Explanation

Application Examples of Data Mining

TAKAMATSU Shinichi

1 Introduction

Data mining is a technology that applies data analytics such as statistics and machine learning to large quantities of data in order to obtain useful findings including potential trends and patterns. When the technology works well, unwritten knowledge noticed by experts through their experience, namely intuition or knack, could be formulated and transformed into formal knowledge, and any correlation between unknown events could be identified and become a foundation for innovation. Data mining literally means to "mine" "data" as if like digging a vein of gold. The technology has become increasingly important in a modern world flooded with data.

Fig. 1 can help you imagine how much this field has been gaining attention easier. In total five terms were selected for research purpose and how many times these terms were searched over the Internet every week in the past five years was counted. The proportion of the search count of the terms is shown in the figure with the maximum value during each period taken as 100. The five terms consist of three terms related to the technology (i.e., "data mining", "data analytics" and "data science"), one term representing KYB's core technology for reference purposes ("hydraulics"), and the other term expressing the impossible dream of mankind ("immortality"). The figure implies the people's trend of being interested in realistic topics. The terms related to the technology have gained more attention than that for mankind's dream, even though we finally live in an age in which we could scientifically discuss the possibility of achieving immortality using iPS cells. The term data mining, which was conceptualised in the 1980s, has received attention to almost the same extent as hydraulics, which is a long-lived technology since the age of the industrial revolution. Attention to synonyms data analytics and data science has dramatically increased. These three data-related terms are gaining much attention, equal to or above that for hydraulics or immortality.

With a focus placed on data mining as a technology for effective application of data, namely information, KYB promotes technology application and human resource development with an aim of achieving company-wide dissemination and establishment of the technology. Although this effort has just begun, I would like to describe the significance of the effort and introduce part of its technical results in this paper.

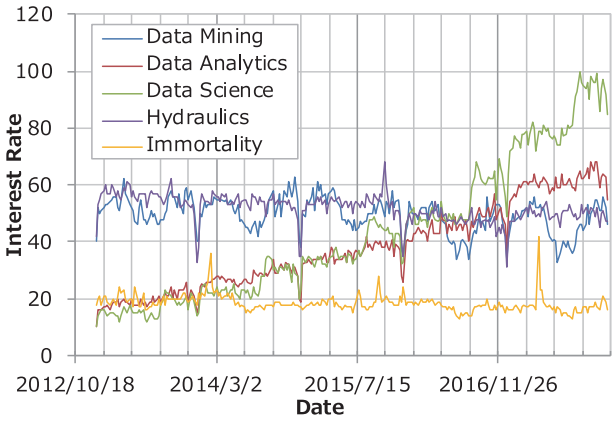


Fig. 1 Trends of word searching with Google (author's own survey)

2 What is Information?

Many of those who take notice of the keyword "data mining" have probably learnt that data scientists are said to have "the sexiest job of the 21st century" [1]. This expression originated in the phrase that "the sexiest job in the next 10 years will be statisticians" [2] said by Hal Varian, chief economist of Google, in 2009. The phrase predicting the coming of an age in which data analysis is critical was instantly propagated worldwide partly because of its vivid sound of words. Now it may be an even depreciated rhetoric with which you could be tired of hearing. Even if you do not know this phrase, you can hardly live your daily life without hearing the words big data, artificial intelligence (AI) and Internet of Things (IoT). There is no doubt that everybody can see or hear these words somewhere in society. A common concept that lies behind these words is information.

The word "information" is translated into Japanese as "jo-ho". "jo-ho" used to be a military term that originated in the "French Infantry On-site Exercise Standard", translated by the General Staff Office of the Japanese Imperial Army in 1876 [3]. The translated version of this Standard included an expression of "teki-jo-ho-koku" ("enemy

information report"), and this expression was abbreviated as "jo-ho" ("information"). The original word "information" means "to give form to the mind" or to inform of details or circumstances of a matter. As you may imagine from the origin of the word, information is any entity that brings the receiver a new idea or criterion, and is defined as "any entity that resolves uncertainty of knowledge on matters" [4].

