*International Journal of Latest Research in Engineering and Technology (IJLRET)*
*ISSN: 2454-5031*
*www.ijlret.com || Volume 05 - Issue 02 || February 2019 || PP. 34-42*

# Recommendation of Data Analysis Method by Vectorizing Text

Takahisa Oe[1], Humiko Harada[2], Hiromitsu Shimakawa[3]
[1](*Graduate School of Information Science and Engineering, Ritsumeikan University, Japan)*
[2](*Connect Dot LTD, Japan)*
[2](*College of Information Science and Engineering, Ritsumeikan University, Japan)*

**Abstract:** In recent years, demand for data analysis has increased in companies and others, but there are limited people who can handle data analysis method correctly. The data analytical method that should be applied differs according to the problems the company has. Therefore, in this research, problems to be solved are expressed as texts as scenarios. It is proposed that method to recommend a data analysis method to adapt by vectorizing the scenario. In vectorization, by detecting a sentence expressing the significant of the scenario, it is increased that the weight of the sentence closer to it and excluded that sentences unnecessary for recommendation such as background knowledge. Experimental results show that the usefulness of the proposed method is different depending on the recommended data analysis method. If this method is applied, it is possible to provide an environment where new learners narrow down the learning field and learn efficiently.
**Keywords:** Data Analysis, Recommendation, Learning Support, Text Mining, Vectorization

## I. INTRODUCTION

Recently, Industry 5.0 that makes the production process more efficient is drawing attention[1].Even in Japan, it is required that worker reforms to advance more to produce more value in a short time[2].In order to improve work efficiency, many people want to analyze their own problems.

With the improvement of computer hardware and software capabilities, data storage has become easy. In order to solve the problem, the company accumulates data related to the problem. To obtain value from the accumulated data, it is analyzed. Analysts can easily analyze data using analysis tools available for free. By studying the data analysis method and understanding the mechanism and the theory, the analyst can obtain various merits[3].For example, they can modify the production model for each purpose, and take appropriate measures when the production efficiency is not good. Conversely, if analysts do not understand the mechanism and theory of analysis, they may make a wrong analysis. For example, they may select models that are inappropriate for the problem and apply incorrect countermeasures to the production site.

In order to solve the problems that it owns, data analysis is a skill that must be worn by every and all people, not only engineers, but also salespersons and personnel managers, in modern society. However, in order to learn data analysis, we need a lot of knowledge of mathematics such as vector, differential, matrix, statistics and probability. Many people are not proficient in knowledge of mathematics and it is very difficult for beginner in data analysis to study them all from the beginning.

In this paper, we focus on the fact that the data analysis method is different depending on the kind of problem to be analyzed. Mathematical knowledge to learn is different if analytical methods are different. Therefore, in this research, we identify the type of problems the users have with less effort. Therefore, this paper aims at narrowing down the mathematical knowledge to be learned by data analysis according to user's problem.

## II. NARROW DOWN THE DATA ANALYSIS METHOD ACCORDING TO THE THEME

### 2.1. Significance of narrowing down the field of study

In this paper, in order to lower the learning hurdle of data analysis, think about narrowing down the learning field by extracting the data analysis method suited to the theme. As one of the obstacles to learning, it is conceivable that there are many items to be learned, and it is in a situation where it is not known what to work on. This paper aims to lower the hurdles of such data analysis learning by narrowing down the fields to be learned.

In addition, narrowing down the field of study allows learners to clarify the goals of learning. To clarify learning goals by learners improves learning efficiency [4].

### 2.2. Existing efforts to narrow down the learning field

In order to narrow down the learning field of data analysis, many reference books and Web pages introduce data analysis method by pattern of problem to be analyzed. In this method, it is necessary to embody the problems that the learners want to analyze to the same level as the patterns introduced in the reference books

and the Web pages. Also, with this method, it is not known whether the problems the learners want to analyze is compatible with the introduced patterns.

It is necessary to have a mechanism to connect the problems learners want to analyze with the data analysis methods. There are conceivable ways to combine problems that the learners want to analyze and examples of data analysis methods introduced in reference books and Web pages by way of Topic Modeling[5] or Doc2Vec[6].However, there are many problems that can be applied to one data analysis method. In Topic Modeling and Doc2Vec, these are measuring semantic similarity of sentences, are not appropriate for as methods for linking data analysis methods and problems learners want to analyze.

