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　３ 年　 ５ 月 ２４ 日現在

機関番号：１０１０１
研究種目：挑戦的研究（萌芽）
研究期間：2019～2020
課題番号：１９Ｋ２２８８８

研究課題名（和文）専門用語の知識保全エコシステムを有する特定研究グループ向け論文・図表DBの研究

研究課題名（英文）Research on an Paper/Figure Database System with Technical Term Management Ecosystem for Particular Research Group

研究代表者
吉岡　真治（Yoshioka, Masaharu）

北海道大学・情報科学研究院・教授

研究者番号：４０２９０８７９

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：本研究では、論文からの情報抽出・知識発見を支援するための専門用語の知識保全エコシステムを有する特定研究グループ向け論文・図表データベース（DB）の提案を目指している。本研究では、分野の研究者が興味を持つ最新の論文を含むデータベースをPDF 解析技術を用いて構築すると共に、そこからの用語辞書を作成する枠組を提供することによって、ユーザが継続的に専門用語に関する知識保全を行うエコシステム(生態系)を持つ論文・図表DBを提案し、実際の論文からDBを構築し、評価を行なった。今後は、作成したシステムを基盤として用いることにより、継続的に論文を追加しながら実運用を行なっていく予定である。

研究成果の学術的意義や社会的意義
本研究は、論文データの活用方法として、大規模なデータベースを構築するのではなく、特定の研究グループが注目する論文に限定して網羅的に収集したデータベースを構築することにより、研究活動に役立つデータベースの提案を行なっている。利用者は、分析に有用な用語の定義を行うことにより、研究グループによりカスタマイズされた分析が行えるだけでなく、興味のある研究分野での研究動向分析が行えるといった支援も可能となる。このように情報科学の研究者と分野の研究者が共同で問題に取り組む枠組みを構築することは、異分野協同研究を進めるという学術的意義だけでなく、新たな論文データ活用方法の提案という社会的意義も大きい。

研究成果の概要（英文）：Objective of this research project is proposal of an paper/figure database system with technical term management ecosystem for particular research group. This system constructs database based on the information (text, figure, caption, and metadata) extracted from the research papers collected by a research group. The system also supports to construct technical term dictionary management system and this dictionary is used for extracting technical terms from the texts. The system also provides a framework to analyze papers using these terms. We implement a system and construct database using research papers of interest and evaluate the effectiveness of the system. We plan to use this database for the ordinal research activities for long term evaluation.

研究分野：知識工学

キーワード：論文データベース　テキストマイニング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

我々は、これまでに、ナノ結晶デバイス開発論文から、実験情報を抽出する方法を提案してきた。この研究を発展させ、論文データベース(DB)を大規模にすることにより、実験条件に注目した類似論文や関連するグラフの閲覧、および、関連するパラメータや材料のリストの生成といった研究動向の情報収集が可能になると考えた。

また、同時期に、CREST「構造理解に基づく大規模文献情報からの知識発見」の研究グループとの交流の中で、論文 PDF 解析を用いることにより、PDF で配布される最新の論文から図表の情報を抽出する技術の紹介を受け、上記のデータベースの拡充が可能であることを確認した。

一方で、論文からこれらの情報を機械学習により適切に抽出するためには、大規模なコーパスなどが必要というコールドスタートの問題を解消することが必要であった。特に、このようなコーパスを分野で現れる様々な用語表現をカバーする形で構築するのは困難である。この問題に対し、専門用語については、大量の文書データが存在すれば、語構成要素などの情報から専門用語の候補生成が出来る事に注目し、比較的、準備コストが低い 専門用語辞書の構築とそれを利用した情報抽出という手法からシステムを立ち上げ、 利用していく中で機械学習のコーパスとして使える情報を収集するようなエコシステムを持つ論文・図表 DB を作成するという構想に至った。

２．研究の目的

本研究では、論文からの情報抽出・知識発見を支援するための専門用語の知識保全エコシステムを有する特定研究グループ向け論文・図表 DB を提案する。このシステムでは、分野の研究者が興味を持つ最新の論文を含むデータベースを PDF 解析技術を用いて構築すると共に、これらの分析に利用する情報抽出に利用可能な概念辞書の内容の更新や訓練データを作成する枠組を提供することによって、ユーザが継続的に専門用語に関する知識保全を行うエコシステム（生態系）を構築する方法を提案する。本手法により、特定分野の研究者は、直接的に、その知識保全の効能を享受することが出来るため、知識提供の強い動機付けになることが期待される。

