論　説

知財情報活用のための データサイエンス手法の研究

情報活用委員会
第４小委員会*

抄　録　ビジネスにおける意思決定は情報分析結果に基づくことが一般的になっている。この情報分析において，「データサイエンス」が近年注目されているが，その実態が正しく認識されないままイメージだけが独り歩きしている，いわゆるバズワード化の感がある。そこで本研究では，データサイエンスの全体像を把握・整理したうえで，知財分野での活用事例の研究を行った。具体的には，データサイエンスの概要をまとめたうえで，データサイエンスの知財課題への適用について整理し，回帰分析・テキストマイニング等による特許件数予測やクラスタ分けを試行した。活用事例研究にはツールベンダーが有償で提供する分析ツール，フリーソフトウェア，表計算ソフトウェアの他，近年個々人の環境や分析目的に合わせて容易に作成が可能となってきた，オープンソースを使ったプログラミングによるツールを使用し，その適用可能性を検討した。

目　次

1．はじめに

様々な情報が容易に入手可能となった近年，ビジネスにおける意思決定は経験や勘ではなく情報分析結果に基づくことが一般的になった。

「データサイエンス」（以下「DS」）は情報分析に関連して注目され，ネット上をはじめとする多くの場所で目にするが，その実態が正しく認識されないままイメージだけが独り歩きしている，いわゆるバズワード化の感があり，我々がその実態を正しく認識しているか懸念がある。

本研究ではDSの基本と最近の手法を中心に全体像を把握し，知財情報を活用した分析手法の研究を行った。

なお，本研究は，2020年度情報活用委員会第４小委員会第１ワーキンググループの高橋祐二（小委員長，リコー），落合昌孝（富士フイルムビジネスイノベーション），川村将郎（テルモ），篠崎直樹（東ソー），峯尾泰（富士フイルム知財情報リサーチ），村松慎吾（帝人）によるものである。

* 2020年度　The Fourth Subcommittee, IP Intelligence Committee

2．データサイエンスの概要

ひとくちにDSと言っても，それが何を示すのか，曖昧模糊とした部分が多い。具体的にどうやって取り組むのか（“How”）の前に，DSとは何か（“What”），その定義の再確認と現状把握を行った。

2. 1 定　　義

DSとは，データを使って問題・課題を解決する手段，すなわちデータから何らかの価値を引き出すことであると一般的には定義されている[1)]。本研究においても，この定義をもとに，関連する分析手法を検討した。

まず，DSの概要についてまとめてみる。

DSには「データ処理」,「データ分析」,「価値創造」の３つの要素が含まれ，図１のようなベン図[2)]として概念的に示すことができる。

「データ処理」とは，データを収集・調整して分析に使える形にし，実装，運用する力を示す。

「データ分析」とは，統計・数学などの情報科学系の知識を理解し，データを使う力を示す。

「価値創造」とはデータの課題背景を理解した上で，ビジネス課題を整理し解決する力を示す。

2. 2 手　　法

つぎに, DSの“What”に次いで, DSの“How”を概観する。

AIに関しては，終焉を迎えつつあるといった論調もあるものの，いまだ第三次ブーム[3)]のただ中にあり，DSにおいて機械学習は一般的なツールとなってきた。ここでは図１の各要素について，以下，文献[4)~6)]の必要箇所やWeb情報を参考に，独自にDSの手法をとりまとめた。その結果を，表１に例示した。

以下，(1) 機械学習，(2) データエンジニアリング，(3) 伝統的な研究に大別して具体的な手法を概説する。

(1) 機械学習

様々な手法が存在し，また日々更新されている領域であるが，大きくは「教師あり学習」と「教師なし学習」に大別できる。「教師あり学習」はさらに連続するデータの予測を行う「回帰」といくつかのクラス分けを行う「分類」に分けられる。また，それぞれの手法を用いた計算を

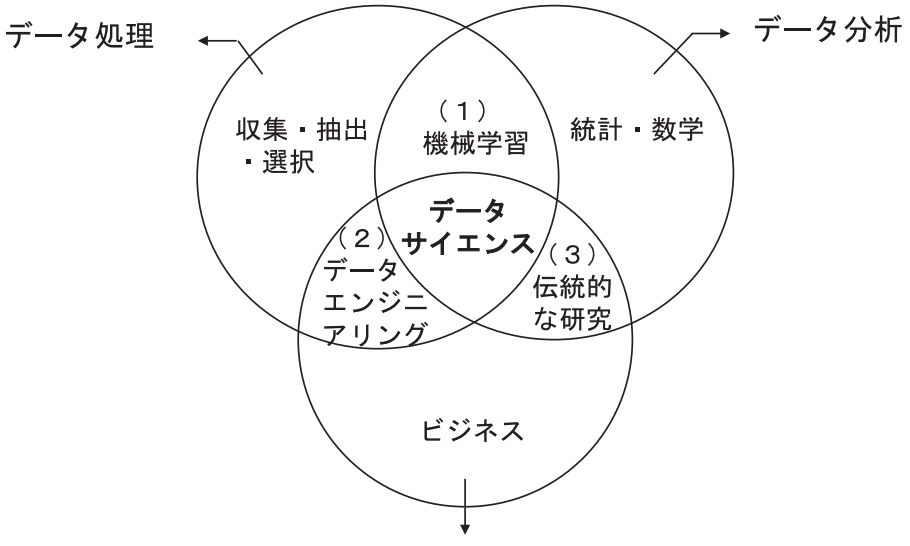

図１　データサイエンスが扱う内容

行うための複数のアルゴリズムが知られている。例えば「教師あり学習」で「分類」を行うためのアルゴリズムとしては，サポートベクタマシンやニューラルネットワークなどが，「教師なし学習」に関してはｋ平均法などが知られている（表１）。

(2) データエンジニアリング

分析に用いるデータの前処理段階として，「データ収集」「前処理」「データ削減」がこの領域に該当する。

「データ収集」は具体的にはスクレイピング等があり，「前処理」は，得られたデータを例えば機械学習に適したデータとするための修正であるクレンジング（知財関連でいえば出願人の名寄せ等）がある。また「データ削減」のための主成分分析（次元圧縮）もこの領域に含まれる。

なお，前処理のみでなく，データの可視化などの後処理段階として，グラフネットワーク化等を含む分類「ネットワーク」も，ここに含めることができる。

(3) 伝統的な研究

AIブーム以前から知られている「データマイニング」手法が該当し，具体的にはTF-IDFやWord2vec等が知られている。

3．知財課題への応用

3. 1 DSを活用した分析に必要な考え方や知識，スキルの整理

分析に先立ち，DSの活用における留意点について情報分析に関する有識者２名およびAIツール開発者１名にヒアリングを行った。以下，ヒアリングにおいて得られた知見を当委員会で要約した概要を説明する。

