熱電特性データベースの構築と実験値マテリアルズ・インフォマティクスへの展開

桂　ゆかり*

1. はじめに

(1) 機械学習の特性と可能性

人工知能や機械学習と言う言葉をよく耳にするようになり，「将来，仕事が人工知能に奪われる」という懸念の声も聞こえる．しかし，これは表計算ソフトや検索エンジンのようなただの便利なツールである．よって，その特性を生かして自分の一部として利用することで，よりよい材料を早く開発して普及させることを目指すことが求められる．

数値計算がコンピュータによる理論科学の補助だとしたら，機械学習はコンピュータによる「長年の勘」の補助である．長年の勘とは，特に理論的裏付けはなくとも，膨大な経験をもとに「こうしたらこうなる」と予測する能力である．優れた勘を持つ人は，予想に反した結果を得るたびに，自分の中の予測式を改善することで予測精度を上げている．機械学習はこれを人間の処理能力をはるかに超えた数のデータに対して行う技術であり，人間が混乱するほどの情報を整理するのに役立つと期待できる．

機械学習では，多数のデータ(目的変数)とその「記述子(関係しうるデータ項目)」のセットを機械に与えることで，プログラムが自動的にデータ間の関係式を導く．このため，細かい場合分けのような明示的なプログラミングを行う必要がないのが利点である．

成功の秘訣は，どれだけ良いデータセットを作り，どんな記述子を指定してプログラムに「考え方」を指示するかというセンスが問われる．百程度のデータ数では意味のある結果が出ないことが多く，数千，数万というデータを解析していくことで初めて，そこそこ良い予測ができるようである．

(2) マテリアルズインフォマティックス(MI)の可能性

材料科学の情報学であるMaterials Informatics(MI)は，アメリカを中心に近年，急速な発展を遂げている．自動第一原理計算によって得られた数万件の電子構造データを世界に公開し，これを基軸に様々な機能材料の特性予測が進められている[1]．日本でも第一原理計算の研究者の一部は，早くからMIの可能性に気づいて近年目覚ましい成果を挙げている．物質・材料研究機構が2015年に立ち上げた情報統合型物質・材料研究拠点(MI^{2}I)では，そんな全国のMIの科学者と，新規にMIに参入する材料系研究者・情報系研究者の相互交流を通して，手探りでのMIプロジェクトが進んでいる．このプロジェクトに筆者も参画させていただいており，第一原理計算による新規熱電材料探索に取り組んだ経験を生かして，新たなMIのフィールドを切り開く研究に取り組んでいる．

(3) 次世代のMI：実験値MIの可能性

現在，世界で行われているMIの多くは第一原理計算データを使用するものである．これは，第一原理計算研究者がプログラミングに強いこと以外にも，自動計算スクリプトによって大量のデータを同じフォーマットで生成することが容易なためである．だが第一原理計算から材料物性のすべてが完全に予測できるわけではない．

MIの世界における未開拓分野は，実験データの活用にある．これがなかなか進まないのは，実験データを大量に取得することが難しいためである．難しいと言っても，データ収集作業そのものは複雑な実験や計算に比べたらはるかに簡単なはずである．難しいのは，データ収集を行うためのインセンティブの確保にある．

そこで，本記事では筆者が日本熱電学会において取り組んでいる，熱電特性データベースの作成プロジェクトについて紹介する．本格的なデータ収集活動はこれからであるが，他の金属材料分野に対しても何らかのヒントになれば幸いである．

* 東京大学大学院新領域創成科学研究科物質系専攻；助教(〒277-8561 柏市柏の葉5-1-5)，国立研究開発法人物質・材料研究機構(NIMS)統合型材料開発・情報基盤部門(MaDIS)情報統合型物質・材料研究拠点(CMI^2)；特別研究員
Construction of Thermoelectric Database and Its Application for Experimental Data Materials Informatics; Yukari Ktasura(*Tokyo University, Kashiwa)
Keywords: *materials informatics, database, thermoelectric materials, thermoelectric properties, web system*
2017年5月31日受理[doi:10.2320/materia.56.560]