３．研究の方法

本研究では、研究グループの興味をもつ研究分野が定常的に参加する国際会議や定期的に購読する雑誌論文により表現されるという前提のもとで、これらの論文を網羅的に収集することで、用語辞書の構築や特定分野の研究動向調査が可能となると考えた。この基本的な考え方の元で、与えられた論文の PDF 群から専門用語辞書を構築するための研究課題として、「1.分類付き専門用語辞書の作成支援」、「2.専門用語抽出ツールの構築」を設定し、分析の基盤となる用語辞書構築の支援の枠組みを構築する。さらに、この用語辞書を用いて論文の検索や研究動向調査を行うための課題として、「3.図表・論文の多観点分析可能な検索インターフェース」、「4.研究動向の時系列分析」を設定し、用語辞書を作成するというコストに対して、ユーザが研究に役立つ情報を得ることができるというサイクルの実現を目指す。このサイクルを具体的な研究分野（ナノ結晶デバイスの開発）に対して、行うことで、その有用性を検討する。

４．研究成果

2019 年度(初年度)は、「2.専門用語抽出ツールの構築」の研究を実施するための、PDF からのテキスト情報抽出に関する整備を行うとともに、専門用語抽出ツールの整備を行なった。

「2.専門用語抽出ツールの構築」については、結晶成長によるナノデバイス開発に関する研究分野の論文を対象として、複数の PDF 解析ツールによる実験を行い、主に、図表については、pdffigure2 (https://github.com/allenai/pdffigures2)[1]を用い、論文メタデータの抽出も含むテキスト抽出については、GROBID(https://github.com/kermitt2/grobid)を利用することとした。また、専門用語抽出ツールとしては、TermExtract(http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract/)[2]を利用した。

この抽出結果である専門用語の候補について、「1.分類付き専門用語辞書の作成支援」のためのインターフェースを作成した（図１）。

図１：分類付き専門用語辞書作成のためのインターフェース

本インターフェースでは、専門用語の候補に対し、用語が実際に使われる文書の検索結果を提示することで、利用者に専門用語としての分類付き登録を支援する。その際に、略称などの異表記の統合や大文字小文字の正規化の可否などの情報を合わせて登録してもらうことで、表記の揺れに対応した論文の検索や、研究動向調査に役立てることが可能となっている。

2020 年度（2 年度目）は、2019 年度に作成した専門用語辞書作成のためのインターフェースを活用した辞書の作成実験を行うとともに、これらのデータを用いた「3. 図表・論文の多観点分析可能な検索インターフェース」の開発と「4. 研究動向の時系列分析」を行う枠組みを提案した。図２に図表・論文の多観点分析可能な検索インターフェースを示す。

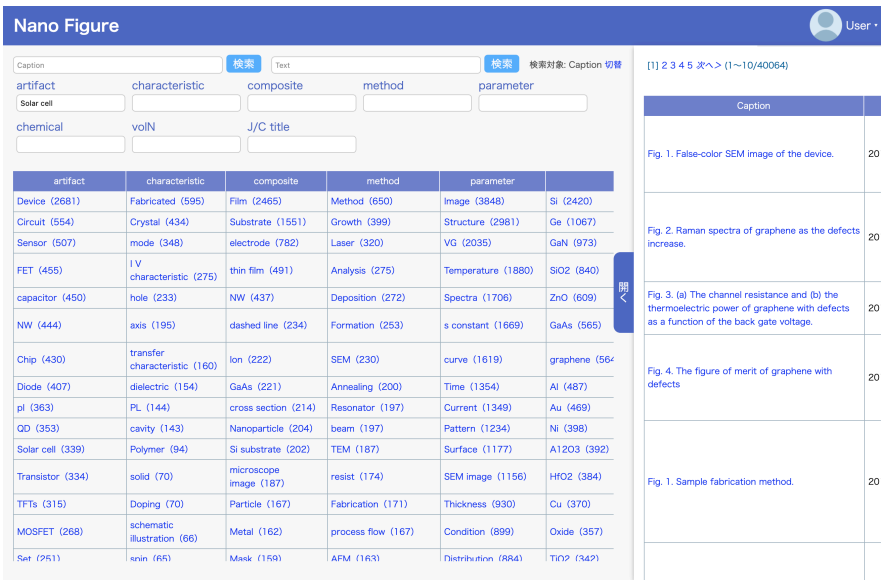


図２：図表・論文データベースの多観点分析可能なインターフェース

本インターフェースでは、「1. 分類付き専門用語辞書の作成支援」の結果として得られた単語辞書を使って、多観点からの絞り込み検索を行うことが可能である。具体的には、論文が含むべき検索語を上記の検索語リストから入力する、もしくは、下の用語リストから選択することにより、絞り込み検索を行う。例えば、artifact の分類の中から、Solar cell を選択することで、Solar cell に関する論文が検索され、その論文の中で利用されている各分類に対応する用語が頻度順に表示される。

