論　説

知財情報分析におけるAI等の活用に関する研究

情報活用委員会
第２小委員会*

抄　録　近年，会員企業各社においてIPランドスケープへの取り組みが本格化している。IPランドスケープでは特許情報と非特許情報を活用し，経営に貢献するための分析を行うことから，特許情報のみの分析に比べ，扱うデータの種類や量および検討項目が多くなり，またアウトプットに対する要求レベルも高くなる傾向にあるため，これを担当する各社の知財部員の負担が増加していると推測される。その一方で各種調査用データベースや分析ツールにおいては，AI搭載などの機能拡充がなされてきているため，各社の知財部員にはこれらの活用による調査・分析業務の効率化も期待されている。そこで本研究では，主にIPランドスケープによる技術動向調査および新用途探索を想定し，AI等を活用した効率化と精度向上の手法，特に非特許情報について，多数のデータの集合を特許の母集団と組み合わせて解析する上での工夫を検討したためこれを報告する。

目　次

1．はじめに

近年，IPランドスケープの概念の定着および方法論の浸透に伴い，会員企業各社においてもその取り組みが本格化しているであろうことは想像に難くない。しかしIPランドスケープは，経営に貢献する提言に向け，特許情報や非特許情報を活用した各種の分析を行うもの[1)]であることから，特許情報のみの分析に比べて，扱うデータや検討すべき項目が必然的に多くなり，またアウトプットに対する要求レベルも高度化する傾向にある。したがって実際の調査・分析を担当する会員企業各社の知財部員にとっては工数的な負担が増加しているものと推測される。

一方で近年は，AIを搭載した特許調査データベースや分析ツール，企業情報やニュースなどの非特許情報の調査に主眼を置いた各種データベースもリリースされてきており，各社の知財部員に対しては，これらを利用することによる調査・分析業務の効率化も期待されている。

すなわち各社の知財部員に対し，「特許情報と非特許情報に基づき，AI等を活用することで，効率的に調査・分析を行い，経営に貢献し

* 2020年度　The Second Subcommittee, IP Intelligence Committee

うる確度の高いアウトプットに繋げること」が求められている，または少なくともそう期待されているものと考えられる。そしてこのような要求への対応としては，AI等を搭載した特許調査・分析ツールの利用，例えば，機械学習を用いた類似特許検索やノイズ除去等による母集団作成，および分類付与やクラスタリング等による動向調査の効率化といったことが挙げられる。実際，これらの機能の活用に関する検討事例[2]も既に報告されている。

ただし，これらは基本的に特許情報の分析を想定したもので，非特許情報と組み合わせての分析にまで深く踏み込んだものではない。

また，「IPランドスケープの事例」としての報告[3]においても，非特許情報は「特許情報の分析結果として得られた知見や仮説の裏付け」となるニュースやレポートを，ピンポイントで抽出して組み合わせているに留まるものが多いと見受けられる。もちろん，その手法や結果を否定するつもりは一切無いが，ピンポイントで抽出した情報を仮説の根拠として採用する際にその妥当性を判断するには，それを行う担当者自身に相応の知識や経験が求められる。よって例えば新事業探索などの目的で，自社の知見や経験が少ない分野を調査する場合においては，「確証バイアス[4]」による恣意的な結論に陥るリスクが懸念される。

そこで本研究では，例えば自社の知識や経験の少ない分野のIPランドスケープを行う場合でも客観性を担保できるように，非特許情報をピンポイントでなくバルクの母集団（データの集合）として特許の母集団と同様に扱うことを前提に，調査・分析の効率と精度を向上すべく，AI等を活用する手法やその課題を検討した。

なお本研究は，2020年度情報活用委員会第２小委員会の青山裕樹(小委員長, ポリプラスチックス)，佐々木俊輔（小委員長補佐，東日本旅客鉄道)，飯村信（第一三共)，久我範夫（マレリ), 座古泰裕（シチズン時計), 日下部真吾（日東電工)，白井一光（日本化薬)，日高輝（フジクラ), 藤原伸城（京セラドキュメントソリューションズ)，安福友浩（セイコーエプソン）によるものである。

