APPLICATION OF SUPPORT SYSTEMS FOR CONSULTING SERVICE TO REAL PROBLEM BY USING A SYNONYM DICTIONARY

Ruriko WATANABE^{*}, Nobutada FUJII^{*}, Daisuke KOKURYO^{*}, Toshiya KAIHARA^{*}, Yoichi ABE^{**} ^{*}Department of Systems Science, Graduate School of System Informatics, Kobe University Graduate School, 1-1 Rokkodai-cho, Nada-ku, Kobe, Japan, tel. +81 78 803 6033, e-mail: watanabe@kaede.cs.kobeb-u.ac.jp

* F&M Co., Ltd.

ABSTRACT

This study aims to build a support method for consulting service companies allowing them to respond to client demands regardless of the expertise of the consultants. With an emphasis on the revitalization of small and medium-sized enterprises, the importance of support systems for consulting services for small and medium-sized enterprises, which support solving problems that are difficult to deal with by an enterprise, are increasing. Consulting companies can respond to a wide range of management consultations; however, because the contents of a consultation are widely and highly specialized, a service proposal and the problem detection depend on the experience and intuition of the consultant, and thus a stable service may occasionally not be provided. Therefore, a support system for providing stable services independent of the ability of consultants is desired. In this research, as the first step in constructing a support system, an analysis of customer information describing the content of a consultation with the client companies is conducted to predict the occurrence of future problems. Text data such as the consultant's visitation history, consultation content by e-mail, and call center content are used in the analysis because the contents explain not only the current problems but also possibly contain future problems. This research proposed method for analyzing the text data by employing text mining. In the proposed method, by combining a correspondence analysis with a DEA (Data Envelopment Analysis) discriminant analysis, words that are strongly related to problem detection are extracted from a large number of words obtained from text data, and variables of the DEA discriminant analysis are reduced and analyzed. This paper describes improved method for the application in the real problem. The method is improved to eliminate the following two problems. First, IDF values are used to extract more general phrases. Second, in order to reduce the number of companies that cannot be identified, it is used standardization and data are expanded with synonym dictionaries. In this study, computer experiments were conducted to verify the effectiveness of the improved method through a comparison with an existing method. The results of the verification experiment are as follows. First, there is a possibility of discovering new factors that cannot be determined from the intuition and experience of the consultant regarding the target problem. Second, through a comparison with the existing method, the effectiveness of the proposed method was confirmed.

Keywords: Text Mining, Correspondence Analysis, DEA Discriminant Analysis, Service Engineering

1. INTRODUCTION

In recent years, with the emphasis on the activation of small and medium-sized enterprises [1], support systems of consulting for small and medium-sized enterprises are regarded as important [2], which support small and medium-sized enterprises to solve problems that are difficult to deal with within their company. Although consulting company can respond to a wide range of management consultation, they have a problem that leading in the special field is not always sufficient [3]. In addition, service suggestions and client companies' problem detection depend on the experience and intuition of each consultant. An auxiliary system is necessary to provide services with stable quality independent of consultant's ability. The purpose of this study is to realize consulting service that does not depend on expertise of consultant. As a first step in constructing an auxiliary system, this study proposes a method to predict occurrence of future problems in client companies by using text mining with data describing communication of consultation matters received from customers.

In previous studies [4], a discriminant was made using real-scale data for cancellation problems and the effectiveness of the proposed method was confirmed. It was confirmed that the difference between the number of cancellation companies and the number of continuing companies affects discriminant analysis, and extending the objective function of DEA discriminant analysis, it was able to suppress the influence of the difference in the number of companies between groups.

In this paper, to solve the problem of the real scale problem, the method is extended to solve the following three problems.

- The amount of text of the prediction data is smaller than that of the training data, and the overlapping of words is small. The method is extended to use words that appear in only one group.
- Although the factor word was extracted using the correlation between the company and the word, with the increase in the number of companies, there was a problem that a word unique to a company with a large amount of text data was also extracted. The method was extended using IDF values to extract more appropriate words.
- Data are expanded with synonym dictionaries in order to reduce the number of companies that cannot be identified.
- A normalization was introduced to normalize the word distribution.

