Text Mining to Support Consulting Services for Client Company State Recognition

Watanabe, Ruriko
Fujii, Nobutada
Kokuryo, Daisuke
Kaihara, Toshiya
Abe, Yoichi

(Citation)
International Journal of Automation Technology,14(5):779-790

(Issue Date)
2020-09-05

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90009504

https://doi.org/10.20965/ijat.2020.p0779

Paper:

Text Mining to Support Consulting Services for Client Company State Recognition

Ruriko Watanabe[*,†], **Nobutada Fujii**[*], **Daisuke Kokuryo**[*], **Toshiya Kaihara**[*], **and Yoichi Abe**[**]

[*]Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan
[†]Corresponding author, E-mail: watanabe@kaede.cs.kobe-u.ac.jp
[**]F&M Co., Ltd., Suita, Japan

[Received April 3, 2020; accepted July 20, 2020]

This study was conducted to devise a method for supporting consulting service companies in their response to client demands irrespective of the expertise of consultants. With emphasis on revitalization of small and medium-sized enterprises, the importance of support systems for consulting services to serve them is increasing. Those systems must support solutions to difficulties that must be addressed by enterprises. Consulting companies can respond to widely various management consultations. Nevertheless, because the consultation contents are highly specialized, service proposals and problem detection depend on the experience and intuition of the consultant. Often, stable service cannot be provided. A support system must provide stable services independent of the ability of consultants. In this study, analyzing customer information describing the contents of consultation with client companies is the first step in constructing a support system that can predict future problems. Text data such as a consultant's visit history, consultation contents by e-mail, and contents of call centers are used for analyses because the contents can explain current problems. They might also indicate future problems. This report describes a method to analyze text data using text mining. The target problem is fraud, which includes uncertainty: cases in which it is not clear whether a fraud problem has occurred with the company. To address uncertainty, a method of using logistic regression models is proposed to represent inferred values as probabilities, rather than as binary discriminated data, because the possibility exists that some misidentified companies might have some difficulty. As described herein, computer experiments are conducted to verify the effectiveness of the proposed method and to compare consultants' forecasted and achieved results. Results of a verification experiment are presented in the following. First, the proposed method is applicable to problems including uncertainties. Secondly, the possibility exists of discovering companies with a fraud problem of which they are unaware.

Keywords: correspondence analysis, DEA discriminant analysis, text mining

1. Introduction

1.1. Current Status of Small and Medium-Sized Enterprises (SMEs) and Consulting Services

In recent years, with emphasis on the activation of small and medium-sized enterprises (SMEs) [1], support systems including consulting for SMEs are regarded as important [2]. They support SMEs to solve problems that are difficult to address in their company. Consultant companies have difficulties that render them as not always sufficient in supporting leading and specialty field [3]. Moreover, service suggestions and client companies' problem detection depend on the experience and intuition of each consultant [4]. A support system must be devised to provide services with stable quality irrespective of consultant ability.

Labor-intensive industries have a large proportion of tasks performed by a human labor force. An important difficulty is that labor-intensive industries have low labor productivity; support methods are established in labor-intensive industries in many diverse fields. Feng Duan proposed a multimodal assembly support system for information and physical aspects of assembly workers while satisfying the actual manufacturing requirements. The system improves the cell production system efficiency, mainly depending on human operators' work performance [5]. Toshihiro Kamma specifically examined animation production efficiency and studied task computerization to improve video work quality [6].

Among labor-intensive services, this study specifically examines consulting services, which are knowledge-intensive services. The quality of services depends greatly on the consultants' intuition. This study was conducted to assess consulting services independently of the expertise of a consultant.

Figure 1 portrays the consulting service workflow. Consultants conduct state recognition, such as what difficulties clients must confront, from contents of surveys with the client company, and judge appropriate strate-

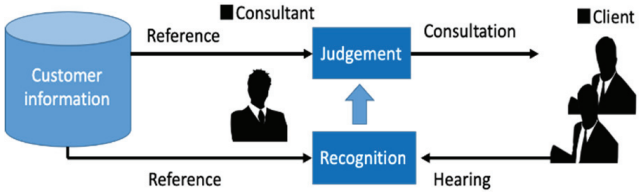


Fig. 1. Workflow of consulting service.

gies based on that consultation. Because the consultant's ability is regarded as particularly dependent on the consultant's recognition and judgment, these are targeted for support in this study. Nevertheless, proper judgement cannot be achieved unless the recognition is correct. Therefore, for constructing a support system particularly supporting consultant's recognition system, some method must be devised to use text mining to predict future problems at client companies using data from communications and consultations received from customers.

1.2. Applying Text Mining to Consulting Services

Text data analysis has attracted attention because of computer processing capability improvements. Analyses of large amounts of data can be done with inexpensive memory. Because the contents included in texts are not prescribed in advance, the possibility exists that the text data might reveal some irregularity that was overlooked in the past. The need for text utilization has increased in recent years [7]. Text mining is used in widely diverse fields [8] such as economics [9], business administration [10], pedagogy [11], nursing science [12], and philosophy [13].

To analyze text data scientifically and to apply it to corporate evaluation, some studies have analyzed correlation between stock prices and stock-related announcements sent by corporate managers to shareholders [14]. Other studies have analyzed long-term market trends using text information from monthly reports of financial data [15]. Because text data used in these studies were standard expressions for reporting information for external use, the data were homogeneous. The contents were limited to those related to the subjects of the respective studies.

This study analyzes text data that describe communications with client companies after accumulation by consulting companies. Because these text data are based on various records such as the contents of the consultation between the consultant and the client company and the contents of the call center, the data are heterogeneous. Because the data are not limited to company information released externally, more detailed information about the company is included. The data quality differs from the research introduced earlier. Because the consultation contents described in the target data of this research are diverse, analyses using all words and phrases appearing in the text data might include wasteful variables and high calculation costs. Therefore, one must extract phrases related to the target problem in advance. For discriminant analysis, consulting services must recognize the cause of the problem occurrence, with an attempt to create a linear discriminant that can analyze factors of inference instead of a machine learning method that has been confirmed as having high inferential accuracy for discrimination problems in recent years.