Topi said the world's interest in data science is strong and creating a new educational program[7].Many attempts have been made, but not enough to educate beginner in data analysis[8][9].

From the above, it is necessary to have a mechanism for recommending data analysis methods tailored to the problems learners want to analyze.

### 2.3. Difference between free description and choice

In this paper, it is considered the mechanism which the system recommends data analysis methods using free descriptions described what learners want to analyze. Methods using items can not cover all patterns of problems. Also, the methods using items are hard to use without the learners can organize the problem to the same concreteness as the items. By describing the content that the learners want to analyze with free description, it is possible to directly deal with the problem that the learners themselves possesses. In addition, learners can describe the contents that they want to analyze by free description so that learners can organize the contents themselves.

It is possible to create an identification model by extracting feature words from text. In this paper, it is proposed the method of creating a identification model by extracting feature words from the content written by learners.

## III. IDENTIFICATION MODEL FOR DATA ANALYSIS METHODS RECOMMENDATION

This paper proposes an identification model for recommending data analysis methods suitable for user's purpose.

### 3.1. Recommendation in significant sentence strengthening method

In this paper, a sentence written about contents which wants to be analyzed is called a scenario. The significant sentence is a sentence within the scenario and shows how the user wants to analyzed the data. It is a condition of the scenario that one or more significant sentences are included. Generally, in a well-written scenario, the problem background is explained first, and the problem is defined using the terms described therein. Then, analysis method of data is presented, and how to solve the prescribed problem is described as significant sentences. Therefore, in this paper, it is assumed that the closer the sentence is to the significant sentences, the more words necessary to define the data analysis that the scenario wants to solve are contained.

In addition, in this paper, it is assumed that words far from data analysis appear, such as explanations of background knowledge for significant sentences, the more distant sentences are in the significant sentences.

In the identification model, a scenario is an explanatory variable, and a method optimal for analyzing a problem written in the scenario is a target variable. An overview of the identification model is shown in the figure 1.

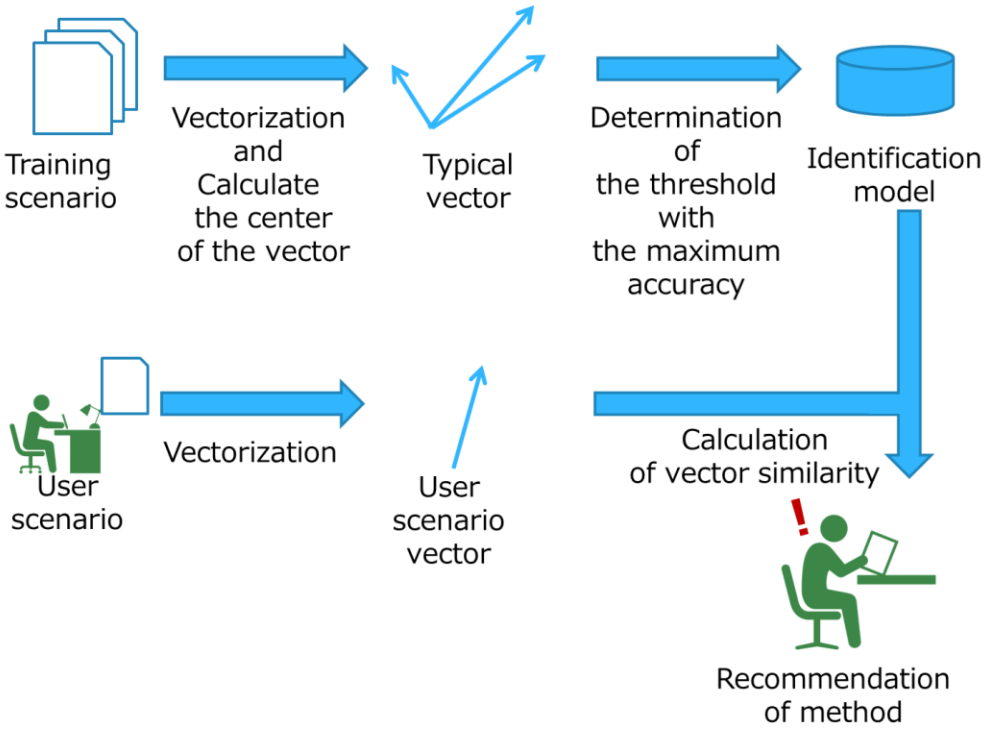


Figure 1: Identification model based on similarity in significant sentence strengthening method

There are zero or more data analysis methods for realizing the contents written in the scenario. In this paper, the scenario written by the user is called the user scenario. The identification model recommends zero or more data analysis methods for one user scenario.

