DISCUSSION PAPER No.175

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

A method of extracting content information from patent documents and comparison of their characteristics by applicant type by using the vector space model of distributed expressions

2019 年 12 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 (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
DOI: https://doi.org/10.15108/dp175

MOTOHASHI Kazuyuki, KOSHIBA Hitoshi and IKEUCHI Kenta (2019) "A method of extracting content information from patent documents and comparison of their characteristics by applicant type by using the vector space model of distributed expressions," *NISTEP DISCUSSION PAPER*, No.175, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/dp175

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

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

本稿では，特許の発明内容を分析するための自然言語処理技術と統計数理手法に基づく新たな手法を提案し，日本の特許データを用いて提案手法の機能可能性を評価した。結果として，特許の発明内容の分布状況の可視化や類似特許の検索において提案手法が期待通りに機能することが確認された。また，本提案手法により，日本では個人や大学等の特許は幅広い分野に分布している一方，企業特許は特定分野に集中的に出願されていることが分かった。

研究開発に関する情報は企業にとって戦略的に重要なものであり，内部情報として企業の内部で秘匿されることが多い。しかし，特許が出願されると，その発明の内容は広く公開される。そのため，特許データは個々の企業や産業，場合によっては国全体の技術トレンドについて分析するための貴重な情報源となっている。また，特許権の構成要件として，当該発明の新規性や進歩性に加えて，産業応用可能性が必要とされる。そのため，科学技術論文として公開される情報と比べて，特許情報には，より産業寄り，言い換えれば新商品などのイノベーションに近い情報が含まれている。

他方，特許の情報はデータサイズが膨大になるため，単純にその内容の類似度で分類することは計算コストの面から難易度が高かった。これらの課題に対応するため，本稿では分散表現などの近年普及してきた自然言語処理手法及び高次元ベクトル近傍探索，次元圧縮などの統計数理手法を用いた特許データの分析を試みた。まず，日本の特許庁の公開公報情報におけるタイトルと要約文を用いた分散表現を通じて，特許内容のベクトル空間モデルを作成した。次に，この特許内容のベクトル空間モデルを用いて，特許のクラスタリングや近傍特許の抽出，特許間の距離の測定を試行した。さらに，これらの情報を用いて出願人タイプ（個人・企業・大学等）による特許の特性を明らかにした。

A method of extracting content information from patent documents and comparison of their characteristics by applicant type by using the vector space model of distributed expressions

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

ABSTRACT

In this paper, we propose a new method based on the latest natural language processing technology and statistical mathematical methods for analyzing patent invention contents, and evaluate the usefulness of the proposed method using Japanese patent data. As a result, the usefulness of the proposed method was confirmed in the visualization of the distribution of the invention contents of patents and the search for similar patents. In addition, the proposed method shows that patents by individuals and universities are distributed in a wide range of fields in Japan, while company patents are intensively applied in specific fields.

Information related to research and development is strategically important for companies, and is often hidden inside the company as internal information. However, when a patent application is filed, the contents of the invention are widely disclosed. For this reason, patent data is a valuable source of information for analyzing technology trends in individual companies, industries and, in some cases, the entire country. In addition to the novelty and inventive step of the invention, industrial applicability is required as a constituent of patent rights. Therefore, compared to information published as scientific and technical papers, patent information contains information that is closer to industry, in other words, closer to innovation such as new products.

On the other hand, since the data size of patents is enormous, it is difficult to simply classify based on the similarity of the contents in terms of calculation cost. In order to deal with these problems, this paper tried to analyze patent data by using natural language processing techniques such as distributed expressions and statistical mathematical techniques such as high-dimensional vector neighborhood search and dimension compression. First, a vector space model of patent contents was created through distributed representations using titles and abstract sentences in the publication information of the Japanese Patent Office. Next, using the vector space model of this patent content, we tried clustering patents, extracting neighboring patents, and measuring the distances between patents. Furthermore, the characteristics of patents by applicant type (individual, company, university, etc.) were clarified using this information.

1. はじめに

研究開発に関する情報は企業にとっても戦略的に重要なものであり，内部情報として秘匿されることが多い。しかし，特許出願が行われた情報は，その発明の内容が出願公開によって明らかになるので，個々の企業や産業，場合によっては国全体の技術トレンドについて分析するための貴重な情報である。また，特許権の構成要件として，当該発明の新規性や進歩性の他に，産業応用可能性も含まれる。従って科学技術論文として公開される情報と比べて，より産業寄り，言い換えれば新商品などのイノベーションに近い情報が含まれている。

例えば，日本特許庁（JPO）はこの特許情報をベースに技術動向調査として，毎年重要な技術分野をいくつか選んで，内外の技術動向に関するレポートをまとめている。また，WIPO（世界知的所有権機関）は近年 AI に関するレポートを取りまとめた[WIPO 19]。これらのレポートにおいては，対象となる分野（例えば AI）に関する特許を抽出することが必要となるが，その際には IPC（国際特許分類）コードをベースとした検索式（IPC より細かい特許庁ごとの技術分類，例えば JPO の FI (File Index)や USPTO の CPC (Cooperative Patent Classification)，やタイトル，要約のキーワード）が作成されている。

最近では，自然言語処理技術を用いて，特許のテキスト情報（タイトル，要約文，請求項など）から発明の内容を把握し，特許分類や技術動向分析に用いるケースも多い。例えば，[Arts 17]は米国特許のタイトル，要約文から特許の内容をベクトル表現化し，特許間の類似度（Jaccard 類似度）を計算した。更に，この結果を IPC コードや引用による特許間の類似度と比べて，より客観的な類似度を表していることを明らかにした。また，[Younge 16]は，やはり米国特許テキスト情報のベクトル空間モデルを作成し，特許間の類似度（cos 類似度）を算出し，その結果を公開している。ベクトル空間モデルを使うことで，①技術分類のバイナリ情報（同じ分類に属しているか否か）と違って，連続変数として特許間の類似度を表現できること，関連して②同一技術分類内における特許集合の中での位置（例えば中心にあるか，周辺か）によらない，技術スペース上での評価が可能となること，などのメリットを挙げている。

本稿は JPO の公開特許公報情報におけるタイトルと要約文を用いた分散表現を通じて，特許内容のベクトル空間モデルを作成した。また，この情報に基づいてクラスタリングと近傍特許の抽出・距離の測定を行い，その内容について考察を行った。更に，その情報から出願人タイプ（個人・企業・大学等）による特許の特性について分析を行った。結果として，個人や大学等の特許は幅広い分野に分布している一方，企業特許は特定分野に集中的に出願されていることが分かった。

2. 提案手法

本章では分析の手法・手続について述べる。

具体的には，分散表現と呼ばれる手法を用いて特許概要文を座標値に変換し，その特許空間上で様々な処理を行うことで，特許空間の特徴を把握する手法・手続を述べる。

2.1. 既存の分析方法

前章においてすでに示したとおり，特許データについてはこれまでにも様々な分析手法の提案・分析がなされている。ここで特許や論文データの分析手法を整理すると大きく 2 つの方法があげられる。ひとつは引用情報を用いた計量書誌学的な分析方法，もうひとつは，記載内容を用いた分析方法である。後者については更に細分化することができ，特許の場合は 1．IPC 分類や F タームなど何らかのキーワード・分類を用いるもの[元橋 18，WIPO 19]，2．概要文など具体の内容を用いるもの[Arts 17, Younge17]，などがあげられる。

■ 引用情報ベースの分析方法とその特徴

引用情報を用いた分析は関係性が明示されていることから信頼性が高く，論文や特許の分析でこれまでにも多くの実績[富澤 06，科学 07，科学 14，科学 16，佐藤 17，科学 18]を有する手法である。ただし，引用関係はグラフ構造であるので，直接引用だけでなく，その先の階層まで関係性を追跡するに従って分析コストが増大してゆくという課題もある。

■ キーワード・分類ベースの分析手法とその特徴

記載内容に基づく分析を行う場合，キーワード・分類を用いると，ある分類の出願傾向を時系列で追うような場面には有効であるほか，分類やキーワードでフィルタリングした上で前述の引用関係を分析することが可能になる。

■ 具体の記載内容ベースの分析手法とその特徴

具体の記述内容を用いた分析手法も提案されている。例えば，古典的な自然言語処理の手法である Jaccard 係数や cos(コサイン)類似度を用い，特許間の相関行列を求めてマッピングする手法などがあげられる[樽松 14, Younge 16，富永 18]。この手法は引用情報に現れない関係性を推定することができるため，補完手法として有用であるが，いくつか課題もある。具体的には，1．単純に相関行列を求める場合，全ての組合せを計算する必要があり，分析コストが大きいこと，2．単純に cos 類似度を用いると，意味内容が反映され難いこと，などがあげられる。

■ 単純な cos 類似度が有する一般的な課題

▽ *cos 類似度の計算方法*

後者について補足するため，まず cos 類似度計算方法を説明する。

cos 類似度の基礎的な発想は「同じような単語が同じような頻度で出てくるものは似ている」というものである。そこで単語それぞれを独立した次元とみなす。すると各文書における単語の出現回数に基づいて，文章を多次元空間上の 1 点にマッピングすることが可能になる。このとき，同じ単語が同じような頻度で使われていると，ベクトル間の内積（cos）が 1 に近づく。一方で単語の重複が無いような場合は 0 に近づく。単語の出現回数は 0 を含む正の整数値であるため，完全に独立ならゼロ，使用されている単語が同じで，その頻度の割合が同じであれば角度が一致するため 1 を取る。以上より，内積（cos）によって「似ていない」「似ている」の類似度を 0 から 1 までの数値で表現できる。

例えば，「みかん」という単語が 3 回，「りんご」が 1 回出てくる文書 A と，「みかん」が 2 回，「りんご」が 3 回出てくる文書 B の類似度を計算したいとする。

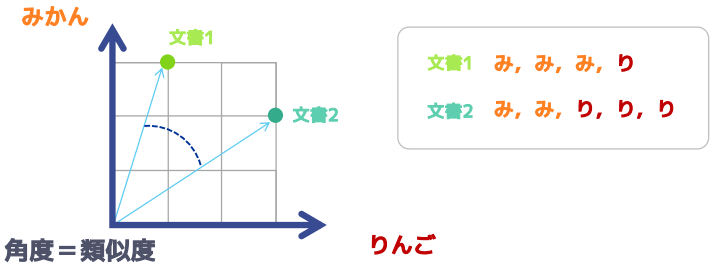


図 2－1： cos 類似度の考え方

このとき，各単語を独立した次元と見なすと，「みかん」次元と「りんご」次元の 2 次元で空間を定義でき，単語の出現頻度と成分とを対応させると文書 A，B をこの次元中の座標値（ベクトル）として表現できる。このとき文書 A，B の内積（cos）を取ると，おおよそ 0.8 程度となり，類似度は約 0.8 若しくは約 80%の類似度といえる。

▽ *単純な cos 類似度が有する課題*

この手法は有用であるものの，課題もある。

例えば，計算機の中では“A”と“a”，それぞれに別々のコードが割り振られ，異なる記号として扱われる。このように計算機にとっての記号と，人間が記号に与えた・記号から読み取る意味は基本的に乖離している。同様に「みかん」と「ミカン」は人にとっては多くの場合で同じような概念を指すと期待できるが，記号として異なっているため計算機上では別物として扱われる。ここで cos 類似度は各「単語」を独立した次元として扱うが，ここでの単語は記号の集合である。したがって「みかん」と「ミカン」は異なる記号の集合となり，類似度はゼロとなる。同様に「細君のバースデーにケーキを購入して帰宅し

た」「妻の誕生日に“いちごショート”を買って帰った」は人間にとっては似たような意味内容を有するが，先に示した単純な cos 類似度の算出手法に従った場合，単語の重複がないため，類似度がゼロとなる。

こうした課題を解決できる手法として近年，分散表現（単語埋め込み，Word Embedding）という手法が提案され，活用されている。

2.2. 分散表現

■ 分散表現の概要

分散表現は深層学習の核となる技術でもあるニューラルネットワークを応用したもので，単語を何らかのベクトル表現（ベクトル空間モデル）に変換してくれる仕組みといえる。

いくつかの手法があるが，イメージとしては「ある単語の前後に同じような頻度で出てくる単語は似ている」という前提で学習をさせるようなものといえる。つまり穴埋め問題を対象に学習させ，ある穴埋め問題が出題されたとき，その穴によく当てはまりそうな単語の集合は似ているとするようなものである。

この分散表現を用いることで，「みかん」と「ミカン」は類似する（意味空間上で近傍に配置される）ことを数値的に表現できるため，分散表現を用いて距離，又は cos 類似度に代表される類似度を算出することで，前述の「みかん」と「ミカン」が独立に取り扱われる問題を回避することができる。