For example, under a cloudless sky, information that it will be still clear one hour later is not valuable because it is quite likely, but information that heavy rain will come one hour later is valuable because it is unlikely to happen. The latter, if obtained beforehand, can be used as a criterion to determine proper action to be taken, such as bringing an umbrella.

That is, information has a higher value when it predicts what is unlikely to happen, or "an event with a low probability of occurrence". The concept of information quantity can be interpreted as the degree of unexpectedness [5]. Therefore, information quantity is highest when it certainly foresees an event by resolving very difficult uncertainty, although such information can't be easily obtained.

3 Trends of Data Mining

As a model clearly presenting the significance of information, this section introduces the DIKW model (Fig. 2) showing the hierarchy of data, information, knowledge and wisdom.

In this model, information as a generic term is divided into "data", "information", "knowledge" and "wisdom" in the form of a hierarchy where a higher-level component is the more important element. The model implies that, for example, numerical "data" such as time, atmospheric pressure and coordinates can be integrated and compiled into "information" expressing a change in pressure distribution with time, to which "knowledge" that low atmospheric pressure causes bad weather can be applied to produce a weather forecast, which may lead to "wisdom" of bringing rain gear for a certain level of probability of precipitation or higher.

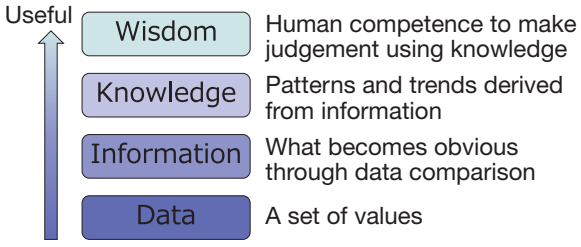


Fig. 2 DIKW Model

Exerting "wisdom" in this way supports an appropriate judgement for bringing about a desired outcome and offers a greater advantage compared to cases of not exerting "wisdom". In order to capture useful "wisdom", "data" and "information" should be organised and accumulated in such a manner that can be refined into available "knowledge". This is the real reason for obtaining information.

This importance of information has always been recognised. However, there were only limited means of effectively using information because both collection and analysis of information were not easy.

With recent rapidly advancing science, the limitations have been substantially alleviated. As shown in Figs. 3 and 4, data processing speed has been dramatically improved with advanced computer systems and miniaturisation technology for integrated circuits has advanced to achieve explosive growth in the data storage size. In addition, the penetration of the Internet and the improvement of sensors now make it possible to accumulate and analyse huge quantities of data that could not be collected or handled before. Today, data mining has gained the spotlight as a technology that can efficiently analyse a number of automatically collected data sets. Engineers engaged in such data mining are wanted as data scientists.

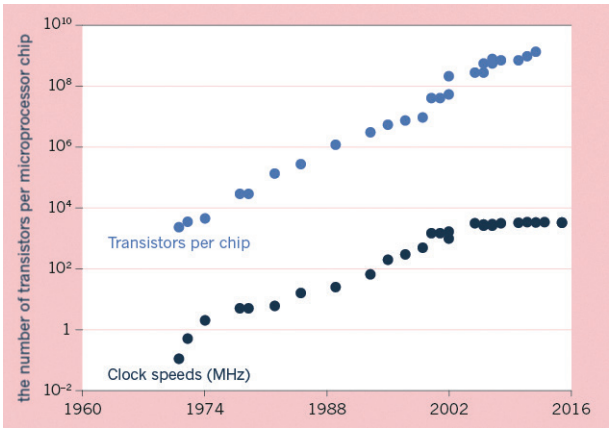


Fig. 3 Improvement of data processing capability of computer systems [6]

