WIAS Discussion Paper No.2012-004

共引用クラスタリングによる研究分野の動的把握に向けた試論

An attempt of analyzing formation dynamics of academic knowledge using co-citation clustering

October 11, 2012

七丈　直弘（早稲田大学高等研究所）

Naohiro SHICHIJO
Waseda Institute for Advanced Study, Waseda University, Tokyo, Japan

1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan
Tel: 03-5286-2460 ; Fax: 03-5286-2470

共引用クラスタリングによる研究分野の動的把握に向けた試論

An attempt of analyzing formation dynamics of academic knowledge using co-citation clustering

七丈 直弘[1*]
Naohiro SHICHIJO[1*]

1 早稲田大学高等研究所
Waseda University、 Waseda Institute for Advanced Study
〒169-8050 東京都新宿区西早稲田1-6-1
E-mail: shichi@acm.org

研究開発競争が加速するなか、有望な研究領域を早期に把握することの重要性が増している。だが、有望な研究テーマは成熟した学術領域ではなく、いまだ学術分野として確立していない萌芽的かつ融合的な領域に多く存在していると考えられる。このような領域を把握する手法を構築するために、本論文では学術分野をより高い粒度で把握し、研究者の論文ポートフォリオの学際性を評価するための手法を提案する。

学術分野把握のため、共引用分析を用いて論文間の近接性を求め、これを関係性として論文をノードとしたネットワークを構成した。さらに、Newman法によりドメイン分割を行うことで、従来論文誌レベルやキーワードレベルでしか把握できなかった研究分野の微細構造の評価が可能となった。この構造を基に、研究分野の発展の時系列把握、研究者の論文ポートフォリオの多様性の評価をした。

While R & D competition accelerates、 the importance of understanding a promising research area at an early stage is increasing. But、 most of promising research subjects are considered to reside in the interdisciplinary fusion area、 that are not yet established as a distinct discipline. To pursue the technique of grasping such promising interdisciplinary research area、 in this paper、 I propose the technique for grasping the scientific field with a higher granularity、 and the technique for evaluating the interdisciplinarity of a researcher's academic portfolio.

I used bibliometric method "co-citation" to evaluate proximity of content of academic writings、 and constructed the network structure using paper as the node and similarity relationship as edge. By applying Newman algorithm to partition those networks、 detailed research area is defined. Using this detailed research area distinction、 evaluation of longitudinal analysis of the diversity and researcher's interdisciplinarity from their research portfolio are achieved.

キーワード: 研究評価、リサーチフロント、 共引用分析、 Newmanアルゴリズム、 学際研究
Keywords: Research evaluation、 Research front、 Co-citation、 Newman algorithm、 Interdisciplinary Research (IDR)

1. はじめに

経済発展の結果、先進国のみならず新興国においても、製造業から知識産業に代表されるようなサービス業へと産業重点領域のシフトが起きつつある。新興国での科学技術研究の進展は著しく、特に、中国では公刊される論文数が指数関数的に増加している[1]。このように世界規模で加速する研究開発競争を背景として、新しい研究領域を自ら創出し、学術的発見を土台として、世界に先駆けてイノベーションを普及させ市場化を通じた収益化を社会全体で実現していくことが、課題となっている。知識生産活動の加速に向け、研究分野発展の動的機構の解明を通じ、新知識発見の可能性が高い領域を見出す方法論が求められている。

de Solla Price[2]は、その萌芽的な計量書誌分析の研究の中で、公刊される論文の中のごく少数が多数の論文から引用されるという現象を発見し、多くの論文から同時に引用（これを「共引用」という）される少数の論文をリサーチフロントと呼んだ。この手法は萌芽的研究領域を同定するために利用できる可能性があることから、トムソン社(Thomson Scientific)は商業サービスとして提供してきた。そこでは、同社の書誌データベース ESI(Essential Science Indicators)に含まれる論文を用い、高被引用論文間の共引用関係がダイアグラムとして可視化され、高被引用論文の中でも特徴的な位置を持つものが Science Watch 誌の中で紹介されてきた。日本国内でもリサーチフロントと同様の手法を用いて俯瞰的に研究領域を先導する特徴的な研究成果をピックアップしようという取り組みが行われている[3]。

