１Ｃ２１

オープンな引用データ(COCI)を用いたサイエンスマップ

○渡邊勝太郎（国立研究開発法人科学技術振興機構）
katsutaro.watanabe@jst.go.jp

1. はじめに

引用情報のデータセットは、Web of Science や Scopus に代表されるような商用製品が主流であるが、近年、COCI[1]のように単純なデータサイズではこれらに比しうる大規模な引用データセットがオープンなライセンスで公開されている。本発表では、まだ実例が少ないオープンな引用データ(COCI)の利用例の開拓を目指す。具体的には、NISTEP のサイエンスマップ調査[2]と同様に、共引用を利用した学術分野全体を地図のように俯瞰することができる可視化分析を行う。可視化結果の WEB ページ、詳細な作成方法、作成したマップのデータを著者のリポジトリ(https://github.com/k2taro/COCI_map2021)で公開しているので、ぜひご参照いただきたい。

なお本発表は著者の所属である科学技術振興機構(JST)の業務として行ったものではない。JST における研究開発戦略立案や競争的研究費を扱う業務等とは一切の関係がないことをご承知おき願いたい。

2. 論文のクラスタリング

マップ対象論文の選定

NISTEP サイエンスマップでは特定の 6 年間に発行された被引用数が Top1%に入る高被引用論文を対象にしているが、本発表では、計算時間の軽減や、最新のマップを作成する目的から、概ね 2019 年以降に発行された Top1%論文となるように以下の手順によって対象論文を選定した。

元のデータセットには COCI の 2022 年 1 月ダンプデータ(https://doi.org/10.6084/m9.figshare.6741422.v13)を利用した。COCI のデータには被引用論文の発行年のデータがそのまま含まれていないため、引用情報作成日が 2019 年 1 月 1 日以降、かつ、引用情報作成日からタイムスパン(引用元論文の発行からの時間経過)を引いた年部分が 2019 以上、との条件で抽出し疑似的に 2019 年以降に発行された論文に対する引用情報を抽出した。結果、46,563,981 の引用情報と被引用数が 1 以上の論文 7,450,676 報が抽出された。

次に、分野ごとに Top1%の高被引用論文を特定するために、まず被引用数 Top5%の論文 390,587 報を特定した。これに対して COCI の API から書誌情報を取得し掲載誌情報の ISSN と Scimago(https://www.scimagojr.com)の雑誌リストを照合し分野を特定した。これを分野ごとにさらに被引用数上位 20%を抽出することにより、最終的に Top1%の高被引用論文 78,026 報を特定した。最初に Top5%論文を特定したのは、API からの書誌情報の取得に膨大な時間がかかるためである。

マップ対象論文のクラスタリング

上記で特定したマップ対象論文を、NISTEP サイエンスマップと同じく共引用数をそれぞれの被引用数で正規化した共引用度(cosine)を用いたクラスタリングを行った。NISTEP サイエンスマップでは第 1 段階は共引用度 0.3 でクラスタリングを行っているが、この閾値はデータセットに応じて変わり得ると考えられるため閾値自体の検討を行った。図１－１は、クラスタのサイズが 5、10、50 以上の数を共