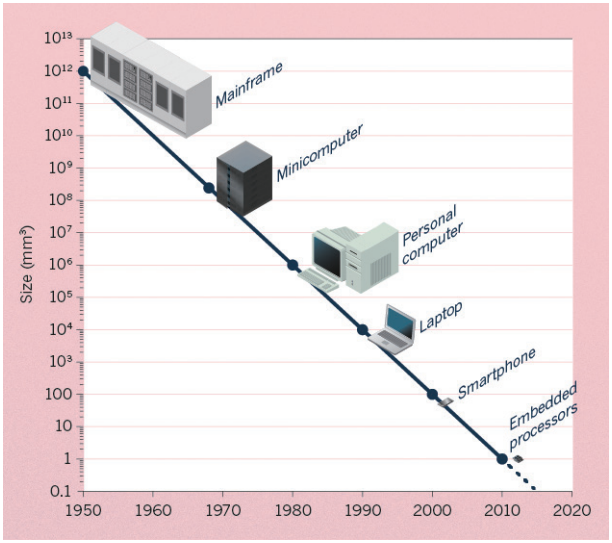


Fig. 4 Shrinking size of integrated circuits [6]

4 Physical and Statistical Models

In conventional product development using engineering means, the object is represented by a physical model and an equation is established, which is parametrically analysed to find an optimal value. This approach can produce more accurate results under simpler conditions. In reality, however, experimental or modelling errors are unavoidable, which directly leads to less accurate analysis. Thus, a physical model established according to textbooks is insufficient and generally modified with various knowhow introduced.

Unlike physical modelling to pursue the true value generation structure, statistical modelling used in data mining is based on a totally different idea. The purpose of statistical modelling is to successfully approximate the relationship between input and output regardless of the mathematical formula structure or the system of units. Equations obtained as a statistical model are not theoretically persuasive like the governing equations that describe phenomena according to physical law. They may be a heretic approach from the standpoint of engineering, but can disclose hidden patterns embedded in complex real phenomena and often work well as a method to dig up new findings.

Physical and statistical models can be compared to each other and evaluated as shown in Fig. 5. In physical modelling, input is subjected to an operation based on the governing equations for deductive processing to derive output. For example, a combination of input and operation as in "(1 + 2)" is used to determine the output "3". On the other hand, statistical modelling uses a set of input and output to estimate an operation that properly associates them with each other as an inductive approach. After the set of input and output "(1 2) = 3" is simply identified, the appropriate operation ("+") that can associate them with each other is estimated.

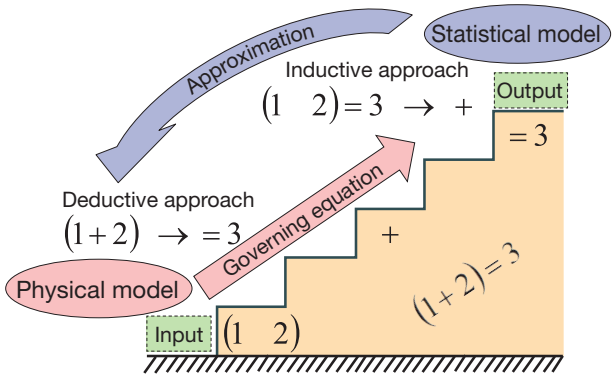


Fig. 5 Physical and statistical models

As stated above, physical and statistical models use totally different approaches to handle phenomena. Conventionally, sociology, medicine, psychology and other fields that handle phenomena with low reproducibility have often used statistical modelling, and the engineering field has applied physical modelling exclusively in many cases since it handles phenomena in which certain patterns can be found. However, the governing equations that form the basis of engineering are just formulas systematised from some of the patterns universally identified in natural phenomena, not a faithful description of actual complex phenomena. With recent leaps in data accumulation and processing technology as described in the previous section, it can be expected to obtain findings categorised in knowhow that can't be included in the governing equations by applying statistical modelling to actual input and output as a secure fact.