■ 文章の分散表現

ここで，単語ではなく文章の類似度を測りたい場合，いくつかの方式が考えられる。直接文章の分散表現を算出する doc2vec などの方法[Le 14, Dai 15, Lau 16]のほか，たとえば，各単語の分散表現を線形加算して文章の分散表現とする方法や各次元の最大値を取る方法，単語の重要度によって重み付けをした上で加算する方法，そもそも重要単語のみに絞り込んで加算する方法，などもある[Shen 18]。

単語のバリエーションが十分に大きい場合は，単語の分散表現を用いる方法を用いると，単語単体の分散表現を得ることも，文章の分散表現を得ることもでき，利便性が高い。ただし，すでに述べたとおり，文章の分散表現獲得に様々な方式が考えられ，それぞれ長短が存在する点には留意が必要となる。

■ 分散表現がもたらすメリット

分散表現を通じて文章（特許）をベクトル化できることで，いくつかのメリットが得られる。ひとつは単純にこれまで述べてきた「みかん」「ミカン」問題の緩和である。二つ目は情報の圧縮にある。前述の通り単純な cos 類似度では単語それぞれを次元と見なす

ため，データセットが巨大になるとベクトルが数十万次元を超すこともあり，かつその多くがゼロでスパースである。一方，分散表現では手法にもよるものの，数百次元で表現でき可用性が高い。さらに，数百次元程度で表現できることにより，高次元ベクトル近傍探索などの手法の適用が容易になり，完全性を求めない場合は全組合せの計算を伴わずに類似文章が取得できるようになったり，実用的な時間で次元圧縮手法を適用できたり，といったことが実現する[小柴 19, 椿 19]。

こうした背景から，論文やファンディング研究課題，国会会議録など科学技術イノベーション政策関連のテキストデータについても分散表現を用いた分析が行われている[小柴 19, 椿 19]。

2.3. 提案する分散表現を用いた特許空間の分析手法

以上の背景より，分散表現を用いることで公開特許公報データについても，意味内容ベースで個別特許間の関係性に基づいて，全体の構造・特徴を理解できる可能性が高いと考え，実験を行った。

■ 分析の手順・手続

分析の大まかな手順・手続は以下の通りである。

1. 特許データから，タイトルおよび概要文を抽出する
2. 形態素解析器にかけ，名詞句のみ抽出する
3. 上記，2. のデータに基づき単語の分散表現を獲得する
4. 獲得した単語分散表現を用い，各特許の分散表現を獲得する
5. 場合により，特許分散表現をもとにクラスタリングを行う
6. 場合により，特許分散表現をもとに次元圧縮を行い 2 次元で可視化する
7. 特許分散表現に対して高次元ベクトル近傍探索用のグラフインデクスを作成しておくことで，任意の特許データに類似する特許データを高速に取得する

データや手続の詳細については後述する。

3. 実験

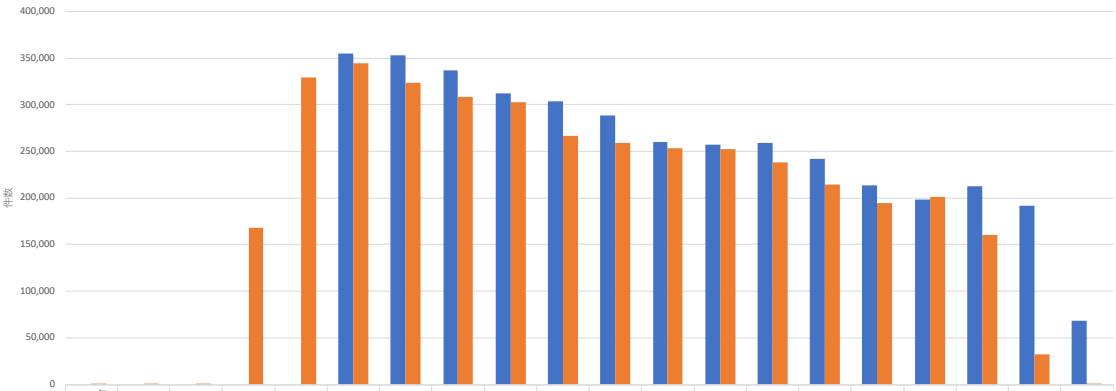
本章では，提案手法を実際に適用した実験の結果について述べる。

3.1. データ

本実験では，特許庁が提供・公開している「公開特許公報」のデータ[1]を用いた。

データの期間は公開日ベースで 2005 年 1 月から 2019 年 4 月末日まで，種別としては「A」公開特許公報，公表特許公報および「S」再公表特許に分類されているものを対象とした[2]。結果，対象となる公開特許公報の件数は 4,069,503 件となっている。

期間毎のデータの件数について，図 3－1，3－2 に示す。

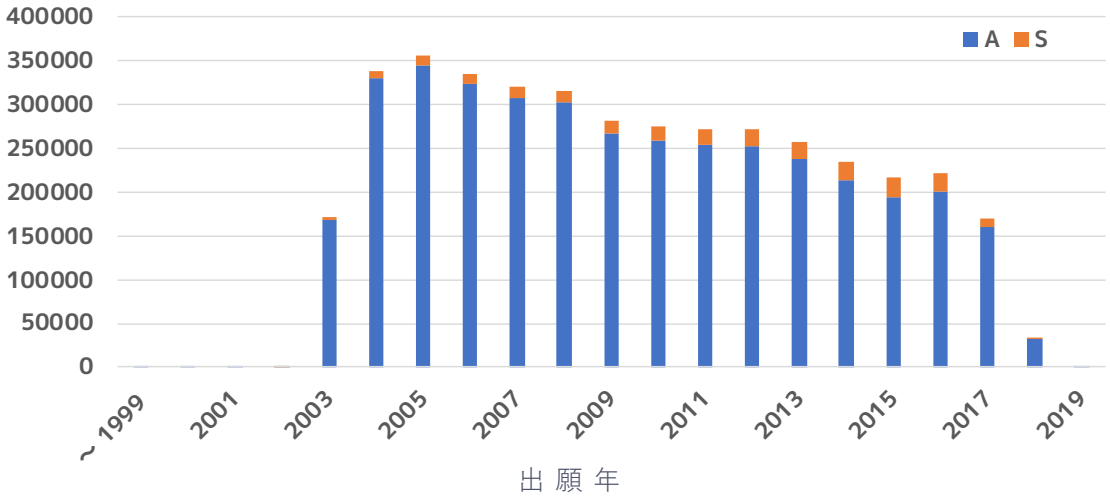


	2000年以前	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
公開年	0	0	0	0	0	354,900	353,100	336,797	312,443	303,486	288,450	259,700	257,455	258,913	241,728	213,433	197,998	212,880	191,516	68,740
出願年	83	233	379	168,300	329,440	344,813	323,560	308,241	303,020	266,966	259,205	253,551	252,460	238,139	214,181	194,095	201,319	160,045	32,755	725

図 3－1： 公開特許公報の出願年・公開年ごとの件数推移

[1] https://www.publication.jpo.go.jp/

[2] データ中には A1（再公表），B1，B2，など様々な種別のものがある。また，「再公表特許」はいわゆる「公開特許公報」ではないが，利便性を考慮して特許庁からは公開特許公報と併せて公開されており，今回の分析対象データに含めた。



	～ 1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	
A	29	83	233	379	168,300	329,440	344,813	323,560	308,241	303,020	
S	0	0	0	476	3,437	8,753	11,094	12,093	12,120	12,052	
	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
A	266,966	259,205	253,551	252,460	238,139	214,181	194,095	201,319	160,045	32,755	725
S	13,949	15,295	17,811	18,836	18,621	19,990	22,033	20,695	10,191	518	0

図 3－2： 公開特許公報の出願年・種別ごとの件数推移

次に，これらの公開特許公報データ（以後，単に特許データという）の概要文について「【課題】」や「【解決手段】」などのラベル文字をルールベースで削除した。

3.2. 単語分散表現の作成

単語の分散表現については Facebook 社が開発・公開している FastText[3] [Joulin 16, Bojanowski 17]を用い，以下の手順で作成した。

まず，手法としては skip-gram を採用し，データ量を勘案して次元数を 300 に設定した。その他はデフォルトのパラメータを採用した。

データについては，前述した整形済み特許データを利用した。

ところで，分散表現作成は前後の単語を元に学習する。したがって，ここでは 1. タイトルについては全てをひとつの単位，2. 概要については読点までをひとつの単位，として，単位毎に学習を行わせることにした。

[3] https://fasttext.cc/

また形態素解析器（MeCab[4] ＋ mecab-ipadic-neologd[Sato15, Sato16, Sato17]）を用い，名詞句のみを抽出し，学習させることにした。

結果，140,638 単語について 300 次元の分散表現を得た。

これらの分散表現について K-means++[Arthur 07]を用いて 16 分類し，UMAP (Uniform Manifold Approximation and Projection) [McInnes 18]によって 2 次元空間上で可視化したものを図 3－3，3－4 に，分類ごとに作成したワードクラウドを，図 3－5 から図 3－20 までに示す。なお，ワードクラウドにおける単語の大きさは元の特許データにおける出現頻度に対応している。

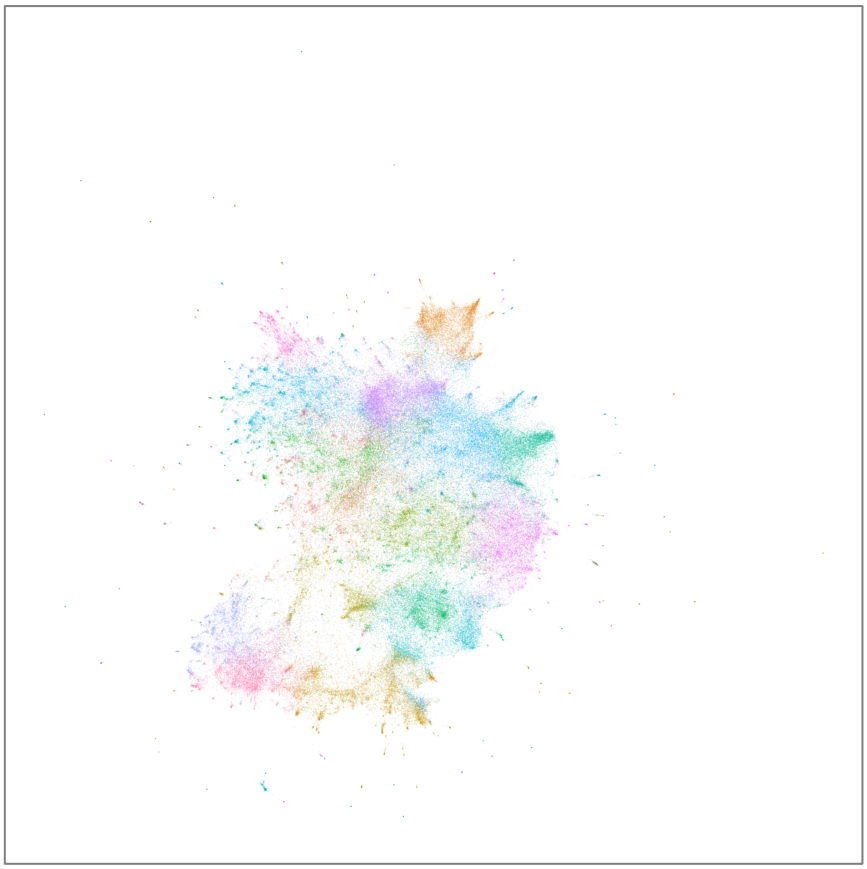


図 3－3： 単語分散表現（300 次元）の 2 次元圧縮表示

[4] http://taku910.github.io/mecab/

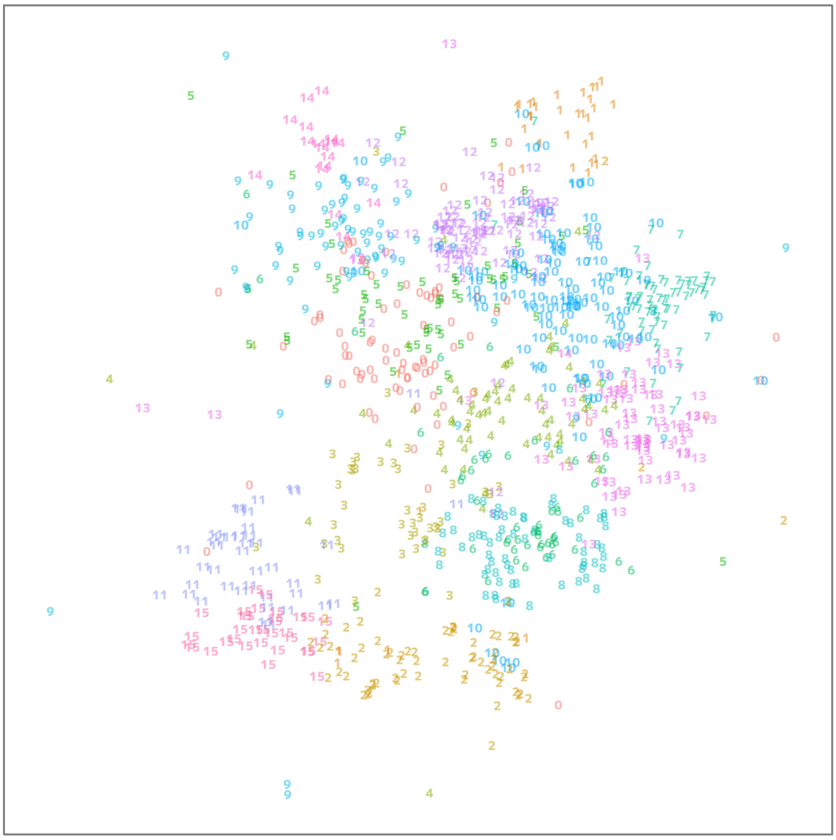


図 3−4： 2 次元圧縮した単語分散表現とクラスタ ID の対応

図 3−5： ID0

図 3−6： ID1

図 3－7： ID2

図 3－8： ID3

図 3－9： ID4

図 3－10： ID5

図　3－11： ID6

図 3－12： ID7

図 3－13： ID8

図 3－14： ID9

図 3－15： ID10

図 3－16： ID11

図 3－17： ID12

図 3－18： ID13

図 3－19： ID14

図 3－20： ID15

図 3－3 から 3－20 をみると，機械系のものや化学系のものなど関連があると思われるものが固まっており，かつ，位置的な関係性についても定性的にある程度妥当と思われる結果が得られている。

