DISCUSSION PAPER No.192

論文・特許のテキストデータを使った科学と技術の連関分析

New indicator of science and technology inter-relationship by using text information of research articles and patents in Japan

2021 年 02 月

文部科学省　科学技術・学術政策研究所

第 2 調査研究グループ

元橋　一之　　小柴　等　　池内 健太

【執筆者】

元橋 一之　第 1 研究グループ 客員研究官
文部科学省科学技術・学術政策研究所

小柴 等　第 2 調査研究グループ 上席研究官
文部科学省科学技術・学術政策研究所

池内 健太　第 1 研究グループ 客員研究官
文部科学省科学技術・学術政策研究所

【Authors】

MOTOHASHI Kazuyuki　Affiliated Fellow / 1st Theory-oriented Research Group,
National Institute of Science and Technology Policy (NISTEP), MEXT

KOSHIBA Hitoshi　Senior Research Fellow / 2nd Policy-oriented Research Group,
National Institute of Science and Technology Policy (NISTEP), MEXT

IKEUCHI Kenta　Affiliated Fellow / 1st Theory-oriented Research Group,
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.

元橋 一之・小柴 等・池内 健太（2021）「論文・特許のテキストデータを使った科学と技術の連関分析」，*NISTEP DISCUSSION PAPER*, No.192, 文部科学省科学技術・学術政策研究所
DOI: https://doi.org/10.15108/dp192

MOTOHASHI Kazuyuki, KOSHIBA Hitoshi and IKEUCHI Kenta (2019) "New indicator of science and technology inter-relationship by using text information of research articles and patents in Japan," *NISTEP DISCUSSION PAPER*, No.192, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/dp192

論文・特許のテキストデータを使った科学と技術の連関分析

文部科学省 科学技術・学術政策研究所
第 2 調査研究グループ

要旨

本稿においては、 1990 年以降に出版された日本の著者による学術論文（約 230 万件）と日本特許庁に対する出願特許（約 1200 万件）のタイトル・要旨のテキストデータを用いて、科学(論文)と技術（特許）の相互連関関係について分析を行った。具体的には、それぞれの文献のタイトルと要約文を用いた分散表現ベクトルを作成し、コサイン類似度を用いた近傍文書の抽出し、論文の近傍特許数と特許の近傍特許数のトレンドや分野別特性を明らかにした。その結果、1990 年代、2000 年代、2010 年代と時代が新しくなるにつれて、論文の近傍特許数は減少し、特許の近傍論文数は上昇するトレンドが見られた。これは、全体として、科学的なフロンティアの拡大が先に進み、技術的な進展が科学的な知見が多い分野をフォローする動きを表していると解釈できる。特許の非特許（論文）引用情報から、科学集約度の高い技術領域の抽出は行われてきているものの、本稿のアプローチによって、この科学→技術の関係に加えて、技術→科学（技術応用可能性が高い論文の学術領域の特定）の双方向の連関分析が可能となることを示した。

New indicator of science and technology inter-relationship by using text information of research articles and patents in Japan

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

ABSTRACT

In this study, the text information of academic papers (about 2.3 million) published by Japanese authors and patents filed with the Japan Patent Office (about 12 million) since 1990) are used for analyzing the inter-relationship between science and technology. Specifically, a distributed representation vector using the title and abstract of each document is created, then neighboring documents to each are extracted using cosine similarity. A time trend and sector specific linkage of science and technology are identified by using the count of neighbor patents (papers) for each paper (patent). It is found that the number of patents in the vicinity of papers decreased over time while the number of papers in the vicinity of patents increased. This can be interpreted as an advance the expansion of the scientific frontier by papers come first, then the technological progress (by patents) follows in the fields with substantial scientific knowledge already existed. The science intensity of technology has been measured by non-patent literature citation by patent. However, the citation information does not give the information of technology’s impact on science. This study shows that our methodology enables both way interlinkage of science and technology.

目次

1. はじめに

イノベーションにおける科学的知見の重要性の高まりが多くの産業でみられるようになっている。科学集約度（サイエンスリンケージ）の高い産業の代表といえる医薬品産業においては，ゲノムサイエンスの進展によって新薬開発プロセスにおける科学の重要性はますます高まっている。電子デバイス産業においては LSI 生産プロセスの微細化が進む中で，ナノスケールの物材特性に関する理解が必要不可欠になった。また，最近ではビッグデータを用いた機械学習（いわゆる AI）の進展によって，製造業のみならず，金融・サービス業を含めた様々な分野においてイノベーションプロセスが進化している。この分野においても重要な役割を担うのは大学や公的研究機関におけるサイエンティストである（Motohashi, 2019）。

これまで，イノベーションとサイエンスの近接性については，特許の非特許文献引用によって計測されてきた（Narin and Norma, 1985; Schmoch 1997）。非特許文献引用は特許として出願された発明が，当該特許が引用している科学的論文における知見をどの程度用いてなされたかを示す指標と考えられ，特許の科学集約度（サイエンスリンケージ）と呼ばれている。ただし，この指標は科学→発明の関係を示したものであり，科学とイノベーション相互の連関関係を示すものではない。文献引用（引用情報）を用いたサイエンスリンケージのアナロジーとして，論文が引用する特許の情報を用いるということが考えられる。しかし，科学論文に求められる引用の性質は，発明の新規性要因を問うという特許における引用とは異なる。すなわち，科学論文においては科学的発展のベースとなる引用文献についても，客観性や再現性といった科学的知見の要件を満たしている科学論文が主として用いられ，新規性があり産業応用可能性があればその原理は問わない特許を引用することは稀である。従って，引用情報によって，発明→科学の関係を，導き出すことはできない。

特許と科学論文の近接性を示すもう一つのアプローチが，同じ知見・発明を表現した論文と特許のペアを探す手法である。これには，同時に発表された特許と論文を抽出する方法（Lissoni et. al, 2013）やテキストマイニングを使って内容の似ている特許と論文を特定する方法（Magerman et. al, 2015）などがあり，アカデミックインベンターの分析（例えば，特許が論文生産性に与える影響）に用いられている。また，論文著者と特許出願人を接続した大規模なデータベースを作成し，同一研究者による特許と論文のペア情報を用いた科学技術集約度の測定を行った研究成果も存在する（Ikeuchi et. al, 2015）。

本研究は，テキストマイニングによって内容の近い論文と特許のペアを抽出する後者のアプローチをとり，日本の科学技術の進展と両者の関係に関する俯瞰的な分析を行った。具体的には，1990 年～2018 年に公表された日本の著者・発明者による論文と特許のタイトル・要約を用いて，内容の類似性が高いものをグルーピングし，科学と技術の進展において，論文→特許と特許→論文の相互の関係性を明らかにした。以下，第 2 章においては分析の手法について述べ，第 3 章においてはデータの概要及びクラスター分析の結果を示す。第 4 章においては論文・特許の引用情報を用いて，本論文で用いるテキストマイニングによる類似性指標の評価を行い，第 5 章においては科学と技術の連関指標を提示して，日本における両者の関係のトレンドを示す。最後に結論と今後の検討課題

について述べる。

2. 分析手法

分析手法は以下のとおりである。

- 分析対象とする論文，特許のタイトル，アブストラクト（英文）から抽出した文書群に対して Facebook 社が開発・公開している FastText (Joulin, 2016; Bojanowki, 2017）を用い，単語の分散表現（300 次元のベクトル表現）を作成。
- 上記の“単語の分散表現”をもちい，論文，特許ごとの文書情報（タイトル，アブストラクト）から“文書の分散表現”（単語分散表現を線形加算して単位ベクトル化したもの）を算出。
- 文書分散表現に対して K-Means++法(Arthur, 2007)を用いたクラスタリングを実施し，クラスタごとの頻出語で Word Cloud を作成。
  - UMAP(McIness, 2018)による次元圧縮（300 次元→2 次元）を用いた分散表現空間の 2 次元可視化
- 文書間の類似度(ベクトルのコサイン類似度)による近傍文書（それぞれの文書に対する近傍 200 文書）の抽出。
  - 近傍文書の取得については高次元ベクトル近傍探索 NGT（Neighborhood Graph and Tree,：岩崎，2013）を用い処理を高速化
- 上記の近傍文書情報を用いた学術分野（論文）と技術分類(特許)のコンコーダンステーブルの作成，科学と技術の連関分析

なお，近傍文書の抽出までの手法は，特許文書情報を用いて発明内容の抽出を行った研究成果（元橋・小柴・池内，2019）を踏襲している。また，コンコーダンステーブルと連関分析については，第 5 章で詳しく述べる。

3. データセットと記述統計

3.1. データセット

本件研究で用いたデータは以下のとおりである。

- 論文情報：Clarivate 社の Web of Science における SCIE(Science Citation Index Expanded)収録論文について，1990 年～2017 年までに出版されたものでかつ日本を所在地とする著者が一人以上含まれているもの。
- 特許情報：PATSTAT2020 Spring Version に含まれている日本特許庁に出願された特許（英語の翻訳された発明の名称と要約情報が入手可能なもの）

文書件数としては，論文 2,342,987 件，特許 12,037,068 件，合計 14,380,055 件である。図 1 に出版年(特許については出願年)別のそれぞれの件数の推移を示した。特許件数は 2000 年以降減少傾向にあり，論文については 10 万件程度で安定的に推移している。

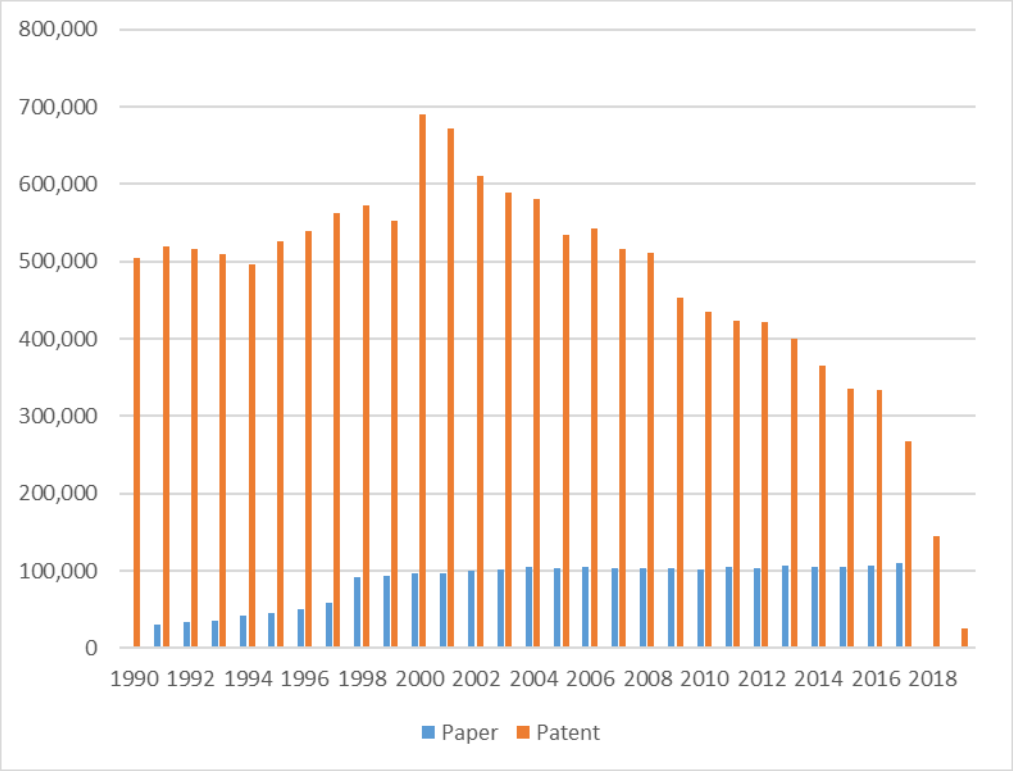


図 1：分析で用いた論文と特許の出版年(出願年)別推移

3.2. 分散表現の結果とクラスター分析

すでに述べたとおり，ここではまず単語の分散表現を作成し，それらを用いて文書分散表現を作成するという，いわゆる SWEM (Simple Word-Embedding-based Methods)-aver (Shen, 2018) を採用している。単語については，論文・特許の合計約 1430 万件のタイトルとアブストラクトに出現する単語を取り出し，レマタイズや stop words, common words, rare words の除去，登場回数の低い語の除去など一通りの下処理を行った上で得られた，合

