粉体および粉末冶金, 第 64 巻 8 号（2017 年）pp. 467-470. *J. Jpn. Soc. Powder Powder Metallurgy*, **64**, 8 (2017) 467-470.
研究論文

実験的熱電特性のデータベース化に向けた論文データ収集 Web システム Starry data の開発

Development of "Starry data" Web System for Data Curation of Published Experimental Thermoelectric Properties

桂　ゆかり[1,2*], 熊谷　将也[3], 郡司　咲子[2], 今井　庸二[2], 木村　薫[1]

[1] 東京大学大学院新領域創成科学研究科物質系専攻, 〒277-8561 柏市柏の葉 5-1-5. [2] 物質・材料研究機構 情報統合型物質・材料研究拠点, 〒305-0047 つくば市千現 1-2-1. [3] 大阪大学大学院工学研究科 環境・エネルギー工学専攻 〒565-0871 大阪府吹田市山田丘 2-1

Yukari KATSURA[1,2*], Masaya KUMAGAI[3], Sakiko GUNJI[2], Yoji IMAI[2], Kaoru KIMURA[1]

[1] Dept. Advanced Materials Science, Graduate School of Frontiers Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan. [2] Center for Materials Research by Information Integration, National Institute for Materials Science, Japan. [3] Dept. Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871 Japan.

*Corresponding Author, Email: katsura@phys.mm.t.u-tokyo.ac.jp

ABSTRACT
Although numerous papers are published each year, most of the experimental data reported in those papers are only available as two-dimensional plot images. Data-driven materials science using the machine learning technologies will be accelerated by gathering those published experimental data into a database. By taking thermoelectric materials as a test case, we attempted to optimize the processes of collection of papers, extraction of numeric data from plot images, and sample-based data storage into a database. By searching with a keyword "thermoelectric", we obtained a list of 47,936 papers. Among these papers, we selected 18,471 papers as possible papers with thermoelectric properties, and succeeded to download 14,835 full-text PDF files. We developed a web system named "*Starry data*", to assist the sequential data extraction from the images contained in those PDF files. This system also assists materials scientists to annotate experimental samples efficiently, to develop a descriptive database that can be used for machine-learning of the complex, sample-dependent materials properties.
KEYWORDS
materials informatics, materials database, data curation, thermoelectric materials

1　緒言

温度差と電位差を相互変換する熱電変換材料は, 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. これらへの応用にはmmスケールの多結晶体バルクが必要であり, 特性改善のため緻密化と微細組織制御が必要であるため、多様な粉末冶金技術が利用されている.

熱電特性はほぼすべての半導体および金属が持つ性質であるため, 熱電材料の候補物質は多いが, どの物質が最も有望であるか予測することは簡単ではない.

熱電発電と冷却の効率は, 無次元性能指数

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

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. (1)式を単純に読み解けば, S と σ が高く, κ_{el} と κ_{ph} が低い材料が有望で

あることになるが，いずれも独立に制御できず，同じ物質でも組成や製法のわずかな違いで S, σ, κ が何桁も変化してしまう．この結果，試料依存性が物質依存性よりも大きくなってしまう．これらのしがらみの中で，ZT が高くなる絶妙な組成や製法を探索する複雑な研究が，熱電材料開発である．

これらの情報整理の難しさを乗り越え，有望な熱電材料を高速に開発する手段として，筆者は機械学習に着目している．機械学習はコンピュータによる「長年の勘」の補助であり，人間の処理能力をはるかに超える膨大なデータから傾向を学習することができる．近年，材料科学に機械学習などのデータ科学を導入した「材料（マテリアルズ）インフォマティクス」が世界で急速に発展している．自動第一原理計算を用いて数万種類の無機/有機化合物の計算データが無償で公開されている．熱電材料に関するデータベースの整備と機械学習も進行しており，2000 種類以上の実在物質の熱電特性予測の第一原理計算結果を掲載した TEDesignLab [1-2]，100 本ほどの熱電特性の文献からデータを抽出して収録した MRL Datamining Chart (Energy Materials Datamining) [3]，これらのデータと外部データベースを融合して機械学習を行い，S, σ, κ, ZT の有望性を物質単位で数値化した Citrination [4]が公開されている．

