2023年5月26日　統計数理研究所　オープンハウス

データによる研究テーマの可視化

濵田　ひろか　　モデリング研究系　特任研究員

1.　研究多様性指標（Research Diversity Index, REDi）の開発

統計数理研究所では、平成二十八年度より大学や研究機関での研究活動およびその成果を可視化し、機関の意思決定に利活用されることを志向する「研究IR（Institutional Research）」[1]の方法論について、統計学的見地からの確立を目指し、公募型共同利用の重点型研究を中心に多角的に取り組んできました。その成果の1つとして、私たちは「論文」の新しい評価指標、**研究多様性指標（Research Diversity Index, REDi）**[2]を開発しました。従来より、イノベーションを引き起こす重要な要素と考えられている「**多様性**」の観点を加えた、論文の分析が可能です。

この指標で論文を評価した場合、REDiが高い論文は「**多様性**が高い」、REDiが低い論文は「**専門性**が高い」と捉えることができます。同じ被引用回数をもつ論文でも、REDiスコアで比較すると、その被引用傾向の違いが可視化されます（表1、図1）。

No	論文のタイトル	論文の掲載ジャーナル	発行年	被引用数	研究多様性
67	Identifying FGA Peptides as Nasopharyngeal Carcinoma-Associated Biomarkers by Magnetic Beads	JOURNAL OF CELLULAR BIOCHEMISTRY	2012	10	26.39
134	Studies on gambogic acid (IV): Exploring structure-activity relationship with I kappa B kinase-beta (IKK beta)	EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY	2012	10	55.03
5	Geochronological, geochemical and Sr-Nd-Hf isotopic constraints on the origin of the Cretaceous intraplate volcanism in West Qinling, Central China: Implications for asthenosphere-lithosphere interaction	LITHOS	2013	10	99.99

表1. REDiによる論文評価対象例

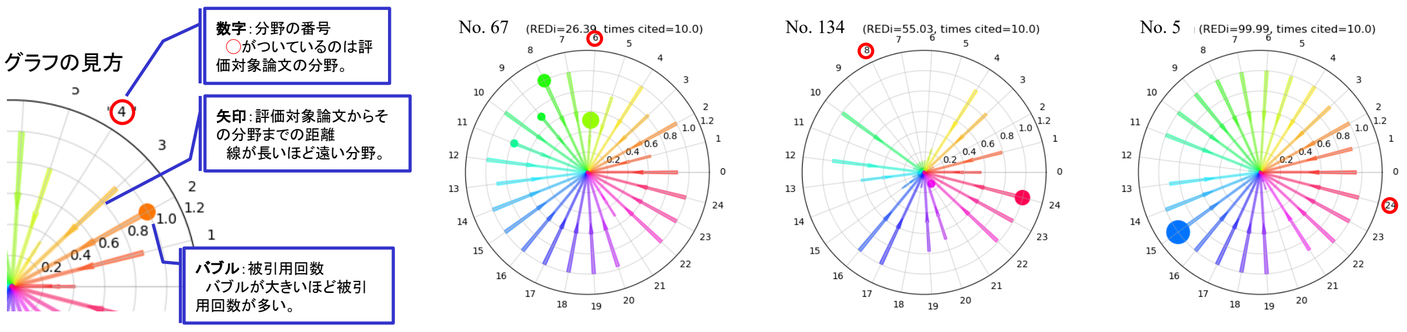

図1. REDi による論文評価例

2.　研究テーマをデータから定義する

現在、私たちはIR実務者の目線からREDiの改良に加えて、研究分析基盤システムの機能拡充を目指しています。今一度研究IRの実務や意思決定の場面に立ち返り、改めて研究活動にフォーカスしてみると、例えば研究者の研究活動について複数年の時系列推移を考慮した場合に、その研究スタイルの多様さに気づかされます。同じ分野の研究者で比較しても、専門性の高い（REDiスコアの低い）論文が多く、掲載ジャーナルが同一クラスタに所属する頻度が高いケースや、REDiのスコアは論文ごとにばらつきが大きく、掲載ジャーナルも様々なクラスタに散らばっているケースなどが存在します。

このように、任意の研究者の研究成果群に対してREDiを活用することによって、その研究スタイルないし研究活動傾向を多様性の観点から可視化することは可能ですが、その研究成果群は実態としてより詳細に研究テーマ単位に分割したうえで評価することが妥当です。

そこで私たちは、研究テーマや研究課題のような研究活動の軸（あるいは中心点）を論文などのデータから定義することで、実務の場面において評価対象となる研究を前提知識なしにより適切に細分化ないし集約し、意思決定に資する形式で提供する手法について検討を行っています。

