Causality model for text data with a hierarchical topic structure

Causality Model for Text Data with a Hierarchical Topic Structure

Takuro Ogawa
Department of Sustainable System Sciences Graduate School of Humanities and Sustainable Systems, Osaka Prefecture University
Japan
saa01052@edu.osakafu-u.ac.jp

Hideyasu Shimadzu
Department of Mathmatical Sciences, Loughborough University
United Kingdom
H.Shimadzu@lboro.ac.uk

Ryosuke Saga
Department of Sustainable System Sciences Graduate School of Humanities and Sustaineble Systems, Osaka Prefecture University
Japan
saga@cs.osakafu-u.ac.jp

***Abstract*—This study describes a method for constructing a causality model from text data, such as review data. Topic modeling is useful to find these evaluation factors from text data. The method based on hierarchical latent Dirichlet allocation is useful because it automatically constructs relationships among topics. However, the depth of each topic in a hierarchical structure is the same even if the contents differ for each topic. Accordingly, the method can generate less important topics that are not worth analyzing. To solve this problem, we construct a hierarchical topic structure with different depths and more important topics by using Bayesian rose trees. In the experiment, the values of the hyperparameters for constructing a hierarchical topic structure are estimated by using evaluation indexes for causal analysis. In addition, the experiment compares the proposed method with related approaches to demonstrate the usefulness of this model.**

Keywords—bayesian rose trees, causality analysis, topic model, hierarchical topic structure

I. INTRODUCTION

In recent years, apps for various services (e.g., Twitter and navigation), the introduction of recommended hotels, and the rise of internet shopping (e.g. Amazon) are rapidly increasing with the advancement of smartphones. Many things can now be performed online. Post evaluation about services and products can be easily conducted, and the amount of evaluation information, such as user reviews and social media for products and services, has considerably increased. Many companies, hotels, and restaurants also post reviews and evaluations about themselves online. SNSs, such as blogs and microblogs, provide evaluations of services and products to other people. Such evaluation information is used not only by consumers but also by producers to improve their services and products. Therefore, analyzing the evaluation of the service and the product is important to improve them.

A user review, as evaluation information, includes text data containing user experience and perception. The evaluation structure of products and services can be understood by analyzing the text data of reviews. Then, the content that has a large effect to evaluation in structure can be understood by causal analysis. Here, text mining is necessary to analyze text data. This mechanism can obtain valuable information from a vast amount of text data [1]. Some methods analyze text data on the basis of word co-occurrence [2]. Other methods analyze emotions [3] from text data through text mining. In addition, a topic model can extract the major theme from a group of text data.

Kunimoto et al. [4] proposed a model that predicts the purchase factors of games from text data by combining hierarchical latent Dirichlet allocation (LDA) (hLDA) [5], which is a topic model with structural equation modeling (SEM) that is used to conduct causal analysis. Their study succeeded in applying SEM to text data. Extracting topics by using hLDA can identify the evaluation factors for each analytical target. However, this previous study did not consider topic granularity. Topic granularity is the richness in content of topic, that is, it is the frequency of the topic in documents. Specifically, this factor is the importance of the topic. Topic granularity generally depends on the content of a topic and differs for each topic. However, the method that depends on hLDA does not consider topic granularity and constructs a structure with the same hierarchy regardless of the topic size. Small topics can generate a low hierarchy and large topics. Fig. 1 shows the hierarchical topic structure of hLDA. The size of the circle in Fig. 1 represents the granularity of the topic. Smaller topics, such as “salmon”, compared with the same hierarchy topics, such as “mammals”, can be generated because hLDA constructs a structure with the same hierarchy. Specifically, causal models can include unimportant and invaluable topics.

This research aims to solve the problem of existing studies that conduct causal analysis by using a hierarchical topic structure. To solve this problem, LDA [6] and the Bayesian rose