On the other hand, in this paper, the scenario for training the model is called the training scenario. Prepare multiple sets of training scenario and data analysis method adapted to it. Using these, the discrimination model is trained so that the accuracy of discrimination is the greatest.

The flow of the method is shown below.

1. The method vectorizes training scenarios.
2. The method creates a typical vectors as a guide for each method of the recommended items.
3. The method obtains a threshold value that maximizes the accuracy of associating the training scenario with the corresponding analysis method.
4. The system vectorizes the user scenario and recommends a method whose degree of similarity with the typical vector is equal to or higher than the threshold value.

**3.2. Vectorization of scenarios**

The method of vectorizing scenarios is shown below.

1. The system morphologically analyzes the scenario and extracts some nouns and verbs. The extracted verb is converted into the verb stem.
2. The system generates vector using the nouns and verbs retrieved and significant sentence strengthening method described later.

In the above, some nouns are pronouns which do not express much the characteristics of sentences, nouns other than non-autonomous words and numerals. As described above, the identification model uses some nouns representing specific targets and verbs representing actions and states as elements of vectors.

The significant sentence strengthening method is explained below.

1. The system calculates the distance between the sentence containing the extracted word and the significant sentence strengthening method.
2. The system weights the words according to the calculated distance.
3. The system generates a vector by the TF-IDF method considering the given weight.

In this paper, it is assumed that the more the sentence closer to the significant sentence contains the words necessary to define the data analysis that the scenario wishes to realize. Conversely, it is assumed that words far from data analysis appear, such as explanations of background knowledge for significant sentence, the

more distant sentences are in the significant sentence. Therefore, according to the significant sentence strengthening method, weights are assigned to words according to the distance to the significant sentence.

The distance to the sentence is 1, with the sentence adjacent to the significant sentence being 1, incremented by 1 each time you leave the significant sentence. However, if the target sentence is a significant sentence, the distance to the significant sentence is 0.If there are multiple significant sentences, the system considers the distance to the significant sentence closest to the target sentence.

The relationship between distance and weight from sentences is shown in the figure 2.

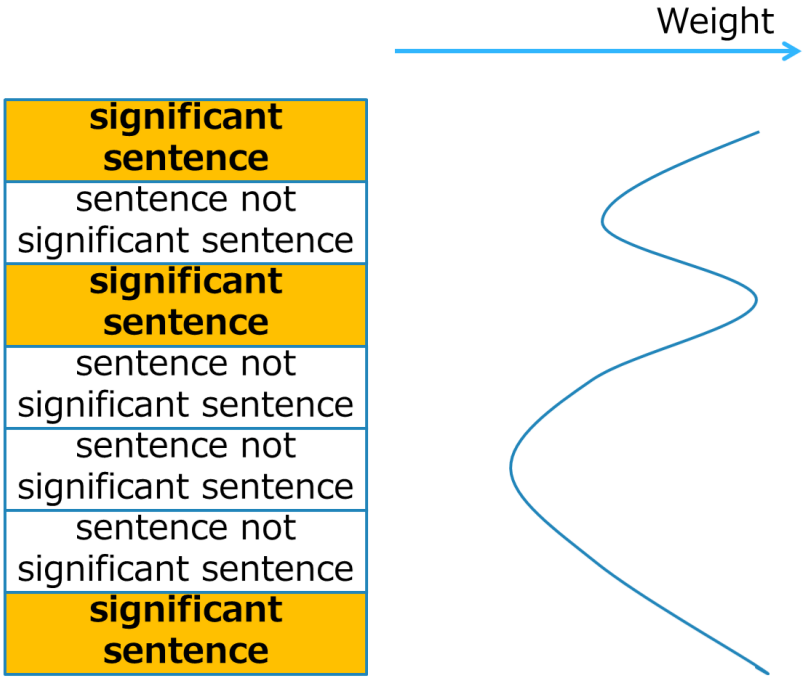


Figure 2: Relationship between distance and weight from sentences

An expression for giving a weight is shown in the formula 1.In formula 1,$W$ is the weight and $d$ is the distance to the significant sentence.

