DEVELOPMENT OF MATERIALS INFORMATICS PLATFORM

Yasumitsu Orii, Ph.D., Shuichi Hirose, Ph.D., Hiroki Toda, Masakazu Kobayashi
New Value Creation Office, Nagase & Co., LTD
Japan
yasumitsu.orii@nagase.co.jp; shuichi.hirose@nagase.co.jp

ABSTRACT
As the use of IT increases importance with big data and AI, the issue of power consumption has been highlighted. Under these circumstances, the development of new materials is more and more important. Materials Informatics (MI) is one of the hottest technologies in the material development field, because of its potential to reduce the time and costs of discovering innovative materials. To achieve this, the key is to collect data that has been accumulated for many years at research institutions and companies, and to make information extracted from the data into knowledge.

This article introduces the development of two methods based on AI: the “cognitive approach”, which reads vast amounts of literature information and digitizes data, and the “analytic approach”, which theoretically estimates the structure and physical properties of chemical substances from predictive models.

INTRODUCTION
Artificial intelligence (AI) has been repeating the hot boom and the cold winter era, but this latest AI boom is said to be the “third AI boom”. We think of the base of this third AI boom is four technological innovations: the explosion of data volume (big data), the evolution of networks, the evolution of algorithms, and the dramatic progress of hardware.

In terms of “data volume”, the total amount of digital information generated worldwide in 2010 is expected to exceed 1 ZB (Zettabyte) per year and reaching 44 ZB in 2020 [(1)]. The Internet of Things (IoT) is key factor of the data volume increasing. Conventionally, humans created data, but these days, machines and sensors create data on behalf of humans.

However, among this vast amount of data, only about 20% of the data used by companies is structured data, and it is easy to manage with conventional computers. On the other hand, unstructured data accounts for about 80% of the total data, and it is expected that this percentage will increase furthermore soon. And that huge amounts of data will not be used. It is no exaggeration to say that AI has emerged to take full advantage of this buried unstructured data.

Regarding the “network”, an increase in communication speed is a major interest. The data created by sensors is stored in various places, but most of them are stored in the cloud environment via wireless communication. At present, the communication band and speed are problems, and there are cases and situations where it is impossible to use, even if it is desired by users. The fifth-generation mobile communication system (5G) is expected as a breakthrough. 5G is expected to be introduced between 2019 and 2020. It is assumed that a maximum transmission speed of 20 Gbps, that is 20 times faster than current LTE, and low latency. Therefore, remote robot operation and surgery in real time can be performed. In addition, the problem of multiple simultaneous connections will be solved. 5G will be the key technology for acceleration of big data.

As for the "algorithm", the adoption of deep learning significantly improved the error rate at the image recognition. It is said that the error rate in human image recognition is about 5%, but when deep learning is used, it is in the 2% range [(2)]. And it can be said that it has far exceeded human ability. Deep learning can turn a lot of big data into value.

As for "hardware", along with Moore's Law, semiconductors have evolved, that is, the performance of computers is one trillion times higher in 60 years and has evolved overwhelmingly. This is exactly following "The Law of Accelerating Returns" proposed by Ray Kurzweil. One important invention is linked to another, it can be shortening the time between the appearance of the next important invention and accelerating the speed of innovation. Thus, the performance of computers has been improved based on the rule of thumb that science and technology advance exponentially instead of linearly. Moore's law in the law of accelerating returns is said to be the fifth paradigm shift. The history of semiconductor starts from punch cards, relays, vacuum tubes, transistors, and integrated circuits. It is exactly that integrated circuits have evolved along with Moore's law. However, following Moore's law has also become more difficult due to the physical limitations of transistor integration. That means the improvement of AI performance requires the sixth paradigm shift.

There is one more thing to consider about hardware. It is a matter of power consumption at data center. The data volume will expand to 163ZB in 2025 [(3)], and already in 2015, the power consumed by data centers worldwide will exceed the power consumption of only UK. If the data volume keeps

increasing accordingly and the use of AI expands, it is predicted that electric power will be insufficient. Therefore, there are two scenarios for reducing power consumption. One is "AI at the Edge" (making the edge intelligent). The other is "Brain inspired Devices," (a new device that mimics the human brain).