In order to verify the effectiveness of the proposed method, prediction is made about cancellation problem and is compared with the prediction results by the consultant.

2. METHODS

2.1. Research subject

By analyzing text data accumulated in consulting companies, possible problems in client companies are predicted in this research. The text data to be used were recorded through various methods, such as the interaction at the time of a visit between the consultant and client company, questions from the client company provided by e-mail, and correspondence with a call center. Examples of such data are shown below.



Fig. 1 Recorded text data

As shown in Fig. 1, the text data consist of the company profile, visitation history, problems currently held by the client companies, and the consulting services in use. Some of the texts appear to describe contents that may be a trigger of future problems. In this study, these text data were analyzed by a computer to support a consulting service. The proposed method supports consultants who need to deal with problems from a wide variety of fields, allowing consultants with little experience to be aware of the problems that can occur in the client companies by predicting them from the accumulated text data.

2.2. Overview of proposed method

This section outlines the proposed method. First, text data are classified based on the problem occurrence, and a discriminant formula is created using a text mining method. Phrases are extracted as factors by extracting the correspondences for each group from among many different words and phrases, and using the extracted phrases as variables of a DEA discriminant analysis to create a discriminant. To verify the effectiveness of the discriminant formula obtained, newly categorized text data are used to judge the presence or absence of a problem. A flow of the proposed method is shown in Fig. 2.



Fig. 2 Flow of proposed method

The proposed method is conducted through the following six steps. These steps are presented in Fig. 2. As described in this paper, the method was improved by the introduction of IDF values in III, the use of words that appear in only one group in IV, using synonym dictionaries in V and the standardization in VI. Detailed descriptions of the respective steps are presented in the following sections.

I.	Classify the text data	(2.3.)
II.	Morphological analysis	(2.4.)
III.	Correspondence analysis	(2.5.)
IV.	Extract words as factors	(2.6.)
V.	Data augmentation	(2.7.)
VI.	Standardization	(2.8.)
VII.	DEA discriminant analysis	(2.9.)

2.3. Classify text data

This study used text data related to communication with the client companies accumulated by a consulting company. The text data are categorized according to the presence or absence of problem occurrence when considering the time series of problem detection. To predict problem occurrence, text data with direct content on the problem are deleted. In addition, past text data for a certain period of time since the problem occurred are applied because the proposed method tries to predict the occurrence of a problem as early as possible by not applying a direct description of the problem in the text data.

2.4. Morphological analysis

A morpheme, the smallest character string, becomes meaningless when it is decomposed further. Decomposing sentences into morphemes and specifying their parts of speech is called morphological analysis. This study used MeCab [5], a morphological analyzer developed by Kudo as a general-purpose analyzer that is independent of dictionaries or text data. It is faster than other analyzers. Words and phrases of the following items, which should be considered noise in analyses, were omitted from the analyzed morphemes.

- Name of consulting service
- Fixed phrase
- Incoherent words

2.5. Correspondence analysis

2.5.1. Overview

Correspondence analysis, proposed in the 1960s by French researcher Jean-Paul Ben-zeccri, is a method to compress information in the rows and columns of a data table into a few components [6]. In this study, words are included in lineitems (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 1 presents two-dimensional data: a_i is the sample score; b_i is the category score. Each is set as a variable.

Word AAA BBB ... KKK ... Company a a_2 a_{κ} A Co. ••• b_1 t_{11} t_{12} t_{1K} B Co. b_2 ... t_{2K} t_{21} t_{22} ٠. ÷ ÷ N Co. $t_{N^{-1}}$ ••• t<u>_NK</u> b_N t_N

Table 1 Data for Correspondence Analysis

Variables are set for samples and categories. Calculations are done to maximize inter-variable correlation. 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 the presence/absence of fraud problem detection. Also, attention is devoted to the word close to the origin. Words which become factors of the respective groups are extracted using the method shown in the next section.