リサーチフロントのような手法が評価される背景には、科学技術が成熟した現在、ブレークスルーとなるような研究は旧来の学術領域の境界を超え、分野横断的な取り組みから生まれる傾向が従来よりも高まってきたという認識がある[4]。分野横断的な研究は、広い科学分野を俯瞰しなければ認識することが出来ない。本論文では、分野横断的な取り組み・萌芽研究分野把握の新手法として、計量書誌的手法による研究分野同定手法を提案する。また、それを基に研究分野の時間変化のダイナミクスをネットワーク構造として可視化する手法を提案する。

2. 学際性の指標化

分野横断的な取り組みが集積し、新分野として確立されれば、新しい論文誌の創刊につながる場合もあり、そのような場合は学際研究の存在を比較的認識しやすい。実際、公刊されている学術論文誌の数は急速に増加している。Ulrich’s Periodical Directory では、1970 年代には 4,634 誌、1980 年代には 5,673 誌、1990 年代には 6,245 誌、2000 年代には 5,264 誌が新規の peer-review による学術誌として登録されるに至っており、この中には分野横断的な取り組みが学術分野に昇華した事例が多く含まれている可能性がある[5]。

だが、一般的には研究の学際性を測ることは難しい。特定の研究が学際的であるかを判断するための基準としては、それが掲載された論文誌に付与されたカテゴリの組み合わせをみ

るという方法がよく用いられてきた[6, 7]。しかし、この手法では、論文誌単位での学際性しか測ることができず、学際性の尺度としては極めて精度が低い。これを補強するため、当該論文が引用している論文が掲載されている論文誌のカテゴリの組み合わせや、当該論文を引用している論文が掲載されている論文誌のカテゴリの組み合わせを考慮することによって発展させ、論文単位や研究者単位で学際性を定量化する手法も提案されている[7、 8]。このような指標は計算が比較的容易であるため、大集団に対して適用し、学術研究のダイナミクスのマクロ的把握ができるが、学際性の判断が既存のカテゴリの区分（研究領域によってカテゴリの細分化に大きな差がある）に依存しているということ、さらに既存のカテゴリ区分がそもそも粗いという根本的な問題を抱えており、既存の学術領域から新たな学術分野が派生していくような、細分化された研究領域ダイナミクスを把握するには向いていない。

これらに対し、Noyons ら[5]は論文誌の分類に関する先験的な情報を用いずに、キーワードや論文分類コード(Journal Classification Codes)の共出現関係のみから学術領域の構造とその時間的変化を導くことで、非経験論的に学術構造を分析できる可能性を示した。論文分類コードの代表例として、JEL コード（経済学)、PACS コード（物理学）などがある。だが、論文分類コードそのものが既存学術領域を想定したものであり、複数分野を跨いだ研究の分類には向かない。また、萌芽的研究ではキーワード選択に揺れが存在するため、手作業によるキーワードの標準化を経なければ正確な評価が難しい。

より安定的に研究の学際性を示すには、論文誌のカテゴリ分類や、キーワード・論文分類コードに拠らない研究分野同定の手法が求められる。本研究では、共引用関係によって得られた論文間の関係性ネットワークをクラスタリング手法によってドメイン分割し、細分化された学術領域として捉える手法を提案する。このドメインに対して、Small が高被引用論文に対して適用した手法[9]を参考にしながら、研究分野となるドメインの時系列的変化についても可視化することで、特定学術領域内におけるミクロな領域間での知識の相互関係を明らかにする。また、研究者ごとに、研究論文のドメイン間分布（論文ポートフォリオ）は大きくことなっていることから、論文ポートフォリオの集中指数と研究パフォーマンスの関係についても考察を試みる。