$$W = \frac{1}{d+1} \cdots \text{formula 1}$$

A formula for calculating the value of a vector considering a given weight is shown in the formula 2.$x_t$ indicates the value of a vector for a certain word $t$.$\sum w_t$ represents the sum of the weights of the word t within the target scenario.$s$ is the sum of words extracted from the target scenario.$N$ indicates the total number of scenarios.$\mathrm{df}(t)$ indicates the number of scenarios in which the word$t$appears.

$$x_t = \frac{\sum w_t}{s} \times \left(\log \frac{N}{df(t)} + 1\right) \cdots \text{formula 2}$$

Based on the value obtained by using the formula 2, the significant sentence strengthening method generates a vector of the scenario. In the vector obtained here, the axes indicate each word and the magnitude of the value of the axis indicates the characteristic of the scenario.

### 3.2. The calculation of typical vectors

A typical vector is a vector serving as an index of each analysis method when determining a data analysis method to be recommended using similarity, and it is obtained for each analysis method to be recommended. In the training scenarios, data analysis experts have attached labels indicating whether or not each of the data analysis methods that are recommendation items can be used or not. When calculating a typical vector, the system gathers all the training scenarios that can use the corresponding data analysis method. The system obtains the vector average of the collected training scenarios as a typical vector. A method of creating a typical vector of a regression which is one of data analysis method is shown in figure 3

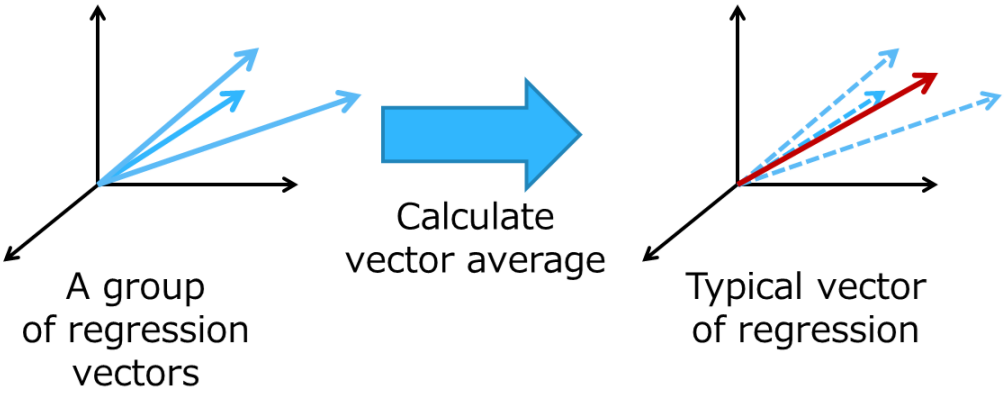


Figure 3: The example of Typical vector creation

**3.4. Determination of threshold**

The system determines the threshold of each typical vector using the typical vectors and the training scenario. The threshold value is used for determining the recommendation of the data analysis method from the similarity between the typical vectors and the vectors of the user scenario.

The method for determining the threshold value is described below.

1. The system calculates the similarity between a typical vector and all training scenarios.
2. The system finds a section of similarity in which the F value in the typical scenario is the maximum from the label on whether or not the data analysis method indicated by the typical vector in each training scenario can be used.
3. The system calculates the median of the intervals of similarity with the maximum F value in the training scenario as a threshold.
4. The system does the above for all typical vectors.

A method for finding the similarity $cos(\vec{q}, \vec{d})$ of the two vectors $\vec{q}$and$\overrightarrow{d}$ is shown in the formula 3.

$$cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} \dots \text{formula 3}$$

**3.5. Determination of recommendation methods**

The method of determining the recommendation method for the user scenario from each typical vector and the threshold value of each typical vector is described below.

1. The system determines the vector of the user scenario.
2. The system obtains the similarity between the vector of the user scenario and each typical vector, respectively.
3. For each degree of similarity, the system recommends all data analysis methods corresponding to items with higher values than the corresponding threshold to the user.

**3.6. Method for detecting significant sentences**

The method of detecting significant sentences is as follows.