2.5.2. Caluculating the word factor strength

The results of correspondence analysis showed, considering all dimensions, distance d_{iG} between word *i*. The origin in each group (G_1, G_2) is calculated using Eq. (1). Here, D_G represents the total number of dimensions, x_{ijG} is the sample score of word *i* in dimension *j*, and C_{jG} denotes the contribution of dimension *j*.

$$d_{iG} = \sqrt{\frac{\sum_{j=1}^{D_G} \{(x_{ijG} * C_{jG})^2\}}{D_G}}$$
(1)

Weighting is performed by multiplying the distance from the origin by the IDF value to extract more commonly used words. The IDF value, called reverse document frequency, is a high value if the word is rare between documents, and a low value if the word appears frequently in many sentences. IDF values are calculated using Eq. (2).

$$idf_i = \log \frac{N}{N'_i} \tag{2}$$

In that equation, N denotes the total number of documents; N'_i represents the number of documents including word *i*.

2.6. Extract words to be factored

2.6.1. Extracting words appearing in both groups

For each word, the d value is updated by the operations of Eq. (3) and (4). For each group, a fixed number of words with small e values are extracted.

when $G_1 < G_2$

$$\begin{cases} e_{iG_1} = d_{iG_1} * idf_i + M - d_{iG_2} * idf_i \\ e_{iG_2} = \infty \end{cases}$$
(3)

when $G_1 > G_2$

$$\begin{cases} e_{iG_2} = d_{iG_2} * idf_i + M - d_{iG_1} * idf_i \\ e_{iG_1} = \infty \end{cases}$$
(4)

2.6.2. Extracting words appearing in only one group

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

- I. From the words appearing in both groups, select the word with the strongest factor: the word with the highest e_{iG} value.
- II. Words that appear in only one group with a distance smaller than the selected words are extracted in order from a word with a 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 the method, words used for creating a discriminant are words that overlap in the learning data group and the prediction data group. Since the data for prediction is smaller than the data for creation, there is a problem that there are many companies that cannot be distinguished because no word used in the discriminant appears. Therefore, this research aims to reduce the number of unidentifiable companies by expanding data by using synonym dictionaries for phrases that appear in the data for prediction. Fig. shows an image of the data augmentation.



Fig. 3 Data augmentation

For data augmentation, Wordnet is used, which is a concept dictionary taking into account the meaning and the thesaurus. Fig. 4 shows an example of a concept dictionary.



Fig. 4 Example of word "月(moon)"

A word has a plurality of concepts, and 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, and the appearance frequency table is extended.

2.8. Standardization

A discriminant is created using the words extracted in section F. Before performing discriminant analysis, standardization is applied to unify the word scale. The standardization is calculable by Eq. (5) 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.

$$z = \frac{x - M}{s}$$
(5)

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

2.9. Dea discriminant analysis

The DEA discriminant analysis method is a discrimination model proposed by Sueyoshi [7]. A DEA discriminant analysis is conducted in two stages. In the first stage, the data are classified into two groups: properly discriminated data and data that cannot be discriminated or cannot be discriminated easily. In the second stage, improving the accuracy of the discrimination, a discriminant analysis is applied to data that are difficult or unable to detect in the first stage. A model of the applied DEA discriminant analysis is shown below. Table II presents variables used for DEA discriminant analysis.

Table 2 Variable using DEA discriminant analysis

Decision variables			
G	Group		
i	Extracted word		
j	Company		

Z _{ij}	No. of word <i>i</i> occurrences in company <i>j</i> text data	
η	Width of overlap	
Dependent variables		
λ	Discrimination coefficient	
d	Discrimination boundary	
$S_{ij}^+, \overline{S_{ij}^-}$	Slack variable	

Stage 1

The objective function of Eq. (11) 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 .