3. データ

本研究では、論文書誌情報として Thomson Scientific 社が作成している学術論文書誌データベースである SCI-EXPANDED を利用した。このデータベースの中から 1970 年から調査時点である 2004 年までに収録されたデータを分析の対象とした。

まず、分析対象となる学術領域を規定するために、以下のようにして論文サンプルの抽出を行った。最初に全文検索により「photocatal*」（語尾の*はワイルドカードであり、光触媒に関連した語である「photocatalyst」や「photocatalysis」等に適合する）を文章中に含む論文を抽出した。その結果抽出された論文の集合を S と定義する。次に、引用情報を用いて S に含まれる論文を引用している論文を抽出し、この集合を C と定義する。S に含まれる論文

は 6、992 本であり、C に含まれる論文は 25、651 本であった。後で説明されるように、C の論文が持つ引用情報によって S がクラスタリングされることになる。

また、研究パフォーマンスの比較を行うために、研究者ごとの論文数の集計を行った。S に属する論文を有する研究者は 6、763 名のうち、1 回以上引用を受けている論文を 5 本以上有する研究者を分析対象とした。このような条件を課すことにより、学生などのように教育の派生的成果として論文を出版した者の除去が可能となり 684 名の研究者が抽出された。このうち 9 名は、他の研究者との共著関係を持たなかったため、研究者間の連携による効果を評価することができないため分析対象外とした。最終的に、1 件以上の共著関係を有し、5 報以上を出版する 635 名の研究者が以下で行われる分析の対象となった。

4. 手法・結果

分野の時系列的な発展を知るために S を対象としたクラスタリングを行う。なお、クラスターの時間的発展を知るために、直近の 5 年間に出版された論文を対象としたスライディングウィンドウを設定し、クラスタリングを行うこととした。具体的には、y 年のクラスタリング対象は、y-4 年から y 年に引用された S に含まれる論文の中で、その後 1 回以上他の論文から引用を受けたものとなる。各年のクラスタリング対象となった論文数の変化を図 1に示す。

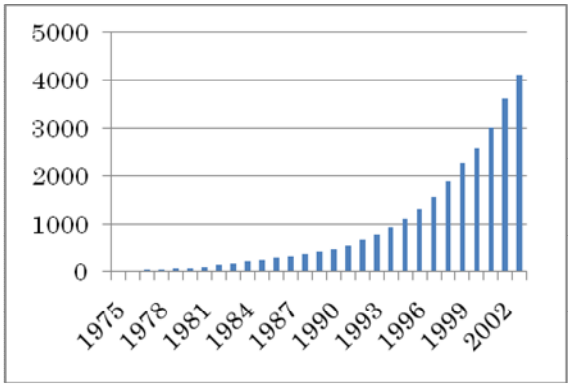

図 1クラスタリング対象となった論文数の年次変化

次にクラスタリングの手法について述べる。

書誌結合分析や語分析で扱うデータは、論文の参考文献リストや本文といった、論文が出版された時点で確定されるデータであり、どの時点で分析を行っても分析結果は変わらない。一方、共引用分析では、新しく引用を受けるたびに分析結果がダイナミックに変化する。この性質は、知識構造の進化プロセスを観察するのに有効である。なお、論文が出版されてから引用されるまでのラグの存在により、共引用分析では、出版年が若い論文間の関係性を過小評価する傾向があることに注意すべきである。

また、共引用分析による分析結果は、当該分野の知識構造に対する研究者の分析時点での認識といえる。研究者の研究戦略を分析するためには、研究者の意思決定には当時の認識が影響を与えているので、過去にさかのぼって共引用分析を行って、当該分野の知識構造に対する研究者の当時の認識を捉える必要がある。このため、全期間を通じたクラスタリングを行うのではなく、過去に溯ってウィンドウを設定した上でクラスタリングを行うことに意味がある。

