文献ファミリーの同定と特許からの引用評価の試み

Feasibility study of identifying Paper-families and their citations from patents

国立研究開発法人科学技術振興機構
山下　泰弘

独立行政法人産業技術総合研究所技術と社会研究センター、山形大学評価分析室等を経て現職。計量書誌学分析に従事。

yasuhiro.yamashita@jst.go.jp

国立研究開発法人科学技術振興機構
吉田　秀紀

株式会社東芝研究開発センターを経て、現在 JST でファンディング業務や分析業務に従事。

エルゼビア・ジャパン株式会社
高坂　香那

ベンチャー企業、コンサルティング会社を経て、現在エルゼビアでデータサイエンス業務に従事。

1　はじめに

一つの研究課題が原著論文[1]の形に結実するまでには、初期段階の発表メディアとして、学会・カンファレンス（国際会議など）が多くの分野で活用されてきた。一方、未審査段階の論文を早期公開するプレプリントが近年では広がりを見せている（小柴 , 林 , 伊藤 ,2020）。特に、新型コロナウイルス感染症のパンデミック下において、感染症研究の成果の早期公開が求められたため、学術論文の審査期間が大幅に短縮されるとともに、プレプリントサーバの役割についても改めて注目を集めることとなった。

通常、科学計量学における研究成果の集計単位とされるのは原著論文であるが、上記のような複数の同一内容の出版物が存在することを考慮するならば、それだけでは不十分である。学界の注目を集めている研究課題であれば、前段階のプロシーディングスやプレプリントの段階で引用されるので、そのインパクトを正確に把握するには、原著論文以前の段階も含めて評価を行う必要がある[2]。

そこで、本稿では、「同一研究者ないし研究グループ」による「同一研究内容の文献群」を分析単位とみなすことを提案する。いわば、特許におけるパテントファミリーの論文版であるので、パテントファミリーにちなんで「文

1　本稿では、レビュー論文も含めて原著論文として取り扱う。

2　arXiv に収録されたプレプリントを対象とした分析において、数学と計算機科学では特に大部分の引用にプレプリントが関与していることが示されている（林 ,2021;MEXT-NISTEP プレプリント調査・検討チーム ,2020）

献ファミリー」と呼称する。

また、学術分野によって中心的な出版メディアが異なることが知られている。例えば、人工知能分野では、権威ある国際会議での発表が重視され、原著論文のウェイトはそれほど高くない。このような分野であっても、研究課題の最小単位ともいえる「文献ファミリー」を集計単位とすることにより、他分野と横並びでの分析が可能になると思われる[3]。

本稿は、文献ファミリーについての取組の第一報であるが、本誌の趣旨に則り、特に特許文献からの引用の視点から報告する。

2 文献ファミリー同定の試み

上述では、プレプリントと原著論文、プロシーディングスと原著論文間のつながりについて言及したが、それ以外にも想定しない組合せが出現し得る。そこで、文献ファミリーの同定にあたっては、初期段階では文献タイプの限定は行わないこととした。

文献ファミリーという概念は、管見の限りでは先行研究が見当たらない。そのため、その同定についても試行錯誤で進めている。プレプリントと論文のマッチングについては、タイトル、第一著者、出版日に基づく手法が提案されているが（Cabanac, Oikonomidi, & Boutron, 2021）、同じ手法・基準が文献ファミリーを形成するすべての文献タイプに対して適用可能かは定かではない。実際、いくつかの事例を見る限りでは、2～3語程度の非常に短いタイトルの場合は、同一の著者が同タイトルで複数文献を発表しているケースが散見された。このような場合は、むしろ別文献であるケースが多いと思われる。

これまでにとったプロセスは下記の通りである。

（1）Scopus において収録文献のタイトルと著者の類似度に基づき文献ファミリーの候補を抽出

初期サンプルとして2017年以降に出版された3万件超の高被引用論文を使用し、エルゼビア社の文献データベースScopusにおいて、サンプル論文と類似した学術文献の抽出を行った。この段階では、抽出される学術文献の出版年については制限しなかったため、長期間にわたっている。

学術文献著者の同定には、Scopus収録文献の全著者に付与されている研究者IDを使用した。「同一著者」による著作物の識別には通常は困難を伴うが、このIDを活用することにより、機械的に識別することができた。

