１Ｃ０９
引用構造のフラクタル次元として定義されるスケール不変な派生 h-index

○藤田裕二（内閣府（当時）政研大（現在)），宇佐美徳隆（内閣府（当時）名古屋大（現在)）

1. はじめに

引用構造の分析は学術政策にとどまらず、複雑系研究一般においても重要な研究対象である。引用構造の重要性は、そのグラフ理論的な含意だけでなく、それが知識拡散を担っているという点にある。知識の流れと広がりは知識生産における重要な要素であり、それゆえ（OECD1996）にいわれる「Knowledge-based economy」の主要な駆動要因でもある。本稿は（Fujita2022）として採録されたスケール不変な h-index 派生指標である h-dimension の紹介である。h-dimension はその設計と h-index 自体の統計的性質によりデータサイズ、分野等の影響をほぼうけることなく知識生産の効率を測定できるという、機関評価指標として望ましい性質を備えている。

h-index（Hirsch2005）はもっとも広く用いられている書誌データに基づく学術評価指標の一つである。その統計的性質については(Pratelli2012)に詳細な検討がある。有力な指標として多数の派生指標があるが、本稿に直接の影響を与えたものとして（Koizumi2018）の h5-index がある。既存の h-index および派生指標にはいずれも、指標がデータセットのサイズと強い相関を持つという性質があり、機関評価指標として用いるには困難がある。提案指標はこの問題を引用構造の自己相似性に着目することで解決するものである。

本研究は著者らが内閣府の総合科学技術・イノベーション会議（以下 CSTI）に所属し、科学技術政策に係る分析機能・データを共有するプラットフォームである e-CSTI の研究と構築に携わっていた折に、業務の一環として行われたものである。e-CSTI で注目する国立大学等の研究機関の論文データサイズには 3 桁以上の差があり、本研究が注目するデータサイズ依存性の問題の解消は重要な課題である。

2. 引用の自己相似性と提案指標 h-dimension

h-index は、ある論文の集合 X について定義される整数であり、h 以上の被引用回数を持つ論文が h 本あるとき X の h-index を h とする。[20, 10, 5, 4, 3]ならば 4 であり[4, 3, 3, 2, 2]ならば 3 である。h 個の要素を持つ、h-index を定義する X の部分集合を H とする。詳細は（Fujita2022）に譲るが、h-index は引用回数の分布関数に基づく不動点となっており、この発想は h-index の確率論的性質の把握を容易にするとともに、これに基づいて線形時間で指標を算出するアルゴリズムが構築可能である。研究者個人も研究機関の場合もともに、時間とともに学術論文はほぼ単調増加するので、引用構造もそれとともに規模と複雑さが増大してゆく。h-index が分布関数の不動点であることから、分布に大きな変化が無いと仮定すると、データ規模との正の相関が直ちに帰結する。

データ規模との相関についてはすでに（Hirsch2005）で対策がとられており、指標の数値そのままではなく研究者として活動した期間で割った値を評価基準としている事実は注目に値する。しかし研究機関にはそれぞれ独自の来歴や規模があり、時間による正規化は機関評価手段として採用するには問題がある。

図 1 は人工的に構成した引用構造の可視化であり、実在の引用構造とその統計的性質などを揃えてある。文献 A が全体の根ノードであり、被引用ノードが下、引用ノードが上に配置されるよう描画した。文献 B は A を引用しており、B の上に構成される引用構造は、図の全体構造の部分グラフである。この部分グラフは全体と引用数の分布がほぼ共通である。このように部分構造が全体構造と統計的性質を共有することを（Mandelbrot1977）では「統計的自己相似性」と呼称しており、本稿もこれに習う。

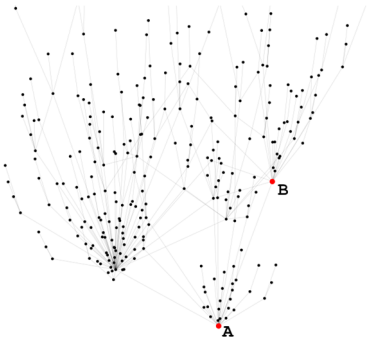

Figure 1: 引用構造の自己相似性

h-index 定義集合である H も同様に、全体構造の中で部分グラフとなっているが、図の中ではもっぱら下部に位置している。H も全体構造と次数分布を共有しているが、相違点もあり、小さい値の分布を欠いている。

実在の引用構造は有限であり、それゆえ統計的自己相似性も、たどれる詳細レベルには上限がある。時刻とともにネットワークには新たな頂点が新規論文として追加され、それとともに詳細レベルも向上する。もし、部分グラフ（ここでは部分集合が定義する部分グラフを便宜上、部分集合と同一視する）H が時間の経過にしたがって、ある定常状態に収束するのであれば、そのような終端状態は時間経過によらない（したがってデータ規模の問題もない）機関評価基準の基礎として望ましいものとなるはずである。

