1J2-4

ナノ知識探索プロジェクト：実験記録からの知識発見 (第 4 報) -キャプションからのメタデータの自動抽出による グラフイメージ検索システム-

Knowledge Exploratory Project for Nano Device Design and Manufacturing: Knowledge Discovery from Experimental Records (4th Report)
Graph Image retrieval System by extracting Metadata from Caption

朱 濤[*1] ZHU Tao　Thaer M. Dieb[*2] Thaer M.DIEB　吉岡 真治[*1] Masaharu YOSHIOKA　原 真二郎[*3] Shinjiro HARA

[*1]北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

[*2]物質・材料研究機構
National Institute for Material Science

[*3]北海道大学 量子集積エレクトロニクス研究センター
Research Center for Integrated Quantum Electronics, Hokkaido University

Knowledge Exploratory Project for Nano Device Design and Manufacturing is a project for extracting useful information from the nano device development information for supporting nano device design and manufacturing. In the previous reports, we have proposed a framework to extract useful metadata from the nano device development reserach papers. In this report, we propose a novel graph image retrieval system that shows characteristic information of graph retrieval results by summarizing multifacet metadata (e.g., source material, experimental parameter). In this system, metadata of the graph image is extracted from the caption of the image by using NaDev that is a automatic metadata extraction system from the nano device development papers.

1. はじめに

ナノ結晶デバイス [Fukui 91, Noborisaka 05] とは、ナノスケールの結晶構造を用いたデバイスであり、半導体や太陽電池といった様々な分野での応用が期待されている。このナノ結晶デバイスの開発プロセスにおいては、第一原理に基づくような計算だけではなく、主に経験に裏打ちされた多くの試行錯誤が必要とされている。

ナノ知識探索プロジェクトは、このようなナノ結晶デバイスの開発の現状を受け、その開発プロセスに関する情報を整理することによって、ナノ結晶デバイスの開発者への情報提供を行うと共に、その知識化を支援する事を目標としたプロジェクトである [吉岡 10]。

我々は、ナノ結晶デバイスの開発を行っている北大の量子集積エレクトロニクスセンターと協力し、これまでに、ナノ結晶デバイス開発の中で、重要な実験条件などの記述を見つけることを目標として、ナノ結晶デバイス開発論文からの、有用なメタデータ (材料、実験パラメータ、評価パラメータなど) の抽出を行うためのコーパスの作成 [Dieb 16] と、そのコーパスを用いたメタデータの自動抽出システム NaDev の開発 [Dieb 15] を行ってきた。

本報では、NaDev を用いて抽出したメタデータの活用方法として、論文中の図とキャプションを利用したグラフイメージの検索システムを提案する。ナノ結晶デバイス開発などの物性系の論文においては、情報系の論文と比較すると、図のキャプションに多くの情報が含まれている。特に、グラフにおいては、そのグラフの縦軸や横軸に相当するパラメータの名前や、実験条件の情報などが記述されることが多い。そのため、単純に論文内に共起するパラメータの組み合わせについての情報を集めるよりも、キャプションの中で共起するパラメータの組合わせなどを調べることは、より関係性の強いパラメータ群の発見などに役立つと考えている。

そこで、本研究では、キャプションに含まれるこれらのメタデータを NaDev により収集し、その情報を利用してグラフイメージを検索すると共に、検索結果に関連するメタデータの情報を提供することで、さらなるグラフイメージの絞り混みを支援するグラフイメージ検索システムを提案する。

連絡先: 吉岡 真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目,011-706-7107, yoshioka@ist.hokudai.ac.jp

2. NaDev：ナノ結晶デバイス論文への自動タグ付けシステム

我々は、ナノ結晶デバイス開発論文を対象として、文書中に存在する実験に関係する情報に対して、自動的にタグ付けを行う NaDev[Dieb 15] を開発している。この NaDev では、以下の情報についてタグ付けを行う。

材料 (SMaterial)： 実験に用いる素材や化合物 (Ga や GaAs など)

物質の特性 (MChar)： 素材や作成された化合物が持つ特性 (結晶の異方性に関する情報など)

実験パラメータ (ExP)： デバイス作成時の実験で用いるための制御パラメータ (圧力や流量など)

実験パラメータの値 (ExPval)： 上記の実験パラメータに対応する値

評価パラメータ (EvP)： デバイスの性質を評価するためのパラメータ (厚みや導電性など)

評価パラメータの値 (EvPval)： 上記の評価パラメータに対応する値

デバイスの作成手法 (Method)： 作成手法の名前 (SAMOVPE など)

最終製品 (TArtifact)： 最終製品の名前 (半導体ナノワイヤなど)