（2）上記文献群を Dimensions 収録文献と突合

現状では、Scopusにおいては、プレプリントに被引用数等の評価指標が付与されていない。そこで、文献ファミリーの評価には、JSTエビデンス分析室で契約しているDigital Science & Research Solutions社のDimensions Analytics APIを活用することとした。原則としてDOIをキーとしてScopusとDimensionsの学術文献を突合したが、一部の学術文献（例えば、arXivのプレプリントなど）にはDOIが付与されていない[4]。DOIが付与されていない文献については、タイトルと第一著者に基づいて検索し、必要に応じて目視確認を行った。

（3）Dimensions による類似度指標の付与

（1）で同定されたファミリー集合はかなり広めにとられているため、さらなるフィルタリングを行うため、Dimensions Analytics APIにより、文献間のアブストラクトとリファレンスの類似度指標を付与した。現在は、指標値の閾値設定を試みている段階である。

以下では、得られた文献ファミリー候補データの中から、文献ファミリーである確証が得られたいくつかの例について、特許文献との引用関係を分析する。

3 文献ファミリーの特許からの引用

本章では、代表的な3つの文献ファミリーを取り上げ、その内訳を分析する。なお、前章のプロセス（1）においてシーズとして用意したサンプル論文は2017年以降のものであるが、事例3では、目視によるフィルタリングでシーズ論文自体がファミリー候補から外されたため、文献ファミリーを構成する学術文献が2016年以前のもののみとなっている。

3　無論、文献以外を主たる成果メディアとする分野には適用できないし、分野によって平均的な文献ファミリーの規模や性質が異なる点には考慮する必要がある。

4　arXivは2022年1月に収録文献へのDOIの付与を開始し、同年2月に全文献への付与を完了しているが（arXiv.org 2022）、2022年8月現在Scopus、DimensionsともarXivの収録文献にDOIを付与していない。

事例 1　プロシーディングスが多く引用される文献ファミリー（文献ファミリー 1）

カリフォルニア大学バークレー校の Darrell らは、完全畳み込みネットワークの提案に関する研究を 2015 年に IEEE Conference on Computer Vision and Pattern Recognition（CVPR）において発表し（Long, Shelhamer, & Darrell, 2015）、それを元に 2017 年に IEEE Transactions on Pattern Analysis and Machine Intelligence 誌で原著論文を出版した（Shelhamer, Long, & Darrell, 2017）。これらの文献は、プロシーディングスが 68 公報（57 パテントファミリー）、原著論文が 22 公報（20 パテントファミリー）から引用されている（表 1）。6 公報（5 パテントファミリー）がプロシーディングスと原著論文の両方を引用しているため、文献ファミリー全体では、84 公報（72 パテントファミリー）から引用を得ていることになる。プロシーディングスがジャーナル論文よりも高いインパクトを示している点は、人工知能分野の典型とも言える。

この文献ファミリーの特徴は、プロシーディングス、原著論文とも並行して特許から引用され続けている点である。特許からの引用を時系列で見ると、原著論文発表前は当然の事ながらプロシーディングスのみが引用されているが、2017 年の原著論文発表後もプロシーディングスが多く引用され、原著論文は後塵を拝する形となっている。とは言え、原著論文への引用を加えることにより、プロシーディングス単体を見た場合と比較して 25% 程度増加することになる。