(1) DSの位置づけ

客観的事実に基づいた意思決定のためには，DS等を通じてデータで語るのは重要であるが，データ分析ですべてが分かるわけではない。DSは意思決定のための材料の一つであり，どこまでDSで明らかにして，どこからDSによる分析結果に基づいて人間が考える必要があるのか，その境界を踏まえたうえでDSを活用すること

表１　データサイエンス手法のまとめ

手法	分類１	分類２	アルゴリズム例
(1) 機械学習	教師あり	分類	決定木，ランダムフォレスト，サポートベクタマシン，ナイーブベイズ，ロジスティック回帰，ニューラルネットワーク
		回帰	線形回帰，非線形回帰，サポートベクタマシン，ｋ近傍法
	教師なし	クラスタリング	ｋ平均法，混合ガウスモデル
(2) データエンジニアリング	データ収集	スクレイピング	－
	前処理	クレンジング	－
		プロファイリング	－
	データ削減	次元圧縮	主成分分析，トピックモデル
	ネットワーク	リンク予測	－
		グラフネットワーク化	－
(3) 伝統的な研究	データマイニング	類似度	TF-IDF，Word2vec
		分類	－

が重要である。分析結果の解釈においては，データに基づく客観性と，意思決定に役立つように分析の目的に沿ってデータを解釈していく主観性を両立させる必要がある（情報分析有識者）。

(2) DSを活用した分析に必要な考え方

どのような手法を用いる場合においても，分析の5W2H：分析の目的（Why），報告テーマ（What），報告期限（When），報告形態（Where），実施者／報告先（Who），ツール（How），コスト（How much）を押さえることは必須である。

DSを活用した分析では①データの前処理（データエンジニアリング），②分析手法の適用，③結果の解釈，の３段階があり，それぞれ別の人が関わることもある。

特許情報は，項目別に構造化されたデータとして提供されているため，その分析においてはそれほど問題として顕在化しないが，構造化されていないデータを用いる場合，①データの前処理を行ってデータを分析可能な形式に変換する作業が必要となる。このような場合は特にデータエンジニアリングが重要となる。

また，最終的に意思決定する経営層に向けて，分析結果を適切に伝えるには，③結果の解釈がDS活用の重要なポイントとなる。経営層が判断できるように，結果が何を意味するのか，解釈まで含めた情報提供を行い，結果的に正解でなかったとしても自分の考えを持って提案することが必要である（情報分析有識者）。

(3) 機械学習を活用した分析に必要な考え方

機械学習を用いる場合，②分析手法の適用において，モデルを作りこんでいく。その際は，分析の目的の理解が重要となる。特に予測結果の報告においては，その根拠を納得してもらう必要があるため，分析手法や予測精度等を説明するが，精度は100％となることはあり得ない。そこで，精度の数字にこだわるよりも，分析の目的に応じて「人がやるよりは良い（精度面・効率面で）」というところで使いこなす，という考え方で臨むのが良い（AIツール開発者）。

3. 2 知財課題と対応するDS手法例の整理

前節に記載のヒアリング内容を参考にして，DS手法の知財分野への活用を検討するにあたり，主要な知財課題を抽出し，それぞれに対応する分析とDS手法を整理した。その結果を表２に示す。なお，表２のDS手法例は文献[7)〜15)]等を参照して独自に取りまとめた。

(1) 開発動向将来予測，市場予測

競合企業の技術開発動向や，自社製品を含む特定技術分野の技術動向を把握することは，特許調査担当者の主要業務の一つである。IPランドスケープの広がってきた近年においては，これまで行われてきた「過去から現在までの技術動向把握」に加え，さらに「将来の技術予測や市場の予測」も知財担当者に求められてきている。

この場合，過去のデータから将来を予測するために，過去のデータを教師データとして学習モデルを作成し，作成されたモデルを用いて将来の予測値を算出するDS手法が考えられる。通常，市場がある時点から拡大または縮小するといった二値的変数や非連続的な変数を予測する場合は分類，特許出願件数といった連続的な変数を予測する場合は回帰と呼ばれ，具体的には表１に記載の手法等を用いる。また株価等の時系列的な周期性を有する連続データを用いる場合は，時系列分析と呼ばれる手法を用いる。

(2) 新規技術探索，用途探索

オープンイノベーションの重要性が高まっている昨今において，社外の有望技術の探索は事業的にも大きな意味を持つ。また技術成熟分野においては，自社技術を既存の市場とは別の市場へ展開することによる市場拡大も事業継続の

表２　知財課題と対応するDS手法例の整理

No.	知財課題	分析内容例	データサイエンス手法例
(1)	開発動向将来予測	・競合企業の技術開発動向を予測 ・特定分野の技術動向を予測	・使用データ：特許，非特許（論文，市場情報，企業情報等） ・手法：回帰，分類 ・分析例(回帰)：過去の競合企業の特許出願動向データを用いて，次年以降の特許出願件数を予測 ・分析例（分類)：過去の市場規模推移データを用いて，市場が今後拡大するか縮小するかを予測
	市場予想	・特定分野(技術，事業)の今後の市場拡大・縮小を予想	
(2)	新規技術探索	・新規な技術を保有するスタートアップ企業を探索 ・自社に適合する新規出願／研究開発テーマを探索	・使用データ：特許 ・手法：テキストマイニング ・分析例：特定分野における特許データのテキストマイニングにより特徴的なキーワードを有する技術を抽出 ・分析例：自社を含む特定分野の特許データをテキストマイニングすることにより自社技術と類似度の高い技術や用途を抽出
	用途探索	・自社技術の新規な用途，展開分野，開発ターゲットを探索	
(3)	基本特許探索	・特定技術分野の基本特許を見つける ・特定特許が基本特許である／となる可能性を判断	・使用データ：特許 ・手法：ネットワーク分析 ・分析例：引用データを用いて引用数，引用分野の広がり等を算出
(4)	特許対応製品情報探索	・特許対応製品情報を見つける ・製品の特許侵害を判定	・使用データ：特許，非特許（webデータ等） ・手法：スクレイピング ・分析例：web製品情報を網羅的に収集し，特許記載と照合

