Selection of Core Words from Textual Patent Data with DEA based on Citation

1st Shigeaki Onoda
Dept. Information Systems Engineering
Toyama Prefectural University
Kurokawa, Imizu, Toyama, Japan
t855005@st.pu-toyama.ac.jp

2nd Mamoru Ota
Dept. Information Systems Engineering
Toyama Prefectural University
Kurokawa, Imizu, Toyama, Japan
ota@pu-toyama.ac.jp

3rd Koji Okuhara
Dept. Information Systems Engineering
Toyama Prefectural University
Kurokawa, Imizu, Toyama, Japan
okuhara@pu-toyama.ac.jp

***Abstract*—The web includes enormous data such as patents. The purpose of this research finds the rule of textual patent data and creates new model. Hence, we suggest new weighted method using DEA to handle unstructured data like patent. Our Proposed method is advantageous because this considers the value of the patent compared with TF-IDF and other weighted methods. Using suggested method, we probe new text-mining in the field of patent.**

***Index Terms*—patent, big data, DEA, natural language processing**

I. Introduction

The need for information in companies or public sectors is increasing. In Ministry of Public Management in Japan, 1) the transparency, 2) reliability and 3) accessibility are as an agenda for boosting Open-Data strategy [1].

For example, some agencies conduct making kinds of hazzard map as API [2]. A wide variety of ICT field is promoted by science and technology, the field of patent is also focused on. Japan Patent Office releases a API service named “Patent Information Platform” which includes patent documents or utility models. Anyone can search any patent in Japan. Using ICT technology or machine learning, we can analyze and survey efficiently than before.

In addition, the AI boom all over the world intrigue scientist and engineer, and there are some application examples of AI technology in the field of the patent. “Doloitte Analytics” based on PLSA powered by Deloitte Tohmatsu Consulting can be considered as an implementing survey tool.

In Japan, FRONTEO and TTDC (Toyota Techno Development) developed a tool called “KIBIT” [3]. However, unstructured data like patent data have no absolute metrics yet.

The research of the patent may be divided into two type as follows; an analysis type which aims at making a patent map and a reasoning type which creates new technology area. Patent map is a visualization of unstructured Data. Patent map is a resource of patent research, and it aims at visualization, arrangement and extraction for human [4].

As an analysis type, Kim adopted k-means to classify some categories from words of a target technology field [5]. They built semantic network from clustered output to visualize.

There is another method by Ota et al. It used as input a small number of patents in the patent field to be investigated, selected a sample matching the object from a set of patents by using a RF (random forest) for it, extracted a task expression by NLP (Natural Language Processing on the character data to create a patent map [4].

Sakai et al., suggested new approach to extract a key phrase in the field of the patent based on NLP methods. They discovered that a phrase includes “can” has important expression in the field of patent [6].

However, This type of research merely consider word in documents and it doesn’t consider patent data such as citation number, cited number and inventor. Thus, Tsumura et al., conducted a RF learning to patent categories as a reasoning type to make a mapped space includes patent category and the term frequency [7].

In this research, we apply DEA (Data Dnvelopment Analysis) to some patent data to evaluate statically unstructured data like patent, and we suggest a weighted model with value. In addition, our model consider the frequency of words and patent value so as to multi-modal method suggested by Tsumura et al.

A research applied DEA in the field of patent is suggested by Seol et al [8]. They aimed at finding similarity among industry area and DMU (Decision Making Unit) in their research, while our purpose is to get a rule among word information and citation. Therefore, we try to probe core-words in the field of patent mapped from that rule.

This research makes an advantage in that it doesn’t require technical knowledge about the field of patent to user. This is because we don’t adopt supervised learning like Tsumura et al.

Concretely, we build patent evaluation model regard each patent as a firm. Results of DEA supply us to weight of each word from statical analysis such as arithmetic mean. It is possible to find the core words in this methods. Thus, this study explores capability and possibility of multi-modal text-mining.

At first, this paper explain fundamental law of DEA, and its visualized method in Section II. Next, we show the way to collect patent data in Section III. In Section IV, there are experimental results and discussion. At last Section V, we put conclusion.

