様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月　　５ 日現在

機関番号：１２６０１
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１４３７９

研究課題名（和文）文献データ収集支援システムの開発による大規模高次元物性データベースの構築

研究課題名（英文）A large-scale high-dimensional material property database by a literature data collection system

研究代表者

岡本　ゆかり（桂ゆかり）（Okamoto (Katsura), Yukari）

東京大学・大学院新領域創成科学研究科・助教

研究者番号：００５５３７６０

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：論文中のグラフ画像から実験データを効率的に収集する世界初のWebシステムStarrydataを開発し、材料インフォマティクスのオープンデータベースを公開した。応用例として、熱電材料についての実験データ収集を進め、約2,500本の熱電材料に関する論文から15,000試料を超える膨大な実験データの収集に成功した。ベストデータに偏ることなく幅広い特性の試料を収録した点、データ一括ダウンロード機能を搭載した点で画期的である。得られた熱電特性データと第一原理計算データを合わせて電子緩和時間の評価に成功した。さらに、化学組成を入力とした熱電特性の機械学習を行い、平均誤差85%程度の熱電特性予測に成功した。

研究成果の学術的意義や社会的意義
あらゆることがデータとしてAIに利用される現代、科学論文に掲載された実験データが、デジタルデータとして全く手に入らないのは時代遅れである。そこで本研究で開発したのは、論文中のグラフ画像から効率的に実験データを収集してデータベースを作成できるStarrydata webシステムである。例として約2500本の熱電材料の論文から約15000試料の実験データを収集して世界最大の熱電材料データベースを作成し、そのデータ解析により優れた熱電材料の特徴を解明した。今後は他の材料科学分野についても同様のデータベースを作成することで、AIを活用した材料開発の進歩に貢献したい。

研究成果の概要（英文）：We developed the first web system named Starrydata to collect experimental data from plot images in published papers, to generate an open database for materials informatics. As an application, we collected experimental thermoelectric properties of over 15,000 samples from over 2,500 papers. Our database recorded data from samples in wide variety of properties, without biases on the best data. By combining the data with a first-principles calculation, we succeeded to evaluate electron relaxation times of samples. We carried out machine learning of thermoelectric properties by using chemical composition as an input, and succeeded in prediction of thermoelectric properties with an average accuracy 85%.

研究分野：材料科学

キーワード：材料インフォマティクス　データベース　熱電材料　第一原理計算　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

1. 研究開始当初の背景

物性科学の世界では、「自分のデータと文献上のデータを直接同じグラフで比較する」という単純な解析は、十分に行われているとは言い難い。これまで膨大な予算を投じて得られた過去の研究データを活用せず、研究者が変わるたびに新しいデータを生産するのは宝の持ち腐れである。実験と理論の相互比較をしにくいことによる弊害や、類似データ、トップデータを検索できる共通の場所がないことによるトラブルもある。

現在、Materials Informatics (MI)という、情報科学を活用した物性研究手法が注目を集めている。アメリカを中心に、自動第一原理計算により数千種類の化合物に関する電子構造のデータベースを構築し、その情報科学的解析による新規機能材料の探索が活発化している。次なる MI のデータソースは膨大な数の論文から取得した実験データであるべきだと考えている。さまざまな方法・条件で得られた多数の実験データには、人工的に得られたデータには存在しない、多くの未知の傾向や情報が眠っていると考えられる。

文献に掲載された物性値を集めてデータベースを作ろうという試みはこれまでにもいくつもあるが、いずれも研究分野のごく一部をカバーするに留まっている。データ量の大幅な不足を補うため、近年、テキストマイニングやグラフの画像解析により自動的に文献データを収録する仕組みの開発も進んでいるが、そのように取得したデータでは、実用上最も重要である補助情報（各試料の作製条件や測定方法など）に欠けた表面的なデータベースとなってしまう。

2. 研究の目的

そこで本研究では、あらゆる文献（論文）に掲載された物性値を、作製条件や計算条件などのメタデータとともに、丸ごと数値データとして登録した文献値データベースを構築することを目指した。これにより軸や単位の違いを気にせずに、文献調査や論文の枠を超えた考察を容易にして、次世代の物性研究スタイルを確立することを目指した。

3. 研究の方法

本研究では、世界中の論文にグラフ画像として公開されている実験データを大規模に集積できるデータベースを構築した。独自の Web システム Starrydata の開発によりデータ収集プロセスの大幅な効率化を目指した。完全自動化を目指すのではなく、論文の取得とデータの選択・判断は人間が行い、登録されたデータの管理をシステムで行う仕組みを選んだ。文献管理ソフトウェアに似せたユーザーインターフェイスと、無駄のないデータ収集画面、登録データの即時グラフ表示、自動単位変換機能により、データ登録の手順を大幅に簡略化した。