具体的なクラスタリングは、以下に述べるように重みつき Newman 法によって行った。

まず、y 年の論文ネットワークの重み付き隣接行列 $R(y)$ は、ij 成分を「論文 i と論文 j が y-4 年から y 年までに同一論文によって引用された件数」とする行列として定義する。ここで、$R(y)$ の対角成分は便宜上すべて 0 とした。

重み付き Newman 法[10]はクラスタリング問題を以下によって定義される量 Q の最小化問題ととらえるものである。

$$Q = \sum_i \left(e_{ii} - a_i^2 \right)$$

ここで e_{ij} はクラスターi とクラスターj を結ぶエッジの重みの和を全エッジの重みの和で割った値であり、$a_i = \sum_j e_{ij}$ はクラスターi に属するノードと結びついたエッジの重みの総和となる。

重み付き Newman 法によりクラスタリングを行うことで得られた結果を図 2に示す。

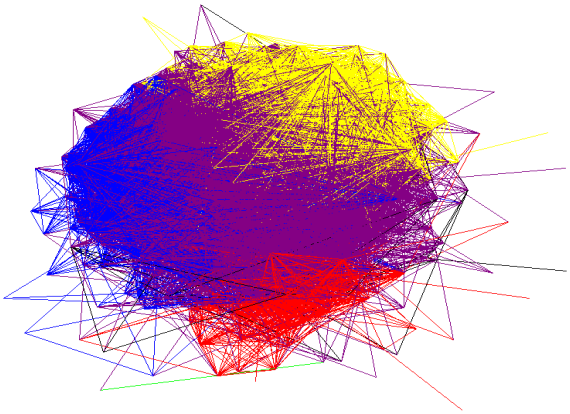

図 2 2003 年の共引用ネットワークのクラスタリング結果（論文は点で示され、論文間を結ぶ辺の色は所属するクラスターを意味する）

次に、クラスターの時系列変化を図 3に示す。図において、丸は各年の各クラスターを示し、その大きさはクラスターに属する論文数を示す。ここでは視認性確保のため、各年、所属論文数の上位 10 クラスターのみを表示した。また、丸と丸を結ぶ線は、各クラスターに属

する論文が次の年にどのクラスターに属するかを示しており、線の太さはその論文数を示している。これは Small による手法[11]を参考にした。

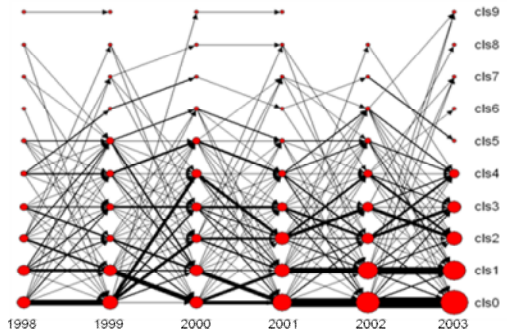

図 3 クラスター間の継承関係の時系列的変化

クラスターの時系列変化の可視化像の観察により、クラスターは複雑に融合したり分裂したりしていることが分かった。そこで、続く分析では、知識の融合と分裂を表すマクロ的指標としてハーフィンダール指数(HHI)を導入し、この値の時系列変化を追う。クラスター数を k、クラスターi に属する論文数を a_i とした場合のハーフィンダール指数(HHI)は以下の式で与えられる。

$$HHI = \frac{\sum_{i=1}^{k} a_i^2}{\left(\sum_{i=1}^{k} a_i\right)^2}$$

この値が大きいほど知識構造が集中している。つまり、少数の大きなクラスターで知識構造が形成されている。逆に、この値が小さいほど知識構造が分散している。つまり、多数の小さなクラスターで知識構造が形成されている。HHI の値を年ごとに算出した結果、研究分野ごとの論文数の HHI の時間変化は図 4のようになった。