ための重要な施策であり，知財・非知財情報を用いて展開先となる分野を探索することもIPランドスケープに求められる内容である。

多くのデータから未知の情報を探索する手法は，データマイニングと呼ばれる手法としてこれまでも発展してきた。

このうちテキスト情報を用いるものは特にテキストマイニングと呼ばれる。特許情報は豊富な技術情報を含むテキスト情報である。特許のテキスト情報にテキストマイニングを行って個々の特許文献の類似度や距離を計算して図示することによって，特許分類等よりも詳細に技術を分類して俯瞰することができる場合がある。また作成された俯瞰図から，自社技術と近いが想定していなかった新たな用途や，自社が気づいていなかった特徴的な他社技術を発見できる可能性がある。

(3) 基本特許探索

新規技術領域に参入しようとするとき，その分野における技術理解及び侵害予防のために当該分野の基本特許を調べる場合がある。

特許の価値については既に多くの研究があり，主として引用情報，特に後続の特許からの被引用件数情報，さらに訴訟情報や審査経過情報，外国出願ファミリ数といった質的情報を加味した価値指標は既にいくつかの特許検索ツールに搭載されている。ここではその他の方法の一例としてネットワーク分析を用いた分析例を挙げる。特許は引用関係を通じて複雑なネットワークを形成する。次数中心性や媒介中心性といったネットワーク指標を用いて中心性の高い特許を特定することで，当該分野において被引用件数から見た価値の高い特許とは異なる中心的な特許を特定できる可能性がある。またネッ

トワーク分析を発明者に対して適用することで，当該分野における重要な発明者とその発明者の特許を特定するという手法も考えられる。

(4) 特許対応製品情報探索

自社保有特許を侵害する他社製品情報の探索は事業担当者や技術者が行う場合もあるが，知財担当者からの情報提供も有用と思われる。

一般的に他社製品の情報はテキスト情報が少なく，画像やカタログ等の，テキストでの検索が困難な情報が多い。そこで分析に用いるデータ収集（データエンジニアリング）の段階においてスクレイピング（クローリング）を用いて特定の条件を満たすデータを機械的に収集する手法が考えられる。ただしスクレイピング利用を禁止する規約を有するwebサイトもあること，また収集したデータそのものを譲渡等することは著作権法に触れること，相手先サーバに過負荷をかけた場合には偽計業務妨害罪となりうること，収集対象が個人情報を含む場合に個人情報保護法に触れる可能性があること，などには注意を要する。

4．分析事例

これまでDSの知財課題への適用について概要を述べてきた。４章では実際に課題に取り組んだ事例を紹介する。

課題は委員の関心の高かった「予測」と「探索」に関するものとした。

課題を解決するための分析手段はツールベンダーが有償で提供するツール（以降「商用ツール」と呼ぶ），フリーソフトウェアや表計算ソフトウェア以外にオープンソースでのプログラミングによる自作分析ツール（以降「自作ツール」と呼ぶ）を使用した。自作ツールに取り組んだ理由は，文献[16]にもあるように最近ではオープンソースによる，個々人の環境や分析目的に合わせたツールの自作が容易に可能になり，手軽に情報分析の幅の拡大や質の向上の実現が期待できるようになってきたことから，この研究を機会にその実力を確認したかったからである。

4．1 特許出願件数予測

(1) 回帰分析による特許出願件数予測

開発動向将来予測のためにまず確認するのは特許出願件数との考えのもと，その件数予測に取り組んだ。具体的には，出願から公開公報発行までの１年半分の未公開出願件数を把握することを目的に，2010～2017年までのデータを教師データとして，回帰分析による2018年の出願件数予測を試行した[17]。なお，後述するように，出願件数のカウントにはDerwent World Patents Index™（以下DWPI）におけるDWPIファミリを採用している。このため，対象国はDWPIに収録されている59か国である。

なお，本事例では，実際のビジネスシーンにおいてDS手法を活用していくための方法論であるCRISP-DM（CRoss-Industry Standard Process for Data Mining）[18]に則っており，図２[19]に示すそのプロセスに沿って具体的な内容を説明する。

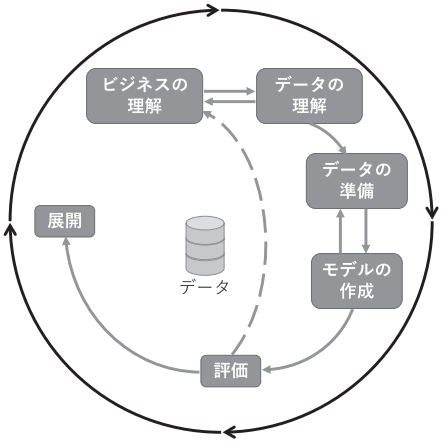

図２　CRISP-DM

なお，分析の目的を明確化するステップである「ビジネスの理解」については，割愛する。

1）データの理解

予測対象となる特許の出願件数のカウント方法には，実数を反映した出願番号ベース以外に，優先権主張で紐づけられたINPADOCファミリ，シンプルファミリ，「発明内容が同じ」と人が判断した発明単位のDWPIファミリ等が存在する。本事例では，開発動向の予測に繋げるために，R&D活動の成果をより反映している可能性の高い発明数が最適との考えから，DWPIファミリによる出願件数カウントを採用した。

また，特許出願は，企業のR&D活動の成果である一方で，競合優位性を築いていくための投資の側面を有している。このため，その年の出願件数は，「①過去の出願件数の推移」といった特許データのみならず，「②開発投資額」，「③会社の規模」，「④会社業績」，「⑤経営戦略」，「⑥ビジネス環境」といった非特許データにも影響される可能性がある。これらの非特許データについては，データ取得の容易性から，「②開発投資額」，「③会社の規模」，「④会社業績」を対象に，欧州委員会が毎年公開しているEU Industrial R&D Investment Scoreboardのデータを活用することにした。

また，特許出願件数の予測には，１社の数年分のデータでは足りないため，複数の会社のデータを用いて，回帰分析に対応できるようにした。一方で，予測したい出願件数は，自社と同一業界の大手競合や新規参入したい業界のメジャー企業であることがほとんどである。このためEU Industrial R&D Investment Scoreboardにある“Industry”を基に，業界ごとに上位の複数企業のデータを用いることとした。

2）データの準備

EU Industrial R&D Investment ScoreboardおよびDWPIのグローバルなデータを基に，目的変数である「当年出願件数」を予測するための説明変数として，表３に示す21種類のデータを準備した。このうち，非特許データについて，過去直近の数値とその変化が出願件数に影響する（例：R&D費について，複数年にわたるR&D活動の成果が出願件数に反映されるまでに最低でも１年のタイムラグがある）との考えのもと，前年，２年前，３年前の数値，および，２年前から前年にかけての成長率（1y-G），３年前から前年にかけての成長率（2y-G）を用いている。また，特許データである出願件数についても同様に，前年，２年前の数値を用いている。