3.3. 特許分散表現の作成

単語分散表現に基づいて，個別の特許データについて特許分散表現を作成した。

ここでは，特許データのタイトルと概要を単位として，単語分散表現と同様の手法で名詞句（単語）を抜き出す。

その後，各単語の分散表現を線形加算し，正規化したものを特許分散表現とする。単語分散表現が 300 次元であることから，特許分散表現も 300 次元のベクトル・座標値として表現されている[5]。

これらの分散表現について K-means++を用いて 16 分類し，分類ごとに作成したワードクラウドと，UMAP を用いて 2 次元に可視化したものを図 3－21，3－22 に示す。なお，ワードクラウドにおける単語の大きさはクラスタ内の特許データ全体に対する単語の出現頻度に対応している。なおワードクラウドのキャプションに付けられた語句は，ワードクラウド全体を代表すると期待される表現を主観により設定したものである。また，UMAP による次元圧縮においては 4,069,503 件全件をそのまま用いて計算することが困難であったため，ランダムサンプリングした 30 万件でモデルを学習させ，当該モデルに基づいて 4,069,503 件全件の配置を行っている。

[5] 正規化を伴っていることから，より具体的には半径 1 の 300 次元球面上に分布している。

図 3－21： 特許分散表現（300 次元）の 2 次元圧縮表示

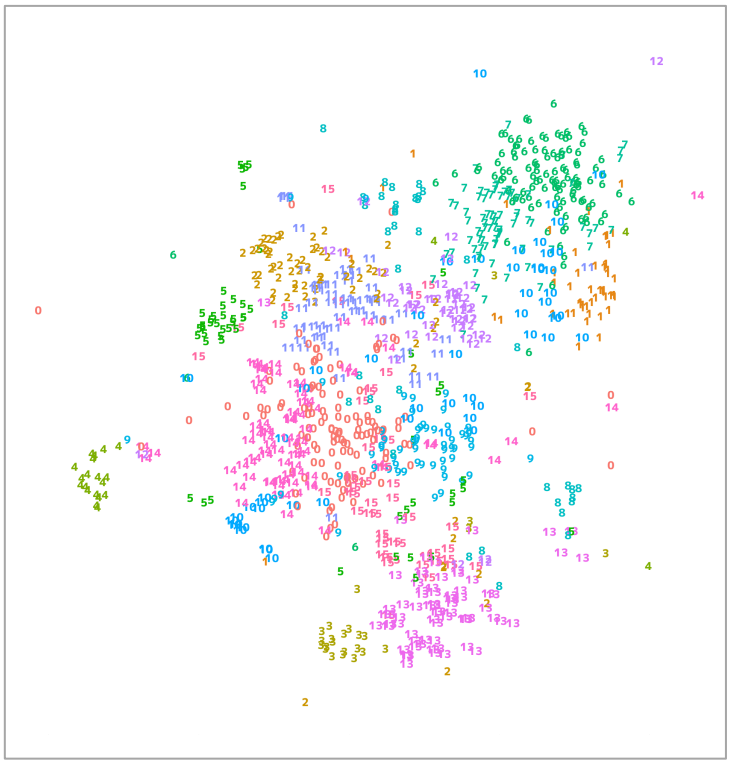


図 3－22： 2 次元圧縮した特許分散表現とクラスタ ID の対応

図 3－23： ID0 加工

図 3－24： ID1 電子回路

図 3－25： ID2 半導体

図 3－26： ID3 バイオ

図 3－27： ID4 ゲーム

図 3－28： ID5 金属

図 3－29： ID6 情報

図 3－30： ID7 画像

図 3－31： ID8 画像

図 3－32： ID9 流体

図 3－33： ID10 車両

図 3－34： ID11 端子

図 3－35： ID12 光学

図 3－36： ID13 化合物

図 3－37： ID14 モータ

図 3－38： ID15 樹脂・膜

図 3—37，3－38 を見ると，「回転」「軸」などモータに関連しそうなクラスタや，「半導体」「基盤」など半導体製造に関連しそうなクラスタなど，ある程度解釈が可能な状態が得られている。これらから，定性的にはある程度妥当と考えられる結果が得られたと考えられる。

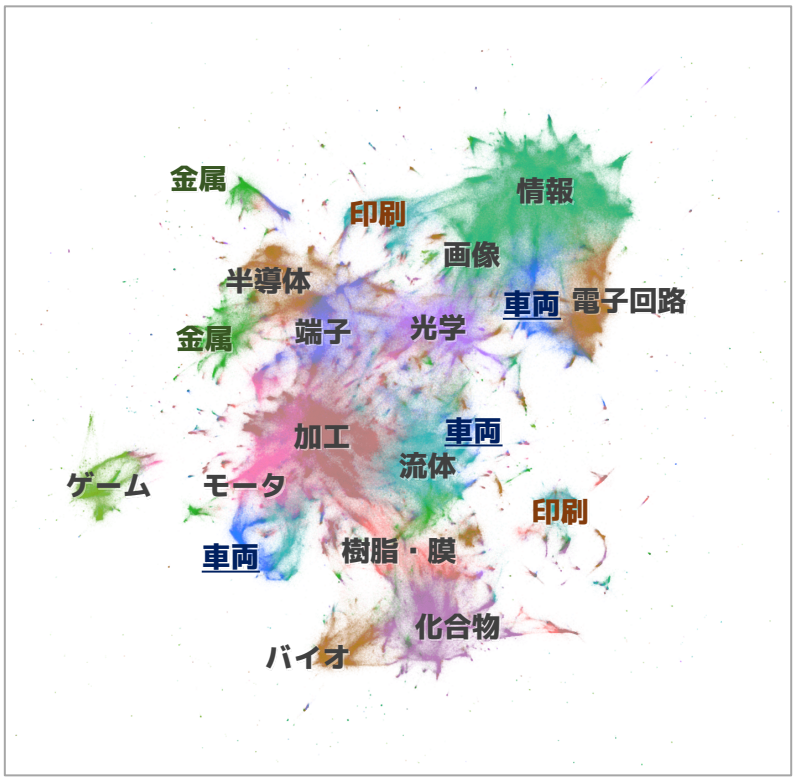


図 3－39： 2 次元圧縮した特許分散表現とクラスタの対応

あわせて，ワードクラウドのラベルと UMAP での 2 次元表現を組み合わせたものを図 3－39 に示す。「車両」や「印刷」「金属」のように複数のエリアにスプリットしているクラスタもあるものの，多くは 2 次元に圧縮した状態でも近くに配置されている。また，「情報」の近くに「画像」や「電子回路」が，「化合物」の近くに「バイオ」や「樹脂・膜」など関連が強いと思われるものが近くに配置されている。他にも例えば複数にスプリットしている車両についても，制御系に関しては「電子回路」と，燃料制御や空力特性などは「流体」と，駆動系は「モータ」と関連が近いと考えられ，全体としてある程度妥当と思われる結果が得られている。ただし，クラスタの名付けは主観的に行われており必ずしも正しく意味内容が反映・表現されているとは限らない。あくまで印象の範囲に留まっている点に注意を要する。

3.4. 特許分散表現空間の特徴

本節では前節までで得られた特許分散表現空間の特徴についてまとめる。

3.4.1. 128 分類のクラスターと IPC クラスの比較

3.3 節では簡単に 16 分類で特許分散表現の特徴を確認した。ところで，特許には IPC 分類や F タームなど，予め人手で分類コードが付与されている。そこで本節では IPC 分類をベースとして特許分散表現との比較を試みる。

IPC 分類は国際的に用いられる分類コードで，複数の階層を有している。ここでは可読性を重視してサブグループまでの 126 分類を採用する。これに類似させ K-means++でも 128 分類を採用して比較を行った。なお 128 分類を採用している意図は 126 に最も近い 2 の乗数で，切りが良いということのみに起因し，他意は無い。

結果を図 3－40 および図 3－41 に示す。

図 3－40： 128 のクラスターと IPC 分類の対応（1）

図 3－41： 128 のクラスターと IPC 分類の対応（2）

図 3－40 および図 3－41 は横軸に IPC 分類を縦軸に K-means++で分類した 128 分類を並べた。なお，ひとつの特許に複数の IPC 分類が割り当てられている場合は最初の 1 件のみを採用している。セルの色の濃さは行単位（128 分類）で正規化した上で，100%をもっとも濃い色で，0%を最も薄い色（白）で表現している。また，128 分類については ID に特に意味は無いため，後述のルールに沿ってソートし，番号を振り直してある。

128 分類のソート手続は以下の通りである。まず 128 分類それぞれの最頻値が IPC 分類の A01 から H99 のどれに属するかを計算する。次に，A01 を 1，A02 を 2，… H99 を 126 として，IPC 分類に数値を割り振る。その上で 128 分類に最頻 IPC 分類のソート用数値を割り付けて，昇順にソートする。このとき，128 分類に割り付けた IPC のソート用数値が同じ場合は，頻度の大きい方が上に来るように調整している。

前述した規則で順序を揃えてあるため，IPC 分類と分散表現からの 128 分類の間に強い相関がある場合，表の対角線が濃くなる。定性的には図からも明らかに IPC 分類とのある程度の相関が伺える。一方で，例えば 128 分類側の番号で 4 から 9 のクラスタのように 128 分類の複数のクラスタが IPC のひとつのクラスタに強く結びついているものもある。また，それとは逆に IPC 分類の B67，68，81，82 のように 128 分類では結びつきが見えないものもある。

頻度ベースで分類尺度間の相関に関する指標であるクラメールの連関係数（Cramer's V）を計算すると 0.314 を示しており，定量的にも上記の観察結果と合致する結果が得られている。

3.4.2. 時系列変化を表現（5 年ごと）

次に，特許分散表現空間の時系列での変化について示す。

ここでは特許データを，公開年ベースで 1．2005 年から 2009 年まで，2．2010 年から 2014 年まで，3．2015 年から 2019 年まで，の 3 区間に分割して示す。2019 年については 4 月末までのデータのため区間 3 のみ他の区間に比べて 8 ヶ月ほど期間が短い点には注意が必要である。また「S」再公表特許のうち公表年が 2005 年以前のもの(6,290 件)は全て 2005 年に計上する。