プロトタイプとして熱電特性データベースの構築を行った。この例として PbTe 系熱電材料の実験データの解析を行った。Starrydata を用いて、3 人のデータ収集者とクラウドソーシングによって、過去の熱電特性の論文から熱電特性の実験データを収集した。

4. 研究成果

データ収集ワークフローの最適化によって、手作業にも関わらず、1 時間平均で論文 1 本、グラフ 5 枚、7 試料、20 カーブ、500 データ点という高速でのデータ収集に成功した。これをコツコツ続けることで、既存の熱電特性データベース(UCSB Thermoelectrics Data)の 50 倍以上となる、2500 論文、15000 試料以上の大規模実験値データベースの作成に成功し、このデータベースを世界に公開した。このデータベースは国内外の学会において高い評価を受け、企業とのコラボレーションによる他材料分野のデータ収集への展開も開始した。

鉛テルル系熱電材料 434 試料の熱電特性の実験データの解析からは、熱電特性の温度依存性に強い試料依存性があることがわかり、熱電特性を1つの代表試料で示すという従来の解析の限界が示された。だが、横軸を電気伝導率、縦軸をゼーベック係数とした Jonker plot を作成することで、電子構造を反映した材料系の特徴が浮かび上がった。これは、特性の良い試料のみならず、多数の特性の悪い試料を含めたことにより、初めて浮かび上がった特徴であった。

この実験データ群を用いて、第一原理計算結果に含まれる未知変数である電子緩和時間を算出した。すると、電子緩和時間は、多くの第一原理計算で仮定されていたような定数ではなく、作成方法によって二桁以上も変化し、無次元性能指数 ZT に大きな影響を与える変数であることがわかった。これは、高特性熱電材料の設計に、実験データと第一原理計算を合わせた電子緩和時間の計算が必要であることを示す結果であった。

さらに、この PbTe の実験データ群を、ニューラルネットワークを用いて機械学習したところ、試料の化学組成のみからゼーベック係数などの熱電特性を高精度で予測することに成功した。これらは実験データと第一原理計算の融合や、機械学習の導入が、物性科学の新発見につながることを示す結果である。Starrydata に収録されたデータの多くは未解析であり、今後他のデータ群についても同様の知見が得られると期待できる。

5. 主な発表論文等

〔雑誌論文〕（計5件）

1. Yukari Katsura, Masaya Kumagai, Takushi Kodani, Mitsunori Kaneshige, Yuki Ando, Sakiko Gunji, Yoji Imai, Hideyasu Ouchi, Kazuki Tobita, Kaoru Kimura, Koji Tsuda, "Data-driven analysis of

electron relaxation times in PbTe-type thermoelectric materials",Sci. Tech. Adv. Mater., (2019) In Press. DOI: 10.1080/14686996.2019.1603885 (査読あり)

2. Yukari Katsura, Hidenori Takagi, Kaoru Kimura, "Roles of Carrier Doping, Band Gap, and Electron Relaxation Time in the Boltzmann Transport Calculations of a Semiconductor's Thermoelectric Properties", Mater. Trans. 59, 7 (2018) DOI: 10.2320/matertrans.E-M2018813 (査読有)
3. 桂ゆかり, 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, まてりあ, 56, 9 (2017) 560-563. DOI:10.2320/materia.56.560 (依頼執筆・査読無)
4. 桂ゆかり, 熊谷将也, 今井庸二, 郡司咲子, 木村薫, 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, 粉体および粉末冶金, 64, 8 (2017) 467-470. DOI:10.2497/jjspm.64.467(査読有)
5. 桂ゆかり, "日本熱電学会 熱電特性データベース WG とデータ収集プロジェクトのご案内", 日本熱電学会誌 12, 3, 16-22 (2016).(査読無)

〔学会発表〕(計 25 件)