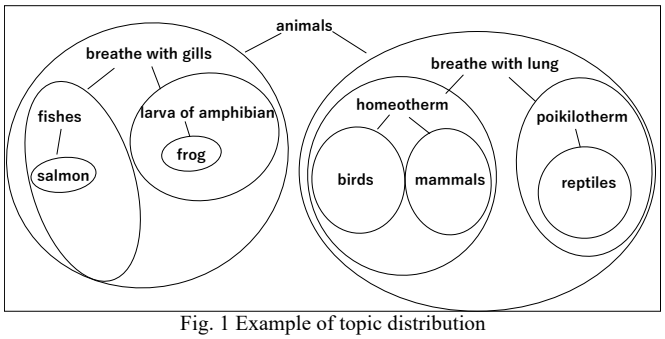

Fig. 1 Example of topic distribution

trees (BRT) [7] are used to generate major topics with a high degree of granularity, and the topics are used to construct a hierarchical topic structure to generate hierarchical relationships by a bottom-up method. Therefore, the topic granularity of the bottom layers is higher than that of hLDA that generates a bottom topic by considering the higher topic. This structure is the image that deleted topics "salmon" and "frog" from Fig. 1, and the bottom topics are "fishes", "larva of amphibian", "birds", "mammals", and "reptiles".

Several important factors and useful evaluation structures could be discovered to improve services and products. In this study, simBRT, one of the BRTs, is used to construct a causal model. A causal analysis is conducted by using SEM.

However, the hyperparameters for constructing a hierarchical topic structure were not discussed in a study of simBRT [8]. These hyperparameters are important factors because the structure of the hierarchical topic depends on them. We estimate the value of each hyperparameter by using indexes for causal analysis (SEM) instead of topic evaluation indexes because this study aims to conduct an accurate causal analysis in the experiment. The proposed approach is compared with causal analysis that uses hLDA and SEM to confirm that the simBRT method can construct better models than the hLDA method and considers topic granularity.

The remainder of this paper is organized as follows: Section 2 presents the existing related research. Section 3 explains the BRT method, which is the core technology. Section 4 describes the analytical experiments by using actual data. Section 5 concludes this work and discusses future studies.

The contributions of this study are as follows:

- This study constructs a causal model that considers topic granularity with a different layer for each topic.
- This study estimates the values of hyperparameters for constructing a hierarchical topic structure by using indexes for causal analysis.
- This study compares the proposed method with existing approaches to confirm the feasibility of the proposed approach in an experiment.

II. Literature Review

A. Topic Models

Topic models are algorithms for determining the major themes that pervade in a large and otherwise unstructured collection of documents. Topic models can organize such collection in accordance with the identified themes [9].

Topic models include different methods, such as latent semantic analysis (LSA) [10], LDA, and hLDA. The LDA assumes a multitopic model in which the document is based on mixed topics. LDA has a 1:*n* relationship between documents and topics, not 1:1, such as LSA. LDA is considered to be a natural model in documents, such as review texts that are written in one document about various aspects. HLDA is an extended method. This method can automatically construct relationships among hierarchical topics.

Identify applicable funding agency here. If none, delete this text box.

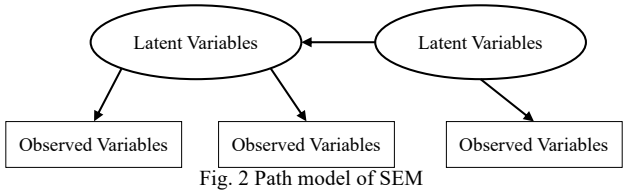

Fig. 2 Path model of SEM

B. SEM

SEM [11] is a technology that is characterized by the use of factor and regression analysis. Factor analysis is a method wherein the observed variables are based on some hidden factors, and the influence of the factor is to be determined by "correlation" (variance/covariance). Regression analysis is a technique for finding the relationship between a variable to be predicted (target variable) and a variable (explanatory and independent variables) that describes the target variable.

The SEM can visually and quantitatively express causal relationships between variables by using a path model (Fig. 2). A path model consists of three elements: latent variables, observed variables, and paths. Latent variables are factors that cannot be observed in actual. Observation variables can actually be observed and are essential for estimating a latent variable. In the path model, the latent variables are represented by ellipses, and the observation variables are represented by rectangles. The causal relationship between such items is represented by the path of the arrow, and the degree of influence is denoted by the path coefficient. Therefore, in the causal analysis for SEM, the manner by which to construct a path model must be determined.

C. Related Work

Several methods can be applied to construct a path model for SEM. SERVQUAL [12] is used to construct a path model. Al-Mhasnah et al. proposed a method that uses SERVQUAL and SEM to examine the effects of the former [13]. Ali et al. improved the SERVQUAL index and analyzed it through SEM [14]. Bivina et al. used the main aspects of a pedestrian label of service (PLOS) [15] to provide a comfortable and safe walking environment. PLOS is a measurement tool for evaluating the degree of pedestrian accommodation on roadways. SEM is used to provide the essential information for interpreting the aspects of the walking environment that influence PLOS [16]. Many indicators are available; however, various services and products are difficult to measure by using one standard because of the many types of services and products, and their characteristics largely differ.

