DISCUSSION PAPER No.186

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

A Trial of early detection system for research trends through the preprints data

— Research status around COVID-19 / SARS-CoV-2

2020 年 6 月

文部科学省　科学技術・学術政策研究所

小柴　等，林　和弘，伊藤　裕子

【執筆者】

小柴 等　第2調査研究グループ

林 和弘　科学技術予測センター

伊藤 裕子　科学技術予測センター

【Authors】

KOSHIBA Hitoshi　2nd Policy-Oriented Research Group,
National Institute of Science and Technology Policy (NISTEP), MEXT

HAYASHI Kazuhiro　Science and Technology Foresight Center,
National Institute of Science and Technology Policy (NISTEP), MEXT

ITO Yuko　Science and Technology Foresight Center,
National Institute of Science and Technology Policy (NISTEP), MEXT

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this paper.

小柴 等, 林 和弘, 伊藤 裕子 「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: http://doi.org/10.15108/dp186

KOSHIBA Hitoshi, HAYASHI Kazuhiro, ITO Yuko, "A Trial of early detection system for research trends through the preprints data — Research status around COVID-19 / SARS-CoV-2," *NISTEP DISCUSSION PAPER*, No.186, National Institute of Science and Technology Policy, Tokyo.

DOI: http://doi.org/10.15108/dp186

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

文部科学省　科学技術・学術政策研究所
小柴 等，　林 和弘，　伊藤 裕子

要旨

近年，査読前の論文草稿であるプレプリントをとりまとめて公開するプレプリントサーバの活用が進んでいる。ジャーナル論文の投稿に先立つというプレプリントの性質上，ジャーナル論文を対象とした研究動向分析と比較して早期に研究動向を把握できる可能性があり，そのため新興の（エマージングな）研究領域の補足に有用と考えられる。

こうした背景のもと，本報ではCOVID−19関連のプレプリントを対象に，自然言語処理を用いたエマージング領域の把握を試行した。

その結果，既存の分析と同様に，疫学調査のステップに合致する動向を得ることができた。それに加えて，通常の査読論文も含めて分析した先行研究では明確には検出ができていなかった，医薬・ワクチン開発に関するトピックを抽出することができた。

今回の試行により，プレプリントを利用したエマージング研究内容のメタ把握が実現できる可能性が示唆された。

A Trial of early detection system for research trends through the preprints data — Research status around COVID-19 / SARS-CoV-2

KOSHIBA Hitoshi, HAYASHI Kazuhiro, ITO Yuko
National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

In recent years, the use of preprint servers, which compile and publish preprints of pre-reviewed drafts of articles before peer review is in progress. Due to the nature of the preprint, which precedes the submission of a journal article, the research on journal articles in comparison with trend analysis, it is possible to understand research trends at an earlier stage, and therefore it can be used to supplement emerging research areas. It is thought to be useful.

In this report, we attempted to understand the emerging regions of COVID-19 related preprints by using natural language processing.

As a result, we were able to obtain the trend consistent with the epidemiological survey step as well as the existing analysis. In addition, we were able to detect the emergent regions, which had not been clearly detected in previous studies that also included ordinary peer-reviewed papers. In addition, we were able to extract topics related to drug and vaccine development.

This trial demonstrated the possibility of realizing a meta-understanding of emerging research content using preprints.

目次

1 はじめに 1

2 データ・手法 2

2.1 対象 2

2.2 分散表現辞書の作成 4

3 結果・考察 4

3.1 PPS 記事の分析結果 5

3.2 既存の分析との比較 8

4 まとめ 9

1　はじめに

EBPM (Evidence based Policy Making) を推進するためには，データの蓄積・把握が重要である．科学技術に関わる政策立案やファンディングの検討のためには研究動向の把握が必要である．

しかし，研究分野を複数またがるような，学際的なエマージング研究を定量的に把握することは容易ではない．現状では例えば，ある研究分野に投稿された論文が，どのような分野の論文から引用されているかに基づいて，エマージングな研究領域を定義したり [Small85, 伊神 09, 治部 12]，その融合度を定義したり [Okamura19]，といった手法が提案されており，これらはある程度有効に機能している．

ただし，これらは引用関係を用いるという特性上，論文の公開から分析までに一定以上の期間を要する．仮に引用情報を用いないとしても，査読という性質上，投稿してから出版までには通常数ヶ月，長ければ 1 年を超える期間を要する．さらに，解析に手間がかかることから多くの場合引用数ベースでトップ 10% や 1% など，上位の論文に絞らなければ解析が難しいという面もある．したがって，例えば 2019 年末から始まった COVID-19 の流行のような緊急性の高い案件について，比較的短期に分野間の融合の状況を見る，といったことには使いづらい側面もある．