1. A data analysis expert manually extracts the significant sentences of the training scenarios.
2. The system vectorizes the significant sentences of the training scenarios with TF-IDF. At this time, the system does not weight by the significant sentence strengthening method.
3. The system calculates the average of the vectors obtained above. The obtained vector is called a significant sentence vector.
4. The system vectorizes each sentence of the training scenario with TF-IDF. At this time, the system does not weight by the significant sentence strengthening method.
5. The system calculates the similarity between the vectors of each sentence of the training scenario and the significant sentence vector.
6. The system finds a section of similarity with which the F value of sentence discrimination is maximum, depending on whether each sentence of the training scenario is a summary sentence or not.

7. The system calculates the median value of the section of the degree of similarity with which the F value of the gist significant sentence discrimination is maximum as the threshold value of the gist of significant sentence vector.
8. The system vectorizes each sentence of the user scenario with TF-IDF. At this time, the system does not weight by the significant sentence strengthening method.
9. The system calculates the similarity between the vector of each sentence of the user scenario and the significant sentence vector and detects the sentence that is equal to or larger than the threshold as the significant sentence of the user scenario. However, when all the similarities are lower than the threshold value, the system detects the sentence with the highest similarity as the significant sentence of the user scenario.

## IV. EXPERIMENT FOR EVALUATION OF IDENTIFICATION MODEL

This experiment was conducted to investigate the usefulness of recommendation of data analysis methods using the significant sentence strengthening method.

### 4.1. Experimental overview

In this experiment, regression, classification, clustering and reinforcement learning, which are basic data analysis methods[10], are recommended items. Subjects in the experiment wrote one or more scenarios. Subjects are not designated scenario theme. Subjects wrote scenarios of various subjects such as their own familiar problems, social problems and so on.

One of the obtained scenarios is shown below.

"I currently work on creating a talk BOT listening to troubles and bitches. It aims encouraging, comforting, or calming down the irritation, when people are depressed or irritated, in a seminar for graduation researches. It must understand the troubles and complaints sent by users. It should identify nuance and avoid inappropriate words. However, it takes a huge amount of knowledge and time to integrate all of them into software. So, I would like to use machine learning to read the opinions of others and words from existing greats. I want to incorporate them into the software so that we can get better achievements from big data."

Other scenarios were of various themes such as medical, baseball and others.

The subjects were 18 undergraduates in the university, and the number of scenarios obtained was 24.The average number of scenarios written by one subject was 1.33, the most frequent subjects wrote 3 scenarios. The ratio of male to female subjects is male: female = 5: 4.

The experiments were evaluated by cross validation in which the experimental scenario group was divided into three, two of which were set as a training scenario group and one was a user scenario group.

### 4.2. Labeling solutions

The raters labeled solution for the evaluation of the experiment result. The labeling solution was done by seven raters to avoid becoming arbitrary by one rater. For one scenario, five raters randomly selected from seven raters labeled by majority vote.

For each scenario, the items labeled are the sum of the position of the significant sentence (one item, one or more significant sentences decided) and whether each recommendation item is recommended (4 items, boolean type) total 5 items.

## V. EXPERIMENTAL RESULT

### 5.1. Feature word

A word with a high vector element value in the created typical vector by significant sentence strengthening method is called a feature word. The feature words of each recommendation item are shown below.

Feature words of regression

- する- do
- データ- data
- 車両- vehicle
- 線路- rail
- もと- based on
- 新た- new
- ダイヤ- time table
- 支援- support
- ツール- tool
- 拍手- applause
- 野球- baseball

Feature words of classification

- 患者- patient
- 診察- consultation
- 肺- lung
- 音- sound
- 異常- abnormal
- 特徴- feature
- 確認- confirmation
- リーダ- reader
- 本人- principal

- 認証- authentication

Feature words of clustering

- 入力- input
- 多く- many
- 推測- guess
- いる- exist
- 服- clothes

Feature words of reinforcement learning

- 利用- use
- 運転- operation
- れる- to be
- データ- data
- 把握- grasp
- 自動車- car
- 方言- dialect
- 入力- input
- ルート- route
- 指導- guidance
- する- do
- 時刻- time

**5.2. Evaluation of identification model**

Table 1 shows the accuracy of cross validation.