Meanwhile, another method uses a topic model to construct a path model. Saga et al. attempted to analyze the factor relationships of the game software market by using a topic model [17]. They proposed a path model generation process for SEM by using LSA and combined the text data of user reviews with the model. This method requires each document to belong to only one topic. Consequently, the model cannot express natural variables and relationships. Saga et al. extended this method to LDA and generated a path model from the topics extracted by using LDA [18]. However, LSA and LDA cannot define the relationships among topics in the learned model. To solve this problem, Kunimoto et al. [4] proposed a model that predicts the purchase factors of games from text data by combining hLDA and SEM, as mentioned in Section 1. Ogawa

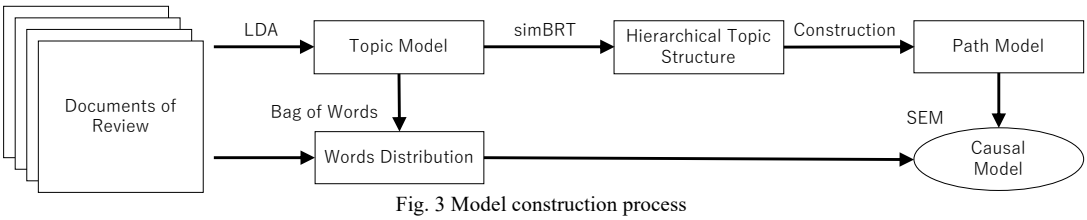

Fig. 3 Model construction process

et al. extended the topic structure of hLDA to model the considered emotional factor [19]. However, all topics have the same depth because these methods depend on hLDA. Therefore, the preceding studies do not consider topic granularity.

Pachinko allocation model (PAM) [20] and hierarchical PAM (hPAM) [21] are also available, in addition to hLDA. PAM documents, a mixture of distributions over a single set of topics, use a directed acyclic graph to represent topic co-occurrences. Each node in the graph is a Dirichlet distribution. HPAM is an extension of PAM. In hPAM, every node is associated with a distribution over the vocabulary. However, the hierarchical topic structures of these methods also have the same hierarchy regardless of topic granularity. A method for extracting a hierarchical topic structure that combines the biterm topic model (BTM) [22] and BRTs is also available [8]. The aforementioned method uses simBRT for considering topic similarity. This study conducts a time series analysis on the basis of the constructed hierarchical structure. Several studies have analyzed time series on the basis of a hierarchical topic structure by using BRTs. However, no study has conducted causal analysis on the basis of a hierarchical topic structure using BRTs.

In this study, causal analysis based on a model constructed using simBRT is conducted to consider topic granularity. Each document should have many words and should be characterized for SEM. Therefore, LDA is used instead of BTM because the document contains many words.

III. Construction Model Using BRTs

In this study, analysis is performed in accordance with the process shown in Fig. 3. First, a topic is extracted using the topic model. Then, the topic is represented in the hierarchical topic structure by using BRTs. Lastly, causal analysis is conducted via SEM on the basis of the model constructed using BRTs.

A. BRTs

A BRT is a probabilistic approach for hierarchical clustering and an extended method of Bayesian hierarchical clustering [23]. BRT greedily predicts a tree structure on the basis of the probability $P(C|T)$ that represents the likelihood of data C given tree T. In this study, the topics are used as data D. All topics $C = \{t_1, t_2, \dots t_k\}$ extracted using LDA are the leaves.

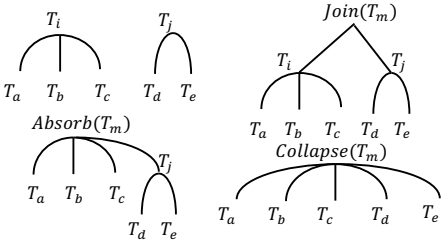

Fig. 4 Three merging operations

First, each topic t_k is regarded as an individual tree $T_i = \{t_i\}$. BRT is repeated to combine two trees that are selected for making a new tree T_m on the basis of three basic operations (join, absorb, and collapse). Fig. 4 shows the three basic operations.