計 258,459 単語について分散表現（単語分散表現）を算出した。なお，分散表現獲得には FastText を用いた。また，その結果については，K-means++法によるクラスタリング分析を行い，意味的に似ている単語が同一クラスターに属していることについて目視チェックを行った。

この単語分散表現の結果を SWEM-aver 形式で文書毎に集計した文書分散表現に対しても単語分散表現と同様に K-means++法によるクラスタリング（16 クラスターの分類）を行った。この 300 次元の文書分散表現について，次元圧縮手法である UMAP を用い 2 次元に圧縮した技術マップを図 2 に示す。なお，それぞれのクラスターの内容については，別添 1 のワードクラウドの結果を参照されたい。

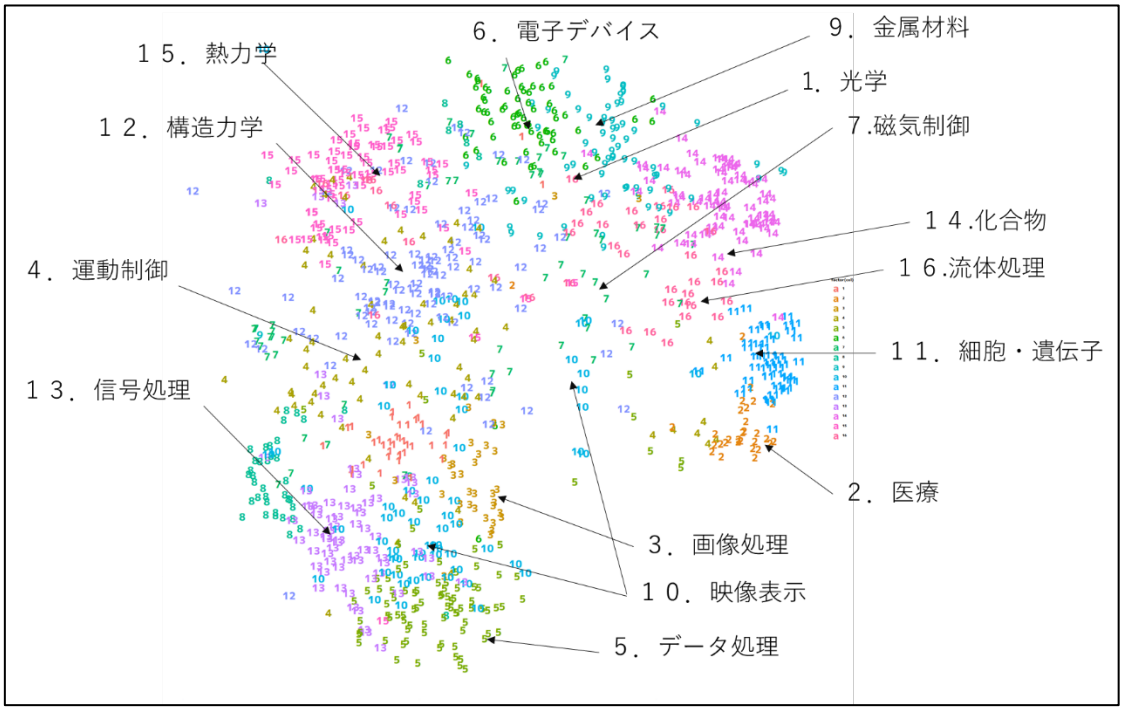


図 2：クラスター分析結果の可視化（UMAP を用いた 2 次元圧縮結果）

3.3.　論文と特許の関係

図 2 は論文と特許の両方を併せた技術マッピングの結果であるが，この両者を識別して，時系列的な変化を見たものが図 3－1～3－3 である。それぞれの図において赤が特許，青が論文の位置を示している。

全体的には，論文の割合が多いのが，ライフサイエンス関係（細胞・遺伝子，医療），化学・材料関係（化合物，金属材料）であり，光学，流体処理，映像表示関係にも論文が分布していることが分かる。一方で，運動制御，構造力学，熱力学などの機械関係，電子デバイス，画像処理関係はほとんど特許文書で占められている。

時系列的な変化については，特に 1990 年代と 2000 年代の間に違いが見られる。文書全体の占める論文の数が大きくなっていることから，論文において，技術分野に広がりが見られるが，その傾向は化学・材料分野（化合物，金属材料）において顕著であると見受け

られる。また，特許のみであった熱力学などの分野においても論文が出てきていることが読み取れる。この分野別の論文と特許の関係については第 5 章で詳しく分析する。

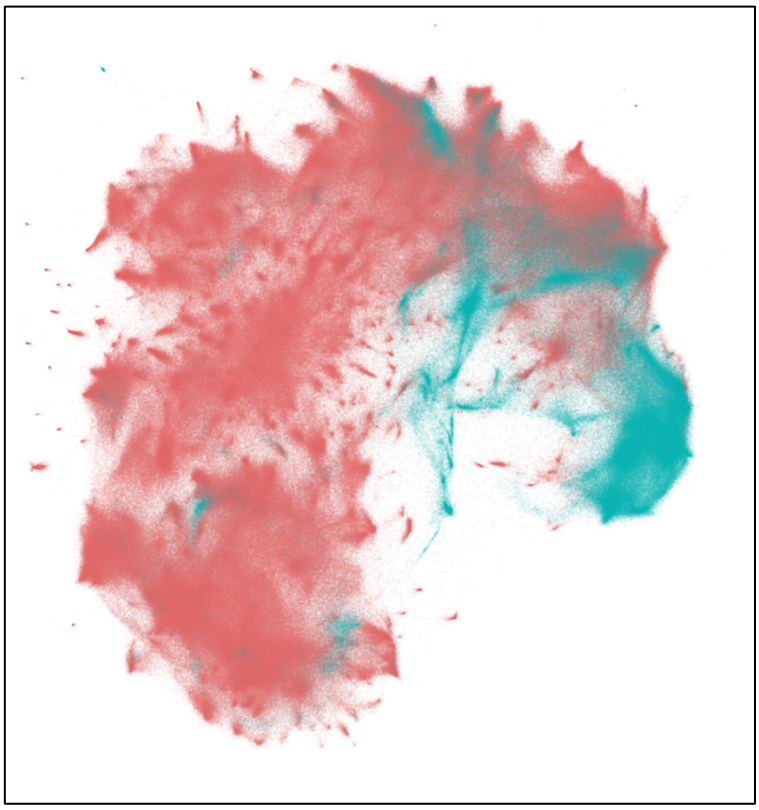


図 3－1：論文・特許の技術マッピング(1990 年代)

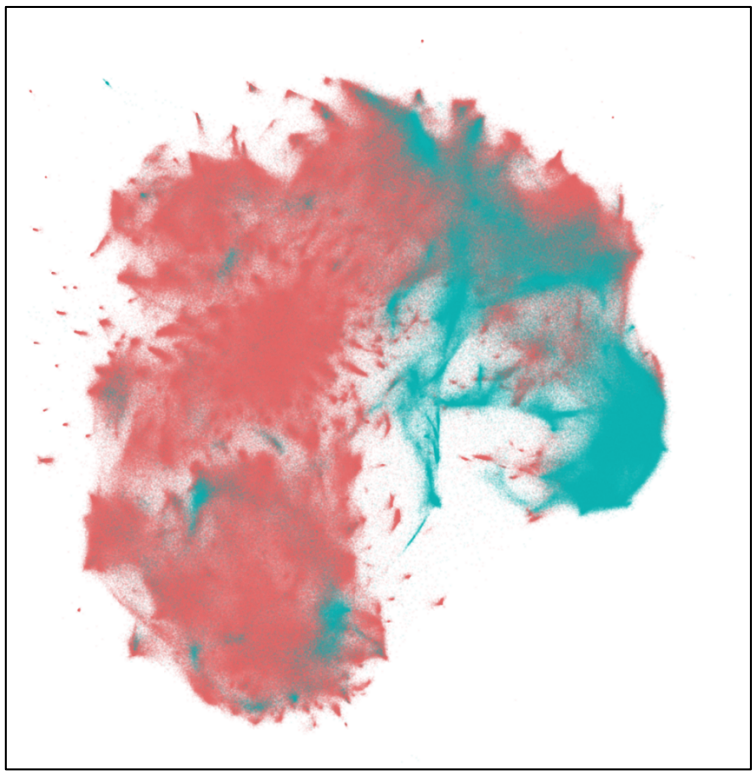


図 3－2：論文・特許の技術マッピング(2000 年代)

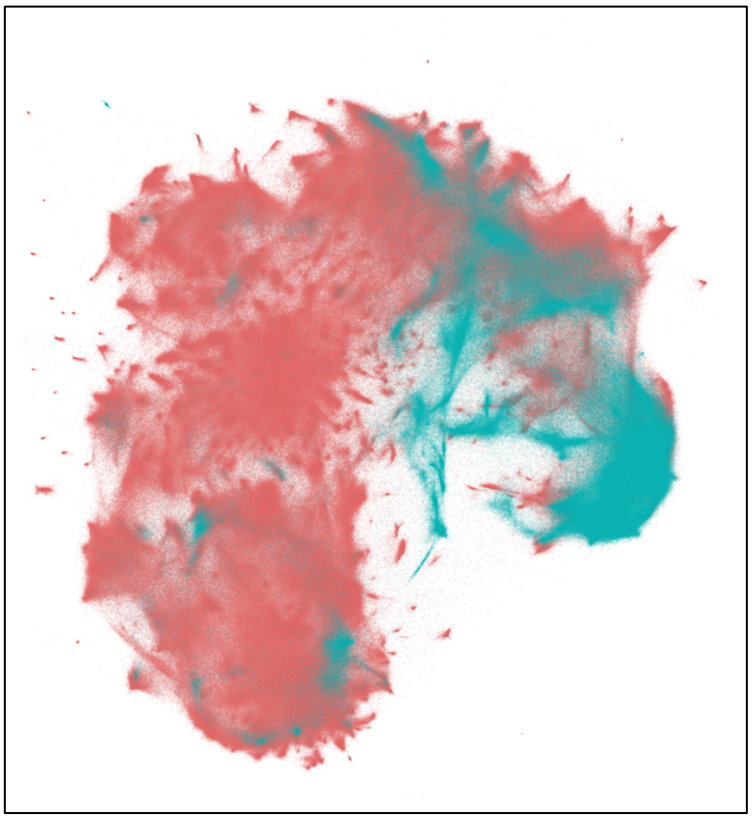


図 3－3：論文・特許の技術マッピング(2010 年代)