2. 熱電特性への実験値 MI 導入の必要性

(1) 熱電材料開発の一般的な手法とその課題

本題に進む前に，熱電材料について簡単に紹介する．熱電材料とは，温度差と電位差を相互変換する材料であり，発電素子やペルチェ冷却素子としての応用が期待されている[2]．比較的高い性能を示す Bi_2Te_3 や PbTe はすでに実用化されているが，希少元素の Te や毒性元素の Pb, Te を用いる必要があるため，環境調和元素から構成された新規熱電材料の発見が望まれている．

実は，熱電材料の候補物質はかなり多い．これは，熱電特性はほぼすべての半導体および金属が示すものであり，作製法によっても特性を大きく変化させられるためである．しかし逆に，どれが有望な材料なのか判断が難しくなっている．1つの物質の有望性を調べるにも，膨大な数の試料を作製して判断する必要があるためである．

熱電発電と冷却の効率は，無次元性能指数

$$ZT=\frac{S^2\sigma T}{\kappa_{el}+\kappa_{ph}} \qquad (1)$$

が大きいほど高くなる．S は熱起電力(ゼーベック係数)，σ は電気伝導率，κ_{el} は電子熱伝導率，κ_{ph} はフォノン熱伝導率である．変換効率が 8%に達する $ZT>1$ が実用化の目安だと長く言われてきた．なお，$ZT\to\infty$でカルノー効率に達する．

式(1)を単純に読み解くと，S と σ が高く，κ_{el} と κ_{ph} が低い材料が有望であるということになる．しかしこれらは物質ごとに決まった値が定義できるわけではない．S, σ, κ_{el} はキャリアドープ量 n によって大きく変化する．たとえば n を大きくすると，σ が上昇し，S が低下する．この結果，ZT は n に対して山なりのカーブを描く．このカーブの全貌を見るため，材料研究者は，不純物ドープによって n の調整を試みる．

式(1)では独立パラメータに見える κ_{ph} も，フォノン散乱頻度によって大きく変化する．このため不純物ドープや結晶粒の微細化，格子欠陥の導入によって κ_{ph} の低減に取り組んでいる．うまく行くと，κ_{ph} は数分の1～数十分の1に低減させることができる．

このようにさまざまな条件で膨大な実験を行った結果，$ZT=1$ を超える明らかな高特性試料が1つでも見つかれば有望だとわかる．だが $ZT=1$ を示すと言われている試料を実際に自分で作ってみても，$ZT=0.1$にしかならなかったということは，熱電分野ではよくある事象である．

(2) 第一原理計算による熱電特性予測の限界

筆者はこれまで，第一原理計算による熱電特性の予測研究に取り組んできた[3]．しかし第一原理計算では一般にバンドギャップの計算誤差が大きく，S の予測値に大きな誤差が出てしまった．また，ボルツマン輸送方程式を使って解くと，σ と κ_{el} の計算に未知パラメータである電子緩和時間 τ_{el} が含まれてしまい，予測が難しかった．熱伝導率の計算も，フォノン緩和時間 τ_{ph} の情報が必要になるため難しい．σ, κ を直接計算する手法も研究が進められてはいるものの，計算できるのは純粋物質であり，不純物元素の多い実用熱電材料に適用するのは困難であった．

(3) 実験値 MI 導入にむけたデータベースの必要性

そんな複雑な熱電材料研究のデータを解析するには，実験データを用いた機械学習が適していると考えられる．熱電材料の実験系論文には，標準的に S, σ, κ の温度依存性のグラフが掲載されている．これらのデータの相互関係は人間の頭で把握するには複雑だが，機械学習であれば対処できる可能性がある．

問題は，過去に発表された熱電特性の実験データのほとんどは，画像ファイルの形で論文 PDF の中に眠っていることである．そこで，大量の論文から実験データを収集する活動に取り組むことにした．

3. 日本熱電学会におけるデータ収集プロジェクト

(1) データベース作成のインセンティブ

グラフ画像や文章などの表現物をコピー・転載することは著作権の侵害となりうるが，科学的な数値データの転載は，著作権には抵触しない[4]．

データベース作成作業は従来「研究成果にならない雑用」と見なされ，研究者自身は行わない文化があったが，Scientific Data などのデータジャーナルの創刊により，きちんと整理されたデータセットがあれば，新発見がなくても論文同等の研究成果にできるようになった．そこで，自発的なデータ収集サイクルの実現のため，作業の面倒くささをゼロに近づけ，データベースに収録されるメリットがそれを上回る仕組みを作ればよいと考えた．