"AI at the Edge" is a method of converting unstructured data into structured data on the edge side. The structured data created by making the edge side intelligent without sending sensor data directly to the cloud. As a result, the cloud side can use the conventional computer processing, and it is contributing to the higher energy efficiency. "Brain inspired Devices" is an effort to reduce the power consumption of semiconductors themselves. IBM's Watson, who won the Quiz Jobapadi in 2011, used 200 kW of power, but it is said that if humans tried to do the same thing, 20W would be enough [(4)]. The development of a self-learning chip that mimics the Synapse in the human brain is progressing all over the world, in order to make a low power system like the human brain.

As described above, hardware supporting AI has two major problems. First, with the end of Moore's Law, the law of accelerating returns in computer performance is beginning to be difficult to continue. The second is the problem of power consumption in data centers. As a technology to solve these two problems, “Brain inspired Devices” is one possible way, but this fusion technology requires knowledge of biology in addition to electronics technology. It can be said that fusion of different field technology has high hurdle to implement. The technology that has emerged under such circumstances is Materials Informatics (MI). In order to bring about the sixth paradigm in the law of accelerating returns, excavation of new materials is an important factor, and MI uses AI as its means.

According to Nature report about Chemical Space, there are 10^9 (one billion) materials known to the world. On the other hand, there are 10^{62} unknown materials [(5)]. From such many undiscovered materials, AI will find materials suitable for new materials for next-generation semiconductors. There are two different AI engines for MI. One is the “cognitive approach,” in which AI reads a vast amount of materials and data on materials, understands and organizes the data, and then proposes new materials required by users. The other is an "analytic approach" that learns the relationship between the chemical structure of a huge number of substances and their physical properties and indicates the "chemical structural formula" of the substance required by the user. This paper describes these two different engines.

DEVALOPMENT OF MI PLATFORM

Utilization of AI in Material development process

The material development process can be roughly divided into three steps: “planning”, “design”, and “evaluation”. In conventional material development, especially the “design” process often depends on unexpected discovery and the empirical knowledge of skilled workers, so material design difficulties and prolonged development period remains as major issues. In addition, expanding the search space for materials is an indispensable task for designing new materials. With the explosion of experimental data by accompanying the automated experimental methods and literature data , the importance of MI in the field of material development has increased dramatically. The future picture of MI is to achieve dramatic time saving and cost reductions, and to develop new materials at a high speed that overwhelms trial and error in conventional experiments.

There are three main reasons why AI uses in the “design” process of material development. The first is to make it possible to identify the optimal experimental conditions through exhaustive process parameter search and material simulations. Although, in traditional approach, it takes a lot of time to repeat the trial and error to find out the optimal experimental conditions, AI technology allows researchers to obtain the desired result at high speed. The second is to enable humans to comprehend data by visualizing vast and multi-dimensional information. Information that was previously too complex to be understood is simplified by machine learning processing so that material researchers can interpret the data. Third, in material design, the merit of machine learning is to discover novel knowledge that humans have not found yet. Machine learning is able to handle a huge amount of data that humans cannot read even for a lifetime and solve problems without prejudice like existing rules in material development, hence, there is a possibility to get unexpected discoveries. For material developers, the discovery of novel knowledge in material design among three prospects may have the highest for AI (MI).

There are two types of knowledge discovery in the filed of material design: deductive and inductive approaches (Figure 1). A deductive method is the way to lead conclusion by applying things to known laws and rules. This is an approach to consider the cause of a function or property from a structure or composition, in order to determine which structure performs a specific function. On the other hand, the inductive method is the way to draw up conclusions by accumulating statistical parameters derived from various facts. This is an approach to create a material and evaluate why properties have been achieved. Then, to design the molecular structure and composition based on the desired functions and characteristic conditions.

