-ipadic-NEologd [17]. We have also used the dependency analyzer Cabocha [18].

Step 4. Graph data is created from the results of analyses in Step 3. Created graphs are, for example, those shown in Figs. 2 and 3.

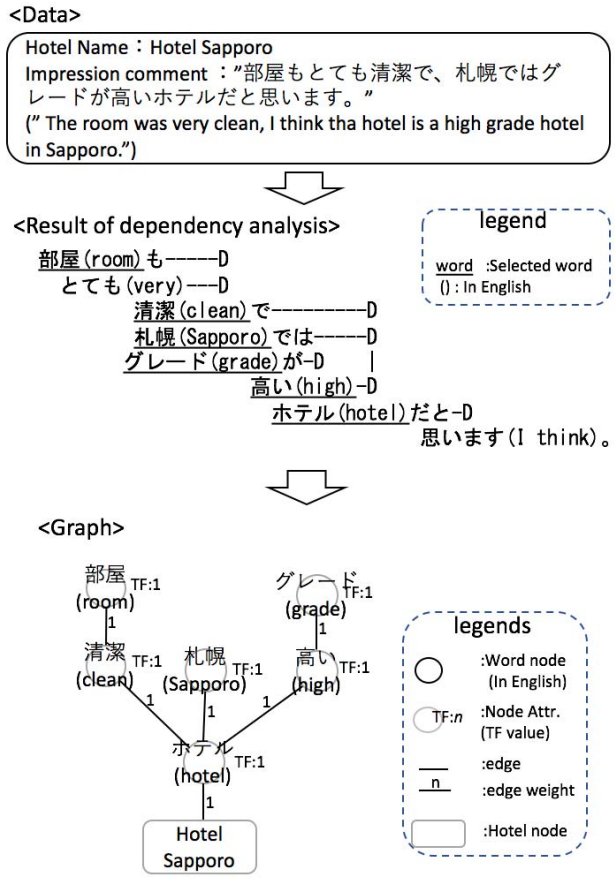

Fig. 2. A result of dependency analysis and the resulting graph

Fig. 2 shows a result of the dependency analysis of impression comments and the resulting graph structure. The result of dependency analysis above are shown in the Tree format, which is the default output format of Cabocha.

In creating a graph structure from the result of dependency analysis, words of specific parts of speech are extracted first. As described in II. A, in the evaluation attribute dictionary, items of evaluation, such as the "large public bath" and the "bed," are set for the <property> and conditions of the items of evaluation, such as "large," for the <evaluation>. Namely, the items of evaluation appear as nouns and their conditions as modifiers in sentences. Thus, we have selected nouns, adjectives and adverbs.

Next, graph data is created from the selected words and their dependency structure. Each selected word is called a node (a term node hereafter) and each dependency relationship an edge. A term ("Hotel" in Fig. 2) which does not have the destination of dependency is connected with the node representing the hotel name under review ("Hotel A" in Fig. 2).

The appearance frequency of each term (term frequency, or TF value hereafter) is set as an attribute of the term node. In addition, the appearance frequency of the dependency relationship between two terms is set as the weight of the edge between the two term nodes.

While Fig. 2 shows an example of creating a graph from an impression comment, multiple impression comments are processed in actual analysis. In such a processing, graphs are merged as shown in Fig.3