しかしながら、図 2 に明らかなように、そのような定常状態が存在するようには見えない。あるいは、観測の詳細レベルが向上するとともに、測定値も際限なく増大してゆくように見える。この問題は海岸線長の測量の問題、すなわち測定のスケールを小さくすると測量結果が大きくなる問題と同一であり、いずれも対象の自己相似性にその原因がある。

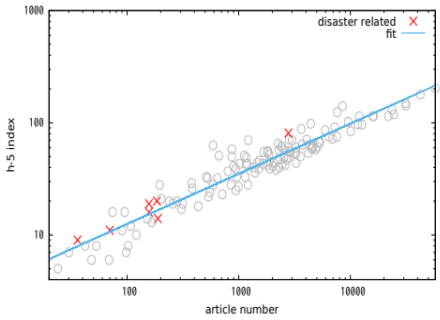

Figure 2: h-index と論文数

これら自己相似的対象の取扱いについては既知の有力な手法、すなわちフラクタル次元が存在している。本研究でもこれを応用して、論文集合 X の h-index が h、引用構造のスケールを s としたとき提案指標 h-dimension （以下h_d）を

$$h_d = \frac{log(h)}{log(s)}$$

と定義する。スケール s は引用ネットワークのサイズであるが、これは対象とする論文集合が持っている被引用数の総和として算出する。

3. h-dimension とその含意

本研究で提案指標を適用するデータは、DigitalScience 社によるデータベース製品 Dimensions の 2014 年から 2018 年および、2016 年から 2020 年の出版年をもつものを用いている。Dimensions では機関の同定を GRID（Global Research Identifier Database）に基づいて行っており e-CSTI の主要な関心対象である国立大学および研究開発法人等の独立行政法人 134 件の GRID について当該年限のデータを API を用いて取得（2014-2018 データ）および、e-CSTI 開発のために導入したバルクデータから SQL 化したものを用いた（2016-2020 データ）。該当する論文件数は 550,602（2014-2018 年）および 815,752（2016-2020 年）であった。

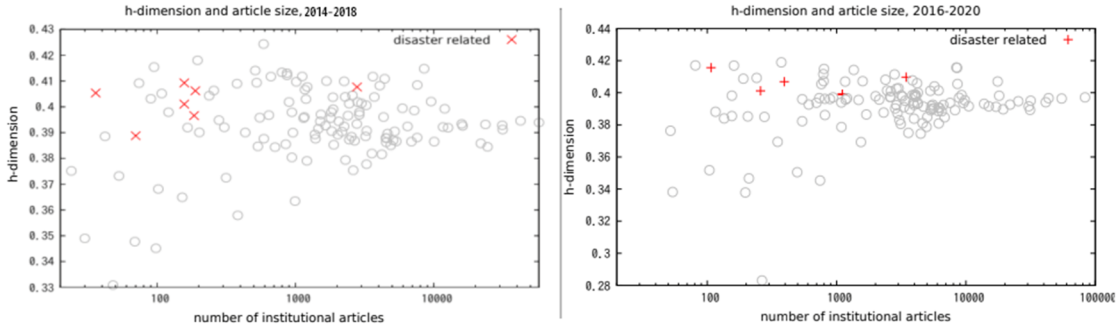

Figure 3: a) 2014-2018 プロット。b) 2016-2020 プロット：h-dimension

図 3 は二つのデータセットについて集計した h-dimension を機関ごとにプロットしたものであり、図同様両対数軸としている。赤い標識は防災関連の研究機関の値である。どちらのデータセットでも一貫して、規模はさほど大きくはないが、比較的良好なスコアを記録しており、多様な自然災害が頻繁に発生するわが国の状況を反映していると考えられる。

図から指標のデータサイズ依存性の問題はほぼ解消した（相関係数+0.07）ことが判るが、ここから直ちに本指標が機関のパフォーマンスを反映していると結論することはできない。対象機関は総合大学、工業大学、特定研究領域に注力する独立行政法人など研究領域でも非常に幅広いものとなっている。（Ahlgren2015）をはじめとして、分野による正規化を目標とする評価指標がこれまで開発されてきた。

そこで、研究領域の相違が h-dimension に及ぼす影響をまず計測する。手法は、研究機関ごとの分野構成比に基づいて当該分野から乱択することで構成される論文集合の h-dimension をコントロールとし、これと測定値を比較する事で行った。

図 4 の横軸がコントロール、縦が測定値である。弱い正の相関は見られるが、残差は 0.88, Kendall 順位相関は 0.2 であり、分野の選択で説明できる測定値の相違は 2 割に満たない。