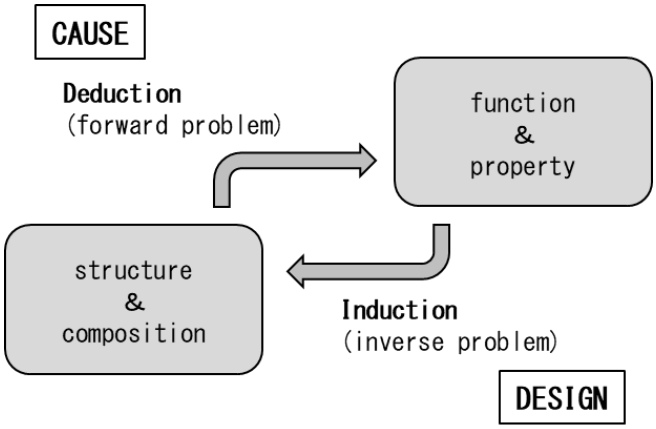

Figure 1. Two approaches for discovering material [6]

In the field of MI, the former is sometimes described as solving a forward problem, and the latter is described as solving an inverse problem. For material developers, the latter approach is more important, since they consider which structure and composition satisfy the desired physical properties. Therefore, the importance of finding novel knowledge by solving the inverse problem will increase in the use of MI in the future.

Cognitive approach and analytic approach

Here, the "cognitive approach" and "analytic approach" in material design are introduced, and their features are summarized in Figure. 2.

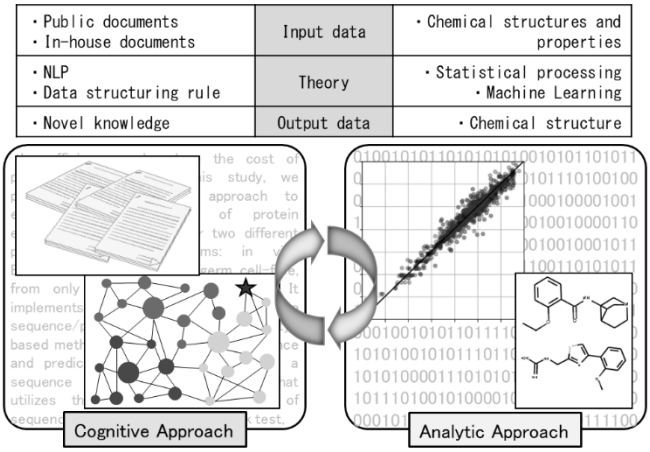

Figure 2. Overview of cognitive and analytic approach

The cognitive approach consists of three steps: "converting document information into structural data", "extracting facts by natural language processing", and "transforming facts into a knowledge graph (KG) expressing their correlation" (Figure. 3).

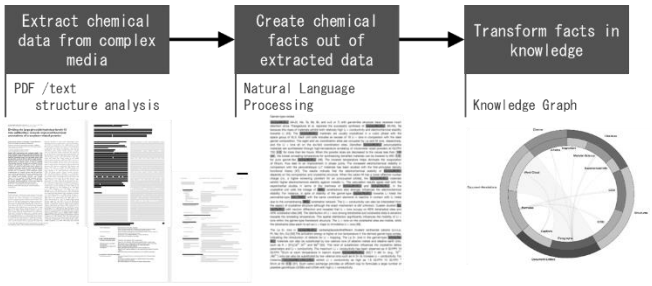

Figure 3. Workflow for cognitive approach

As a first step, the document information, which is unstructured data, is converted into structured data so that the computer can process it. For example, scientific papers, patents, and in-house documents are usually stored in pdf format, therefore those data are often difficult for computer to process. One of the solutions for document processing is the Corpus Conversion Service (CCS) reported by Staar et al. [7]. CCS can recognize titles and authors (labels) in scientific papers and extract text information along with the labels. In the next step, facts (entities) related to materials and their relationships (relation) are extracted from the text. Here, in addition to commonly used natural language processing, a dictionary that describes technical terms such as material names and physical property values and the relevance of those terms in the field of material is required to get entity. The final step is to develop the KG by joining the relations extracted from the document. The KG can represent data obtained from various resources as one graph structure.

As an application example of this method, Manica et al. reported the search for carbohydrate-related enzymes [8]. One KG constructed from public databases dealing with compounds, enzymatic reactions, and proteins, and an exhaustive search for enzymes involved in trehalose production has been conducted. While the cognitive approach has the advantage of handling volumes that humans can hardly read, it can only extract facts in documents. In the future, it is expected that cognitive approach will infer new knowledge not written in documents by combining KG and machine learning. That is, we call Inference Model (IM).