ここで，近年プレプリントサーバ (Preprint Server, PPS) の活用が進みつつあることに着目する．“プレプリントとは，主に査読付きジャーナルに投稿する前の草稿原稿のこと” であり “このプレプリントを掲載して誰でも読めるようにする” サービスが PPS である [林 20]．投稿前の草稿という特性上，査読論文に比べてその信頼性は必ずしも担保されないが，代わりに最新の情報を得られる可能性がある．PPS は有名なものでは 1990 年代初めから運用されている arXiv などが存在し，近年では，医学や化学などの分野でも PPS が開設され始めている．こうした情勢によりプレプリントだけでもある程度の情報量を確保することが可能となりつつあり，有識者の間でもプレプリントの動向把握をしておくことの重要性を指摘する声がある [文科 19]．特に COVID-19 を機に関連するプレプリントの登録数は飛躍的に増大しており，動向把握のための情報源として重要な位置を占めつつある．

このようにプレプリントを用いることで，査読付き論文と比較するとより最新の情報が得られるものの，動向把握についても別手法が必要となる．すなわち，引用・被引用関係からの動向把握を行う限り，分析に資するだけの引用がなされるまで数年の期間が必要であり，プレプリントの先行性は誤差の範囲にとどまる可能性が高くなるためである．そこで，引用情報を用いない研究動向自体の把握については，自然言語処理を用いて実現する．具体的にはプレプリントの各記事に付与されたタイトルや概要を手がかりとして，ここからトピックを抽出し，それらをベースとして分析する．具体的な手法は文献 [小柴 20a] を踏襲する．

この手法は前述した先行研究 [Small85, 伊神 09, 治部 12, Okamura19] と異なり，数値的な判断が困難であるほか，トピックの意味解釈を要する点に難点があるが，その一方で，最低限一定量のタイトル・概要のみがあれば分析ができるため，引用関係の分析に比べて短期で，かつ様々なデータソースを横断的に分析できる点に利点がある．したがって，先行研究と並列する別種の評価情報として機能することが考えられる．分析可能なデータ数の面においても，例えば文献 [小柴 19] では同様の手法で約 5.6 千万件の文献を分析しており，十分に機能することが期待される．

以上より本稿では，COVID-19 に関するプレプリントの各記事を用いて内容の近さで論文をマッピング・分類することにより，エマージング研究のメタ把握を試みた．

COVID19 は，2019 年 12 月に中国武漢で大流行した重症急性呼吸器症状を特徴とする新型コロナウイルス（SARS-CoV）による感染症であり，その後世界中に感染拡大しパンデミックとなった．2020 年 6 月 2 日現在

で，216 の国や地域で約 620 万人が感染し，37 万人以上が亡くなっている．

結果，医療のみならず経済の面においても大きな影響を与えており，同一の現象に対して様々な分野から取り組みが行われている様子や，融合している様子が観察できることが期待できる．

2　データ・手法

データおよびその詳細は文献 [小柴 20b] の通りであるが，以下でも改めて解説する．

2.1　対象

arXiv, medRxiv, bioRxiv, chemRxiv, SSRN (Social Science Research Network) という 5 つの PPS を対象にした．

これらについては文献 [林 20] でまとめられているとおり，それぞれ図 1 の分野に強みを有する．

名称	創設年	2020年1月現在の運営母体	分野	システム	DOI
arXiv	1991	コーネル大学	物理学に始まり，情報学，経済学等多分野に広がる	オリジナル	×
SSRN	1994	Elsevier	社会科学に始まり多分野に広がる	オリジナル（ColdFusion）	○
BioRxiv	2013	コールド・スプリング・ハーバー研究所	生命科学を中心とした分野	HighWirePress	○
ChemRxiv	2017	米国化学会，英国化学会，ドイツ化学会，日本化学会，中国化学会	化学を中心とした分野	figshare	○
MedRxiv	2019	米イエール大学，コールド・スプリングハーバー研究所，BMJ(British Medical Journal)	医学を中心とした分野	HighWirePress	○

（ 文献[林20] 図表3をもとに作成 ）

図 1　PPS と主要分野

さらに，これらの PPS の多くは今回の COVID-19 / SARS-CoV-2 に関連する記事について，独自にとりまとめたリンク集を生成している（図 2 参照）．

chemRxiv については，データ収集を行った 2020 年 5 月 25 日時点で，chemRxiv 独自のリストは見当たらなかったが，論文等文献検索サービスである Dimensions [1)]がとりまとめて公開している COVID-19 関連のデータセット[2)]に chemRxiv 上の記事が出てくるため，これを利用した[3)]．

その上で各リストに掲載された各記事の投稿日，タイトル，概要などの書誌情報を収集し，分析することにした．

ここで，SSRN については図 3 に示すとおり，“Preprints with THE LANCET” との表示がついた記事も散見される．

Lancet は医学系の著名雑誌の一つであり，相対的に人文社会系のプレプリントが多いと考えられる SSRN

---

1) https://app.dimensions.ai/

2) https://dimensions.figshare.com/articles/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063

3) chemRxiv は figshare というシステムを採用しているが，Dimensions，figshare の運用母体は両方とも “Digital Science” 社である．