2．検討概要

2. 1　検討対象業務の選定

前述の通り本研究では対象業務としてIPランドスケープを想定し検討を行うが，一口にIPランドスケープと言ってもその目的は様々であり, それにより用いる情報や手法も変わりうる。

そこで当小委員会メンバーの関心の高い業務として，「技術動向調査」および「新用途探索」について検討することとした。

2. 2　分析手法検討の前提

上記の対象業務の検討において，特許情報と非特許情報を組み合わせた分析手法を検討するにあたり，非特許情報を「ピンポイントで抽出した１文書または少数の文書のみ」にて用いるのではなく，特許の母集団と同様，「多数の文書からなるデータの集合」として扱うことを前提とした。

ただし，特許情報では請求項や明細書に記載された文章以外にも，IPCやFIといった各種の技術分類や, 審査経過, 権利維持の状況など種々の項目が，共通様式で提供されているため，それらを容易に分析に利用することが可能だが，非特許情報では，提供されている項目や様式はソースによって異なる。このため，特許情報と組み合わせた分析に利用することができるのは，発信者や公表時期等，ソースに関わらず共通で得られる項目の他は，実質的に「文章そのもの」のみ[5]ということになる。

そこで本研究では，自然言語処理を利用したテキストマイニングを主軸として，特許情報と

非特許情報を組み合わせた母集団のテキストを分析する際のAI等の活用を検討した。

3．非特許情報の入手

会員企業各社の知財部員にとって，特許情報の入手は日頃から行っていることであり，分析に必要なデータを収集する上で特に障害はないと思われるため，本章では非特許情報の入手についての検討結果を紹介する。

3. 1　ソース探索

技術動向調査に利用する非特許情報としては，例えば論文，企業情報，業界情報，技術開発に向けた投資・助成金に関する情報，規格・規制その他政策に関する情報などが挙げられる。

これらのデータを母集団として分析しやすい形で効率的に入手するには，対象の非特許情報が収録されたデータベースを利用するのが有用だが，そのようなデータベースには，収録されている情報の種類や量，検索や絞込の容易さ，データのダウンロードの可否およびその形式，利用料金などの点で様々なものが存在する。

そこで初めに非特許情報のソースとしてどのようなものがあるのか調査し，その中でどれを利用するのが良いかを検討した。具体的には，まず当小委員会メンバーでの探索および当委員会メンバーへのヒアリングにより非特許情報のソースの候補となるデータベース等を抽出し，当小委員会メンバーでその収録内容等を確認した。調査したソースの一部を表１に例示する。

3. 2　推奨ソース

前述の通り本研究ではテキストマイニングを用いて検討を行うこととしたため，非特許情報のソースとしては，解析に利用しやすいように表形式でテキストデータをダウンロードできること，また会員企業各社が検討しやすいように無料で利用できること，が望ましい。

以下では上記の観点から本研究での検討手法における利便性が高いと考えられた推奨ソースを幾つか紹介する。

(1) 科学研究費助成事業データベース

文部科学省と日本学術振興会が交付する科学研究費助成事業に関するデータベースであり，助成金を受けて行われた各種研究テーマの情報が収録されている。当然ながら助成金の交付を受けるには事前審査を通過する必要があるため，収録されるのはいわゆる「スジが良い」情報と推測される。

表１　本研究で検討した非特許情報ソースの例

カテゴリ	データベース名	概略	URL
論文	arXiv	物理学，数学，計算機科学等の論文	https://arxiv.org/
	Google Scholar	各分野の論文	https://scholar.google.com/
	JDreamⅢ	各分野の論文【有料】	https://jdream3.com/
	PubMed	医学，生物学の論文	https://pubmed.ncbi.nlm.nih.gov/
	Web of Science	各分野の論文【有料】	https://login.webofknowledge.com/
企業・業界	EDINET	各社の有価証券報告書	https://disclosure.edinet-fsa.go.jp/
	JPubb	各社のプレスリリース	http://www.jpubb.com/
	MARKLINES	自動車分野の各種情報【有料】	https://www.marklines.com/
その他	科学研究費助成事業データベース	各分野の助成事業情報	https://kaken.nii.ac.jp/ja/
	経済レポート情報	各国，各分野の経済情報	http://www3.keizaireport.com/
	情報通信統計データベース	総務省による情報通信分野の各種統計	https://www.soumu.go.jp/johotsusintokei/