(2) 熱電特性データベース WG の設立

文献からのデータ収集は，一人で行うには膨大過ぎる作業であるが，実現すれば多くの研究者の役に立つ．そこで筆者は2015年に，日本熱電学会において熱電特性のデータベース化プロジェクトを提案し，理事会での協議を経て，日本熱電学会熱電特性データベース WG(ワーキンググループ)を設立した[5]．この概要を**図1**(次頁)に表した．

このチームで行うデータベース作成作業は，学会の事業とはせず，学会に所属する研究者による「共同研究プロジェクト」と位置づけた．これにより，研究のためのさまざまな制度(電子ジャーナル，研究費，エフォートなど)を利用できるようにして，各参加者の研究成果につなげられるようにした．これは，事業化による有償化やアクセス制限などによる使い勝手の低下を防ぐ効果もあると考えた．

最初の取り組みとして，2016年3月に日本熱電学会の研究会「計算＆データ研究会」を1泊2日の日程で開催した．厳密にはデータ科学には第一原理計算は関係ないのだが，プログラミングに強い研究者に興味を持ってもらえる思える研究会にすることを目指してテーマを設定した．この結果，本データ収集プロジェクトについて知っていただけるきっかけにすることができた．

この後，全国各地の大学や研究機関で本 WG の説明会を

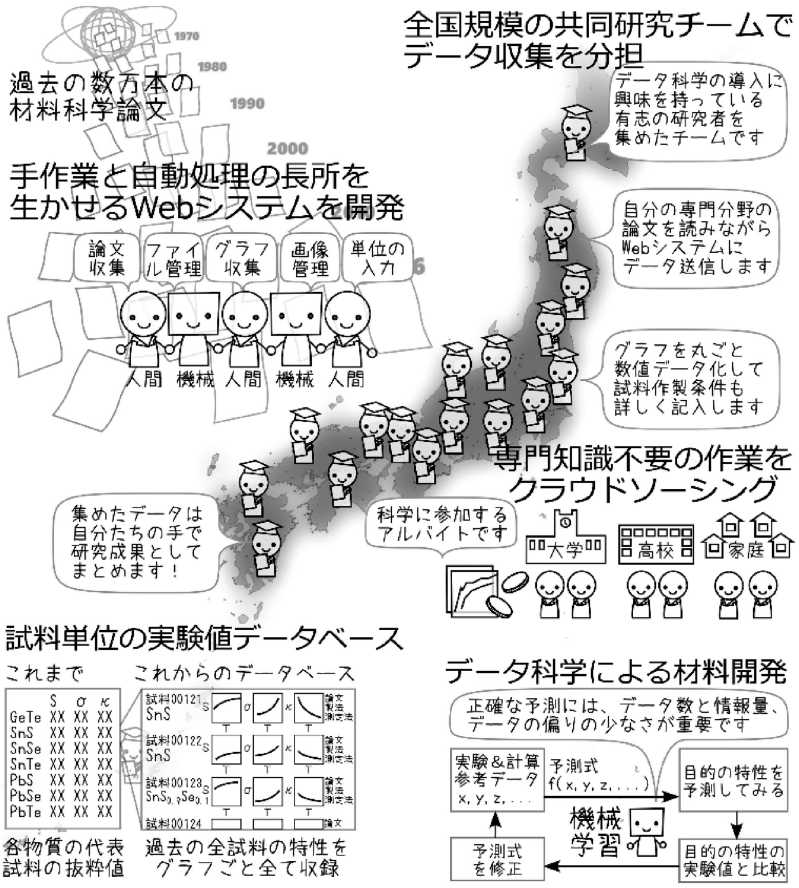


図１　本プロジェクトにおける文献データ収集のアプローチ．

開催し，興味を持っていただける研究者や学生さんを集めた．この時，できるだけ強制にならず，各個人の自発的な意志で参加するよう心がけることで，良い雰囲気のグループの構築に努めた．半年間の説明会で47名の WG メンバーを集めることができた．各メンバーが個々の研究で忙しいことを考慮すると，決して十分な人数とは言えないものの，面白いチームになっていくことを期待している．