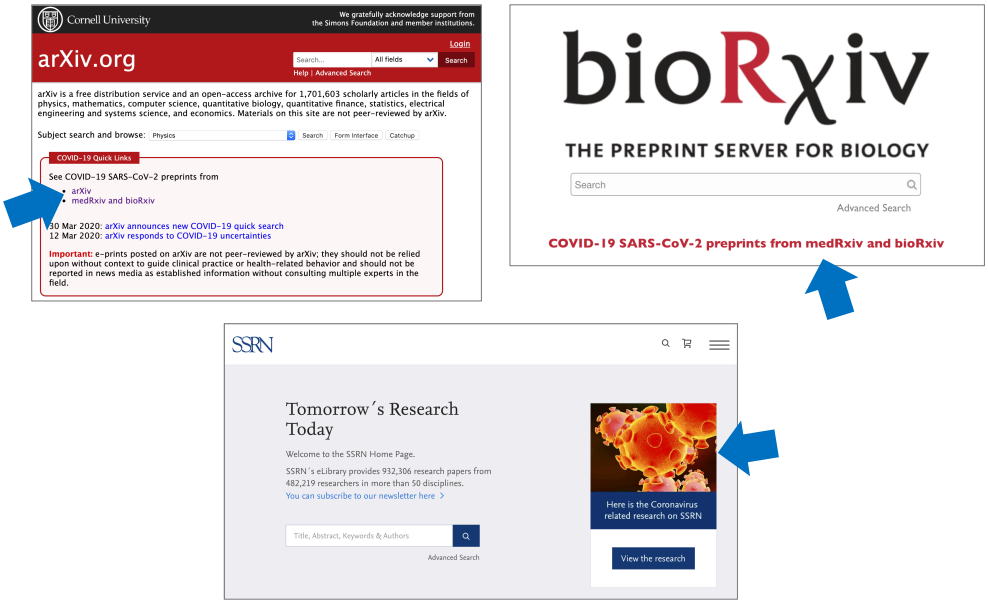


図 2　収集対象

の中では異質と言える．そこで本報告においては，“Preprints with THE LANCET” との表示がついた記事について，これを “SSRN Lancet” と切り分けて扱うことにした．

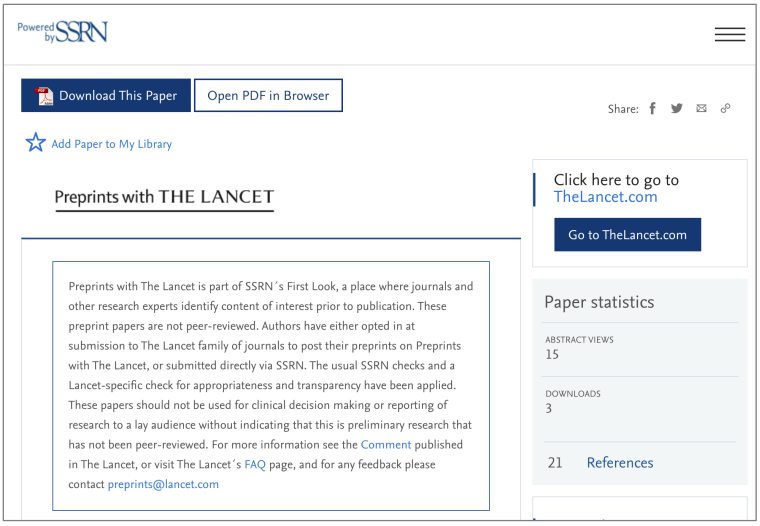


図 3　SSRN における Lancet のプレプリント

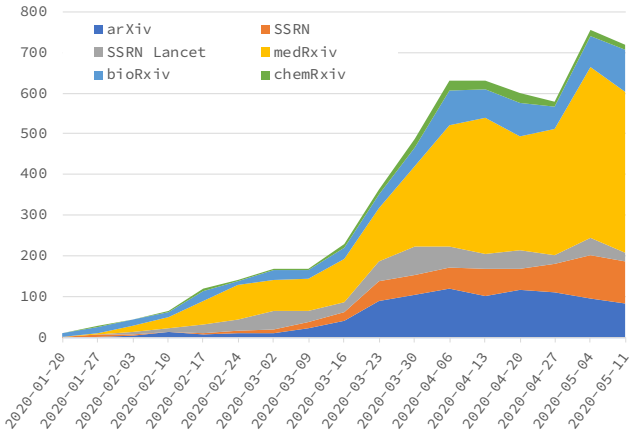
収集した記事の中には 2018 年など古い情報も見られたため，記事を 2020 年 1 月以降のものに限定し，結果として，それぞれ表 1 に示した記事数を得た．

これら記事に関する 2020 年 第 4 週以降の週次投稿数推移を図 4 に示した．

PPS	Num
arXiv	936
bioRxiv	716
chemRxiv	175

PPS	Num
medRxiv	2837
SSRN	612
SSRN Lancet	496

表 1　PSS ごとの記事数 (2020 年 5 月 25 日時点)



