Text-based Causality Modeling with a Conceptual Label in a Hierarchical Topic Structure Using Bayesian Rose Trees

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

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

Abstract

This paper describes a method for constructing a causality model from review text data. Review text data include the evaluation factors of rating, and causality model extraction from text data is important for understanding the evaluation factors and their relationships. Several methods are available for extracting causality models by using a topic model. In particular, the method based on hierarchical latent Dirichlet allocation is useful for hierarchically comprehending causality structure. However, the depth of each topic in a hierarchical structure is forcefully pruned even if granularities differ for each topic. Thus, interpreting a hierarchical topic structure is difficult. To solve these problems, we construct a hierarchical topic structure with different depths by using Bayesian rose trees. Furthermore, we use conceptual labeling to add explicit semantics for each topic for interpretation. An experiment confirms that this model is accurate and interpretable using actual data.

1. Introduction

The amount of evaluation information, such as from user reviews and social media for products and services, has increased considerably in recent years. For example, many websites provide product reviews and some social networking services (SNSs) also review hotels and restaurants. At present, many companies, hotels, and restaurants post reviews and evaluations about themselves online. SNSs, such as blogs and microblogs, provide evaluations of services to other people. Such evaluation information is used not only by consumers, but also by producers to improve their products and services and develop new ones.

As a piece of evaluation information, a user review 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. Text mining is necessary for text data analysis. Moreover, text mining can obtain valuable information from a vast amount of text data [1]. Some methods analyze text data according to word co-occurrence [2]. Other techniques 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 (hLDA) [5], i.e., a topic model with structure equation modeling (SEM). Their study succeeded in applying SEM to text data. Extracting hierarchical topic structure by using hLDA can identify the evaluation factors for each analysis target. However, this previous study disregarded topic granularity. Topic granularity is the richness in content of topic, in other words it is frequency of the topic in documents. That is, it is the importance of the topic. Topic granularity usually 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 structures with the same hierarchy regardless of topic size. That is, minor topics can generate low hierarchy as well as major topics. The image of Figure 1 shows the hierarchical topic structure of hLDA and the size of the circle represents the granularity of the topic. Smaller topics

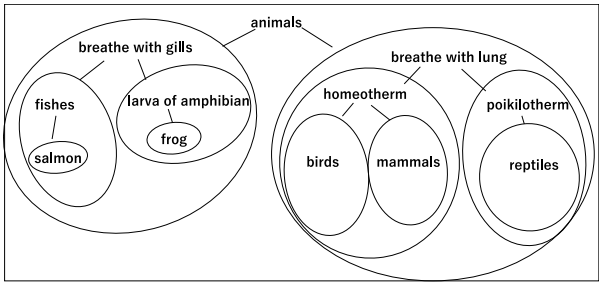

Figure 1. Example of topic model of hLDA

URI: https://hdl.handle.net/10125/70745
978-0-9981331-4-0

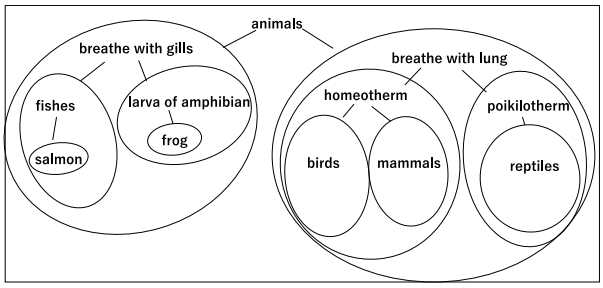

Figure 2. Example of topic model of BRTs

such as ‘salmon’ compared to same hierarchy topics such as ‘mammals’ can be generated because hLDA constructs structure with same hierarchy. That is, causal models can include unimportance and invaluable topics.

In addition, a topic is a bag of words without explicit semantics. Humans usually find it difficult to interpret a topic without explicit semantics. In particular, interpreting a hierarchical topic structure than a general topic is more challenging the former has more complexity.

The objective of the current research is to solve the problem of existing studies that conduct causal analysis using a topic structure. This work describes a method for conducting causal analysis by using evaluation factors from text data, such as user reviews. A topic model is employed to perform causal analysis. Latent Dirichlet allocation (LDA) [6] is adopted as the topic model, and a hierarchical topic structure is found on the basis of Bayesian rose trees (BRTs) [7]. The image of Figure 2 shows the hierarchical topic structure of combining LDA and BRT. LDA generates bottom topics ‘fishes’, ‘larva of amphibian’, ‘birds’, ‘mammals’ and ‘reptiles’, and the BRTs generates hierarchical relationships by a bottom-up method. LDA is used to generate major topics with a high degree of granularity. Therefore, the topics granularity of the bottom layers is higher than that of hLDA that generates bottom topics by considering the higher topic. In this way, several more important factors and more useful evaluation structures can be discovered. User reviews also quantitatively analyze the impact of each factor by applying the constructed topic structure to SEM. Furthermore, we must provide explicit semantics for each topic that involves a bag of words. That approach is useful for understanding each topic. Here are a few examples:

1. breakfast, dinner, dog, cat → meal, pet
2. Japan, America, Canada, China → country
3. guitar, piano, soccer, golf → instrument, sport

For giving explicit semantics, we use hierarchical conceptual labeling (HCL) [8]. HCL can generate a semantic conceptual label set from a bag of words using Microsoft concept graph (MCG) [9] as a knowledge base. Moreover, HCL can effectively delete noise words from a bag of words. We can therefore construct an interpretable causal model using HCL.

The remainder of this paper is organized as follows: Section 2 presents the existing related research. Section 3 explains the BRT and HCL methods, which are the core technologies in this work. Section 4 describes analysis experiments using actual data.

The contributions of this study are the following.

- This work constructs a model with a different layer for each topic to analyze causality.
- This research improves interpretability for understanding hierarchical topic structure.

2. Literature Review

This section describes the methods that our approach uses and presents the existing related research.

2.1. SEM

SEM is a technology that is characterized by the use of factor and regression analyses [10]. In factor analysis, the observed variables are based on some hidden factors, and the influence of a factor is determined via “correlation” (variance/covariance). Regression analysis is a technique for determining the relationship between the variable to be predicted (i.e., the target variable) and a variable that describes the target variable (e.g., explanatory or independent variables).

SEM can visually and quantitatively express causal relationships between variables by using a path model. A path model consists of three elements: latent variables, observed variables, and paths. Latent variables are factors that cannot be actually observed. Observation variables are observable and are essential for estimating latent variables. The latent and observation variables in a path model are represented by ellipses and rectangles, respectively. The causal relationship between such items is represented by the path of an arrow, and the degree of influence is depicted by the path coefficient.

2.2. Topic Models

Topic models are algorithms for determining the major themes that pervade an extensive and otherwise unstructured collection of documents. Topic models can organize such collection in accordance with the identified themes [11].

Topic models include various methods, such as latent semantic analysis (LSA) [12], LDA, and hLDA. LDA assumes a multi-topic model in which the document is made on the basis of mixed topics. LDA exhibits a *1:n* relationship between documents and

topics instead of a *1:1* relationship, such as that found with LSA. LDA is regarded as a more natural model for documents and reviews text written in one document about various aspects.

2.3. Related Work

Several methods are available for constructing a path model for SEM using text data. Saga et al. attempted to analyze the factor relationships of the game software market by using a topic model [13]. They proposed a path model generation process for SEM using LSA and then combined the text data of user reviews with the model. This technique requires each document to belong to only one topic. Thus, the model cannot express natural variables and relationships. Accordingly, Saga et al. extended this method to LDA and generated a path model from the resulting topics extracted [14]. 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 et al. extended the topic structure of hLDA to model the considered emotional factor [15]. Nevertheless, all topics have the same depth because these methods depend on hLDA. Therefore, the preceding studies do not consider topic granularity.

Another approach involves producing reliable models by integrating multiple path models and extracting frequent elements in a path model [16]. This method fails to restrict the number of latent variables. Therefore, an identification problem can arise in which the coefficients of a path model cannot be estimated.

Aside from hLDA, a method for extracting a hierarchical topic structure that combines the biterm topic model (BTM) [17] and BRTs is available [18]. The current study extracts topic by using BTM and constructs a hierarchical structure by employing BRTs. Moreover, this study uses simBRT for considering topic similarity. This study conducts a time series analysis on the basis of the constructed hierarchical structure. This method can develop a hierarchical topic structure according to BRTs. 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.

For topic labeling, Nalasco et al. proposed a technique of automatic labeling using a new candidate selection algorithm and three scoring methods [19]. Bhatia et al. proposed a neural embedding approach that involves automatic topic labeling using Wikipedia article titles [20]. Mao et al. proposed auto labeling for a hierarchical topic structure [21]. This approach employed original topic terms as candidate labels and calculated the evaluation scores with importance provided to candidate labels in documents related or not to the topic and to the appearance rate of a candidate label in topics. With this approach, higher-hierarchy topics must have bags of words.

Several methods that use a knowledge base are also available for topic labeling. Xiangyan et al. proposed conceptual labeling (CL) which aims at generating a minimum set of conceptual labels that best summarize a bag of words using Probase [22]. Haiyun et al. proposed HCL that extended CL [8]. HCL uses MCG as knowledge base. HCL effectively deletes noise words from a bag of words and hierarchically labels a clean bag of words using IsA relations in MCG and the idea of BRTs.