※「概略」欄に【有料】と記載されているものは検索画面にアクセスする時点で有料の会員登録が必要。

また本データベースでは「詳細検索」機能を用いて，研究ステータス（採択，交付，完了，中断など）や，報告種別（研究概要，研究進捗評価，研究成果報告書，実績報告書など）での絞込による効率的な調査が可能である。

さらにデータ取得についても，検索結果からチェックボックスで対象を選択し，出力形式（XML，CSV）を指定して「実行」のボタンを押すだけで，各情報の作成日時，タイトル，概要の他，キーワード（特徴語）等のリストを容易にダウンロードすることができる。なお，このキーワードのリストも取得できる点は後述する一般的用語の除外において有用と言える。

(2) EDINET

金融庁が運営している電子情報開示システムであり，金融商品取引法に基づいて提出された各事業者の有価証券報告書等が収録されている。業績や財務状況といった投資家や株主にとって関心の高い情報だけでなく，研究開発活動等の技術動向に関連する情報も含まれるものであり，かつ各事業者が自身で提出した情報であることから信頼性が高いソースと考えられる。

データについては提出者の名称や業種，書類種別や提出時期等での絞込の上，XBRL形式でダウンロードすることができる。XBRL形式は企業の財務データを記述する標準書式の一種であるが，EDINETのサイトでは「XBRLからCSVへの変換ツール[6]」も提供されているため，容易にテキストデータを抽出可能である。

なお本研究における効率化検討の一環として，指定した業種および期間の有価証券報告書から「経営上の重要な契約等」と「研究開発活動」のテキストデータを抽出するツールを作成した。これを本稿の付録[7]として提供するため，適宜活用されたい。

(3) JDreamⅢ

本データベースは有料ではあるが，国内外の各分野の論文等が日本語抄録とともに多数収録されており，キーワードでの検索や概要確認が容易で使い勝手が良いため，導入している会員企業も少なくないと思われる。

データとしては，論文の全文テキストは個別に入手する必要があるが，タイトル，発行日，抄録，シソーラス用語・準シソーラス用語等[8]を含むリストをタブ区切りのテキストファイルでダウンロード可能である。

(4) その他

上記３つのソースは，技術分野の範囲が広く，データをテキスト形式で容易に出力することができるため，分析に用いる非特許情報の母集団を作成する上での汎用性や利便性が高い。一方，分野や言語等は限定されるが，これらの他にも例えば下記のようなソースも適宜用いることができる。

１）PubMed

医学・生物学分野の論文等が収録された無料データベースである。公式ページの検索は英語のみだが，関連サービスの「PubMed CLOUD」からであれば機械翻訳による日本語での検索も可能である。

データ取得は，検索結果から取得希望の文献のチェックボックスをオンにして，「CSV形式」と「PubMed形式」の２種類のフォーマットでそれぞれのテキストファイルをダウンロード後，ExcelのINDIRECT関数等を用いて文書IDの「PMID」をキーに「CSV形式」の書誌情報と「PubMed形式」の要約文を紐付けすることで解析に適したテキストデータを得ることができる。

２）Web of Science

JDreamⅢと同様に，各分野の論文等が収録された有料データベースである。ただし，収録されているのは外国文献が中心であり，検索は英語で行う。

データ取得は，検索結果から取得希望の文献を選択し,「レコードの出力」を行うことにより，タイトル，抄録等のリストをExcelファイルやタブ区切りのテキストファイル等としてダウンロードすることが可能である。

会員企業各社の知財部員であれば，上述したソースについては既に知っているか，あるいは既に利用しているものばかりかも知れないが，これらはあくまで例であり，ここで挙げたもの以外にも有用な非特許情報のソースは多数存在するものと思われる。特に業界情報のソースは有料・無料含め各社で把握・利用しているものがあろうかと思われるため，後述する本研究の分析手法の実施においては，それらのソースも適宜含めて検討されたい。