Week		arXiv	SSRN	Lancet	medRxiv	bioRxiv	chemRxiv
2020-01-20	(04)	0	1	0	0	9	1
2020-01-27	(05)	2	5	1	3	16	1
2020-02-03	(06)	4	5	5	16	13	1
2020-02-10	(07)	13	2	8	27	11	5
2020-02-17	(08)	7	4	20	59	24	6
2020-02-24	(09)	10	6	28	84	9	3
2020-03-02	(10)	10	10	45	76	24	5
2020-03-09	(11)	24	13	27	81	22	2
2020-03-16	(12)	42	20	26	106	27	10
2020-03-23	(13)	91	46	49	133	31	13
2020-03-30	(14)	105	50	68	198	47	19
2020-04-06	(15)	120	51	51	300	85	25
2020-04-13	(16)	102	68	35	334	70	21
2020-04-20	(17)	116	52	45	280	84	24
2020-04-27	(18)	111	69	23	309	54	13
2020-05-04	(19)	95	107	43	420	76	14
2020-05-11	(20)	84	103	22	394	104	12

図 4　PPS ごとの週次投稿数推移

2.2　分散表現辞書の作成

収集した記事のタイトル，概要をベースとして単語ベースの分散表現辞書を構築し，記事単位での分散表現を構築した．

参考にした文献 [小柴 20a] では，pubMed のデータに基づいて単語ベースの分散表現辞書を構築していたが，今回は arXiv や SSRN などの記事も対象としているため，医療系に閉じない多様な話題が含まれることが想定され，pubMed のデータでは十分な表現獲得が行われない可能性がある．そこで今回は収集した記事のタイトル，概要をベースとして単語ベースの分散表現辞書を構築した．手法には FastText[Bojanowski17, Joulin16] を用い，記事数が 6 千件程度であることから，次元数を 100 として構築した．

また，文献 [小柴 20a] では各記事の概要の記述量に差があることから TF-IDF を用いて特徴語の上位 20 件までをもちい，記事ごとの分散表現を構築していたが，今回はストップワードを除外した上で，全単語を用いて記事ごとの分散表現を構築している．

これらから，単純に文献 [小柴 20a] の結果と比較ができない点には注意が必要である．

3　結果・考察

分析の結果について以下に述べる．

3.1　PPS 記事の分析結果

構築した記事ごとの分散表現について，UMAP[McInnes18] で 2 次元に圧縮し，PPS ごとに可視化したものを図 5 に示す．

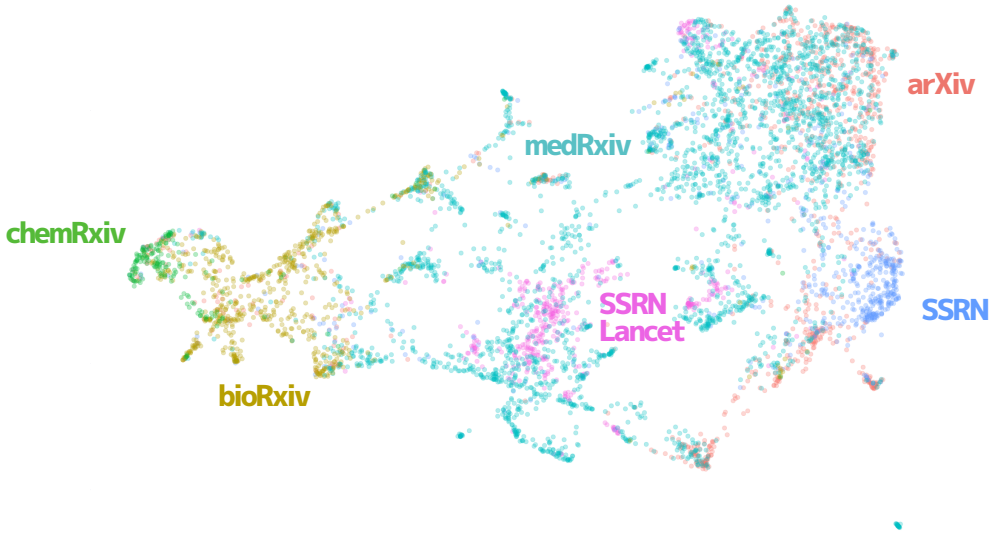


図 5　論文の分布（PPS 単位）

図 5 を見ると，medRxiv 以外は PPS の種類ごとに近くに固まっている様子が見て取れる．SSRN,SSRN Lancet, chemRxiv, bioRxiv それぞれにおいて，塊が読み取りやすく，それぞれの PPS の持つ分野のプレプリントが集まっていることが示唆される．一方，medRxiv は他の 5 種類の PPS を繋ぐように広がっている様子が示された．このことは，medRxiv は対象としている分野が幅広いことを示し，また，COVID-19 関連については分野横断的で学際的な研究が含まれることを示唆している可能性がある．

仮に PPS と分野，若しくはトピックが対応しているとすると，そうした分野・トピックの違いが用語の違いに表れて，このような偏在を示した可能性がある．そこで，文献 [小柴 20a] と同様に，k-means++[Arthur07] を用いて記事を 16 分類し，それらの分類ごとに頻出語のワードクラウドを作成することでトピックの抽出を試みた．

