１Ｃ０４

# 重点分野分析のための論文マップの作成

○七丈直弘（政研大／一橋大）※，寺田好秀，加瀬豊（政研大）
※n.shichijo@r.hit-u.ac.jp

## 1. はじめに

17 世紀に近代科学が成立して以来、現在に至るまで科学研究の成果の数は指数関数的に増加しており、例えば現在と 20 年前を比較すると約 10 倍の数の成果が生み出されている。また、科学は単にその成果の量的側面で拡大しているだけでなく、その構造も複雑化している[1]。新発見に伴い、新しい分野が生み出されたり、複数の分野が融合して新しい分野になったり、あるいは単一分野が複数の分野に分化したりすることで、その構造が変化している。このようにダイナミックに変化する知の構造を把握することで、科学技術政策、とりわけ重点分野ターゲティングに活用しようという動きがでてきている。内閣府総合科学技術・イノベーション推進事務局では、科学技術・イノベーション政策立案に向けて、論文書誌データベースに含まれる情報を分析対象として、科学研究の変化を把握可能とする試みを行ってきている[2]。本稿では、その一環として作成された、科学研究の状況を把握可能とすることを目標に作成された論文マップの作製およびそれを用いた分析事例について紹介する。

## 2. 先行研究

論文間の引用関係に基づいて論文から把握可能な科学の構造を把握しようとする試みは、H. Small による先駆的研究[3, 4]を端緒とし、現在では多数の研究者によって行われるに至っている。日本では NISTEP により、Small の手法を踏襲する形で「サイエンスマップ」が作成され[5]、現在にいたるまでその更新が行われている。Small のアイデアは論文引用関係（直接引用、共引用、書誌結合）によって論文間の意味的近接性を定義し、それを一定の論文集合に対して適用することで論文間の近接性をグラフ構造によって表現可能というものであった。当初は少数の論文集合に対して分析が行われたが、科学全体の発展傾向を知りたいというニーズが高まったこと、またその処理に必要なデータの入手可能性の向上・処理に必要なコンピューティング資源の入手が容易になることによって、より多くの論文によって構成される論文集合に対して引用関係に基づく関係性の可視化が行われるようになっていった[6-9]。

論文間の関係が定義された後に、グラフレイアウトアルゴリズムによって 2 次元平面内にその構造が表現される。論文間の構造は巨視的構造と微視的構造を伴い、そのディテールには科学発展に関する豊富な情報を含む可能性がある。可視化を通じて視覚的に認知された結果は、分析者による仮説形成に供され、得られた仮説は後に行われる定量的分析や統計的推定によって評価される。グラフレイアウトアルゴリズムには多様なアプローチが存在するが、引用関係に基づく論文近接性で構成されたグラフは自明な構造を持たないことから、そのレイアウトには「ばねモデル」[10][11]を用いられることが多い。しかし、ばねモデルでは対象論文の 2 乗程度の関係性を考慮する必要があること、グラフの規模が大きくなると最適化が困難であることから、様々な工夫が行われてきた[12]。巨大なネットワークに対してはグラフ全体ではなくそのスパニングツリーのみを抽出して描画されることも多い[13]。

また論文近接性に基づくグラフ構造は可視化によって得られる視覚的特徴のみに基づく分類には限界がある。可視化とは別に、書誌情報に基づく論文近接性によって構成されたグラフをグラフ分割アルゴリズムによって分割することで、関連性が高い論文集合をクラスターとして抽出し、萌芽的研究領域の同定に使用している研究がある[14]。グラフクラスタリング手法の選択によって把握可能性は異なる。2000 年代にはグラフ一般に適用可能な Newman 法[15]が使用されていたが、最近では計量書誌分析に適した Leiden 法[16]が使用されることが多い。このようにして、新興分野の同定を目的とした可視化は Elsevier 社の SciVal や Clarivate 社の InCites など研究 IR 用システムでも利用可能である。

## 3. 論文マップ作成の手法そのその結果

上述したように既存の論文マップやそれを通じた分析が可能なシステムは存在しているが、科学技術動向を把握可能とするには既存マップには無い拡張性が求められる。例えば、論文書誌データベースから入手可能な情報だけでなく、他の情報源から得られた情報（例えば、内閣府においては e-Rad や標準

化データ[17]や他の調査によって得られた結果）をマップ上に付与したい、欧米の論文書誌データベースには日本語文献のカバレッジが低いためその追加ができるようにデータソースを拡充可能としたい、等のニーズが存在する。また、論文マップの作成においては数多くの任意性（arbitrariness）が存在しており、結果の解釈にはその過程に関する深い理解が求められる。以上の理由から、内閣府では調達した論文書誌データベース（世界バルクデータ、Scopus および Scopus に J-STAGE の文献情報を追加し、相互の引用関係を同定したもの、および Dimensions）を基にしたマップを作成中である。