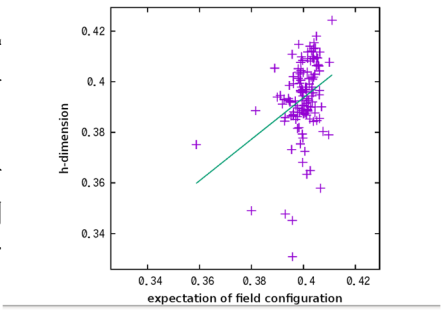

Figure 4: 分野構成によるコントロールと測定値

そこで、研究領域ごとの h-dimension を測定すると、ほぼ 0.39 前後に集中しており、そもそも領域ごとの相違が非常に小さい事が判った。詳細については（Fujita2022）に譲るが、これは、h-dimension が順位統計としての側面をもち、かつ、その測定対象となる論文の部分集合が、比較的多数の被引用数を獲得したものに限定されること、更にその領域では被引用数の分布の分野による相違が小さいためである。

スケール依存性がなく、分野の相違も小さいフラクタル次元によって、何が測定されているのかは必ずしも自明ではない。図 5 は同じ論文数であって異なる h-dimension を持つよう人工的に構成されたランダムな引用構造を可視化したものであり、リンク方向のレイアウトは図と同様、被引用ノードが下にくるよう布置されている。色彩はグラフ構造に頂点が追加された順、すなわち論文出版時刻によって赤、黄、緑を経て青とし、円のサイズは被引用数を反映している。左の低い h-dimension を示す引用構造では、互いに疎に結合した複数の部分ネットワークに分断されており、それらの境界を跨いだ知識生産に支障が生じている。これに対して高い h-dimension の論文集合では共通の知識基盤に相当するものが存在し、これが比較的大規模な h-index 定義集合の存在につ

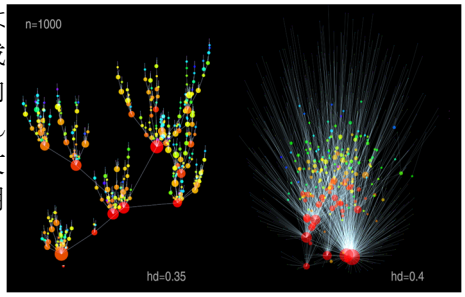

Figure 5: h-dimension は異なるが頂点数が同一の構造

ながっている。

4. 結論

本研究ではスケール不変かつ分野の相違にほぼ影響されない h-index のフラクタル次元として定義される機関評価指標 h-dimension を開発、提案した。研究活動の多面性や人類の適応力を考えると、単一の数値によって研究活動評価の全てが可能になることはなく、また、この指標の前提とする評価基準を迂回する戦略が考案されると思われる。3 章末尾で h-dimension が測定しようとする引用構造の意味内容を考察したが、ここでは別の角度から、すなわちできるだけ少ない労力で h-dimension を向上させる敵対的戦略を検討することで当該指標の性質を考察し、結論にかえたい。

h-dimension を向上させるには定義の式の分母を小さく保ったまま分子を増大する必要がある。したがって、h-index 定義集合 H の被引用数が、論文集合の被引用数と一致するときに h-dimension は理論上の最大値 0.5 となる。この引用構造の性質を hd-最適と呼ぶことにすると、h-index に寄与しない被引用は hd-最適の阻害要因である。すなわち、一旦集合 H に算入された論文が、それ以上引用されることは hd-最適から遠ざかることになり、H に算入されていない論文が引用されることもまた hd-最適とは反する。これを徹底することで h-dimension を向上させることが可能になる。

これは、引用された文献が他の研究者の目にとまりやすくなり、再び引用されやすくなる、という一般的かつ強力な仕組みに制限を加えることを意味する。以上の戦略は非常に素朴で単純なものであるが、これを実践するのは自己相似的な引用構造という世界に止まる限り容易ではないように思われる。

今後は国際比較などを通して h-dimension の性質と可能性を検討していきたいと考えている。

参考文献

Ahlgren P, Sjögårde P (2015) Formal definitions of field normalized citation indicators and their implementation at kth royal institute of technology http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-267026

Fujita Y., Usami N. (2022) “Fractal dimension analogous scale-invariant derivative of Hirsch’s index”, Applied Netowork Science, https://doi.org/10.1007/s41109-021-00443-x

OECD “The knowledge-based economy” 1996

Hirsch JE (2005) An index to quantify an individual’s scientific research output. PNAS. https://doi.org/10.1073/pnas.050765510

Koizumi A (2018) Kenkyuuryoku-no-hakarikata (in Japanese). Gakujutsu-no-doukou (in Japanese) 23(12):64–67. https://doi.org/10.5363/tits.23.12_64

Mandelbrot B (1977) Fractals: form, chance and dimension. W H Freeman and Co

Pratelli L, Baccini A, Barabesi L, Marcheselli M (2012) Statistical analysis of the Hirsch index. Scand J Stat 39(4):681–694