4. 文書分散表現データの特性

ここでは論文，特許のテキスト情報から作成した文書分散表現の特性に関する分析を行う。第 3 章で示したように単語分散表現についてはクラスタリング分析の結果を用いて，目視による内容の評価（似ている単語が同一クラスターに集まっていることの確認）を行った。ここではよりフォーマルな定量的な評価を試みる。具体的には，論文と特許の引用情報を用いて，引用ペア（引用文献と被引用文献のペア）間の類似度が有意に高くなることを確認する。また，日本学術振興会(JSPS)の科学研究費助成事業（学術研究助成基金助成金／科学研究費補助金，いわゆる科研費）成果報告の情報を用いて，同一科研費プロジェクトから生まれた論文と特許の類似性が高いことについても確認する。最後に高次元ベクトル近傍探索（NGT）による近傍文書の情報を用いて，技術スペースの分布状況によって文書間のコサイン類似度が影響をうけるかどうかについても検討する。[1]

まず，引用ペアは同一研究プロジェクトの成果間のコサイン類似度が有意に高いことを示すための対比サンプルとして，論文-論文，論文-特許，特許-特許の 3 つのパターンについて，ランダムに 10000 ペアを抽出して，それらのコサイン類似度の分布を見た。図 4 はそれぞれのペアに関する十分位の値をプロットしたものである。中央値（ミディアン, P50）を見ると，論文-論文の値が最も高く 0.73，その次が特許-特許（0.70），最後に論文-特許（0.69）となっている。また，10 パーセンタイル（P10）を見ても，それぞれ 0.6 程度となっており，ランダムに抽出したサンプル間のコサイン類似度は比較的狭い領域（10 パーセンタイルから 90 パーセンタイルの幅が 0.2 程度）に分布していることが分かる。これは各単語を独立次元とした一般的な単語ベクトルの代わりに，分散表現を用いているため，そもそも単語間にある程度の相関関係が存在することによる。ただし，日本語の特許文書を用いて同様の作業を行った結果（ランダムに抽出したコサイン類似度の中央値が約 0.5）と比べると（元橋・小柴・池内，2019），コサイン類似度が高くなっており，今回用いた英語の文献の前処理（ストップワードの除去，テクニカルタームの n-gram 化など）において改良の余地があることを示している。

[1] 当然ながら近傍とは距離の値が小さいものを指す。一方，類似度は値が大きいほど似ていることになるため概念・数値表現としては逆になる。ここでコサイン類似度は一般に 0 から 1 の範囲の値をとる。従って，距離として（1-コサイン類似度）を採用することで近傍を探索した。また NGT を通じて得られる結果は近似解であって，必ずしも正確な結果ではない。

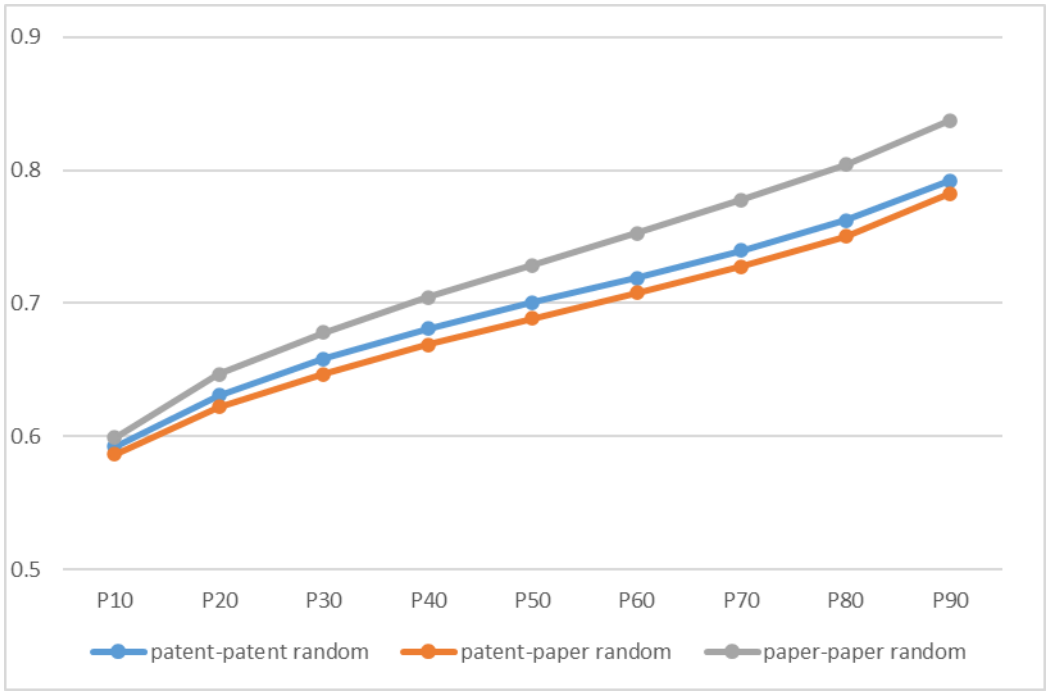


図 4：ランダムサンプルにおけるコサイン類似度分布(十分位値)

次に図 4 の分布と引用ペア及び科研費同一プロジェクト成果における文献間のコサイン類似度とを比べる。図 5－1，図 5－2 及び図 5－3 はそれぞれ，論文-論文，論文-特許，特許-特許のペアに関する状況を見たものである。すべてのペアにおいて，引用ペアと同一プロジェクト成果間のコサイン類似度はランダムペアと比較して高くなっており，文書分散表現の妥当性が確認された。

また，文献タイプによる違いについてみると，論文間の引用ペア，同一プロジェクトペアは均質性の高い情報提供している（10 パーセンタイル値でも 0.8 以上)。一方で，それ以外のペアにおいては 10 パーセンタイル値が 0.7 を切っているものもあり，ランダムペアの中央値よりも低い値となっている。また，引用ペアと同一プロジェクトペアの分布は，特許間のものを除いてほぼ同一の分布となっている。一方，特許間のペアについては，科研費同一ペアのバラつきが，引用ペアより大きくなっている。