既に Dimensions を使用した論文マップについては、以下のようにして作成している。Dimensions に含まれる全ドキュメントタイプの文献情報について、2010～2019 年の 10 年間に出版された論文のうち、論文分野(FOR)ごと、年ごとに上位 10%の被引用数を有する論文を対象とした。また、論文間の関係性については共引用を用いて算出した。共引用によって構成されたグラフは重みづけ Leiden アルゴリズムによってクラスタリングを行った。また、得られたクラスターは再度クラスタリングを行った。分析対象となった全論文(上述のトップ 10%論文)の相互の近接性を 2 次元レイアウト(node2vec によってグラフを高次元空間に埋め込んだけっかを t-SNE で 2 次元に縮退)し、クラスター、サブクラスター、論文の各々の単位で評価に活用可能な量（論文数、被引用数、引用関係の国際性、引用関係の多様性等）を算出し、Tableau によって表示し、探索的な分析を可能とした。作成したマップの特徴と NISTEP のサイエンスマップ 2018」を表 1 に比較した。また、作成されたマップを図 1 に示した。図 1 では論文間の局所的構造（関連性が強い論文が近接配置されること）と大域的構造（大括りで分野間の関係が適正であること）があることが確認されたが、個々のクラスターにはかなり幅広い研究分野が含まれており、融合分野や新興分野がクラスター内に埋没していたことから、マップ作成対象を個別のクラスターに限定し、そのクラスターに属する論文間の関係性をマップ化した(図 2)。

**表 1　本研究で作成された論文マップ(本マップと表示)と「サイエンスマップ 2018」の比較**

| | サイエンスマップ 2018 | 本マップ |
|---|---|---|
| 使用データ | ESI (Clarivate) | Dimensions (Digital Science) |
| 対象文献種別 | Article, Review | Article (Review を含む), Chapter, Proceeding, Preprint, Monograph, Edited Book |
| 分析対象とした文献 | Top 1% (ESI), 2014～2018 年 | Top 10%, 2010～2019 年 |
| 近接性の算出 | 共引用度 | 共引用度（補正付き） |
| クラスタリング手法 | 2 段階(リサーチフロント→研究領域)<br>→902 個のクラスターに分割 | 2 段階 (Leiden アルゴリズムによる)<br>→1,524 個のクラスター(うち 1,076 を使用)<br>→12,445 個のサブクラスターに分割 |
| 可視化手法 | ScienceMap visualizer (経年変化を考慮した spring model) | グラフ分散表現と t-SNE の組み合わせ |
| 規模（可視化対象論文数） | 28,824 件 (コアペーパのみがネットワークに貢献するためその数) | 2,224,645 件<br>(サイエンスマップ 2018 の約 77 倍) |

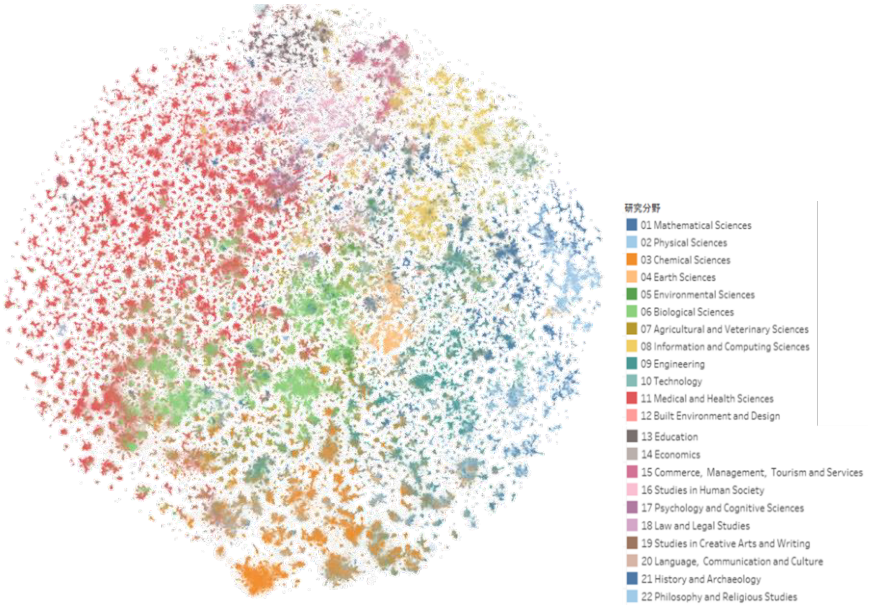


**図 1 共引用によって定義した論文間の近接性によって構成されるネットワークのネットワーク分散表現を t-SNE により 2 次元化したもの。点は論文を示し、点の色は分野(FoR22 分類)を示す**


**図 2 特定クラスター(リチウムイオン電池を含む二次電池に関連するクラスター)に属する論文の共引用による近接性を可視化したもの。図の点は論文を意味し、点の色は当該論文が所属するサブクラスター(クラスターを Leiden アルゴリズムによって再分割したもの)を示す。**

クラスター単位でマップを作成することで、クラスター内の研究分野としての構造とその相互の関係について洞察を深めることが可能となった。図 2 はリチウムイオン電池を含む二次電池に関連したクラスターと推定されるが（この推定は論文集合に含まれる論文のタイトルおよびアブストラクトから得られた特徴語によって推定している）二次電池の構造（正極材料、陽極材料、セパレーター、電解質等）やその物質系によってサブクラスターが生成されていること、またその相互の関係はおおむね整合的であることが判明した。