結果を図 6 から 8 に示した．

また，図 7,8 には頻出語に加え，著者のうち，医学・薬学に知見を持つものがワードクラウドから推測できるトピックのラベル（解釈）を付与した．

これら，トピックの解釈を含めた結果について，図 9 に示した．

図 9 を見ると，左端に治療薬やワクチンの島が隣接しており，さらに，治療薬やワクチンの開発に必要なゲノム解析や，感染機構に関連すると思われるトピックが並んでいる．右に目を向けると，社会・経済・政策や肺画像診断のトピックがあり，両者に関連しそうな情報・データ分析がそれらの中間に位置している．これらの結果を鑑みると，トピックの分類はある程度は納得がゆくものである．なお，WHO 論文分析の文献 [小柴 20a] と比較すると，そこは示されなかったトピックや，内容がさらに明確になったトピック等，全体の 1/3 程度は異なっている．

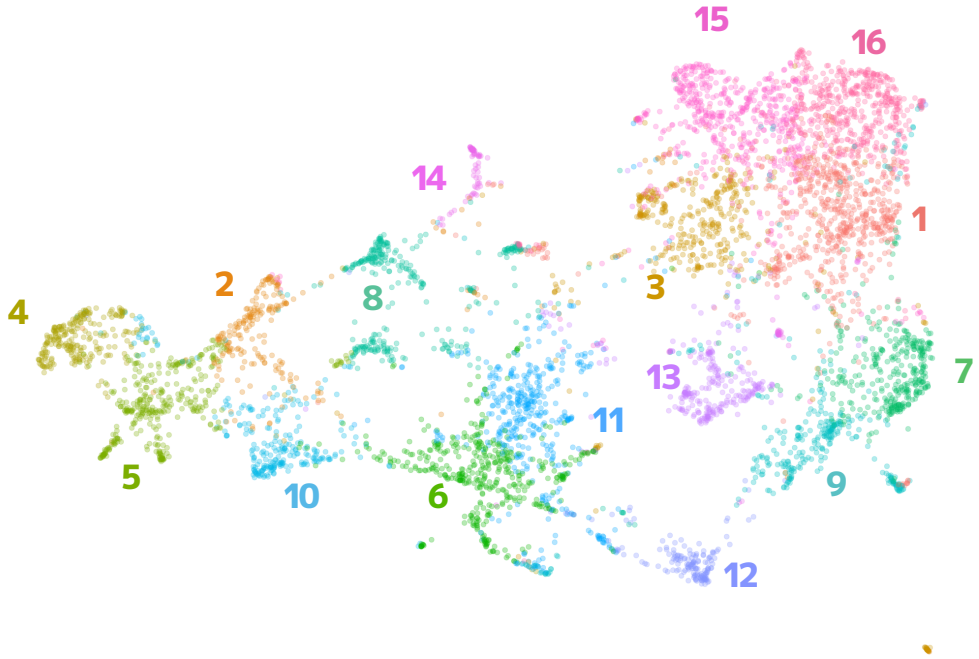


図 6　論文の分布（トピック単位）

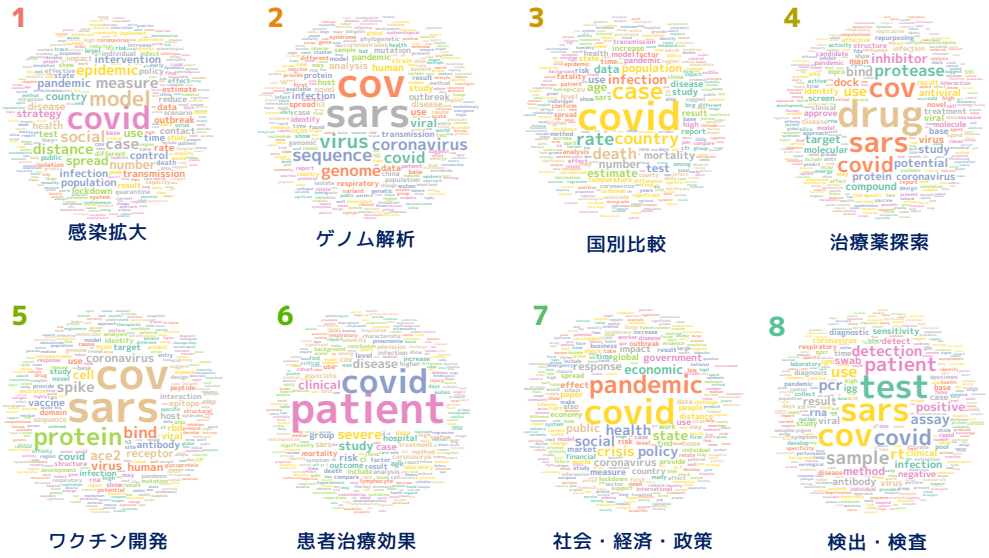


図 7　トピックとワードクラウド (1/2)

そこで，当初の疑問に立ち返り，トピックと PPS の対応について，図 10 にまとめた．

図 10 を見ると，たとえば chemRxiv は治療薬探索のトピックに強く結びついていることが分かる．また，SSRN は社会・経済・政策と，SSRN Lancet は患者病状と，bioRxiv はワクチン開発と強く結びついている事が分かる．逆に medRxiv はもともとの記事数の多さとも相まって，多少の濃淡があるものの広く様々なトピックをカバーしていることが分かる．arXiv もやや幅広ではあるが，こちらは中身を見ると感染モデルや情