In the present study, causal analysis on the basis of a model constructed using simBRT is conducted to consider topic granularity. Each document should have many words and be characterized for SEM. Therefore, LDA is used instead of BTM because the document contains numerous words. Moreover, we utilize HCL as it can comprehensively generate a topic. Furthermore, the number of labels is automatically decided and the correspondence between a word and its concept is represented.

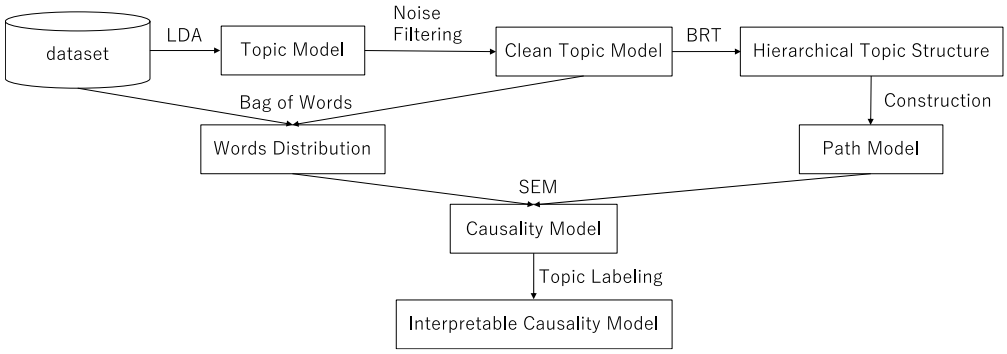

Figure 3. Process of Model Construction

3. Construction Model and Topic Labeling

In this study, analysis is performed in accordance with the process shown in Figure 3. First, a topic is extracted employing the topic model. The noise of the topic is deleted by noise filtering. Then, the topic is represented in the hierarchical topic structure by using BRTs. Next, causal analysis is conducted via SEM according to the model constructed using BRTs. Lastly, HCL adds labels to the causal model to interpret the topic.

3.1. 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 probability $P(D|T)$ that represents the likelihood of data D given tree T. In this study, topics are used as data D. All topics $C = \{t_1, t_2, \dots, t_K\}$ extracted using LDA are the leaves. Each topic t_K is regarded as an individual tree. First, each topic is considered an individual tree $T_i = \{t_i\}$. BRT is repeated to combine two trees that are selected for formulating a new tree T_m according to three basic operations (join, absorb, and collapse). Figure 4 shows the three basic operations.

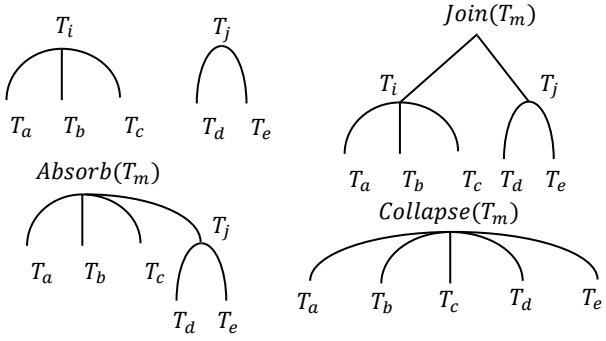

Figure 4. Three merging operations

(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, for T_i in Figure 4, $ch(T_i)$ is $\{T_a, T_b, T_c\}$. The "join" operation is the traditional one in a binary tree. Meanwhile, the "absorb" and "collapse" operations cater to a multi-branch tree. The three operations are conducted in all combinations of trees, and the combination with the maximum probability ratio is selected as follows:

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

where $C_m = T_i \cup T_j$ are 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:

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

where π_{T_m} is the prior probability that all the topics in T_m are maintained in the same partition 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 the coarse partitions of the data as opposed to fine partitions. $f(C_m)$ of (2) is the marginal probability of C_m and can be modeled by the Dirichlet compound multinomial model [24] distribution which 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 the keyword j included in the topic i, V is the total number of the vocabulary, and $\alpha = (\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(j)})$ is the hyperparameter. 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 employed as the data of a tree. Therefore, topic similarity should be considered. SimBRT takes into account the similarity of topic distribution in (1). Topics are distributed over vocabularies. Thus, the Kullback–Leibler (KL) divergence can be employed to measure the similarity between two topics. The similarity of topics z_i and z_j is defined as follows:

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

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

where $KLD(z_i||z_j)$ is the KL divergence between topics z_i and z_j. To obtain topic similarity in tree construction, simBRT defines the weighted topic distribution in each operation.