II. Methodology

A. DEA

DEA is a nonparametric method coined by Charnes and Cooper [9] and it is used to apply unit's evaluation of absolute scale [10].

DEA can measure productivity of DMU in an organization as a target by using promotion of profits and assets. You can calculate efficiency of DMU as $(\text{sales/employee number})$ when you know sales and employee number of your shop.

However, the real world has not only 2 parameters but many parameters, and some parameteres such as the number of visitors or quantity of stock combined complicated. DEA is a better solution if a problem doesn't have absolute scale.

DEA makes a hint which resource is deficit compered with efficient DMU and inefficient one. Solving efficiency θ_k of a branch named k is the following FP (Fractional Liner Problem).

$$\begin{aligned} \text{max :} \quad & \theta_k = \frac{\sum_{n=1}^{N} v_{kn} y_{kn}}{\sum_{m=1}^{M} u_{km} x_{km}} \\ \text{subject to :} \quad & \frac{\sum_{n=1}^{N} v_{kn} y_{sn}}{\sum_{m=1}^{M} u_{km} x_{sm}} \leq 1 \ (s = 1, 2, 3, ..., K) \\ & u_{km} \geq 0 \ (m = 1, 2, 3, .., M) \\ & v_{kn} \geq 0 \ (n = 1, 2, 3, .., N) \end{aligned} \tag{1}$$

K is the number of DMUs, x and y show inputs and outputs. M and N show the number of input variable and the output one respectively. u_{km} and v_{kn} show each weight of direction, so s is a number of target DMU. The FP (1) shows multi-input/multi-output, weighted DMU to maximize efficiency for k and efficiency of DMU_k is calculated from max efficiency (DMU_k promotion of multi-input and multi-output).

In other words, each DMU advantageously evaluates itself and the smaller input is better in FP (1).

To solve FP (1) as LP, we convert a denominator or a numerator to scalar number "1". This method is called as CCR. It is considered as a practical method to solve DEA [10]. We explain applied CCR model in this research.

B. CCR Model

CCR model has 2 types for this: input-oriented model and output-oriented model. That is depends on whether we set a denominator or a numerator to scalar number "1". Input-oriented model is shown as LP (2), while output-oriented model is shown as LP (3).

Input-oriented model can suggest improvement plan for output because its maximize efficiency equal to its maximize multi-output. By contrast, output-oriented model can suggest improvement plan for input because its maximize efficiency equal to its maximize multi-input. Which model would be better is up to the situation.

$$\begin{aligned} \text{max :} \quad & y_k = \textstyle\sum_{n=1}^{N} v_{kn} y_{kn} \\ \text{subject to :} \quad & \textstyle\sum_{m=1}^{M} u_{km} x_{sm} - \sum_{n=1}^{N} v_{kn} y_{sn} \geq 0 \\ & \textstyle\sum_{m=1}^{M} u_{km} x_{km} = 1 \ (s = 1, 2, 3, ..., K) \\ & u_{km} \geq 0 \ (m = 1, 2, 3, ..., M) \\ & v_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \end{aligned} \tag{2}$$

$$\begin{aligned} \text{min :} \quad & x_k = \textstyle\sum_{m=1}^{M} u_k x_{km} \\ \text{subject to :} \quad & \textstyle\sum_{m=1}^{M} u_{km} x_{sm} - \sum_{n=1}^{N} v_{kn} y_{sn} \geq 0 \\ & \textstyle\sum_{n=1}^{N} v_{kn} y_{kn} = 1 \ (s = 1, 2, 3, ..., K) \\ & u_{km} \geq 0 \ (m = 1, 2, 3, ..., M) \\ & v_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \end{aligned} \tag{3}$$

Fig. 1 shows the concepts of DEA. The frontier of efficiency in this figure is bordered of efficiency or inefficiency. The closer DMU to this line is more efficient DMU and efficiency of DMUs on the line takes maximum 1.