Table1: Accuracy with vector similarity

| Recommended item | The F value which signicant sentence strengthening method is not applied | The F value which signicant sentence strengthening method is applied |
|---|---|---|
| regression | 0.48 | 0.48 |
| classication | 0.45 | 0.78 |
| clustering | 0.33 | 0.41 |
| reinforcement learning | 0.37 | 0.74 |

Among the feature words for classification and reinforcement learning in Section 5.1, the words that are considered to be frequently appearing in each data analysis method are shown below.

Classification

- 異常- abnormal
- 確認- confirmation
- 認証- authentication
- 特徴- feature
- リーダ- reader

Reinforcement learning

- 利用- use
- データ- data
- 入力- input

Among the words other than the above, nouns appear only in one scenario, and the TF-IDF value became high, so it is considered that it became a feature word.

**5.3. Usefulness of identification model**

From the results of chapters 5.2, it can be said that classification and reinforcement learning can be recommended by using vector similarity in significant sentence strengthening method. From the results in Section 5.2, it can be said that classification and reinforcement learning can be recommended by using identification model. However, recommended items should be increased in the future since there is weak empathy when there are few recommended items.

## VI. DISCUSSION

**6.1. Features of erroneous identification scenario**

An example in which correct discrimination was not made in this experiment and its consideration are shown below.

“Currently I am a third year college student, and I have a lot of subjects to report in my classes and I use my computer a lot. When writing a report, it is necessary to write the reference document and the author together, but they are often unpredictable and cannot be converted, especially if it is strong specialist, sometimes struggling to input.”

In this scenario, as a result of labeling solutions by raters, it does not apply to the reinforcement learning, but the model was identified as a reinforcement learning. The erroneous identification of this scenario seems to be caused by "入力" meaning inputting, which is a feature of reinforcement learning, three times in the scenario. In the method of recommending by vector similarity, misjudgment occurs when feature words happen

to appear like this time.

The recommendation with vector similarity is a method in which discrimination depends on the presence or absence of a feature word. However, with this method, when a feature words appear in a scenario that should not be recommended, the similarity with the typical vector is high, so that erroneous discrimination may occur.

### 6.2. Suggestion from features

The following two patterns can be seen in the regression scenario.

- Quantitative expressions are used during the scenario, such as "～値" meaning number of sizes.
- Quantitative expressions are not used in the scenario, but the raters determine that the scenario wants to obtain a quantitative value and is labeled to the regression.

Also, the following two patterns can be seen in the regression scenario.

- Clustering itself is aimed.
- Clustering is an intermediate process and the final service is a identifier.

As described above, there are multiple patterns in regression and clustering, and they are impossible to uniquely determine the feature of typical vectors. Therefore, the recommendation with vector similarity was not useful for the regression and clustering.

Conversely, in this experiment, each data analysis method used the same way of using classification and reinforcement learning scenario. From this point it is suggested that there is bias in the subjects perception of data analysis.

### 6.3. Introduction of method examples

It is estimated that many university students including subjects are accustomed to the problems of supervised learning and reinforcement learning by recent AI boom etc. It was suggested that data analysis method which users are familiar with and easy to perceive can be recommended because classification and reinforcement learning can be said to be distinguishable.

On the other hand, before the labeling solution, the raters were presented with an example of a scenario using each recommended item. As a result, the raters understood the respective data analysis method and carried out correct labeling. From this fact, it is suggested that users can understand the outline of the method by showing simple examples of the methods to the user. From the above, it is thought that by introducing simple examples of methods to the user, letting the scenario written after understanding the outline of the method, it is possible to recommend more methods to the user (Figure 4).