表 1　文献ファミリー 1 の特許からの引用状況（出願日順）

引用元特許			引用先文献	
公報番号	パテント ファミリー ID	出願日	CVPR（プロシーディングス）	IEEE Trans. Pattern Anal. Mach. Intell.（原著論文）
US-9786036-B2	55967426	18-Sep-15	○	
EP-3391290-A4	59055596	16-Dec-15	○	
EP-3493149-A1	56688725	19-Feb-16	○	
EP-3166075-A1	56360184	29-Jun-16	○	
EP-3144851-A1	56943353	14-Sep-16	○	
EP-3391284-A4	59057840	16-Dec-16	○	
EP-3355270-A1	57914817	27-Jan-17	○	
EP-3217332-A1	58277170	09-Mar-17	○	
EP-3465174-A4	60412548	24-May-17	○	
US-11010610-B2	64660880	13-Jun-17		○
EP-3506200-A4	61245431	12-Jul-17	○	
US-10713816-B2	64999555	14-Jul-17		○
EP-3438929-A1	59702525	04-Aug-17	○	
WO-2018091486-A1	60452621	15-Nov-17	○	
EP-3330898-A1	60543392	29-Nov-17	○	
US-10936938-B2	67059739	28-Dec-17		○
US-10861185-B2	62783250	04-Jan-18		○
US-10957068-B2	62781884	04-Jan-18		○
EP-3511861-A1	60972092	12-Jan-18	○	
WO-2018138104-A1	57914817	24-Jan-18	○	
JP-2020509466-A	63037815	05-Feb-18	○	
EP-3410344-A1	61622454	12-Mar-18	○	
US-10140544-B1	64315465	02-Apr-18		○
JP-2020516427-A	58744797	11-Apr-18	○	

US-10706499-B2	68982017	21-Jun-18		○
WO-2020003434-A1	68986315	28-Jun-18	○	
EP-3627379-A1	63683080	24-Sep-18	○	
EP-3754593-A4	67620971	28-Sep-18	○	
US-11320508-B2	60269642	22-Oct-18	○	
JP-2021502627-A	66247744	24-Oct-18	○	
JP-6997309-B2	66247744	24-Oct-18	○	
US-11216988-B2	66247744	24-Oct-18	○	
EP-3480740-A1	64051355	26-Oct-18	○	
US-10956793-B1	64604786	15-Nov-18	○	
US-10910094-B2	60480180	20-Nov-18	○	
KR-20200063349-A	71089290	22-Nov-18		○
US-10977923-B2	66665058	30-Nov-18	○	
WO-2020129176-A1	71101149	19-Dec-18	○	
US-10740896-B2	67057727	21-Dec-18		○
WO-2019152144-A1	65269066	08-Jan-19	○	
WO-2019138074-A1	60972092	11-Jan-19	○	
WO-2019158442-A1	65409074	08-Feb-19	○	
WO-2019179889-A1	65904382	15-Mar-19	○	
US-11337432-B1	81656264	03-Apr-19	○	
US-11170200-B2	62242750	31-May-19	○	
US-10657379-B2	68840014	19-Jun-19		○
FR-3098001-A1	68072758	27-Jun-19	○	
EP-3599607-A1	67437899	17-Jul-19	○	
EP-3608902-A1	67438937	25-Jul-19	○	
EP-3608903-A1	67438939	25-Jul-19	○	
WO-2020105225-A1	70773995	01-Aug-19	○	
WO-2020064253-A1	63683080	28-Aug-19	○	
WO-2020074035-A1	68242223	12-Sep-19	○	
DE-102019218177-A1	73544161	25-Nov-19	○	
DE-102019218186-A1	73544168	25-Nov-19	○	
DE-102019218187-A1	73544169	25-Nov-19	○	
DE-102019218192-A1	75784300	25-Nov-19	○	
WO-2020121996-A1	71076887	09-Dec-19	○	
WO-2020137745-A1	71131569	18-Dec-19	○	
US-10991097-B2	71123021	31-Dec-19	○	○
EP-3951710-A4	72668087	16-Jan-20		○
US-11037325-B2	62781884	03-Feb-20		○
EP-3866113-A1	69645893	17-Feb-20	○	
US-10957041-B2	71944990	25-Mar-20	○	○
EP-3901898-A1	70464874	24-Apr-20	○	
FR-3111268-A1	72709498	10-Jun-20	○	
WO-2020254448-A1	71111416	17-Jun-20		○
WO-2020257756-A1	71528078	22-Jun-20		○
US-11341722-B2	67623182	07-Jul-20	○	

US-11308623-B2	74102434	09-Jul-20		○
US-11244450-B2	74646342	18-Aug-20	○	
WO-2021105006-A1	73544161	20-Nov-20	○	
WO-2021105014-A1	73544168	20-Nov-20	○	
WO-2021105015-A1	73544169	20-Nov-20	○	
WO-2021105017-A1	73544174	20-Nov-20	○	
US-11348239-B2	76091811	31-Dec-20	○	○
US-11348240-B2	76091812	31-Dec-20	○	○
US-11348661-B2	76091684	31-Dec-20	○	○
US-11263748-B2	71944990	05-Feb-21	○	○
WO-2021178605-A1	75223460	04-Mar-21	○	
WO-2021214032-A1	70464874	20-Apr-21	○	
WO-2021250091-A1	72709498	09-Jun-21	○	
WO-2022046725-A1	80269654	24-Aug-21		○
WO-2022090483-A1	73776606	29-Oct-21	○	