(1) Join: $T_m = \{T_i, T_j\}$
(2) Absorb: $T_m = \{ch(T_i), T_j\}$
(3) Collapse: $T_m = \{ch(T_i), ch(T_j)\}$

Here, $ch()$ denotes a tree's set of children. For example, in T_i in Fig. 4, $ch(T_i)$ is $\{T_a, T_b, T_c\}$. The join operation is the traditional operation in a binary tree. Meanwhile, the absorb and collapse operations cater to a multibranch tree. The three operations are conducted in all combinations of trees. The combination and the operation with the maximum probability ratio are selected as follows:

$$\frac{p(C_m|T_m)}{p(C_i|T_i)p\left(C_j\middle|T_j\right)}, \tag{1}$$

where $C_m = T_i \cup T_j$ are the topics under the tree structure T_m. $p(C_m|T_m)$ is the likelihood of topic C_m under T_m. $p(C_m|T_m)$ can be calculated using a dynamic programming paradigm, as follows:

$$p(C_m|T_m) = \pi_{T_m} f(C_m) + (1 - \pi_{T_m}) \prod_{T_i \in ch(T_m)} p(C_i|T_i), \tag{2}$$

where π_{T_m} is the prior probability that all the topics in T_m are maintained in the same partition, and π_{T_m} is defined as follows:

$$\pi_{T_m} = 1 - (1 - \gamma)^{n_{T_m} - 1}, \tag{3}$$

where n_{T_m} is the number of children of T_m, and γ $(0 < \gamma < 1)$ is a hyperparameter of the model that controls the relative proportion of partitions of the data. The $f(C_m)$ of (2) is the marginal probability of C_m, which can be modeled by the Dirichlet compound multinomial model [24] distribution and is defined as follows:

$$f(D) = \prod_i^n \frac{\sum_j^V x_i^{(j)}!}{\prod_j^V x_i^{(j)}!} \cdot \frac{\Delta(\alpha + \sum_i x_i)}{\Delta(\alpha)}, \tag{4}$$

$$\Delta(\alpha) = \frac{\prod_{j=1}^V \Gamma(\alpha^{(j)})}{\Gamma(\sum_{j=1}^V \alpha^{(j)})}, \tag{5}$$

where $x_i^{(j)}$ is the frequency of keyword j included in topic i, V is the total number of vocabulary, and $\alpha = (\alpha^{(1)}, \alpha^{(2)}, \dots \alpha^{(j)})$ is the hyperparameter that specifies the distribution over the probability simplex. Γ is the gamma function, and $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.

In addition, simBRT is used to consider topic similarity. Here, topic distribution is used as the data of a tree. Therefore, topic similarity should be considered. SimBRT considers the

similarity of topic distribution in (1). Topics are distributed over vocabularies. Thus, the Kullback-Leibler (KL) divergence can be used to measure the similarity between two topics. The similarity of topics z_i and z_j is defined as follows:

$$topic\_sim(z_i||z_j) = \frac{1}{\frac{KLD(z_i||z_j)+KLD(z_j||z_i)}{2}+1}, \quad (6)$$

$$KLD(z_i||z_j) = \sum_{k=1}^{V} \phi_{ik} \log(\phi_{ik}/\phi_{jk}), \quad (7)$$

where $KLD(z_i||z_j)$ is the KL divergence between topics z_i and z_j, and ϕ_{ik} is the topic distribution in this work. simBRT defines the weighted topic distribution WT in each operation to obtain topic similarity in tree construction.

$$\text{Join: } WT = \frac{avg(C_i)p(C_i|T_i)+avg(C_j)p(C_i|T_i)}{p(C_i|T_i)+p(C_j|T_j)}, \quad (8)$$

$$\text{Absorb: } WT = \frac{avg(C_i)p(C_i|T_i)+\sum_{T_a \in ch(T_j)} avg(C_{T_a})p(C_{T_a}|T_a)}{p(C_i|T_i)+\sum_{T_a \in ch(T_j)} p(C_{T_a}|T_a)}, \quad (9)$$

$$\text{Collapse: } WT = \frac{\sum_{T_a \in ch(T_i)} avg(C_{T_a})p(C_{T_a}|T_a)+\sum_{T_b \in ch(T_j)} avg(C_{T_b})p(C_{T_b}|T_b)}{\sum_{T_a \in ch(T_i)} p(C_{T_a}|T_a)+\sum_{T_b \in ch(T_j)} p(C_{T_b}|T_b)}, \quad (10)$$

where $avg(C_m)$ is the average of the topic distribution of the topics that are included by C_m. Specifically, $avg(C_m)$ is the topic distribution that is simply calculated on the basis of the average. Here, the final merged topic distribution under T_m is $avg(C_m)$. Then, the topic similarity between the weighted topic WT and the final merged topic $avg(C_m)$ is added into the primitive objective function in (1). Thus, (1) can be rewritten as follows:

$$\frac{p(C_m|T_m)}{p(C_i|T_i)p(C_j|T_j)} * topic\_sim(WT||avg(C_m)), \quad (11)$$