The analytic approach includes two processes. The one is called a forward problem that analyzes a correlation between chemical structures and their physical property values, and the other is called an inverse problem that generates a chemical structural formula satisfying a target property value (Figure. 4).

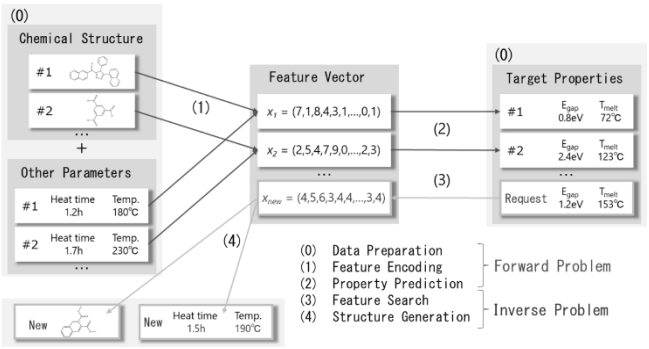

Figure 4. Workflow for analytic approach

First, the chemical structural formulas read as input information are converted into feature vectors using by a set of descriptors. For instance, the number of atoms, the number of partial structures, and so on are employed for descriptors. They need to be able to capture features of molecular structure exactly, and to generate a chemical structural formula from the features.

Next, model for predicting a physical property value is constructed. Generally, regression model is applied for such cases. The model with high prediction accuracy can be constructed, then it is possible to proceed to the next process called the inverse problem. In this step ,a feature vector that satisfies a target property value is explored from a prediction model using a particle swarm optimization (PSO) method or other similar methods at the beginning.

Next, based on the given feature vector, chemical structural formulas that can exist mathematically are generated employing graph theory, and so on. As an application example of this method, Takeda et al. reported the design of drug candidate compounds [9]. In this research, candidate compounds which meet the target values for two contradictory property values (LogP(Octanol/water partition coefficient), TPSA(polar envelope)) are designed, some of them were not included in the input data set. The potential of analytic approach to discover novel materials is of great interest.

However, it is difficult for the inverse problem to design material with large molecular weight due to their limitation of the calculation cost of search space. In this case, designing sub-structure having a high correlation with the target property value is one of the solutions.

While analytic and cognitive approaches seem at first glance to be entirely different approaches, combining the two approaches is expected to further accelerate material discovery.

Overview of Platform Services for MI System

Our Materials Informatics (MI) system is intended to be provided as SaaS (Software as a service). Since this system uses the service built on the server via the Internet, multiple authorized members can use the system in the same state. Software is regularly updated by cloud administrators and is always up to date. Since the used data is not stored in the terminal, the risk of data leakage can be reduced.

In Figure 5, it shows an overview of the MI system. The MI system is managed in the cloud as a domain for each application. Each domain can only be accessed by users who have been granted an account. Since the users' private data are ingested by the users, the private data cannot be viewed by anyone other than the users, including the cloud administrator. Public data such as patents can be viewed only by the users when the data are ingested. However, if there is a request from the users, a service for collecting and ingesting public data will be considered in the future.

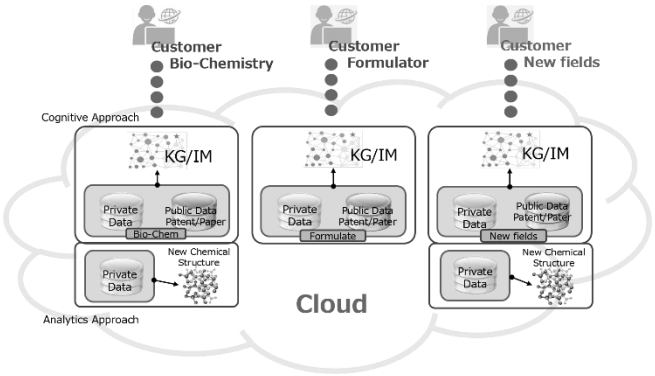

Figure 5. Schematic image of Materials Informatics

Figure 6 summarizes the possibilities of the MI system. It roughly summarizes two application examples. The first is the search for new materials discovery. In addition to shortening the development period for finding new materials, it is possible to reduce the number of experiments for finding alternatives. Another example of application is the proposal of new manufacturing conditions and the composition of a blended product. The offer of a new manufacturing process could reduce the number of trials and errors for optimizing manufacturing conditions. In addition, there is also expected a possibility of presenting a composition having a new function and improving the yield.