3. 3　データ取得における課題と対策

前節では非特許情報の推奨ソースを示したが，今回探索・調査したソースには，データの取得または利用が困難なものも多く存在した。以下ではそれらの主な課題を挙げる。

(1) 課　題

1）ダウンロード対象がURLリスト

特にウェブブラウザからアクセスするウェブサイト上のデータベースに多く見られる課題として，ダウンロードできるのがテキストデータ自体ではなく，当該テキストデータの保存先にリンクするURLのリストである場合がある。

この場合，テキストデータを取得するには，URLからリンク先を開いてテキストデータを入手するという作業を，各URLに対して行う必要があり，解析の母集団とする多数のデータの取得にはそれに応じた多くの工数が発生してしまうことになる。

2）ダウンロード対象がPDFファイル

上述のURLと同様に，ダウンロードできるのがテキストデータではなく，各報告書のPDFファイルという場合がある。これは，特に市場レポートやプレスリリース，各社技報等に多い仕様だが，これらを解析の母集団に用いるには，ダウンロードされたPDFファイルをテキストに変換してリスト化するという作業が発生し，工数面の負担が大きい。

また，PDFファイルのセキュリティ設定次第では，そもそもテキストが抽出できなかったり，ファイル中の図表，注記，段組，ページ番号等のレイアウトによっては本文のテキストだけを区別して抽出することができなかったりという技術的な課題も存在しうる。

3）テキストデータの区切りの不統一

一部有料データベースでも確認された課題として，検索結果をテキストファイルとして取得することはできても，当該テキストファイル内の各データ間，あるいは各データ内の各項目間の区切り方が不規則な場合がある。

取得したテキストデータが，カンマやタブといった特定の記号等で区切られており，さらにその順序や間隔等が決まっているのであれば，それをもとに各項目のデータを割り当てることも可能だが，データの区切りが不規則な場合，適切なデータ入力が困難となり，そのまま解析しても本来の結果が得られないか，または解析処理自体ができないという課題が生じうる。

(2) 対策検討

本研究では上述の課題への対策も検討した。ただし後述する通り，全ての場合を解決できるという訳ではなく，むしろ解決できない場合の方が多いようにも感じられたため，ケースバイケースで必要に応じベンダー等とも相談することを推奨する。

1）URLリストからのテキストデータ取得

この課題にはいわゆるウェブスクレイピングを利用し，URLのリンク先のデータを自動的に取得することが効率的と考える。

紙幅の都合上詳細は割愛するが，Octoparse[9]などの専用ツール，GoogleスプレッドシートのIMPORTXML関数やExcelのWEBSERVICE関数，Pythonのrequestsライブラリ等，無料で利用できる範囲でも各種の手段が提供されているため，各社のIT環境に応じ検討されたい。

ただしデータベースによっては利用規約等でスクレイピングを禁止している場合もあるため注意が必要である。

２）PDFファイルからのテキストデータ抽出

PDFファイルの内容がテキストのみならば，Pythonの外部モジュールPDFminer等により指定フォルダ内の複数のPDFファイルからのテキスト抽出も可能だが，レイアウトが複雑なPDFファイルの場合，適切な範囲を選択してのテキスト抽出は基本的に困難である[10]。

なお有料ツールでは，図表の混在等ある程度複雑なレイアウトのPDFファイルでも画像とテキストを別々に抽出可能[11]とされているが，どこまで複雑なレイアウトに対応できるのかは未知数であり，上述した通りセキュリティ面で抽出を禁止している場合もあるため注意が必要である。

３）テキスト内のキーワードによる処理

データ間の区切りが不規則な場合であっても，各データに共通で記載されるキーワードが存在する場合，そのキーワードからの文字数や行数をもとに項目の区切り位置を特定できる可能性がある。例えばExcelのCOUNTIF関数にて当該特定キーワードを含むセルを抽出し，FIND関数にてそのセルの中の何文字目に当該キーワードが含まれるかを特定し，そこを起点にLEFT関数，RIGHT関数，MID関数により必要な範囲の文字列を抽出するといった方法が挙げられる。