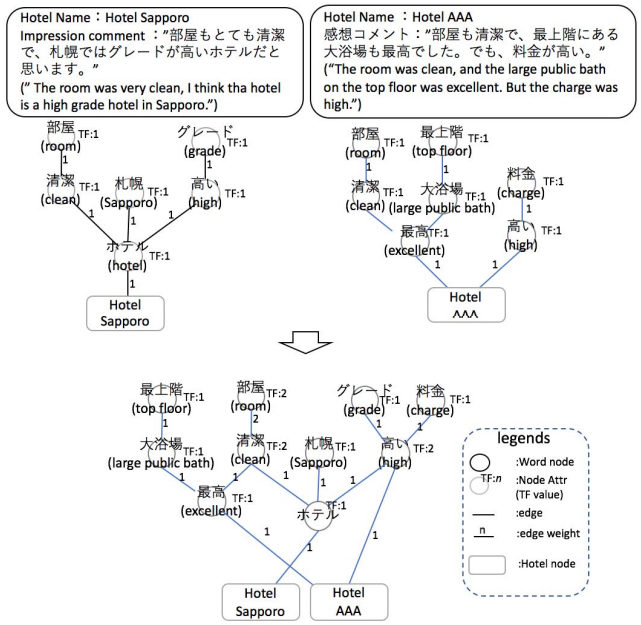

Fig. 3. Merging of the graphs of two impression comments

Fig 3. shows the merging of the graph of the impression comment shown in Fig. 2 with the graph of another impression comment. If the same term node exists in the two graphs, they are shared by the graphs, in which the node attributes, or the TF values, of the nodes are added together. Since "Room," "Clean" and "High" are used in both impression comments in Fig. 3, the TF value of each term node has become 2 in the merged graph. Furthermore, if the edge of the same set of term nodes exist in the graphs, the weights of the edges are added together. Since the dependency relationship between "Room" and "Clean" are used in both impression comments, the weight of the edge between the two term nodes has become 2 in the merged graph.

The processing described above is done for all impression comments, and the clustering of the resulting graphs are conducted in Step 5.

In the clustering of Step 5, we have used the method of R. Lambiotte et al. [19]. In this method, graph clustering is done considering the weights of edges.

In Step 6, feature terms are extracted from each cluster in the obtained result of clustering. Specifically, nodes having high TF values, which are attributes of nodes, are extracted.

IV. RESULT OF EXTRACTING FEATURE TERMS

The following shows the result of extracting feature terms from the impression comments in hotel review data using the method described in III.

We have used the same hotel review data used in the study by Tsujii et al. [6]. The reviewed hotels are 50 hotels in a business area, and the number of impression comments is 4,986.

Table I shows characteristics of the graph data created in Step 4.

TABLE I. CHARACTERISTIC QUANTITIES

Nodes	8,302
Edges	56,071
Average Degree	13.508
Diameter	9
Average Path Length	3.217
Average Clustering Coefficient	0.300

Clustering the data above in Step 5 has produced 17 clusters.

Fig. 4 shows the plots of the relation between the number of term nodes and the average TF value in each cluster.

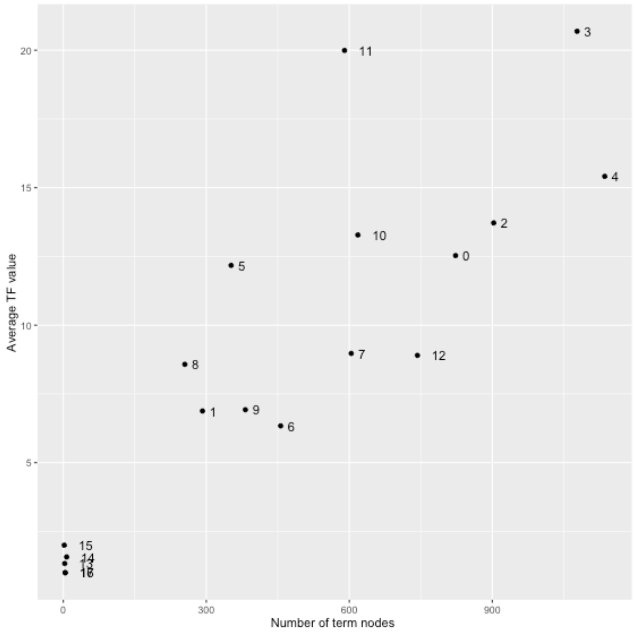

Fig. 4. The number of term nodes and the average TF value in clusters

As can be seen, the number of term nodes and the average TF value in clusters No. 0 to 12 are relatively high, and those in clusters No. 13 to 16 are all extremely low. The following show the details of clusters No. 0 to 12. It is noted that clusters No. 13 to 16 have been found to contain terms which are related to topics of specific shops and events.

Table II shows the top 20 feature terms of high TF values in each cluster.