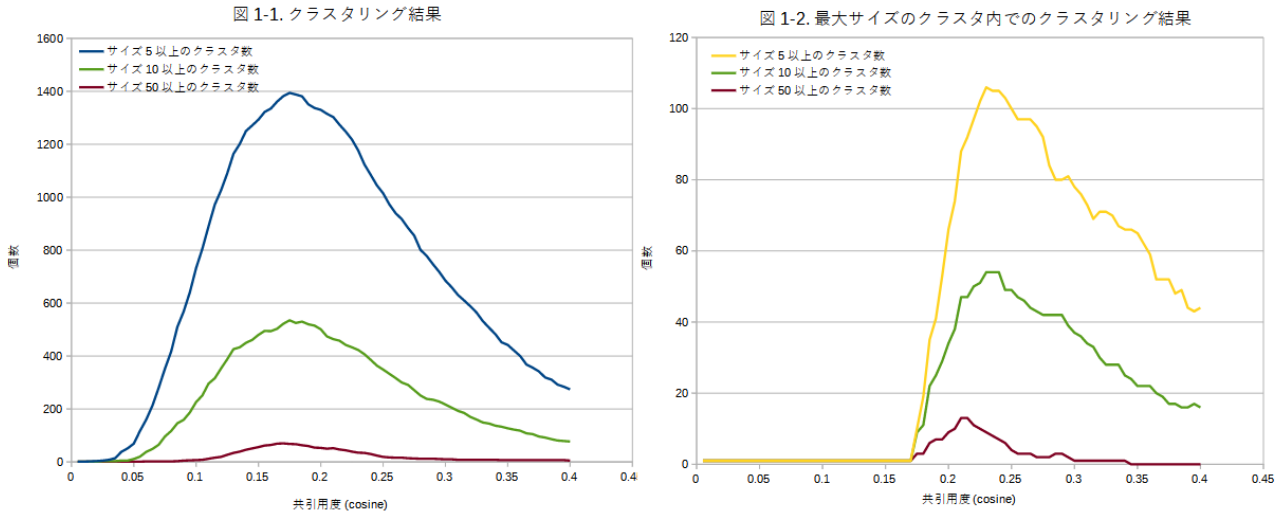

図１．共引用度と特定サイズ(5～50)のクラスタ数の関係

引用度ごとにプロットしたものである。どのサイズの数も共引用度 0.17 付近にピークがあることがわかる。共引用度 0.17 の閾値でクラスタリング(最小クラスタサイズ 10 以上)したところ、サイズが 1,000 を超える巨大なクラスタもできてしまった。図１－２は最大のサイズ(3,452)のクラスタに対して同様にグラフを作成したものだが、ピークは 0.22 付近となり更に高い閾値で再クラスタリングを行うのが適当であることが示唆された。よって、極端にサイズの大きい 3 クラスタ(それぞれ、3,452、1,991、730)については更に 0.19 の閾値で再クラスタリングを行うことで最終的なクラスタリング結果として全部で 527 クラスタが得られた。なお、閾値 0.17 の共引用度でクラスタサイズ 10 未満の論文がマップ不可となり、マップ対象論文は 21,836 報となった。

3. 可視化

地図上の可視化(クラスタの 2 次元空間へのマッピング)

2 節で作成した各クラスタを 2 次元空間へ配置するため、各クラスタ間の共引用を再計算した。この共引用を力学モデルのグラフ描画アルゴリズムの一種である Fruchterman-Reingold アルゴリズムを用いて 2 次元空間への配置を計算した。共引用数や共引用度をエッジの重みとして計算せずとも見通しの良い結果が得られたため、単純な共引用関係のみで計算している。

各クラスタを構成する論文は 2 節のマップ対象論文の選定をした際に付与した分野情報があるため、クラスタ内の構成論文のうち最も多い分野をそのクラスタの分野とした。最多分野の割合が 50%を超えたクラスタは 241、30%超えたクラスタは 480 クラスタとなっており多くのクラスタでは支配的な分野が存在することがわかる。

分野ごとに色分けを行って可視化したものが図２である。論文数の多い医学(Medicine)、材料科学(Material Science)が目立つが、医学分野に近い位置に生化学(Biochemistry)が位置されているなど、分野の関連を考えても妥当と思われる可視化がされている。

クラスタ構成論文の可視化

各クラスタの内容をわかりやすく表示するために、構成する論文の著者、タイトルに含まれる単語の頻度、共起関係、また、分野ごとの割合(Scimago の分野はの 2 階層で構成されるためこの構造も反映)を可視化した詳細なページも作成した。著者は名前を名寄せされず単純な表記を取ったのみ、またタイトル語も単純な 1 単語ごとの切り出しのため簡易な内容ではあるが、クラスタよっては 100 報を超える論文の概要を掴むには便利な可視化を実現している。