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

Absorb: $WT =$

$$\frac{avg(C_i)p(C_i|T_i)+\sum_{T_a\in ch(T_j)} avg(C_{T_a})p(C_{T_a}|T_a)}{p(C_i|T_i)+\sum_{T_a\in ch(T_j)} p(C_{T_a}|T_a)} \quad (9)$$

Collapse: $WT =$

$$\frac{\sum_{T_a\in ch(T_i)} avg(C_{T_a})p(C_{T_a}|T_a)+\sum_{T_b\in ch(T_j)} avg(C_{T_b})p(C_{T_b}|T_b)}{\sum_{T_a\in ch(T_i)} p(C_{T_a}|T_a)+\sum_{T_b\in ch(T_j)} p(C_{T_b}|T_b)} \quad (10)$$

In particular, the final merged topic distribution under T_m is $avg(C_m)$, which is simply calculated on the basis of the average. Then, the topic similarity between the weighted topic WT and the final merged topic $avg(C_m)$ is calculated by (6) and added to the primitive function in (1). Accordingly, simBRT can construct a rose tree that considers topic similarity.

3.2. Conceptual Labeling

HCL uses MCG as a knowledge base for conceptual labeling. MCG was created by data-driven approaches with a very extensive scale. That is, MCG contains 5,376,526 unique concepts which are the glue that holds our mental world together, 12,501,527 unique entities which are instances of the concepts, and 85,101,174 IsA relations between entities and concepts. Figure 5 expresses the conceptual labeling using HCL. In upper side of figure 5, assume that they are four words "beef", "chicken", "carrot" and "lettuce", "beef" and "chicken" are conceptualized by "meat", and "carrot" and "lettuce" are conceptualized by "vegetable". Furthermore, "meat" that has "beef" and "chicken" and "vegetable" that has "carrot" and "lettuce" are conceptualized by "food". In bottom of figure 5, assume that they are five words "soccer", "golf", "cricket", "rose" and "flower", "soccer", "golf" and "cricket" are conceptualized by "sport", and "rose" and "flower" are conceptualized by "flower". Here, these concepts do not merge anymore because "sport" and "flower" have no relationship.

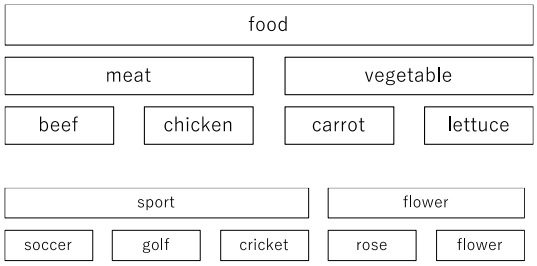

Figure 5. Conceptual labeling using HCL

To measure the semantic relevance between an entity that is instance and a concept, we introduce the typicality score. Typicality is defined as

$$p(e|c) = \frac{n(c,e)}{\sum_{e_j} n(c,e_i)} \quad p(c|e) = \frac{n(c,e)}{\sum_{c_i} n(c_i,e)}, \quad (11)$$

where e is an entity, c is a concept, and $n(c,e)$ is the frequency of c and e occurring in a syntactic pattern for an IsA relation.

In addition to typicality, we must also define the probability of a concept or an entity. That probability is defined as

$$p(c) = \frac{\sum_{e_i} n(c,e_i)}{\sum_{(c_j,e_i)} n(c_j,e_i)} \quad p(e) = \frac{\sum_{c_j} n(c_j,e)}{\sum_{(c_j,e_i)} n(c_j,e_i)}. \quad (12)$$

The bag of words generated by many scenarios usually contains noise. For example, each document in topic modeling is modeled as a probability distribution over topics, and each topic is represented by a probability distribution over words. However, some words may be semantically unrelated to the corresponding topic, thereby disturbing us to better understand the topics as well as the documents. These words are, in fact, noise and should be filtered out.

As an approach to noise filtering, let D be the input bag of words, and d_i (d_j) be the i-th (j-th) entity in D. We take $p(c|d_i,d_j)$ to measure how well the concept c conceptualizes the semantics of two entities d_i, d_j. On the assumption that all entities in D are independent to each other, $p(c|d_i,d_j)$ is computed by Bayesian theorem as follows:

$$p(c|d_i,d_j) = \frac{p(d_i,d_j|c)p(c)}{p(d_i,d_j)} = \frac{p(d_i|c)p(d_j|c)p(c)}{p(d_i)p(d_j)}. \quad (13)$$

Here, when $c = d_i$ and $c \neq d_j$, $(d_i|c)\,p(d_j|c)$ is computed as $p(d_j|c)^2$. For example, when $c = pet$, $d_i = pet$, and $d_j = dog$, $p(pet|pet)\,p(dog|pet)$ is computed as $p(dog|pet)^2$. Furthermore, assuming that all the entities in D have equal prior probabilities, i.e., $p(d_k) = \tilde{p}(\forall d_k \in D)$, then

$$p(c|d_i,d_j) = \frac{1}{\tilde{p}^2} p(d_i|c)p(d_j|c)p(c). \quad (14)$$

The prior probability $p(c)$ measures the popularity of c. That is, popular concepts will have large probabilities. Intuitively, a larger $p(c|d_i,d_j)$ indicates that d_i and d_j can be well summarized by c. Thus, d_i and d_j have strong semantic relevance. $p(d_k|c)$ and $p(c)$ are estimated using the knowledge in MCG.