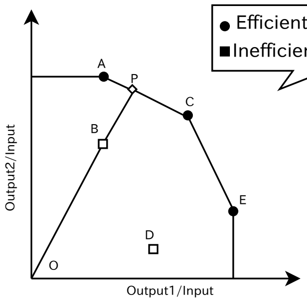

Fig. 1 Concept figure of DEA

To apply DEA, DMUs which take 1 are efficient and other DMUs are inefficient. the existence of inefficient DMU_k is composed of dominant DMUs. For example, point B in Fig. 1 references A and C. That means the existence of A and C judge B as inefficient DMU. To fix inefficient one, you improve input resource as DMU_B be more close to point P.

Reference promotion of a reference set for DMU_k in DMUs is gotten to solve the dual problem of CCR model. LP (4) is dual problem the case of input-oriented, while LP (5) is dual problem of output-oriented model.

$$\begin{aligned} \text{min :} \quad & \theta_k \\ \text{subject to :} \quad & \lambda_s \geq 0 \ (s = 1, 2, 3, .., K) \\ & \theta_k x_{km} - \textstyle\sum_{s=1}^{K} \lambda_s x_{sm} \geq 0 \ (m = 1, 2, 3, ..., M) \\ & \textstyle\sum_{s=1}^{K} \lambda_s y_{sn} - y_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \end{aligned} \tag{4}$$

$$\begin{aligned} \text{min :} \quad & \theta_k \\ \text{subject to :} \quad & \lambda_s \geq 0 \ (s = 1, 2, 3, .., K) \\ & \textstyle\sum_{s=1}^{K} \lambda_s y_{sn} - \theta_k y_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \\ & x_{km} - \textstyle\sum_{s=1}^{K} \lambda_s x_{sm} \geq 0 \ (m = 1, 2, 3, ..., M) \end{aligned} \tag{5}$$

where λ_s shows reference rate of DMU_s for DMU_k. If DMU has no reference, λ_s takes 0. λ is used in the next section to visualize.

C. Visualization of DEA

In a DEA problem, we can visualize like Fig. 1 in case of a few parameters, but it is difficult for human to draw and realize DEA results in case of multi input and multi output. In that case, correspondence analysis is useful to visualize [11].

Correspondence analysis is a descriptive method analyzing table include rows and columns criteria [12]. This sorts element between rows and columns so as to maximize correlation between each row and column, displays table formed data to 2-dim space visually. This result of it is caused by structure or similarity of each category [13].

Toyozumi at el., solved efficiency θ_k and reference set by using DEA to make visible. Table 1 shows cross-tabulation from DEA λ. The row items of table include efficient DMU_{n1}, and columns include inefficient DMU_{n2}. Here, $n1 + n2 = n$, and n show the number of target DMU.

λ equals to λ_s in LPs (4) and (5) which is the element of the reference set. We took this methodology for visualization.

TABLE I
SUBSTITUTION TABLE TO CORRESPONDENCE ANALYSIS

		Efficient DMU		
		DMU_7	...	DMU_{n1}
Inefficient DMU	DMU_1	$\lambda_{1,7}$	...	$\lambda_{1,n1}$
	DMU_2	$\lambda_{2,7}$	...	$\lambda_{2,n1}$
	.	.	...	.
	DMU_{n2}	$\lambda_{n2,7}$	...	$\lambda_{n2,n1}$

III. COLLECT PATENT DATA FROM OPEN DATA

A. Collecting Method

The noted open data such as J-Platform is sufficient to search a topic manually, but it is difficult for human to browse whole patent data as big data. There are some problem for example it adopts PDF as a format, and it doesn't synthesize patent documents and the number of cited. It is not organized well.

Google Patents browse html format, it is easier for gathering data than unstructured data like PDF. In this study, we use Patents - Google[1] as a resource to analysis patent steadily.

Moreover, other platform powered by Google is Google Patents[2]. This has original domain, and is difficult for Patents -Google in interfaces and search outputs.

Firstly, we have collected some data which includes 1) patent, 2) inventor, 3) title, 4) grant date, 5) citation number, 6) cited by number, 7) words in the patent documentation.

Moreover, I also extract only noun from the documents as well as Tsumura [7]. We picked up Japanese patents which is ease to understand for us.