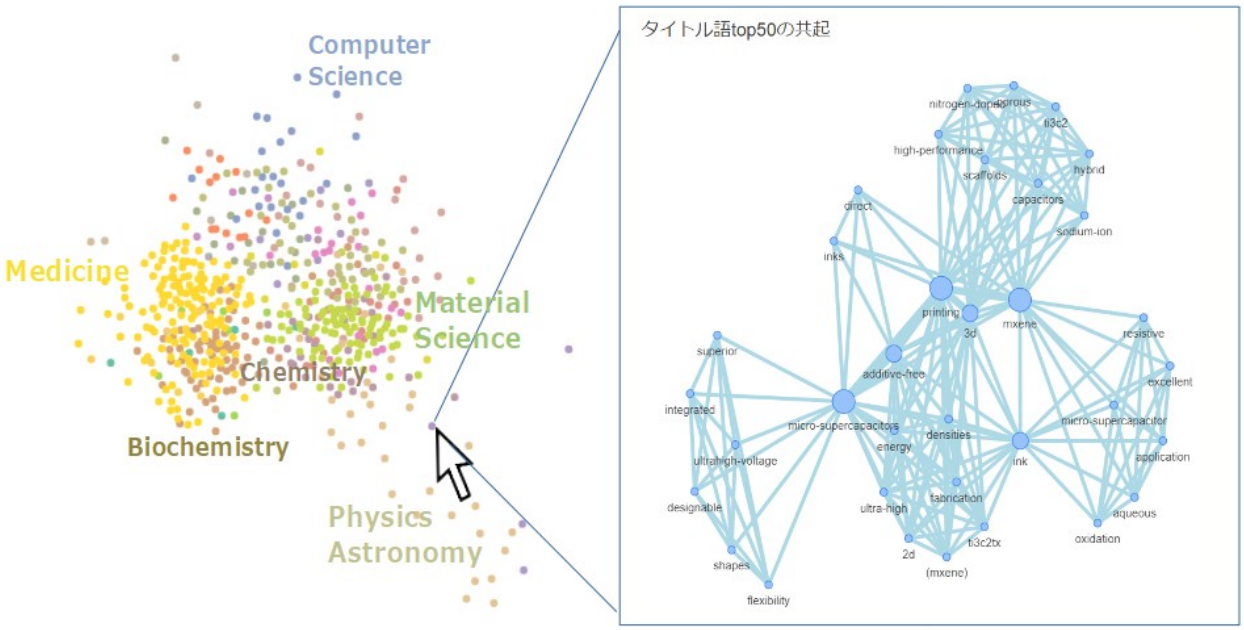

図２．可視化のイメージ(※https://k2taro.github.io/COCI_map2021/page/ で公開している)

4. 具体事例の検討

具体事例 1：COVID-19 関連の研究の広がり

2019 年以降、学術論文に関わらず Covid-19 は世界的に注目をで浴びたトピックである。今回は単純にマップ対象論文のタイトルに「covid」という文字列が含まれる 2,314 報を関連研究の論文であると判断し作成したマップにオーバーレイを行った(図３)。Covid-19 関連論文は全 84 クラスタに渡っており、

このうち 63 クラスタは医学分野(Medicine)でありオーバーレイでもその様子がよくわかる。この結果は当然のものと思われるが、医学分野以外のクラスタにおいて支配的なトピックとなる特徴的なクラスタがいくつか存在しており、Covid-19 の分野を広げた影響を示していると考えられるため表 1 にまとめた。人の移動・行動と行動規制に関するもの、歯科・歯科医師への影響(クラスタ:4372_。総論的な論文も多くあり特定分野のホットトピックであったと考えられる)、大気汚染と気温が Covid-19 の感染に与える影響に関するもの、消費者の買占め行為に関するもの、行動変容からおこる電力・エネルギー需要の変化に関するもの、など、各分野によって様々文脈の研究が行われていたことが推察できる。

表 1．Covid-19 関連の特徴的なクラスタの例

クラスタ	クラスタサイズ	Covid-19 論文	シェア	分野	タイトル語(Covid-19 は除く)
17229_	16	16	100%	3300:Social Sciences	Travel, Impact, Behavior, Restriction, Transport
4372_	22	19	86%	3500:Dentistry	Dental, Dentistry
517_	69	65	94%	2300:Environmental Science	Air Pollution, Temperature, Transmission
15294_	15	14	93%	1400:Business, Management and Accounting	Panic Buying, Consumer
6755_	16	16	100%	2100:Energy	Energy, Electricity, Demand,

具体事例 2：科研費成果論文とのオーバーレイ

マップにおける日本の存在感を確かめるために、科学研究費助成事業(https://kaken.nii.ac.jp)の成果論文をマップにオーバーレイ(図３)を行った。対象となる科研費成果論文は、研究成果情報 2019 年以降で種別が雑誌論文とされているものから DOI が付与されているものとした。対象論文数は 242,470 報(重複排除。2022/9/1 時点)あり、オーバーレイした結果 502 報の論文が 151 クラスタにわたって存在することがわかった。

さらに詳細を見るために、クラスタごとに科研費成果論文のシェアが大きいものに注目した。シェアが 10%を超えるものは全部で 41 クラスタあり、多い分野は生化学(Biochemistry)9、物理・天文学(Physics and Astronomy)9、医学(Medicine)7 であった。中でも特にシェアが大きいものを小規模(サイズ 50 未満)、大規模(サイズ 50 以上)に分けてそれぞれ上位 5 クラスタを示したものが表２である。元のクラスタサイズが大きい傾向のある医学分野に比較して、物理・天文学の分野がより目立つ形になった。

表２．科研費成果論文のシェアが大きいクラスタ

小規模クラスタ(シェアが大きいもの上位 5)

クラスタ	クラスタサイズ	科研費成果論文	シェア	分野	タイトル語
26500_	17	8	47%	3100:Physics and Astronomy	Skyrmion, Magnetic, Lattice, Topological
4076_	32	13	41%	1900:Earth and Planetary Sciences	Asteroid(101955)Bennu, White Dwarf, Surface
5963_	13	5	38%	2700:Medicine	Influenza, Baloxavir, Susceptibility, Endonuclease
3912_	11	4	36%	3100:Physics and Astronomy	Modular A4/S4, Mixing, Neutrino, Invariance
576_	29	9	31%	2700:Medicine	Neuromyelitis Optica Spectrum Disorders, glycoprotein Antibodies, MOG