NaDev は、材料としてよく用いられる化学物質名の認識システム、パラメータなどで用いられる物理量に関する単語のリストの情報、形態素解析の結果などのフィーチャを用いて、ナノ結晶デバイス開発論文についてタグ付けされた文書から学習したモデルを用いて、上記の情報に対応する項目の抽出を行う。

3. グラフイメージ検索システム

3.1 グラフを中心としたパラメータ間の関係分析

ナノ結晶デバイスの作成のためには、デバイスの持つ構造を設計するだけではなく、そのデバイスの構造を製造するための適切な条件を決定する必要がある。しかし、どの様なパラメータが最終的なデバイスの性能に大きく影響を与えるのかについては、実験の初期段階では必ずしも確定しておらず、実際の実験中の試行錯誤により、有用なパラメータ間の関係について検討を行い、その結果が、論文の中にまとめられるという状況が存在している。そのため、論文中に存在するグラフに現れるパラメータ間の関係に関する情報は、類似した実験を行う際にどの様なパラメータについて考慮した方が良いのかといった有用な知識となることが期待される。

また、ある程度、関係しそうなパラメータが思い浮かぶ場合においても、これらの関係を表現したグラフの多くを閲覧し、通常と異なった関係を持つようなグラフを発見することは、パラメータ間の関係をコントロールするような別の実験条件の発見へとつながることが期待される。

本研究では、このようなナノ結晶デバイスの開発者に対して、論文のグラフの検索を支援すると共に、関連するパラメータの情報などを提供することができるグラフイメージ検索システムの構築を目標としている。

3.2 図表の検索システムの機能的要件

論文中の図表を検索するシステムとしては、Elsevier の Science Direct[*1] などの論文検索システムにおいても、キャプションを対象とした検索システムが提供されている。ここでは Science Direct の図表の検索システムを例にとり、その一般的な機能について説明する。この検索システムでは、キャプション中の文字列に加え、論文タイトル、著者といった書誌情報に関するメタデータを利用した検索が行えると共に、検索結果から掲載されている論文誌や発表年度といった主に書誌情報から得られるようなメタデータの集約結果が表示され、さらなる図表の絞り混みを支援することが可能になっている。

本研究では、このような図表の検索システムとは異なり、ナノ結晶デバイス開発論文に代表されるような物性系の実験などに関する図表を対象とした検索システムの構築を目的としている。具体的には、グラフのキャプションに記述されることが多い、3.1 節で述べたようにパラメータ間の関係や、実験条件に関連するような材料や手法の名前といったメタデータを抽出し、Science Direct などで提供する書誌情報に関するメタデータではなく、実験条件に関するメタデータの集約結果を表示することにより、関係するパラメータの発見や実験条件に関する知見を支援するグラフイメージ検索システムの提案を行う。

その実現方法としては、図表のキャプションを対象として、NaDev を用いることにより、実験条件に関連する情報のタグ付けを行い、その結果をメタデータとして利用する。ユーザは、メタデータのタイプを考慮した検索を行う事ができるだけでなく、検索結果に表示される他のメタデータの集約結果を見ることにより、目的のパラメータと同時に現れることの多い他のパラメータのリストなどを見ることが可能となり、3.1 節で述べたナノ結晶デバイス開発における利用方法に即した検索が実現できると考えている。

∗1 http://www.sciencedirect.com

4. プロトタイプシステムの構築と評価

4.1 プロトタイプシステムの概要

前節で述べたグラフイメージ検索システムのプロトタイプを作成した。データとしては、Creative commons のライセンスで公開されている Nano 分野のオンラインジャーナルである Beilstein Journal of Nanotechnology [*2]から、集めた図 6,133 枚に対してカラムストア型データベースであり、全文検索の機能を持つ groonga[*3] を用いてデータベースを構築した。本来は、全ての図ではなく、グラフイメージのみに限定してデータベースを構築することを考えていたが、現時点では、グラフイメージとそれ以外を区別するための処理が準備できていなかったため、全ての図を用いてデータベースを作っている。

各々の図に対しては、以下のデータが付与されている。

- キャプション：図のキャプション (テキスト)
- 要旨：論文の要旨 (テキスト)
- タイトル：論文のタイトル (テキスト)
- メタデータ：2. で述べた 8 種類のメタデータの各々について、その抽出された語のリスト

また、図 1 にプロトタイプシステムのスクリーンショットを示す。

本システムでは、上部の検索条件の設定部分においてキャプション、タイトル、アブストラクトに加え、メタデータに対応する条件を設定して、図のデータの検索を行うことができる。また、検索結果について、全てのメタデータについて集約操作を行うことにより、検索結果群に含まれるパラメータ名のリストがサマリーとして表示される。また、ユーザは、これらのキーワードをクリックすることで、検索条件に追加することにより、検索結果の絞り込みを行うことができる。