図 8　トピックとワードクラウド (2/2)

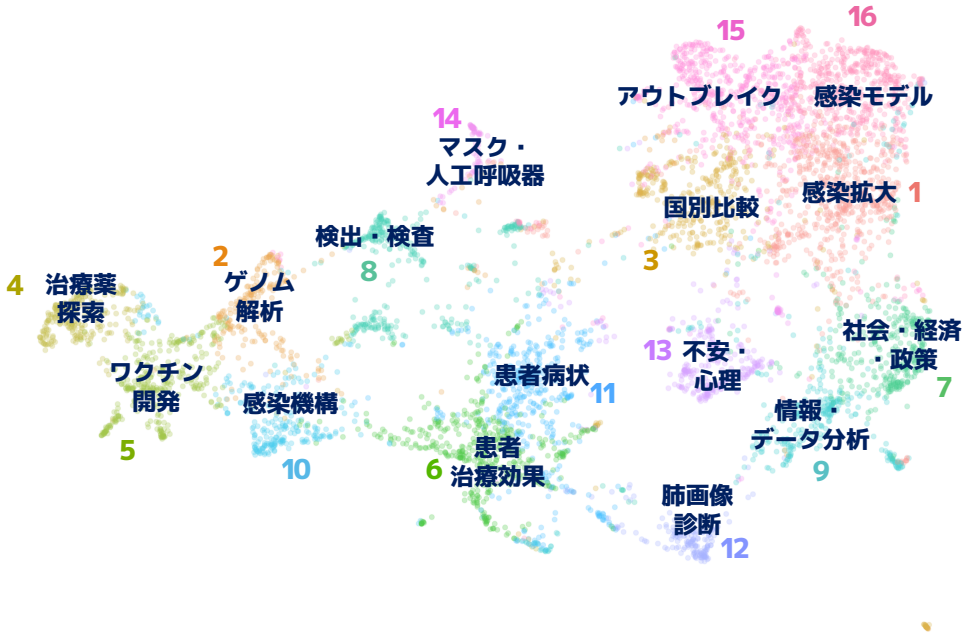


図 9　論文の分布（トピック（解釈）単位）

報・データ分析など，メインは数理系に偏っている．

これらを鑑みるとやはり，PPS とトピックの間にはある程度強いつながりを見て取ることができ，それらは PPS のもともとの分野と関係している事が分かる．

COVID-19 という同じ対象であっても，当然，研究分野ごとにそれぞれの専門性を活かして作業に当たることになるため，このこと自体には何ら驚きはない．一方で，社会・経済・政策などの話題にまで広がりがある

Kind	Total	1 感染拡大	2 ゲノム解析	3 国別比較	4 治療薬探索	5 ワクチン開発	6 患者治療効果	7 社会・経済・政策	8 検出・検査	9 情報・データ分析	10 感染機構	11 患者病状	12 肺画像診断	13 不安・心理	14 マスク・人工呼吸器	15 アウトブレイク	16 感染モデル
arXiv	936	16.5%	1.3%	3.3%	2.8%	3.0%	0.0%	5.0%	1.6%	19.0%	1.6%	0.0%	10.5%	0.7%	1.8%	7.9%	25.0%
bioRxiv	716	0.1%	19.6%	0.0%	8.9%	40.2%	0.0%	0.0%	9.8%	2.1%	17.5%	0.0%	0.3%	0.1%	0.4%	1.0%	0.0%
chemRxiv	175	0.0%	1.1%	0.6%	75.4%	17.1%	0.0%	0.0%	1.1%	1.1%	1.7%	0.6%	0.0%	0.0%	1.1%	0.0%	0.0%
medRxiv	2837	15.2%	2.3%	11.3%	0.7%	0.5%	13.5%	0.6%	7.8%	3.3%	2.9%	6.6%	2.7%	6.7%	2.7%	13.6%	9.8%
SSRN	612	10.3%	2.5%	7.7%	2.0%	1.5%	2.0%	49.3%	0.8%	7.4%	2.1%	1.8%	0.0%	2.0%	1.3%	6.0%	3.4%
SSRN Lancet	496	2.6%	0.2%	1.6%	0.6%	0.0%	12.9%	1.2%	2.0%	1.4%	2.0%	48.4%	0.0%	12.1%	0.0%	13.9%	1.0%

図 10　トピックと PPS

こと，arXiv が仮に「データサイエンス」を意味すると考えた際に，感染拡大や肺画像診断，社会・経済・政策など，様々な分野に進出し，COVID-19 研究がメタサイエンスの地位をある程度確立している様子が観察できたこと，などは今回の発見と言える．