結果を図 3－42 から 3－44 までに示す。また，3 区間を重ねたものを図 3－45 に示す。

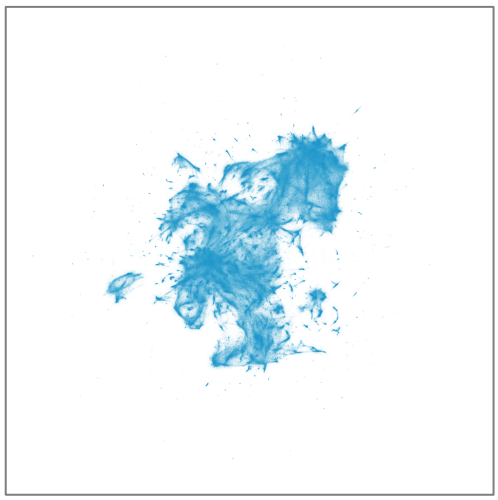


図 3－42： 2009 年までの特許分散表現

図 3－43： 2014 年までの特許分散表現

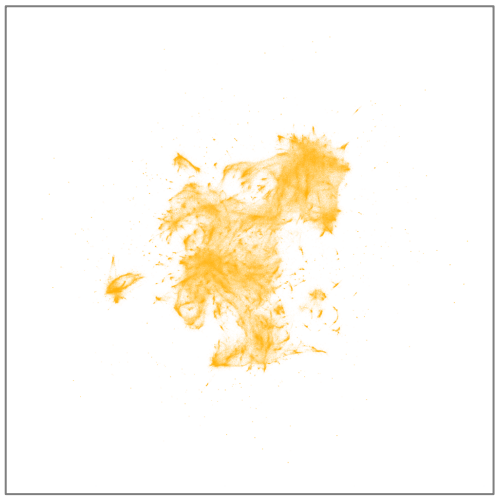


図 3－44： 2019 年までの特許分散表現

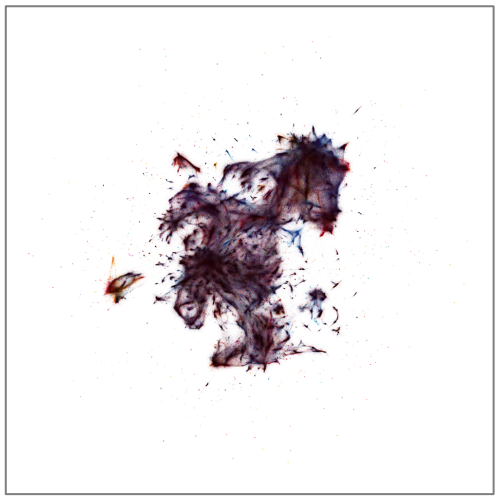


図 3－45：3 区間を重畳した特許分散表現

3 区間はそれぞれシアン，マゼンタ，イエローに対応しているため，重畳して 3 色が重なると黒で表現される。3 区間の各図や重畳図を見ると，今回の対象とした 3 区間において，基本的な構造自体には変化が見られていない。重畳図では強いて言えば，相対的に青が目立つため，2009 年までとそれ以降で多少違いが生じている可能性が示唆される。

3.5. 高次元ベクトル近傍探索

最後に，類似特許データの探索に供するため，作成した分散表現について Yahoo!JAPAN 社が開発・公開している高次元ベクトル近傍探索（NGT: Neighborhood

Graph and Tree for Indexing High-dimensional Data）[6][岩崎 13]を用いてインデクスも作成した。

すでに述べたとおり，単純に特許間の類似度を計算すると組み合わせ爆発により，計算量やストレージ，検索の面から大量の計算コストを要する。そこで，NGT を用いてこれらの課題を解決する。NGT を用いることにより，任意の特許について近傍 n 点の特許データを取得する。といった操作を高速に行うことができる。

なお，NGT は近似手法であるため，必ずしも近傍 n 点を正確に取得できていない可能性があり，その点に注意が必要である[7]。一方，本手法を採用することで全特許間の類似度を算出する必要がなくなるため，計算機コストを飛躍的に圧縮することができる。例えば，[Younge 2016]では米国特許約 530 万件特許間の距離を計算し約 300TB のデータになったとしているが，我々の試行では計算結果を保持しないためその容量を圧縮することができ，約 400 万件の特許に対してインデックスのサイズは約 5.4GB，任意の 1 特許の近傍 200 件を取得するために要する時間は約 7.5msec と，リーズナブルな結果になっている。

検出精度については以下の通り検証した。特許分散表現からランダムに 1000 件を抽出し，この 1,000 件に限定した上で全件に対して cos 類似度を算出した。その上で，この cos 類似度で取得した類似特許分散表現の上位 200 件の ID と，NGT で取得した類似特許分散表現の上位 200 件の ID とを比較した。結果は一致率 98.27% となっており，これは 200 件のうち，3-4 件が漏れている程度といえる。なお，1,000 サンプルから全件に対する cos 類似度のデータは概ね 120GB となった。

[6] https://github.com/yahoojapan/NGT

[7] 近傍 10 点を取得させた際，データセットとして 10 点は確実に返却されるが，その際，データセット中には「実際に存在する 4 番目に近いはずのデータ点」が含まれておらず，4 番目を除く 11 番目までの近傍点が返却される，といった状態が生じる可能性がある。

4. 近傍（距離）データの評価

発明内容のベクトル表現情報（300 次元）から得られる特許間の距離（1－cos 類似度）の評価を行った。ここで対象とした約 470 万件の特許からランダムに抽出した 10 万ペア（10 万件×10 万件）の cos 類似度を計算した。図 3－1 はその分布状況を見たものである。平均値は 0.449，中央値は 0.455，0.05 間隔のヒストグラム上の最頻値は 0.45～0.50 の間にあることが分かった。[Arts 17]や[Younge 16]などのこれまでの研究成果は，単語単位の TF-IDF モデルをベースとしたベクトル表現が用いられている。この情報から算出された特許間の類似度は，特許間の単語の重なりが見られないことから cos 類似度（[Arts 17]は Jaccard 類似度)）は大半の特許ペアでほぼ 0（距離とすると 1）となる。一方，本研究においては単語の分散表現をベースとしていることから，特許ベクトル間のある程度の見せかけの相関が表れることを示している。

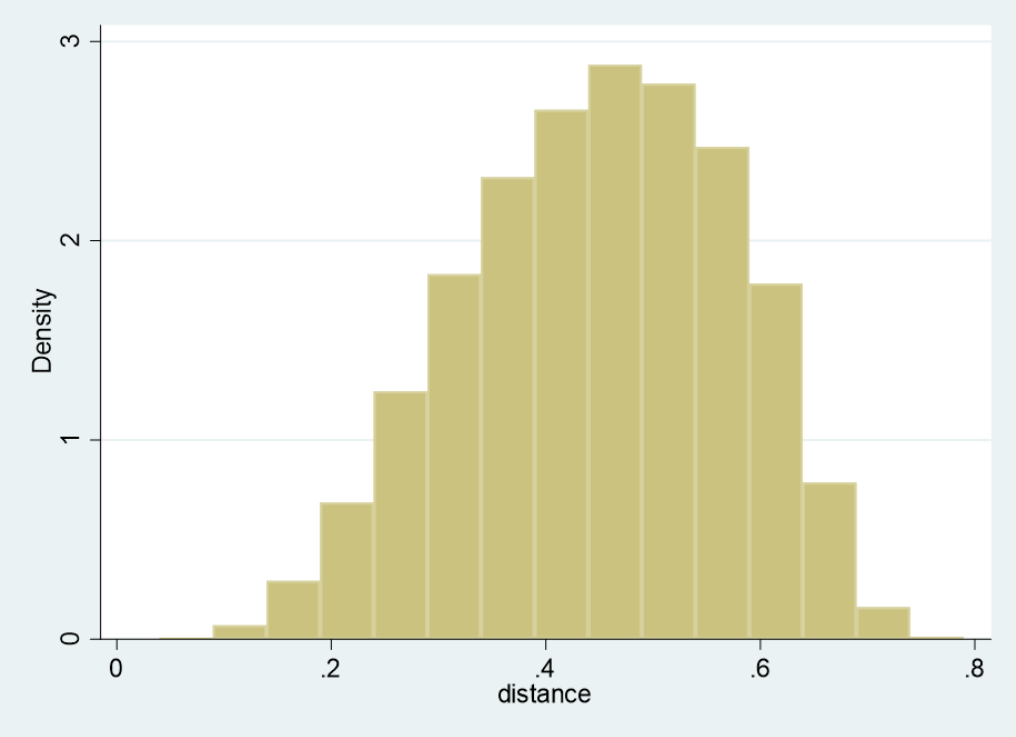


図 4－1：特許間距離のヒストグラム

次に IPC 分類体系を使った特許間距離の評価を行った。特許に関する国際的な技術分類である IPC(International Patent Classification)，特許の技術分類について，セクション（A～H）＋クラス（2 桁の数字）＋サブクラス（1 桁のアルファベット）＋グループ＋サブグループというような階層構造になっている。ここではセクションからグループまでの各レベルにおいて同じ分類に属する特許間の距離を算出した。なお，セクションレベルで 5,000 件，クラスレベルで 3,000 件，サブクラスレベルで 2,000 件の特許をランダム

算出して，ペアワイズの距離を見ている。なお，それぞれの分類内における特許数が上記の閾値に達しない分類は分析対象から外している。グループレベルについては，それぞれの分類内で 2 件以上の特許を持つものをすべて対象として，特許数が 1,000 件を超えるものについては 1,000 件をランダムに抽出して分析に用いた。

図 4－2 はセクションごとのペアワイズ距離のボックス図である。同一セクション内で見ることで，全体の状況（図 4－1）と比べて平均的な距離が小さくなるはずであるが，A（生活必需品），G(物理学)，H（電気）などは距離の中央値が 0.4～0.5 の間となり，図 4－1 の状況と大きな差異が見られない。しかし，C(化学；冶金)や E（固定構造物）は中央値が 0.3 と小さく，これらの分類については，比較的均質な内容の特許で構成されていることが分かった。

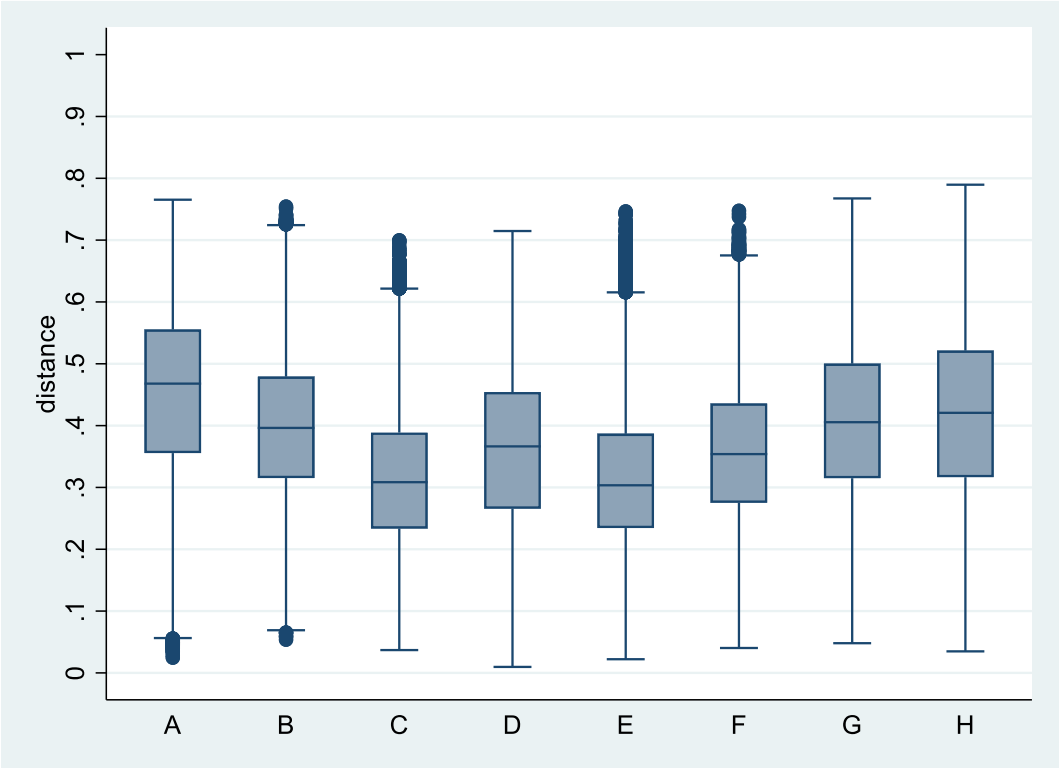


図 4－2：IPC セクションレベルの距離分布

図 4－3 は IPC 分類のレベル毎に距離の中央値を算出し，その値の四分位値を見たものである。なお分類数はセクションレベルで 8，クラスレベルで 90，サブクラスレベルで 273，グループレベルで 6,125 となる。前述したように各分類において一定数以上の特許が存在するものを取り上げて分析を行っているので，上記の数は IPC それぞれのレベルにおける分類数と一致しないことに留意されたい。分類が細かくなるほど距離が小さくなる（特許間の同質性が高まる）ことが確認できた。なお，グループレベルで見ると第 1 四分位が 0.16，第 2 四分位（中央値）が 0.19，第 3 四分位が 0.23 となっている。同質性の高い分類（平均距離が小さい）と異質性が高い分類（平均距離が大きい）が混在してい