Let C'_i and C'_j be the concept sets of d_i and d_j in MCG, respectively. $C'_{i,j} = C'_i \cap C'_j$ denotes the shared concept set of d_i and d_j. We describe the denoising algorithm as follows. Given a word $d_i \in D$, for any other word $d_j \in D(d_j \neq d_i)$, d_i is treated as noise if we

cannot find an appropriate concept in $C'_{i,j}$ to conceptualize d_i and d_j. That is,

$$max_{d_j \in D, c \in C'_{i,j}} p(c|d_i, d_j) < \delta, \quad (15)$$

where δ is a pre-given threshold. Considering that $1/\tilde{p}^2$ is equal for all the words in D, we take the following simplified form to filter out the noise in D:

$$max_{d_j \in D, c \in C'_{i,j}} p(d_i|c)p(d_j|c)p(c) < \delta. \quad (16)$$

For each d_i in D, if equation (8) is established, then we delete it from D.

Next, we describe HCL using a filtered clean bag of words. HCL is conducted on the basis of BRT described in Section 3.1. Instead of (1) and (2), this section uses the following formulas:

$$L(T_m) = \frac{p(D_m|T_m)}{p(D_i|T_i)p(D_j|T_j)}, \quad (17)$$

$$p(D_m|T_m) = \pi_{T_m} f(D_m) + (1 - \pi_{T_m}) \prod_{T_k \in ch(T_m)} p(D_k|T_k), \quad (18)$$

where $L(T_m)$ is the criterion for maximizing the likelihood ratio and $f(D_m)$ is the marginal probability of data D_m.

$f(D_m)$ in BRT denotes the probability that all data points in D_m are generated by the same probabilistic model. D_m can also be considered to be generated by concepts in MCG. That is, any concept c in the shared concept set of D_m is a probabilistic model that could generate D_m with a certain probability described as $p(D_m|c)$.

Specifically, let C'_m be the shared concept set of all the entities in D_m. Then, D_m can be generated by any concept in C'_m. We argue that each concept $c_i \in C'_m$ is selected with a certain probability which is proportional to $p(c_i)$. Thus, we define the selection probability of c_i as follows:

$$p_s(c_i) = \frac{p(c_i)}{\sum_{c \in C'_m} p(c)} \quad (19)$$

Then, $f(D_m)$ is computed as

$$f(D_m) = \sum_{c \in C'_m} p_s(c) p(D_m|c). \quad (20)$$

On the basis of the independence assumption, $p(D_m|c)$ is calculated as

$$p(D_m|c) = \prod_{d_i \in D_m} p(d_i|c). \quad (21)$$

A larger $|C'_m|$ indicates that the words in D_m are more similar in semantics, thereby deriving a larger $f(D_m)$. $C'_m = 0$ indicates that no shared concept exists for D_m. Consequently, $f(D_m) = 0$ and implies that the words in D_m cannot be generated by a single model and should be partitioned into multiple clusters.

The original BRT algorithm will eventually generate a single tree. That is, all data in the dataset will eventually be put into one cluster. By contrast, the cluster operation should be stopped when no appropriate label exists to well conceptualize the current cluster. We introduce a threshold β and stop clustering when $L(T_m) < \beta$.

Finally, we select an appropriate conceptual label to well conceptualize each cluster D_m. The following criterion is used to select the most appropriate conceptual label:

$$c_m^* = argmax_{c \in C'_m} p(c|D_m) = argmax_{c \in C'_m} p(D_m|c)p(c). \quad (22)$$

Here, we described how to generate a topic label for a hierarchical topic structure. When several trees are merged, a node is regarded as a large topic that comprises several topics. Therefore, we generate new conceptual labels from the conceptual labels of the trees merged as shown in Figure 6 In other words, the conceptual labels of each tree are regarded as bags of words that comprise a large topic.