The operation that maximizes (11) is conducted at each step to construct a hierarchical topic structure.

B. Construction of Path Model

The topics that cannot be directly observed are considered latent variables that function as correspondence between the SEM and a topic model. The keywords that comprise a topic are the observation variables because they are words that actually exist in reviews. The idea of a topic model is characterized by the generation of words by topics. Each topic is regarded as a factor, and the path from topics is drawn to the keywords to which topics are related.

Subsequently, the representation of a hierarchical topic structure is described. Some factors are considered in merging topics. When two trees are merged, a node shown as a factor is regarded as a large topic that includes two topics. This node is regarded as a topic and used as a latent variable. The paths between topics are drawn from the upper topics to the lower ones on the basis of the idea that large topics generate small topics. A path is also drawn from the top topic to rate the numerical evaluation of review data and understand the relation between topic structure and actual numeric data. Therefore, a dataset must have text data and rating evaluation expressed by a numerical value to apply this method. In addition, the number of review documents and the length of text data must not be particularly small to apply LDA.

IV. Experiment

This experiment aims to confirm the feasibility of the proposed method by constructing the model described in Section 3 and to compare the proposed approach with existing studies that used hLDA. In addition, this section considers the experimental results and discusses the hyperparameters.

A. Dataset, Indexes for Evaluation, and Hyperparamaters

In this experiment, the data should ideally have as many review data as possible to apply the topic model. The text of a review datum must include many words to characterize the statistical data on the basis of the concept of bag of words. The user reviews in the datasets published online by Kaggle and Github and Amazon are used. The reviews of airports, hotels, apps for shops, electronic services for purchasing clothes, and musical instruments are selected. Each review has a review text with a rating between 1 and 5 or 1 and 10. A review text is also regarded as a document. Only documents expressed with more than 30 words are used to ensure that the topics and the appearance frequency of the described feature words are included in each document. The app analyzes information from randomly extracted data. The number of reviews after this preprocessing is provided in Table 1.

The goodness-of-fit index (GFI), adjusted GFI (AGFI), and root mean square error of approximation (RMSEA) are adopted as indexes to evaluate the result. GFI indicates how well the total variance in the saturation model can be explained by the estimation model. A value between 0 and 1 is considered, and that close to 1 denotes a good model. A value of 0.9 or higher is desirable. GFI is unconditionally improved in fitness as a model's degree of freedom decreases. AGFI corrects the shortcomings of GFI and penalizes models with many parameters and high complexity. The same value as that in GFI is considered, and a value close to 1 indicates a good resultant model. If the model is not complex, t then the values of GFI and AGFI are close to each other. RMSEA is an index that expresses the difference between the model distribution and the actual distribution. Fit is good with a value of 0.05 or less, and if the value is 0.1 or higher.