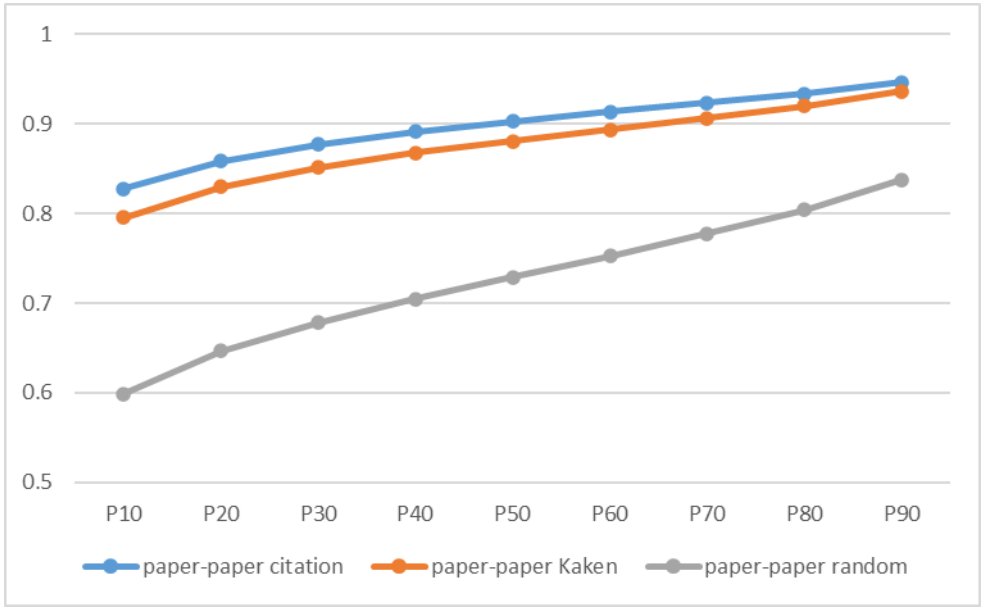


図 5－1：ランダムペア，引用ペア，同一プロジェクトペアの比較（論文間）

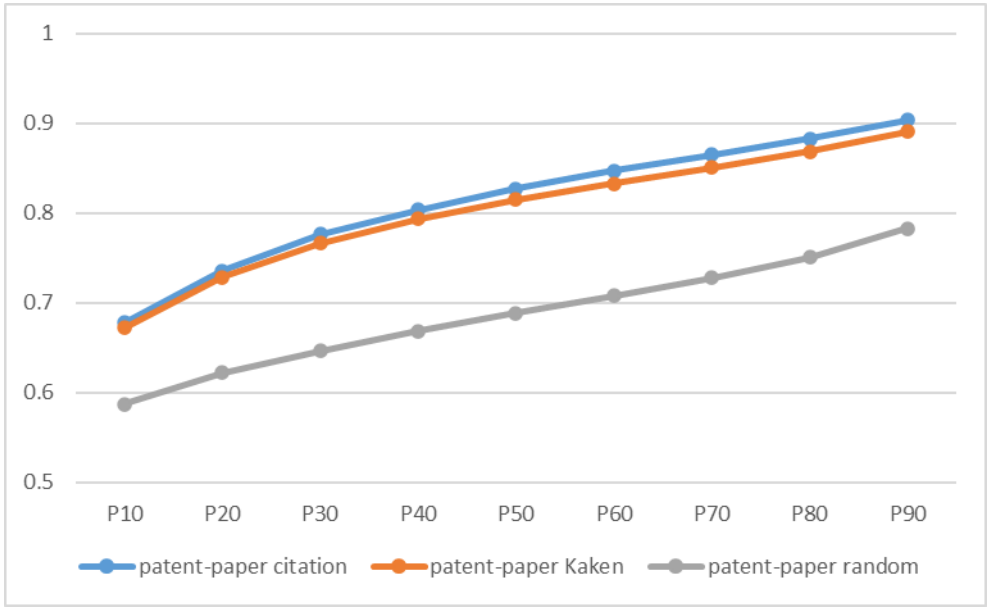


図 5－2：ランダムペア，引用ペア，同一プロジェクトペアの比較（論文・特許間）

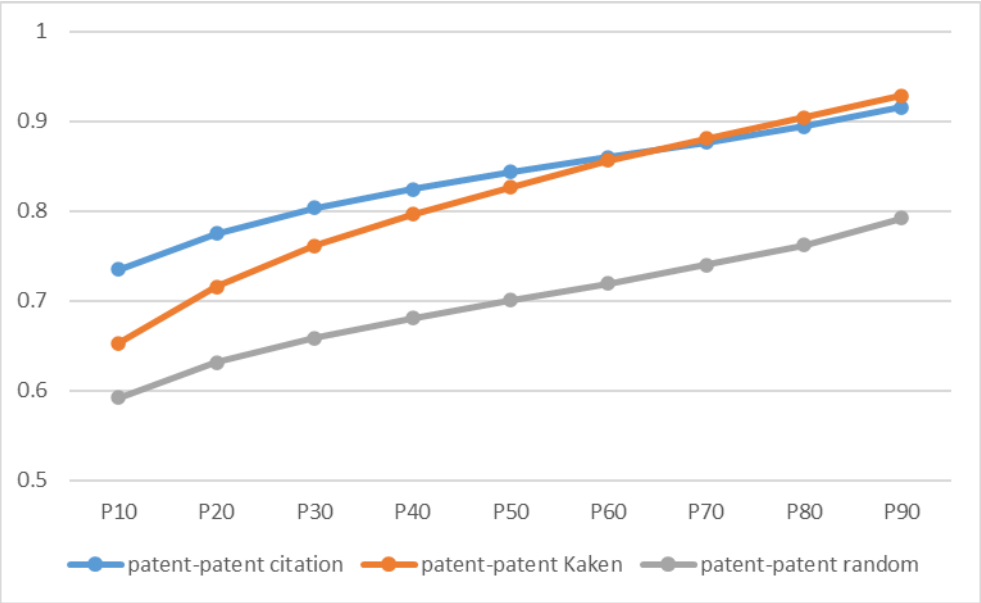


図 5－3：ランダムペア，引用ペア，同一プロジェクトペアの比較（特許間）

最後に NGT による近傍文書とのコサイン類似度の特性についてみる。NGT は数百～数千次元程度の高次元ベクトル空間において，任意のベクトルの近傍に存在するベクトルを近似的ながら効率的に探索するアルゴリズムである。距離関数には L2（ユークリッド）距離などいくつかの指標が選択できるが，今回はコサイン類似度により近傍 200 件の文書の抽出を行った。表 1 は，100 番目と 200 番目の文書とのコサイン類似度の分布を見たものである。まず，100 番目の文書と 200 番目の文書とのコサイン類似度はほとんど変わらないことが分かる（例えば，それぞれの中央値が 100 番目で 0.899，200 番目で 0.893）。これは文書分散表現のベクトルが 300 次元と次元数が高く，単位ベクトル化を行っていることによる（300 次元の超球体の半径と体積の関係）。また，200 番目の文書とのコサイン類似度の中央値である 0.9 は，引用ペアでみると論文間では 60 パーセンタイル，論文－特許で 90 パーセンタイル，特許間で 80 パーセンタイルの近接度に対応しており，内容的にかなり近いにある状況を示している。

	100th	200th
1%	0.843	0.834
5%	0.870	0.863
10%	0.881	0.875
25%	0.899	0.893
50%	0.916	0.911
75%	0.932	0.928
90%	0.944	0.941
95%	0.951	0.948
99%	0.961	0.958

表 1：100 番目，200 番目の近傍文書とのコサイン類似度の分布

なお，200 番目の近傍文書とのコサイン類似度の違いは，論文・特許が分布している技術空間における分布密度の違いによるものである。200 番目の文献のコサイン類似度が大きな文書は，その周辺により密に論文・特許が分布していることを示している。この技術空間密度の状況によって，引用ペアのコサイン類似度も影響を受けることが考えられる。技術空間密度が高いところに位置する文献は，よりコサイン類似度が高い文献が引用される可能性が高いからである。

図 6 は，近傍 200 番目の文献とコサイン類似度によって，全体を 4 つのグループ（コサイン類似度が低いもの，つまり技術空間密度が疎である文献から Q1 から Q4 まで）に分けて，それぞれのグループにおける文献の引用ペアとのコサイン類似度の分布（十分位値）を見たものである。仮説どおり技術空間における密度が高い場所に位置する文献（例えば Q4）については，引用文献とのコサイン類似度も高くなっている。なお，空間密度の影響は，密度が疎であるグループ（例えば Q1）においてより大きくなっている。

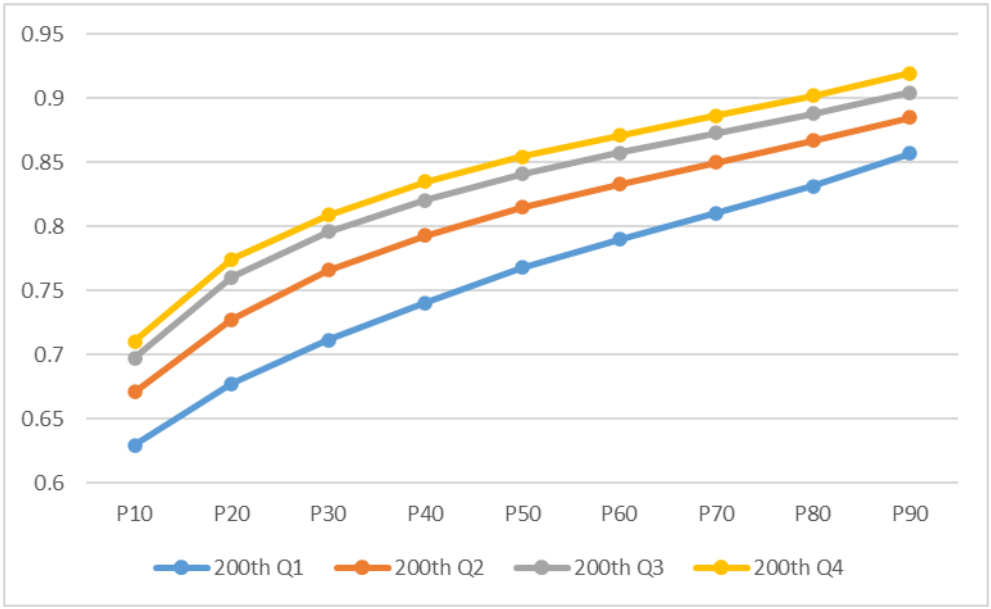


図 6：技術空間密度と引用ペアのコサイン類似度

5. 近傍文書情報による科学と技術の相関分析

特許と論文のそれぞれについて，技術空間の近傍に位置する論文や特許の数をカウントすることによって，科学(論文)と技術(特許)の相関関係を分析することが可能である。ここでは，論文の学術領域別（WOS における学術分類をベースとした 22 分類)，特許の技術分野別（WIPO における 35 分類）のそれぞれについて，平均近傍特許数を求めた。なお，近傍文書については，上位 200 番目以内かつコサイン類似度が 0.9 以上のものを選択した。なお，特許と論文の文書分散表現に基づくコサイン類似度が 0.9 以上というのは，特許-論文の引用関係ペア，あるいは同一科研プロジェクトの成果としてのペアのコサイン類似度の上位 10%タイルの値である（図 5－2 の P90)。つまり内容的に類似度がかなり高いものを抽出していることになる。また，表 1 で見た通り，200 番目の近傍文書のコサイン類似度が 0.9 以上のものは，全体の半数以上存在する（200 番目近傍文書のコサイン類似度中央値は 0.911)。これらの特許は比較的技術空間における文書密度が高いところに位置していると考えられる。NGT を実行するためには近傍文書数（近傍何件までを取得するか）をあらかじめ決めておく必要があり，今回の研究においては 200 番目までの近傍文書を抽出した。従って，データ制約というプラクティカルな理由によって，200 番目でうちきることとしたが，コサイン類似度が 0.9 以上の近傍文書をすべて取り上げるとするとその文書数が膨大になるものが存在しうる。全体に有意な影響を与える外れ値を取り除くためにも，1 つの文書に対する近傍文書数に閾値を設けることはいずれにしても必要と考える。

まず，図 7 に科学と技術の相関関係に関する全体的なトレンドを示す。このグラフは，論文（特許）の出版年（出願年）によってサンプルを 3 分割（1990 年代，2000 年代，2010 年代）し，それぞれの論文（特許）の平均特許（論文）数を技術分野（学術領域）別に集計し，更に，その分野間の平均値をとったものである。特許から見ると近傍論文の数が上昇している（赤線）のに対して，論文から見ると近傍特許の数は低下している（青線)。科学（論文）の近傍特許数の減少は，最近の論文になるほど周辺に産業応用性のある特許が出願されにくい分野に出版されている傾向を示す。一方で特許から見ると科学論が出版されている領域の出願が増えている。つまり，全体として，科学論文が技術スペースのフロンティアを開拓し，技術（特許）がそれらの科学をベースに発展してきている様相を反映したものと考えられる。

