Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 67 (2018) 569 – 573

www.elsevier.com/locate/procedia

11th CIRP Conference on Intelligent Computation in Manufacturing Engineering, CIRP ICME '17

A study on support method of consulting service using text mining

Ruriko Watanabe[a,*], Nobutada Fujii[a], Daisuke Kokuryo[a], Toshiya Kaihara[a], Yoshinori Onishi[b], Yoichi Abe[b], and Ryoko Santo[c]

[a]*Kobe University,1-1,Rokkodai,Nadaku,Kobecity,Hyogo,657-8501,Japan*
[b]*F&M, co.,ltd.,1-23-38,Esaka-cho,Suitacity,Osaka,564-0063,Japan*
[c]*The New Industry Research Organizetion,6-1,Minatojimanaka-cho,Tyuoku,Kobecity,Hyogo,650-0064,Japan*

* Corresponding author. Tel.:81-78-803-6520; fax: +81-78-803-6520. *E-mail address:* watanabe@kaede.cs.kobe-u.ac.jp

Abstract

This study aims to build a supporting method for consulting service companies so that the companies can respond to client's demand regardless of the expertise of consultants. Occurrence of future problems in client companies is predicted by using text mining with data taken from a consulting company; correspondence analysis and DEA discriminant analysis are employed. Computer experiments are conducted to verify the effectiveness of the proposed method.

Peer-review under responsibility of the scientific committee of the 11th CIRP Conference on Intelligent Computation in Manufacturing Engineering

Keywords: Text mining, DEA Discriminat analysis, Correspondence analysis

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], creation of discriminant formula concerning fraud problem detection and verification of validity of this discriminant formula have been carried out. The influence on the text data by the business type of client company was confirmed. Compared with statistical discriminant analysis, the effectiveness of DEA discriminant analysis used in the proposed method was confirmed. In this paper, the relationship between the number of extracted words and the discrimination rate was examined in order to confirm the effectiveness of extracting the words as factors of the proposed method.

2. Method

First, classify text data by the occurrence of fraud problem, and create a discriminant formula using text mining method. In order to verify the effectiveness of the obtained discriminant formula, newly categorized text data is judged the presence or absence of fraud problem. The following list is the flow of proposed method. It is explained in detail

Peer-review under responsibility of the scientific committee of the 11th CIRP Conference on Intelligent Computation in Manufacturing Engineering
doi:10.1016/j.procir.2017.12.262

below.

- Classify the text data
- Morphological analysis
- Correspondence analysis
- Extract words to be factors
- DEA discriminant analysis
- Verification using the predictive data

2.1. Classify the text data

In this research, text data that describes communication with client companies accumulated in a consulting company are used. Text data are categorized according to the presence or absence of fraud problem, taking into account the time series of fraud problem detection. Fig. 1 and Fig. 2 indicate the relationship between the time series of fraud problem detection and group classification. In case that fraud problems are discovered during consultations, their text data are classified into a group of fraud problem. The text data after the fraud problem discovery is excluded from the analysis targets; this is because the purpose of this study is to predict the occurrence of fraud problems. Even if there is a description of fraud problem detection, when the detection is before the start of consultation, they are classified into a group of normal.

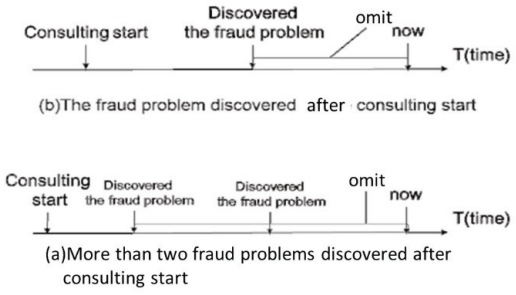

Fig. 1. Fraud Problem Group.

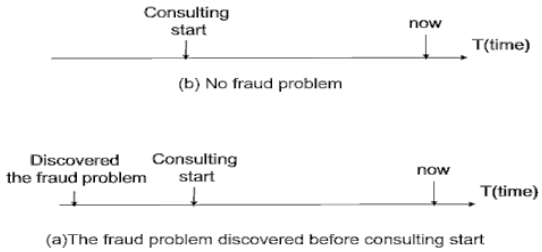

Fig. 2. Normal Group.

2.2. Morphological analysis

A morpheme is the smallest character string which becomes meaningless when it is decomposed further. Decomposing sentences into morphemes and specifying their part of speech is called morphological analysis. In this research, MeCab which is published as a morphological analyzer developed by Kudo is used [5]. MeCab has been developed with a general-purpose that does not depend on dictionaries and text data, and it is faster than other analyzers. From the analyzed morphemes, words and phrases of the following items which should be noise of analysis are omitted.

- Name of consulting service
- Fixed phrase
- Words not making sense

2.3. Correspondence analysis

Correspondence analysis is a method proposed in 1960s by French researcher Benzeccri and is a method to compress information contained in rows and columns of data table into a small number of components [6]. In this research, words are included in line items (sample), and company names are included in column items (category). Two-dimensional data of the appearance count t_{ij} of the word j of the company i are targeted. Table 1 shows an example of two-dimensional data.

Table 1 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}

Variables are set for samples and categories, and calculations are performed to maximize the correlation between these variables. 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. Meanwhile, general companies and words in the target data appear near the origin. In this study, correspondence analysis is carried out for each group divided by presence / absence of fraud problem detection, and attention is paid to the word close to the origin. Words which become factors of each group are extracted by the method shown in the next section.

2.4. Extract words to be factor

From the result of correspondence analysis, considering all dimensions, the distance d_{oG} between word i and the origin in each group (G_1, G_2) is calculated by Eq. (1). D_G is the total number of dimensions, x_{ijG} is the sample score of word i in dimension j, and C_{jG} is the contribution of dimension j.

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

For each word, d value is updated by the operations of Eq. (2) and (3).