1.3. Research Objective

This study uses correspondence analysis to extract words used as variables before analyzing state recognition of client customers. To analyze state recognition, a discrimination method able to discriminate the presence or absence of a problem is used. It is not a general statistical discriminant analysis method, but a data envelopment analysis (DEA) discriminant method. Because it need not make assumptions about the distribution of the population, it is expected that DEA discriminant analysis can accommodate heterogeneous text data for which it is difficult to judge the distribution. Furthermore, because it is possible to obtain a solution merely by solving linear programming problems, it presents advantages such as being able to address large scale data. A support method combining the correspondence analysis described above and DEA discriminant analysis is proposed for this study [16].

Moreover, for application to real problems, the method has been improved in three respects: standardization, use of IDF values, and using words that appear in only one group. The improved method reduces the number of companies for which results cannot be determined. Comparison with forecast results produced by consultants suggests that the support system captures characteristics of the company for reference when the consultant predicts [17].

As described herein, the support method is applied to other problems. The target problem is more complex and uncertain: the problem can occur without knowledge of the consultant or the client company. To address uncertainty, a method using logistic regression models to represent inferred values by probabilities is proposed. Binary discrimination is discarded because the possibility exists that some misidentified companies might actually have some problem. Therefore, it is necessary to identify companies to be supported by consultants. The effectiveness of prediction for a problem that includes uncertainty cannot be verified solely by the discrimination rate. For this study, grouping is done to create text data describing communication between the client and the consultant for use in discrimination, but there might be companies in which problems actually occurred even though such companies were classified as having no problems. Therefore, additional interviews with client companies and estimation of problems by consultants are conducted by comparing those results with inferences made using the proposed method. Then the effectiveness of this method is verified for problems of uncertainty.

Table 1. Problems of which consultants must be aware.

	The problem is certain.	The problem has uncertainty.
Problem structure	Problems the client recognizes. Problems the consultant recognizes later.	Problems within the company the client might not be aware of.
Example	Cancellation issues; corporate performance.	Embezzlement and personnel issues.

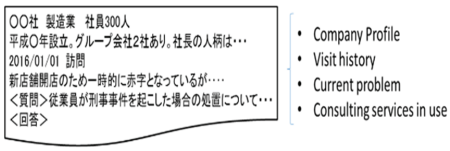


Fig. 2. Text data.

2. Proposed Method

2.1. Research Subject

By analyzing text data accumulated by consulting companies, possible difficulties faced by client companies were inferred using the method assessed for this study. Data used for these analyses were recorded using various methods such as contents of interaction during a consultant's visit to the client company, e-mail questions sent from the client company, and call center correspondence. Examples are presented below. **Table 1** shows problems of which consultants must be aware.

As shown in **Fig. 2**, text data consist of details of the company profile, visit history, problems held by client companies, and consulting services in use. Apparently some texts describe contents that are expected to be a trigger of future problems. For this study, these text data were analyzed using a computer to support consulting services. This proposed method supports consultants who must address problems of widely various fields. Using it with accumulated text data, consultants with little experience can be aware of difficulties that can occur in client companies by inferring difficulties occurring in client companies.

Consultants must be aware of diverse problems of many types at client companies, but they are broadly classifiable into two types.

Issues that are certain have been emphasized: whether client companies will continue or cancel consulting services. This paper presents specific examination of a more complex problem: fraud. The system can infer whether fraud is occurring within the client company. Fraud is an act of embezzlement by taking advantage of a position, such as embezzling a company's resources or sales revenues, or inflating a billing for payment. **Fig. 3** presents the cancellation problem structure. **Fig. 4** presents the fraud problem structure.

In the cancellation problem, the correct answer data for creating the prediction formula coincides with whether or not to cancel a consulting service. However, in the case of fraud, the presence or absence of the description of the fraud problem in the text data is used as correct answer

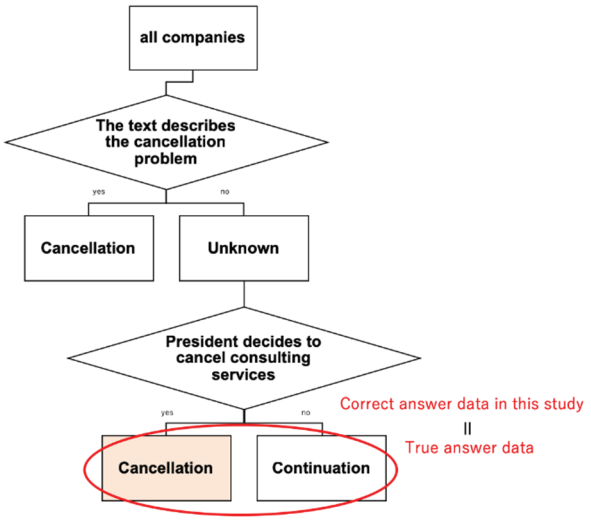


Fig. 3. Cancellation problem.

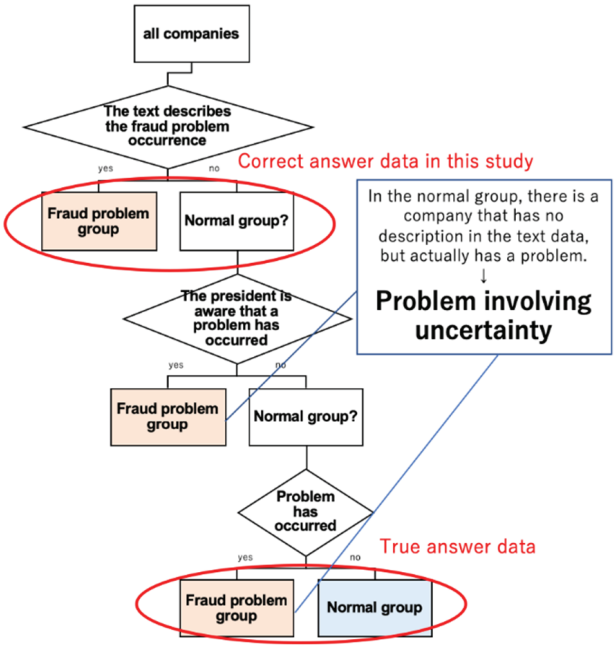
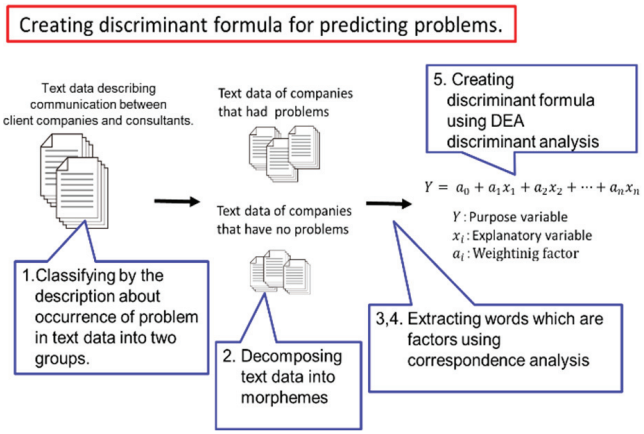


Fig. 4. Fraud problem with uncertainty.