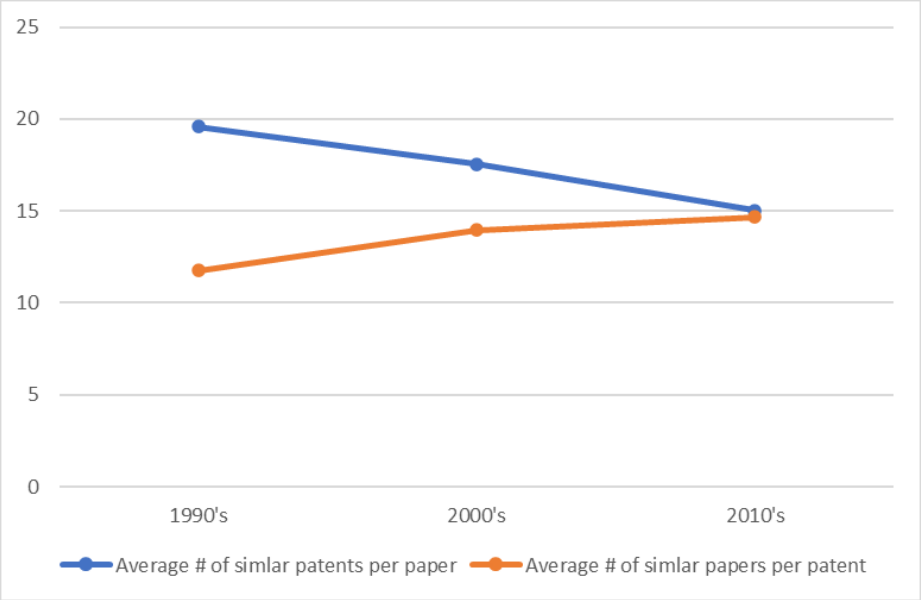


図 7：近傍特許（論文）数の推移

次にこの動向を論文の学術領域別，特許の技術分野別に見たものを，図 8－1，図 8－2 にそれぞれ示す。

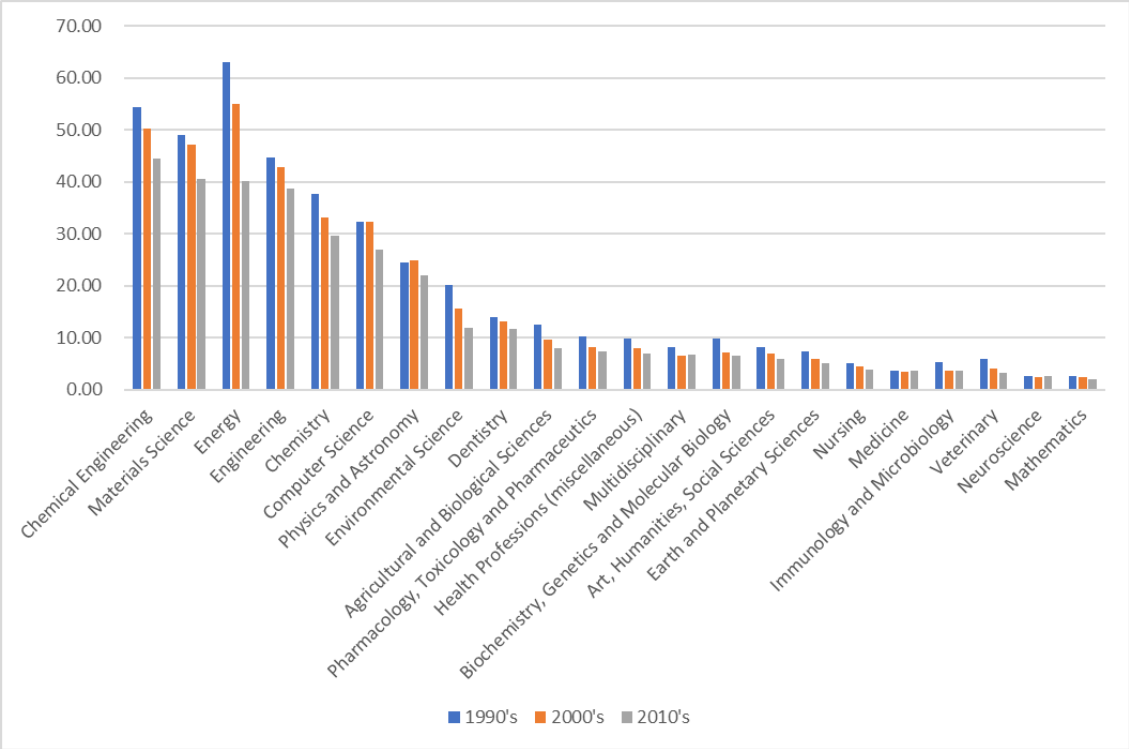


図 8－1：学術領域別に見た論文の近傍特許数

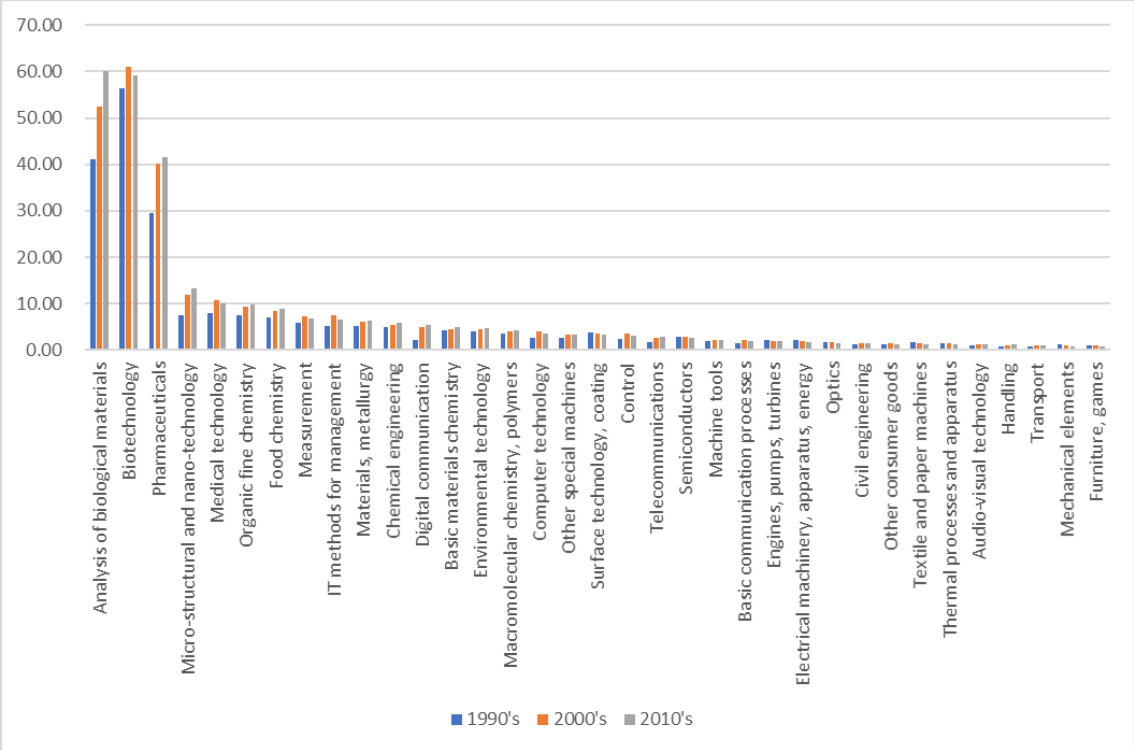


図 8－1：技術分野別に見た特許の近傍論文数

近傍特許数の多い学術領域としては，化学工学，材料工学，エネルギー，エンジニアリング，化学，コンピュータ科学などとなっているが，近傍特許数の減少はすべての学術領域でみられる。つまり，科学論文が技術スペースにおけるフロンティアを切り開いていく様相は学術領域を超えた一般的な現象としてとらえることができる。

一方で，近傍論文数の多い技術分野としては，生物材料，バイオテクノロジー，医薬品が突出している。なお，これらの技術領域は，引用データによるサイエンスリンケージが高い分野としても認識されている（Schmoch, 1997）。トレンドを見ると全体として近傍の科学論文数の増加が見られるが，減少傾向にある分野も見受けられる。

近傍文書が対象となる文書の前に公表されたものである場合は，当該文書は比較する文書をベースとして生まれたものであり，逆に比較する近傍文書が後に公表されたものだとすると，当該文書がその近傍文書の影響を与えた，と解釈することができる。図 9－1 と図 9－2 は，この近傍文書における公表タイミングの前後のバランスを分野別に見たものである。図 9－1 を見ると，学術領域によっては AFTER が BEFORE より相対的に大きい領域（コンピュータ科学）と逆のパターンとなっている領域（材料工学，化学）が見られる。前者については，科学的進展が技術（特許）に影響をおよぼす傾向が強いもの，後者については，技術的進展が見られる分野において，更なる科学的発展が見られる傾向が強いものと解釈することができる。一方で，技術分野別の近傍論文数については，ほとんどの分野で BEFORE と AFTER のバランスが取れている状況にある。

