２Ｂ１６

文献ファミリーの提案と同定・分析

○山下泰弘（JST），吉田秀紀（JST），高坂香那（エルゼビア・ジャパン）
yasuhiro.yamashita@jst.go.jp

1. はじめに

新型コロナウイルス感染症のパンデミックを端緒として、研究成果の早期把握が喫緊の課題となり、論文の出版サイクルが加速されるとともに、プレプリントが注目を浴びるようになった。プレプリントは、論文審査のタイムラグを埋める速報メディアであるが、プレプリントサーバには論文化を出口としない文献も多く収録されており（林・小柴 2020)、それ自体が独立した研究成果公開メディアとしての側面も有している。

一方、口頭発表を経て論文出版をする形も、従来から標準的な成果発表の道筋として、広く用いられている。分野によっては、権威ある国際会議での発表がジャーナル論文以上に重視されるケースもある。

研究課題のインパクトを定量的に計測する際には、論文が単位とされる場合が多いが、上記のように同一内容の文献が多様な形態で併存し、それぞれが個別に利用されている状況、分野により重視されるメディアが異なることなどを考慮すると、正当に評価されていない可能性がある。

そこで、本稿では、特許におけるパテントファミリーと同様に、同一研究者による同一内容の文献群を「文献」ファミリーとして定義し、その同定を試行する。

2. 文献ファミリー

我々が提案する「文献ファミリー」は、必ずしもプレプリント―原著論文のペアのみとは限らない。何らかの理由で出版された複数の同一内容文献の組み合わせである。これまでに我々が見出した例としては、複数ジャーナルから共同出版される専門家コンセンサス文書などの「原著論文」同士の組み合わせがある（山下・吉田・高坂 2022)。

3. 文献ファミリー抽出の試み

候補集合の作成

文献ファミリーを構成する文献の可能な組み合わせについて、網羅的な把握はなされていない。網羅的に文献ファミリーを抽出するには、本来であれば文献データベースの収録文献の組み合わせすべてについて同一性をチェックする必要があるが、計算量が膨大になる。本研究では、試行的に、2017 年以降の高被引用論文（約 30,000 報）を核として、それらとタイトルが類似し、かつ著者が共通している論文を Scopus より抽出する形をとった。この段階では、閾値を緩めに設定し、文献ファミリー候補を広めに抽出し、以降の処理で絞り込むこととした。

上記候補から文献ファミリーを同定するに当たり、タイトルと著者のみでは判定が困難なため、第 2 段階では、ファミリー候補文献間のアブストラクト及びリファレンスの類似度を算出し、判定の一助とした。この処理には、プレプリントに対しても他の文献タイプと同様の処理を行うことが容易な Dimensions Analytics API を使用した。

本稿では、文献ファミリーの可能な構成についての知見を得るために、文献ファミリーである確度が高い集合に絞って精査する方法をとることとした。そのために、アブストラクト及びリファレンスの類似度の閾値を高く設定し（ともに 0.9 以上)、絞り込んだ文献ファミリー候補集合（127 クラスタ）を作成した（図１)。

識別と特徴の分析

現在は、上記処理により作成した文献ファミリー候補に対し、目視によって、判定の正しさと、ファミリーの性質について確認を行っている段階である。図１の各ノードは文献を表し（文献タイプごとに着色)、エッジで相互接続されているクラスタが文献ファミリー候補である。一部自己ループを形成しているが、これらのうち６件の内訳を確認したところ、いずれも Scopus 上で異なる２レコードが存在するが、DOI が一つしか付与されていないため、Dimensions と接続した際に１文献に束ねられたもので

あった。これまでに全体の約半数の文献ファミリー候補の精査が完了しているが、文献ファミリーに該当しないものは１件（図１[106]）のみである[1]。一部、ノード間に複数エッジが張られているケースがあるが、文献データの重複によるものであり、今後プロセスの修正により排除される見込みである。

図１　文献ファミリー候補集合（アブストラクト・リファレンスともコサイン類似度 0.9 以上）

4. 文献ファミリーの内訳

現在は、図１の各ファミリー候補を精査している段階であり、全体の傾向を網羅するものではないが、表１に現在までに見出された文献ファミリーの特徴を示す。最も多いパターンは「プレプリント経由」による、いわば通常パターンでの出版であり、47 クラスタが該当する。これらは全件文献ファミリーであることを目視確認済みである（類型①）。