TABLE II. ORTHOGRAPHICAL VARIANTS OF JAPANESE WORDS

#	Top 20 feature terms
0	Convenient, Station, Near, Location, Place, Excellent, Convenience store, Food, Hakata Station, Busy street, Condition, Subway, Cars, Osaka Station, Transportation, Shopping, Neighborhood, Walk, Direct access, Ginza
1	Amenity goods, Perfect, Towels, Air cleaner, Shampoo, Amenity goods, Set, Flagrance, Equipped, Strong, Shortcoming, Equipped, Air, Aromatic, Air freshener, Replace, Hair conditioner, Toothbrushes, Clean, Goods
2	Breakfast, Satisfied, Very, Tasty, Plan, Buffet, Many, Tasty, Especially, Kinds, Free, Bread, Restaurant, Abundant, Menu, Few, Appreciate, Japanese food, Food, Store
3	Room, Large, Room, Bath, Bed, Neat, Small, Comfortable, Cleanliness, Clean, Room, Tidy, Bath, Night view, Bathroom, Guidance, Enough, Large, Lovely, Shower
4	Use, Stay, Business trip, First time, Trip, Occasion, Nagoya, Always, Sapporo, Definitely, Osaka, Children, Often, Sightseeing, Tokyo, Definitely, Alone, Fatigue, Fukuoka, Two people
5	Charge, Charge, Low, High, Parking lot, Fairly, People, Charge, Economical, Reasonable, Cost performance, Point, Discount, Charge, Parking, Grade, Setting, Honest, Normal, Additional
6	Service, Pleasing, Mood, Impressed, Impressive, Quite, Luxurious, Imperial Hotel, Cakes, Surprised, Heart, Surprising, Very thankful, Newspaper, If possible, Perfect, Yukata (robe), Rich, Satisfied, Personal
7	Check-in, Reservation, Checkout, Luggage, Request, Late, Early, Arrival, Smooth, Confirm, In advance, 12 o'clock, Typhoon, Earlier, In a hurry, Thankful, As wished, Satisfied, Midnight, Jalan (hotel reservation)
8	Elevator, Secure, Card, Floor, Necessary, Key, Crowded, Security, Key, Lounge, Care, Club, System, Door, Security, Elevator, Executive, Locker, Ladies only, Soundproof
9	Large public bath, Hot spring, Pleased, Top floor, Sauna, Open-air bath, Underground, Taking a bath, 1st floor, Appealing, Possible, Narrow, Natural, Undressing, Partly, 24 hours, Open-air, Signs, City hotel, Tatami mat
10	Hotel, Impression, Good, Other, New, Atmosphere, Business hotel, Old, Ordinary, As expected, Building, Reviews, Pleasant, Guests, Itself, Facilities, Evaluation, Interior, Fancy, Like
11	Good, Service, Front desk, Good, Staff, Service, Courteous, Very, Kind, Lady, Wonderful, Pleasant, Fantastic, Really, Access, Impression, Feeling, Face, Employees, Comfortable
12	Unavailable, Disappointed, Unavailable, Intend, Bad, Expectation, Smell, Problem, I, Inconvenient, Anymore, Unsatisfactory, Explanation, Water, So so, One thing, All right, Unavoidable, Feet, Tobacco

Cluster No. 0 (expressed as #0 hereafter) includes terms related with the location of a station, such as "Station" and "Near," and terms which mean the objectives of actions, such as "Food," "Busy street" and Shopping." This cluster has many feature terms related with access to a hotel. The first two terms in #1 are "Amenity goods" and "Perfect," followed by specific nouns for amenity, such as "Towels," "Air cleaner" and

"Shampoo." The cluster includes a collection of feature terms related with amenity. It should be noted that "Amenity goods" appear twice among the terms in #1, and the repetition of some terms is also seen in other clusters. This is because, as shown in Fig. 5, the original Japanese terms have different orthographies. We have confirmed that orthographical variants having the same meaning appear in a cluster: the attachment and the nonattachment of the prolonged sound notation, the notation of a word in Chinese and hiragana characters and a Japanese word and the katakana notation of its English counterpart.