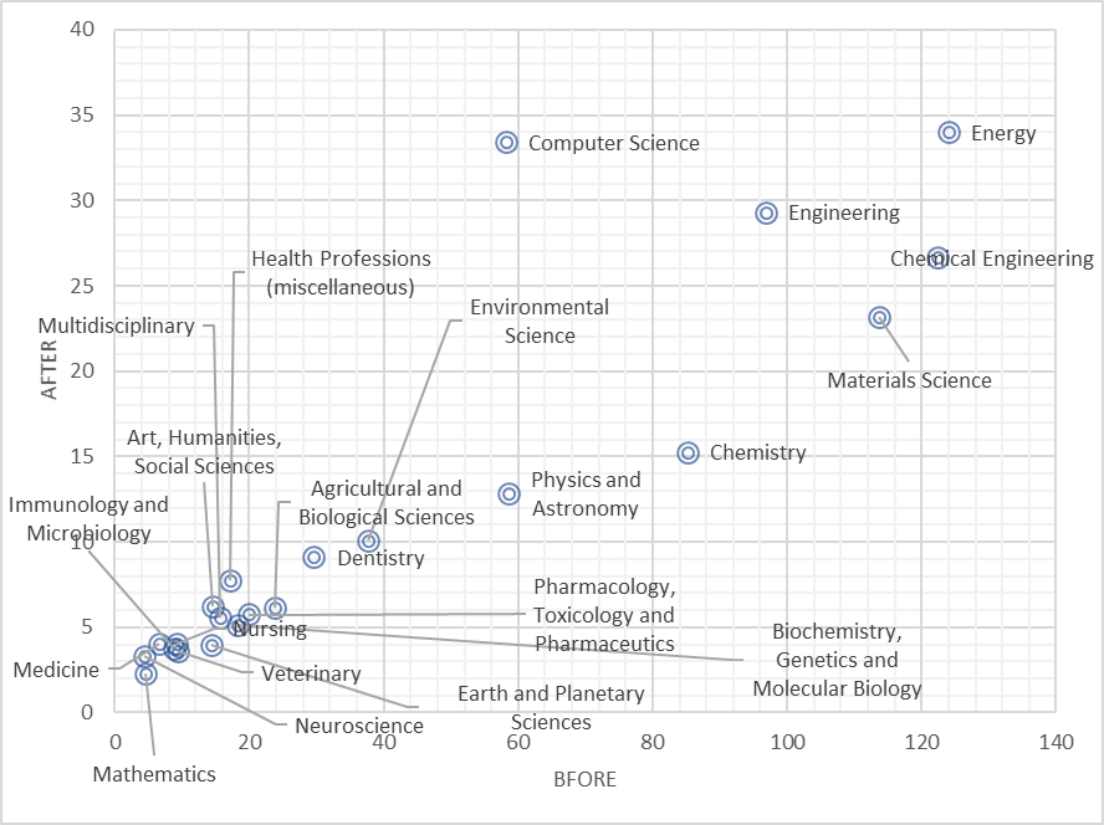


図 9－1：学術領域別，公表前特許数（BEFORE）と公表後特許数（AFTER）

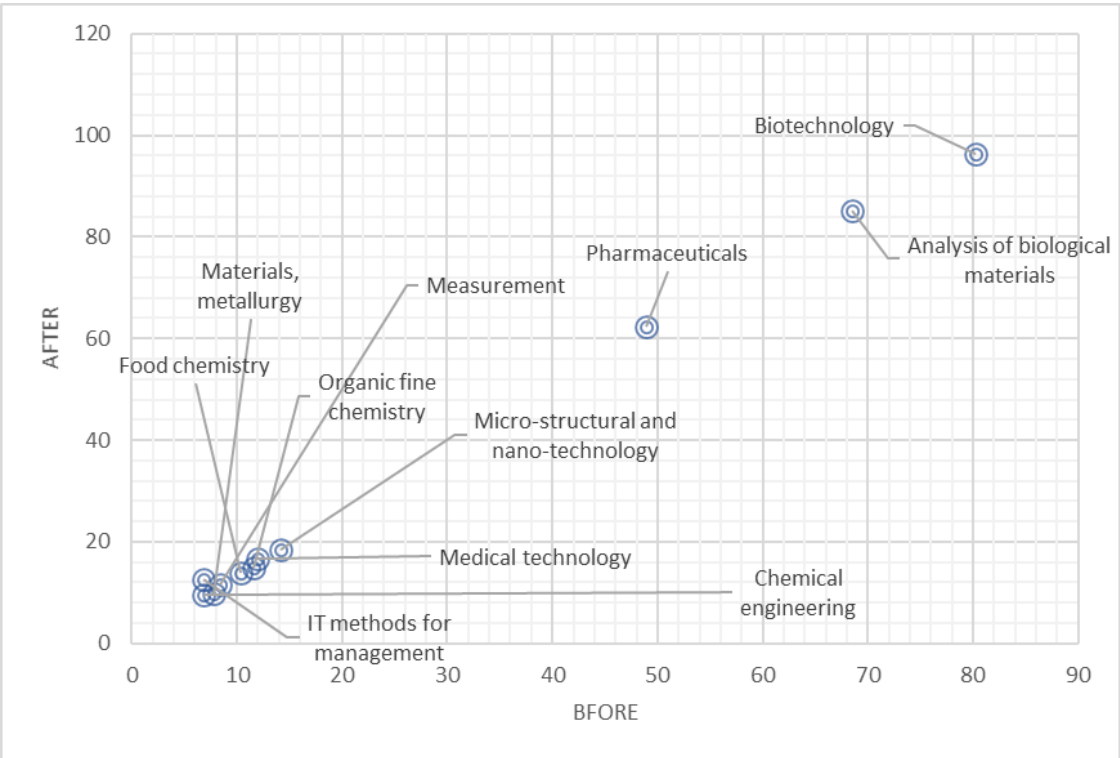


図 9－2：技術分野別，公表前論文数（BEFORE）と公表後論文数（AFTER）

6. まとめの今後の研究課題

本稿においては， 1990 年以降に出版された 230 万件の論文と出願された 1200 万件の特許のテキストデータ（タイトルおよび概要）を用いて，科学(論文)と技術（特許）の相互連関関係について分析を行った。具体的には，それぞれの文献のタイトルと要約文を用いた分散表現ベクトルを作成し，コサイン類似度を用いた近傍文書（類似度が高いものトップ 200 でかつコサイン類似度が 0.9 以上）の抽出を行った。論文と特許の連関については，論文（特許）それぞれの近傍特許(論文)の数で定量化した。

その結果，1990 年代，2000 年代，2010 年代と時代が新しくなるにつれて，論文の近傍特許数は減少し，特許の近傍論文数は上昇するトレンドが見られた。これは，全体として，科学的なフロンティアの拡大が先に進み，技術的な進展が科学的な知見が多い分野をフォローする動きを表していると解釈できる。また，技術と連関が深い科学の学術領域は，化学工学，材料工学，エネルギー，エンジニアリング，化学，コンピュータ科学などで，科学と連関が深い技術分野としては，生物材料，バイオテクノロジー，医薬品などであることが分かった。後者については，特許の非特許文献(論文)引用によって過去の文献においても明らかになっていたが，前者については既存研究にない新しい知見といえる。

今回の研究は，日本の科学技術の進展について俯瞰的なトレンドを示すことを目的としたものであるが，大量の論文・特許文献の分散表現情報は，今後様々な応用研究に生かすことが可能である。日本においては，2001 年に国立試験研究所の独立行政法人化，2004 年の国立大学の国立大学法人化といった大きな制度改革が 2000 年代に行われた。この制度改革が今回示した科学と技術の連関トレンドに対して影響を与えていることが予想される。論文著者や特許出願人の所属機関と両者の関連性について分析することで，これらの制度改革のインパクトについて有益な知見が期待できる。

また，最新の自然言語処理手法を取り入れることでより精度の高い内容表現を獲得し，科学技術のトラジェクトリーに応用する研究も今後の課題として有望である。今回の研究においては，単語単位の分散表現を得て，それをドキュメント単位に集計する BoW(Bag of Words)のアプローチをとっているが，近年急速に利用進んでいる BERT（Bi-directional Encoder Representation with Transformation）などの手法によると，単語の意味に加えて，文章中の単語のコンテクスト情報も分散表現の中に埋め込むことが可能である。例えば“核”という単語の分散表現について「原子」という単語が周囲に出てくるときと，「細胞」が出てくるときで異なるものにできる。大量なパラメーターを持つ深層学習モデルである BERT で分散表現を学習させるためには，膨大なコンピュータ資源を必要とするが，近年，Google チームによって世界の特許文献をベースとした BERT の学習モデルが公開された（Srebrovic and Yonamine, 2020）。ここでの推計結果をベースとして，内容の近接性の詳細に踏み込んだ研究についても今後検討していきたい。

参照文献

Arthur, D. and Vassilvitskii, S.: K-means++: The Advantages of Careful Seeding, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, pp. 1027–1035, Philadelphia, PA, USA (2007), Society for Industrial and Applied Mathematics

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.(2017), Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146, arXiv:1607.04606

Goto, A. and K. Motohashi (2007) "Construction of a Japanese Patent Database and a first look at Japanese patenting activities," Research Policy, 36(9), 1431-1442.

Ikeuchi, K. Motohashi, R. Tamura and N. Tsukada (2017), Measuring Science Intensity of Industry using Linked Dataset of Science, Technology and Industry, RIETI Discussion Paper, 17-E-056

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

Lissoni, F, F. Montabio and L. Zirulia (2013) "Inventorship and authorship as attribution rights: an enquiry into the economics of scientific credit," Journal of Economic Behavior and Organization, 95, 49-69.

Magerman, T., B.V. Looy and K. Debackere (2015) "Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology," Research Policy, 44(9), 1702-1713.

McInnes, L., Healy, J., and Melville, J (2018).: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv preprint (2018), arXiv:1802.03426

Motohashi K. (2019) "Science and Technology Co-evolution in AI: Empirical Understanding through a Linked Dataset of Scientific Articles and Patents;, RIETI Discussion Paper 20-E010

Narin, F. and E. Noma (1985) "Is technology becoming science?" Scientometrics, 7, 368-381.

Schmoch, U. (1997) "Indicators and relations between science and technology," Scientometrics, 38(1), 103-116.

Shen, D., Wang, G., Wang, W., Renqiang M., Su, Q., Zhang, Y., Li, C., Henao, R. and Carin, L. (2018), Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms, ACL

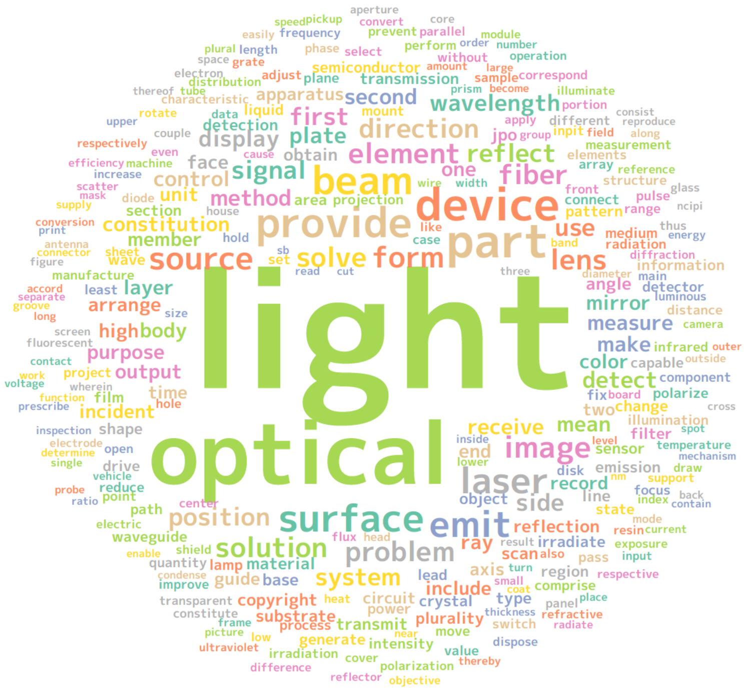
Srebrovic, R. and Yonamine J. (2020), Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery, November 2020, Google Cloud Blog, How AI improves patent analysis | Google Cloud Blog

元橋 一之 小柴 等 池内 健太 (2019), "特許文書情報を用いた発明内容の抽出と 出願人タイプ別特性比較 ", NISTEP Discussion Paper No. 175, 文部科学省科学技術・学術政策研究所