## 4. 論文マップの評価

以上のようにして作成された論文マップには、クラスターやサブクラスター単位で算出した様々な指標を表示可能とし、それを用いて研究分野の発展状況や、国別論文アウトプットの比較等を行えるようにしている。実際に、量子技術、二次電池、サイバーセキュリティ等の分野について専門家による評価を行い、おおむね専門家が把握する各分野の構造を反映していることが確認されている。一方で、論文マップは共引用によって形成しているため、相対的に引用数が少ない近年出版された論文の構造把握は不得意であること、また全論文の 10%のみの把握であって、分野全体の傾向を把握しているものでは必ずしもないことが指摘された。また、クラスターやサブクラスターの把握において、都度論文集合を構成する論文を仔細に読み解くことは現実的ではないことから、自然言語処理（特徴語抽出アルゴリズム）を用いて、特徴的な語を自動抽出し、分野把握の補助に供しているが、その精度が必ずしも高くなく、専門家の再解釈抜きには分野特徴を把握しにくいという点も課題であった。

## 5. おわりに

以上のようにして作成した論文マップは有用性があり、その活用方法によっては政策立案や政策評価に活用可能な情報を抽出可能であることが判明した。今後は、日本語文献の拡充(Scopus＋J-STAGE)、特許データの追加（全特許をパテントファミリー単位で引用関係を構築してマップ化)、近接性算出手法の変更（共引用以外に、書誌結合、直接引用を加える、またテキストデータから得られるドキュメント間の近接性を論文間の近接性算出に導入する）などを加え、より有用性が高いデータを作成し、またその操作を Tableau を用いてインタラクティブに可能とし、多様な分析者のニーズに応えることができるシステムを作成したいと考えている。

## 謝辞

本研究は内閣府令和４年度科学技術基礎調査等委託事業「エビデンスデータベースの利活用高度化に関する調査」の成果の一部である。

**参考文献**


1. Dimensions (httpp://www.dimensions.ai/)に基づく分析(2022 年 8 月 27 日)
2. 内閣府総合科学技術・イノベーション推進事務局が行うエビデンスに戻るく政策立案に関連した活動はウェブサイト e-CSTI (http://www.e-csti.go.jp/)に掲載されている。(2022 年 8 月 27 日)
3. Small, H., *Co-citation in the scientific literature: A new measure of the relationship between two documents.* Journal Of The American Society For Information Science, 1973. **24**(4): p. 265-269.
4. Small, H. and B.C. Griffith, *The Structure of Scientific Literatures I: Identifying and Graphing Specialties.* Science Studies, 1974. **4**(1): p. 17-40.
5. NISTEP による「サイエンスマップ」は「サイエンスマップ 2004」(2007 年 3 月)として最初に公刊され、その後現在まで作成が継続されている(最新は 2020 年に公刊された)。
6. Boyack, K.W., C. Smith, and R. Klavans, *A detailed open access model of the PubMed literature.* Sci Data, 2020. **7**(1): p. 408.
7. Börner, K., et al., *Design and Update of a Classification System: The UCSD Map of Science.* PLoS ONE, 2012. **7**(7): p. e39464.
8. Boyack, K.W., R. Klavans, and K. Börner, *Mapping the backbone of science.* Scientometrics, 2005. **64**(3): p. 351-374.
9. Boyack, K.W. and R. Klavans, *Creation of a highly detailed, dynamic, global model and map of science.* Journal of the Association for Information Science and Technology, 2014. **65**(4): p. 670-685.
10. 論文を仮想的に質点として捉え、質点間には仮想的に論文近接性に比例したばね定数を持ったばねが存在し、相互作用しているものとして、運動方程式を解き安定的な配置を得るという方式である。
11. Kamada, T. and S. Kawai, *An algorithm for drawing general undirected graphs.* Information Processing Letters, 1989. **31**(1): p. 7-15.
12. Martin, S., et al. *OpenOrd: an open-source toolbox for large graph layout*. in *Visualization and Data Analysis 2011*. 2011. SPIE.
13. Adai, A.T., et al., *LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks.* Journal of Molecular Biology, 2004. **340**(1): p. 179-190.
14. Jarneving, B., *Bibliographic coupling and its application to research-front and other core documents.* Journal of Informetrics, 2007. **1**(4): p. 287-307.
15. Girvan, M. and M.E.J. Newman, *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences, 2002. **99**(12): p. 7821-7826.
16. Traag, V.A., L. Waltman, and N.J. Van Eck, *From Louvain to Leiden: guaranteeing well-connected communities.* Scientific Reports, 2019. **9**(1).
17. 内閣府政策統括官（科学技術・イノベーション担当）付「研究力の分析に資するデータ標準化の推進に関するガイドライン」（平成 31 年 4 月 5 日）
https://www8.cao.go.jp/cstp/evidence/guideline_honbun.pdf