3.2　既存の分析との比較

ここでは文献 [小柴 20a] との比較を通じて，PPS の特徴について見る．

まず，文献 [小柴 20a] では WHO の文献リストと，一部の PPS を対象としている．結果，期間について多少の差異があるものの，bioRxiv, medRxiv の記事については大部分が共通していると考えられる．ただし，2.2 節で述べたとおり，分散表現辞書の作成方法等にも差があり，単純に比較できない点には留意を要する．

まず本分析におけるトピックの時系列推移を図 11 に示した．

Week	Total	1 感染拡大	2 ゲノム解析	3 国別比較	4 治療薬探索	5 ワクチン開発	6 患者治療効果	7 社会・経済・政策	8 検出・検査	9 情報・データ分析	10 感染機構	11 患者病状	12 肺画像診断	13 不安・心理	14 マスク・人工呼吸器	15 アウトブレイク	16 感染モデル
2020/01/20 (04)	11	0.0%	45.5%	0.0%	9.1%	9.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	36.4%	0.0%
2020/01/27 (05)	28	3.6%	14.3%	0.0%	17.9%	14.3%	0.0%	0.0%	0.0%	3.6%	3.6%	3.6%	0.0%	0.0%	0.0%	35.7%	3.6%
2020/02/03 (06)	44	6.8%	15.9%	2.3%	6.8%	13.6%	0.0%	0.0%	0.0%	2.3%	9.1%	2.3%	0.0%	2.3%	0.0%	34.1%	4.5%
2020/02/10 (07)	66	6.1%	7.6%	0.0%	9.1%	6.1%	3.0%	1.5%	3.0%	4.5%	3.0%	6.1%	1.5%	1.5%	0.0%	36.4%	10.6%
2020/02/17 (08)	120	4.2%	6.7%	0.8%	4.2%	12.5%	5.8%	0.8%	4.2%	2.5%	5.8%	11.7%	0.8%	4.2%	0.0%	30.8%	5.0%
2020/02/24 (09)	140	2.9%	2.9%	2.1%	2.9%	5.0%	15.7%	0.0%	8.6%	2.1%	4.3%	20.7%	0.7%	2.9%	0.0%	21.4%	7.9%
2020/03/02 (10)	170	4.1%	5.9%	2.9%	2.9%	7.6%	13.5%	2.4%	5.3%	1.2%	3.5%	20.0%	1.2%	5.3%	0.6%	18.8%	4.7%
2020/03/09 (11)	169	5.3%	3.0%	6.5%	5.3%	6.5%	9.5%	3.0%	4.7%	1.8%	3.0%	20.7%	1.8%	5.3%	0.6%	18.3%	4.7%
2020/03/16 (12)	231	8.2%	5.2%	5.6%	4.8%	7.4%	6.1%	3.5%	6.5%	4.3%	3.0%	10.4%	1.7%	4.3%	1.7%	16.5%	10.8%
2020/03/23 (13)	363	9.4%	4.7%	6.1%	4.7%	6.9%	9.1%	7.2%	3.9%	6.3%	2.5%	11.0%	4.7%	3.3%	1.4%	11.6%	7.4%
2020/03/30 (14)	487	13.8%	3.1%	3.9%	4.5%	6.2%	7.2%	6.6%	4.9%	5.5%	3.9%	7.6%	3.5%	3.9%	1.6%	11.7%	12.1%
2020/04/06 (15)	632	12.5%	2.1%	7.3%	6.8%	7.1%	7.0%	6.0%	5.5%	6.6%	3.8%	8.4%	3.5%	3.2%	2.1%	7.8%	10.4%
2020/04/13 (16)	630	13.0%	2.9%	8.4%	5.4%	4.8%	7.3%	6.3%	4.3%	6.8%	5.4%	4.6%	3.7%	3.7%	3.3%	9.4%	10.8%
2020/04/20 (17)	601	12.5%	4.5%	7.3%	4.7%	6.8%	5.3%	7.2%	6.0%	6.3%	5.2%	7.3%	3.7%	4.5%	3.2%	5.3%	10.3%
2020/04/27 (18)	579	14.9%	5.7%	9.3%	3.5%	5.0%	8.1%	6.6%	7.4%	8.5%	3.3%	4.0%	2.2%	4.1%	1.4%	5.2%	10.9%
2020/05/04 (19)	755	12.6%	3.3%	8.2%	2.9%	4.2%	7.5%	10.2%	7.2%	6.2%	4.8%	5.3%	3.7%	7.2%	1.3%	6.0%	9.4%
2020/05/11 (20)	719	12.5%	3.5%	10.0%	2.9%	7.2%	10.6%	8.1%	5.0%	6.4%	5.1%	4.0%	3.1%	7.1%	2.1%	5.0%	7.4%

図 11　トピックの推移（週次）

感染拡大やゲノム解析が早く，その後，検出・検査方法や臨床などの話題が続く点では，既存の分析 [小柴 20a] でも示された疫学調査のステップ[4]．とほぼ同じ傾向を示していると言える．今回は，迅速査読を経て公開さ

[4] 疫学調査の基本ステップ，国立感染症研究所 https://www.niid.go.jp/niid/images/idsc/kikikanri/H28/13-7.pdf (ac-

