Creating New Value using Mass Text Data
Text Mining for Performance Improvement

KAZUKI OZAKI[*1]　　SHUN YAMAGATA[*2]

HISASHI NISHIKI[*2]　　TAKASHI TAGUCHI[*3]

Text data, such as incidents and their causes relating to incompatibilities that occur during manufacturing, troubleshooting during facility operation, patent information, etc., is stored in general as unstructured data, unlike numerical data which expresses physical quantities, etc., in table form that can be structured. Therefore, the development of text data analysis has been quite slow, hindering its utilization. If such text data could be quantified and structured, classifying sentences and searching for similar ones would be possible by utilizing an analytical approach to numerical data, from which the effective use of text data such as a quick response to newly occurring incidents could be expected. This paper introduces text mining technology and application examples where text data could be quantified and structured.*

* *Structuring is facilitating the handling of large-volume, complex incidents by sorting them into smaller categories in phases.*

1. Introduction

Mitsubishi Heavy Industries, Ltd. (MHI) has been working on predicting future energy demand utilizing data analysis techniques as one of our solutions to respond to the changes in the energy environment that have become increasingly significant due to global environmental issues and power deregulation, and to achieve more efficient energy management. The data subject to analysis here consists of various types of numerical data such as physical quantities measured by various sensors attached to factory equipment, calendar information, weather information and time information linking them together, which is structured data expressed in table form.

Meanwhile, unstructured data including text data is generally thought to account for 80% of all data owned by companies. MHI also owns a large quantity of text data, including measures against incompatibility incidents occurring at the time of manufacture, expertise for efficient facility operation, patent documents, etc., which is unstructured and has seen very little progress in terms of analysis such as evaluating text correlation in a quantitative manner, where its utilization is still a long way off.

MHI successfully quantified and structured such unstructured text data by applying the text mining technology described below where we have achieved clustering (classification) of an enormous number of sentences and searches of similar sentences, in an attempt to extract and utilize useful information from the text data we have collected. This attempt has allowed us to create measures against incompatibility incidents, to offer troubleshooting during facility operation and to achieve increased efficiency in document checking.

2. Analytical procedure utilizing text mining technology

We are going to introduce in 4 steps a method of converting mass text data to numerical data, structuring a huge volume of unstructured text data and performing clustering (classification) and

*1　EPI Department, ICT Solution Headquarters

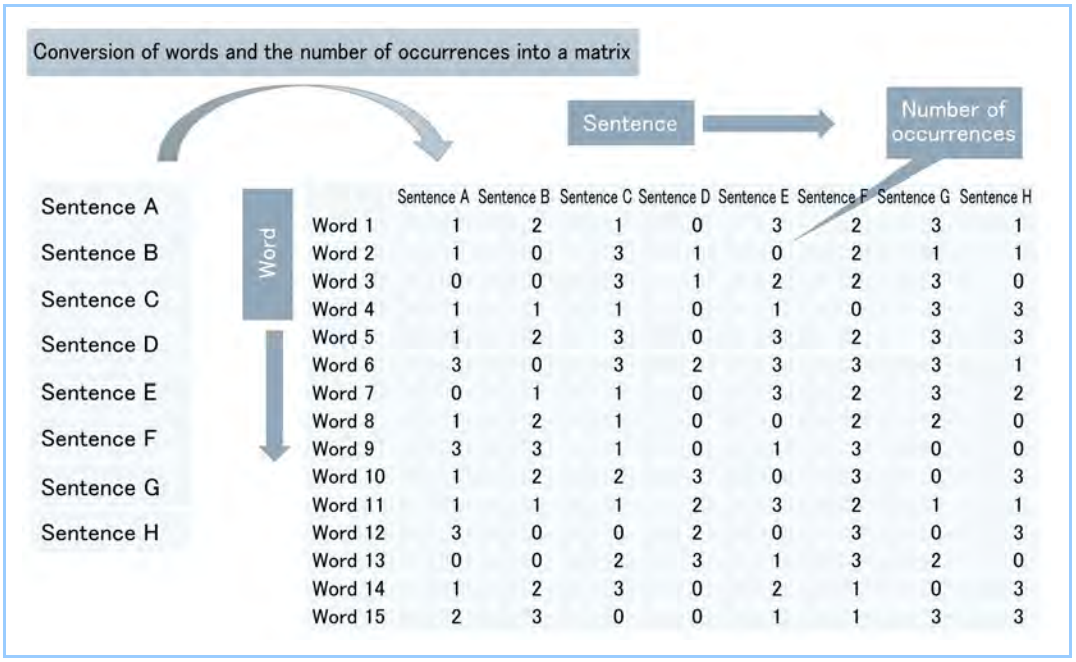
*2　Program Execution Manager, Power & Energy Solution Business Planning Department, Power & Energy Solution Business Division, Power Systems

