E-001

新素材研究開発分野における文献からの研究支援情報抽出技術の提案
Technology to Extract Research Clue from Scientific Papers for New Materials Developments

照屋絵理† 小野寛太‡ 竹内理† 森田秀和† 林貴之†
Eri Teruya Kanta Ono Tadashi Takeuchi Hidekazu Morita Takayuki Hayashi

1. はじめに

材料開発分野では国際的な開発競争の激化に伴い、短時間・低コストでの材料開発が必須となっていることを背景に、マテリアルズインフォマティクス（MI）が活発化してきた[1]。MI は物質・材料の物理的・化学的性質に関する膨大なデータと AI などの IT 技術を駆使して、材料開発を効率化する取り組みである。近年では、過去の実験データ等の数値データの解析により求める特性を持つ新規材料の原料候補を絞り込むことで、新規材料開発にかかる実験回数を削減する試みが盛んに行われており、実験回数を最大 10 万分の 1 に削減したという事例も発表されている[2]。

上記のように、従来は実験データ等の数値データに対する MI が主流であった。しかし、MI の有用性を示す事例が増えるにつれ、更なる材料開発の効率化に向け、数値データのみならずテキストデータを活用した MI への期待が高まっている。特に、テキストマイニング技術により、科学文献から材料開発を支援する情報（研究支援情報）を抽出したいというニーズが増大している。例えば、文献から物質の未知の性質やある物質の類似物質を抽出する、過去に特定の物質に対してよく適用された技術(実験や解析手法など)を抽出する、過去に特定の物質の研究を実施した研究者や所属の情報を抽出するなどである。これらのニーズに対し、Tshitoyan ら[3]は、科学論文に対して word embedding 技術[4]を用いて単語をベクトル化し、単語間の距離を見ることで、材料研究の支援となる物質とその物質の性質との対応関係が抽出可能であることを示している。

彼らの試みにより、材料分野における word embedding 技術への期待が高まっている。しかし、材料研究実業務での本技術の活用に向けては、彼らの研究内容に加え、材料研究者が着目する観点に関する情報の検索や可視化の仕組みが求められる。加えて、材料研究時に研究者がよく参考とする、過去に特定物質を研究した研究者やその所属等に関する関係性の抽出も求められている。

そこで、本研究では、材料研究では物質名等の特定のキーワードや、著者や所属等の科学文献に関する特定の情報およびそれらの関係性が研究を支援する上で特に重要であることに着目し、これらの情報および関係性を研究支援情報として文献から自動抽出し、さらに関係性の視認性が良いグラフ構造で可視化、検索を可能とする技術を提案する。以降では、2 章で提案技術の概要を説明する。3 章で提案技術の初期評価および考察を実施し、4 章でまとめを述べる。

2. 研究支援情報抽出・可視化技術の提案

本研究では、研究支援情報を科学文献から半自動で抽出、さらに情報の絞込み、検索、可視化を可能とする研究支援情報抽出・可視化技術を提案する。本技術の開発に先立ち、材料研究者と議論する中で、研究支援情報として、物質名とその物性値の詳細情報、技術名、物性値名に関するキーワード、および論文情報(journal や出版年などを含む)、著者名、所属名とそれらの関係性を求めていることが分かった。そこで、これらの情報を科学文献から固有表現抽出技術[5]、word embedding 技術、科学文献出版社が提供する Web からの論文情報取得の仕組み(文献取得 API)を用いて半自動で抽出する。さらに、材料研究者は関係する複数の情報から物質の性質等を連想することがあり、関係性を一目で把握できることを求めていることに着目し、視認性のグラフ形式で情報を可視化する。

図 1 に提案技術の概要を示す。提案技術では情報抽出部にて研究支援情報を抽出し、可視化部にて抽出した情報を可視化する。下記にてそれぞれについて説明する。

2.1 情報抽出部

情報抽出部では、文献からキーワードとして物質名、技術名、物性値名を固有表現抽出技術を用いてキーワード種別毎に抽出する。さらに word embedding 技術を用いてキーワード間の関係性を抽出する。また、論文情報、著者名、所属名の文献のメタデータとそれらの関係性を抽出する。