ここから、さらに、興味のあるパラメータなどで絞り込むことが可能であり（図３）、一定程度絞り込んだ段階で、検索結果に含まれる図表を確認する（図４）ことができる。さらに、必要に応じて、論文データを確認することによって、より詳細な情報を得ることができる。

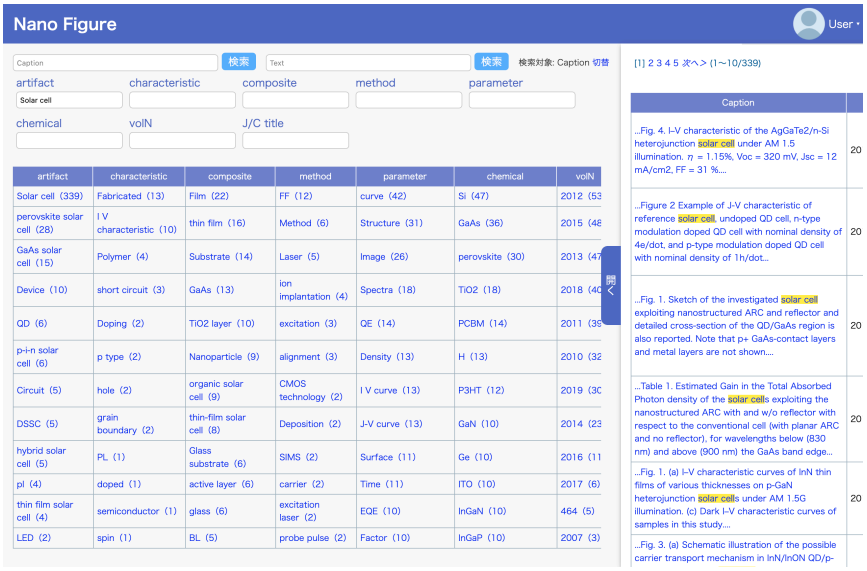


図３：検索結果と関連用語の提示

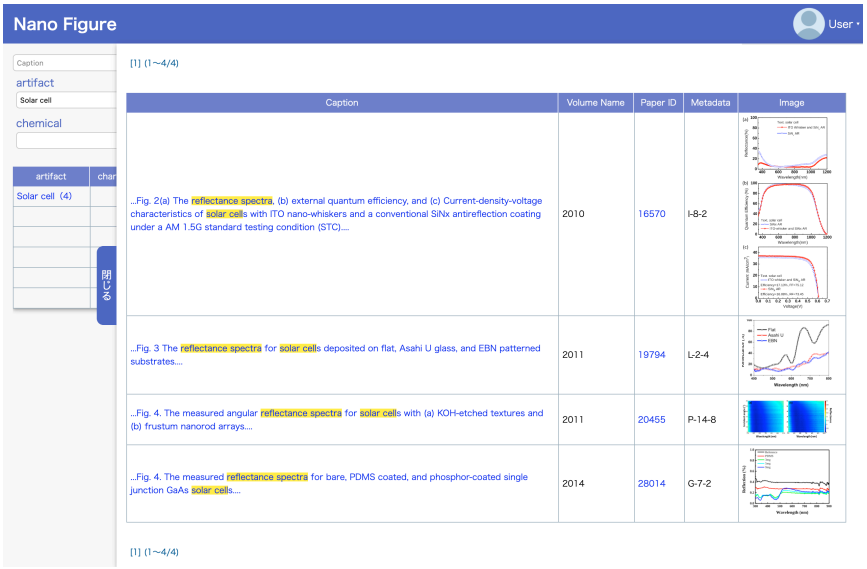


図４：検索結果の表示

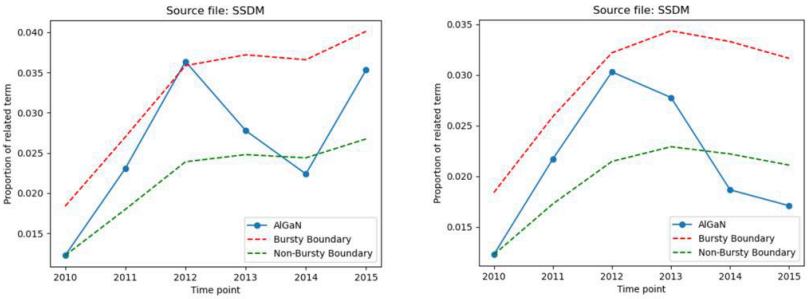
また、「4. 研究動向の時系列分析」の手法として、ここで得られたデータベースを対象として、時系列データ分析であるバースト解析[3]を実行することで、その国際会議における研究動向調査を行う方法を提案した。表１に、全論文に対して、１年ごとのそれぞれの用語を含む論文数に注目したバースト解析の結果を示す。