ただ，機械学習の精度は元のデータセットとその整理の仕方に大きく依存する．そこで現在よりもはるかに多くの実験データを集めたデータベースを，試料依存性を明示的に取り込んで構築すれば，より本質をとらえた熱電特性予測が可能となると期待できる．

実験データをグラフから数値データとして抽出すれば，出版社の著作権には抵触しないと考えられる．著作権とは表現物（文章・画像などの創作物）としてのオリジナリティに付随するものであり，文章やグラフ画像に著作権は存在しても，数値データそのものには著作権はない [5]．また，科学論文として投稿して受理されている時点で，その内容自体は創作物ではなく普遍的事実であるという前提であり，普遍的事実に著作権は主張できない．そして実験データそのものについては，研究が行われた機関に帰属している．

論文からのデータ収集作業が何十年も進まない最大の原因は，インセンティブの不足と，作業工程の最適化の不十分さにあると考えている．1 本の論文からデータを抽出するだけであれば，目的のグラフ画像を保存して，WebPlotDigitizer [6]などのグラフトレースソフトでデータ点の x, y 値を読み取ることで簡単に元の数値データが取得できる．だが論文数が増えていくに従い，グラフトレース以外の前後処理：たとえば対象論文・試料の選定，書誌情報・数値データ・試料情報の管理，単位換算などが複雑に絡み合ってくる．グラフトレースを自動化しても，前後処理まで自動化しなければ作業は煩雑なままであり，人間によるデータ確認作業まで加わることでさらに面倒になってしまう．よって論文からの大量データ収集には，論文選定からデータ保存までの一連のプロセスを最適化し，作業者が迷うことなく進められる環境の構築が第一である．

2 方法

そこで本研究では，電子ジャーナル利用規約や著作権を侵害せずに大規模なデータ収集を行うため，論文中のグラフから実験データを効率的に収集するプロセスを設計した．データ収集対象は，熱電特性の論文に標準的に掲載されている S, σ (または電気抵抗率 ρ), κ, 性能指数 $P=S^2\sigma$, ZT の温度依存性のグラフとした．文献検索システム Scopus [7]を用いて論文リストを入手後，熱電特性を含む可能性の高い論文を抽出して，ダウンロードを行った．複雑な論文データ管理作業を自動化することで，論文 PDF から効率的にデータを抽出できる Web システムを，Ruby on Rails [8]フレームワークを用いて開発した．

3 結果

3.1 データ収集対象論文リストの作成

Scopus より “Thermoelectric”というキーワードで約 47,936 件分の論文リストを取得し，表計算ソフトでそれらのタイトルを閲覧したところ，約半数程度はシステム応用，伝熱シミュレーションなどの非材料系論文であると推測された．そこで，論文タイトルを手掛かりに材料系論文と非材料系論文を分類することを試みた．材料系論文はタイトルに物質名（化学式，元素名，鉱物名など）を含むという傾向に着目し，このような単語の有無を自作のスクリプトで判定した．物質名の可能性が高い単語の特徴として，下付きタグを含む単語，元素名，-ide, -ite, -ium で終わる単語のうち一般的な単語を除いたもの，2 文字目以降に大文字がある単語，元素記号の組み合わせと数字・算術記号・x, y, z だけで構成される単語を指定した．この結果，一割未満の判定ミスは残ったものの，材料系および非材料系の論文を大まかに分類でき，本研究に関係しうる論文 18,471 件のリストを作成できた．

3.2 論文フルテキストのダウンロード

論文 PDF は，必要な論文ごとに Web ブラウザで出版社のページにアクセスすることでダウンロー

ドした. 論文リストは表計算ソフトでジャーナル名順に並び替え, 進捗管理のため各論文に 5 桁の通し番号を付与した. ダウンロードページの URL を記載し, PDF の保存先ファイル名をあらかじめ定義しておくことで, 大量ファイルの管理を容易にした. この結果, 14,835 本の論文 PDF を収集することに成功した.

3.3 データ収集 Web システム *Starry data* の開発

論文 PDF からのデータ収集を行う Web システム *Starry data* を, 熊谷が中心となって開発した. ソースコードは公開していないが, URL は準備が整い次第公開し, 本論文または今後出版する本システムについての英語論文の引用を条件に無償利用を許可する予定である.