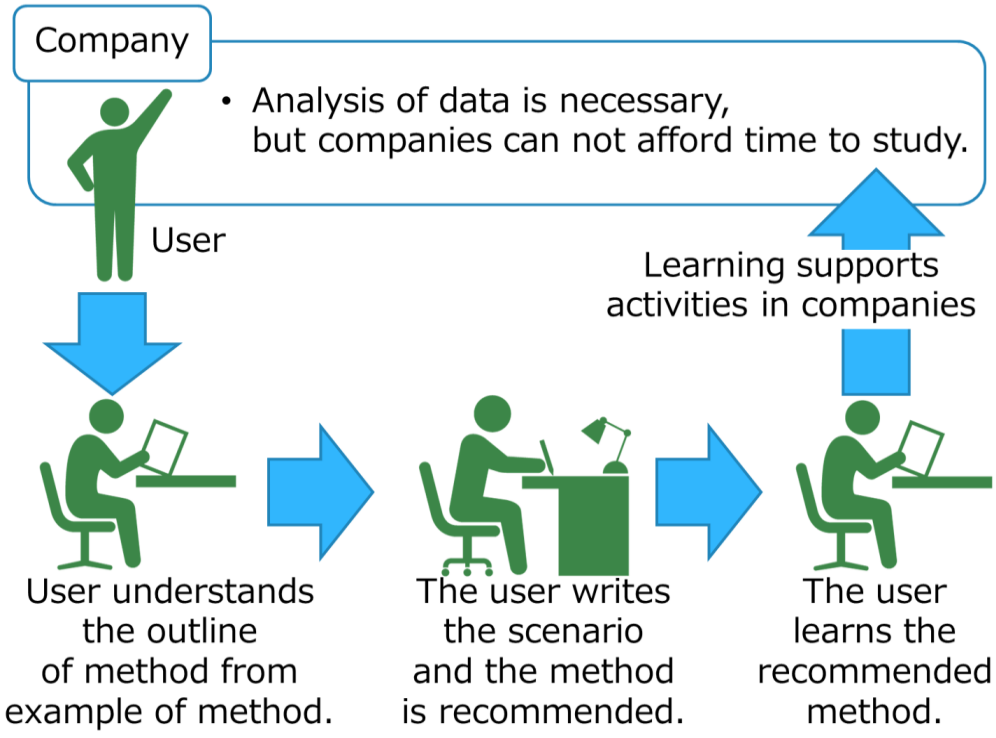


Figure 4: Operation of identification model using example of methods

## VII. CONCLUSION

In this paper, to clarify the purpose of learning and to improve the learning efficiency, a method to recommend the data analysis method suitable for the scenario to the user was proposed. As a result of the experiment, it turned out that it was possible to recommend the classification and reinforcement learning. In addition, it was suggested that beginner in data analysis seize the outline of the method by introduced a simple example of the method. By introducing a simple example of the method and writing the scenario, it can be expected that the accuracy of the recommendation can be improved.

## VIII. ACKNOWLEDGEMENTS

We express our gratitude to all members of Data Engineering Laboratory from Ritsumeikan University who supported us. We gratefully appreciate the help of students of Ritsumeikan University and Kyoto Women's University cooperated with the experiment.

## REFERENCES

[1] IEEE, "INDUSTRY 5.0", 2018, URL: http://sites.ieee.org/futuredirections/
[2] Ministry of Health, Labor and Welfare in Japan, "Towards realization of "Workstyle reform" ", 2018, URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000148322.html [accessed: 2018-01-17].
[3] L. Carine and K Vincent. "“ How Could an Intranet be Like a Friend to Me? ”- Why Standardized UX Scales Don’t Always Fit." Proceedings of the European Conference on Cognitive Ergonomics 2017. pp. 9-16. 2017.
[4] Ministry of Justice Ministry of Japan, "Relationship between learning motivation and learning process", 2015, URL: http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/gijiroku/__icsFiles/afieldfile/2015/09/29/1362371_2_2_4_3.pdf [accessed: 2018-01-17].
[5] D. Ramage, S. T. et al., "Characterizing microblogs with topic models", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp.130-137.2010.
[6] A. Suglia, et al., "A Deep Architecture for Content-based Recommendations Exploiting Recurrent Neural Networks", UMAP 2017 Full Paper. pp.202-211. 2017.
[7] M. Marttila-Kontio, et al., "Advanced Data Analytics Education for Students and Companies", ITiCSE '14. pp.249-254. 2014.
[8] C. Geigle, et al., "CLaDS: A Cloud-Based Virtual Lab for the Delivery of Scalable Hands-On Assignments for Practical Data Science Education", ITiCSE 2018. pp.176-181. 2018.
[9] L. Chen, and A. Dubrawski, "Accelerated Apprenticeship: Teaching Data Science Problem Solving Skills at Scale", L@S '18. 2018.
[10] I. Makoto, "New machine learning textbook learned by moving in Python (written in Japanese)". SHOEISHA, 2017, 408pz