ただしこの方法を用いることができるのは，当然そのようなデータ間での共通のキーワードが存在する場合に限定される。

これらの点を踏まえ，非特許情報のソースは，テキストデータの入手しやすさや，その利用のしやすさを考慮して選定する必要がある。

4．分析事例

以下，２章１節で触れた「技術動向調査」と「新用途探索」について，検討事例を紹介する。

4．1　技術動向調査

まず技術動向調査について，「自動運転車」をテーマとした検討事例を紹介する。このテーマは当小委員会メンバーの関心が高く，かつ特許情報の母集団作成において日本国特許庁による平成29年度　特許出願技術動向調査報告書の「自動走行システムの運転制御」（以下,「技動」）を参照できるという理由から選定した。

(1) 分析ツール

２章２節で述べた通り，テキストマイニングでの検討を行うにあたり，当小委員会メンバー全員が共通で利用できる無料のツールとして「KH Coder[12]」を用いることとした。

(2) 分析手法

１）母集団作成

下記の①特許情報と②非特許情報を合わせて１つの母集団とした。なお，各情報に含まれる個々のテキストデータには，それぞれの時期に関する情報（特許情報は最先出願日または最先優先権主張日，非特許情報は当該情報の作成日または発行日）を紐付けてリスト化した。

①特許情報

特許情報の集合として，上記の「技動」にて分析対象とされていた特許文献リストのうち，日本にファミリがある4,159件を用いた。なお，ここに含まれる特許文献は，最先出願日または最先優先権主張日が2010年から2015年までのものであった。

②非特許情報

３章２節で挙げた推奨ソースの「科学研究費助成事業データベース」にて,「自動運転」と「車」で全文をAND検索した結果から，研究期間が2016年以降，かつステータスが「交付」または「完了」のものを抽出して，「キーワード」欄のデータを非特許情報の集合に利用した。

また，もう一つの推奨ソースの「EDINET」を用いて書類提出者業種が「輸送用機器」である2016年以降の有価証券報告書を取得し，「経営上の重要な契約等」欄および「研究開発活動」欄を抽出したデータを非特許情報の集合に追加した。

２）IPC分類の学習

下準備として，日本国特許庁サイトにある「IPC分類表[13]」を,図１の通り「記号」欄の「／（スラッシュ）」以下を削除した表に加工し,「IPC分類のメイングループ」ごとの技術概要を表すリストを作成した。

図１　各メイングループの技術概要リスト準備

上記にて作成したリストをKH Coderに取り込んで,「ツール」→「文書」→「ベイズ学習による分類」→「外部変数から学習」を実行し，IPC分類の学習結果ファイル（拡張子「.knb」）として保存した。

３）学習結果による自動分類付与

１）で作成しておいた母集団をKH Coderに取り込み,「ツール」→「文書」→「ベイズ学習による分類」→「学習結果を用いた自動分類」を実行し，２）で作成したIPC分類の学習結果ファイルによる１）の母集団のデータへの自動分類付与を行った[14]。

４）対応分析によるキーワード分布

自動分類付与後の母集団に対し，KH Coderにて「ツール」→「抽出語」→「対応分析」を行った。また，外部変数として１）で用意した時期に関する情報，および３）で自動付与したIPC分類を設定した。これにより各キーワードに対応するIPC分類を，その時期による推移とともに把握することができる。

(3) 分析結果

前項の手順に従い,「自動運転車」に関する2015年以前の特許情報と2016年以降の非特許情報に基づく母集団について，KH Coderにてテキストマイニング（対応分析）を行った結果を図２に示す。

１）対応分析の読み方

図２において，プロットされた点の隣の単語（車両，速度，位置，検知等）は，母集団中に含まれるキーワードであり，図中での単語間の距離が近いほど，母集団のテキスト内における各単語の出現位置も近いことを意味する。

また外部変数の「時期に関する情報」は□でプロットし，2013年以前のものには「～2013」,2014年以降2017年以前には「2014～2017」,2018年以降には「2018～」のラベルを付している[15]。これは例えば図２において「2014～2017」の□の近くにプロットされたキーワード（認識，通信等）は，この時期の特許等に多く記載されていたと見ることができる。