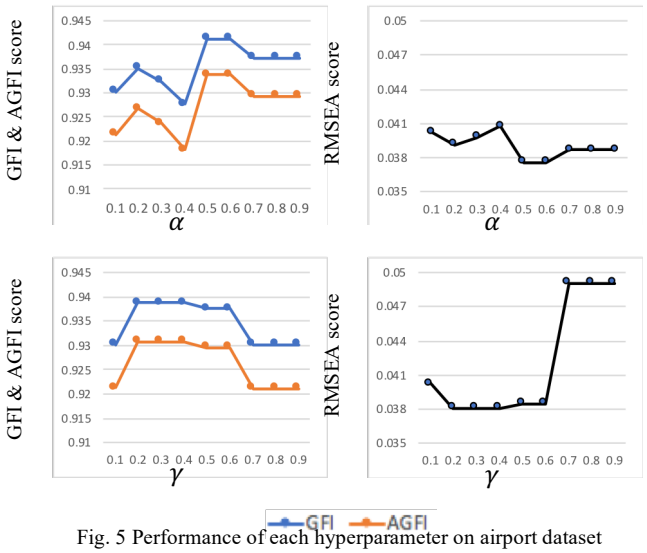

Fig. 5 Performance of each hyperparameter on airport dataset

TABLE I. Dataset and Results of evaluation indexes

Dataset	Number	method	GFI	AGFI	RMSEA
app	5588	simBRT	0.9614	0.9548	0.03179
		hLDA	0.9451	0.9390	0.03498
hotel	8201	simBRT	0.9720	0.9686	0.02339
		hLDA	0.9528	0.9480	0.02996
airport	13466	simBRT	0.9437	0.9365	0.03677
		hLDA	0.9443	0.9342	0.04143
instrument	8538	simBRT	0.9478	0.9374	0.04175
		hLDA	0.9082	0.8967	0.04797
e-commerce	19599	simBRT	0.9768	0.9722	0.02750
		hLDA	0.9715	0.9680	0.02564

TABLE II. The evaluation indexes for each γ and α

γ	α	GFI	AGFI	RMSEA
0.2	0.5	0.9429	0.9357	0.03693
0.2	0.6	0.9437	0.9365	0.03677
0.3	0.5	0.9415	0.9341	0.03734
0.3	0.6	0.9415	0.9341	0.03734
0.4	0.5	0.9415	0.9341	0.03734
0.4	0.6	0.9415	0.9341	0.03734

γ , α, the number of topics and words comprise a topic are the hyperparameters. In this experiment, the number of bottom topics is 10, and that of words that comprise a topic is 5. We change hyperparameters γ and α to observe the influence of these parameters on the performance.

Several package and libraries, namely, Python's genism for LDA [25], Mallet package for hLDA [26], and the SEM package of R for SEM analysis, are used in this experiment [27].

B. Results

Fig. 5 shows the result of the evaluation indexes for each hyperparameter for the airport data and analyzed models. We fix one of the indexes to 0.1 to observe the influence of one parameter and understand that the evaluation indexes are best when $\gamma = 0.2, 0.3, 0.4$ and $\alpha = 0.5, 0.6$. Table 2 shows the result of the evaluation indexes of these parameters, and the best performance is $\gamma = 0.2$ and $\alpha = 0.6$.

Table 1 presents the calculation results of the evaluation indexes for each data and analyzed models when γ is 0.2 and α is 0.6. In the table, all the models have a GFI and AGFI of over 0.9. Moreover, the models have RMSEA values of less than 0.05. The BRT results of all the models, except for the RMSEA of e-commerce and the GFI of airport, have higher values than the hLDA results. In the hLDA model, 10 topics are obtained from the bottom layer and the keywords that comprise a topic at the upper layer are deleted to achieve the same situation as that in the BRT model.

For example, Fig. 6 shows the analysis result model of the airport dataset. The causal relation between keywords that comprise a topic is presented similar to that in Fig. 7.

The words at the bottom of the model are those that make up the identified topics from the text data of the review by using the topic extraction with LDA. Here, the contents of the topics (latent variables) are estimated by authors from the words that make up each topic. For example, "access" is estimated by different access features, namely, "car", "taxi", and "train". The causal relationships can be analyzed by paying attention to the arrow and values calculated by SEM between topics or between topics and words at the bottom of the model.

This study focuses on the topics "airport service" and "country". Accordingly, we can understand that these topics have topics of deep hierarchy, such as "flight" and "Asian country". Topics "flight" and "empathy" have topics of deeper hierarchy, such as "procedure". In this way, this method can construct a hierarchical structure with different hierarchies for each topic. Next, we focus on the path from "airport" to understand an important factor. "Airport condition" and "airport service" are important factors because they have large path coefficients. This study focuses on the path from these topics to understand many detailed important factors. Accordingly, the important factors that have a large effect to evaluation are understood by focusing on the path and path coefficient.