大規模クラスタ(シェアが大きいもの上位 3。シェア 10%を超えるものは 3 つだった)

クラスタ	クラスタサイズ	科研費成果論文	シェア	分野	タイトル語
3898_	59	12	20%	3100:Physics and Astronomy	Primordial Black Hole, Gravitational Waves, Dark Matter
3906_	64	11	17%	3100:Physics and Astronomy	Muon Anomalous Magnetic Moment, Hadronic Vacuum Polarization, Anomalies, Model
1369_	427	47	11%	3100:Physics and Astronomy	Twisted Bilayer Graphene, Topological, Non-Hermitian

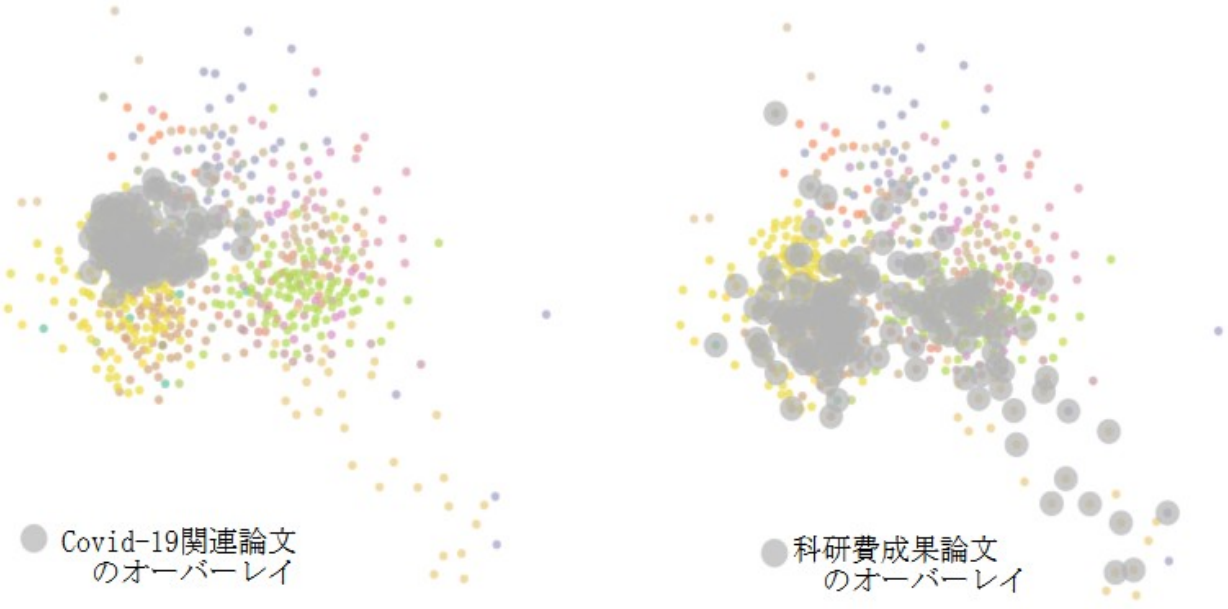

図３．Covid-19 関連論文、科研費成果論文のオーバーレイ

5. 議論

4 節の具体事例で見た通り、COCI を利用して意味のある議論が可能な結果が得られた。しかし、データソースの限界(著者、所属機関の名寄せ)により、NISTEP サイエンスマップで行っているような国際比較ができないなどの分析の限界は存在する。著者の名寄せについては COCI の API から得られる書誌情報では一部に ORCID が付与されており、マップ対象論文の著者の総数 305,038 人(延べ人数)のうち 45,339 人の ORCID 情報が含まれていた。ORCID の普及による今後の情報拡充に期待したい。

一方で、今回の可視化分析では 2019 年のある一時点の情報しか対象にできていないため、過去に遡り同様の可視化分析を行うことによる時系列変化も可能と思われるし、NISTEP サイエンスマップの Sci-Geo チャートのように、クラスタを論文の発行年や他のクラスタとの共引用度の度合などの分野以外の多面的な観点から分類することも考えられる。本発表の内容は著者の個人的な活動ではあるが、今後も取り組みを継続していきたい。

参考文献

[1] Heibi,I., Peroni,S. & Shotton,D; Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. Scientometrics 121,1213-1228(2019); DOI:https://doi.org/10.1007/s11192-019-03217-6

[2] 科学技術・学術政策研究所科学技術・学術基盤調査研究室; サイエンスマップ 2018―論文データベース分析（2013-2018 年）による注目される研究領域の動向調査― (2020); DOI:http://doi.org/10.15108/nr187