*3　Power & Energy Solution Business Planning Department, Power & Energy Solution Business Division, Power Systems

searching, utilizing text mining technology. This method is widely used and is applied to the demonstration of text mining described in the next chapter.

Step 1: Quantifying text data

Divide sentences into words by morphological analysis where text data is converted into numerical data by calculating the number of occurrences of each word. Expression in table form comprising the number of occurrences of sentences and words facilitates processing by computers (**Figure 1**).

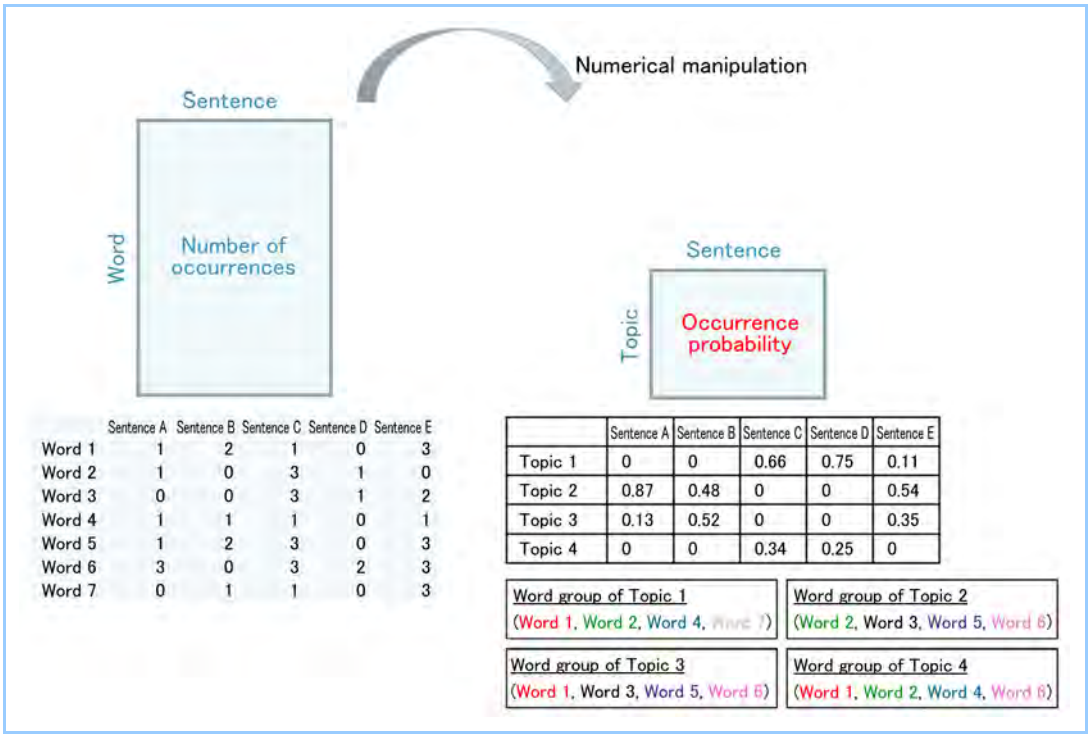


	Sentence A	Sentence B	Sentence C	Sentence D	Sentence E	Sentence F	Sentence G	Sentence H
Word 1	1	2	1	0	3	2	3	1
Word 2	1	0	3	1	0	2	1	1
Word 3	0	0	3	1	2	2	3	0
Word 4	1	1	1	0	1	0	3	3
Word 5	1	2	3	0	3	2	3	3
Word 6	3	0	3	2	3	3	3	1
Word 7	0	1	1	0	3	2	3	2
Word 8	1	2	1	0	0	2	2	0
Word 9	3	3	1	0	1	3	0	0
Word 10	1	2	2	3	0	3	0	3
Word 11	1	1	1	2	3	2	1	1
Word 12	3	0	0	2	0	3	0	3
Word 13	0	0	2	3	1	3	2	0
Word 14	1	2	3	0	2	1	0	3
Word 15	2	3	0	0	1	1	3	3