業界としては，特許出願の件数やその傾向，委員の所属企業との関連性を考慮して，以下４つを選定した。

①Technology Hardware & Equipment
（APPLE, Intel, A社など）
②Electronics
（Samsung Electronics, Siemens, B社）
③Automobiles
（Volkswagen, C社, D社など）
④Chemicals
（DuPont de Nemours, E社など）

なお，大型買収があった場合は，買収企業の情報を出願件数含めてデータをマージした（例：IntelによるAlteraの買収）。また，新興国企業で非特許データが揃わない場合，20位以下の企業と置き換えた（例：Tata Motors→Tesla）。

また，R&D費は３～５桁の範囲であるのに対して，従業員数は４～６桁の範囲にあるように，これらの21種類の説明変数の中には，互いの数値範囲が大きく異なるものが含まれている。一般的に，機械学習では数値範囲の異なる複数の説明変数（特徴量）をそのまま学習に用いると，予測精度が悪くなる。このため，データの前処理として，各説明変数の平均を０，分散を１にする標準化[20]を行った（表３）。

3）モデルの作成

回帰分析の手法として，EXCELのTREND関数（線形回帰），Python[21]を用いたサポートベ

表３　21種類の説明変数と実データの一例

No.	データ区分	データ名称	データの一例：Apple社の2017年	
			実データ	標準化後データ
1	①過去の出願件数の推移	前年出願件数	2,257（件）	0.02371
2		２年前出願件数	2,272（件）	0.07233
3	②開発投資額	前年R&D費	9,529（€mn）	2.596
4		２年前R&D費	7,410（€mn）	2.113
5		３年前R&D費	4,976（€mn）	1.265
6		R&D費1y-G	1.286	0.9489
7		R&D費2y-G	1.384	1.773
8		R&D/売上	0.04658	－1.536
9	③会社の規模	前年従業員数	116,000（人）	0.7747
10		２年前従業員数	110,000（人）	0.6434
11		３年前従業員数	92,600（人）	0.3590
12		従業員数1y-G	1.055	0.3947
13		従業員数2y-G	1.119	1.069
14	④会社業績	前年売上	204,600（€mn）	4.875
15		２年前売上	214,700（€mn）	5.910
16		３年前売上	150,600（€mn）	4.927
17		売上1y-G	0.9529	－0.5659
18		売上2y-G	1.166	0.3943
19		前年営業利益	56,940（€mn）	4.783
20		２年前営業利益	65,430（€mn）	6.278
21		３年前営業利益	4,324（€mn）	5.198

クタマシン（以下SVM）[22]，商用の予測分析ツール（以下「商用ツールＡ」）を用いた。また，非特許データを用いた予測精度の向上を確認するために，21種類の説明変数を用いた以下の５つのモデルを作成した。

①１変数：前年出願件数のみ
②２変数：前年出願件数，２年前出願件数
③３変数：②＋当年出願件数との相関係数（※）が最も高い説明変数
④４変数：③＋当年出願件数との相関係数（※）が２番目に高い説明変数
⑤21変数：全21種類の説明変数

※商用ツールＡでは，相関係数の代わりに，各説明変数がどの程度予測に寄与するか（影響を与えるか）を示す指標である「寄与度」を使用。

具体的には，2013～2017年のデータを教師データとしてモデルを作成した。例として４変数のモデル作成時の教師データを表４に示す。

なお，紙幅の都合上，具体的なモデル作成の紹介は，前述の業界のうち，Technology Hardware & Equipmentのみとした。また，それ以外の３つの業界についてはそれぞれ，21種類の説明変数の当年出願件数との相関係数を求め，後述する改善策において活用した。

４）評　価

作成したモデルに対して，2018年のデータを用いてその予測精度評価を行った。具体的な評価方法について，EXCELのTREND関数（線形回帰）を用いて４変数で作成したモデルを例に，表５を用いて説明する。

表４　４変数でのモデル作成時の教師データ（※例示した３社以外は省略）

企業	年	説明変数				目的変数
		前年 出願件数	２年前 出願件数	前年 従業員数	２年前 従業員数	当年 出願件数
Apple	2013	0.05396	－0.3208	－0.004671	－0.1778	1,971
Intel	2013	2.645	2.755	2.235	2.196	8,629
A社	2013	0.2392	0.2413	0.6327	0.5200	3,301
Apple	2017	0.02371	0.07233	0.7747	0.6434	2,114
Intel	2017	1.123	1.127	0.5943	0.5959	7,993
A社	2017	2.972	3.475	2.248	2.043	4,293

表５　TREND関数を用いた４変数でのモデルの予測精度評価
（※例示した３社以外は省略）

企業	説明変数				目的変数		精度
	前年 出願件数	２年前 出願件数	前年 従業員数	２年前 従業員数	当年 出願件数 予測値	当年 出願件数 実値	絶対 誤差率
Apple	－0.04283	0.06496	0.9009	0.7489	2,221	2,246	1.10%
Intel	0.9711	1.226	0.5347	0.5730	4,199	3,620	16.00%
A社	2.693	3.178	2.250	2.185	7,873	8,400	6.27%
					平均絶対誤差率		15.40%

①作成したモデルに2018年のデータを読み込ませ，企業別に2018年の出願件数の予測値を得た。
②得られた予測値について，企業別に2018年の出願件数の実値に対する絶対誤差率を求めた。
絶対誤差率＝|実値－予測値|÷実値
③企業別の絶対誤差率から平均絶対誤差率を求め，モデルの精度を評価した。

※目的変数である「当年出願件数」の数値が選定した20社（特に件数上位と下位）で大きく異なることから，評価指標として平均絶対誤差率を用いた[23)]。

EXCELのTREND関数（線形回帰），および，Pythonを用いたSVMで作成したモデルの予測精度を表６に示す。いずれの手法でも，４変数モデルにおいて，特許データのみを用いた１変数モデル，２変数モデルよりも，精度向上が見られた。また，商用ツールＡを用いたモデルでも，４変数モデルにおいて，特許データのみを用いた１変数モデル，２変数モデルよりも，精度向上が見られた。その一方で，最も予測精度の良いEXCELのTREND関数を用いた４変数モデルでも，予測値の誤差率が75.7％と極端に悪い企業が含まれていることもあり，平均絶対誤差率は15.4％止まりであった。

５）考　察

①特許出願件数予測における非特許データの活

表６　モデルの予測精度（平均絶対誤差率）

モデル名	モデル				
	１変数	２変数	３変数	４変数	21変数
EXCELのTREND関数	16.1％	15.5％	15.9％	15.4％	18.7％
PYTHONを用いたSVM	16.8％	16.9％	17.2％	16.5％	20.8％