Fig. 5. Orthographical variants of Japanese words

#2 includes a collection of terms related with food, such as "Breakfast," "Satisfied" and "Tasty," and #3 a collection of terms related with the room, such as "Room," "Bath," "Bed," "Large," "Neat" and "Cleanliness." Furthermore, #5 has a collection of terms related with the charge, such as "Charge," "Low," "High" and "Cost performance," and #7 a collection of terms related with check-in and checkout, such as "Check-in," "Reservation," "Checkout," "Late" and "Early." The extracted feature terms in #9 are terms related with the public bath and/or the hot spring, such as "Large public bath," "Hot spring," "Sauna" and "Open-air bath," and those in #11 are terms related with staff service, such as "Good," Service," " Front desk," Staff" and "Service."

In #4, there appear "Use" and "Stay" followed by "Business trip," "Trip," "Occasion" and place names as well as "Alone" and "Children." This seems to be a result of extracting terms related with occasions, which mean staying at a hotel with whom and in what occasions.

In #6, there are terms such as "Service," "Pleasing," "Mood," "Impressed" and "Impressive." These terms come from pleasant experiences of guests, and it seems to be a result of extracting terms related with customer experience.

The extracted terms in #8 seem to be those related with safety and security, such as "Elevator" which cannot be used without a "Card" "Key" and "Locker," which are terms related with "Security," and "Ladies only" and "Soundproof."

The extracted terms in #10 seem to be those concerning the overall evaluation of each hotel, such as "Hotel" and "Business hotel" together with "Good," "Atmosphere" and "Old."

The extracted terms in #12 seem to be those concerning unsatisfactory things, such as "Unavailable," "Disappointed," "Bad," "Problem" and "Inconvenient" together with "Smell," "Water" and "Tobacco."

Table III shows the topic expressed by each cluster, which is presumed based on the terms included in each cluster.

TABLE III. PRESUMED TOPICS OF CLUSTERS

#	Topic
0	Access
1	Amenity
2	Food
3	Room
4	Occasion
5	Charge
6	Customer experience
7	Check-in and checkout
8	Security and safety
9	Bath and hot spring
10	Overall evaluation
11	Staff service
12	Unsatisfaction

The study by Tsujii et al. [6] lists room, bath, food, staff service and location as the <property> in reputation analysis. These properties correspond to #3, #9, #2, #11 and #0 in Table III. While Tsujii et al. also list "Cleanliness" as a <property>, terms related with cleanliness have also been extracted together with those related with the room in #3, as shown in Table II. Furthermore, the study by Tanabe et al. [4] uses, in categorizing and analyzing hotel reviews, the viewpoint of occasion in addition to the viewpoints of room, food, bath and staff service, which are common to the study by Tsujii et al. This viewpoint corresponds to #4 in Table III.

Thus, while the <subject> of evaluation and its constituent <property> and <evaluation> are necessary in the analysis of hotel reviews, as described in II. A, Table III corresponds to the <subject>, and Table II to the <property> and the <evaluation>. As described above, the <property> is a noun and the <evaluation> is a modifier in Japanese. Although we have listed terms in Table II without grouping parts of speech, it is thus possible to group them into the <property> and the <evaluation> according to the parts of speech. In addition, since we have used data of graph structures, it is possible to easily extract relationships between the <properties> and the <evaluations>.

As shown in Table III, we have succeeded in extracting viewpoints of analysis, which are not set in the studies by Tsujii et al. or Tanabe et al., such as #2 Amenity, #3 Check-in and checkout and #8 Security and safety.

V. CONCLUSION

Previously, the preparation of a thesaurus was necessary in the analysis of impression comments included in hotel reviews. The preparation of a thesaurus, however, had problems such that it was laborious and requires occasional revisions. In addition, the thesaurus was not prepared for the subjects not

under consideration, and thus the evaluation concerning such subjects was difficult.

In the present study, we have extracted feature representations from the clusters obtained by graphing groups of impression comments using co-occurrence restrictions and dependency structures of words. Consequently, we have succeeded in extracting feature representations on evaluation from impression comments in hotel reviews without setting up subjects of evaluation or preparing a thesaurus.