るが，距離でいう 0.2 が，概ね IPC グループレベルで同じ分類の近さを示していることが分かった。

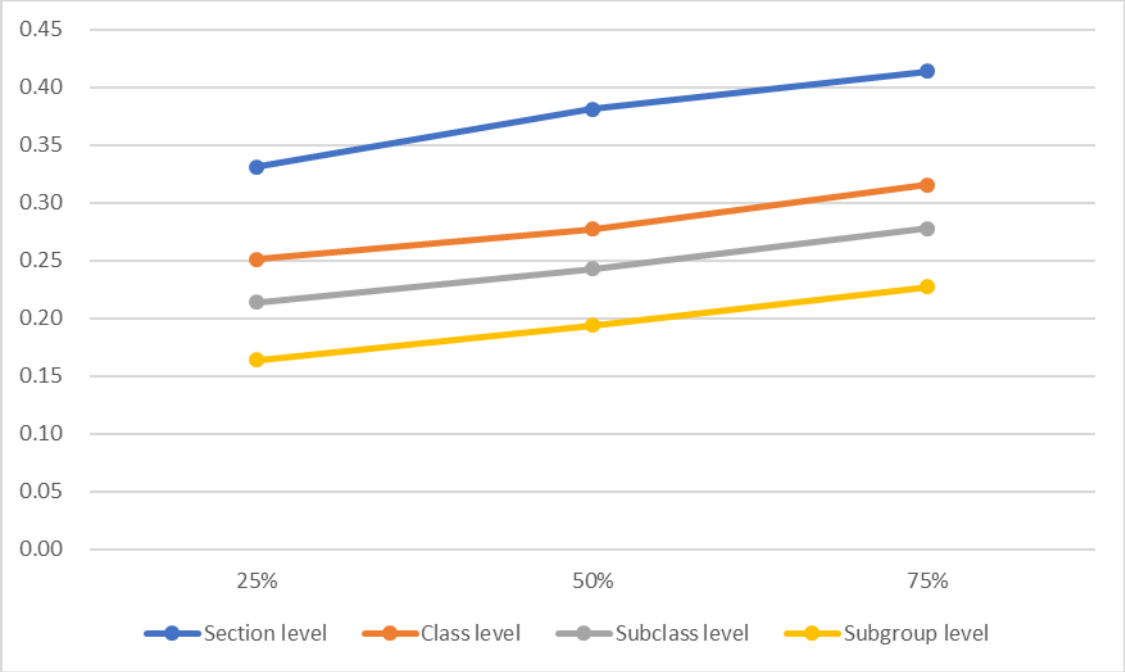


図 4－3：IPC レベル毎の距離中央値の四分位値

図 4－4 は，IPC 分類レベル毎に中央値の分類を箱ひげ図にしたものである。同じレベルの技術分類でも分類の粒度が大きく異なることが分かった。この傾向はグループレベルで顕著であり，同じグループ内の特許でもその距離の中央値がセクションレベルよりも大きい(異質性が高い)ものが存在することが分かった。その一方で中央値がほぼ 0 であるものも存在し，前述した 0.2（グループレベルの均質性）という数字はあくまで中央値を示したものであることに留意が必要である。

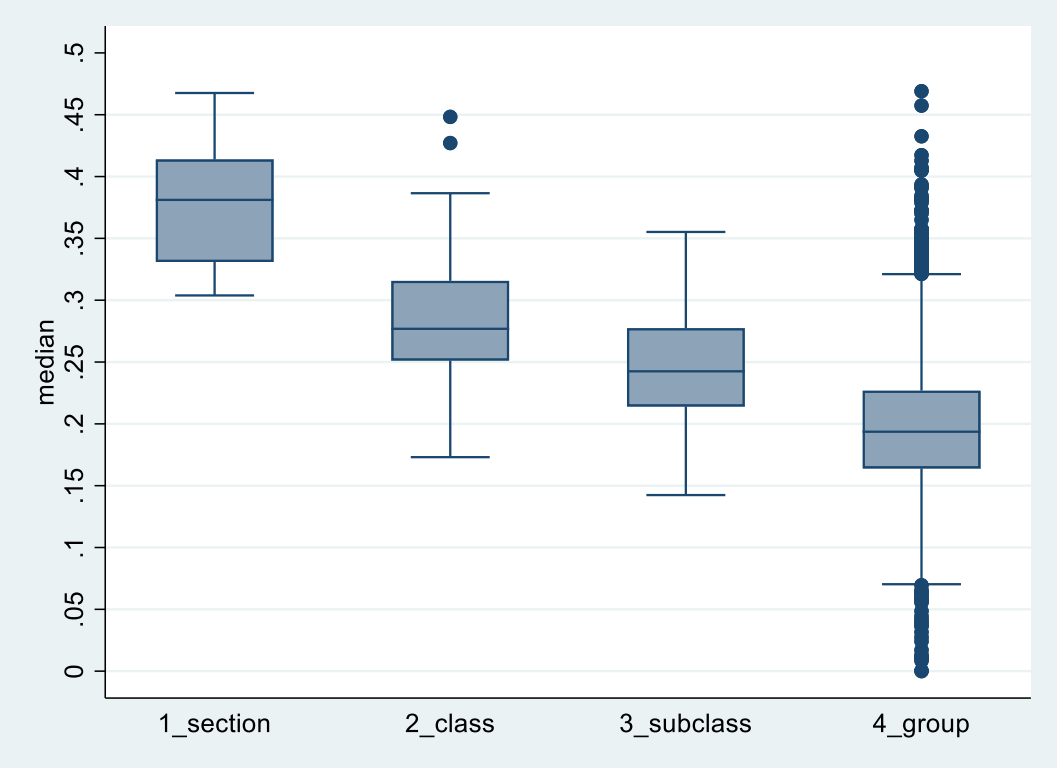


図 4－4：IPC レベル毎の距離中央値の分布

特許の引用・被引用ペアを用いて本研究で算出したベクトル情報の評価を行ったものが図 4－ 5 である。PATSTAT 2019 Spring Version から JPO の引用情報を取り出して，本研究における対象特許と接続した引用・被引用ペア 3,453,953 件について距離の四分位値を算出した。なお，IPC による評価結果と比較するために図 2－1 のグループレベルによる値を再掲している。更に引用・被引用ペアについて同じ IPC サブクラスに属するもの（1,316,546 件）と違う分類に属するもの（2,137,352 件）に分けて四分位値を算出したものも掲げている。結果については，引用・被引用ペアの距離は，IPC の最も詳細な分類による同一技術特許ペアよりさらに小さくなっている。JPO における引用情報は，審査官引用（特許審査官が審査過程において特許の新規性を判断するために先行特許文献を抽出したもの）であり，出願特許との内容の類似度が高いものを選別した結果といえる。従って，この結果は，本研究で算出した特許間の距離の有効性（Validity）をサポートするものといえる。なお，引用・被引用ペアについて IPC 分類の重なりの有無を見たものについて，重なり有のペアは距離がさらに小さくなるが，その差はそう大きくないことが分かった。

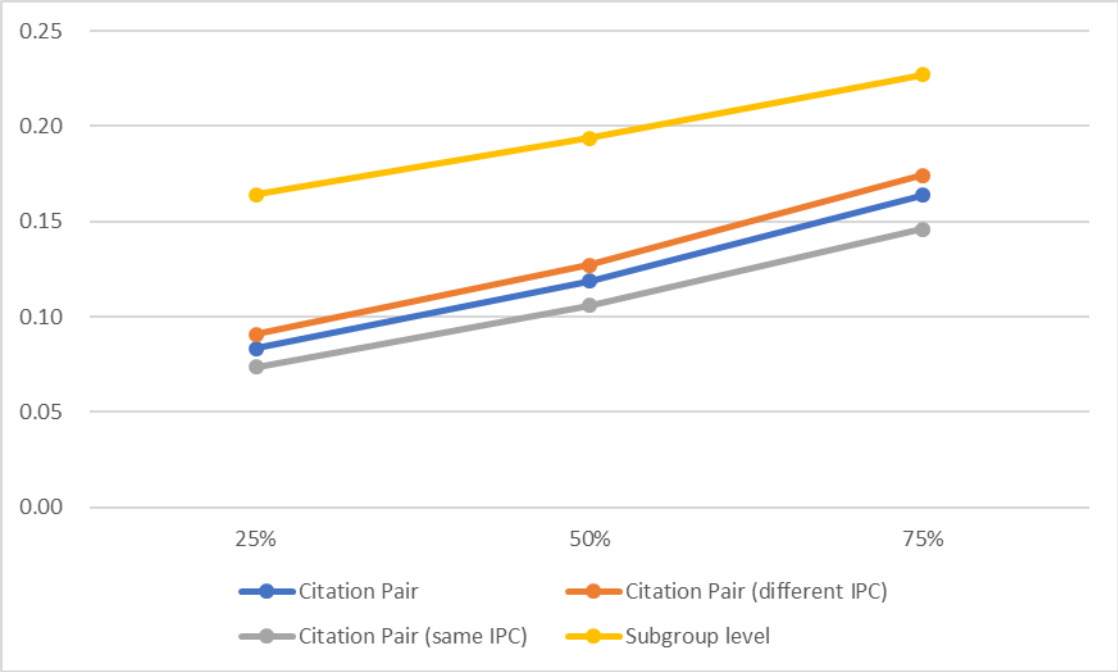


図 4－5 ：引用・被引用ペアの距離四分位値

次に NGT によって得られた近傍 200 特許（約 470 万件のそれぞれの特許について距離で 200 番目までのもの）の状況について見た。図 4－6 は近傍 200 特許のうち，距離が最も近いもの（1），10 番目のもの（10）及び距離が最も遠い（200 番目のもの，200）との間の距離について十分位値を見たものである。例えば，最も近いものとの距離については，全体の 10%（約 470 万特許のうち約 47 万件）のものが 0.007 以下であることを示している。同様に 200 番目のものの第一十分位値は 0.058 なので，近傍 200 番目までの距離を算出することで，全体の 90%の特許がその特許から距離 0.058 以下のものをすべて抽出できているということである。なお，200 件というのは 470 万件の 0.004%なので，図 4－1 の 10 万ペアのランダムサンプルでいうと約 4 件が 200 番目までの近傍特許の対象となる期待値である。ちなみに図 4－1 の 10 万サンプルペアの距離で 0.058 以下のものは 3 件となっており，ほぼこの期待値と合致した結果となっている。

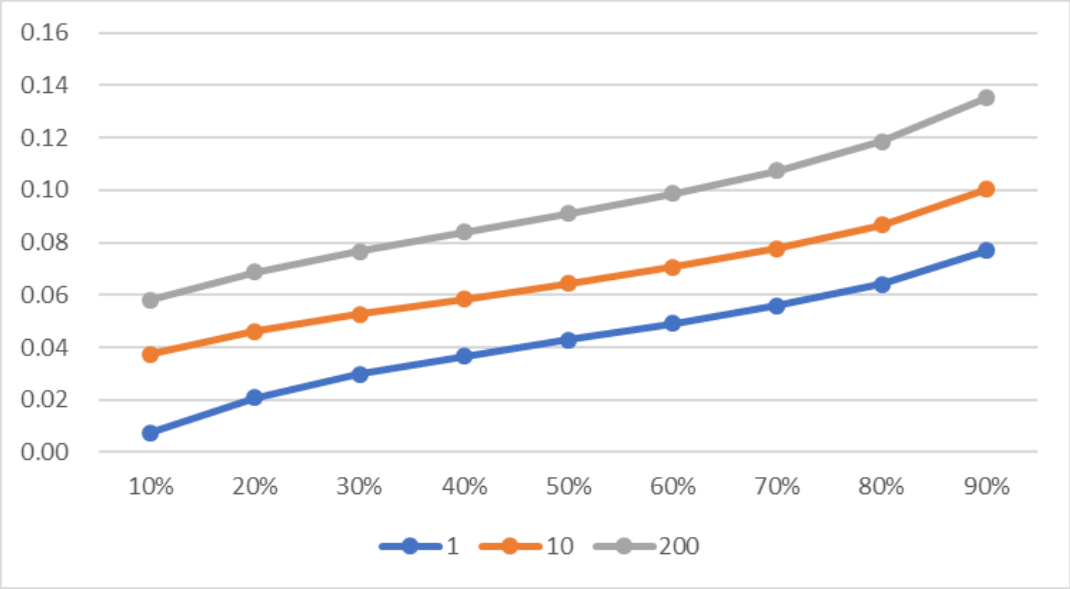


図 4－6：近傍 1，20，200 番目の距離十分位値

図 4－7 は近傍特許の距離の分布が同一出願人に属するものか否かによってどのように異なるのか見たものである。距離が最も近い特許との距離は，同一出願人によるものかどうかによってその値が大きく異なっている。出願人による特許出願内容の特性の影響を大きく受けることが分かった。ただし，200 番目の特許との距離については，やはり同一出願人による特許との距離が近い結果になっているが，両者の違いはかなり小さくなっている。