表１：材料に関するバースト上位の用語のリスト

2010	2011	2012	2013	2014	2015
S21	precursor	molecule	DSA	SiO2	SiO2
p-type layer	Fe	semiconductors	NPs	ions	O2
HfO2	PBS	BCP	No	NPs	DNA
IGZO film	Co	DNA	OH	nanoparticle	MoS2

また、より詳細な分析を行うために、特定の国際会議に注目した分析を行なった実験結果を分野の専門家である協同研究者と分析したところ、開催場所や時期の関係で特定のグループが参加したことによる影響などがバーストとして検出されることが指摘された。

この問題に対し、各年度の専門用語を含む論文の件数を単純に数えるのではなく、著者が共通するものは論文１件として数えるという正規化を行う方法により、上記の問題に対応することとした（図５）。このように初期に計画していたシステムについて開発を行い、辞書を作るという作業に対して、有用な分析結果が得られるというサイクルを回すことができることを確認した。



The trend of AlGaN based on papers from SSDM. **(Before)**

The trend of AlGaN based on papers from SSDM. **(Now)**

図５：バースト解析の結果（著者による影響を考慮する以前と以後）

一方で、分類付き専門用語辞書の作り方については、少し、検討が必要な点があることが判明した。例えば、図２において、NW（ナノワイヤ）については、artifact(製品)と composite（生成物）という二つのカテゴリに属しており、どちらの項目にも検索結果として表示されている。これは、NW（ナノワイヤ）がその製造を目的とした製品である側面と、中間生成物として作成される生成物としての側面を持つためである。ただ、現在のシステムでは、辞書による字面マッチングによる抽出を行なっているため、この区別が行えない。この区別を行わない段階では、検索結果の表示という観点からは、どちらかの分類に限定して表示させるといったことを検討することも含めて、その取扱について考えていく必要があることがわかった。

今後は、作成したシステムを基盤として用いることにより、継続的に論文を追加しながら実運用を行なっていく中で、本研究の提案による用語辞書のメンテナンスを行なっていき、より良いエコシステムをもつシステムへの発展へとつなげていく予定である。

また、上記のシステム開発と並行し、化学物質名の機械学習による自動抽出に関する研究を並行して行なった。本研究では、化学物質名の表現が、特定の接尾・接辞・ハイフンなどを用いた用語の結合といったパターンの組み合わせにより構成されていることに注目し、従来型の化学物質名の抽出システムが論文中の文脈に依存して判定するという方針とは異なり、化学物質名をサブワードと言われる高頻度で現れる文字列により分割し、その文字列のシーケンスに対して、化学物質名か否かを判断するという方法を提案した。本システムのメリットは、化学物質データベースから、大規模な正事例を収集することができる点にある。

本手法の有用性を検証するために、化学物質名データベース ChemIDPlus (https://chem.nlm.nih.gov/chemidplus/) と WordNet の一般語を用いた化学物質名の認識の実験を行った。単純な、化学物質名データベースと WordNet との分類問題としては、SentencePiece(https://github.com/google/sentencepiece)を用いたサブワードのシーケンスを深層学習による言語解析モデルである Bidirectional Long-Short Term Memory(Bi-LSTM)[4]により学習させることで、F 値で 0.995 と高い制度で分類できることを確認した。一方で、この学習させたモデルを、化学物質名抽出の分野で利用されるコーパスである CHEMDNER に現れる論文中に現れる化学物質名に適用したところ、F 値が 0.370 となり、データベースに特有な表記に特化した学習をしていることが判明した。実際に、コーパスで用いられる用語の一部を訓練データに追加した実験を行うことで、F 値が 0.862 となり、用語の表記のバリエーションに対応した多様な用語データが必要であることが確認された。また、実際のコーパスでは、略称のように、かならずしも文字列からだけでは判断ができない用語の存在も確認され、このような用語を扱うためには、論文全体で、略称とオリジナルの表記を対応づけるようなマクロな分析が必要であることも確認された。

[1] Christopher Clark and Santosh Divvala. "Pdffigures 2.0: Mining figures from research papers." 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL). IEEE, 2016.: “PDFFigures 2.0: Mining Figures from Research Papers”. 143-152, 2016. 10.1145/2910896.2910904.