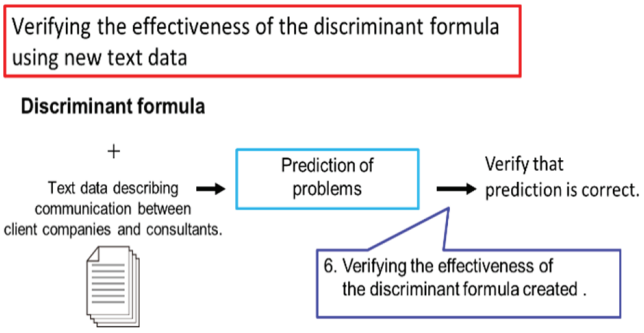
data to create a prediction formula. For some companies, a fraud problem has occurred even when the text data do not describe a fraud problem; therefore, the problem structure includes uncertainty. Fraud problem is an uncertain problem because there is no one-to-one correspondence between occurrences and discovered. As described herein, the proposed method is applied to this problem including uncertainty. Then its effectiveness is verified. Furthermore, by using the problem's occurrence probability as a predicted value, even a problem that includes uncertainty can be addressed.

2.2. Overview of the Proposed Method

This section outlines the proposed method. First, text data are classified by the occurrence of problem. A dis-



(a) First stage



(b) Second stage

Fig. 5. Overview of the research.

criminant formula is created using text mining. Phrases are extracted as factors by extracting correspondences for each group from among many words and phrases and by using the extracted phrases as variables of DEA discriminant analysis to create a discriminant. To verify the effectiveness of the obtained discriminant formula, newly categorized text data are used to judge the problem presence or absence. An overview of the proposed method is presented in **Fig. 5**. In the first stage depicted in **Fig. 5(a)**, a discriminant is created for predicting the occurrence of a problem. In the second stage presented in **Fig. 5(b)**, prediction is performed using new text data to verify the discriminant effectiveness.

The proposed method is executed in the following six steps. These steps are shown in **Fig. 6**, with detailed description of the respective steps in later sections.

The flow of the method examined for this study includes data shaping, feature extraction, and creation of discriminants. As described herein, we propose a method to calculate the probability of problem occurrence using a logistic curve to target problems including uncertainty.

2.3. Text Data Classification

Text data of client communications accumulated at a consulting company are used for this study. Text data are categorized according to the presence or absence of problem occurrence considering the time series of fraud problem detection. To predict the occurrence of future problems, text data are deleted with direct contents related to fraud problems. In addition, past text data are used for a certain period of time since the problem occurred because the proposed method is designed to predict problems as early as possible by avoiding the use of direct descriptions of problems in the text data.

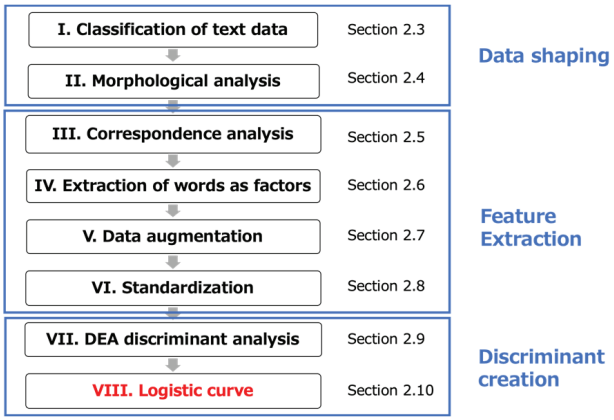


Fig. 6. Flow of the proposed method.

2.4. Morphological Analysis

A morpheme is the smallest character string that retains some meaning: it becomes meaningless when it is decomposed further. Decomposing sentences into morphemes and specifying their respective parts of speech is called morphological analysis. This study uses the MeCab morphological analyzer developed by Kudo [18]. MeCab, a general-purpose program, is independent of dictionaries and text data. It is faster than other analyzers. Words and phrases related to many items, regarded as noise, are omitted from the analyzed morphemes.

Both groups share many common occurrences of morphemes irrespective of whether a problem has occurred. The following words are deleted because they introduce noise into phrase extraction in correspondence analysis.

- Name of consulting service:
  Aptitude diagnosis service.
- Fixed phrase:
  FAX sent.
- Words making no sense:
  Month, year, case.

2.5. Correspondence Analysis

Correspondence analysis, proposed in the 1960s by Benzeccri, compresses information contained in rows and columns of data tables into a few components [19]. For this study, words are included in line items (sample). Company names are included in column items (category). Two-dimensional data of the appearance count t_{ij} of word j of company i are targeted. **Table 2** presents an example of two-dimensional data.

Table 2. Two-dimensional data for correspondence analysis.

	Word	AAA	BBB	⋯	KKK
Company		a_1	a_2	⋯	a_K
A Co.	b_1	t_{11}	t_{12}	⋯	t_{1K}
B Co.	b_2	t_{21}	t_{22}	⋯	t_{2K}
⋮	⋮	⋮	⋮	⋱	⋮
N Co.	b_N	t_{N1}	t_{N2}	⋯	t_{NK}

Table 3. Variables used for correspondence analysis.

i $(i = 1,2,\ldots,K)$	Extracted word
j $(j = 1,2,\ldots,N)$	Company
t_{ij}	Number of occurrences of word i in text data of company j
a_i	Weight used as coefficient of word i
b_j	Weight used as company j coefficient

Table 3 represents variables using correspondence analysis.