研究や課題の最小単位の一つとして、論文が考えられます。また、私たちが新たに定義したい、「データから明らかにした研究テーマ」は、論文を適切にクラスタリングすることによって得られると仮定し、オープンアーカイブに登録された論文を対象に予備実験を実施しました。

より幅広い分野について同様の手法によってデータから研究テーマが定義できることを重視し、2億6千万件を超える様々な科学分野の書誌を持つ Semantic Scholar（ https://www.semanticscholar.org/ ）に着目。Semantic Scholar が保有するデータを対象とした検索を実施できる、Semantic Scholar Academic Graph API（S2AG API）[3] を活用します。

まず始めに、論文が有する属性情報（図2）から、リレーションシップの観点によってノードとエッジを定義し（図3）、一人の任意の研究者を評価対象としたリレーショナルネットワークを作成します（図4）。

次に、対象の研究者に紐づく研究テーマを適切に分類する目的で、ネットワーククラスタリングを行います。リレーショナルネットワークをクラスタリングする手法としては、Community Detection [4]を選択します。これは、エッジの接続によるクラスタリング手法であり、接続の粗密によってコミュニティを抽出する手法です。

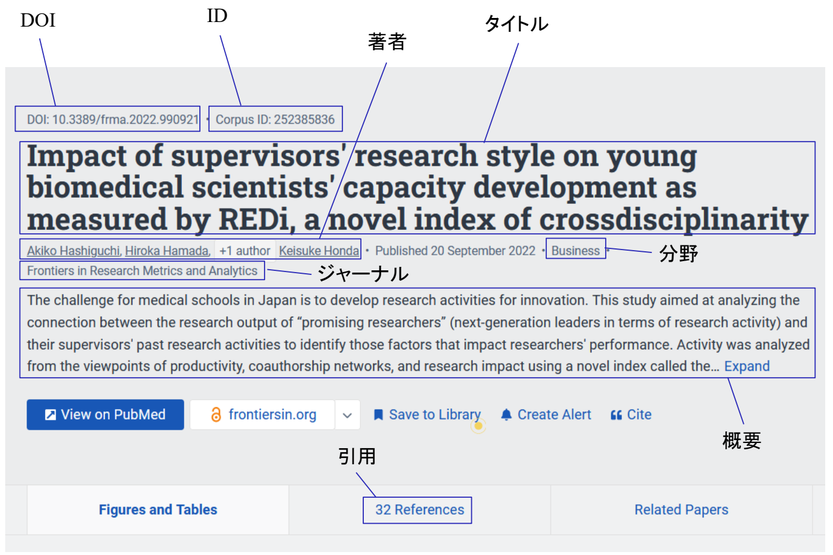

図2. 論文が有する属性情報

図3. 論文の属性情報に関するリレーショナルネットワーク表現

作成した任意の研究者のリレーショナルネットワークを、Community Detectionによってクラスタリングすると、例えば図5の結果を得ることができます。また、分析に使用した研究者の論文群に対して、クラスタリング結果を付与したものが表2です。例えば Cluster 6 のように、著者名の表記揺れによって同一著者グループと判断されない場合にも、その他の属性情報のつながりによって、同一研究テーマクラスタと判断されます。また、Cluster 10 のように、共著者の増減があるケースでも、派生研究や発展研究といった周辺の研究内容については、同一の研究テーマクラスタとしてクラスタリングされました。このように、リレーショナルネットワークを活用することで、効率的に研究を集約できる可能性が示唆されました。

各研究テーマクラスタの自然言語的解釈については現在検討中ですが、例えば、ひとつの研究テーマクラスタに属するすべての論文のタイトルと概要からワードクラウドを作成すると、そのクラスタにおいて出現頻度の高い単語を得ることができます。