1. (Invited) Yukari Katsura, Masaya Kumagai, Riku Sato, Takushi Kodani, Mitsunori Kaneshige, Yuki Ando, Sakiko Gunji, Yoji Imai, Hideyasu Ouchi, Kazuki Tobita, Kaoru Kimura, Koji Tsura, "Materials informatics of thermoelectric materials using big literature data", TMS2019 148th Annual meeting and exhibition, Henry B. González Convention Center (口頭, San Antonio, TX, USA, 2019/3/12-15).
2. Yukari Katsura, Masaya Kumagai, Mitsunori Kaneshige, Yuki Ando, Sakiko Gunji, Yoji Imai, Riku Sato, Takushi Kodani, Hideyasu Ouchi, Kazuki Tobita, Kaoru Kimura, Koji Tsuda , "Starrydata: a plot mining web system for experimental materials informatics, "PRESTO International symposium on materials informatics, 東京大学本郷キャンパス (ポスター, 東京都文京区, 2019/2/8-10).
3. (招待講演)Yukari Katsura, Masaya Kumagai, Mitsunori Kaneshige, Yuki Ando, Sakiko Gunji, Yoji Imai, Riku Sato, Takushi Kodani, Hideyasu Ouchi, Kazuki Tobita, Kaoru Kimura, Koji Tsuda, "論文からの実験データ収集 Web システム Starrydata の開発と熱電材料研究への応用", MaDIS シンポジウム 2019 AI で加速する材料開発とデータプラットフォーム戦略, 東京ビッグサイト会議棟 (口頭, 東京都港区, 2019/1/30).
4. 桂ゆかり, 熊谷将也, 小谷拓史,佐藤陸,金重光則, 安藤有希,郡司咲子,今井庸二,木村薫,津田宏治, "熱電材料の大規模実験データを用いた Materials Informatics", 第 39 回日本熱物性シンポジウム, 愛知県産業労働センター ウインクあいち 11 階 (口頭, 愛知県名古屋市, 2018/11/13-15).
5. 桂ゆかり, 熊谷将也, 安藤有希, 郡司咲子, 今井庸二, 木村薫, "論文からの熱電特性データベース構築プロジェクトの課題と展望", 第 15 回日本熱電学会学術講演会(TSJ2018), 東北大学青葉山キャンパス (口頭, 宮城県仙台市, 2018/9/13-15).
6. Masaya Kumagai, Yukari Katsura, Mitsunori Kaneshige, Takushi Kodani, Hideyasu Ouchi, Sakiko Gunji, Yuki Ando, Yoji Imai, Kaoru Kimura, Koji Tsuda, "A web application "Starrydata" for collecting and sharing plot data on published papers", 37th International Conference on Thermoelectrics, Congress Center (口頭, Caen, France, 2018/7/1-5).
7. Masaya Kumagai, Yukari Katsura, Riku Sato, Mitsunori Kaneshige, Takushi Kodani, Yuki Ando, Sakiko Gunji, Yoji Imai, Hideyasu Ouchi, Kaoru Kimura, Koji Tsuda, "Data-driven materials design from large-scale experimental data", 37th International Conference on Thermoelectrics, Congress Center (口頭, Caen, France, 2018/7/1-5).
8. Takushi Kodani, Yukari Katsura, Masaya Kumagai, Yuki Ando, Yoji Imai, Sakiko Gunji, Kaoru Kimura, "Development of high-performance thermoelectric materials guided by large-scale experimental data", 37th International Conference on Thermoelectrics, Congress Center (口頭, Caen, France, 2018/7/1-5).
9. (Invited) Yukari Katsura, Masaya Kumagai, Mitsunori Kaneshige, Takushi Kodani, Hideyasu Ouchi, Sakiko Gunji, Yuki Ando, Yoji Imai, Kaoru Kimura and Koji Tsuda, "Starrydata2: 論文中のグラフからの材料特性データ収集システム a web system for materials data collection from published plots,"Japan Open Science Summit, 学術総合センター(口頭, 東京都千代田区, 2018/6/18-19).
10. 佐藤陸, 小谷拓史, 桂ゆかり, 熊谷将也, 今井庸二, 郡司咲子, 木村薫, "ニューラルネットワークに基づくPbTe 系熱電材料の特性予測", 第 65 回応用物理学会春季学術講演会, 早稲田大学西早稲田キャンパス (口頭, 東京都新宿区, 2018/3/17-20).
11. 熊谷将也, 桂ゆかり, 小谷拓史, 大内秀恭, 郡司咲子, 安藤有希, 今井庸二, 木村薫, "論文内グラフデータを収集・共有できる WEB システムの開発", 第 65 回応用物理学会春季学術講演会, 早稲田大学西早稲田キャンパス (口頭, 東京都新宿区, 2018/3/17-20).