用

説明変数として非特許データを取り入れることによる予測精度の向上が，いずれの手法でも認められることから，特許出願の件数予測において，非特許データ活用の有用性が示された。また，いずれの手法においても最も予測精度が高かった４変数モデルの説明変数に「前年従業員数」が含まれることから，Technology Hardware & Equipmentでは，「③会社の規模」が出願件数に影響することが示唆された。さらに，いずれの手法においても，21変数モデルよりも４変数モデルの方が，予測精度が高いことから，相関係数などを指標に，出願件数に影響する関連因子のみを説明変数として加えることが重要であることが示された。

②回帰分析手法による違い

本事例では，EXCELのTREND関数（線形回帰）で作成したモデルの予測精度が最も高かった。これは，過去の出願件数の当年出願件数に対する相関係数が0.9以上と非常に高く，単純な線形回帰の方がその影響を予測に反映させ易いことが要因と考えられる。

③モデルの実用性

出願件数の予測として実際に活用するためには，感覚的ではあるが，平均絶対誤差率は10％以下であることが必要と考える。前述の業界ごとのモデルでは，最も良くて15.4％であることから，教師データや回帰分析手法などを工夫して精度向上を図る必要がある。このうち，教師データに着目した改善策について，以下で紹介する。

６）改善策について

Technology Hardware & Equipment以外のElectronics, Automobiles, Chemicalsについても，「当年出願件数」と21変数との相関分析を行って対比した。その結果，事業の特性上，R&D費が相関係数として特許出願件数に与える影響の大きさが異なり，それによって，企業を以下の３つに区分できることが分かった。

①「部品メーカー」：半導体，自動車部品など，最終製品の部品を製造するメーカー

- ✓ Intel，Ｄ社など。
- ✓ R&D費の相関係数が全て0.7以上と高い。
- ✓ 求められる性能がある程度決まっているため，特許で競合優位性を確保する意識が高く，そのためR&D費が特許出願件数に影響する可能性が高いと考えられる。

②「装置・サービス・材料メーカー」：オフィス用プリンタ，ネットワーク機器，化学品などを製造するメーカー

- ✓ Ａ社, Cisco Systems, DuPont de Nemoursなど。
- ✓ R&D費の相関係数が「部品」よりも低い一方で，従業員数の相関係数が全て0.6と比較的高い。
- ✓ 技術サポートなどを通じて得た顧客ニーズから特許出願する傾向が強いため，R&D費が特許出願件数に影響する可能性が部品メーカーよりも低いと考えられる。

③「完成品メーカー」：コンシューマー向け製品を製造するメーカー

- ✓ Apple，HUAWEI，Ｃ社など。
- ✓ 売上，営業利益の相関係数が全て0.6～0.7と比較的高い一方で，R&D費の相関係数が全て0.4以下と低い。
- ✓ 売上・利益が伸びることで，次の製品開

発と特許出願が加速する傾向が強いと考えられるため，R&D費が特許出願件数に影響する可能性が低い一方，売上が特許出願件数に影響する可能性が高いと考えられる。

そこで，この事業特性による企業区分（以下，事業特性区分）を基に，教師データを再度準備し，予測精度が向上するか確認した。

結果として，表７に示すように，EXCELのTREND関数（線形回帰），および，Pythonを用いたSVMでは予測精度の向上が確認できた。このことから，線形回帰やその派生であるSVMなどの回帰分析手法では，目的変数である「当年出願件数」との相関係数を参考にした事業特性区分が有効であると考えられる。ただ，予測精度は，平均絶対誤差率16.3％止まりで，目標とする10％以下までには到達しなかった。

なお，商用ツールＡでは，このような事業特性区分では，予測精度の向上は見られなかった。これは，商用ツールＡで作成されるモデルは，ランダムフォレストなど線形回帰以外の回帰分析の手法を取り入れていることに関連しているものと思われる。

７）まとめ

以上のように，誤差が大きく，特許出願件数予測自体について，今回は実務で使えるまで予測精度を高めることはできなかった。しかし，改善策で示したようにCRISP-DMのサイクルを回すことで精度向上が見られたことから，このサイクルを繰り返すことで実務的に使えるレベルにまで精度を高められる可能性が示唆された。

また，機械学習を使った特許出願の件数予測において，非特許データの活用が有用であることが示せた。特許は，企業の事業活動と密接に関連していることから，特許出願以外の例えば特許スコアの予測においても，このような非特許データの活用が有用であると考えられる。

表７　教師データの企業区分による予測精度比較

モデル名	モデル	
	業界区分	事業特性区分
TREND関数	18.7%	16.0%
SVM	20.8%	20.3%

(2) 時系列分析による特許出願件数予測

次に，周期性に着目した時系列分析による特許出願件数予測を試行した。

１）データの準備

予測分析として，日本国特許庁ホームページから，日本国の「特許出願等統計速報」の「月次出願数」のデータ[24]に基づく時系列分析を検証した。「月次出願数」は日本における月次の出願数を単位とするものであり，10年以上のデータが蓄積されている。初めに，これを俯瞰することで長期的な傾向（Trend）と周期性（Seasonally）の有無を確認した。次に，そういった特徴を考慮したモデルを作成し，予測を行った。

図３は2011年１月～2020年12月の「月次出願数」の推移を示したものである。

まず，年度末の３月に大ピーク，四半期末の６月，９月，12月に小ピークが確認できる。つまり，３ヵ月，12ヵ月の周期性を示すと言える。さらに，大ピークのピークトップと，ベースラインの推移に注目する。視覚上，緩やかな減少傾向を示しているように見受けられるので，長

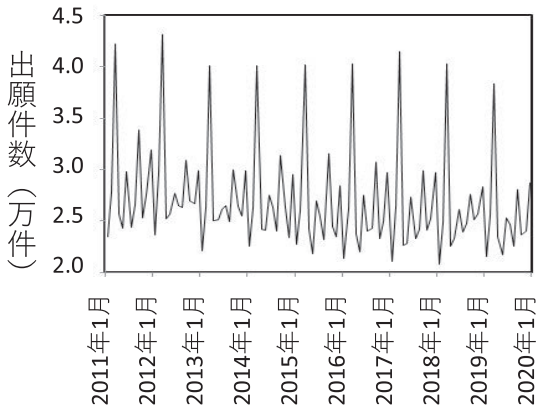

図３「月次出願数」の推移

期的な傾向があると言える。

長期的傾向，周期性が確認されたことから，次に，Pythonのライブラリ“Statsmodels”を用いてモデルを作成し，予測分析を行った。

２）モデルの作成

Statsmodelsのコードはインターネット上のブログを参考にして作成した[25]。

まず，作成したコードを基に，2011年１月～2019年12月の出願数データを教師データとしてモデルを作成した。図４は，前記図３のデータを，Statsmodelsを用いた処理により，“Trend”，“Seasonally”，“Residual”に分解した結果である。ここで，“Trend”は長期的な増減傾向成分，“Seasonally”は周期性成分，“Residual”は残差成分（元のデータから“Trend”と“Seasonally”のデータを引いた後の数値を示す）を元データから抽出したものである。図３を俯瞰した際に，データが周期性を示すこと，長期的に緩やかな減少傾向を示す可能性を確認したことと図４の，“Trend”と“Seasonally”は一致している。