5 Case Examples

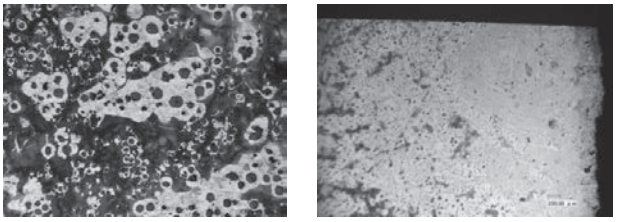
With consideration given to the significance of applying data mining technology to its efforts in the manufacturing industry, KYB promotes an activity to ensure company-wide dissemination and establishment of the technology. This activity includes applying the technology to various issues picked up from different departments in the company, discussing the results with engineers from relevant departments to obtain new findings efficiently, and refining the findings as they can be converted into the company's original technology.

Some of the activity cases to which the technology was applied guided personnel to resolve issues that would have otherwise not been clarified with conventional technology. Some others contributed to early development of proposed countermeasures. From these successful cases contributing to business operations, some examples are extracted and introduced below:

5.1 Measures against abnormal tissue in castings

Castings are one of the products that can't easily ensure stable quality due to several intertwined factors, including material components, shape and temperature. KYB has introduced a Quality Traceability System (QTS) installation line making full use of IoT into the KYB-YS Co., Ltd. plant, which is one of KYB's affiliates, establishing a system to accumulate several hundreds of data sets for each casting product.

KYB-YS had frequently had castings with abnormal tissues (Photo 1) in a period, and had got into a situation



(a) Normal tissues (nodular graphite)

(b) Abnormal tissues (flake graphite)

Photo 1 Normal/abnormal tissues in castings

where the cost incurred for disposal of these castings was too much to be overlooked.

The company had technical discussions about this quality failure with foundry experts. In order for graphite to form normal tissues, graphite must be spheroidised. The experts suspected the presence of an element that would prevent the spheroidising because the abnormal tissues had been found only in limited locations around where combustion gas was generated. Some other hypotheses based on their technical knowledge were also developed. However, measures derived from these hypotheses did not finally solve the failure at the root, although they brought about a certain effect. Then, they singled out the data mining technology to take a bird's eye view of the overall QTS data.

During various analyses conducted at the initial stage, QTS data for various information items, including the temperature of melt (molten metal) during casting, and the quantity of the elemental components were processed and organised into 110 variables that can be subjected to quantitative analysis. However, no obvious cause of the failure was found in this analysis. Casting is so subsensible that can be said to be like a creature. Technicians have to additionally tune the regular casting conditions according to variations in temperature, humidity or other factors of the day. With this feature taken into account, another way of thinking was emerging. The data was then stratificated by time and by product and applied with the decision tree and other analysis approaches that can examine the branching condition to determine whether failure occurs or not. As a result, the condition for abnormal tissues to occur was gradually revealed. That can be shown in Fig. 6 if permitted to be simplified with the detailed conditions omitted. This scatter diagram shows the casting temperature and time of casting process on the two axes for a product whose data has been stratificated for a certain period. Faulty parts are likely to relatively come together in the upper left area of the diagram compared to the satisfactory parts. This implies that a shorter casting time at a higher temperature leads to a higher failure rate.