Figure 1 Quantifying text data

Step 2: Extracting the potential semantic information from text data

Apply mathematical manipulation (probability/statistical manipulation) to a matrix of the words described above and the number of occurrences, to build a conversion model where a sentence can be expressed with the potential semantic information (topic) comprising multiple groups of words and their occurrence probability rather than a matrix of words and frequencies (**Figure 2**) .



	Sentence A	Sentence B	Sentence C	Sentence D	Sentence E
Word 1	1	2	1	0	3
Word 2	1	0	3	1	0
Word 3	0	0	3	1	2
Word 4	1	1	1	0	1
Word 5	1	2	3	0	3
Word 6	3	0	3	2	3
Word 7	0	1	1	0	3

	Sentence A	Sentence B	Sentence C	Sentence D	Sentence E
Topic 1	0	0	0.66	0.75	0.11
Topic 2	0.87	0.48	0	0	0.54
Topic 3	0.13	0.52	0	0	0.35
Topic 4	0	0	0.34	0.25	0

Figure 2 Extracting potential semantic information

Step 3: Clustering (classifying) sentences

Perform sentence classification based on hierarchical clustering utilizing inter-vector distance in terms of the occurrence probability vector of the topic obtained for each sentence in Step 2 (**Figure 3**).

Step 4: Searching for sentences

Perform Steps 1 and 2 on any given sentence, whereby the occurrence probability vector of the topic is calculated in the same manner as it would be for a sentence to be retrieved, and search for/extract similar sentences according to the degree of similarity calculated from a comparison with the occurrence probability vector of a sentence group to be retrieved (**Figure 4**).

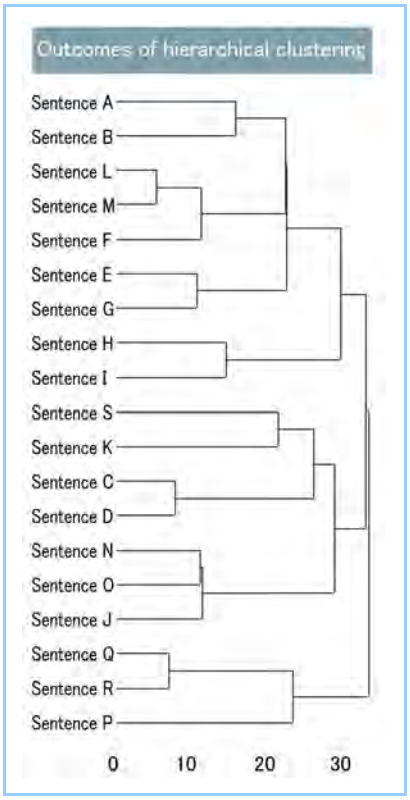


Figure 3 Hierarchical clustering

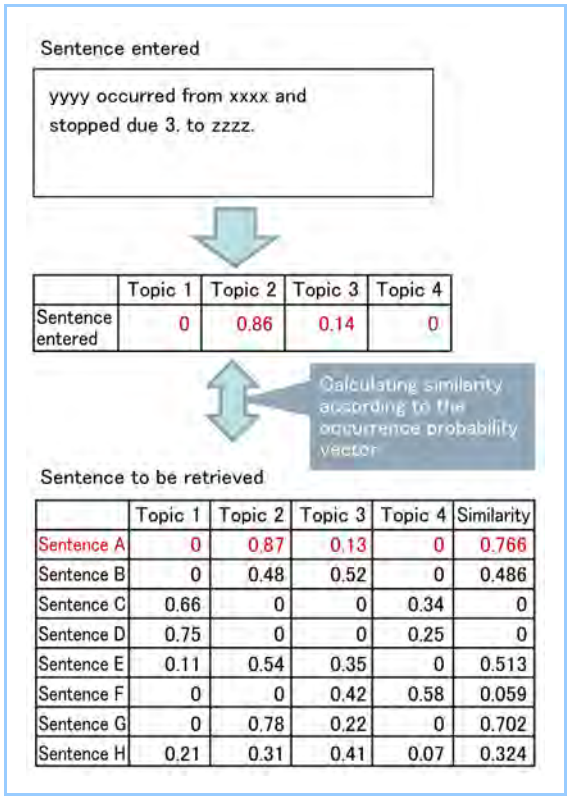


Figure 4 Calculating the degree of similarity to the sentence entered

3. Embodiments of text mining

This chapter will introduce 3 actual applications of text mining.