Basic statistics are calculated as shown below.

$$\text{Average: } \bar{a} = \frac{\sum_{i=1}^{K} a_i}{K}, \quad \ldots\ldots\ldots\ldots (1)$$

$$\text{Dispersion: } V_a = \frac{\sum_{i=1}^{K} \left\{ \left(\sum_{j=1}^{N} t_{ij} \right) (a_i - \bar{a})^2 \right\}}{KN-1}, \quad \ldots (2)$$

$$\text{Covariance: } V_{ab} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} t_{ij} (a_i - \bar{a})^2 \left(b_i - \bar{b}\right)^2}{KN-1}, \quad \ldots (3)$$

$$\text{Correlation coefficient: } R = \frac{V_{ab}}{\sqrt{V_a}\sqrt{V_b}}. \quad \ldots\ldots (4)$$

With correspondence analysis, sample scores and category scores that maximize the correlation coefficient are obtainable under the condition of $\bar{a}, \bar{b} = 0$, $V_a, V_b = 1$. Here, Eq. (5) can be established using the Lagrangian multiplier method.

$$G = V_{ab} - \frac{\lambda}{2}(V_a - 1) - \frac{\mu}{2}(V_b - 1). \quad \ldots\ldots (5)$$

When Eq. (5) is partially differentiated with the sample score and the category score, the following simultaneous equations are obtained.

$$\begin{cases} \dfrac{\partial G}{\partial a_1} = 0, \\ \quad \vdots \\ \dfrac{\partial G}{\partial a_N} = 0, \end{cases} \quad \begin{cases} \dfrac{\partial G}{\partial b_1} = 0, \\ \quad \vdots \\ \dfrac{\partial G}{\partial b_N} = 0. \end{cases} \quad \ldots\ldots (6)$$

By deforming Eq. (6) with $\lambda = \mu$, the following eigenvalue problem is obtained.

$$\begin{bmatrix} X_{11} & \cdots & X_{1K} \\ \vdots & \ddots & \vdots \\ X_{K1} & \cdots & X_{KK} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix} \quad \ldots (7)$$

($X_{11}, \ldots, X_{KK}$: arbitrary number)

Eigenvalues are determined by solving Eq. (7). Sample scores and category scores are calculated for each eigenvalue. The eigenvalue obtained here is called a dimension. Using the operation above, many samples and categories are compressed to a few components.

By mapping sample scores and category scores corresponding to each axis on the scatter diagram, correspondence between these variables can be visualized. Characteristic companies and words in the target data appear as they move away from the origin of the scatter plot. General companies and words in the target data appear near the origin. For this study, correspondence analysis is conducted for each group divided by presence or absence of a detected fraud problem. Also, attention is devoted to words that are near the origin. Words which become factors of the respective groups are extracted using the method explained in the next section.

From correspondence analysis results, considering all dimensions, distance d_{iG} between word i and the origin in each group (G_1, G_2) is calculated using Eq. (8). Also, D_G represents the total number of dimensions, x_{ijG} stands for the sample score of word i in dimension j, and C_{jG} denotes the contribution of dimension j.

$$d_{iG} = \sqrt{\frac{\sum_{j=1}^{D_G} \left\{ \left(x_{ijG} * C_{jG} \right)^2 \right\}}{D_G}}. \quad \ldots\ldots (8)$$

Weighting is performed by multiplying the distance from the origin by the IDF value to extract more commonly used words. The IDF value, known as the reverse document frequency, is a high value if the word is rare between documents; it is a low value if the word appears frequently in many sentences. IDF values are calculated using Eq. (9).

$$idf_i = \log \frac{N}{N_i'}. \quad \ldots\ldots (9)$$

2.6. Extracting Words as Factors

2.6.1. Extracting Words Appearing in Both Groups

For each word, the d value is updated by the operations of Eqs. (10) and (11). For each group, words with small e values are extracted to a fixed number.

$$\text{When } G_1 < G_2, \quad \begin{cases} e_{iG_1} = d_{iG_1} + M - d_{iG_2}, \\ e_{iG_2} = \infty, \end{cases} \quad \ldots\ldots (10)$$

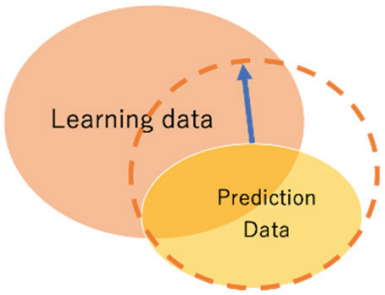


Fig. 7. Data augmentation.

$$\text{when } G_1 > G_2,$$
$$\begin{cases} e_{iG_2} = d_{iG_2} + M - d_{iG_1}, \\ e_{iG_1} = \infty. \end{cases} \quad \ldots\ldots\ldots \quad (11)$$

2.6.2. Extracting Words Appearing in Only One Group

For each group, the following heuristic rules are used to extract words as factor while also considering words and phrases that appear in only one group.

I. From words appearing in both groups, select the word with the strongest factor: the word with the highest e_{iG} value.

II. Words appearing in only one group with distance less than the selected words are extracted in order from a word with smaller distance. It ends when the set number of extracted words is reached. Exit when the set number of extraction words is reached.

III. Extract the word selected in I. Go to I if it is less than the set number of extraction words. Exit when the set number of extraction words is reached.

2.7. Data Augmentation

In this method, words used for creating a discriminant are words that overlap in the learning data group and the prediction data group. Because fewer data are used for prediction than for creation, a problem exists that many companies are indistinguishable because no word is used in the discriminant appears. Therefore, a method was applied to reduce the number of unidentifiable companies by expanding data using synonym dictionaries for phrases that appear in the data used for prediction. **Fig. 7** portrays an image of data augmentation.

WordNet, a concept dictionary incorporating the meaning and thesaurus, is used for data augmentation. **Fig. 8** presents an example of a concept dictionary.

A word has a plurality of concepts. Each concept is associated with broader words and lower words. This method does not consider the meaning of a concept in which a word appears. Therefore, words associated with all concepts are regarded as synonyms. Synonyms are regarded as having appeared the same number of times. The appearance frequency table is therefore extended.