The objective function of Eq. (6) is designed to minimize false discrimination. When λ_i^*, d^* are the optimal solutions obtained in stage 1, the companies are classified into five categories according to the following discrimination criteria: R_0, R_1, R_2, C_1 , and C_2 .

$$R_1 = \left\{ j \in G \mid \sum_{i=1}^k \lambda_i^* z_{ij} \ge d^* + \eta \right\}$$
(7)

$$R_{0} = \left\{ j \in G | d^{*} + \eta > \sum_{i=1}^{k} \lambda_{i}^{*} z_{ij} > d^{*} \right\}$$
(8)

$$R_2 = \left\{ j \in G | \sum_{i=1}^k \lambda_i^* z_{ij} \le d^* \right\}$$
⁽⁹⁾

$$C_1 = \{ j \in R_1 | j \in G_1 \}$$
(10)

$$C_2 = \{ j \in R_2 | j \in G_2 \}$$
(11)

Therefore, C_1 and C_2 are classified correctly. Here, $G_1 \cap R_2$, $G_2 \cap R_1$ is a dataset that was misjudged. Set R_0 is made up of data in the overlapped region.

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

Stage 2
min
$$\sum_{j \in G_1 \cap (R_0 \cup R_2)} S_{1j}^+ + \sum_{j \in G_2 \cap (R_0 \cup R_1)} S_{2j}^-$$
s.t.
$$\sum_{\substack{i=1\\i=1}^k} \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = c$$

$$(j \in G_1 \cap (R_0 \cup R_2))$$

$$\sum_{\substack{i=1\\i=1}^k} \lambda_i z_{ij} + S_{2j}^+ - S_{2j}^- = c$$

$$(j \in G_2 \cap (R_0 \cup R_1))$$

$$\sum_{\substack{i=1\\i=1}^k} \lambda_i z_{ij} \ge d^* + \eta (j \in C_1)$$

$$\sum_{\substack{i=1\\i=1}^k} \lambda_i z_{ij} \le d^* (j \in C_2)$$

$$\sum_{\substack{i=1\\i=1}^k} |\lambda_i| = 1$$

$$S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \ge 0$$
(12)

Data correctly determined in stage 1 are controlled using a constraint expression to ensure the results of stage 1. In the objective function, the slack variables of data correctly discriminated in stage 1 are excluded. Therefore, the sum of the slack variables that occurs when the data are erroneously discriminated in stage 1 and also discriminated erroneously in stage 2 is minimized. In addition, the discrimination boundary value *c* is set as between *d* and $d + \eta$. It is possible to discriminate data existing in the overlapped region. If the optimal solution of stage 2 is c^*, λ_i^* , then data are judged according to the following criteria. In Eq. (13), company *j* is judged to belong to G_1 . In the case of Eq. (14), company *j* is determined to belong to G_2 .

$$\sum_{i=1}^{k} \lambda_i^* \, z_{ij} \ge c^* \tag{13}$$

$$\sum_{i=1}^{k} \lambda_i^* \, z_{ij} < c^* \tag{14}$$

3. Computer experiments

3.1. Experiment overview

In computer experiments, actual data are analyzed using the proposed method as a target of the problem of cancellation. Specifically, a discriminant that predicts client consulting service cancellation is created. To solve a real-scale problem, the method was extended to solve the following three points. The validity of the extended method was verified.

- Using words that appear in only one group
- IDF value
- Data augmentation
- Standardization

To verify the effectiveness of the proposed method, prediction is conducted for a cancellation problem. Results are compared with the prediction results from a consultant.

3.2. Target data

Describe data when predicting a company for which n month d year is the update month. Figure 3 shows data (learning data) for discriminant preparation. Data of the company's prior nine months were used. However, the earlier three months were deleted from the renewal month because the study is aimed at making a cancellation projection before the consultant approaches the cancellation, which requires a three month period.



Fig. 3 Learning data

Fig. 4 shows data (prediction data) for target month prediction. It accomplishes prediction using text data of the company which is the target update month. Similarly to data for creation, we delete the data of the three months before the update month.