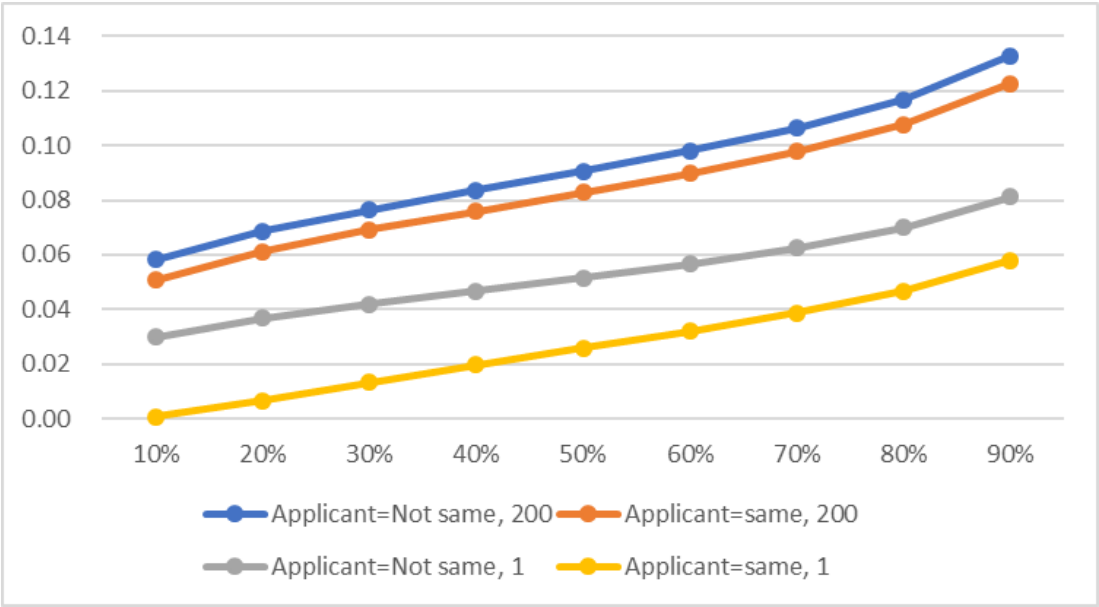


図 4－7：同一出願人か否かの違い

5. 近傍 200 特許を用いた出願人タイプ別の分析

本章では，近傍 200 件の特許のデータを用いて，出願人タイプ別の特許の特徴について分析をおこなう。図 5 － 1 は出願人タイプ（個人出願人，企業，公的研究機関及び大学）の違いによる 200 番目の特許との距離の分布を見たものである。全体として，個人出願人の距離がもっとも大きく，公的研究機関と大学がその次でほぼ同様の値，企業における距離が最も小さくなった。企業における出願特許は特定の技術スペースに集中しているのに対して，個人出願人はよりスパースな技術スペースに出願する傾向がある（公的研究機関・大学はその中間）ことを示している。

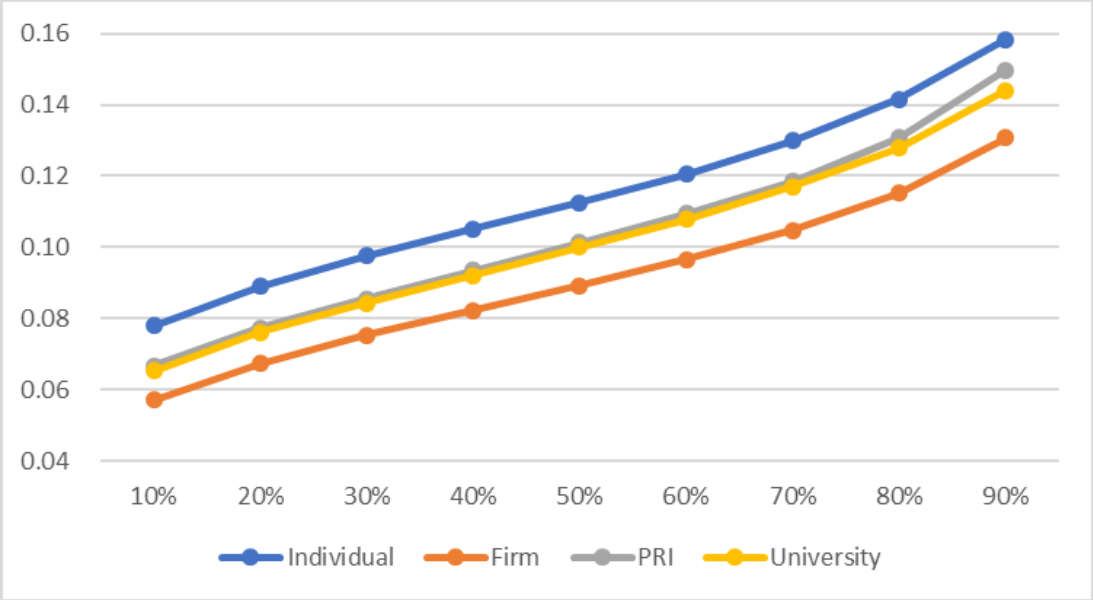


図 5－1：出願人タイプ別の違い

更に，各特許を個人（IND），企業（COM），公的研究機関（PRI），大学（UNI）に加えて産学連携（IUC）の 5 つのタイプに分類し，出願人のタイプ別に近傍特許との距離の分布を比較した。なお，各特許の出願日より前の 5 年以内に出願された特許との距離と出願日の後の 5 年以内に出願された特許との距離を比較する。そのため，比較の基準の特許は 2010 年に出願された特許に限定した。

図 5－2 は出願人タイプ別の出願前の 5 年以内の近傍特許との距離の分布を示している。企業の特許の距離の分布は左に寄っており，企業は出願時点で類似した特許が既に多く存在するスペースに特許を出願する傾向がある。一方，個人の特許の距離の分布は比較的右側に寄っており，個人発明家は類似した特許が比較的少ないスペースに特許を出願する傾向がある。他方，大学や公的研究機関，産学連携特許は企業と個人の中間に位置している。

図 5－3 は出願後の 5 年以内に出願された近傍特許との距離の分布を示している。出願前と比較して，産学連携特許の距離が短い特許が多く分布していることが特徴的である。

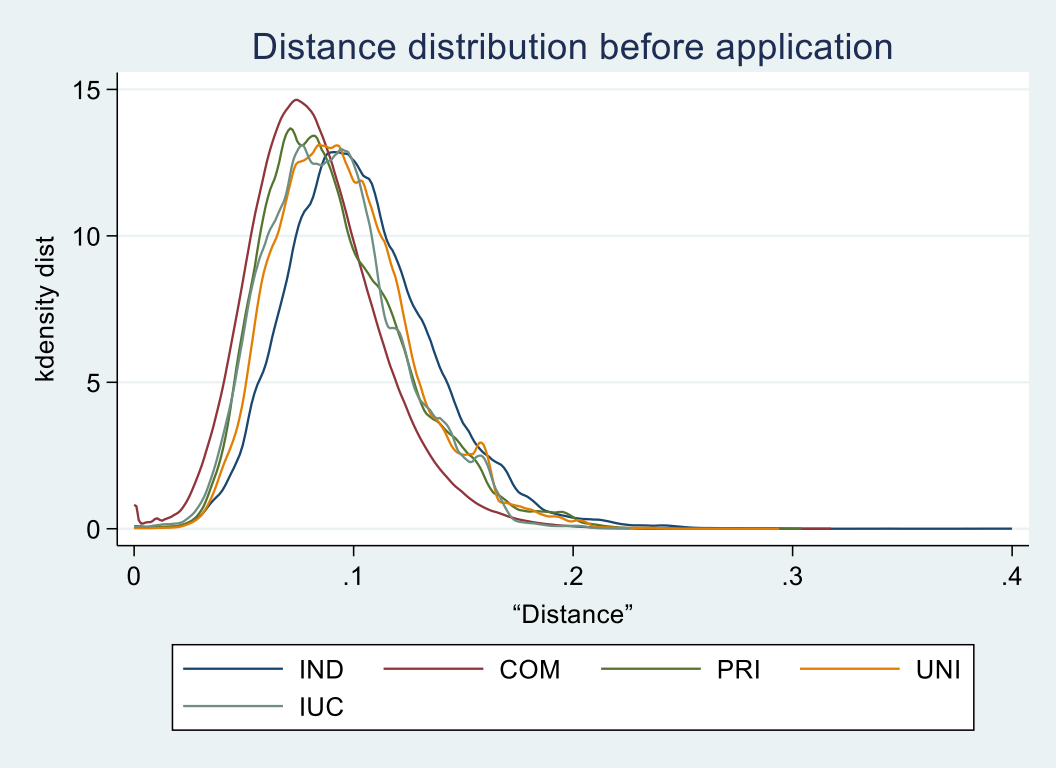


図 5－2：出願人タイプ別の出願前 5 年以内の近傍 200 特許との距離の分布

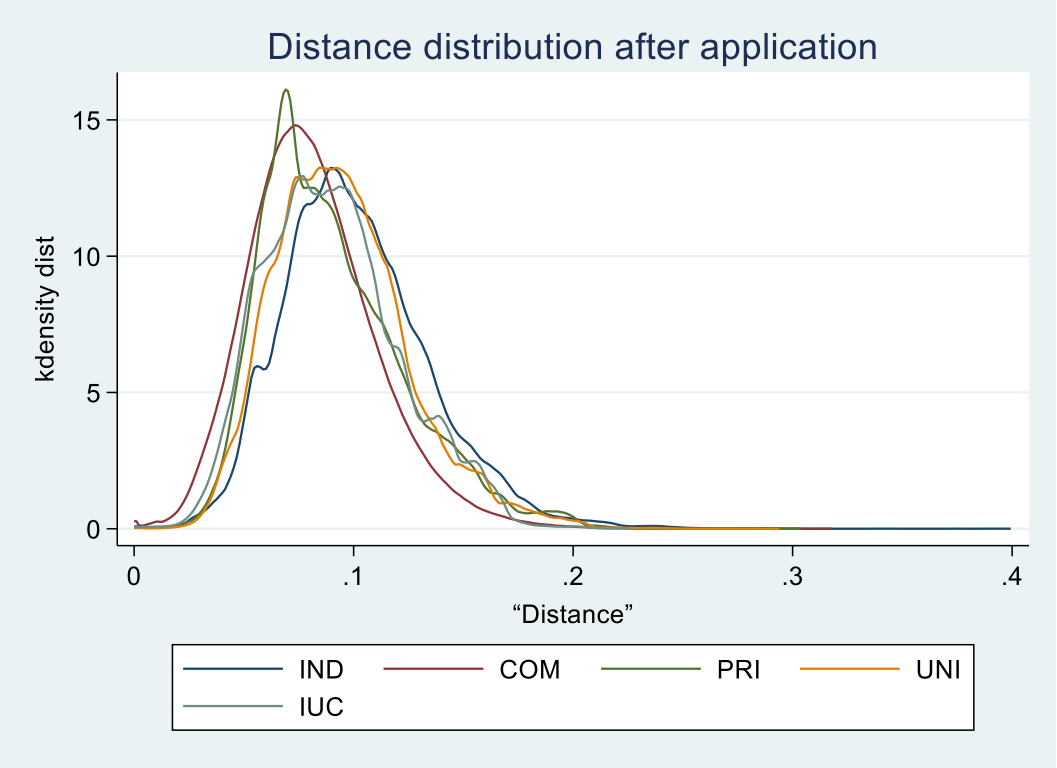


図 5－3：出願人タイプ別の出願後 5 年以内の近傍 200 特許との距離の分布

近傍距離を図 5－4，図 5－5 に示す。図 5－4，図 5－5 は出願前後 5 年以内に出願された特許のうち，距離が 0.05 及び 0.1 以内の近傍特許数の分布を出願人タイプ別に比較している。企業の特許は近傍特許の数が最も多く，個人の特許は最も近傍特許の数が少ないことがわかる。個人の次に近傍特許の数が少ないのは大学の特許であり，次に公的研究機関と産学連携特許は大学と企業の中間に位置している。