同一のメディア（プレプリントサーバ、Cochrane Database Syst. Rev.）の同一文献の複数バージョンが Scopus 及び Dimensions/に個別に収録されているケースもある（類型②）。同一文献とはみなされておらず、バージョンごとに引用が計測されている。

また、複数の主体のコラボレーションの結果、複数の書誌事項が付与されるケースもあった（類型③）。例えば、複数の学協会・出版社が連携してガイドラインや声明を出版しているケースである。図１の中で３文献以上から形成されているクラスタは、プレプリントを含む１件（[86]）を除き、すべてがこれに該当する。これは、策定したガイドライン等を広く普及させるために、３者以上が関与するケースが多いためと考えられる。ガイドライン等に関する文献ファミリーは、Article、Review、Editorial など多様な文献タイプを含むが、実態としてはすべて同じ文献である。また、学会の２つの分科会のジョイントセッションのプロシーディングスが、分科会ごとに書誌事項を付与されて刊行されているケースも見られた（[107]）。このケースも、プロシーディングは２報出ているが、実態としては口頭発表１件である。

---

[1] 文献ファミリーに該当しないケースは、ほぼ同じタイトル・アブストラクトで３年間隔を空けて１つのトピックの進捗についての口頭発表を行ったもの。

出版社が多様な形態で出版することによって、文献ファミリーが形成されるケースもある。図１では、同一内容の論文を２言語（英語・ドイツ語）のジャーナルで出版したケースと、複数の書籍に論文を収録したケースが見られた。

最後に、重複論文の取下げ、掲載号の移動、リプリント発行により、データベース中に複数の書誌事項が併存しているものもあった（類型⑤）。本稿で見出されたのは、出版社のミスにより、掲載方法に齟齬が生じたケースであるが、広義で捉えるならば、不正に重複投稿されたケースなども考えられる。

なお、当初主眼に置いていたカンファレンスペーパーと原著論文（Article, Letter, Note, Review）の組み合わせは、図１では見出されていない[2]。これは第一段階で抽出されたカンファレンスペーパーの数が相対的に少ない（プレプリント 3,674 報に対しカンファレンスペーパー1,288 報）こともあるが、原著論文化に当たって加筆修正されていることも影響している可能性がある。今後、カンファレンスペーパー・原著論文から構成されるファミリーが抽出可能な条件について検討を進めたい。

表１　現時点で見出されている文献ファミリーの構成パターン

類型	内訳	主な組合せパターン	備考
①プレプリント経由での VoR の出版	プレプリント経由の原著論文出版	Preprint-Article, Review	図 1[5][6][9]等 47 クラスタ
②同一メディアのバージョン違い	プレプリントの複数バージョン	Preprint-Preprint	図 1[62]
	Cochrane Database Systematic Review の複数バージョン	Review-Review	図 1[48]
③コラボレーション	複数学会・出版社によるガイドライン等の共同出版	Article, Review, Conference Paper, Editorial 等の組み合わせ	図 1[1][3][4]等 3 報以上のクラスタの多くが該当。
	ジョイントセッションのプロシーディングスを個々に出版	Conference paper-Conference paper	図 1[107]
④出版社による複数形態での出版	同一論文の英語版とドイツ語版	Review-Article	図 1[56]
	同一出版社の複数書籍	Chapter-Chapter	図 1[7]
⑤出版社による重複出版・削除	出版社による重複掲載取下げ・出版号の移動	Review-Review	図 1[15][82]
	出版社によるリプリント	Article-Article Review-Review	図 1[12][13][27][36][50][86][99]

5. 文献ファミリー引用の特徴

以下では、ファミリー単位で捉えた場合、被引用数がどのように変わるか、表１の類型ごとに事例を示す（表２）。

最も多いパターンのプレプリント経由での VoR の出版（類型①）の事例としては、COVID-19 ワクチン接種のグローバルなデータベースに関する論文が挙げられる。Dimensions によると、medRxiv に投稿されたプレプリントが 60 回、VoR として Nature Human Behavior 誌に掲載された論文が 636 回引用されており、うち両者に共通する引用が 30 回である。試しにそのうち 5 件について、原報を当たって確認したところ、いずれも実際には VoR しか引用されていなかった。同一内容の複数のファミリー文献を同時引用することは、ファミリー文献間の差異を比較する以外の目的で発生するとは考えにくいので、以降の事例も含めて、実際には重複引用はもっと稀なものと考えられる。