Fig. 4 Prediction data

3.3. Common experiment conditions

The experiment conditions are shown below. In this study, although machine learning is not used, data used for discriminant creation are designated as learning data. Data used for verifying the discriminant are designated as prediction data. Table 3 expresses the condition.

Table 3 Experiment conditions

	Cancellation	Continuation
	Group: G_1	Group: G_2
Learning data	315	4048
Prediction data	349	91
No. of extracted	500	
words		
Width of overlap		0.5

3.4. Effectiveness of method improvement

3.4.1. Data augmentation

When data Augmentation was performed on 500 words extracted to create a discriminant, the number of words increased to 1832 words. Table4 shows the data augmentation of five words having a strong factor in the cancellation group.

Extracted words	synonyms
ris	-
phrase	phrase, idiom, music phrase
lose weight	slim up, reduce weight, degenrate
now	soon, presently, shortly, before long, presently, currently, just, just now,
defualt	nonpayment, nonremittal

 Table 4
 Data Augmentation

There are more synonyms for conjunctions and adverbs than for nouns. "Risu" is a word whose meaning is incomprehensible because a part of the name of the risk management service is incorrectly morphologically analyzed. It is noteworthy ,however, that risk management services have been identified as the strongest terms in the cancellaiton group.

3.4.2. Learning data

Table 5 presents the discrimination rate of learning data and the number of companies that cannot discriminate. Here, the undeterminable company is a company with words used in the discriminant which do not appear in the text data and which cannot be discriminated in either group. In other words, it is not the number of misclassified companies. Both methods have discrimination rates close to 100% for both groups. The existing method showed about half the number of undetermined companies existed. However, using the improved method, the number of unidentifiable companies became zero. This result is regarded as attributable to the fact that "the phrase does not appear" also contributes to the determination factor because the words are standardized and the occurrence frequency is assigned a value of 0.

Method before improvement			
	Cancellation	Continuation	
Discriminant ratio	96.8	97.6	
Undetermined	62.5	68.4	
companies ratio			
Improved method			
	Cancellation	Continuation	
Discriminant ratio	99.0	99.9	
Undetermined	0	0	
companies raito			

Table 5 Discriminant ratio of learning data

3.4.3. Prediction data

Table 6 shows the discrimination rate of prediction data and the number of companies that cannot be discriminated.

 Table 6
 Discriminant ratio of prediction data

Method before improvement				
	Cancellation	Continuation		
Discriminant ratio	72.0	66.7		
Undetermined companies raito	54.5	60.1		
Improved method				
	Cancellation	Continuation		
Discriminant ratio	72.2	68.9		
Undetermined companies ratio	7.3	8.2		

In the existing method, the discrimination rate of the continuation group is 100%, but the discrimination rate of the cancellation group is 0.9%. Predicting the company to be cancelled is not possible. In addition, most companies have become undetermined companies. However, using the improved method, the termination rate was 72%. The continuation rate was 66.7%. It was easier to create discriminants without bias among groups than by existing methods. Although the undetermined companies were about half of all companies, they were fewer than under the existing method. The effectiveness of the improved method was confirmed from the above.

3.5. Comparison with consultants' forecasted

Fig. 5 and Fig. 6 respectively present the forecasted rates of updating by consultants and the number of firms in the results that were forecast using this method.



Fig. 5 Cancellation group



Fig. 6 Continuation group

The vertical axis presents the rate at which the consultant predicts that the target company will continue consulting services. The horizontal axis shows the number of companies.

The support system correctly predicted many companies for which consultants' forecasts were correct. The results confirmed that it is difficult for the support system to determine cancellation for a company for which the consultant predicted an update as likely to be high but for which cancellation occurred. Results suggest that the support system also captures characteristics of the company when the consultant predicts.