事例２　特許から引用されたプレプリントを含む文献ファミリー（文献ファミリー２）

プレプリントは、Scopus には 2017 年以降のものが登録されており、通常原著論文はプレプリントの後で出版されるため、プレプリントと原著論文からなる文献ファミリーは必然的に新しい原著論文を含むことになる。このグループの文献ファミリーとしては、Okba らによる新型コロナウイルス感染症の血清学的アッセイ開発に関する論文（Okba et al., 2020a, 2020b）が挙げられる。この文献ファミリーは、プレプリントが 26 公報（22 パテントファミリー）、原著論文が 7 公報（5 パテントファミリー）の引用を特許から得ている。2 公報（2 パテントファミリー）がプレプリントと原著論文の両者を引用しているため、文献ファミリー全体では 31 公報（25 パテントファミリー）からの引用を得ている（表 2）。この論文は、2020 年 3 月に medRxiv にプレプリントが投稿され、同年 7 月に Emerging Infectious Disease 誌で出版されている。原著論文が出版されるまでに出願された特許については前の例と同様プレプリントを引用しているが、原著論文がオープンアクセスであるにも関わらず、原著論文の出版以降もプレプリントの方が多く引用されていることがわかる。少なくとも、このケースにおいては、プレプリントは、原著論文が出版されるまでの繋ぎの存在ではなく、独立した出版物としての地位を得ていると言えよう。

表２　文献ファミリー２の特許からの引用状況（出願日順）

文献元特許			引用先文献	
公報番号	パテント ファミリー ID	出願日	medRxiv （プレプリント）	Emerging Infect. Dis. （原著論文）
EP-3734286-A1	70775260	15-May-20	○	
WO-2021169167-A1	77489852	24-Jul-20	○	
EP-3889604-A1	71943928	31-Jul-20	○	
DE-202020005492-U1	72290786	27-Aug-20	○	○
EP-3800473-A1	72290786	27-Aug-20	○	
DE-202021100842-U1	74591943	19-Feb-21		○
EP-3978927-A2	74591943	19-Feb-21		○
WO-2021202812-A1	77855816	31-Mar-21	○	

WO-2021198503-A1	77930276	01-Apr-21	○	
WO-2021203034-A3	77929241	02-Apr-21	○	
WO-2021203093-A1	77929078	05-Apr-21	○	
WO-2021207281-A2	75747059	06-Apr-21	○	
WO-2021216276-A1	78269863	06-Apr-21	○	
WO-2021211331-A1	75675023	07-Apr-21	○	
WO-2021211332-A3	75954246	07-Apr-21	○	
WO-2021209463-A1	75438806	13-Apr-21	○	
WO-2021211688-A1	78084791	14-Apr-21	○	
WO-2021222315-A3	78374237	27-Apr-21	○	
WO-2021222316-A3	78374237	27-Apr-21	○	
WO-2021226200-A1	76181224	05-May-21	○	
EP-3910332-A1	71994825	11-May-21	○	
WO-2021231659-A1	78513191	12-May-21	○	
EP-3855186-A3	70775260	14-May-21	○	
WO-2021229078-A1	70775260	14-May-21	○	○
EP-3855186-A2	70775260	14-May-21		○
WO-2021236790-A1	76444610	19-May-21	○	
WO-2021257695-A3	79268349	16-Jun-21	○	
WO-2022022927-A1	76845200	28-Jun-21	○	
WO-2022036330-A1	80247446	16-Aug-21	○	
WO-2022084672-A1	73005325	20-Oct-21		○
WO-2022132625-A1	80121655	13-Dec-21		○

事例３　複数の原著論文から構成される文献ファミリー（文献ファミリー３）

３つ目のパターンは、やや特殊な例になるが、複数の学協会が連携して策定した声明やガイドライン等である。このような場合、同一の論文が複数のジャーナルに投稿されている。多くの同僚研究者に向けてのメッセージとして出される性質から、論文から多くの引用を集めやすいものと思われる。