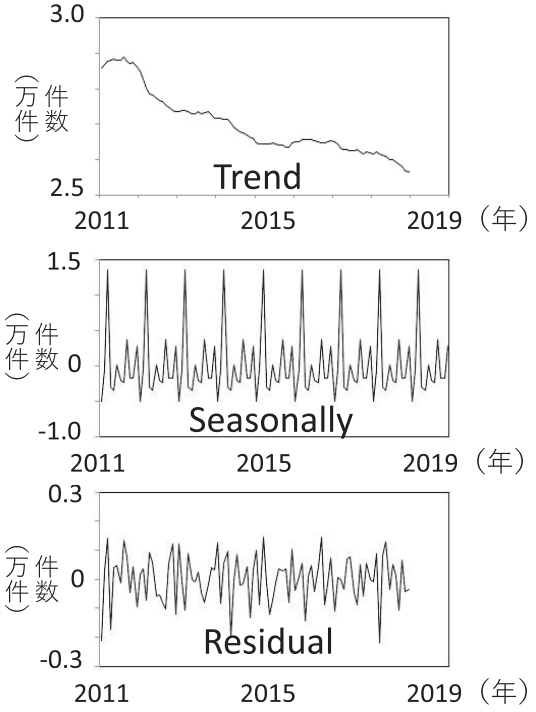

図４　“Trend”，“Seasonally”，“Residual”の分解

“Residual”は，モデルの妥当性の検証，モデル内のパラメータ調整に利用できるが本稿ではその点は行っていない。

また図５は，前記図３のデータの自己相関関係を示したものである。横軸はタイムラグを，縦軸は相関強度を示す。

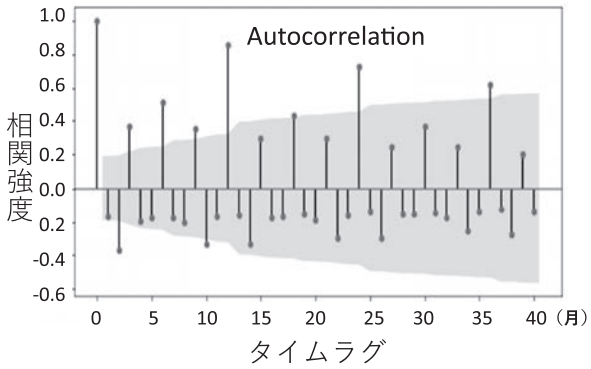

図５　自己相関関係（Autocorrelation）

タイムラグが０というのは，すなわち，自己を指すものであるため，その相関強度は１である。以降，タイムラグ12，24，36，すなわち，12ヵ月の倍数で前後するデータとは強い相関強度を示し，タイムラグ３，６，９，すなわち，３ヵ月の倍数で前後するデータとは弱い相関強度を示している。また，薄墨色の領域は95％信頼区間を示し，この領域外である場合，統計的に優位な強度と判断できる。したがって，12ヵ月で強い周期性を，３ヵ月で弱い周期性を示すデータであると判断できる。

３）予測結果

最後に，作成したモデルを用いて予測分析を行った。2011年１月～2019年12月のデータを教師データとし，2020年１月～2021年12月の出願数を予測した。結果を図６に示す。

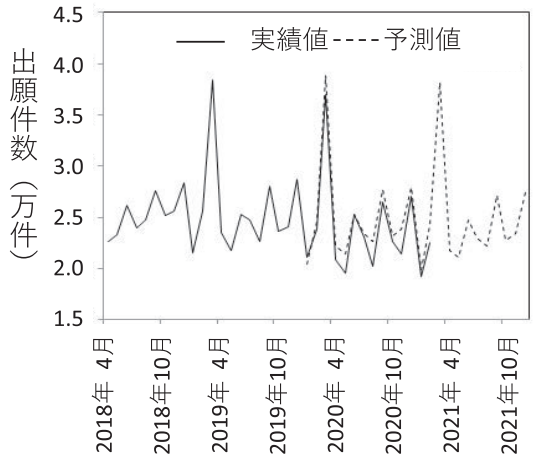

図６　「月次出願数」の実数と予測結果

４）まとめ

2020年１月～2020年12月のデータは，実数と予測値の比較が可能な期間である。図６の視覚的な印象では，まずまずの一致と言える。また，この期間におけるズレ（実数－予測値）は100件以内に収まるものであった。

このことから，本予測結果を何らかの判断材料として用いる場合，参考にできるものと考えられる。

4. 2　クラスタ分け

用途探索を目的とした「教師なし学習」でのクラスタ分けを実施した。

１）データ準備

用途探索をする題材として，「炭素繊維」をとりあげた。比較的新しい分野であり，実践的と考えたからである。

具体的には，以下の条件で母集団を作成した。

・検索ツール：DerwentInnovation
・検索条件：IPC＝C04B35/83，検索されたコレクション：DWPI and DPCI，検索日：2021年５月２日
・検索結果：3,510件。これを母集団とした。

２）分析手法

母集団を用途ごとの小集団（クラスタ）に分ける方法は色々ある。本研究では，以下の３つの手法を合わせて実施した。これは，新規用途探索は分析者にとって未知の新たな用途を見つけるという宝探しのようなものであり，複数手段で分析することがひとつの有効な方法だからである。

手法①　テキストマイニングによる，特許出願単位のクラスタ図（鳥瞰図）（商用ツールを使用）

手法②　テキストマイニングによる，キーワードの共起ネットワーク（KH Coder3を使用）[25), 26)]

手法③　トピック分析(Latent Dirichlet Allocation：LDA)[27)～29)]による，キーワードのワードクラウド（自作ツールを使用）

手法①はすべての特許出願をクラスタに分類するため，各クラスタの特許の内容や，主要語句を確認することにより，すべての用途を漏れなく抽出できるという特徴を持つ。手法②と手法③は，出現頻度が高く，関連性がある語句（群）を抽出できるので，それらの語句を確認することにより，主要な用途を容易に把握することができるという特徴がある。なお，手法③は，一つの語句が複数のトピックに重複して出現することを許しているため，他の手法とは異なった結果が期待できる。