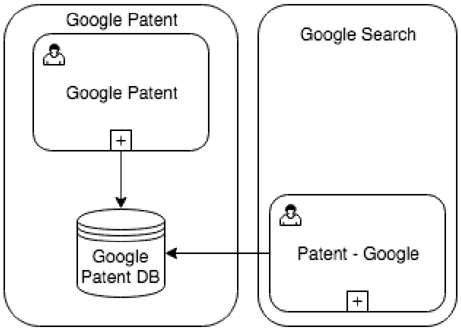

Fig. 2 Patent Search Platforms powered by Google

B. DataBase

I have to construct a data base to analyze. Collection data could divide 7 types, and each word datum is different among sorts of words. I apply NoSQL as DB because of scalability of it.

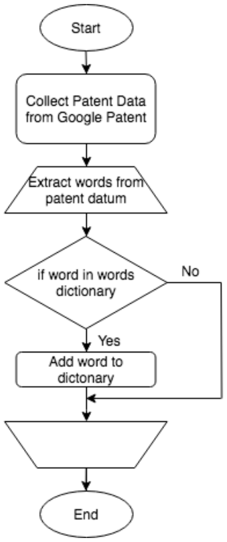
Start
Collect Patent Data from Google Patent
Extract words from patent datum
if word in words dictionary
No
Yes
Add word to dictonary
End

Fig. 3 Process of making word dictionary

Fig. 3 depicts the process of making word dictionary. The dictionary has word and its frequency. First, we use Google Patents as target platform. We save the noted 7 information after we extract them.

Then, we constructed DataBase. Next, we search word data recursively for it and put data to the dictionary.

IV. NUMERICAL EXPERIMENTS

A. Applying DEA model

Applying the DEA method to patent data is studied by Seol et al [8]. They found a new business area from patent data, but we will shave the input data to word frequency matrix. That can predict a valued patent and valued word in patent. This method aims at finding core words in the domain "patent". Next, we will show concrete rule to apply DEA.

This study assigns the words included in patents and the frequency to multi-input for DEA, while assigns citation number and cited by number to multi-output. It is considered by Goto that citation information is important in patent data,

[1] https://www.google.co.jp/?tbm=pts

[2] https://patents.google.com

and we adopt that [14]. That means, the more cited the number, the more valuable a patent.

Table 2 is a conceptual table of input / output for DEA. This table expresses a word frequency f_{ki} of w_i with each Patent_k as a input.

Normally, DEA defines that small multi-input and large multi-output are more efficient than large, and vice versa. However, it is not clear that the rule is eligible for analysis because this study use words as inputs. Thus, we make two rules and apply it to DEA.

- **Rule1:** we use words: w_i in patent i as an input, and input sets scalar Z if no word: w_i occur in patent i
- **Rule2:** we set frequency distribution as inputs. It is called Plain input.

Rule 1 follows an analogy in that small input is better, and we use it to construct high value patent model. Moreover, the model enables us to predict a word distribution in a good patent. Rule 2 is based on Luhn's model, which define a high frequency word is more important than a low one [15].

Specifically, it is a model that evaluates it as efficient if it is a larger output. This is opposite to an original DEA model, that is considered as higher efficiency as input is less and output is larger. We conduct Rule 2 to compare with Rule 1. Applying 2 rule enable us to find feature of words between 2 rules.

We assign input to 1 with Rule 1, or Z with Rule 2. Besides, Z is scale number 2 in this study and this value is from word dictionary in patents.

We apply input-oriented model of DEA for validation. As a target category, we chose G06 (COMPUTING; CALCULATING; COUNTING) and H05B (ELECTRIC HEATING; ELECTRIC LIGHTING NOT OTHERWISE PROVIDED FOR), then we analyzed them with 2 rules and uncompressed or compression version applying PCA (Principal Component Analysis) to reduce input dimensions. In short, there are 4 pattern environments.