4. CONCLUSION

This study was aimed at producing a method for supporting consulting service companies so that companies can respond to client demand irrespective of the company expertise. Occurrence of future difficulties in client companies is predicted using text mining with data taken from a consulting company. Also, correspondence analysis and DEA discriminant analysis are used. As described in this paper, the method has been improved to apply real problems in three respects: using words that appear in only one group, an IDF value, and standardization. In computer experiments conducted to verify the effectiveness of the proposed method, prediction is made about cancellation problems. The results are compared with the prediction results achieved by the consultant. The obtained results are summarized below.

- The improved method reduced the number of companies for which results can not be determined.
- The improved method made it possible to create discriminants without bias among the groups, yielding respective discrimination rates of 72% and 66.7% in the cancellation and continuation groups.
- Comparison with forecast results produced by consultants suggests that the support system also captures characteristics of the company to be referred to when the consultant predicts.

This study examined a method to support state recognition in client companies of consultants, but a support method for the consultant's judgment system is also necessary to construct supplementary systems.

REFERENCES

- [1] Small and Medium Enterprise Charter, 2010. http://www.chusho.meti.go.jp/sme_english/index.ht ml
- [2] White Paper on Small and Medium Enterprises in Japan, 2017.
- [3] WATANABE, R. FUJII, N. KOKURYO, D. KAIHARA, S. – ABE, Y. – ONISHI, Y. – SANTO, R.: A Study of Supporting Method for Consulting Service using customer information, International Conference on Serviceology, November 2018.
- [4] http://taku910.github.io/mecab/ (in Japanese)
- [5] SHIDA, S. MAEDA, T. YAMAZAKI, M.: Statistical Input for Language Research Gate, Kuroshio Publishing, 2010. (in Japanese)
- [6] SUEYOSHI, T.: DEA-discriminant analysis in the view of goal programming, European Journal of Operational Research, 115, 564-582, 1999.

Received March 3, 2020, accepted May 21, 2020

BIOGRAPHIES



Ruriko Watanabe was born on 20.1.1992. In 2015 she graduated Department of Computer Science and Systems Engineering, Kobe University. Her main works title was "A Study on Support Method of Consulting Service Using Text

Mining". Her current research interests include the service engineering, data mining.



Nobutada Fujii. Received the Master of Engineering degree from the Kobe University, and the Ph.D. degree in Engineering from Tokyo University. He is currently an Associate Professor at the Graduate School of System Informatics, Kobe University, Japan. His research work has covered areas

ranging from self-controlled and distributed production system, service engineering. N. Fujii was the Advances in Production management Systems special session organizer for conferences (2013, 2014 and 2015). He is a number of The Japan Society for Precision Engineering, The Japan Society of Mechanical Engineers, and the program committee of Japan Society for Serviceology.



Daisuke Kokuryo. Received the Master of Engineering degree from the Kobe University, where received the Ph.D. degree in Engineering. He is currently an assistant professor at the Graduate School of System Informatics, Kobe University, Japan.

His research works has covered areas from Medical systems, Biomaterials and Biomedical engineering. He is a number of the Japan Society of computer Aided Surgery, Japanese Society for magnetic resonance in medicine, and the Institute of Systems.



Toshiya Kaihara. Received the Master of Engineering degree from the Kyoto University, and the Ph.D. degree in Engineering from Imperial College London. He is currently a Professor and vice dean at the Graduate School of System Informatics, Kobe University, Japan. He has been

involved with research in multi-agent system which is applied to production, service system and social system, management information and service engineering. He has authored 250+ refereed journals, conference proceedings, and book chapters. He is the president of the institute of Systems, Control and Information Engineers, the manager of the production system department of The Japan Society of Mechanical Engineers, the director of Japan Society for Serviceology.



Yoichi Abe. In 1999, he joined F&M Co., ltd. In 2018, he assumed office as a general director of Japan Cryptocurrency Tax Association (JCTA). He has been involved with "How to leave smart money" that the presidents of the construction industry want to know - Grant · Fund

procurement · Management matters review · Labor management measures