12. 小谷拓史, 桂ゆかり, 熊谷将也,今井庸二, 郡司咲子, 木村薫, "実験値データベースの応用による高性能 SnTe 熱電変換材料の開発", 第 65 回応用物理学会春季学術講演会, 早稲田大学西早稲田キャンパス (口頭, 東京都新宿区, 2018/3/17-20).
13. (招待講演)桂 ゆかり, "熱電特性の実験値マテリアルズインフォマティクス", 第二回かけはし研究会「温度変化で発電するモバイル発電器」, (筑波大学, つくば, 2018/1/22)
14. 桂 ゆかり, 熊谷 将也, 小谷拓史, 郡司 咲子, 今井 庸二, 大内 秀恭, 飛田 一樹, 木村 薫, "熱電特性実験データの大規模収集とマテリアルズインフォマティクスへの応用", 第 38 回日本熱物性シンポジウム OS9-I (産業技術総合研究所, つくば, 2017/11/8)
15. 桂 ゆかり, 小谷 拓史, 北原 功一, 熊谷 将也, 郡司 咲子, 今井 庸二, 大内 秀恭, 木村 薫, "熱電材料の大規模文献データと第一原理計算に基づく電子緩和時間の推定", 第 14 回日本熱電学会学術講演会, S2A3 (大阪大学豊中キャンパス, 2017/9/11-13)
16. (招待講演)桂 ゆかり, 小谷 拓史, 熊谷 将也, 郡司 咲子, 今井 庸二, 木村薫, "熱電特性グラフデータの収集による実験値マテリアルズインフォマティクス", 第 78 回応用物理学会秋季学術講演会, 5a-S21-7(福岡国際会議場, 2017/9/5-8)
17. (Invited) Yukari Katsura, "Theoretical considerations for enhancing thermoelectric properties", 2nd Asian Association of Thermoelectrics (AAT) Summer School, Oral (Kyoto University, Kyoto, Japan, 2017/8/27).
18. Yukari Katsura, Takushi Kodani, Koichi Kitahara, Masaya Kumagai, Yoji Imai, Sakiko Gunji, Hideyasu Ouchi, Kazuki Tobita, Naoki Sato, Kaoru Kimura, "Data-driven evaluation of effective relaxation times for real thermoelectric materials", The 36th International Conference of Thermoelectrics (ICT2017), Oral (Pasadena Convention Center, Pasadena, CA, USA, 2017/7/30-8/3).
19. Yukari Katsura, Kaoru Kimura, "Estimation of effective electron relaxation times in real thermoelectric materials", American Physical Society March Meeting 2017, (Oral; New Orleans Convention Center, New Orleans, LA, USA, 2017/3/13-17).
20. (招待講演)桂 ゆかり, "熱電変換技術の最新理論と材料インフォマティクスの熱電への応用～電子構造エンジニアリング＆材料インフォマティクス入門 ～", CMC リサーチセミナー, (口頭発表;ちよだプラットフォームスクウェア, 2016/12/8).
21. (招待講演)桂 ゆかり, 熊谷 将也, 今井 庸二, 郡司 咲子, 木村 薫, "熱電特性の実験値のデータベース化に向けたデータ収集 web システムの開発", 粉体粉末冶金協会 平成 28 年度秋季大会(第 118 回講演大会), 東北大学青葉山キャンパス, (口頭, 2016/11/9-11).
22. 桂ゆかり, 熊谷将也, 郡司咲子, 今井庸二, 小谷拓史, 木村薫, "熱電特性の実験値データベース作成に向けた文献データ収集システムの構築", 第 13 回日本熱電学会学術講演会, S5B-5 (口頭発表; 東京理科大学葛飾キャンパス, 2016/9/5-7)
23. (Invited) 桂ゆかり, "電子構造エンジニアリング入門" (チュートリアル講演), 第 13 回日本熱電学会学術講演会, S5B-5 (口頭発表; 東京理科大学葛飾キャンパス, 2016/9/5-7)
24. 佐藤陸, 小谷拓史, 桂ゆかり, 熊谷将也,今井庸二, 郡司咲子, 木村薫, "ニューラルネットワークに基づく PbTe 系熱電材料の特性予測", 第 15 回日本熱電学会学術講演会(TSJ2018), 東北大学 青葉山キャンパス (口頭, 宮城県仙台市, 2018/9/13-15).
25. 小谷拓史, 桂ゆかり, 熊谷将也, 安藤有希, 郡司咲子, 今井庸二, 木村薫, "実験値データベースと第一原理計算による高性能 PbTe 系熱電変換材料の探索", 第 15 回日本熱電学会学術講演会(TSJ2018), 東北大学 青葉山キャンパス (口頭, 宮城県仙台市, 2018/9/13-15).

〔図書〕(計 1 件)
1. 桂ゆかり, "熱電材料のマテリアルズインフォマティクス", *マテリアルズ・インフォマティクスによる材料開発と活用集-データベースの構築、記述子の設計法、モデル作成-*, 第 4 章 1 節, 技術情報協会(2019) ISBN:978-4-86104-732-9

〔産業財産権〕
○出願状況(計 0 件)
○取得状況(計 0 件)

〔その他〕
ホームページ等
Starrydata web システム https://www.starrydata2.org/
Starrydata 日本語ブログ http://starrydata.hatenadiary.jp/
Starrydata 英語ページ https://starrydata.wordpress.com/
個人ホームページ Starrydata の紹介 https://sites.google.com/site/yukarisearch/starrydata

6. 研究組織

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や

研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。