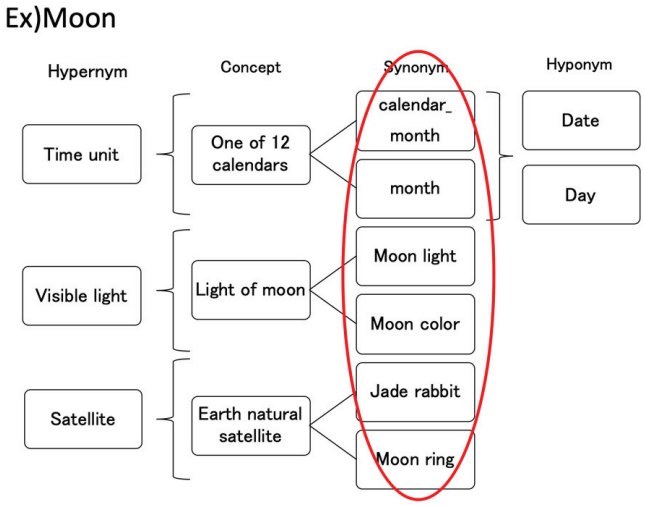


Fig. 8. Example of the word "月" (moon).

2.8. Standardization

A discriminant is created using the words extracted in Sections 2.5–2.7. Before performing discriminant analysis, standardization is applied to unify the word scale. The standardization is calculable using Eq. (12) when the average frequency of occurrence of a word is M and the standard deviation is s. The average of each word is 0. The variance is 1.

$$z = \frac{x - M}{s}. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (12)$$

Results show that, by standardization, a word having several appearances of 0 is also assigned a value. It can be determined.

2.9. DEA Discriminant Analysis

DEA discriminant analysis was first proposed by Sueyoshi [20]. In the present study, DEA discriminant analysis is adopted because the method has the following features.

- No assumption about the population distribution is necessary: text data from which it is difficult to grasp the actual distribution can be accommodated as-is.
- Simply solving a linear programming problem can lead to the answer: large-scale data can also be accommodated.
- No precondition is necessary for the sample size: problems with differences in the amounts of sentences between groups can be accommodated.

The first and the second features are suitable for this study, which deals with large amounts of text data. The third feature is suitable for this study because different amounts of sentences are used between groups: the number of companies with problems differs from the number of companies without problems.

Table 4. Variables used for DEA discriminant analysis.

Decision variables	
G	Group
i	Extracted word
j	Company
z_{ij}	Number of occurrences of word i in text data of company j
η	Width of overlap
Dependent variables	
λ	Discrimination coefficient
d	Discrimination boundary
S_{ij}^+, S_{ij}^-	Slack variable

DEA discriminant analysis is conducted in two stages. First, data are classified into two groups: properly discriminated data and data that are difficult to discriminate. Second, to improve the discrimination accuracy, discriminant analysis is applied to data that are difficult to discriminate in the first stage. The DEA discriminant analysis model is presented below. **Table 4** shows variables used in DEA discriminant analysis.

Stage 1

$$\begin{cases} \min \sum_{j\in G_1} S_{1j}^+ + \sum_{j\in G_2} S_{2j}^- \\ \text{s.t.} \sum_{i=1}^{k} \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = d+\eta \quad (j\in G_1), \\ \quad \sum_{i=1}^{k} \left(\lambda_i^+ - \lambda_i z_{ij}\right) + S_{2j}^+ - S_{2j}^- = d \quad (j\in G_2), \\ \quad \sum_{i=1}^{k} |\lambda_i| = 1, \\ \quad S_{1j}^+,\ S_{1j}^-,\ S_{2j}^+,\ S_{2j}^- \geq 0. \end{cases} \tag{13}$$

The objective function of Eq. (13) is designed to minimize false discrimination. When λ_i^*, d^* are the optimal solutions obtained in stage 1, companies are classified into five categories according to the following discrimination criteria: R_0, R_1, R_2, C_1 and C_2.

$$R_1 = \left\{ j\in G \;\middle|\; \sum_{i=1}^{k} \lambda_i^* z_{ij} \geq d^* + \eta \right\}, \tag{14}$$

$$R_0 = \left\{ j\in G \;\middle|\; d^* + \eta > \sum_{i=1}^{k} \lambda_i^* z_{ij} > d^* \right\}, \tag{15}$$

$$R_2 = \left\{ j\in G \;\middle|\; \sum_{i=1}^{k} \lambda_i^* z_{ij} \leq d^* \right\}, \tag{16}$$

$$C_1 = \{ j\in R_1 \mid j\in G_1 \}, \tag{17}$$

$$C_2 = \{ j\in R_2 \mid j\in G_2 \}. \tag{18}$$

Therefore, C_1 and C_2 are classified correctly. $G_1\cap R_2$, $G_2\cap R_1$ are the datasets that became misjudged. Set R_0 are data in the overlap region.

Misidentified data and data existing in the overlapping area are handled in stage 2. Variable c is a new discrimination boundary existing between d^* and $d^*+\eta$.

Stage 2

$$\begin{cases} \min \sum_{j\in G_1\cap(R_0\cup R_2)} S_{1j}^+ + \sum_{j\in G_2\cap(R_0\cup R_1)} S_{2j}^- \\ \text{s.t.} \sum_{i=1}^{k} \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = c \quad \left(j\in G_1\cap(R_0\cup R_2)\right), \\ \quad \sum_{i=1}^{k} \lambda_i z_{ij} + S_{2j}^+ - S_{2j}^- = c \quad \left(j\in G_2\cap(R_0\cup R_1)\right), \\ \quad \sum_{i=1}^{k} \lambda_i z_{ij} \geq d^*+\eta \quad (j\in C_1), \\ \quad \sum_{i=1}^{k} \lambda_i z_{ij} \leq d^* \quad (j\in C_2), \\ \quad \sum_{i=1}^{k} |\lambda_i| = 1, \\ \quad S_{1j}^+,\ S_{1j}^-,\ S_{2j}^+,\ S_{2j}^- \geq 0. \end{cases} \tag{19}$$

Data correctly determined in stage 1 are controlled by a constraint expression to ensure the result obtained in stage 1. In the objective function, slack variables of data correctly discriminated in stage 1 are excluded. Therefore, the sum of slack variables when data are erroneously discriminated in stage 1 is minimized in stage 2. In addition, the discrimination boundary value c is set as between d and $d+\eta$. Also, it is possible to discriminate data existing in the overlap region. If the optimal solution of stage 2 is c^*, λ_i^*, then it is judged according to the following criteria. In Eq. (20), company j is judged as belonging to G_1. In the case of Eq. (21), company j is determined as belonging to G_2.