While we have used data of hotel reviews in business areas in the present study, we plan to further evaluate the present method by applying it to review data such as that in leisure areas, which seems to have different review characteristics. Furthermore, while we have presumed the topic expressed by each cluster based on the terms included in it, we intend to study the automation of the presumption. For this, it is possible, for example, to use indices such as centrality of each node. Furthermore, it is also possible to analyze which topics of clusters are important, using indices of the nodes.

REFERENCES

[1] Japan Association of Travel Agents 2013 ,"The sale 2013 in a tourist industry"

[2] Tripadvisor, 'Research: What makes a helpful review', http://www.tripadvisor.co.uk/TripAdvisorInsights/n2617/research-what-makes-helpful-review [Accessed 2 April 2015].

[3] Chen Yubo, and Jinhong Xie. "Online consumer review: Word-of-mouth as a new element of marketing communication mix." Management science 54.3 ,2008, pp. 477-491.

[4] Atsusi Tanabe and Masayuki Goto. " A Consideration on User Review Analysis to Support Building Strategies of Accommodation." Journal of the Center for Information Studies 9, 2008, pp.91-101.

[5] Koichi Tsujii, Masakazu Takahashi, and Kazuhiko Tsuda. "Feature Extraction from Numerical Evaluation in Online Hotel Reviews." Procedia Computer Science 60 , 2015, pp.1138-1145.

[6] Koichi Tsujii and Kazuhiko Tsuda. "Method of extracting attention information from hotel reviews using text mining" Journal of Digital Practices. 3 , 2012, pp.289-296.

[7] Nozomi Kobayashi , Ryu Iida , KentaroInui , Yuji Matsumoto. " Extraction of attribute-evaluation pairs and opinion information using anaphora analysis method, NLP2005, 2005, C2-6.

[8] Schutze, H. and Pedersen, J.O.: "A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval" International Journal of Information Processing and Management, Vol.33, No.3, 1997, pp.307–318.

[9] Miller, G.A. "WordNet: A Lexical Database for English" Comm. ACM, Vol.38, No.11, 1995, pp.39–41.

[10] Crouch, C.J. "A Cluster Based Approach to Thesaurus Construction" Proc. ACM SIGIR,1988, pp.309–320.

[11] Chen, Z., Liu, S., Wenyin, L., Pu, G. and Ma, W.Y. "Building a Web Thesaurus from Web Link Structure" Proc. ACM SIGIR, 2003, pp.48–55.

[12] Kotaro Nakayama,Takahiro Hara and Syojiro Nisio. "Wikipedia Mining to Construct a Thesaurus" IPSJ Journal 47.10, 2006, pp.2917-2928.

[13] Akiko Murakami, Tetsuya Nasukawa."Mining Synonymous Expressions usingPersonal Stylistics Variatios" IPSJ SIG TechnicalReport,July 2004 , pp.117-124.

[14] Hidekazu Nakawatase. "A Method for Extraction of Similar Words from Compound Nouns based on Complete Bypatite Graph", Information Processing Society of Japan ,SigDC, Vol.32,No.6,March 2002,pp.39-46,

[15] Yuji Ueno, et al. "A Method for Extraction of Similar Expression using Bipartite Grapf of Word Dependency and Co-occurrences" IPSJ SIG TechnicalReport, 2003-NL-159, 2004, pp.169-176.

[16] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. "Applying Conditional Random Fields to Japanese Morphological Analysis" Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), 2004, pp.230-237.

[17] mecab-ipadic-NEologd : Neologism dictionary for MeCab, https://github.com/neologd/mecab-ipadic-neologd [Accessed 4 April 2017].

[18] Kudo, Taku, and Yuji Matsumoto. "Japanese dependency analysis using cascaded chunking", proceedings of the 6th conference on Natural language learning, August 31, 2002, pp.1-7.

[19] Lambiotte, Renaud, J-C. Delvenne, and Mauricio Barahona. "Laplacian dynamics and multiscale modular structure in networks." arXiv preprint arXiv:0812.1770 ,2008.