TABLE II
DEA MODEL OF THIS STUDY

	inputs			outputs	
No.	w_1	...	w_i	citation	cited by
Patent_1	f_{11}	...	f_{1i}	c_1	cb_1
Patent_2	f_{21}	...	f_{2i}	c_2	cb_2
.	.	...	.	.	.
Patent_k	f_{k1}	...	f_{ki}	c_k	cb_k

B. Experimental Result and Discussion

First, we apply DEA to G06 field. G06 which is one of IPC (International Patent Classification) show information field. Table 3 shows Rule1 and Rule2. This field has 10873 words in 94 articles. The arithmetic mean of this field is almost 1.

The reason is an input dimension is too sparse to make difference among each article. For this input data, we apply PCA to compress. It is bigger S.D (Standard Deviation) than previous one.

We consider that the smaller dimensions are, the larger S.D is from the mentioned environments. Next, we select H05B field which has 74 articles and 6252 words as resources.

As well as G06 field, almost DMU is evaluated as efficient one. So that its arithmetic mean is close to 1. We didn't find difference among each DMU. However, S.D in H05B field is larger than G06 field, and the mean is smaller.

A result of doing dimensional reduction is shown in Table. 3. It is not more ambiguous than before. Furthermore, all efficiency distribution is almost similar. Rule 2 affects S.D and mean so as to be small. We saw a large S.D case from the PCA results.

TABLE III
MEAN AND STANDARD DEVIATION OF EACH RULE AND EACH THE FIELD OF PATENT

	uncompressed		compressed	
No.	mean	s.d	mean	s.d
G06 with Rule1	0.976	0.147	0.905	0.228
G06 with Rule2	0.961	0.176	0.896	0.225
H05B with Rule1	0.973	0.164	0.812	0.307
H05B with Rule2	0.949	0.191	0.767	0.326

Table 4 and 5 show top 10 words in each weight of DMU. Table 4 based on Rule 1 seemed to be having popular words in this field while table 5 based on Rule 2 seemed to be having niche words. Both rules's inputs are almost Z (Rule 1: 0, Rule 2: 1) because the number of words is larger than the number of DMUs. First, we show Rule 1 result in table 4.

Rule 1 include low frequency words. Especially, the word "anthracene" only occurred three times. It is considered a core word in H05B field, and we call a word like this "E-word". Namely, words in which the input (frequency of word) is enough small and what good output (citation and cited by number) is available are weighted highly as we expected.

However, the result include noisy word like "Unknown (pronounced Sa in Japanese)" due to pre-processing for example stemming and morphological analysis.

Next we show Rule 2 result in table 5. Results with Rule 1 include popular words such as "invention", "figure" and "coil". This means high frequency word is highly weighted against Rule 1. Thus, it is seemed to extract general (it call "G-word") word in the patent field.

Comparatively speaking between table 1 and 2, we find the mean of "anthracene" or "aluminium" is 0 with Rule 2.

By contrast, the mean of highly weighted words like "invention" and "figure" is respectively 0.020 and 0 with Rule 1. We observed a word in which both models evaluate high weight, but couldn't find a special one.

This result indicates each rule can extract E-word and G-word.

Fig. 8 and Fig. 9 show the visualization of the last noted results by using Toyozumi method. It maps each position in each patent from DEA results.