$$\sum_{i=1}^{k} \lambda_i^* z_{ij} \geq c^*, \tag{20}$$

$$\sum_{i=1}^{k} \lambda_i^* z_{ij} < c^*. \tag{21}$$

2.10. Logistic Curve

Logistic regression is a multivariate analytical method that explains and predicts the success probability of one categorical variable (binary variable) using multiple explanatory variables. The probability of problem occurrence is calculated using the sigmoid function from the discrimination score obtained in the previous section. As a result, one can identify a company that should devote attention to a problem, even given a degree of uncertainty. Eq. (22) is used to calculate the probability from the discriminant equation.

$$p = \frac{1}{1+\left(\dfrac{1-\pi_1}{\pi_1}\right) * \exp(-z)}. \tag{22}$$

Table 5. Experiment conditions.

	Fraud problem group: G_1	Normal group: G_2
Number of companies	40	90
Paragraph	3731	3946
Sentences	11978	14550
Total number of words	236880	269980
Number of different words	8592	9074
Learning data	20	70
Prediction data	20	20
Width of overlap	0.5	

In that equation, p stands for the probability of belonging to group 1, π_1 denotes the prior probability of group 1, and z expresses the discrimination score.

3. Computer Experiments

3.1. Experiment Overview

Through computer-based experimentation, actual data are analyzed using the proposed method as a target of the problem of cancellation. Specifically, a discriminant is created that predicts a client that has fraud problems. To solve a problem having actual scale and scope, the method was extended to solve the following three points. To confirm the effectiveness of the proposed method, it is compared with an existing method. To verify the validity of extending it probabilistically using the logic regression model, it is compared with a consultant's knowledge and actual survey results.

3.2. Experiment Conditions

The experiment conditions are presented below. Although machine learning is not used in this study, data used for discriminant creation are called learning data. Data used for verifying the discriminant are called prediction data. **Table 5** shows the conditions.

Experiment environment

- OS: Windows™ 8.1
- CPU: Intel® Core™ i5-4460S CPU@2.90 GHz
- Memory: 8.0

3.3. Effectiveness of Proposed Method

Table 6 presents the average of discrimination rates using the proposed method. An indistinguishable company is a company for which no words used in the created discriminant appear in the text data or can be discriminated. The discrimination rate of the fraud problem group is 75.2%; that of the normal group is 72.0%. The rate of unidentifiable firms with fraud issues is 12.0%. The rate of unidentifiable firms in the group without fraud problems is 27.5%. The fact that the group without fraud problems has a lower discrimination rate than the group with fraud problems is attributable to the uncertainty that a fraud problem might be occurring in the group without fraud problems. Also, the percentage of undetermined companies in the group without fraud problems is higher than that in the group with fraud problems, suggesting that firms with fraud problems have characteristics related to problem occurrence.

Table 6. Average discrimination ratio.

Improved method		
	Fraud problem	Normal
Discriminant ratio	75.2	72.0
Undetermined company ratio	12.0	27.5

The proposed method [17] enabled creation of discriminants without bias among the groups, yielding respective discrimination rates of 75.2% and 72.0% in the fraud problem group and normal group. These results confirmed the proposed method is applicable to fraud prediction.

When considering application to actual data, companies that the consultant should approach are those companies which have been determined as having a fraud problem within the group without fraud problems. However, some companies that have been identified as having fraud problems are those which have a high probability of actually having fraud problems, while others are difficult to distinguish and are misidentified. Supporting companies with a high probability of fraud problems must be done efficiently because consultants have limited resources. This study specifically addresses the issue of fraud problems, which entails uncertainty. Many companies among the companies which are determined to have no fraud problems might actually have fraud problems. To identify companies that must take the approach, the possibility of fraud problems is calculated from the discrimination score using a logistic regression model.

Tables 7 and **8** present results obtained from calculating the fraud probability in the best solution. **Table 7** presents results for the fraud problem group. **Table 8** presents results obtained for the normal group.

Regarding fraud, companies that the consultant must support are those companies determined to have a fraud problem in the normal group (**Table 8**). It is difficult to determine which of the five companies that have been identified as having fraud problems in the binary classification should be assigned priority. By calculating the fraud problem occurrence probability, one can determine which company should be supported. As described hereinafter, after verifying the effectiveness of the support method which introduced the occurrence probability, the results are compared with forecast results given by consultants.

Table 7. Probability of fraud (fraud problem group).

Company ID	Discrimination by binary classification		Prob. of fraud
	Judgment	No. companies	
1	Normal	4	7.32
2			7.42
3			8.41
4			9.06
5	Fraud	10	9.63
6			9.95
7			10.22
8			12.64
9			13.34
10			13.87
11			16.60
12			24.79
13			32.20
14			100.0

Table 8. Probability of fraud (normal group).

Company ID	Discrimination by binary classification		Prob. of fraud
	Judgment	No. companies	
21	Normal	11	0.35
22			0.44
23			1.62
24			2.06
25			2.25
26			2.76
27			3.79
28			4.17
29			4.50
30			4.84
31			5.21
32	Fraud	5	7.61
33			11.07
34			13.72
35			23.35
36			100.0

3.4. Comparison with Consultant Forecast and Survey Results

To verify the effectiveness of this method for companies, even considering uncertainties, consultants conducted a forecast of fraud problems and interviewed client and companies.

Table 9. Extracted words.

Paid
Part time job
Allowance
Number
Main rule
Trial calculation
Profit
Postscript
Settlement
Applicable

Table 10. Consultant's reason of inference.

No fraud has been reported yet. Cash management and **profit** management are poor. Therefore, one would not notice even if someone were fraudulent.
In a family business, accounting is for relatives only. Because a human relationship is involved, one feels that no injustice occurs, but there is an environment that allows unjust acts.
Although each employee works separately, no management system exists; employee education is done solely through on-the-job training.
Taxi businesses handle cash and might be fraudulent.
It seems that the risk is high because there are many foreign (**part time job**) employees.

