Overview of Material Research by Information Integration Initiative (MI2I)

K Terakura
(NIMS)

米国での Materials Genome Initiative (MGI) に刺激されて、世界の多くの国で所謂マテリアルズ・インフォマティクスのプロジェクトが始まっている。我が国でも昨年より、JST のプロジェクトとして、物質・材料研究機構（NIMS) を拠点とした情報統合型物質･材料開発イニシアティブ（MI^2I) が始まった [1)]。主な目的は、データ科学と物質・材料科学の連携により、物質・材料開発を加速することである。本プロジェクトでの重要な出口課題の一つとして、磁石・スピントロニクス材料を設定しており、その枠における一つの具体的な成果として、希土類元素と 3d 遷移金属元素からなる磁性体のキュリー温度の実験データをつかって、機械学習によりキュリー温度の予測をした。機械学習を用いて、望みの性質を持つ物質・材料を探索する仕組みを説明し、いくつかの具体的な例を紹介する。

Reference
1) http://www.nims.go.jp/research/MII-I/index.html (Accessible on 2016/06/01)

データ科学手法による磁性材料探索

小口多美夫
大阪大学産業科学研究所
物質・材料研究機構

Data-Science Approach to Magnetic Materials Exploration

T. Oguchi
Institute of Scientific and Industrial Research, Osaka University, Ibaraki 567-0047, Japan
National Institute for Materials Science, Tsukuba 305-0047, Japan

Data-science approaches with rapidly growing data have recently brought a new trend of research and development to a variety of fields in science and technology. In materials science, it is now widely called ”Materials Informatics (MI)”, as often seen in several related world-wide projects[1–5]. The key strategy is to integrate data-science techniques with experimental, theoretical, and computational ones. Especially big data generated by computational simulations together with existing experimental databases are the target of data-science methods such as data mining and machine learning interleaved with appropriate physical modeling and descriptors. In MI, first-principles density-functional-theory calculations among the computational approaches play an important role for supplying data and knowledge on materials complemental to the experimental databases. This is one of the characteristic features of MI contrast to the preceding ”Bioinformatics”. In this talk, I shall introduce some fundamental issues of the data-science approaches to the exploration of magnetic materials in our research project MI^2I.

References

1) Materials Genome Initialtive (MGI): https://www.whitehouse.gov/mgi
2) Materials Design at the Exascale (MAX): http://www.max-center.eu
3) Novel Materials Discovery (NOMAD): http://nomad-coe.eu
4) An e-infrastructure for software, training, and consultancy in simulation and modeling: http://cordis.europa.eu/project/ rcn/198333_en.html
5) Materials Research by Information Integration Initiative (MI^2I): http://www.nims.go.jp/eng/research/MII-I/index.html

Computational exploration of new permanent magnet compounds

Takashi Miyake[1,2]
[1] CD-FMat, National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8568, Japan
[2] CMI[2] and ESICMM, National Institute for Materials Science, Tsukuba 305-0047, Japan

I will discuss current status and challenges for permanent magnet research by information integration. Strong magnet compounds such as $Nd_2Fe_{14}B$, $Sm_2Fe_{17}N_3$ and $NdFe_{12}N$ consist of three elements, namely rare-earth, iron and the third element. A natural question is: What is the best third element, and what about the fourth in a quaternary compound? This is an issue to be tackled by computational screening. As an example, we will present first-principles calculations of $ThMn_{12}$ type iron-based compounds. However, brute-force search based on first-principles calculations is computationally demanding even if using supercomputer facilities, since the number of combinations of chemical composition increases rapidly as the number of elements in a compound is increased. Machine learning is a possible solution to improve the efficiency drastically. It is found that Gaussian process regression using 7 descriptors accurately reproduces the Curie temperatures of bimetal alloys composed of transition-metal and rare-earth elements. This technique can be utilized for virtual screening. Another issue is exploration of crystal structure. Saturation magnetization is expected to be larger as the iron content increases. Hence, the crystal structure of new iron-rich phases is of particular interest. Crystal structure prediction is a hot topic in computational materials science in the past decade, and various efficient algorithms have been developed. Recent progress and applications will be reviewed.