ここで挙げる事例は、欧州心臓学会（European Society of Cardiology: ESC）、米国心臓学会（American College of Cardiology Foundation：ACCF）、米国心臓協会（American Heart Association：AHA）、世界心臓連合（World Heart Federation：WHF）が共同で策定した心筋梗塞の基準についての専門家コンセンサス文書である。ESC、AHA、ACCF のジャーナル（European Heart Journal: EHJ、Circulation、Journal of the American College of Cardiology: JACC）からそれぞれ同一の内容（Thygesen et al., 2007a, 2007b, 2007c）で出版されている。なお、Scopus では EHJ で出版された文献のみプロシーディングス（Conference Paper）で、他はレビュー論文として索引されているが、前者が特定の会議で発表されたという記述を当該文献のウェブページ及び文献中で見出すことはできなかったため、３文献とも原著論文として扱って良いものと思われる。

３文献の特許からの引用状況を表３に示す。EHJ、Circulation、JACC 掲載文献は、それぞれ６公報（４パテントファミリー）、20 公報（17 パテントファミリー）、６公報（６パテントファミリー）から引用されており、３報で 31 公報（25 パテントファミリー）から引用を得ている。３文献の中では Circulation 誌掲載文献が最も早期から継続的に引用されている。１パテントファミリーからの引用に止まるが、EHJ 掲載文献は欧州特許から引用されており、欧州の学会誌から出版したことが、欧州地域の関係者への認知度を高めた可能性はあろう。

表３　文献ファミリー３の特許からの引用状況（出願日順）

引用元特許			引用先文献		
公報番号	パテント ファミリー ID	出願日	Eur. Heart J.（プロシーディングス）	Circulation（原著論文【レビュー】）	J. Am. Coll. Cardiol.（原著論文【レビュー】）
US-8298772-B2	39157469	06-Mar-09		○	
US-8507209-B2	41065428	12-Mar-09		○	
US-9630985-B2	41065427	12-Mar-09		○	
WO-2009113880-A1	41065428	12-Mar-09		○	
US-9116155-B2	41415538	10-Jun-09		○	
US-9492437-B2	42542846	22-Jul-10	○		
US-8501420-B2	42968402	30-Aug-10			○
US-9925265-B2	43974647	10-Nov-10		○	
EP-3100739-A1	44226817	30-Dec-10	○		
EP-3266462-A1	44226817	30-Dec-10	○		
EP-3443977-A1	44226817	30-Dec-10	○		
US-9182413-B2	42045944	21-Jan-11		○	
US-8663941-B2	41623938	24-May-12	○		
WO-2013049278-A1	47045163	27-Sep-12		○	
US-10426368-B2	47846087	14-Jan-13		○	
WO-2013111031-A1	47846087	14-Jan-13			○
US-10114028-B2	49223053	15-Mar-13		○	
WO-2013150105-A1	48141938	04-Apr-13		○	
US-9255930-B2	39157469	15-Apr-13		○	
WO-2014004987-A2	49784031	28-Jun-13		○	
US-10376532-B2	50475863	09-Oct-13		○	
US-9841430-B2	52625870	10-Sep-14		○	
US-9994631-B2	45497426	20-Apr-15		○	
WO-2016166627-A1	55702037	29-Mar-16			○
US-11351187-B2	48797710	02-Aug-16	○	○	
WO-2017035639-A1	58186392	26-Aug-16			○
US-11147498-B2	59965328	31-Mar-17		○	
WO-2018096162-A1	60937677	28-Nov-17			○
WO-2018096163-A1	60702634	28-Nov-17			○
US-11147879-B2	43974647	08-Feb-18		○	
US-11260071-B2	62842064	22-Jun-18		○	

4　まとめ

本稿では、パテントファミリーに擬えた文献ファミリーという概念を提案し、ファミリー単位で学術文献を捉えることにより、特許からの引用の見え方がどのように変わるかを、３事例について提示した。