3.4.1. Comparison of Extracted Words and Factors for Predicting Fraud: Consultants

Table 9 presents 10 words with strong factors in the group with fraudulent problems extracted using this method. **Table 10** shows reasons why a consultant predicted that a problem occurred in fraud problem prediction. Words that are bold also appear in the consultant's reason of inference. A "part-time job" is consistent with the consultant's finding that part-time jobs are associated with the occurrence of fraud problems. The word "profit" is also associated with fraud problems in client companies for which profit management is poor.

This study is intended to discover factors related to the occurrence of fraud problems of which the consultant is unaware. However, results confirmed that the proposed method can extract words and phrases matching the consultant's knowledge.

3.4.2. Probability of Fraud Using a Logistic Curve

Figure 9 shows a graph obtained by converting the discrimination scores of Section 3.3 into probabilities using logistic regression. The horizontal axis shows the company numbers. The vertical axis shows the probability that fraud has occurred.

As for company numbers, 1–18 are of the fraud problem group; 19–34 are from the normal group. Accord-

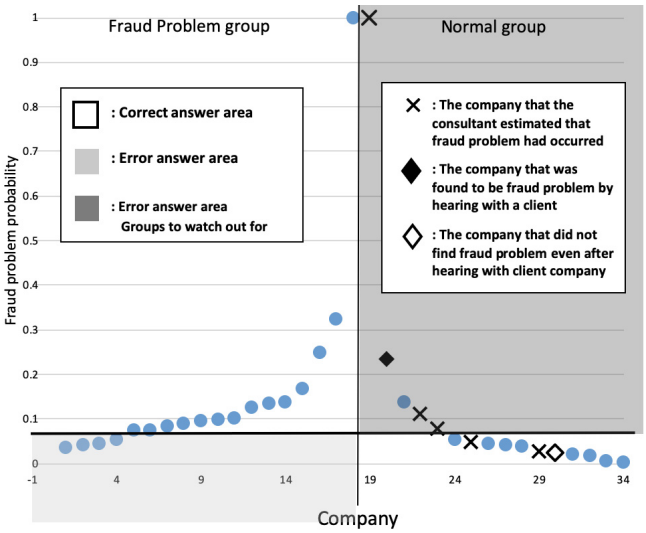


Fig. 9. Probability of fraud.

ingly, the white area is a company for which the determination is correct. The light gray and dark gray areas are companies for which the determination was incorrect. However, because the fraud problem includes uncertainty, consultants predicted and conducted interviews with client companies about companies that did not describe the fraud problem in the text data. Crosses show companies for which the consultant estimated the possibility of fraud as high because of the company characteristics. Diamonds denote companies that were interviewed directly. Black ones represent companies that had a fraud problem. White ones represent companies that found no fraud problem.

Even for companies that have no description of a fraud problem in the text data, this method was able to find that a company that is likely to have a fraud problem based on a consultant's guess or interview is likely to have a fraud problem. Additionally, it was also confirmed that companies determined to have a low probability of fraud problems found no fraud problems through interviews. Results suggest that the proposed method is useful to identify companies that have found a fraud problem, although the text data include no description of the discovery of a fraud problem.

A support system that predicts problem occurrence is realized using this method. It can engender the provision of stable quality of service and support the rapid development of new employees.

4. Conclusions

This study was conducted to devise a method to support consulting service companies so that the companies can respond to client demands irrespective of the company expertise. Occurrence of future problems at client companies is predicted using text mining with data obtained from a consulting company. To those data, correspondence analysis and DEA discriminant analysis are then applied. This report presents a proposal of extending a method to predict problems that include uncertainty. As described herein, computer experiments were conducted to verify the effectiveness of the proposed method by comparison with consultants' knowledge and results of interviews conducted with client companies.

The obtained results are summarized below.

- The proposed method made it possible to create discriminants without bias among the groups, yielding respective discrimination rates of 75.2% and 72.0% in the fraud problem group and normal group. Results confirmed that the group without fraud problems had a lower discrimination rate than the group with fraud problems because some companies actually had fraud problems.
- This method identified that a company which is likely to have a fraud problem attributable to a consultant's guess or interview is likely to have a fraud problem, even for companies that have no description of a fraud problem in the text data. Results suggest that the proposed method is useful to infer companies that have found a fraud problem, even when the text data include no description of a fraud problem discovery.

As described herein, the method to support state recognition in consultants' client companies was proposed, but a support method for the consultant's judgment system is also necessary to construct supplementary systems. Application of machine learning and extension of support methods remain as tasks for future study.

References:

[1] "Small and Medium Enterprise Charter," 2010.
[2] http://www.chusho.meti.go.jp/sme_english/index.html [Accessed August 13, 2020]
[3] "White Paper on Small and Medium Enterprises in Japan," 2017.
[4] K. Hori, "What is consulting," PHP, 2011.
[5] F. Duan, M. Morioka, J. Too, C. Tan, and T. Arai, "Multi-Modal Assembly Support System for Cell Production," Int. J. Automation Technol., Vol.2, No.5, pp. 384-389, 2008.
[6] T. Kamma, T. Saito, and S. Abe, "Analysis and Adaptation for Exaggeration Types of Animation Motion," Graphic Science, Vol.47, pp. 13-23, 2013.
[7] T. Nasukawa, "Technology using text mining / Technology to make – essence and application derived from basic technology and application examples," Tokyo Denki University Press, 2006 (in Japanese).
[8] H. Wakimori, "Text Mining Techniques for Analyzing Big Data," UNISYS Technology Review, Vol.115, pp. 337-349, 2013.
[9] H. T. P. Thanh and P. Meesad, "Stock Market Trend Prediction Based on Text Mining of Corporate Web and Time Series Data," J. Adv. Comput. Intell. Intell. Inform., Vol.18, No.1, pp. 22-31, 2014.
[10] J. Yano and K. Araki, "Performance Evaluation for a Method of Generating Business Reports from Call Center Speech Dialogues Using Inductive Learning," Information Processing Society of Japan SIG Technical Report, 2007.
[11] H. Takase, H. Kawanaka, and S. Tsuruoka, "Supporting System for Quiz in Large Class – Automatic Keyword Extraction and Browsing Interface –," J. Adv. Comput. Intell. Intell. Inform., Vol.19, No.1, pp. 152-157, 2015.
[12] M. Nii, K. Takahama, and S. Miyake, "Rule Representation for Nursing-Care Process Evaluation Using Decision Tree Techniques," J. Adv. Comput. Intell. Intell. Inform., Vol.18, No.6, pp. 918-925, 2014.