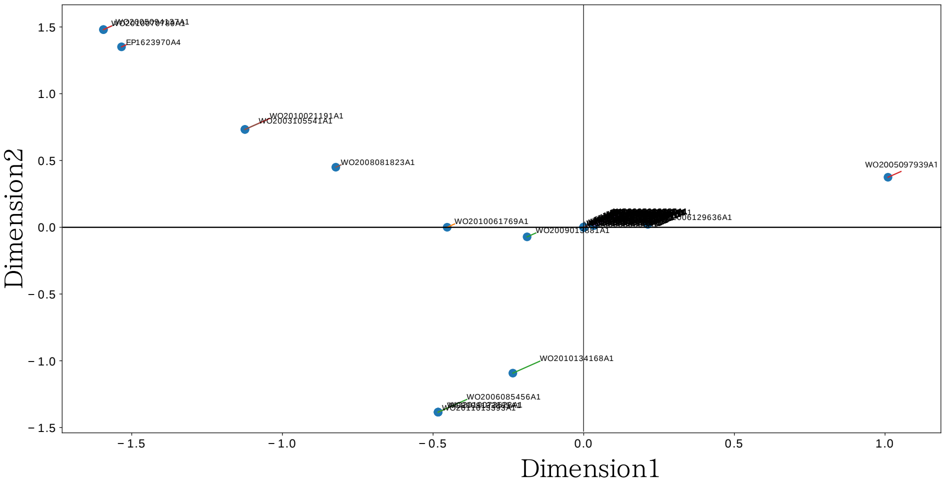

Fig. 8 Visualized Efficient DMU on Rule2

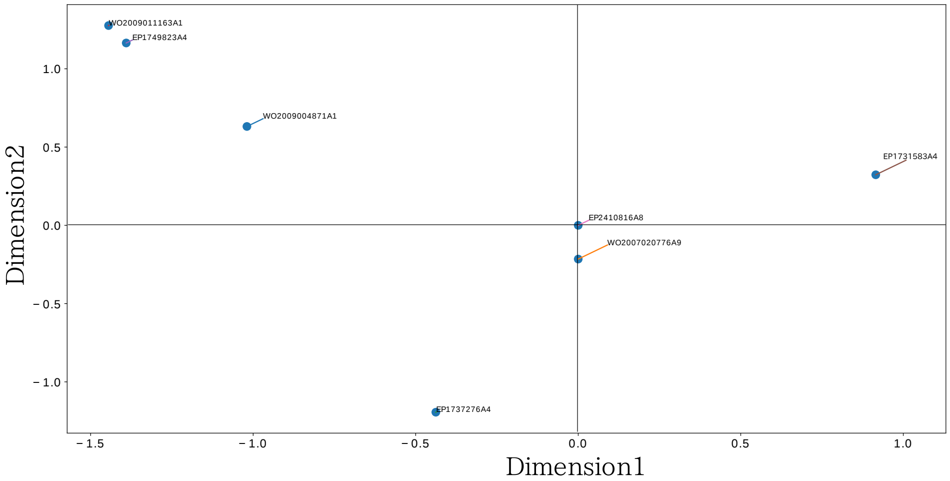

Fig. 9 Visualized Inefficient DMU on Rule2

For instance lower-left cluster elements (WO2005094137A1 or WO2010070789A1) in Fig. 7 belong to H05B41 (Circuit arrangements or apparatus for igniting or operating discharge lamps).

This category belongs to the subclassification of H05B. The center cluster (WO2008081823A1 or WO2010134168A1) is seemed to be organic EL category, and which category belongs to H05B33 (Electroluminescent light sources).

This visualization enables us to catch a certain tendency in the field of patent. The patents of the origin (0, 0) failed to correlate well because the reference set of its DMU was 0. This occurs when input data is occupied by same value Z.

V. Conclusion

A. Interpretation of Results/Discussion

This study made a evaluation model of the patent area using the DEA model. Thanks to that, we will conduct a trend analysis in a technological field. This method is different from conventional methods which use supervised learning by specialists, in that it enables people who doesn't have expertise of patent to excute.

From the results of the evaluation, noted 2 rules are selected depending on what problem you have. For example, Rule 1 is better to extract terminology in a target field, while Rule 2 is better to extract a word sounds patent.

TABLE IV
TOP 10 HIGH WEIGHTED WORD (RULE 1)

Word	Weight	Frequency
アントラセン (anthracene)	5.855055	3
下方 (belowdown)	2.343041	26
クラック (crack)	1.785240	10
さ (unknown)	1.721771	117
光 (light)	1.603364	948
アルミニウム (aluminium)	1.405620	33
システム (system)	1.345943	70
エネルギー (energy)	1.070877	31
アノード (anode)	0.999145	29
お呼び (invitation)	0.920917	218

TABLE V
TOP 10 HIGH WEIGHTED WORD (RULE2)

Word	Weight	Frequency
図 (figure)	0.656393	2404
発明 (invention)	0.047136	1545
コイル (coil)	0.226496	108
実施 (enforcement)	0.111100	1188
調 (style)	0.055942	348
所定 (stated)	0.051680	414
回路 (circuit)	0.047136	1545
層 (layer)	0.046728	1694
チューブ (tube)	0.044257	8
原子 (atom)	0.040127	21

When we deal the number of input-dimensions as the number of a word sort, the result attributes sparse matrix like most of efficiency is 1. Therefore, we should reduce dimensions of inputs using not PCA but NMF (Non-negative Matrix Factorization) or LDA (Latent Dirichlet Allocation). Those methods can analyze each word because of the restorability.