れた論文もあること，そもそも一部のデータが共通していることもあって，時系列的な推移に大きな差はなかったと考えられる．

一方で，トピック自体については差も認められる．例えば，「社会・経済・政策」のトピックは，前回の研究では検出できていなかった．また，治療薬やワクチン開発に関するトピックも明確には検出できていなかった．前者に関しては SSRN という人文社会系に強い PPS を含めた結果，後者も chemRxiv を含めた結果と解釈できる．

ただし，治療薬やワクチン開発に関するトピックについては医療系が中心と想定できる既存の分析で検出できていても不思議はない．これについては，PPS は査読論文に比較して，現場で判明した治療結果や分析結果が短時間で（正誤や質は度外視で）報告する事が可能，といった特徴が関連していることも想定できる．

4　まとめ

本報では COVID-19 関連のプレプリントを対象に，自然言語処理を用いたエマージング領域の把握を試行した．

その結果，既存の分析と同様に，疫学調査のステップに合致する動向を得ることができた．それに加えて，通常の査読論文も含めて分析した先行研究では明確には検出ができていなかった，医薬・ワクチン開発に関するトピックを抽出することができた．

今回の試行により，プレプリントを利用したエマージング研究内容のメタ把握が実現できる可能性が示唆された．

今後，他のエマージング研究でも今回の方法が適応可能か，また，従来の引用数ベースの分析結果の先行指標として機能するのか，それとも，研究者の興味関心などを表す既存手法とは別の指標として機能するのか，といった点について検証を行いたい．

cessed: 2020-04-30) および 文献 [柳川 18]

参考文献

[Arthur07] Arthur, David and Vassilvitskii, Sergei：K-means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp,1027–1035, 2007. http://dl.acm.org/citation.cfm?id=1283383.1283494

[Bojanowski17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017. arXiv:1607.04606

[Joulin16] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T.: FastText.zip: Compressing text classification models, *arXiv preprint*, 2016. arXiv:1612.03651

[McInnes18] Leland McInnes, John Healy and James Melville：UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint,* 2018. arXiv:1802.03426

[Okamura19] Okamura, K. ：Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Commun*, vol.5, no.141. 2019 https://doi.org/10.1057/s41599-019-0352-4

[Small85] Small, H; Sweeney, E; Greenlee E.：Clustering the science citation index using co-citations. II. Mapping science. *Scientometrics*, vol. 8, no. 5-6, p. 321-340. 1985

[伊神 09] 伊神正貫, 阪彩香 ：サイエンスマップによる科学研究の動的変化の観測　手法と応用. **情報管理**, vol.52, no. 5, p. 255-266. 2009 https://doi.org/10.1241/johokanri.52.255

[小柴 19] 小柴 等, 池内 健太, 元橋 一之：日米の特許データと論文データを用いた Mapping Patents の試行. **人工知能学会「社会における *AI* 研究会」** Vol.35, No.8, pp.1–8, Nov 2019. http://id.nii.ac.jp/1004/00010441/

[小柴 20a] 小柴 等, 伊神 正貫, 伊藤 裕子, 林 和弘, 重茂 浩美：COVID-19 / SARS-CoV-2 に関する研究の概況. *Discussion Paper,* DP181, NISTEP, May 2020. http://doi.org/10.15108/dp181

[小柴 20b] 小柴 等, 林 和弘：COVID-19/SARS-CoV-2 関連のプレプリントに関する分散表現データセット—2020 年 05 月 17 日版—. **データ・資料**, NISTEP, June 2020. http://doi.org/10.15108/data_covid19_2020_5

[治部 12] 治部眞里, 松邑勝治, 斉藤隆行 ：J-GLOBAL foresight の構築について. **情報管理**, vol. 54, no.10, p. 639-651. 2012 https://doi.org/10.1241/johokanri.54.639

[林 20] 林 和弘：MedRxiv, ChemRxiv にみるプレプリントファーストへの変化の兆しと オープンサイエンス時代の研究論文. *STI Horizon 2020* **春号**, Vol.6, No.1, Mar 2020. https://doi.org/10.15108/stih.00205

[文科 19] 文部科学省：「海外の最新科学技術動向に係る新興・融合領域に関する調査分析業務」業務成果報告書. **平成 *30* 年度科学技術調査資料作成委託事業**, 2019. https://www.mext.go.jp/a_menu/kagaku/kihon/1404334.htm

[柳川 18] 柳川 洋：臨床研究と疫学. **月刊地域医学**, Vol.32, No.9, pp.804(54) – 812(64), 2018.

DISCUSSION PAPER No.186

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

2020 年 06 月

文部科学省 科学技術・学術政策研究所
小柴 等, 林 和弘, 伊藤 裕子

〒100-0013 東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-3581-2391 FAX: 03-3503-3996

A Trial of early detection system for research trends through the preprints data
— Research status around COVID-19 / SARS-CoV-2

June 2020

KOSHIBA Hitoshi, HAYASHI Kazuhiro, ITO Yuko
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

http://doi.org/10.15108/dp186

https://www.nistep.go.jp

NATIONAL
INSTITUTE OF
SCIENCE AND
TECHNOLOGY
POLICY