類型②の事例としては、Cochrane Database Systematic Review に掲載されたレビュー論文の異バージョンが挙げられる。第 4 版と 5 版が Scopus に収録されており、Dimensions における被引用数はそれぞれ 229、93 で、両者に共通する引用は 14 である。ただし、重複引用のうち 5 件を確認したところ、片方のバージョンしか引用されていなかった。

類型③の事例としては、1,000 以上のジャーナルに支持されていたとされる動物実験のレポーティングガイドライン（ARRIVE）の改訂版である。この文献ファミリーは、PLoS Biology 誌掲載版を母体とし、その他 6 誌からも刊行されている。最も多く引用されている文献は、母体となった PLoS Biology 掲載版（1,207 回）であるが、その他の文献の被引用数も 100 以上のものが 3 報とかなり多く、内容の

[2] 図中に Conference Paper-Article から構成されるクラスタが２つあるが、いずれも口頭発表の形跡が見出されない。

表２　表１の各類型ファミリーの被引用数

①プレプリント経由での VoR 出版（COVID-19 ワクチン接種 DB 論文 図１[6]）

Dimensions ID	掲載誌等	被引用数（Dimensions）		
		文献	ファミリー	重複分
pub.1136701133	medRxiv (preprint)	60	664	32
pub.1137877686	Nat. Hum Behav.	636		

②同一メディアのバージョン違い Cochrane DB Sys Rev.の異バージョン 図 1[48]）

Dimensions ID	掲載誌等	被引用数（Dimensions）		
		文献	ファミリー	重複分
pub.1126661832	Cochrane Database Syst. Rev. (4 版)	229	308	14
pub.1127615948	Cochrane Database Syst. Rev. (5 版)	93		

③コラボレーション（複数学会・出版社によるガイドライン等の共同出版 図 1[1]）

Dimensions ID	掲載誌等	被引用数（Dimensions）		
		論文	ファミリー	重複分
pub.1129355107	PLoS Biol.	1,207	1,877	24
pub.1129359330	J. Cereb. Blood Flow Metab.	269		
pub.1129326696	Br. J. Pharmacol.	147		
pub.1129353555	J. Physiol.	110		
pub.1129322981	BMC Vet. Res.	70		
pub.1129501814	BMJ Open Sci.	54		
pub.1129351066	Exp. Physiol.	44		

④出版社による複数形態での出版（同一論文の英語版とドイツ語版 図 1[56]）

Dimensions ID	掲載誌等	被引用数（Dimensions）		
		論文	ファミリー	重複分
pub.1103785567	J. Neuroinflammation （英語）	420	437	11
pub.1107273050	Nervenarzt （ドイツ語）	28		

⑤出版社による重複出版・削除（掲載号を誤ったために移動されたケース 図 1[15]）

Dimensions ID	掲載誌等	被引用数（Dimensions）		
		論文	ファミリー	重複分
pub.1084061546	Biol. Control （新）	178	189	0
pub.1074204222	Biol. Control （削除）	11		

インパクトを評価する上で無視すべきではないと思われる。

類型④の事例としては、同一内容の論文が英語とドイツ語のジャーナルで発表されたものがある。英語版の方が 15 倍多く引用されているが、ドイツ語版も 28 回引用されており、ローカルなインパクトを考慮するならば、加味して評価すべきであろう。

類型⑤については、出版社が特集号に掲載する論文を誤って通常号に掲載したケースがある。出版社は新たに出版した特集号の論文を引用するよう指示を掲載しているが、通常号掲載論文も 11 回引用されている。正当に評価するのであれば、両者の被引用数は合算されるべきであろう。

6. まとめ

本稿では、同一内容の文献を包含する概念として文献ファミリーの提案をし、その抽出と類型化を試みた。この概念は、特許におけるパテントファミリーと同様に、重複のない正味の文献数や被引用数を把握する上で有用と考えられる。図１の 127 文献ファミリー候補の約半数を精査したが、文献ファミリーに該当しないものは１件のみであり、現状でも高い精度で抽出されている。ただし、それと引き換えに、カンファレンスペーパーを含むファミリーなどが抽出されにくくなっていると考えられる。今後、第一段階（タイトル、著者に基づくファミリー候補抽出）も含めて最適化を進め、より多様なタイプの文献ファミリーの抽出を行う必要がある。

本研究は、進行中の取組であり、口頭発表時にはその後の進捗も加味した内容を報告したい。

参考文献

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

山下 泰弘，吉田 秀紀，高坂 香那. (2022). 文献ファミリーの同定と特許からの引用評価の試み. Japio YEAR BOOK 2022. （印刷中）.