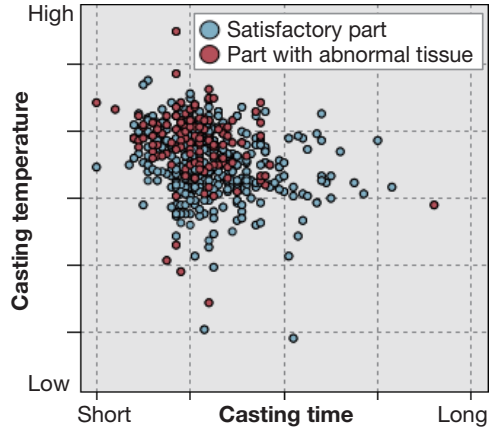


Fig. 6 Trends of occurrence of abnormal tissues

However, this finding alone does not directly lead to an effective measure. At the actual manufacturing site, changing a manufacturing condition will always cause a reciprocal phenomenon resulting in failure. It is essential to determine the manufacturing conditions by striking a balance among various phenomena. For decision tree analysis, it is possible to make analysis with consideration given to the occurrence of reciprocal phenomena caused by changing the condition. That is, the abnormal tissues could be reduced without affecting other factors. In fact, the decision tree analysis worked effectively in this case. The casting time and temperature as well as any given important factors were efficiently screened out. An equilibrium point of changing the conditions was sought in collaboration with the experts, which successfully eliminated failure.

5.2 Determining the concrete mixer state

Concrete mixer trucks (hereinafter "concrete mixers"), which are one of the products sold by KYB, play an indispensable role in the construction industry. Ensuring efficient operation will bring great merits. For the purpose of enhancing the operation efficiency of concrete mixers, KYB has promoted an effort to grasp the operating status of concrete mixers by using measurement signals retained by themselves.

In this effort, the sensors originally mounted on the concrete mixer were used to collect sensor signals on drum rotation speed, drum driving pressure and other parameters. The collected signals were processed into a time series data set consisting of 27 variables. The data was analysed to comprehensively interpret the relative behaviour of the variables. A challenge for interpretation is to determine the state of the concrete mixer at a specific time. Table 1 shows the outline of the data:

Table 1 Data used to determine the concrete mixer state

Time	State			Variable		
T	Y_1	Y_2	…	X_1	…	X_{27}
t_0	TRUE	FALSE	TRUE	…	…	…
t_1	FALSE	FALSE	TRUE	…	…	…
…	…	…	…	…	…	…

Since the concrete mixer state can be freely manipulated during testing, it is possible to put correct data in the columns under "Y" for state determination in Table 1. For actual operation, such data can't be obtained and only the variable X acquired from the sensor signals has to be used to estimate the actual state of the concrete mixer. However, the relationship between variables may substantially vary by the loading amount of fresh concrete of the concrete mixer, or by the load quantity handled by the operator, even if the mixer itself is in a same state. It was thus difficult to establish a universal criterion to determine the mixer state. None of the conventional methods successfully established a criterion for accurate evaluation.

Then, data mining was applied to find an evaluation criterion. As a result, it is now possible to determine the state at an accuracy of about 90% for linear discrimination, or at nearly 100% for non-linear discrimination by machine learning. The linear discrimination implies which variable is important for evaluation, so it can be used for discussion about technical issues. On the other hand, the machine learning approach does not provide an obvious evaluation criterion while offering high discrimination accuracy. So, this approach is not suitable for discussion from an engineering aspect. In this case, to contribute to engineering-based technical advance while ensuring high-accuracy determination, the different approaches have been finally combined in a way suitable for the purpose of solving the issue.

5.3 Pursuing optimal manufacturing conditions for oil seals

In the development of oil seals applied with new manufacturing technology, the new manufacturing process not included in the conventional technical knowledge led to unstable product quality, leaving many nonconforming parts among injection moulded prototypes. The number of quantities of state related to manufacturing conditions, as many as 50, such as temperature and pressure of different parts, were recorded in the injection moulding machine along with accumulated status data related to many faulty prototypes. These data sets were found to have no simple trend from which any relationship could be derived by thinking alone. The development team had difficulty determining the manufacturing conditions for ensuring stable quality.

Then, several analyses were made on the occurrence/non-occurrence of failure. The results are shown in Table 2. In total four analysis approaches created their own model for determining the occurrence or non-occurrence of failure. The table summarises, among the 50 kinds of quantity of state, which one the models focus on as a critical variable. All the four analysis approaches derived their model from different statistics as a basis. Since which statistic was suitable for the basis was unclear in this case, all the analyses were generally conducted to see what happened. Even though it was an ad hoc approach, the variables selected in each of the four analyses were compiled and the results were reviewed as a whole, which naturally presented some variables on which importance was placed by all the analysis approaches.