4.　具体的なデータ収集の方法

データ収集作業の効率化に最も有効なのは，作業そのものの自動化ではなく，データ収集作業の前後にかかる無駄な時間の削減であると考えている．このため，グラフ上の点をクリックする作業を自動化するよりも，その前後でやらなければならない面倒なファイルとデータの管理を自動化することが有効である．もし今後その作業を代替しうる人工知能プログラムができた場合も，いきなりプログラムに任せるとチェックが必要になり二度手間になるので，作業者の入力補助という形で導入するのがベストであると考えている．

(1)　実験値 MI 専用データベースの設計

従来の材料特性データベースには，ほしい情報を簡単に調べるためのレファレンスとしての役割が期待されていた．このため，実験条件などは詳しく書かず，１つの物質の特性を１試料に代表させ，その中の測定値を１点記録しておけば良かった．

だがこのようなデータベースを使い，どの物質が最も良いか比較しても，物質間のフェアな戦いは実現できない．最高特性試料が選定された物質と，純粋すぎて特性は良くない単結晶試料が選定された物質と，研究され始めたばかりでまだ特性の悪い試料しか報告されてない物質が，同じ土俵で戦う状況になるためである．

これを防ぐ単純なアプローチは，データの選定は試みず，目的の物性が載った論文があれば，そこに掲載されているすべての試料を収集対象とすることである．化合物単位ではなく，試料単位で収録することで，互いに矛盾するデータがあっても気にする必要がなくなる．これは，データ収集者の心理的負担を大きく減らすことにもつながる．

もちろんデータ選定をしないことで，信頼できないデータが混入する可能性もある．だが，信頼性の有無は主観で判断するよりも，データに基づいて判断する方が科学者として真摯な態度であるとも言える．このため，明らかに間違っているものを除いて，基本的には収録するべきデータだと考えている．

(2)　データ収集対象論文の選定

過去の全文献をデータベース化するには，大量の論文ファイルを入手する必要がある．そこでまず，対象論文のリストを作成した．Scopus[6]を利用して，論文の書誌情報を一括ダウンロードした．これを表計算ソフトで開き，著者やタイト

ル，雑誌名やDOI(Digital Object Identifier)，フルテキストのダウンロードページへのリンクなどの情報を一括取得した．

はじめに，“Thermoelectric”というキーワードで検索をかけ，約48000件分の論文リストを取得した．だが，これだけでは材料特性以外の論文が多く含まれてしまい，検索キーワードによるそれらの除外が困難であった．そこで，タイトル中に何らかの物質名(化学式，元素名，鉱物名など)を含む論文を自作のスクリプトで抽出することにより，論文のリストを約22000件まで縮めた．

⑶ 論文フルテキストのダウンロード作業

このリストに基づき，研究業務員2名の力を借りて，手作業で1本ずつPDFファイルをダウンロードした．進捗管理のため各論文に独自のIDを設定し，保存時のファイル名も予め決めておくことで，ダウンロード作業が効率化できた．一部，所属機関からアクセスできない論文もあったが，9か月で約15000本弱の論文を収集できた．

なお，プログラムによる論文の自動ダウンロードは，テキストマイニングAPIを除き，出版社から禁止されていることが多い．時間当たりのダウンロード数を厳しく監視している出版社もあり，万一不自然なアクセスだと認識されると，自分のIPアドレスからのアクセスがブロックされるリスクがある．今回のダウンロード作業ではブロックされることはなかったものの，ダウンロード作業が速すぎないように注意した方が良い．一括ダウンロード機能を搭載する論文検索システムであっても，同様の注意が必要である．