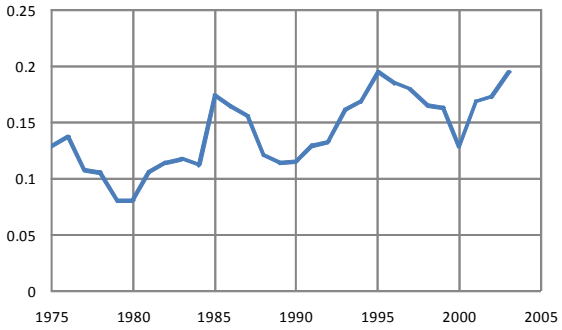

図 4　知識の集中度の時間変化

図 4から、ハーフィンダール指数は下降と上昇を繰り返していることが認識される。この現象は、知識の細分化や融合は一過性の行為ではなく、時期によりアジェンダを変えながら、繰り返し起きている可能性を示唆する。

上記の手法によって可能となる学術領域の動的把握は、研究領域全体の集中・分散だけでなく、個々の研究者の研究開発戦略、連携戦略の分析に対して新たな視点をもたらす。概念実証の実例として、研究者の論文生産性をその研究者が有する研究ポートフォリオの多様性、自らと連携相手の間のポートフォリオの近接性によって説明してみよう。ここで研究ポートフォリオとは、学術領域の細領域の各々に対する個々の研究者の貢献をベクトルとしてとらえたものである。

まず、分析の対象とされた 635 名の研究者に対して、年（1975 年から 2003 年）ごとに論文ポートフォリオを作成する。

研究者 i の y 年の論文ポートフォリオは、前章で作成した y 年の各論文クラスターに含まれる論文のうち、著者欄に i が含まれる論文の本数を並べたベクトルで表す。例えば、y 年のクラスター数が 3 で、研究者 i の論文が 3 つのクラスターにそれぞれ、3、2、0 本ずつ含まれているとした場合、研究者 i の y 年の論文ポートフォリオベクトルは $\mathbf{P}_i(y)=(3、2、0)$ となる。

また、各研究者の研究内容の分散度を表現する指標として、１から研究分野ごとの論文シェアの自乗和を減じた値を用いる。論文ポートフォリオベクトルの成分を $P_{i、k}$ とすると、研究者 i の y 年の研究内容の分散度 $HHI_i(y)$ は、

$$HHI_i(y)=1-\sum_k \frac{\vec{P}_{ik}^2(y)}{\vec{P}_i^2(y)}$$

と表わすことができる。この値が大きいほど、研究内容が幅広いことを示す。

研究内容の遠い研究者と共同研究を行うことは、新たな知識の獲得に有効であり、研究内容の近い研究者と共同研究を行うことは、知識の深化に有効であると考えられる。そこで、共著者間の研究内容の遠近が共著以降の研究パフォーマンスに与える影響をみるため、共著者間の論文ポートフォリオの類似度を算出する。y 年に研究者 i と j とが共著論文を出した場合の共著者間の論文ポートフォリオの類似度 $S_{ij}(y)$ は、y－1 年時の 2 者の論文ポートフォリオの余弦で表す。この値が大きいほど、2 者の研究内容は近いことを示す。

$$S_{ij}(y)=\frac{\vec{P}_i(y-1)\cdot\vec{P}_j(y-1)}{\left|\vec{P}_i(y-1)\right|\cdot\left|\vec{P}_j(y-1)\right|}$$

また、研究者によっては 1 年間に複数の研究者と共著論文を出す者もいる。その場合、共著者の中で最も研究内容の遠い者との論文ポートフォリオの類似度 $Similarity_i(y)$ を説明変数として用いる。

$$Similarity_i(y)=\min_j S_{ij}(y)$$

類似度が小さいほど、離れた研究者と共同研究を行っているといえる。表 1に被説明変数を y+1 年における論文数とし、説明変数として過去 5 年間における論文数、総被引用数、共著