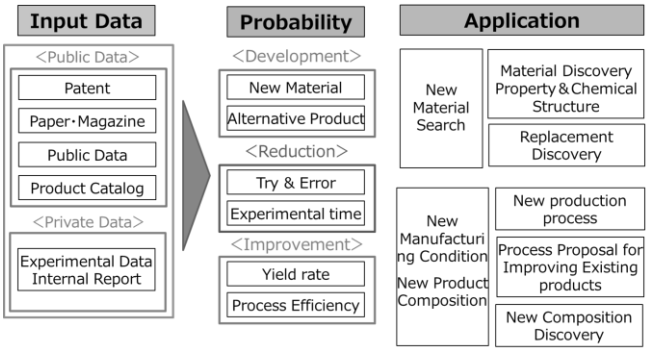

Figure 6. Summary of possibilities of Materials Informatics system

SUMMARY AND CONCLUSION

In this article, the use case of AI and the future demand of Materials Informatics are described. With the evolution of IoT, the advancement of 5G, and the growth of energy consumption due to the enormous amount of data processing on servers, the increasing needs for new material development and intensifying competition in material development are inevitable. MI technology will be an indispensable technology for improving the speed of development in the search and design of such new materials, not only organic materials but also inorganic materials. And also, MI will certainly be the source of future corporate competitiveness. MI can search for target materials directly from the database which is based on conventional method depending on the intuition and experience of researchers. And it also enables efficient or direct way to develop materials to meet the needs of users. There are many advantages of MI, however we should do more for improving the systems for commonly using. For example, further improvement of data analysis technology, accumulation of data, and establishment of support system for MI introduction to users, is needed. The training of data scientists who can analyze data is also urgent. The future direction of MI evolution is the accumulation of unstructured data collected from sensing devices, combined with clustering processing by AI, and new material simulation using quantum computers as accelerators. The field of quantum chemical calculation is also expected to be actively developed with MI field, as a more effective way to find out materials properties compare to conventional calculation methods[10].

REFERENCE

(1) EMC Digital Universe with Research & Analysis by IDC, "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things", April 2014

(2) Matsuo Yutaka (2015), Does AI exceed humans?, KADOKAWA Sensho

(3) David Reinsel, John Gantz, John Rydning, Data Age 2025, IDC Apr 2017

(4) Koji Hosokawa, Yasutaka Katayama, Yasumitsu Orii, Cognitive Chip, ProVISION No.83/Fall 2014

(5) Kirkpatrick R.S, C.McMartin W.C Guida Chemical Space, Nature.432(7019):823-865, 2004

(6) Iwasaki Yuma, "Materials・Informatics", Nikkan Kogyo Shimbun, 2019

(7) Peter Staar, Michele Dolfi, Christoph Auer, Costas Bekas : "Corpus Conversion Service: A machine learning platform to ingest documents at scale", the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018

(8) Matteo Manica, Christoph Auer, Valery Weber, Federico Zipoli, Michele Dolfi, Peter Staar, Teodora Laino, Castas Bekas, Akihiro Fujita, Hiroki Toda, Shuichi Hirose, Yasumitsu Orii : " An Information Extraction and Knowledge Graph Platform for Accelerating Biochemical Discoveries", Workshop on Applied Data Science for Healthcare at KDD, 2019

(9) Seiji Takeda, Hsiang Han Hsu, Toshiyuki Hama, Toshiyuki Yamane, Koji Masuda, Daiki Nakano: "New Material Search Method by Artificial Interagency", The Japan Society for Artificial Intelligence, 2018,

(10) Hidetoshi Nishimori, Masayuki Ozeki (2016), " Quantum computer accelerates artificial intelligence", Nikkei Business Publications, Inc.