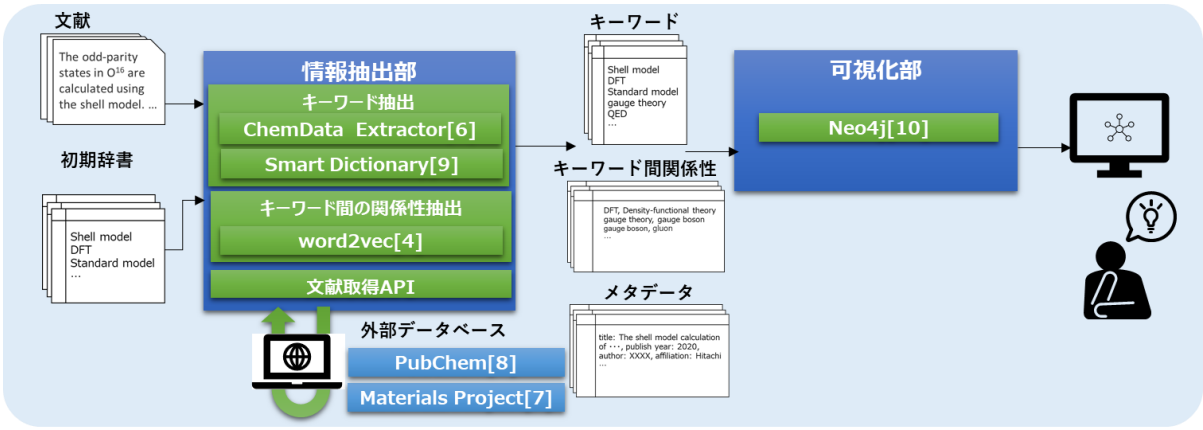

図 1　提案技術概要

† (株) 日立製作所 Hitachi, Ltd.
‡ 高エネルギー加速器研究機構 KEK

キーワード抽出では、物質名は物質名に特化した固有表現抽出技術である ChemDataExtractor[6]を用いて抽出する。また、物質の詳細情報として物質 ID や band gap 等の物性値情報を付与する。これらの情報は材料分野でメジャーなデータベースである PubChem [7]および Materials Project [8]の API を用いて取得する。技術名および物性値は、我々のグループで開発した弱教師学習を用いた汎用的な固有表現抽出技術である Smart Dictionary[9]を用いて抽出する。また、キーワード間の関係性抽出では、キーワード抽出にて抽出されたキーワードのうち、関連度の高いキーワードのペアを抽出する。ペアの抽出手法としては、word embedding 技術を用いて文献内のキーワードをベクトル化し、ベクトル間の cosine 距離が大きいペアを抽出する。word embedding 技術として、今回は word2vec[4]を用い、関連キーワードとして cosine 距離が 0.6 以上のペアを抽出した。cosine 距離の値は材料研究者が関連性が適切に抽出されていると判断したこの値を用いた。さらに、文献のメタデータとそれらの関係性を、多くの科学文献出版社が提供する、文献取得 API を用いて抽出する。

2.2 可視化部

可視化部では、2.1 節で抽出したキーワードとキーワード間の関係性、および文献のメタデータとそれらの関係性をグラフ形式で可視化する。可視化の機能として、Graph 構造を用いて情報を格納・可視化可能なオープンソースの NoSQL 系 DB である neo4j [10]を用いた。図 2 に可視化のイメージを示す。 情報抽出部で抽出した物質名、技術名、論文情報、著者名、所属を neo4j のノードとして可視化する。物質名の詳細情報は物質名ノードのプロパティとして付与する。また、論文の詳細情報である出版年等も論文情報ノードのプロパティ情報として付与する。また、キーワード間、キーワードと論文情報間、論文情報と著者名間、著者名と所属間の関係性を neo4j のリレーションとして可視化する。図 2 では、丸で示す情報がそれぞれのノードであり、色が物質名、著者名などのノードの種別を表す。ノード間を繋ぐ線が関連性を示している。また、それぞれの情報は neo4j の検索クエリを用いて検索が可能である。