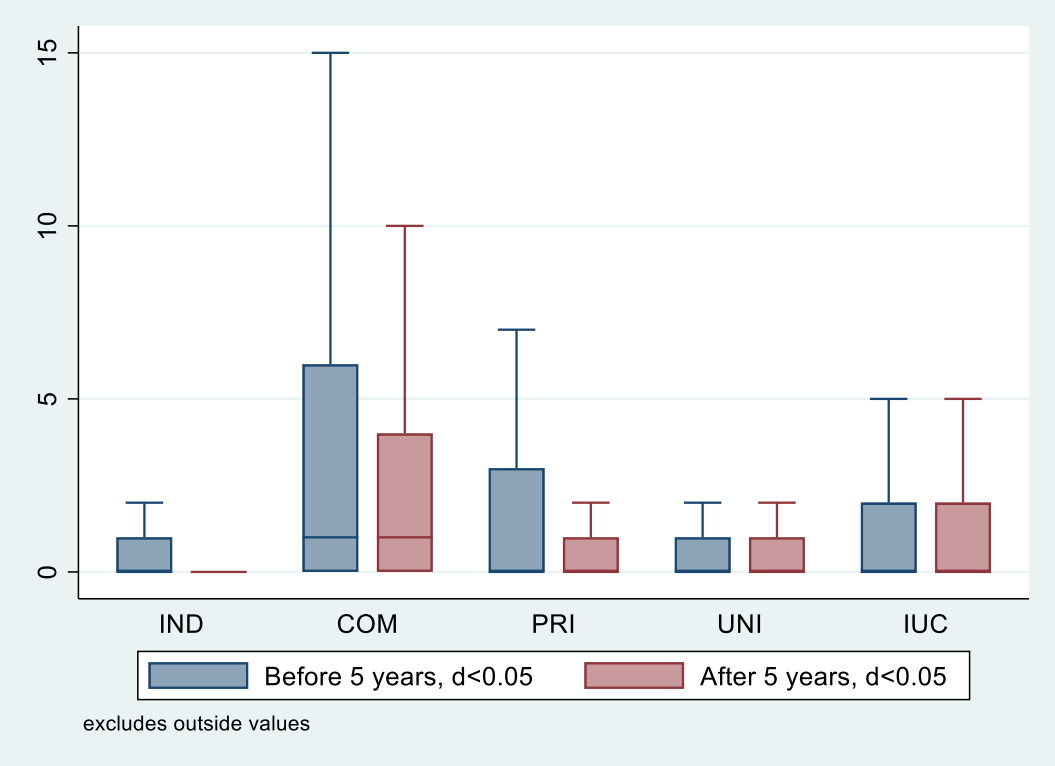


図 5－4：出願人タイプ別の出願前後 5 年以内の距離 0.05 以内の特許数

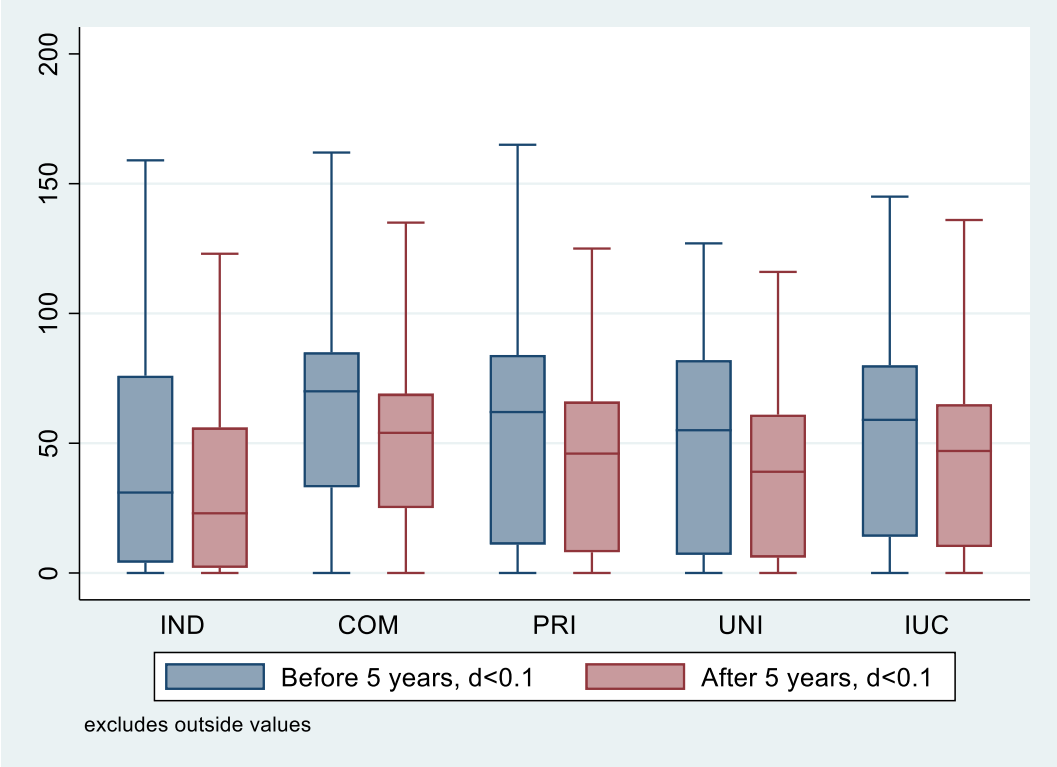


図 5－5：出願人タイプ別の出願前後 5 年以内の距離 0.1 以内の特許数

表 5－1 は距離が一定以内の近傍特許の数を従属変数とするポアソン回帰分析の結果を示している。説明変数は，出願人のタイプのダミー変数，登録ダミー（granted），IPC サブクラスのダミー変数である。出願人タイプのダミー変数の基準は大学としたため，その係数は大学出願特許と比べて近傍特許数が何倍かを示している。近傍特許の距離の基準を 0.01 未満と小さくした場合，IND（個人），COM（企業），公的研究機関（PRI），IUC（産学連携）の係数はいずれも正で有意となり，大学の特許は最も近傍特許の数が少ない傾向にあることがわかる。一方，近傍特許の距離の基準を 0.05 未満または 0.1 未満と大きくすると，INDIV の係数が負で有意となり，個人の方が大学よりも近傍特許の数が少ない傾向がみられる。この傾向は，出願前の 5 年以内の特許との距離の場合と出願後の 5 年以内の特許との距離の場合とで共通している。

また，COM の係数が最も大きく，技術分類の効果をコントロールしても，企業の特許の近傍特許数が最も多い傾向がみられる。IUC の係数は COM の係数に次いで大きく，産学連携特許は企業の特許に次いで近傍特許数が大きい。PRI（公的研究機関）の係数は近傍特許の距離の基準が 0.01 及び 0.05 の場合には正で有意だが，近傍特許の距離の基準が 0.1 の場合には負で有意となった。

加えて，近傍 200 件の特許を抽出しているため，近傍 200 特許での最大距離が近傍特許の距離の基準を下回っている場合，近傍特許数が下方にバイアスがかかってしまう（近傍特許数が右側切断されてしまう）。そのため，近傍特許数が右側切断されている特許を分析から除いた推定結果を表 5－2 に示し，推定結果の頑健性を確認した。表 5－2 の推定結果は表 5－1 と整合的であり，上記の結果の頑健性は認められる。

表 5－1：前後 5 年以内に出願された近傍特許数のポアソン回帰分析結果

	Before 5 years			After 5 years		
	d<0.01	**d<0.05**	**d<0.1**	**d<0.01**	**d<0.05**	**d<0.1**
IND	0.938***	-0.439***	-0.280***	1.051***	-0.432***	-0.316***
	[0.120]	[0.015]	[0.004]	[0.140]	[0.017]	[0.004]
COM	1.689***	0.765***	0.132***	1.599***	0.876***	0.180***
	[0.103]	[0.010]	[0.003]	[0.122]	[0.011]	[0.003]
PRI	0.322***	0.084***	-0.013***	0.527***	0.083***	-0.028***
	[0.124]	[0.013]	[0.004]	[0.151]	[0.016]	[0.004]
UNI	0.000	0.000	0.000	0.000	0.000	0.000
	[.]	[.]	[.]	[.]	[.]	[.]
IUC	0.400***	0.424***	0.037***	0.594***	0.528***	0.055***
	[0.127]	[0.012]	[0.004]	[0.146]	[0.013]	[0.004]
granted	1.118***	0.055***	-0.006***	0.868***	0.119***	0.050***
	[0.009]	[0.001]	[0.001]	[0.013]	[0.001]	[0.001]
Constant	-22.740	1.176***	2.982***	-24.500	0.674**	2.705***
	[13580.500]	[0.209]	[0.078]	[45769.200]	[0.268]	[0.091]
IPC subclass	Yes	Yes	Yes	Yes	Yes	Yes
N	254,651	254,651	254,651	254,651	254,651	254,651

Standard errors in brackets

* p<0.1, ** p<0.05, *** p<0.01

表 5－2：前後 5 年以内に出願された近傍特許数のポアソン回帰分析結果
（近傍特許数が切断されているケースを除く）

	Before 5 years			After 5 years		
	d<0.01	**d<0.05**	**d<0.1**	**d<0.01**	**d<0.05**	**d<0.1**
IND	0.938***	-0.462***	-0.147***	1.051***	-0.467***	-0.242***
	[0.120]	[0.019]	[0.008]	[0.140]	[0.021]	[0.009]
COM	1.689***	0.798***	0.284***	1.599***	0.863***	0.274***
	[0.103]	[0.013]	[0.007]	[0.122]	[0.014]	[0.007]
PRI	0.322***	0.305***	0.042***	0.527***	0.274***	-0.030***
	[0.124]	[0.017]	[0.009]	[0.151]	[0.019]	[0.010]
UNI	0.000	0.000	0.000	0.000	0.000	0.000
	[.]	[.]	[.]	[.]	[.]	[.]
IUC	0.400***	0.454***	0.135***	0.594***	0.507***	0.156***
	[0.127]	[0.016]	[0.009]	[0.146]	[0.017]	[0.010]
granted	1.118***	0.031***	0.005***	0.868***	0.120***	0.071***
	[0.009]	[0.002]	[0.001]	[0.013]	[0.002]	[0.002]
Constant	-22.740	1.200***	2.849***	-24.500	0.708***	2.631***
	[13580.500]	[0.209]	[0.079]	[45769.200]	[0.268]	[0.092]
IPC subclass	Yes	Yes	Yes	Yes	Yes	Yes
N	254,651	240,941	92,783	254,651	240,941	92,783

Standard errors in brackets

* p<0.1, ** p<0.05, *** p<0.01

上記の分析の結果，近傍特許の距離の基準によって，大学と個人及び公的研究機関の近傍特許の傾向が異なることがわかった。そこで，以下では最近傍特許との距離に注目して分析をおこなった。図 5－6 は 2010 年に出願された特許について，出願時点での最近傍特許との距離の分布を出願人タイプ別に比較している。大学（UNI）の特許の分布が右側に寄っており，局所的にみると大学の特許は比較的スパースな領域に出願されている傾向がみられる。

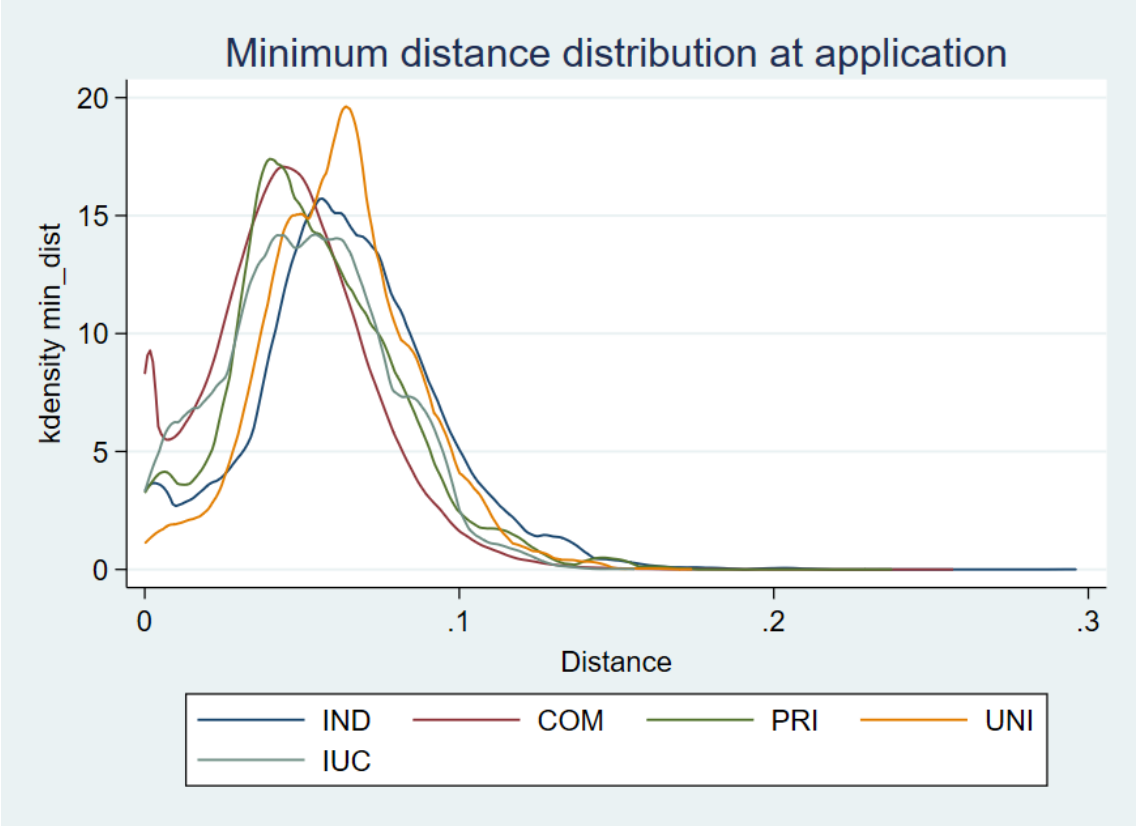


図 5－6：出願時点での最近傍特許との距離の分布

上記の傾向をより厳密に分析するため，出願時点での最近傍特許との距離を従属変数とし，出願人タイプのダミー変数，登録特許ダミー，IPC サブクラスのダミー変数を説明変数とする回帰分析をおこなった。表 4－3 の第 1 列は 2010 年に出願された特許のデータを用いた結果を示しており，大学以外の出願人タイプを示すダミー変数の係数はすべて有意に負であり，大学の特許は他の出願人タイプと比較して最近傍特許との距離が最も遠い傾向が確認される。出願人タイプの中で COM の係数が最も小さく，最近傍特許との距離が短いのは企業の特許であることがわかる。次いで，公的研究機関（PRI）と産学連携（IUC）の特許の最近傍特許との距離が短い。