本稿で示した３事例は、いずれも同一内容の文献であることが明確であることに加え、すべての文献が特許から多く引用されている点において、やや特殊と言える。文献ファミリー集合を俯瞰的に見た場合には、本稿で示した個別の特殊事例とは異なる傾向が顕現するものと考えられるので、指標の閾値設定を踏まえて、文献ファミリー集合の統計的な性質についても把握を進めたい。

現段階では、上記指標の閾値設定が完了しておらず、明確かつ効率的な文献ファミリーの識別法を確立できていない。今後、サンプルを個別に参照しながら、抽出基準を見定めていく必要がある。

また、本稿では原著論文（レビュー論文を含む）が

含まれる文献ファミリーのみを取り上げたが、プロシーディングスやプレプリントが最終的なアウトプットとなるケースも考えられる[5]。今後、想定されえる文献ファミリーの構成について、網羅的に把握・分析を進めていく必要がある。

なお、「特許からの学術文献引用」を分析する観点からは、Scopus に収録されているプレプリントが2017年以降のものに限られる点が制約となった。より正確な知見を得るためには、Scopus に収録されたプレプリントを含む文献ファミリーが、特許から十分な引用を得る期間として、もう２～３年程度を担保した方が良いように思われる。

参考文献

arXiv.org.（2022, February 17）. *New arXiv articles are now automatically assigned DOIs. Update: As of Feb 2022, all arXiv articles now have DOIs*. arXiv.org blog. https://blog.arxiv.org/2022/02/17/new-arxiv-articles-are-now-automatically-assigned-dois/

Cabanac, G., Oikonomidi, T., & Boutron, I.（2021）. Day-to-day discovery of preprint–publication links. *Scientometrics*, 126, 5285–5304. DOI: 10.1007/s11192-021-03900-7.

林 和弘，小柴等．（2020）．arXiv に着目したプレプリントの分析．NISTEP DISCUSSION PAPER，No.187，文部科学省科学技術・学術政策研究所．DOI: 10.15108/dp187.

林 和弘．（2021）．COVID-19 で加速するオープンサイエンス －プレプリント分析にみる学術情報流通の変容－．*STI Horizon*，7（1），40-45. DOI: 10.15108/stih.00249.

小柴等，林 和弘，伊藤裕子．（2020）．COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析 .NISTEP DISCUSSION PAPER，No.186，文部科学省科学技術・学術政策研究所 .DOI: 10.15108/dp186.

Long, J., Shelhamer, E., & Darrell, T.（2015）. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）*，3431-3440. DOI: 10.1109/CVPR.2015.7298965.

MEXT–NISTEP プレプリント調査・検討チーム .（2020）. プレプリントをめぐる近年の動向及び今後の科学技術行政への示唆 . https://www.mext.go.jp/content/20201026-mxt_jyohoka01-000010684_2.pdf

Okba, N. M. A., Müller, M. A., Li, W., et al.（2020a）. *SARS-CoV-2 specific antibody responses in COVID-19 patients*. medRxiv. DOI: 10.1101/2020.03.18.20038059.

Okba, N., Müller, M. A., Li, W., et al.（2020b）. Severe Acute Respiratory Syndrome Coronavirus 2 − Specific Antibody Responses in Coronavirus Disease Patients. *Emerging Infectious Diseases*, 26, 1478-1488. DOI: 10.3201/eid2607.200841.

Shelhamer, E., Long, J., & Darrell, T.（2017）. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 640-651. DOI: 10.1109/TPAMI.2016.2572683.

Thygesen, K., Alpert, J. S., White, H. D. et al.（2007a）. Universal definition of myocardial infarction, *European Heart Journal*, 28, 2525–2538. DOI: 10.1093/eurheartj/ehm355.

Thygesen, K., Alpert, J. S., White, H. D. et al.（2007b）. Universal definition of myocardial infarction, *Circulation*, 116, 2634-2653. DOI: 10.1161/CIRCULATIONAHA.107.187397.

Thygesen, K., Alpert, J. S., White, H. D. et al.（2007c）. Universal definition of myocardial infarction, *Journal of the American College of Cardiology*, 50, 2173-2195. DOI: 10.1016/j.jacc.2007.09.011.

5　林らの分析では、必ずしも原著論文化されないプレプリントが arXiv 中に多数含まれることが確認されている（林，小柴，2020）。