別添資料

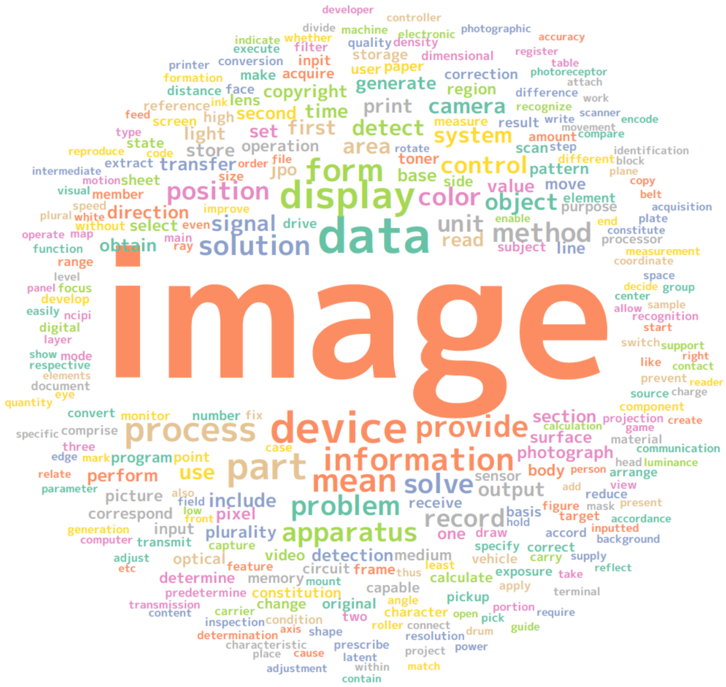
別添１：文書分散表現のクラスター毎の内容（ワードクラウド）

クラスター1：光学



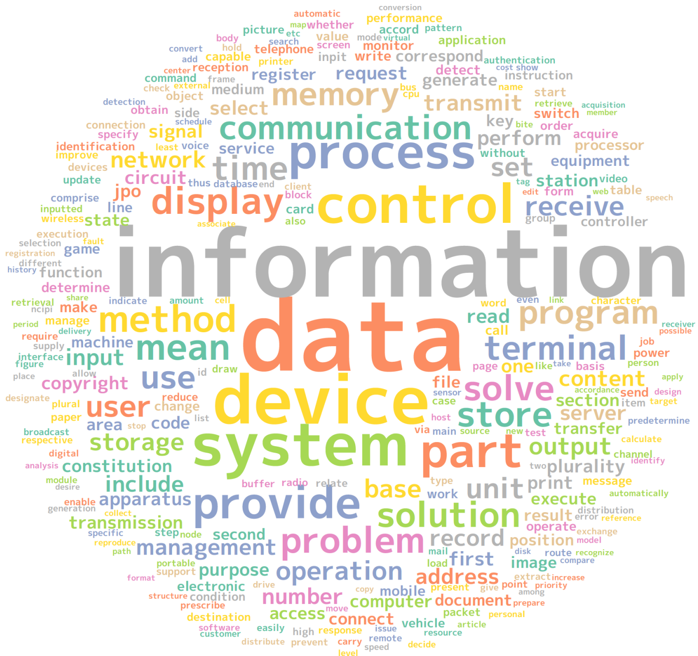
クラスター2：医療

クラスター3：画像処理

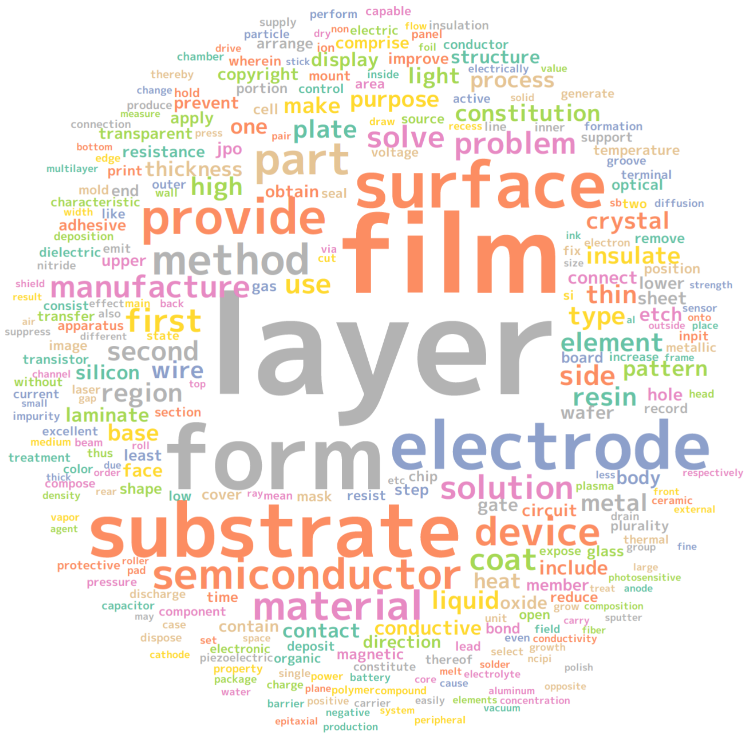


クラスター4：運転制御

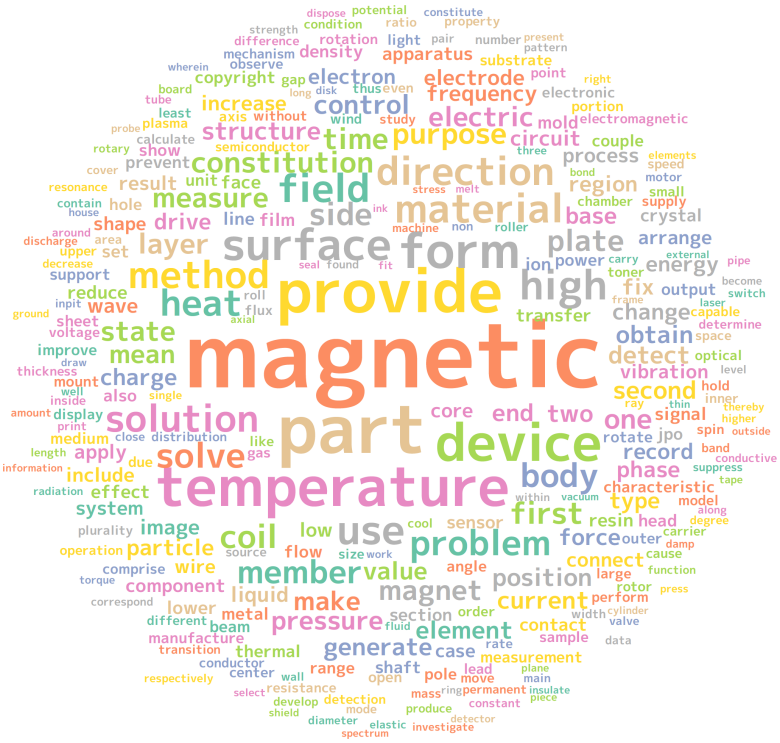
クラスター5：データ処理



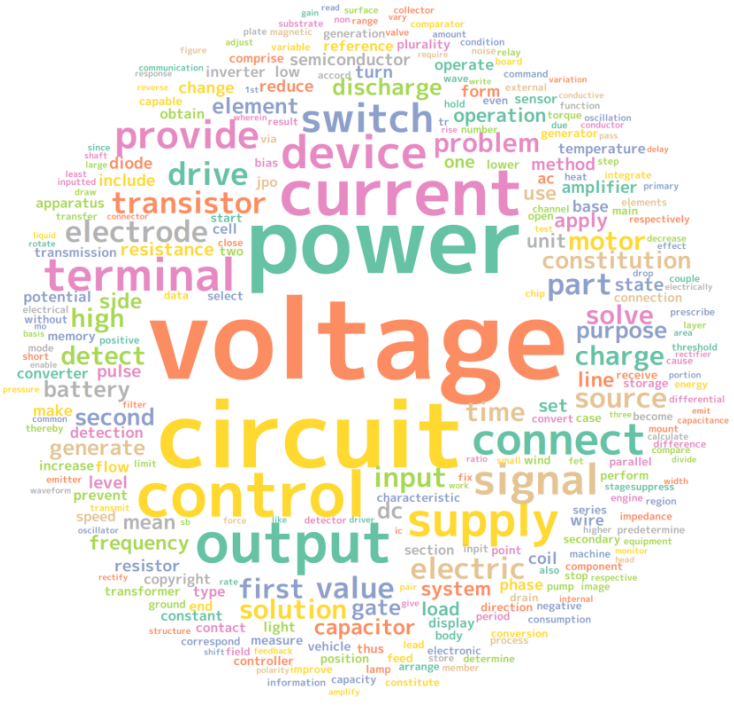
クラスター6：電子デバイス



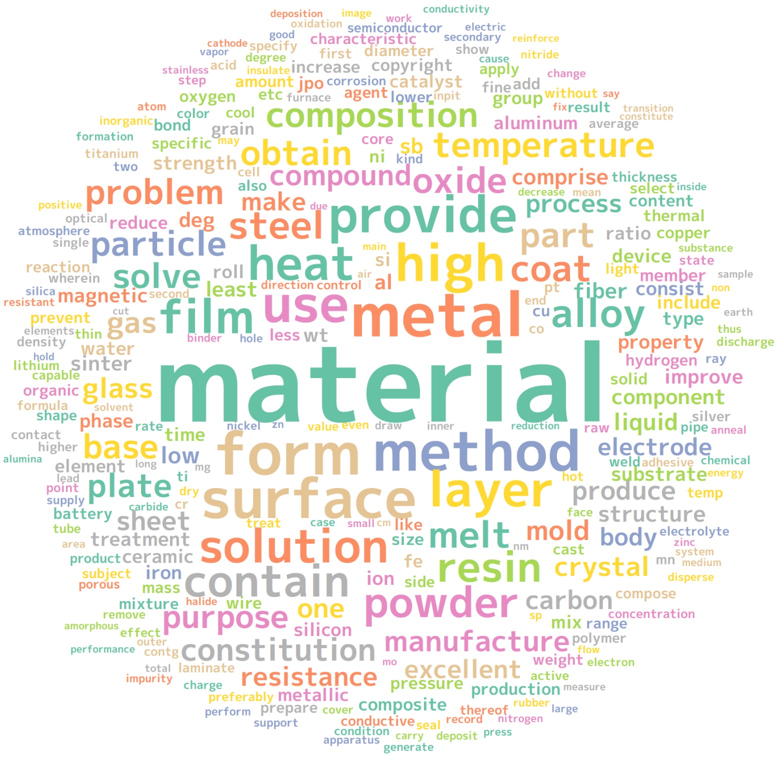
クラスター7：磁気制御



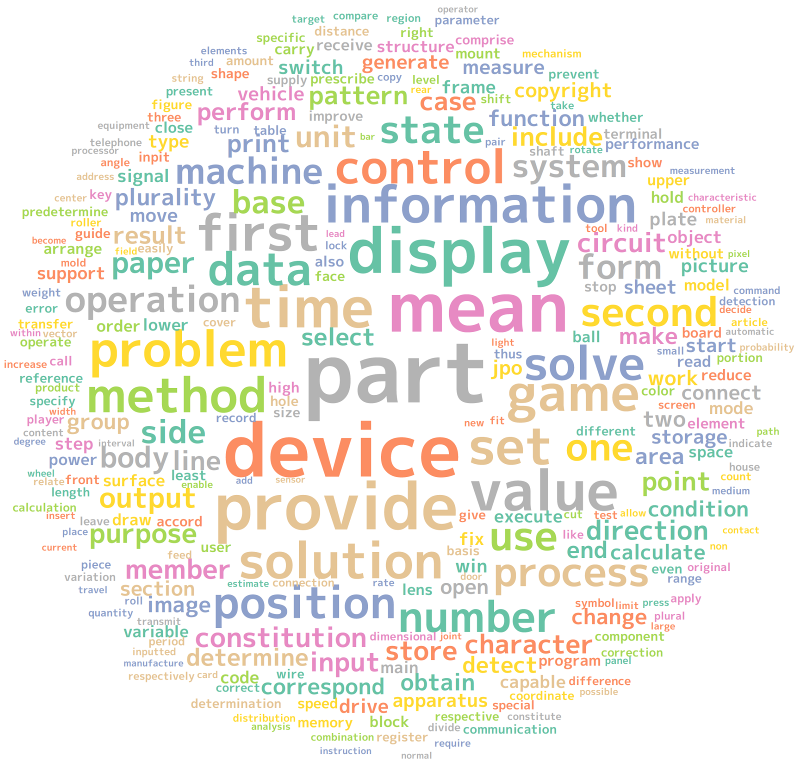
クラスター8：パワー回路



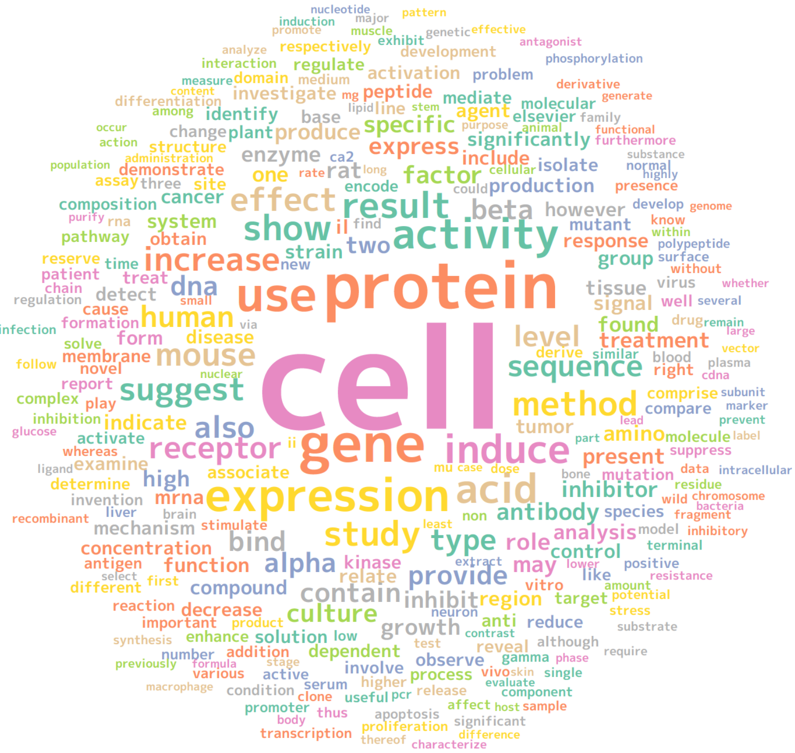
クラスター9：金属材料



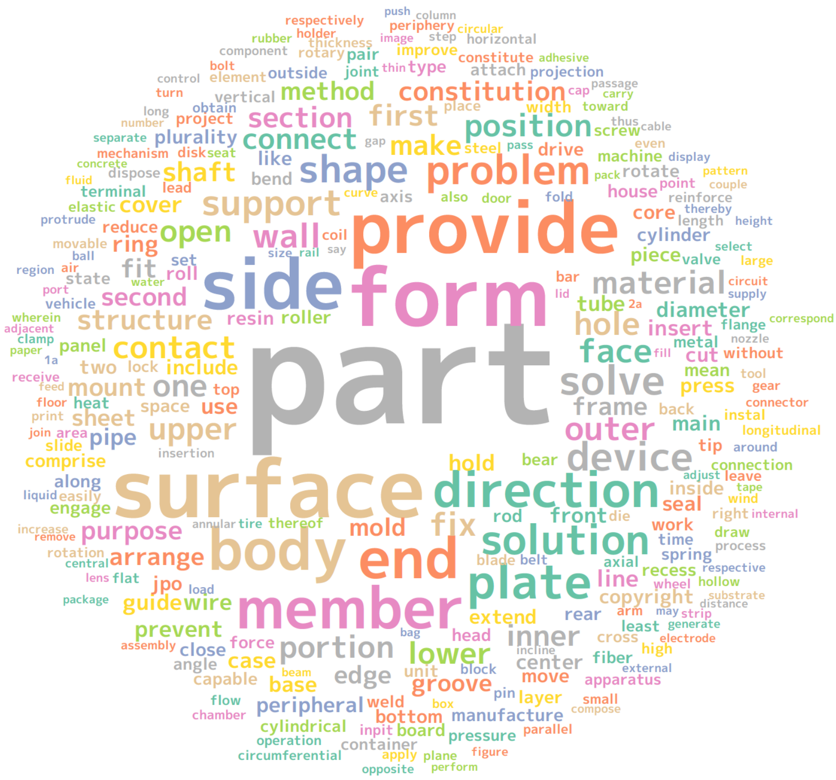
クラスター10：映像表示



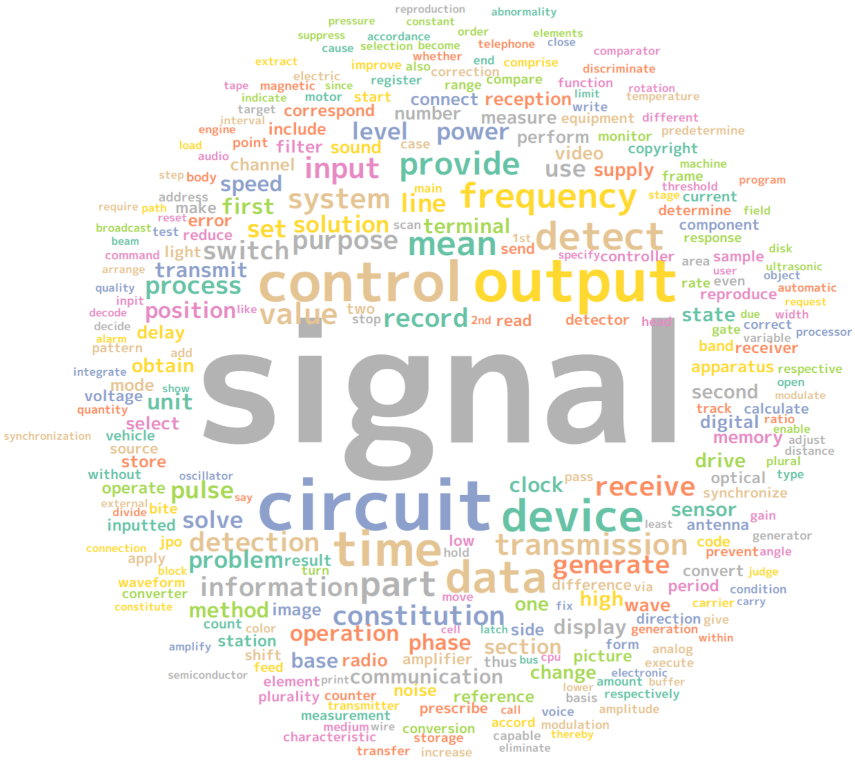
クラスター11：細胞・遺伝子



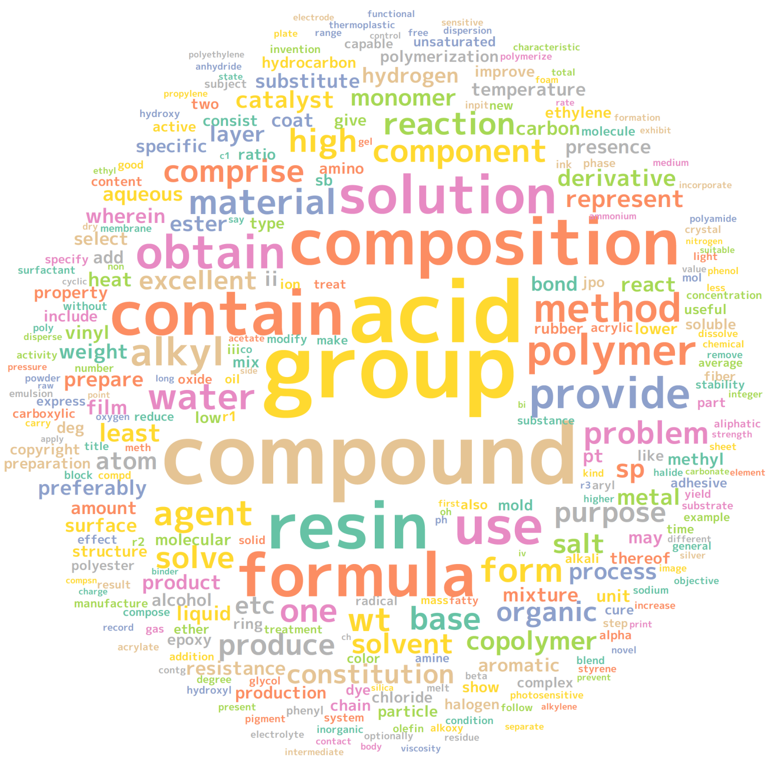
クラスター12：構造力学