本 Web システムは以下のように動作する. Web ブラウザでトップページにアクセス後, メールアドレスとパスワードを入力すると各ユーザーのページにログインできる. Fig.1 に示した「取得」タブでは, データ抽出を行う PDF ファイルのアップロードと, その PDF ファイルから検出された画像からの不要画像の削除, 画像からグラフをトレースして集めた数値データの閲覧ができる. 著者, 雑誌名などの書誌情報は PDF ファイル中に書かれた DOI (Digital Object Identifier)を手掛かりに, 事前に本 Web システムに登録しておいた書誌情報と自動照合することで取得される. 「分類」タブでは PDF ファイルから自動抽出した画像に対して, 各軸の物理量, 単位, 掲載試料の情報などを入力できる. 「抽出」タブでは, グラフ画像からのデータ点のトレース作業を, 簡単なマウスクリックによって行うことができる. また分類, 抽出タブでは, 入力を補助するため論文のアブストラクトの閲覧, 出版社の Web ページへのアクセス, アップロードした PDF の閲覧を可能にしている. なお, 電子ジャーナル利用規約の遵守のため, PDF ファイルはアップロードした本人のみしか閲覧できない仕様とし, 画像を利用する際も引用元を示した上で利用している.

4 考察

4.1 対象論文リストの入手と分類

今回の論文分類法は, タイトルから物質名を認識できなかった材料系論文（特に有機系論文）を見落としており, 改善が必要である. 材料系論文として分類された中にも, 理論計算のみの論文など実験データを含まない論文も多く, これらの選別が必要である. アブストラクトや本文の文字列を利用した自動選別法を開発すれば, より高速に対象論文を判別できると期待できる.

4.2 論文フルテキストのダウンロード

今回の論文 PDF の収集には時間がかかりすぎており, 改善が必要である. プログラムによる自動ダウンロードは電子ジャーナルの利用規約において禁止されているが, 出版社と所属機関の間でテキストマイニング API (Application Programmable Interface)の利用を契約すれば可能である. この契約は出版社によって費用が異なる上, 利用規約による用途制限が多く, 入手できない文献もまだ多いという課題もある. ただし, 論文数に関する制限はほとんどないため, これを有効活用することで高速な論文収集も可能となると期待できる.

4.3 データ収集 Web システム Starry data

PDF 中に DOI が書かれていない論文や, 書誌情報がデータベースに未登録である最新論文への対応が必要である. ファイル形式が出版社や年代, 著者により大きく異なるため, 画像の自動抽出が困難な論文もあり, 代替プロセスの用意が必要である. データの自動補間と, 一括データダウンロードの効率化, 実験方法の入力フォーマットの最適化も必要である.

5 結言

熱電材料開発というひとつの学術分野全体の論文を収集することで, そこに掲載されている実験データを数値データの形で大量に収録した試料単位データベースの構築に取り組んだ. 独自開発の Web システムでデータを一括管理することで, 大量の論文からのデータ収集の効率化に成功した. ただし, 現状のシステムでは対応できない論文や時間のかかる工程もあり, さらなる改善が必要である. 今後は, 収集した論文からのデータ収集作業に取り組む. このような論文上の実験データの収集は熱電材料開発以外にも有用だと期待できるため, ここで得られたノウハウを他の材料科学分野にも生かすことができれば幸いである.

6 謝辞

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構築支援事業「情報統合型物質・材料開発イニシアティブ (MI^{2}I)」から支援を受けたものです. 本プロジェクトに関してあたたかいご協力をいただきました, 日本熱電学会熱電特性データベース WG のメンバーの皆様と同学会理事の先生方, MI^{2}I のメンバーの皆様に深く感謝申し上げます.