論文数、共著者数、論文ポートフォリオの HHI、共著者との研究ポートフォリオの相違度を用い、パネル回帰分析を行った結果を示す。

表 1 パネル回帰分析による推計結果

被説明変数: 論文数	
説明変数:	
論文数(-1)	0.186***
	(0.022)
総被引用数	-0.001***
	(0.000)
共著論文数	-0.047*
	(0.025)
共著研究者数	0.031***
	(0.008)
HHI	-0.349**
	(0.159)
Similarity	-0.170**
	(0.085)
定数項	0.467***
	(0.093)
χ^2	1066.61
N	4012

パネル回帰分析の結果、自らの研究領域が幅広いこと自体は論文数に負の影響を与え、研究内容が異なる研究者と共同研究を行うことが論文数に正の影響を与えることが判明した。この結果は[8]と整合的である。

5. 終わりに

本研究によって、特定研究領域におけるドメインのミクロ構造を共引用分析によって合理的に行えることが判明した。また、Small[9]の手法を適用することで、ドメイン間のダイナミクスを可視化することができた。また、得られたドメイン構造から、研究者の研究ポートフォリオを定義し、研究パフォーマンスとの比較をパネル分析によって行った。手法上の限界として、今回の分析では光触媒という限定された領域内での分析であるため、観測の対象外となった関連領域との間の相互作用（外部性）については考慮していない。また、光触媒という分野そのものが、研究の進展の結果確立したものであって、対象とした論文の中の比較的古いものは、外部性の影響を大きく受けていることが予想される。この問題を克服するためには、query expansion などの手法で、できるだけ関連領域を広く含めるように工夫する必要がある。また、リサーチフロントが対象とするようなマクロレベルでのダイナミクスと、本論文が対象としたミクロレベルでのダイナミクスとの連関についても今後考察したい。

謝　辞

本研究は、科学技術研究費補助金若手(B)　#21730290 による支援の成果が含まれる。また、加毛 誠氏、馬場靖憲教授から研究内容についてアドバイスを受けた。

引用文献

1. Zhou, P. and L. Leydesdorff, *The emergence of China as a leading nation in science.* Research Policy, 2006. **35**(1): p. 83-104.
2. Price, D.J.D., *NETWORKS OF SCIENTIFIC PAPERS.* Science, 1965. **149**(3683): p. 510-515
3. 阪彩香, 伊神正貫, 桑原輝隆, *サイエンスマップ 2006―論文データベース分析（2001 年から 2006 年）による注目される研究領域の動向調査―*, 2008, 科学技術政策研究所.
4. Metzger, N. and R.N. Zare, *Science policy - Interdisciplinary research: From belief to reality.* Science, 1999. **283**(5402): p. 642-643.
5. Noyons, E.C.M. and A.F.J. van Raan, *Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research.* Journal of the american society for information science, 1998. **49**: p. 68-81.
6. van Leeuwen, T. and R. Tijssen, *Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows.* Research Evaluation, 2000. **9**: p. 183-187.
7. Rinia, E., et al., *Measuring knowledge transfer between fields of science.* Scientometrics, 2002. **54**(3): p. 347-362.
8. Qin, J., F.W. Lancaster, and B. Allen, *Types and levels of collaboration in interdisciplinary research in the sciences.* Journal Of The American Society For Information Science, 1997. **48**(10): p. 893-916.
9. Small, H., *Tracking and predicting growth areas in science.* Scientometrics, 2006. **68**(3): p. 595-610.
10. 榊剛史, 松尾豊, 石塚満. *制約付きクラスタリングを用いた論文分類.* 人工知能学会全国大会*(第 20 回)論文集*. 2006.
11. Small, H., *Citation Structure of an Emerging Research Area: Organic Thin Film.* Proceedings of ISSI, 2007: p. 718-725.