表 5－3 の第 2 列は近傍 200 特許に出願前の特許が含まれない特許を除いた 2010 年出願特許の分析結果，第 3 列・第 4 列は 2004 年以降の出願特許を全て含んだ分析結果を示しているが，上記第 1 列の結果と概ね一致した傾向を示している。

表 5－3：出願時点での最近傍特許との距離に関する回帰分析結果

	[1] year=2010	[2] year=2010 cens==0	[3] year>=2004	[4] year>=2004 cens==0
IND	-0.095***	-0.088***	-0.070***	-0.058***
	[0.028]	[0.028]	[0.008]	[0.007]
COM	-0.391***	-0.377***	-0.354***	-0.335***
	[0.023]	[0.023]	[0.006]	[0.006]
PRI	-0.268***	-0.275***	-0.138***	-0.135***
	[0.032]	[0.032]	[0.009]	[0.009]
UNI	0.000	0.000	0.000	0.000
	[.]	[.]	[.]	[.]
IUC	-0.178***	-0.178***	-0.170***	-0.168***
	[0.032]	[0.031]	[0.009]	[0.009]
granted	-0.086***	-0.084***	-0.116***	-0.110***
	[0.005]	[0.005]	[0.001]	[0.001]
Constant	-2.881***	-2.888***	-2.325***	-2.347***
	[0.357]	[0.348]	[0.081]	[0.078]
IPC subclass	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
N	249,754	236,791	3,471,632	3,293,477

Standard errors in brackets

* p < 0.1, ** p < 0.05, *** p < 0.01

上記の結果をまとめると，まず企業の特許は大学や個人の特許に比べて出願時点において内容が他の特許との類似性が比較的高く，出願後も近傍に多くの特許が出願される傾向が一貫して観察される。一方，最近傍エリアを局所的にみれば個人や公的研究機関に比べて，大学の特許は内容の類似した特許が少ないスパースな技術スペースに出願している傾向がみられる。しかしながら，距離空間をやや大局的にみると大学の特許よりも個人や公的研究機関の特許の方がスパースな技術スペースに位置している傾向がみられた。

6. まとめ

本稿では，特許情報をベクトル表現し，それらに統計数理手法を適用することで，網羅的な分析を行う方法と，分析の結果について述べた。

結果として，提案手法は有効に機能し，分析を行う上で必要なデータを得ることができた。実験・分析には JPO の公開特許公報情報におけるタイトルと要約文を用いて，特許内容のベクトル空間モデルを作成した。また，この情報に基づいてクラスタリングと近傍特許の抽出・距離の測定を行い，その内容について IPC 分類や引用情報を用いて検証を行った。

ベクトル空間モデルを用いた特許間の距離は，同一の技術分類内で比較対象特許を抽出することで小さくなるはずである。本稿で作成された特許間の距離は，より詳細な IPC 分類内で見るほど小さくなっており，ベクトル空間モデルの妥当性を確認することができた。なお，IPC 分類の最も詳細な分類であるグループレベル（分類数：約 6 万）内での特許間の距離は中央値で約 0.2 であるが，一方で NGT による近傍特許検索による 200 番目の特許までの距離は 90%タイル値で見ても 0.14 となっている。つまり，IPC 分類は数百万件に及ぶ特許の技術的特性を示すものとして非常に粗いものにすぎず，特許の内容に関する細かな分析を行う際にはベクトル空間モデルを活用することが有用であることを示唆している。また，特許の引用・被引用の関係にある特許については，特許ペアが同一 IPC グループに属するか否かにかかわらず，IPC グループ内全体を見た距離より小さい値に収まっていることが確認できた。この点については，ベクトル空間モデルによる値は IPC 分類と比べて特許の類似性をより的確に示しているとする[Arts 17]の結果をサポートするものである。

一方で，特許の要約文を用いたベクトル空間モデルは，出願人(要約文の作成者)の用語の使い方などの「書き方」の影響を受ける可能性がある。この点については，単語の分散表現を用いることである程度は補正されていると考えられるが，出願人が同一か否かで特許距離の違いをみたところ，最も近い特許については両者の距離分布が大きく異なることが分かった。だだし，この違いは 200 番目の特許を見るとかなり小さくなる。同一の出願人の場合，一般的に内容的にも似通った発明が多いことが想定されるので，距離分布が異なること自体がバイアスであるとは言えない。また，非常に似通った特許を大量に出願することで権利の強化を図る行動（パテントフェンス）が，一部の業界で確認されている。その場合，最も距離的に近い特許においてより距離分布の差が大きいことも頷ける。ただし，本稿では，表現方法の違いによってベクトル空間値がどのような影響を受けているのかについて，システマチックな分析を行っておらず，この点については今後の検討課題としたい。

本稿で作成した特許テキスト情報のベクトル表現モデルや NGT モデルによる近傍特許検索結果は様々な研究テーマに活用可能な基盤的な情報といえる。ここでは，技術空間

における個々の特許の分布密度に着目して，特許出願人のタイプ（個人，企業，大学，公的研究機関）による特許内容の違いについて分析を行った。具体的には，NGT による近傍 200 特許検索結果を用いて，上記の 4 種類の出願人タイプと産学連携特許をあわせた 5 種類の特許の近傍特許の分布状況を見た。ここで近傍特許が対象となる特許の出願前のものか，出願後のものかでインプリケーションが異なる。前者の場合は，その密度が大きい（小さい）場合は，技術的に混雑した（空いた・スパースな）領域に出願されたものであると解釈できる。一方で後者の場合は，その密度が大きい(小さい)場合は，当該特許の近くに多くの（少ない）特許出願がなされたインパクトの大きい（小さい）ものと解釈できる。

その結果，出願前を見た場合でも，出願後を見た場合でも，平均的な距離の順番としては，企業＜産学連携＜公的研究機関＜大学＜個人となることが分かった。企業は出願が盛んにおこなわれるメインストリームの技術領域に出願する傾向があり，大学・公的研究機関や個人は独自性の高い領域に出願するものの出願後のインパクトも小さいという全体的な傾向を示している。これは教員や研究室単位で curiosity driven な研究を行う大学や，その傾向がより強い個人発明家の成果として納得のいく結果といえよう。ただし，近傍特許の中でもより距離の小さい局所的な分布をみると，大学は個人発明家と比べてよりスパースな領域に出願していることが分かった。大学における研究成果は，学術論文として広く公開されることが一般的である。さらに，既存研究と比較して，新規性の高い研究を行うことが求められる。一方個人発明家は，基本的には個人の思い付きがベースになるので，たまたま非常によく似た発明がなされることが少なくないと想定される。一方で，大学においてもある程度研究内容の流行り（廃れ）があることは間違いない。従って，ある程度探索領域を広くとると，大学の特許の方がより密度が高いところに出願される傾向になると考えられる。ただし，ここで見た出願人タイプ別の特許特性は，あくまで全体的な動向を示したものであり，今後その内容を細かく見ていくことが必要と考える。

参考文献

[Arthur 07] 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

[Arts 17] Arts, S., Cassiman, B., and Gomez, J. C.: Text Matching to Measure Patent Similarity, Strategic Management Journal, Vol. 39, (2017)

[Bojanowski 17] 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

[Dai 15] Dai, A. M., Olah, C., and Le, Q. V.: Document embedding with paragraph vectors, CoRR, Vol. abs/1507.07998 (2015)

[Joulin 16] 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

[Lau 16] Lau, J. H. and Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation, CoRR, Vol. abs/1607.05368 (2016)

[Le 14] Le, Q. V. and Mikolov, T.: Distributed representations of sentences and documents, CoRR, Vol. abs/1405.4053 (2014)

[McInnes 18] McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv preprint (2018), arXiv:1802.03426

[Sato 15] Sato, T.: Neologism dictionary based on the language re- sources on the Web for Mecab (2015), https://github.com/neologd/mecab-ipadic-neologd

[Sato 16] Sato, T., Hashimoto, T., and Okumura, M.: Operation of a word segmentation dictionary generation system called NEologd, in Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL), pp. NL–229–15, Information Processing Society of Japan (2016), (in Japanese)

[Sato 17] Sato, T., Hashimoto, T., and Okumura, M.: Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval, in Proceedings of the Twenty-third Annual Meeting of the Association for Natural Language Processing, pp. NLP2017–B6–1, The Association for Natural Language Processing (2017), (in Japanese)

[Shen 18] Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y., Li, C., Henao, R., and Carin, L.: Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms, arXiv preprint (2018), arXiv:1805.09843

[WIPO 19] WIPO: Artificial Intelligence, WIPO Technology Trends 2019 (2019)

[Younge 16] Younge, K. A. and Kuhn, J. M.: Patent-to-Patent Similarity: A Vector Space Model, SSRN (2016)

[岩崎 13] 岩崎 雅二郎:商品画像検索へのグラフ構造型インデックスの適用, 画像電子学会誌, Vol. 42, No. 5, pp. 633–641 (2013)

[元橋 18] 元橋 一之:AI におけるサイエンスとイノベーションの共起化:米国における論文・特許データベースを用いた分析, NISTEP DISCUSSION PAPER, No. 160 (2018)

[小柴 19] 小柴 等, 森川 想:議事録を用いた我が国における議会・行政の関係性分析手法, 人工知能学会論文誌, Vol. 34, No. 5, pp. E–J47_1 – 10 (2019)

[椿 19] 椿 光之助, 小柴 等, 赤池 伸一:STI for SDGs に関する政策レビュー及び研究助成との関連づけへの人工知能（AI）関連技術の試行的活用, NISTEP DISCUSSION PAPER, No. 174 (2019)

[科学 07] 科学技術政策研究所 科学技術動向研究センター:サイエンスマップ 2004, NISTEP REPORT, No. 100 (2007)

[科学 08] 科学技術基盤調査研究室:サイエンスマップ 2006, NISTEP REPORT, No. 110 (2008)

[科学 14] 科学技術・学術基盤調査研究室:サイエンスマップ 2010&2012, NISTEP REPORT, No. 159 (2014)

[科学 16] 科学技術・学術基盤調査研究室:サイエンスマップ 2014, NISTEP REPORT, No. 169 (2016)

[科学 18] 文部科学省 科学技術・学術政策研究所:サイエンスマップ 2016, NISTEP REPORT, No. 178 (2018)

[佐藤 17] 佐藤貢司,安井基陽,田中厚子,中村昭博,中田守: 被引用情報を用いた重要特許抽出方法の検証:, 情報プロ フェッショナルシンポジウム予稿集, Vol. 2017, pp. 61 – 65 (2017)

[富澤 06] 富澤宏之,林隆之,山下泰弘,近藤正幸:有力特許に引用された科学論文の計量書誌学的分析, 情報管理, Vol. 49, No. 1, pp. 2–10 (2006)

[富永 18] 富永 泰規, 久々宇 篤志:特許文献への分類付与における付与根拠箇所推定, 情報の科学と技術, Vol. 68, No. 7, pp. 338 – 342 (2018)

[槫松 14] 槫松 理樹:特許構成を考慮した文書類似度に基づく特許からの課題分類・手段分類推定システム, 人工知能学会 全国大会論文集, Vol. JSAI2014, pp. 1A32 – 1A32 (2014)

DISCUSSION PAPER No.175

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

2019 年 12 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

A method of extracting content information from patent documents and comparison of their characteristics by applicant type by using the vector space model of distributed expressions

Dec 2019

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

https://doi.org/10.15108/dp175

https://www.nistep.go.jp