３）分析結果

図７に手法①の結果を示す。このようにクラスタを読み解くことにより用途が把握できる。また図８に示す手法②の結果においては「航空機，自動車」，「電池」，「水・汚染・フィルタ」などの多くの用途が抽出できている。

図９には手法③の結果を示す。用途を表す語句として，手法②と同様に「自動車（航空機）」，「セラミック・ブレーキ」，「耐熱・繊維（航空・宇宙）」，「電池」が抽出されている（図９のＡ部）。さらに，同図を眺めると「摩擦」（図９のＣ部）という語句が目に付く。これは「ブレーキ」，「クラッチ」，の上位概念であり，「摩擦」を利用す

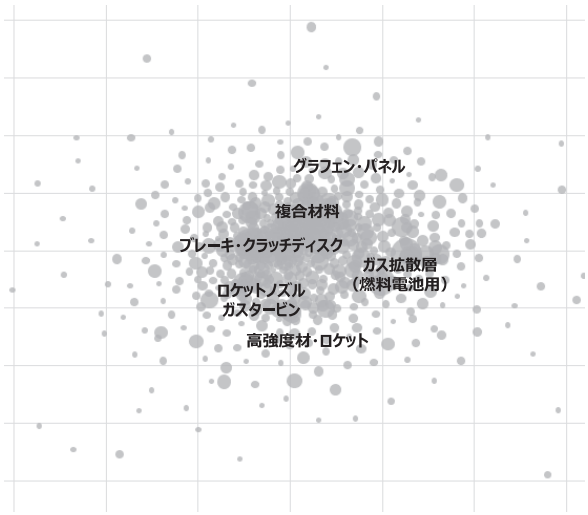

図７　手法①（クラスタ図）の結果

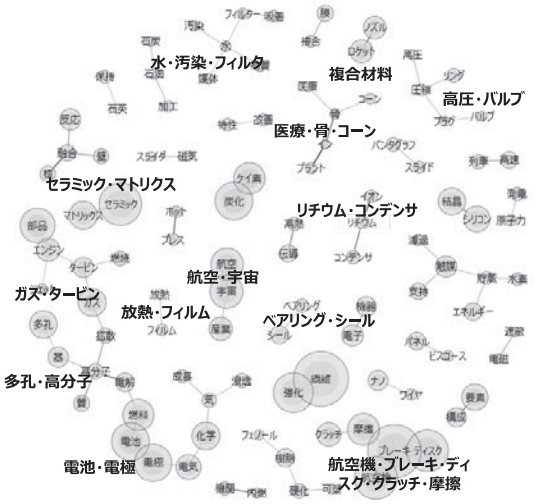

図８　手法②（共起ネットワーク）の結果

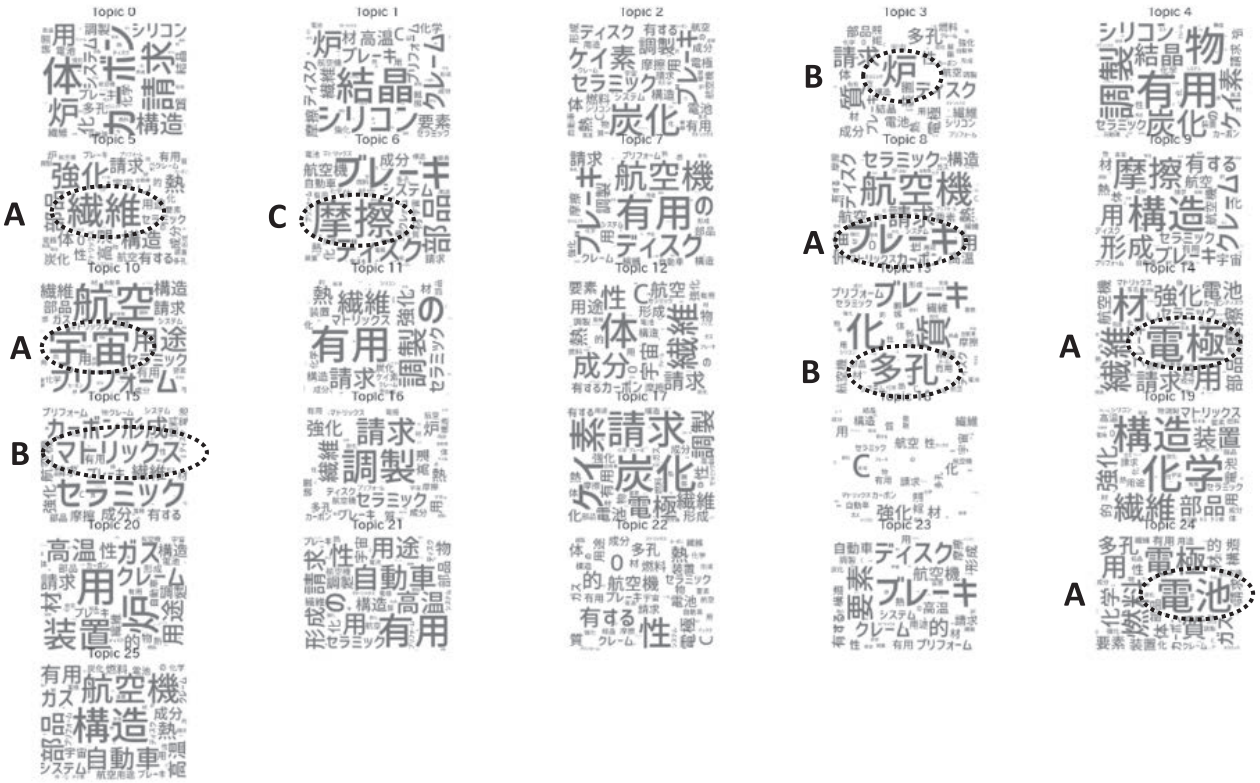

図９　手法③（ワードクラウド）の結果

る他の用途へ発想を広げるきっかけになる。また，「炉」，「多孔」，「マトリックス」の語句も目につく（図９のＢ部）。これらの語句と炭素繊維の関連をWebで調べてみると，「炉」は炭素繊維を製造するために重要な装置であり，「多孔」は新規炭素繊維に関するキーワードであり，「マトリックス」はCFRP（炭素繊維強化プラスチック複合材料）製造のキーワードのひとつでありそれぞれ重要語であることがわかる。