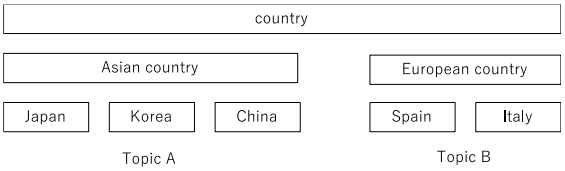

Figure 6. HCL for hierarchical structure

3.3. Construction of the Path Model

Topics that cannot be observed directly are considered latent variables that function as the correspondence between SEM and a topic model. Keywords that comprise a topic constitute the observation variables because these terms 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 is drawn from the topics 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. Therefore, when several trees are merged, a node shown as a factor is regarded as a large topic that includes several topics. Therefore, this node is regarded as a topic and applied as a latent variable. The

Table 1. Dataset and Result of Various Criteria for each method

Dataset	Number	Method	GFI	AGFI	RMSEA
hotel	8201	our method	0.9687	0.9633	0.03028
		hLDA	0.9534	0.9486	0.02976
airport	13466	our method	0.9665	0.9600	0.03472
		hLDA	0.9426	0.9322	0.04179
instrument	8538	our method	0.9395	0.9256	0.04820
		hLDA	0.9082	0.8967	0.04796
e-commerce	19599	our method	0.9833	0.9786	0.02849
		hLDA	0.9716	0.9680	0.02562

paths between topics are drawn from the upper topics to the lower ones on the basis of the idea that large topics generate small ones. In addition, a path is drawn from the top topic to rate the numerical evaluation of the review data, thereby clarifying the relation between topic structure and actual numeric data. Therefore, a dataset must have text data and a rating evaluation expressed by a numerical value for the application of this method. Moreover, the number of review documents and the length of text data must be of the correct magnitude for LDA application.

4. Experiment

This experiment aims to confirm the feasibility of the proposed method by constructing the model described in Section 3 and the interpretability of the causal model by using HCL. This section also considers the experiment results.

4.1. Dataset, Criteria, and Hyperparameters

In this experiment, the dataset should ideally have as many review data as possible to apply the topic model. Moreover, the text of one review datum must