The airport structure can be reviewed by examining the results of the analysis of airport data. Airports are evaluated using certain topics, such as "airport condition" and "airport

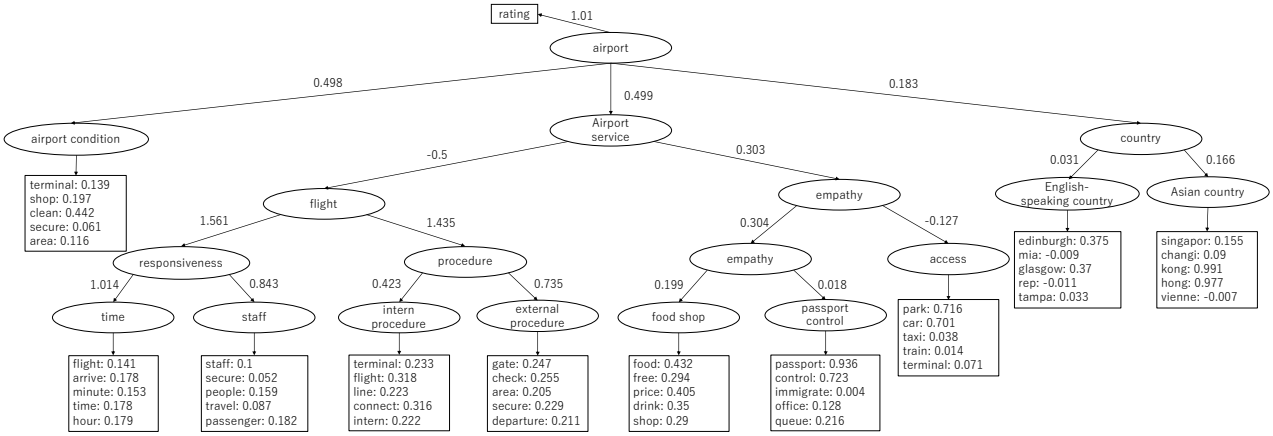

Fig. 6 Analysis result of airport based on BRT

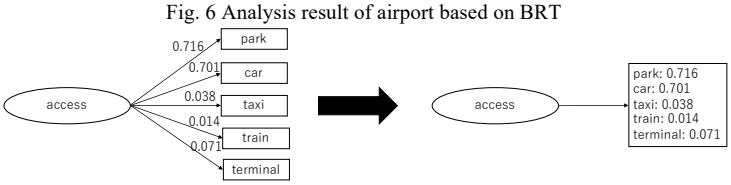

Fig. 7 Expression of a path from the latent variable to the observed

service". This study focuses on low hierarchy to analyze the details of the evaluation factors, such as "intern procedure" and "access".

V. Conclusion

In this study, a hierarchical topic structure was represented by BRTs on the basis of a topic extracted from text data. A path model of SEM was also constructed on the basis of this hierarchical topic structure, and a causal analysis was conducted. In the experiment, the values of hyperparameters γ and α are estimated by using evaluation indexes for the SEM. The value of our proposed method was demonstrated by the result of an experiment that used reviews of apps, hotels, airports, musical instruments, and e-commerce.

In existing causal analyses that used hLDA, all topics have the same depth of layer, and this method cannot consider topic granularity. On this basis, a hierarchical topic structure is constructed using simBRT on the basis of the topics of LDA. This method can construct a hierarchical topic structure with different hierarchies for each topic by using the topics that have large granularity. In the experiment, several services and products were analyzed to confirm the feasibility of the proposed method, and the values of hyperparameters are discussed. In each index, satisfaction values were found for all the datasets. The result of the developed method was compared with that of the analysis that used hLDA. The former was found to be a better model. In addition, the services and products can be visually and quantitatively evaluated using the proposed model (Fig. 6).

In the future, a topic is defined as a bag of words without explicit semantics. In this study, the contents of the topics are estimated using the words that compose them. However, the topic model loses objectivity. We, we can use topic labeling to address this issue.

References

[1] M. Hearst, "What is Text Mining?", www.sims.berkeley.edu/~hearst/text-mining.html, [retrieved: March, 2020]

[2] Y. Matsuo and M. Ishizuka, "Keyword Extraction from a Single Document Using Word Co-occurence Statistical Information", International Journal on Artificial Intelligence Tools, vol. 13, No. 01, pp. 157-169, 2004.

[3] C. J. Hutto and E. Gilbert, "VADER: a parsimonious rule-based model for sentiment analysis of social media text", Proceedings of the Eighth International AAAI Conference om Web and Social Media, pp. 216-225, May 2014.

[4] R. Kunimoto and R. Saga, "Causal Analysis of User's Game Software Evaluation Using hLDA and SEM", The Institute of Electrical Engineers of Japan Transactions on Electronics Information and Systems, vol. 135, Issue 6, pp. 602-610, 2015.