The last, this study will help us to do a multi-modal learning includes words information and value information like the field of academic thesis. We consider this method useful in NLP or text mining tasks using weights of words from this study. Moreover, to visualize them helps engineer and manger to decide something.

B. Future Works

As a subject in the future, we would like to increase the number of patents to be acquired and further verify the behavior when reducing the ratio of input dimension number and DMU number, and analyze the whole patent document without limiting the field. Also, we conduct a extraction of terminology or condense dimension of word vector by NMF method to decrease computational complexity.

REFERENCES

[1] Ministry of Public Management, “Boost Open Data Strategy,” URL: http://www.soumu.go.jp /menu_seisaku/ictseisaku/ictriyou/opendata/, Accessed: May 5, 2018.

[2] NTT Communications, “Report of Deploying Information Infrastructure of Hazzard Data,” No. 1.0, 2013 URL:http:www.soumu.go.jpmain_content000317915.pdf.

[3] T. Kiriyama, T. Ando, “Patent Information and AI: Outline,” *Information Science and Technology*, Vol. 67, No. 7, pp. 340-349, 2017.

[4] T. Ota, “Patent Research using Machine Learning Methods,” *Information Science and Technology*, Vol. 67, No. 7, pp. 366-371, 2017.

[5] Y. G. Kim, J. H. Suh, S. C. Park, “Visualization of patent analysis for emerging technology,” *Expert Systems with Applications*, vol. 34 pp. 1804
UTF20131812, 2008.

[6] H. Sakai, Hirohumi Nonaka, Shigeru Masuyama, “Extraction of Information on the Technical Effect from a Patent Document,” *Journal of the Japan Society for Artificial Intelligence*, Vol. 24, No. 6, *I*, pp. 531-540, 2009.

[7] T. Tsumura, F. Saito, S. Ishizu, “Knowledge Extraction from Textual Patent Data Using a Random Forest,” *J Jpn Ind Manage Assoc*, vol. 68, No. 3, pp. 161-170, 2017.

[8] H. Seol, S. Lee, C. Kim, “Identifying new business areas using patent informatin: A DEA and text mining approach,” *Expert Systems with Applications*, vol. 38, pp. 2933-2941, 2011.

[9] A. Charnes, W. W. Cooper, E. Rhodes, “Measuring the Efficiency of Decision Making Units,” *European Journal of Operational Research*, Vol. 2, , pp. 429-444, 1978.

[10] T. Sueyoshi, “DEA: Business Analysis method,” Asakura, 2001.

[11] K. Toyozumi，Y. Nishiuchi，S. Aoki，H. Tsuji，“Correspondense Analysis based Position Visualization for DEA,” SCI’06, No. 50, Session ID: 2W1-2, 2006.

[12] K.Yamashita，H. Li，“Mining Open A nswer in Questionnaire Data,” *IEEE Intelligent Systems*，Sept./Oct.，pp.58-63, 2002.

[13] “How To Analyze Simple Two-Way and Multi-Way Table, Correspondence Analysis,” URL:http://www.statsoft.com/Textbook/Correspondence-Analysis, Accessed Jun 1 2018.

[14] A. Goto, K. Genba, Jun Suzuki, Schumpeter Tamada, “Classification Index of Important Patentrq,” *RIETI Discussion Paper Series*, 06-J-018, pp. 1-17, 2018.

[15] H. P. Luhn, “The Automatic Creation of Literature Abstracts,” *IBM Journal of Research and Development*, vol. 2, No. 2, pp. 159-165, 1958.