The results above represent a trend found only through data analyses with the statistical models applied, not based on any technical knowledge related to injection moulding. Still, this approach squeezes the 50 kinds of quantity of state down to as low as 14 and even gives them a degree of importance.

A group of squeezed-down number of kinds of quantity of state prioritised according to importance like this can be individually reviewed from an engineering viewpoint. Then, each quantity of state was subjected to evaluation by engineers, which revealed that there was no inconsistency in failure mechanism among them. These several kinds of quantity of state were assigned to an orthogonal array according to the design of experiments for further research. As a result, truly important manufacturing conditions were identified and the failure rate was substantially reduced. This is a successful case in which a combination of data mining and design of experiments refined an analytically found data trend into technical knowledge, resulting in efficient problem solving.

Table 2: Matrix of importance of 50 variables

Reference No.	Selected variable	Analysis approach				Importance (number of circles)
		Discriminant analysis	Decision tree ①	Decision tree ②	Decision tree ③	
1	X_1	○	○	○	○	4
2	X_3	○			○	2
3	X_4	○	○	○		3
4	X_5			○		1
5	X_7			○		1
6	X_8		○	○	○	3
7	X_{10}				○	1
8	X_{31}	○	○	○	○	4
9	X_{32}		○	○		2
10	X_{33}				○	1
11	X_{46}		○	○		2
12	X_{47}	○	○	○	○	4
13	X_{48}	○		○	○	3
14	X_{49}		○	○		2

6 In Closing

6.1 Concluding remarks

Qin Shi Huang, who was the first emperor of a unified China, and also famous for his search for an elixir of immortality, arrogated the first-person Chinese pronoun "朕" for his exclusive use. Since then the word "朕" has become the first-person pronoun exclusively used by absolute monarch. One who calls himself/herself "朕" means one who governs the whole nation. The word "朕" also means a "sign". It is interesting to see a significative relationship between the two words: a sign governs the whole.

Data mining may be a technology to search the "sign" efficiently. The technology is used to pick up subtle signs from a sea of information, not to find an untouched treasure. What is found needs to be refined until it is available as a pattern. Particularly for a manufacturer like KYB, it is indispensable to have the capability to technically interpret the result obtained as a statistical model, to figure out what it really implies, and to transform it into a physical model. If we neglected these, we could certainly be able to apply unknown patterns to practical fields for a short time, but would not be able to obtain an extensive applied science elucidating the full facts, eventually failing to

accumulate true expertise, and even inhibiting development in the long view.

The amount of information available in this modern age is too much for us to handle. A large volume of information is discarded without being processed at all (Fig. 7).

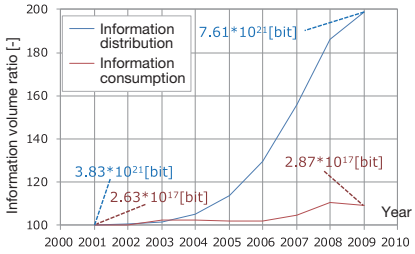


Fig. 7 Changes in information volume using the value in 2001 as a base of 100 [7)]

Information like this that never sees the light of day may include embedded patterns or rules with high repeatability, although they are too complex to be interpreted easily. Particularly when different information sets that have been independently handled in separate fields are integrally analysed, it is highly expected that any unknown relationship can be identified. This was originally a long-established pattern of invention or discovery. Therefore, I believe it is essential to develop a base to collect information, including IoT, and establish a flow of processes, including from accumulating information, allowing efficient discoveries to take place through data analysis, and refining them into reliable manufacturing technology. From this standpoint, KYB still has many apparent inadequacies indeed. First of all, I would like to begin the first steps toward the dissemination and establishment of data mining by promoting an awareness raising activity to open the door.