検索結果としては、図を表示するだけでなく、キャプション、タイトル、メタデータの情報をあわせて表示すると共に、検索キーワードのハイライトを行っている。また、グラフをクリックすることで、アブストラクトの情報も含む全ての情報を閲覧することが可能となる。

4.2 ナノ結晶デバイス研究者による評価

本システムを、共著者の一人であるナノ結晶デバイス研究者 (原) に提供し、簡単な検索を行ってもらい、システムの利点ならびに問題点に関する議論を行った。この評価の際には、より、実際の利用状況に近づけるため、共著者の興味を反映して集められた論文 916 件から作成した 3142 枚の図から構成されるデータベースを作成した。このようなデータの構築については、著作権的な問題があるため、研究グループが持っている論文データなどから作る方法についても並行して検討している。

以下に、指摘された主な利点と問題点を列挙する。

- 利点

∗2 https://www.beilstein-journals.org/bjnano/
∗3 http://www.groonga.org/

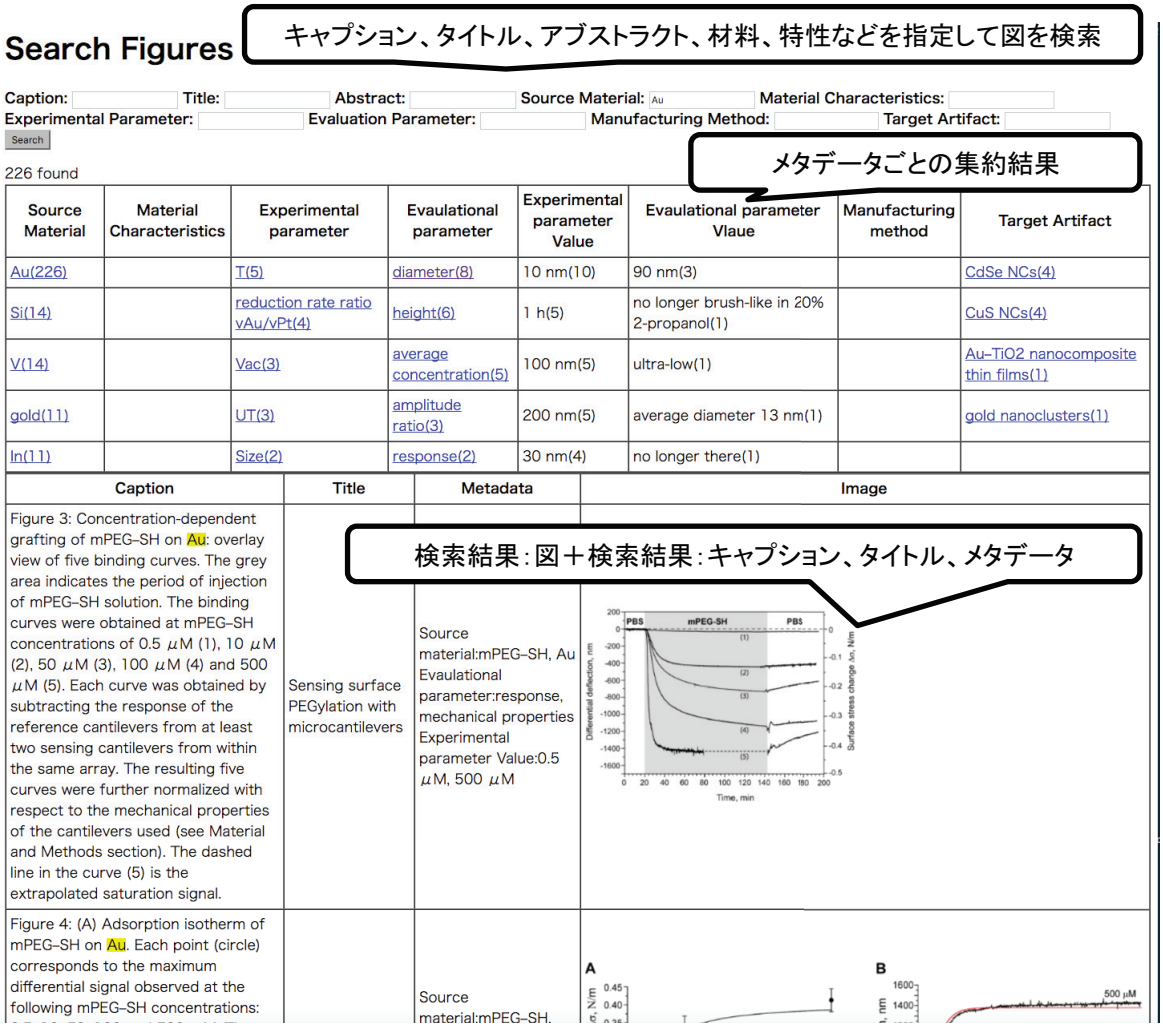

図 1: グラフイメージ検索システム

- 目的や構造などのキーワードから出発して、表示される関連するパラメータなどを用いてグラフを絞り混んでいくという考え方は面白い。
- パラメータを 2 つキーワードに与えて、それらの関係を示したグラフをまとめてみるのは面白い。