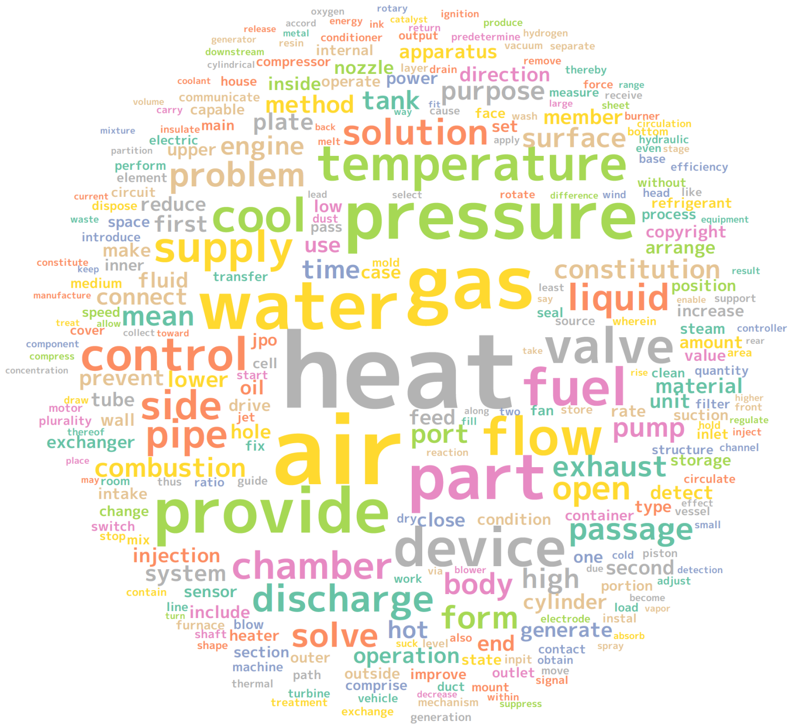
クラスター13：信号処理



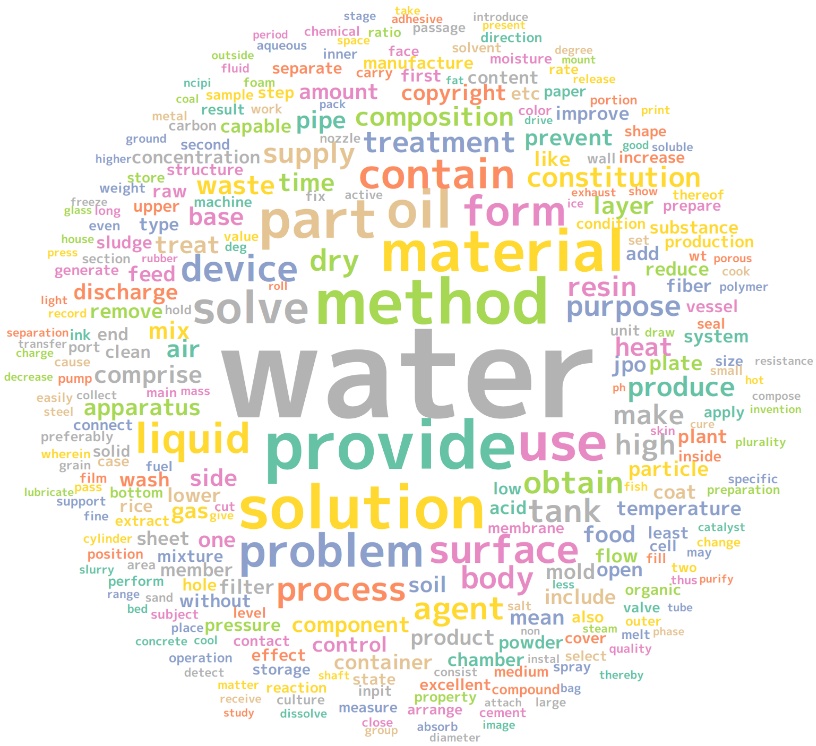
クラスター14：化合物



クラスター15：熱力学



クラスター16：流体処理



DISCUSSION PAPER No.192

論文・特許のテキストデータを使った科学と技術の連関分析

2021 年 02 月

文部科学省 科学技術・学術政策研究所 第 2 調査研究グループ
元橋 一之・小柴 等・池内 健太

〒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: 03-3581-2419 FAX: 03-3503-3996

New indicator of science and technology inter-relationship by using text information of research articles and patents in Japan

Feb. 2021

MOTOHASHI Kazuyuki, KOSHIBA Hitoshi and IKEUCHI Kenta

2nd Policy-Oriented Research Group
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

http://doi.org/10.15108/dp192

https://www.nistep.go.jp