(1) Application to category presentation in incompatibility information management

Our Quality Assurance Department manages past incompatibility information using a system that registers, at the time of incompatibility, the details, causes, countermeasures, etc., while working on recurrence prevention activities utilizing such information. Such activities are crucial as they are closely related not only to product quality, but also to cost, delivery, safety and customer satisfaction.

Incompatibility information is entered into the system by individual departments, but for various reasons, such as how to describe what happened not being standardized and varying widely depending on the person, the information is not necessarily organized sufficiently to be utilized easily for recurrence prevention activities.

Accordingly, MHI has created a model that automatically performs the clustering of various similar expressions in text by applying the text mining technique to past incompatibility information where the individual clusters are automatically classified into multiple incompatibility reasons which are linked to respective countermeasures. Utilization of this model has allowed us to easily classify the expressions in text which have been quite random depending on who wrote them, increasing the quality of recurrence prevention activities.

(2) Application to troubleshooting in large-scale facility test run

When problems occur in the operation of large equipment/plants, especially in a test run at the time of start-up where troubleshooting takes time, losses such as delays in the start of operation, etc., are huge. Meanwhile, our Test Operation Department implements Fault Tree Analysis (FTA) of problems that have occurred in the past, and compiled the details, causes, measures, etc., into "Combat Lessons" which are utilized in troubleshooting and training young engineers.

However, since the details of problems are diverse and consist of text data including complicated technical terms, the conventional classification method would require the experience and expertise of veteran engineers for selection from the Combat Lessons suitable for the problems currently occurring, which is time-consuming.

Accordingly, the details of the problems, their causes and countermeasures compiled in past "Combat Lessons" were first broken down into smaller parts and then made into a database where they were subsequently classified into multiple phases applying the text mining technique (Steps 1 to 3 in Chapter 2) according to the level of relevance of each type of information. As a result, when a similar incident occurs, by entering text, for example, "Y of Machine X stopped at the signal of Z," similarities to past incidents will be immediately identified in the method of Step 4 in Chapter 2, allowing swift searches for an appropriate response/treatment. Conventionally, implementing FTA itself used to require advanced knowledge and experience. However, hierarchical classification using the text mining technique now allows us to effectively utilize a large volume of records in text format.

(3) Application to patent search support in Intellectual Property Department

In patent searches, which are closely related to technical trends, etc., the scope of searches for patents owned by other companies used to be narrowed down by the patent classification code given to each technical field and keywords related to products and equipment in the field of invention. However, when searching for new technologies where it is difficult to select an appropriate patent classification code for abstract keywords like "AI&IoT," it is impossible to narrow the search in an effective manner, and the scope of search becomes too extensive.

Accordingly, to efficiently narrow down the scope of search, the text mining technique is utilized to automatically classify abstract keywords within the search scope into relevant technical fields, which allows us to increase efficiency in later phases in the process including detailed search by experts in individual technical fields and to boost the accuracy of the search.

4. Conclusion

This paper introduced a method of performing clustering (classification) and searches by utilizing text mining technology, where the quantification and structuring of mass text data is achieved, as well as the positive outcomes of the application of this method to text data such as incompatibility information, Combat Lessons and patents, based on actual examples.

By establishing a general-purpose analysis method for text data, the types of data available other than numerical data increases, which indicates the potential for offering new solutions that utilize classification/search in addition to prediction to achieve efficient facility operation and improved lead time in the work process.

In the future, MHI will strive to acquire new knowledge by integrating auxiliary information to text data such as time of occurrence and the enormous amount of numerical data accumulating in day-to-day operations, as well as to expand the range of solutions we offer.