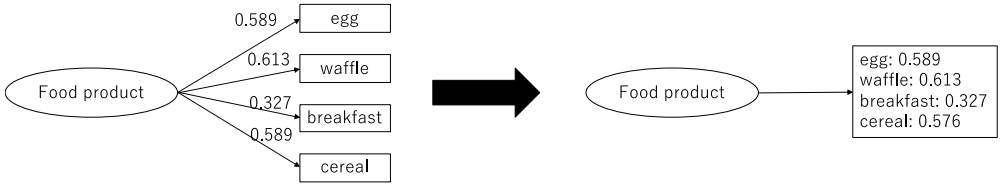

Figure 7. Expression of a path from the latent to the observed variable

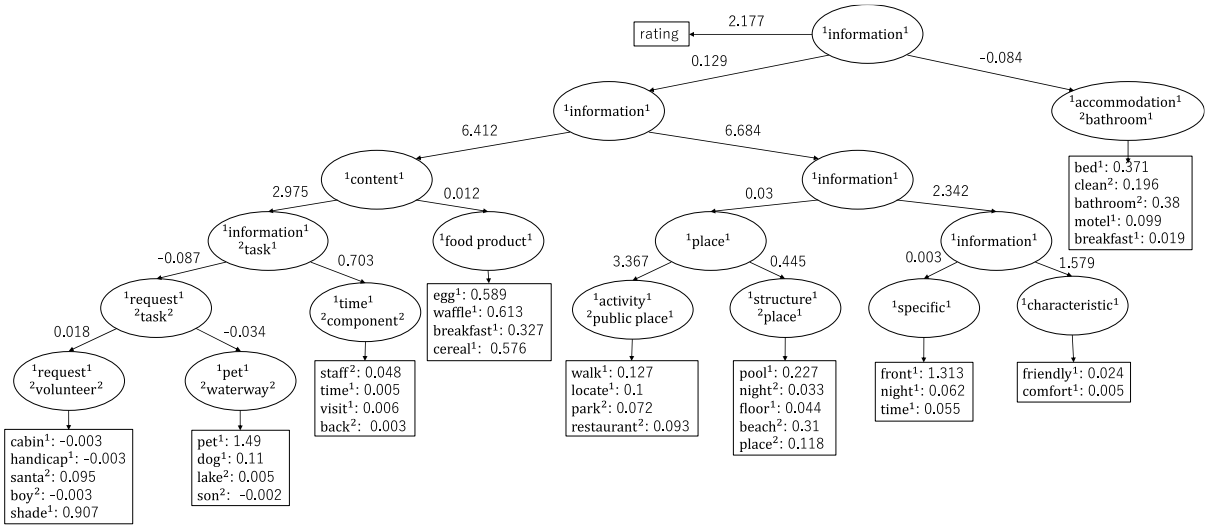

Figure 8. Analysis result of a hotel using BRT and HCL

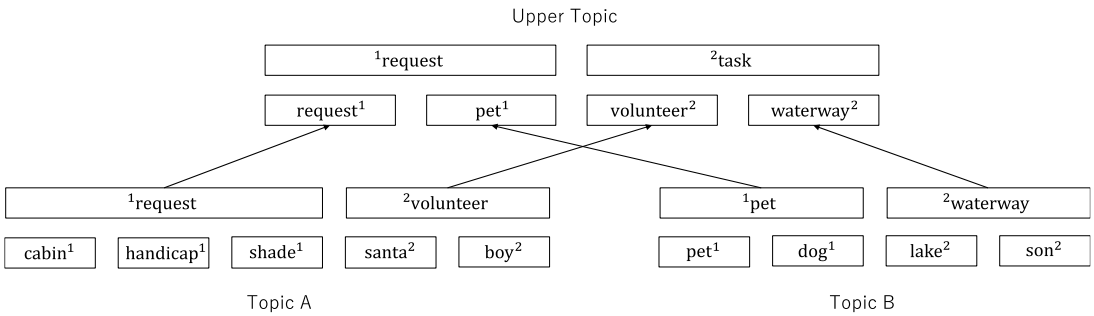

Figure 9. Example of HCL

include numerous words to characterize the statistical data according to the concept of a bag of words. The user reviews in the datasets published online by Kaggle, Github, and Amazon are employed. In particular, reviews of airports and hotels (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 utilized to ensure that the topics and the appearance frequency of the described feature words are included in each document. 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 the criteria to evaluate the result. The 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 a value closer to 1 denotes a better model. A value of 0.9 or higher is desirable. The GFI is unconditionally improved in fitness as a model's degree of freedom decreases. The 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 closer to 1 indicates a better resultant model. If the model is not complex, then the values of GFI and AGFI are close to each other. RMSEA is an index that expresses the difference between model distribution and actual distribution. A fit value of 0.05 or less is satisfactory and is otherwise if the value is 0.1 or higher.

As hyperparameters, γ values are used by BRT, α values are employed by simBRT, and δ and β are utilized by HCL. The number of topics and words comprise a topic. This experiment uses $\gamma = 0.1$, $\alpha = 0.3$, $\delta = 2.5e - 10$, and $\beta = 0.9 + 10e - 15$. The number of bottom topics is 10, and the number of words that comprises a topic is 5. In this experiment, we do not use concepts such as "noun," "adjective," "factor," and "topic" in MCG that cover many words but are not explicitly semantic.

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

4.2. Results

Table 1 presents the calculation results of the evaluation indexes for each data and each method. All the models in Table 1 have a GFI and AGFI of over 0.9. Moreover, the models have RMSEA values of less than 0.05. Further, our method results of all models, except for the RMSEA of hotel, instrument and e-commerce, have higher values than the hLDA results. In the hLDA model, 10 topics were obtained as topics of bottom layer, 5 keywords that comprise the topic were selected and keywords that comprise the topic at the upper layer were deleted to achieve the same situation as that in the BRT model.

For example, Figure 8 shows the analysis result model of the hotel dataset. A topic {odor, cash, Canada, wyndham, lay} is deleted because all values of δ between each word falls below the threshold. Therefore, the causal model is constructed by nine topics. The contents of each topic and the impact on all the topics can be assumed by focusing on the value of the arrow. The causal relation between keywords that comprise a topic is presented similarly to the depiction in Figure 7. Figure 9 describes a part of HCL in hotel analysis according to BRT. In Figures 8 and 9, the left number of concept cover the input the bag of words that has same number on the right.

4.3. Computation time

The computation time is also important factor for evaluating our method. Our method combines LDA and hLDA. That is, the computation time of our method is sum of computation time of LDA and that of simBRT. The computation time of simBRT is important compared to traditional topic models such as LDA and hLDA as new contents. In simBRT, as the number of topics increases, the number of tree combinations that is

Table 2. The Computation time of simBRT

The number of topics	10	20	30
Computation time	0.3388s	3.5020s	14.3746s

considered for finding the combination which has the maximum probability increases significantly. Firstly, to combine two topics from n topics, the number of the combination is considered is ${}_nC_2$ to combine two topics from n topics. Further, to finally merge into one tree, the number of the combination that is considered is $\sum_{k=2}^{n} {}_kC_2 = \frac{1}{6}(n^3 - n)$. Therefore, as the number of topics increases, the computation time increases significantly. Here, we calculated the real computation time of simBRT for each number of topics (10, 20, 30) and found how much computation time would be required in additional to the computation time of LDA. The computation time is the averaged time of 10 times running on hotel data, and the calculations were conducted in a computer with Intel Core i5 CPU@1.8GHz and 8 GB RAM. We used the result of LDA before filtering by HCL. Table 2 shows results of the computation time. From Table 2, in an analysis such as this study, we can understand that the computation time of simBRT is very short and the computation time is mostly determined by the difference between LDA and hLDA.