文献

1) P. Gorai, D. Gao, B. Ortiz, S. Miller, S. Barnett, T. Mason, Q. Lv, V. Stevanovic, and E. S. Toberer:

Comp. Mat. Sci. **112** (2015) 368-376.
2) J. Yan, P. Gorai, B. Ortiz, S. Miller, S. Barnett, T. Mason, V. Stevanovic, and E. S. Toberer: Energy Environ. Sci. **8** (2015) 983-994.
3) M.W. Gaultois, T. D. Sparks, C.K.H. Borg, R. Seshadri, W.D. Bonificio, and D.R. Clarke: Chem. Mater., **25** (2013) 2911–2920.
4) M.W. Gaultois, A.O. Oliynyk, A. Mar, T. D. Sparks, G.J. Mulholland, B. Meredig, arXiv:1502.07635 (2015).
5) R. Elliott, Learned Publishing **18** (2005) 91–94.
6) A. Rohatgi, WebPlotDigitizer (2011), URL: http://arohatgi.info/WebPlotDigitizer/app
7) Scopus, Elsevier, URL: https://www.scopus.com/
8) Ruby on Rails, URL: http://rubyonrails.org/
9) J.H. Kim, N.L. Okamoto, K. Kishida, K. Tanaka, H. Inui, Acta Materialia, **54** (2006) 2057-2062.

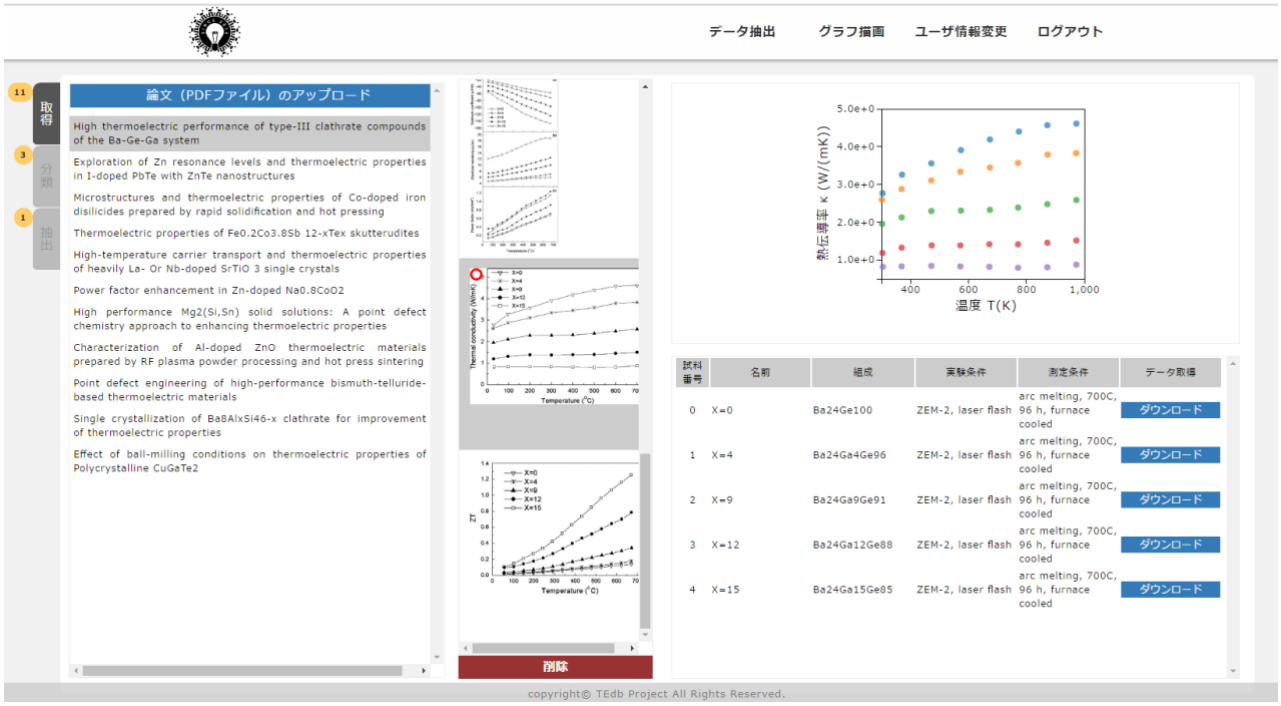

Fig.1 A screen shot of the browsing tab of Starry data web system, showing the list of the processed papers, the list of the images extracted from the selected paper[9], and the experimental datasets collected from the selected image.