４）まとめ

以上のように，土地勘がない分野で用途探索する場合，アルゴリズムや表現手段が異なる複

数の手法で分析した結果を活用するのは効果的である。そして分析のバリエーションを増やすには，商用ツールやフリーソフトウェアの活用だけでなく，オープンソースを使った自作の分析ツールの活用も有用であることが確認できた。

5．おわりに

本研究ではDSの全体像を把握したうえで，有識者の知見を得ながら，知財分野での活用事例の研究を行った。

特に，有識者から得た具体的なアドバイスを踏まえ，回帰分析，時系列分析，テキストマイニングなどを体験したことで，手法やプロセス（データ収集・モデル作成・分析・結果活用）の考え方や留意点等について理解が深まったことは有意義であったと考える。

なお，本研究では，初期段階において予測分析ツール「Prediction One」[30]を使って予測分析を実践的に体験したことで，事例研究の立ち上げをスムースに行うことができた。プログラミングスキルが不要ですぐ導入でき，応用範囲も広いため有用な分析ツールだといえる。

最後に，今回いくつかの分析に自作ツールを利用したが，調査分析担当者が日々進化する分析手法や事例に網を張って最新の情報を収集し，自力でその手法を習得することができれば，その手法を搭載したツールのリリースを待たず，他社に先んじた解析ができる可能性がある。もちろんそのためには相応の自己研鑽が必要でありハードルは低くはないが，自作ツールの将来について期待が持てる研究結果となった。本稿が知財情報活用の参考や，スキルアップのきっかけになれば幸いである。

注　記

1） 濱田悦生，データサイエンスの基礎，pp.1～3，（2019），講談社
2） 前掲注１）p.2　図１－１を参考に作成

3） 特許庁，AI関連発明の出願状況調査 報告書，pp.4～5
https://www.jpo.go.jp/system/patent/gaiyo/sesaku/ai/document/ai_shutsugan_chosa/hokoku.pdf
4） Foster Provost, Tom Fawcett, 戦略的データサイエンス入門，（2014），オライリー・ジャパン
5） Joel Grus，ゼロからはじめるデータサイエンス第２版－Pythonで学ぶ基本と実践（2020），オーム社
6） 石井大輔，漆畑充，及川大智，大下健史，オング優也，現場のプロが伝える前処理技術　基礎から実践まで学ぶテーブルデータ／自然言語／画像データの前処理（2020），マイナビ出版
7） ウォルター・エンダース，実証のための計量時系列分析，（2019），有斐閣
8） 橋本洋志，牧野浩二，データサイエンス教本Pythonで学ぶ統計分析・パターン認識・深層学習・信号処理・時系列データ分析（2018），オーム社
9） 山内長承，Pythonによるテキストマイニング入門（平成29年），オーム社
10） 2018年度情報検索委員会第３小委員会，知財管理 Vol.69 No.10「テキストマイニング技術の活用に関する研究」，pp.1426～1440（2019）
11） 野守耕爾「特許文書データに人工知能技術を応用した競合分析と技術の新規用途探索」，（2017）
https://www.msi.co.jp/userconf/2017/pdf/muc17_501_3.pdf
12） アジア特許情報研究会，tokugikon no.298「AI系基盤技術と，オープンソースを用いた機械学習による特許文書解析」pp.25～37（2020）
http://www.tokugikon.jp/gikonshi/298/298tokusyu3.pdf
13） 海北大輔，中核的な特許出願の特定方法に関する調査研究［課題研究報告書］，pp.35～42（2011）
https://dspace.jaist.ac.jp/dspace/bitstream/10119/9651/5/paper.pdf
14） 渡部俊也，小林徹，藤原綾乃，Japio YEAR BOOK 2013，pp.218～221ネットワーク理論の知財情報への応用
https://www.japio.or.jp/00yearbook/files/2013book/13_3_01.pdf
15） 加藤耕太，Pythonクローリング&スクレイピング　データ収集・解析のための実践開発ガイド

（2017），昭和情報プロセス株式会社
16） 前掲注12）
17） 予測分析の際の時系列分析か回帰分析かの選択について，トレンドや周期を見るのに適したのが時系列分析であるので，周期性やトレンドが無いなら回帰分析が向いている可能性は高い。実務的には両方トライして精度の良い方を用いると良い。特許出願件数の予測については，（時系列的な要素を加えるために，）１年前の出願件数，２年前の出願件数，といった情報も含めて数値分析をすると良い（AIツール開発者のコメント）。
18） 前掲注４）pp.29～31
19） 前掲注４）p.31　図２－２，
CRISP-DMヘルプの概要
https://www.ibm.com/docs/ja/spss-modeler/18.1.0?topic=dm-crisp-help-overview
を参考に作成
20） エンジニア向け機械学習スクール
codexa，正規化・標準化を徹底解説
（Python前処理サンプルコード付き）
https://www.codexa.net/normalization-python/
21） Python Japan,プログラミング言語Pythonの紹介
https://www.python.jp/pages/about.html
22） 栗田多喜夫,サポートベクターマシン入門
https://home.hiroshima-u.ac.jp/tkurita/lecture/svm.pdf
23） データサイエンス情報局，適切な誤差指標の選び方
https://analysis-navi.com/?p=2875
24） 特許庁，特許出願等統計速報
https://www.jpo.go.jp/resources/statistics/syutugan_toukei_sokuho/index.html
25） ものきとデータ解析をはじめよう
https://momonoki2017.blogspot.com/search/label/Python%E6%99%82%E7%B3%BB%E5%88%97%E5%88%86%E6%9E%90
匿名のブログではあるが，参考にしたコード自体は当小委員会で検証し信頼に足るものであることを確認した。
26） 樋口耕一，社会調査のための計量テキスト分析－内容分析の継承と発展を目指して－（第２版）（2020），ナカニシヤ出版
27） 末吉美喜，テキストマイニング入門ExcelとKH Coderでわかるデータ分析（2019），オーム社
28） 奥村学，佐藤一誠，トピックモデルによる統計的潜在意味解析pp.25～36（2015），コロナ社
29） WordCloudとpyLDAvisによるLDAの可視化について
https://ie110704.net/2018/12/29/wordcloud%E3%81%A8pyldavis%E3%81%AB%E3%82%88%E3%82%8Blda%E3%81%AE%E5%8F%AF%E8%A6%96%E5%8C%96%E3%81%AB%E3%81%A4%E3%81%84%E3%81%A6/
匿名のブログではあるが，参考にしたコード自体は当小委員会で検証し信頼に足るものであることを確認した。
30） ワンクリックで高度な分析が可能な予測分析ツール。詳細はPrediction One紹介ページを参照。
https://predictionone.sony.biz/
（URL参照日は全て2021年９月22日）

（原稿受領日　2021年11月１日）