[13] H. Murai and A. Tokosumi, "Network Analysis of the Four Gospels and the Catechism of the Catholic Church," J. Adv. Comput. Intell. Intell. Inform., Vol.11, No.7, pp. 772-779, 2007.
[14] M. J. Jones and M. Clatworthy, "Financial reporting of good news and bad news: evidence from accounting narratives," Accounting and Business Research, Vol.33, pp. 171-185, 2003.
[15] K. Izumi, T. Goto, and F. Matsui, "Long-term market trend estimation using economic text information," Information Processing Society of Japan J., Vol.52, No.12, pp. 3309-3315, 2011 (in Japanese).
[16] R. Watanabe, N. Fujii, D. Kokuryo, T. Kaihara, Y. Abe, and R. Santo, "A Study of Supporting Method of Consulting Service using text mining," Int. J. Automation Technol., Vol.12, No.4, pp. 482-491, 2018.
[17] R. Watanabe, N. Fujii, D. Kokuryo, T. Kaihara, and Y. Abe, "Application of support systems for consulting service to real problem using a synonym dictionary," Acta Electrotechnnica et Informatica, Vol.20, No.2, pp. 3-10, 2020.
[18] http://taku910.github.io/mecab/ (in Japanese) [Accessed August 13, 2020]
[19] S. Shida, T. Maeda, and M. Yamazaki, "Statistical Input for Language Research Gate," Kuroshio Publishing, 2010 (in Japanese).
[20] T. Sueyoshi, "DEA-discriminant analysis in the view of goal programming," European J. of Operational Research, Vol.115, pp. 564-582, 1999.

Name:
Ruriko Watanabe

Affiliation:
Ph.D. Student, Department of Systems Science, Graduate School of System Informatics, Kobe University

Address:
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

Brief Biographical History:
2015- Department of Computer Science and Systems Engineering, Kobe University

Main Works:
• "A Study on Support Method of Consulting Service Using Text Mining," Int. J. Automation Technol., Vol.12, No.4, pp. 482-491, 2018.

Membership in Academic Societies:
• Japan Society of Mechanical Engineers (JSME)
• Japan Society for Precision Engineering (JSPE)
• Institute of Systems, Control and Information Engineers (ISCIE)

Name:
Nobutada Fujii

Affiliation:
Associate Professor, Department of Systems Science, Graduate School of System Informatics, Kobe University

Address:
1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

Brief Biographical History:
1998- JSPS Research Fellow, Kobe University
2000- Research Associate, Department of Mechanical Engineering, Kobe University
2002- Research Associate, Research into Artifacts, Center for Engineering (RACE), The University of Tokyo
2005- Invited Associate Professor, RACE, The University of Tokyo
2007- Associate Professor, Graduate School of Engineering, Kobe University
2010- Associate Professor, Graduate School of System Informatics, Kobe University

Main Works:
• "An EOQ Model for Reuse and Recycling Considering the Balance of Supply and Demand," Int. J. Automation Technol., Vol.9, No.3, pp. 303-311, 2015.
• "A Study on Support Method of Consulting Service Using Text Mining," Int. J. Automation Technol., Vol.12, No.4, pp. 482-491, 2018.

Membership in Academic Societies:
• Japan Society of Mechanical Engineers (JSME)
• Japan Society for Precision Engineering (JSPE)
• Society for Serviceology (SfS)
• Institute of Systems, Control and Information Engineers (ISCIE)

Name:
Daisuke Kokuryo

Affiliation:
Assistant Professor, Graduate School of System Informatics, Kobe University

Address:
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

Brief Biographical History:
2004-2007 Ph.D. Student, Graduate School of Science and Technology, Kobe University
2008-2014 Postdoctoral Fellow / Researcher, National Institute of Radiological Sciences
2015- Project Assistant Professor / Assistant Professor, Graduate School of System Informatics, Kobe University

Main Works:
• "Evaluation of a combination tumor treatment using thermo-triggered liposomal drug delivery and carbon-ion irradiation," Translational Research, Vol.185, pp. 24-33, 2017.
• "Value Co-creative Manufacturing with IoT based Smart Factory for Mass Customization," Int. J. Automation Technol., Vol.11, No.3, pp. 509-518, 2017.

Membership in Academic Societies:
• Institute of Systems, Control and Information Engineers (ISCIE)
• Japanese Society for Magnetic Resonance in Medicine (JSMRM)
• International Society for Magnetic Resonance in Medicine (ISMRM)
• Japanese Society for Computer Aided Surgery (JSCAS)

Name:
Toshiya Kaihara

Affiliation:
Professor / Deputy Dean, Graduate School of System Informatics, Kobe University

Address:
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

Brief Biographical History:
1985- Mitsubishi Electric Corporation
2001-2004 Associate Professor, Kobe University
2004- Professor, Kobe University

Main Works:
• "Optimization and Simulation of Collaborative Networks for Sustainable Production and Transportation," IEEE Trans. on Industrial Informatics, Vol.12, No.1, pp. 417-424, 2016.
• "Optimisation of freight flows and sourcing in sustainable production and transportation networks," Int. J. of Production Economics, Vol.164, pp. 351-365, 2015.
• "Cloud-based Additive Manufacturing and Automated Design: A PLM-enabled Paradigm Shift," Sensors, Vol.15, No.12, pp. 32079-32122, 2015.

Membership in Academic Societies:
• College International pour la Recherche en Productique (CIRP)
• International Federation for Information Processing (IFIP)
• Institute of Electrical and Electronics Engineers (IEEE)
• Japan Society of Mechanical Engineers (JSME)

Name:
Yoichi Abe

Affiliation:
F&M Co., Ltd.

Address:
1-23-38 Esaka-cho, Suita, Osaka 564-0063, Japan

Brief Biographical History:
1999- F&M Co., Ltd.
2018- General Director, Japan Cryptocurrency Tax Association (JCTA)

Main Works:
• "How to leave smart money" that the presidents of the construction industry want to know
• Grant/fund procurement
• Management matters review
• Labor management measures