- 問題点
  - 適切なメタデータの付与が必要
    検索結果をみたところ、本来パラメータなどのメタデータが付与されるべき単語にメタデータが付与されていない事があるため、絞り混みの情報として不十分である。
  - 同義語や複合語の取り扱い
    量子ドットのように、Quantum dots, QD, QDs などで表される概念について、取りまとめをして欲しい。また、メタデータを付与する単語の単位の問題から、the gas temperature と、temperature of the gas などをまとめることができないので、これについても検討が必要である。

4.3 今後の改良方針

プロトタイプシステムとして具体的なシステムイメージを提示することにより、ナノ結晶デバイスの研究者からのフィードバックをもらい、全体としての方向性の有用性については、確認した。しかし、現状のシステムでは、NaDev によるメタデータの抽出精度の問題もあるため、さらなる改善が必要である。

具体的には、これまでの論文の全文を対象にしたコーパスだけではなく、キャプションを対象にした訓練用のコーパスの作成を行うと共に、化学物質の認識システムとして汎用の化学物質の抽出システムである ChemSpot[Rocktäschel 12] などを利用すると共に、ナノ結晶デバイスの論文で用いられるパラメータのリストなども充実させることにより、網羅的なメタデータの抽出が行えるように NaDev の改善を行う。

また、パラメータの種類を実験パラメータと評価パラメータに分けていることにより、パラメータ全体としての抽出精度を下げている場合があることが確認されている。実験パラメータと評価パラメータを確実に抽出できるのであれば、そちらの方が良いが、抽出漏れが多いようであれば、単にパラメータとして抽出することについても検討したいと考えている。

その他にも、専門用語リストなどを用いた用語の表記の正規化などについても検討していきたい。

5. おわりに

本報ではナノ結晶デバイス開発の研究論文と、その活用方法を踏まえたグラフイメージ検索システムの提案を行った。実際にナノ結晶デバイス開発を行っている研究者に試用してもらい、利点や問題点についてのコメントを得た。今後は、メタデータのより適切な付与を行うための NaDev の改良を含み、これらの問題点を解決する事により、実際の研究活動に利用可能なシステムの構築を目指す予定である。

謝辞

また、本研究の一部は、科研費挑戦的萌芽 26540165 により行われた。ここに記して、謝意をあらわす。

参考文献

[Dieb 15] Dieb, T. M., Yoshioka, M., Hara, S., and Newton, M. C.: Framework for automatic information extraction from research papers on nanocrystal devices, *Beilstein Journal of Nanotechnology*, Vol. 6, pp. 1872–1882 (2015)

[Dieb 16] Dieb, T. M., Yoshioka, M., and Hara, S.: An Annotated Corpus to Support Information Extraction from Research Papers on Nanocrystal Devices, *Journal of Information Processing*, Vol. 24, No. 3 (2016), (to appear)

[Fukui 91] Fukui, T., Ando, S., Tokura, Y., and Toriyama, T.: GaAs Tetrahedral Quantum Dot Structures Fabricated using Selective Area Metalorganic Chemical Vapor-deposition, *APPLIED PHYSICS LETTERS*, Vol. 58, pp. 2018–2020 (1991)

[Noborisaka 05] Noborisaka, J., Motohisa, J., Hara, S., and Fukui, T.: Fabrication and characterization of freestanding GaAs/AlGaAs core-shell nanowires and AlGaAs nanotubes by using selective-area metalorganic vapor phase epitaxy, *APPLIED PHYSICS LETTERS*, Vol. 87, (2005)

[Rocktäschel 12] Rocktäschel, T., Weidlich, M., and Leser, U.: ChemSpot: a hybrid system for chemical named entity recognition, *Bioinformatics*, Vol. 28, No. 12, pp. 1633–1640 (2012)

[吉岡 10] 吉岡 真治, 冨岡 克広, 原 真二郎, 福井 孝志：ナノ知識探索プロジェクト：実験記録からの知識発見, 2010 年度人工知能学会全国大会 (第 24 回) 論文集 (2010), CD-ROM 1B3-3