"If you want to achieve a triumph, you must work hard to do the small things. Many drops make a shower. Generally, nobodies always want to succeed and neglect doing the small things. They feel sad about being unable to do what is difficult and do not do what is easy. That is the reason why they can't finally achieve in anything". [8)] These words are really a wisdom.

6.2 Acknowledgement

On this occasion I would like to thank the cross-departmental team members involved in the technology dissemination activity and their supervisors as well as many those who extend support by providing information about actual cases to which the technology was applied. Without their cooperation, I think this activity would still be groping in the dark. I would like to appreciate here that the activity has taken root as a wide-spread profound effort.

6.3 Additional note

Apart from natural science including technology, this additional section introduces what the phrase "significance of a sign" sometimes suggests in the humanities, which may reflect the significance of finding relationships by integrating different fields. I hope the reader will discuss the issue based on the concept of data mining that derives patterns from signs.

The elixir for immortality Qin Shi Huang obtained was mercury. He is said to have died from elixir poisoning. Mercury had been considered as a mysterious material from ancient times and was also positioned as a very important material in Western alchemy. Alchemy is the study aimed at sublimating material, body and mind to a complete existence as like gold, and was started with a stimulus from gold's physical rarity and stability. Mercury has the physical property that it can evaporated from gold amalgam to produce gold plating. Ancient people with no idea of specific weight must have been dazzled by the property of mercury that implies a sign of completeness.

A famous novel on alchemy [9)] presents a theme that one who listens to signs will open a way to realise a dream. This novel is a literary work of enduring value often is compared with The Little Prince [10)] whose theme is "the precious things can't be find by eyes", and whose precept is "the significance of looking at things with one's heart."

Do you understand what I mean? The humanities also have unreachable goals such as completeness and dreams. Like natural science, it is full of challenges and failures. Many masterpieces that arouse people's sympathy have one thing in common, that, to reach these goals that are generally difficult to find, one must have an attitude of trying to look at what you can't see with your eyes. This way of thinking is similar to the concept of serendipity [11)] in natural science, that deriving something from failures will probably bring a big discovery.

Finally, even though it may be an unnecessary addition, I would like to conclude this article by referring to the word "significance". The prefix of this word is "sign".

References

1) Thomas H. Davenport, D.J. Patil: Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review, (2012/10).
2) Steve Lohr: For Today's Graduate, Just One Word: Statistics, New York Times, (2009/8).
3) Takumi Sato: Literacy of Modern History - Cosmos of books, Iwanami Shoten, Publishers, (2012).
4) Telecommunication Glossary, Corona Publishing Co., Ltd., (1984).
5) Yukito Iba: 13 Chapters on "Information" - Privately Printed Information Science Primer, Bussei Kenkyu, 78-2, (May 2002).
6) M. Mitchell Waldrop: More Than Moore, Nature, Vol. 530, (February 2016).
7) Ministry of Internal Affairs and Communications: Result of Research on Measurement of Information Communication Volume and Actual Status of the Information Communication Market in Japan, (2011).
8) Sontoku Ninomiya: Hotoku Yoten, Naigai Shobou, (1934).
9) Paulo Coelho: O Alquimista - A Boy Who Traveled in a Dream, Kadokawa Corporation, (1997).
10) Antoine de Saint-Exupery: The Little Prince, Iwanami Shoten,

Publishers, (2000).
11) Kuniyoshi Sakai: What scientists do - How do they have creativity? Chuokoron-Shinsha Inc., (2006).

Author

TAKAMATSU Shinichi

Joined the company in 2007.
Mechanical Component Engineering Sect., Basic Technology R&D Center, Engineering Div.
Professional Engineer (P. E. Jp)
Vibration control expert engaged in studies on hydraulic shock absorbers and promoting applying statistical data analysis to operation.