3. 初期評価・考察

提案技術の初期評価として、計算材料科学の論文誌である Computational Material Science[11]に掲載された論文 10205 件に対し提案手法を適用し、材料研究者の視点で有用性を考察した。論文は Elsevier API[12]を用いて取得した。
まず、キーワード抽出およびキーワード間の関連性抽出について確認した。計算材料科学分野で論文中に出てくるキーワードは多岐にわたるが、物質名、技術名、物性値名などのカテゴリに分類される。また、キーワードには Density functional theory 、 DFT 、 Density-functional theory 、 first-principles density functional theory 、 Density functional calculations などの表記ゆれが見られる。これらのキーワードのカテゴリ分けやアノテーションには専門知識を要するため、多岐にわたるキーワードを手動で抽出し分類することは現実的ではない。しかし、提案技術ではそれぞれ約 100 件の初期辞書を与えるだけで、Smart Dictionary を用いて初期辞書から、キーワードの周辺単語等の特徴を学習することで、初期辞書に含まれないキーワードについても自動抽出することに成功した。得られたキーワードおよびキーワード間の関係性について、材料研究者による手動でのアノテーションと比較し、ある物質名とその物質に良く適用される技術名の関係性が抽出されるなど、全く遜色なく抽出できていることを確認した。また、可視化機能について、図 2 に示すように関係性を可視化することにより、研究者が着目する物質名や技術名について、著者、所属情報を含めて直感的に把握することが可能となり、本技術が材料研究を行う研究者にとって有用であるとの意見を頂いた。

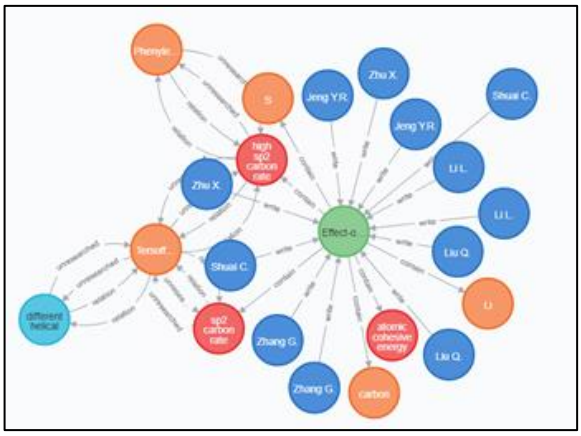

図 2　可視化イメージ

4. まとめ

本研究では、材料開発の効率化に向け科学文献から研究支援情報を抽出し、さらに情報の絞込み、検索、可視化を可能とする研究支援情報抽出・可視化技術を提案した。初期評価として、提案技術を計算科学の論文誌に適用した結果、研究者が着目するキーワードについて、研究を進める上で重要な情報を直感的に把握できることがわかった。今後評価を進め、本技術で材料開発研究の促進が可能かを検証する。

参考文献

[1] (株)日立総合計画研究所, ”研究員解説 マテリアルズ・インフォマティクス”,2016,http://www.hitachi-hri.com/keyword/k109.html
[2] 日経 XTECK, MI で先陣を切る住友化学、材料開発で驚きの効率,https://xtech.nikkei.com/atcl/nxt/column/18/00026/00043/
[3] Tshitoyan, V., Dagdelen, J., Weston, L. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98 (2019).
[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. NIPS, 2013
[5] Archana Goyal, Vishal Gupta, Manish Kumar,"Recent Named Entity Recognition and Classification techniques: A systematic review",Computer Science Review,Volume 29,2018,21-43
[6] Swain, M. C., & Cole, J. M. "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature", J. Chem. Inf. Model. 2016, 56 (10)
[7] PubChem, https://pubchem.ncbi.nlm.nih.gov/
[8] The Materials Project, https://materialsproject.org/
[9] 照屋絵理, 業務改善に向けたダークデータ活用技術, 第 19 回情報科学技術フォーラム (2019)
[10] neo4j, https://neo4j.com/
[11] Computational Materials Science, https://www.journals.elsevier.com/computational-materials-science
[12] Elsevier API , https://dev.elsevier.com/