Mining magnetic materials data

DAM Hieu Chi
(Japan Advanced Institute of Science and Technology)

The most important underlying hypothesis of materials researches is that the features of the structure of materials, as well as its derived physical properties has strong multivariate correlations. The task of materials design is to make these correlations clear and to determine a strategy to modify the materials to obtain desired properties. However, such correlations are usually hidden and difficult to uncover or predict by experiments or experience.

For dealing with this issue, data mining methods which can extracting meaningful information and knowledge from large data sets, are attracted a great deal of interest. Motivated by using data mining to solve data-intensive problems in materials science, we develop a method to quantitatively model the multivariate correlations between physical properties of materials and their structures by using sparse modeling. The key idea of our method is to use advanced statistical mining algorithms, in particular multiple linear regression and non-linear regression regularized least-squares [1, 2] to solve the sparse approximation problem on the space of structural and physical properties of materials. We use cross-validation to consistently and quantitatively evaluate the conditional relations of physical properties to all the structural features of the materials in terms of prediction. We apply the method to a data set of more than four thousand transition rare-earth metal alloys. We demonstrate that the obtained sparse model is not only significant for the comprehension of the physics relating to the materials, but also valuable for the guidance of effective material design.

Reference

1) R. Tibshirani, J. R. Statist. Soc. B 58, 267 (1996). B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Annals of Statistics 32, 409 (2004).
2) C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press (2006).

Expectation for Materials Informatics in Magnetic Material Research

磁性材料研究におけるマテリアルズ・インフォマティクスへの期待

T. Shoji
Advanced Material Engineering Division, Toyota Motor Corporation, Susono 410-1193, Japan

概要

急激な計算機の計算速度の高速化と記憶媒体の高密度化に伴い、大量のデータの利活用が可能になり、様々な領域へ情報科学（Informatics）を活用した取り組みが波及している。物質・材料の研究開発においても、情報科学の利活用の潮流は確実に押し寄せており、2011 年に開始されたアメリカの Material Genome Initiative[1]を皮切りに世界レベルで物質・材料にかかわるデータを活用した新材料の探索、新たな法則の探求といった取り組みが本格化しつつある。単純に Big Data を活用するといってもデータそのものだけでは何も得ることはできず、そこに情報科学的なアプローチで解析するということが必須となる。得られた結果をデータとして蓄積し、解析を行うことで、データの持つ意味を最大化し、新たな情報への変換や新たな知見を抽出することが材料科学（Materials Science）へ情報科学（Informatics）を適用することへの期待である。

一方、自動車メーカーの先端材料技術に携わる観点から見たとき、現在の電磁気活用を想定した磁性材料を取り巻く状況は、アプリケーション面では拡がりを見せているといえる。例えば、駆動用モーターや電圧変換、直流交流変換など、従来の自動車には搭載されていなかった電磁気部品がハイブリッド車をはじめとする駆動系にモーターを搭載している次世代車では欠くことのできないものとなっている。駆動用モーターを搭載した車両の年間の販売台数も、ハイブリッド車への参入を果たす自動車メーカーが増えてきたことも相まって、加速度的に増加している。現在のところ、NdFeB 系の磁石が駆動用モーターに用いられる磁石としては主流であり、希土類の低減や重希土類フリー化などの課題は依然として解決していない。また、従来車両にも用いられている部品においても、小型補機モーターやスピーカーなど目立たないところにも多量の磁性材料が用いられており、性能とコストをバランスさせた磁石の開発についても、軽量化を目的としてニーズが高い。

講演では、自動車メーカーの材料技術の技術者から見た自動車用途を想定した磁性材料の研究への期待と、研究の深化・加速・拡大に対して情報科学（インフォマティクス）が果たしうる役割についての所感と期待について述べる。

Reference

1) https://www.mgi.gov/

Opportunities and Challenges for Inorganic Material Informatics from a View Point of Big Data Analytics

Yuzuru Tanaka
tanaka@meme.hokudai.ac.jp
(Graduate School of Information Science and Technology, Hokkaido University)

1. Opportunities for Data-driven Sciences

While “big data” in general is characterized by 3V, i.e., the volume, the velocity and the variety of the target data set and/or data stream, by 4V, adding the veracity of data, or by 5V, adding the value of the analysis result, “big data” in applications, especially in cutting-edge science, symbolizes the paradigm shift from mission-driven research to data-driven research, where the volume may not be the major property of the target data set in the current situation. Recent development of big data core technologies including analysis algorithms and high performance data management and analysis platform technologies, together with the development of automatic measurement instruments and/or large-scale high-performance computer simulation technologies, are currently strongly promoting this paradigm shift to data-driven research in varieties of domain sciences, which is gradually allowing us to conduct scientific research studies completely in cyber worlds after having obtained all the required data sets, or through the real-time receiving of data streams. This trend will further allow us to easily share and exchange not only data sets but also analysis and visualization tools and services, analysis scenarios, and meta knowledge about them, and will definitely lead us to what we call open science.