⑷ 論文からのグラフ画像とデータの抽出

論文中のグラフ画像を独立した画像ファイルとして保存すると，データの抽出と管理が容易になる．そこで，自動と手動の2種類の画像抽出法を用意した．自動抽出では思うようにいかない論文(ロゴマークや文字など本文に関係のない画像の多い論文，短冊状や格子状に分割された画像から構成される論文，図表を構成する円や線などの図形パーツが独立に検出されてしまう論文，目的のグラフ画像がなぜか検出されない論文などである．手動抽出ソフトは，論文PDFのスナップショット画像からクリップボードを経由してPNG形式の画像ファイルを保存するソフトを開発した．論文やグラフ情報を含めたファイル名の設定を補助することで，手動画像抽出も効率化できた．

⑸ データ収集Webシステムの開発

本研究では，グラフからのデータ収集作業を全国規模で共有・分担できるWebシステムを開発した[7]．Webシステムとすることで，作業者側でのソフトウェアの更新が不要となり，Windows, Macintosh, Linux以外にもiOSやAndroidなどのタブレットの利用が可能となった．

開発は，日本熱電学会の会員である熊谷将也氏(2015年当時博士課程2年)が引き受けてくれた．彼は実験と計算による熱電材料開発の研究をしていたが，データ科学とWebプログラミングの知識も持っている．論文をまたがって同じ作業を繰り返せる仕様とすることで，最小限のクリック数でデータを収集できるようにすることを目指した．試料情報を論文単位で管理することで，別のグラフに同じ試料が登場した場合も，同じ試料のデータとして関連付けて登録できる．

⑹ 試料情報の記入フォーマットの設計

実験値データベースのフォーマットには，まだ世界標準が存在しないため，自分たちで設計する必要がある．収集した論文をテキストマイニング技術で解析し，情報を体系化するところから始める予定である．学会内の幅広い分野の材料研究者で検討して，「自分たちの研究に最もしっくりくるフォーマット」を設計することが重要である．

5. 終 わ り に

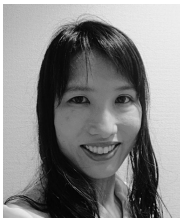
過去の大量の論文上の実験データをデータベース化するという仕事を誰よりも早く達成すれば，それを用いたデータ科学研究という形で，材料科学のフロンティアを切り開ける可能性がある．熱電材料のデータ収集をモデルとしてデータ収集方法が確立できたら，他の材料科学分野でのデータ収集に広げていきたいと考えている．

もし他の金属材料分野で，同様のデータベース化を進めたい方がいらっしゃれば，技術やノウハウ，データの共有を通して，ぜひ連携して進めたいと思う．

文献データ収集Webシステムの開発に大きくご貢献して下さった熊谷将也氏と，論文PDFのダウンロードを手伝っていただきました今井庸二氏と郡司咲子氏に深く感謝申し上げます．熱電特性データベースWGのメンバーと，日本熱電学会理事会の先生方の温かいサポートに，この場を借りてお礼申し上げます．本研究は，科学研究費補助金(挑戦的萌芽研究16K14379)と，科学技術振興機構(JST)のイノベーションハブ築支援事業「情報統合型物質・材料開発イニシアティブ(MI2I)」から支援を受けたものです．

文 献

（1）A. Jain, K. A. Persson and G. Ceder: APL Mater., **4**(2016), 053102.
（2）G. J. Snyder and E.S. Toberer: Nat. Mater, **7**(2008), 105.
（3）桂ゆかり：日本金属学会誌，**79**(2015), 633–637.
（4）R. Elliott: Learned Publishing, **18**(2005), 91–94.
（5）桂ゆかり：日本熱電学会誌，**12**(2016), 16–22.
（6）Scopus, Elsevier (2012). URL: https://www.scopus.com/
（7）桂ゆかり，熊谷将也，今井庸二，郡司咲子，木村 薫：粉体および粉末冶金，**64**(8)(2017).



桂 ゆかり

★★★★★★★★★★★★★★★★★★★★★★
2009年3月 東京大学大学院工学系研究科博士課程修了
2009年4月–2012年8月 理化学研究所 基礎科学特別研究員
2012年9月–2015年7月 東京大学 特任研究員
2015年8月 現職
2015年7月 物質・材料研究機構 特別研究員(兼任)
★★★★★★★★★★★★★★★★★★★★★★