これらの手法を組み合わせ、研究テーマクラスタおよびそのクラスタの特徴をデータから明らかにし、研究テーマを定義する手法を確立します。

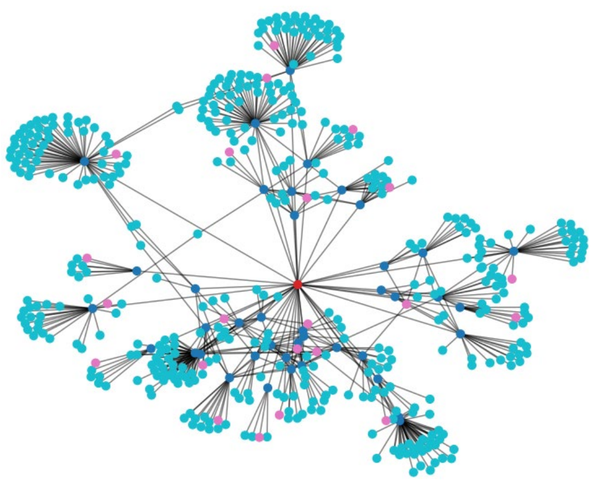

図4.任意の研究者を中心としたリレーショナルネットワーク表現

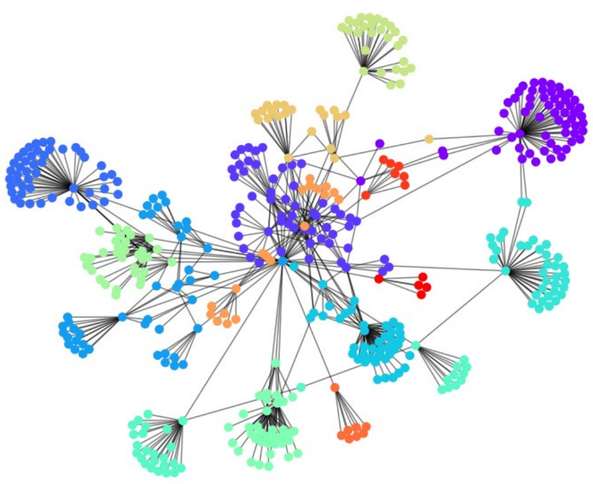

図5. 研究テーマクラスタサンプル

Cluster	PaperID	Title	Authors
6	2009885c3a56b901309ce68ca0503a2fd6d5a308	A generative model of article citation networks of a subject from a large-scale citation database	[A, B, T]
6	29ae3ba80abddb48a98dcb8a1de8ea674a848eef	A New Metric for the Analysis of the Scientific Article Citation Network	[A', B', T]*
7	f34491bd6d1794a05c756227225c54a04478fc9a	$${¥mathcal{M}}$$ -decomposability and symmetric unimodal densities in one dimension	[C, T]
7	58892e05c73912ae0b51e495cddc50247694799b	And Applications to Clustering and Kernel Density Estimation	[T]
7	f131ce3a6625a10cfdfe964ee7e6b09a762b4649	M-decomposability, elliptical unimodal densities, and applications to clustering and kernel density estimation	[C,T]
10	b6eb9d4ef48fa43cc46852a031fdd10948bcec2e	Programming Statistical Data Visualization in the Java Language	[T,D,H]
10	ed62334f2075dab65b52eff5eeed3119144e53c3	3 dimensional parallel coordinates plot and its use for variable selection	[H,T]

表2. 任意の研究者の研究テーマクラスタサンプル
（*: 著者A'，著者B'は著者A，著者Bの表記揺れ）

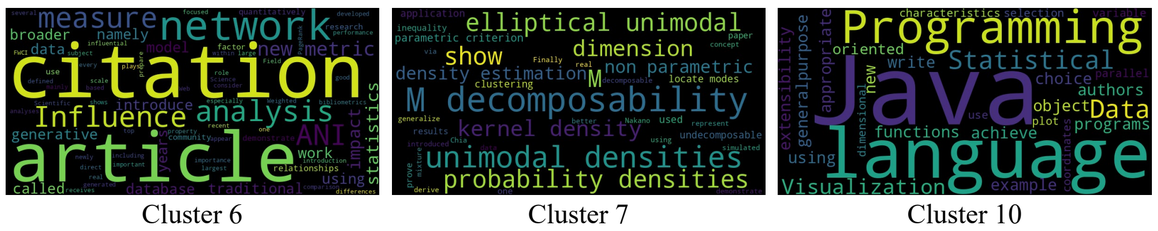

図6. 任意の研究者における研究テーマクラスタのワードクラウドサンプル

3.　今後の展望

本研究によって、データから定義した研究テーマクラスタを得ることができれば、事前知識なしに研究者の研究活動について、その成果物情報から客観的に研究テーマを可視化することが可能になります。研究者ごとに複数もつケースもある研究テーマのそれぞれの位置づけや、分野特化型、分野横断型、などの研究スタイルのように、これまでは自然言語情報や専門的な事前知識がなければ分析が困難であった研究者の研究活動について、理解を支援することが可能です。

適切な研究評価には、従来通りのインパクト偏重な評価活動を改め、その研究活動を多角的に捉えていくことが大切です。本研究成果は、その一助となります。

<参考情報>

[1] 本多 啓介 (2020).「特集 Institutional Researchと統計科学」について, 統計数理 68号(2), 195-196.

[2] Hashiguchi, A., Hamada, H., Takahashi, S., & Honda, K. (2022). Impact of supervisors' research style on young biomedical scientists' capacity development as measured by REDi, a novel index of crossdisciplinarity. Frontiers in Research Metrics and Analytics 7.

[3] The Allen Institute for Artificial Intelligence. Semantic Scholar Academic Graph API. https://www.semanticscholar.org/product/api（最終アクセス 2023年5月10日）

[4] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E 70(6).