5. Conclusion

In this study, a hierarchical topic structure was represented by BRTs on the basis of a topic extracted from text data. Then, the noise of each topic is deleted by a filtering method to generate a clean topic. Moreover, a path model of SEM was constructed on the basis of this hierarchical topic structure and causal analysis was conducted. Lastly, a causal model is added conceptual label to interpret the hierarchical topic structure by HCL. The value of our proposed method was demonstrated by the result of an experiment that employed the user review of 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. Therefore, topics function as evaluation factors from the text of review data, and a hierarchical topic structure is constructed using simBRT on the basis of this topic. This method can consider topic granularity and construct a hierarchical topic structure with different hierarchies for each topic. Conversely, a topic is a bag of words without explicit semantics. Thus, interpreting a hierarchical topic structure is difficult. To solve this problem, topic labeling is conducted by HCL that uses MCG. The causal model facilitates hierarchical topic structure interpretation. In addition, clean topics are generated by noise filtering that employs MCG.

In the experiment, several services and a product were analyzed to confirm the feasibility of the proposed method. For each criterion, satisfactory values were found for the entire dataset. In addition, the services and a product can be visually and quantitatively evaluated using the proposed model as shown in Figure 8.

In future work, another topic model can be used to improve accuracy. Several conceptual labels that are difficult to interpret are generated by HCL, as shown by "component" in Figure 8. Moreover, several conceptual labels that do not have relevance to the analysis target (such as a hotel) are generated, as shown by "characteristic" in Figure 8. Therefore, more appropriate conceptual labels may be generated by considering the relationship analysis target and concept. Specifically, we formulate a concept label by considering the similarity between a concept that includes words of the target of analysis (such as "hotel service") and bag of words.

6. 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, 2004, vol. 13, No. 01, pp. 157-169.

[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, May 2014, pp. 216-225.

[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, 2015, vol. 135, Issue 6, pp. 602-610.

[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, December 2003, pp. 17-24.

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

[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, July 2010, pp. 65-72.

[8] H. Jiang, Y. Xiao and W. Wang, "Explaining a bag of words with hierarchical conceptual labels", World Wide Web, 2020, vol. 23, pp. 1693-1713.

[9] L. Ji, Y. Wang, B. Shi and D. Zhang, "Microsoft Concept Graph: Mining Semantic Concepts for Short Text

Understanding", Data Intelligence, 2019, vol. 1, Issue 3, pp. 238-270.
[10] C. M. Stein, N. J. Morris and N. L. Nock, "Structural Equation Modeling", Statistical Huma Genetics: Methods and Protocols, Methods in Molecular Biology, January 2012, vol. 850, pp. 495-512.
[11] D. M. Blei, "Probabilistic Topic Models", Communications of the ACM, April 2012, vol. 55, No. 4, pp. 77-84.
[12] 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, 1990, vol. 41, Issue 6, pp. 391-407.
[13] 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, 2013, pp. 222-230.
[14] R. Saga and R. Kunimoto, "LDA-based Path Model Construction Process for Strucutre Equation Modeling", Artificial Life Robotics, 2016, vol. 21, Issue 2, pp. 155-159.
[15] 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, November 2019, pp. 15-20.
[16] R. Kunimoto and R. Saga, "Path model integration method for strucutural equation modeling by OR and probability concepts" 2013 IEEE Internaional Conference on Systems, 2013, Man, and Cybernetics, pp. 3073-3078.
[17] 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, May 2013, pp. 1445-1456.
[18] J. Zhu, X. Li, M. Peng, J. Huang, T. Qian, J. Huang, T. Qian, J. Huang, J. Liu, R. Hong and P. Liu, "Coherent Topic Hierarchy: A Strategy for Topic Evolutionary Analysis on Microblog Feeds", Proceedings of 16th International Conference on Web-Age Information Management, June 2015, pp. 70-82.
[19] D. Nolasco and J. Oliveira, "Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data", Proceeding of the 49th Hawaii International Conference on System Science (HICSS), 2016, pp. 358-367.
[20] S. Bhatia, J. H. Lau and T. Baldwin, "Automatic labeling of topics with neural embeddings", Proceedings of the 26th International Conference on Computing Linguistics: Human Language Technologies, 2011, pp. 1536-1545.
[21] X. L. Mao, Z. Y. Ming, Z. J. Zha, T. S. Chua, H. Yan and X. Li, "Automatic labeling hierarchical topics", ACM International Conference Proceeding Series, 2012, pp. 2383-2386.
[22] X. Sun, Y. Xiao, H. Wang and W. Wang, "On Conceptual Labeling of a Bag of Words", Proceeding of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), 2015, pp. 1326-1332
[23] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering", ICML'05: Proceedings of the 22nd international conference on Machine learning, August 2005, pp. 297-304.
[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, August 2005, pp. 545-552.
[25] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, May 2010, pp. 45-50.
[26] R. Ihaka and R. C. Gentleman, "The R Project for Statistical Computing", https://www.r-project.org, [retrieved: January, 2020]