2. Challenges for Data-driven Sciences

Bioinformatics has made the first big success among data-driven sciences to encourage other sciences to follow. Personalized medicine and material informatics are example followers. However, their researchers are gradually recognizing the difficulties to fill in the gap between varieties of available data analysis methods and the goals to find out new meaningful personalized treatments or new functional materials. This gap has two major causes.

In these data-driven sciences, most of the target systems are complex systems of systems in which more than one subsystem with different mechanisms interact with each other, and each of them is also a heterogeneous system, i.e., a mixture of more than one subsystem following either different mathematical models or the same model with different parameter values. In the machine learning of such a system, the learning data set inherently consists of more than one subset that follow different mathematical models or the same model with different parameter values. It is necessary to appropriately segment the learning data set into homogeneous subsets before applying the machine learning separately to each subset. Such segmentation is generally not an easy task. Furthermore, the size of each homogeneous data subset may often become too small for statistically meaningful analysis. Personalized medicine aims to find out a personalized treatment that works best for a specific patient, but not necessarily well for the others. The learning data set of patients is inherently a mixture of different types of patients with different chemo-responses. Each existing large-scale database of inorganic natural materials is also a mixture of different types of materials consisting of different atoms arranged in different structures. The total number of the learning data for a certain type of inorganic natural materials for which we can assume the same physical model for simulation and/or the same regression model for analysis may be in the order of 10^3, or 10^4 at most, which is definitely small for machine learning, and definitely not sufficient for the deep learning.

Besides the first cause of the gap, i.e., the heterogeneity of the learning data set and the comparatively small size of each homogeneous data subset, it is often difficult to define sufficient number of appropriate explanatory variables in providing the learning data set through measurement and/or simulation. In bioinformatics, “genome” constitutes substantial portion of explanatory variables. In material informatics, we also need its counterpart, i.e., “materials genome”. For proteins and peptides, a web server called PROFEAT computes structural and physicochemical features from amino acid sequence to systematically define a sufficient number of explanatory variables. It is a challenge, especially in inorganic material informatics, to systematically define a sufficient number of appropriate explanatory variables, i.e., inorganic materials genome.

3. Proposed Action Plan for Inorganic Material Informatics from a Computer Scientist’s View Point

In order to increase the size of each homogeneous subset of the learning data set, we may focus more attention on artificial inorganic materials than on natural ones. Examples may include those with amorphous structures and those with higher-order crystal structures of atom clusters. Such higher-order nanostructures and/or mesoscopic structures may increase not only the design parameters but also the value space spanned by these design parameter variables. An amorphous material, for example, may introduce two more design parameters, i.e., the average and the variance of its crystalline diameters. A super crystal of atom clusters may introduce the design parameters of both each atom cluster and the super crystal structure. These design parameters may work as explanatory variables of the learning data set, which may be provided by the simulation based on the first-principle-calculation modeling of the artificial materials and by databases of related physical properties of the involving atoms and crystal structures. We can compute only a sufficiently large finite number of simulations to calculate some functional properties of our concern. These functional properties of the materials may include conductivity, magnetic property, optical property, interfacial activity, catalytic activity, and bulk modulus. The machine learning for the regression using the simulation result as the learning data set will estimate the values of such physicochemical properties for arbitrary value combinations of explanatory variables for which the simulation is still missing.

It is not always possible to mathematically model the total system with all the physicochemical and structural parameters taken into account as explanatory variables for estimating some functional properties of our concern. The original idea of machine learning was to give a solution to this problem. Instead of assuming the knowledge about the underlying mechanism of the total system, it uses the observation records of the relation between a sufficiently large set of aspects and each functional property of the system as its learning data set to estimate this functional property value for an arbitrary new value combination of aspects. The success of machine learning heavily depends on the quality and the quantity of such aspects of the target system. Each aspect defines explanatory variables as parameters of its mathematical modeling. In the simplest case, an aspect defines a single explanatory variable.