[2] Hiroshi Nakagawa, Tatsunori Mori. "Automaic Term Recognition based on Statistics of Compound Nouns and their Components", Terminology, Vol.9 No.2, pp. 201-209, 2003

[3] Jon Kleinberg. "Bursty and hierarchical structure in streams." Data mining and knowledge discovery 7.4 (2003): 373-397.

[4] Alex Gravesand, Mohamed Abdel-rahma, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.

[5] Martin Krallinger, et al. "The CHEMDNER corpus of chemicals and drugs and its annotation principles." Journal of cheminformatics 7.1 (2015): 1-17.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Yoshioka Masaharu、Hara Shinjiro	４．巻 1040
２．論文標題 Construction of an In-House Paper/Figure Database System Using Portable Document Format Files	５．発行年 2019年
３．雑誌名 Information Search, Integration, and Personalization: 10th International Workshop, ISIP 2018	６．最初と最後の頁 41～52
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-30284-9_3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 吉岡真治，大久保好章,尹磊,原真二郎,鈴木晃,高山英紀,石井真史
２．発表標題 更新可能な用語抽出機能を持つ小規模研究グループ向け論文・図表データベースの構築
３．学会等名 第80回応用物理学会秋季学術講演会 19-a-B01-9
４．発表年 2019年

１．発表者名 鈴木晃，石井真史
２．発表標題 マテリアルズ・インフォマティクスのための材料辞書群の構築
３．学会等名 第80回応用物理学会秋季学術講演会，19a-B01-6
４．発表年 2019年

１．発表者名 町光二郎，吉岡真治
２．発表標題 無機化合物を対象とした論文に対する化学物質名抽出システムの性能分析
３．学会等名 言語処理学会第26回年次大会発表論文集，G5-4
４．発表年 2020年

１．発表者名
鈴木晃，石井真史

２．発表標題
マテリアルズ・インフォマティクスのための材料辞書の構築とデータベース化検討

３．学会等名
第67回応用物理学会春季学術講演会，15a-A205-4

４．発表年
2020年

１．発表者名
Kojiro Machi and Masaharu Yoshioka

２．発表標題
Word-Level Chemical Named Entity Recognition Based on Subword Sequence Analysis

３．学会等名
Proceedings of Fourth International Workshop on SCIentific DOCument Analysis (SCIDOCA2020)（国際学会）

４．発表年
2020年

１．発表者名
町光二郎，吉岡真治

２．発表標題
化学物質データベースを訓練データに用いた化学物質名識別システムに関する実験的分析

３．学会等名
研究報告情報基礎とアクセス技術（IFAT），2020-IFAT-139-3

４．発表年
2020年

１．発表者名
Lei Yin, Masaharu Yoshioka and Shinjiro Hara

２．発表標題
Construction of In-house Paper/Figure Database System Supporting Research Trend Analysis

３．学会等名
第68回応用物理学会春季学術講演会，19p-Z32-9

４．発表年
2021年

１．発表者名
鈴木晃，石井真史

２．発表標題
材料辞書データベースを使った論文からの大量データ抽出：材料用語の階層化による体系的自動タグ付け

３．学会等名
第81回応用物理学会秋季学術講演会，9p-Z09-17

４．発表年
2020年

１．発表者名
鈴木晃，石井真史

２．発表標題
材料辞書データベースを使った論文からの大量データ抽出：用語間関係性抽出の自動化検討

３．学会等名
第68回応用物理学会春季学術講演会，19p-Z32-11

４．発表年
2021年

１．発表者名
鈴木晃，石井真史

２．発表標題
磁石物性データ大量取得のためのテキスト処理要素技術の開発

３．学会等名
日本金属学会2021年春期(第168回)講演大会，S1.7

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
Knowledge exploratory project
http://nanoinfo.ist.hokudai.ac.jp/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	原　真二郎 (Hara Shinjiro) (50374616)	北海道大学・量子集積エレクトロニクス研究センター・准教授 (10101)	
研究分担者	鈴木　晃 (Suzuki Akira) (50799723)	国立研究開発法人物質・材料研究機構・統合型材料開発・情報基盤部門・NIMS特別研究員 (82108)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関