[5] D. M. Blei, T. L. Griffiths, M. I. Jordan and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process", Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 17-24, December 2003.

[6] D. M. Blei, A. Y. Ng, J. B. Edu and M. I. Jordan, "Latent Dirichlet allocation", The Journal of Machine Learning Research, No. 3, pp. 993-1022, 2003.

[7] C. Yee, Y. W. Teh and K. A. Heller, "Bayesian Rose Trees", UAI'10: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, pp. 65-72, July 2010.

[8] J. Zhu, X. Li, M. Peng, J. Huang, T. Qian, J. Huang and et al., "Coherent Topic Hierarchy: A Strategy for Topic Evolutionary Analysis on Microblog Feeds", Proceedings of 16th International Conference on Web-Age Information Management, pp. 70-82, June 2015.

[9] D. M. Blei, "Probabilistic Topic Models", Communications of the ACM, vol. 55, No. 4, pp. 77-84, April 2012.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis", Journal of The American Society for Information Science, vol. 41, Issue 6, pp. 391-407, 1990.

[11] C. M. Stein, N. J. Morris and N. L. Nock, "Structural Equation Modeling", Statistical Huma Genetics: Methods and Protocols, Methods in Molecular Biology, vol. 850, pp. 495-512, January 2012.

[12] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality", Journal of Retailing, vol. 64, No. 1, pp. 12-40, 1988.

[13] A. M. Al-Mhasnah, F. Salleh, A. Afthanorhan, and P. L. Ghazali, "The relationship between services quality and customer satisfaction among Jordanian healthcare sector", Management Science Letters, vol. 8, Issue 12, pp. 1413-1420, 2018.

[14] M. Ali and S. A. Raza, "Service quality perception and customer satisfaction in Islamic banks of Pakistan: the modified SERVQUAL model", Total Quality Management & Business Excellence, vol. 28, Issue 5-6, pp. 559-577, November 2015.

[15] K. E. Zannat, D. R. Raja, and M. S. G. Adnan, "Pedestrian Facilities and Perceived Pedestrian Level of Service (PLOS): A Case Study of Chittagong Metropolitan Area, Bangladesh", Transportation in Developing Economies, vol. 5, Issue 2, pp. 1-16, April 2019.

[16] G. R. Bivina and M. Parida, "Modelling perceived pedestrian level of service of sidewalks: a structure equation approach", Transport, vol. 34, No. 3, pp. 339-350, May 2019.

[17] R. Saga, T. Fujita, K. Kitami and K. Matsumoto, "Improvement of Factor Model with Text Information Based on Factor Model Construction Process", Proceedings of the 6th International Conference on Intelligent Interactive Multimedia Systems and Services, pp. 222-230, 2013.

[18] R. Saga and R. Kunimoto, "LDA-based Path Model Construction Process for Strucutre Equation Modeling", Artificial Life Robotics, vol. 21, Issue 2, pp. 155-159, 2016.

[19] T. Ogawa and R. Saga. "Text-based Causality Modeling with Emotional Information Embedded in Hierarchic Topic Structure", Proceedings of the Ninth International Conference on Social Media Technologies, Communication, and Informatics, pp. 15-20, November 2019.

[20] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations", ICML'06: Proceedings of the 23rd international conference on Machine learning, pp. 577-584, June 2006.

[21] D. Mimno, W. Li and A. McCallum, "Mixtures of hierarchial topics with Pachinko allocation", ICML'07: Proceedings of the 24th international conference on Machine learning, pp. 633-640, June 2007.

[22] X. Yan, J. Guo, Y. Lan and X. Cheng, "A biterm topic model for short texts", WWW'13: Proceedings of the 22nd international conference on World Wide Web, pp. 1445-1456, May 2013.

[23] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering", ICML '06: Proceedings of the 22nd international conference on Machine learning, pp. 297-304, August 2005.

[24] R. E. Madsen, D. Kauchak and C. Elkan, "Modeling word burstiness using the Dirichlet distribution", ICML'05: Proceedings of the 22nd international conference on Machine learning, pp. 545-552, August 2005.

[25] R. eh and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50, May 2010.

[26] A. Kachites, "Mallet: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu, [retrieved: January, 2020]

[27] R. Ihaka and R. C. Gentleman, "The R Project for Statistical Computing", https://www.r-project.org, [retrieved: January, 2020]