Aspect modeling is different from the total-system modeling. It may use a simple model that may explain the specified aspect of the system. In naive application of machine learning to materials data, some material properties become difficult to estimate accurately. Material properties such as lattice constant and magnetic moment can be accurately estimated from simple descriptors, i.e., explanatory variable, using basic machine learning methods [1]. However, in the experiments, machine learning did not work well to estimate the material bulk modulus (the resistance to compression of the material). After adding new explanatory variables such as bond type, energy difference in compression and expansion, and density for the aspect modeling of the material bulk modulus, and calculating, for each record in the learning data set, the values of these added explanatory variables through the simulation of this aspect modelling, the bulk modulus could be well estimated.

Some aspect of our concern may be defined as a function of already defined explanatory variables. Depending on the types of machine learning, such an aspect may require the explicit introduction of a new explanatory variable as a derived variable, i.e., a function of other variables. In linear-regression machine learning, derived variables defined as linear combinations of other explanatory variables need not be explicitly introduced as new explanatory variables. They are implicitly considered by the algorithm if necessary. However, such a derived variable as x/y should be explicitly introduced as a new explanatory variable. Some indices obtained as analysis results such as cluster ids or pattern ids may sometime work as new explanatory variables for further segmentation and analysis. We call such explanatory variables marker variables or, simply, markers.

It should be noticed that the design of appropriate explanatory variables and the process of segmentation and analysis are both by their nature exploratory processes. This implies the importance of the development of an integrated exploratory visual analytics platform for data-driven sciences. A further shift toward open science requires not only the sharing of platform systems, but also a shared repository of data sets, analysis and visualization tools and services, analysis scenarios, and meta knowledge about them in reusable forms. Meme media and meme pool architectures [2] as well as their web-based implementation Webble World will answer these requirements.

Reference

1) K. Takahashi and Y. Tanaka, “Material synthesis and design from first principle calculations and machine learning,” Computational Materials Science, vol. 112, pp. 364–367, 2016.

2) Y. Tanaka, Meme Media and Meme Market Architecture. Piscataway; NJ; USA: IEEE Press, 2003.

Comments on Materials Informatics from a Researcher in Industry

Takeshi Nishiuchi
Magnetic Materials Research Laboratory, Hitachi Metals, Ltd., Osaka 618-0013, Japan

Over 30 years passed since invention of an Nd-Fe-B magnet, there are strong demands of “new materials” exhibiting characteristics more excellent than this magnet. To realize this matter, for example, there are many efforts to find out a new compound with better magnetic properties than $Nd_2Fe_{14}B$.

“Materials Informatics” is an approach which combines material sciences and data sciences, and has great possibility to change a way of development of new materials in industry in the future. Several national projects are promoted in Japan, and magnetic materials, especially permanent magnets, are one of the important targets of them.

In this talk, I will give personal comments on application of “Materials Informatics” for research and development of permanent magnets based on my own experiences in industry.

Perspective／展望

Satoshi ITOH／伊藤 聡
(Japan Science and Technology Promotion/JST)

A new national project concerning the materials informatics (MI) research has been started from July 1st 2015 in the NIMS; which called MI^2I (Materials research by Information Integration Initiative). In this project, a new data will be added to the materials database operated by NIMS, the tools required in the MI research will be developed, and a data-platform for materials research will be constructed. By using this platform, the effectiveness of the MI approach will be demonstrated in the development of magnetic materials including spintronics materials. Considering that many practical magnetic materials are multi-component compounds, we have to develop a more advanced searching system. A recent development in AI technology will play an important role in that way.

The MI approach will significantly reduce the time to discover, develop and manufacture new magnetic materials; in which a key issue is open and easy accessible database of the materials. The materials database contains crystal structure, composition rate, etc., but it is not enough. That is, in addition to materials data of the ideal state such as a perfect crystal, information of manufacturing processes in the actual material should be gathered in the materials database. However, production or manufacturing process usually is concealed as know-how. In order to promote the MI study, a policy regarding the handling of materials data including the know-how has become extremely important.