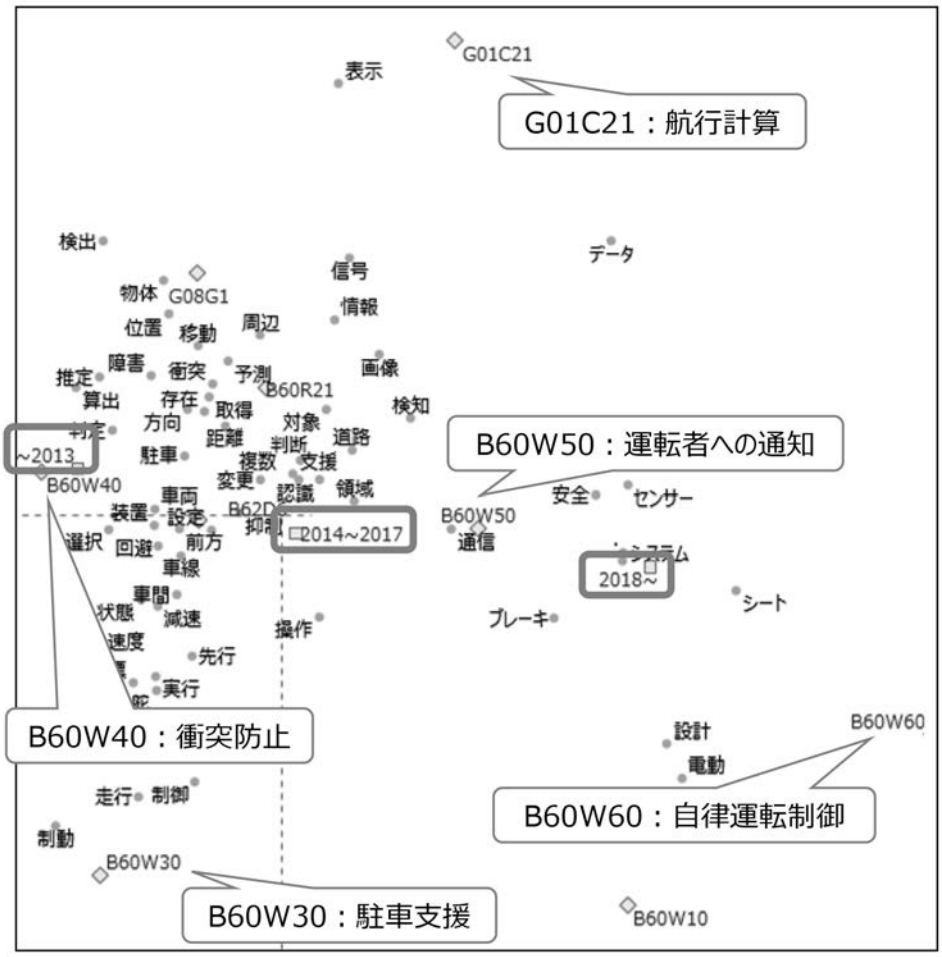

図２　自動運転車に関するテキストマイニング（対応分析）の結果[16)]

同様にもう一つの外部変数「IPC分類（自動付与されたもの）」は，◇でプロットしてIPC分類メイングループの記号（例えば航行計算に関するG01C21等）のラベルを付している。

２）対応分析の結果の考察

図２において，時期に関する情報を見ると，向かって左側に「～2013」，中央付近に「2014～2017」，右側に「2018～」が配置されている。すなわち図の左側から右側に推移している。

これを踏まえて図の左側から右側に向かってIPC分類のラベルを見ていくと，「衝突防止」や「駐車支援」から，「運転者への通知」，「自律運転制御」へと推移しており，実際の市販車両において，自動ブレーキやパーキングアシストから，車線逸脱警報，自動操舵といった機能が搭載されてきた経緯とも感覚的に一致する。

よって，この分析では特許情報と非特許情報を組み合わせた母集団に，学習結果を用いてIPC分類を自動付与し，テキストマイニングを行うことで，技術動向調査としてある程度妥当な結果が得られているものと考えられる。

(4) 実務への適用

今回は「答え合わせ」ができるように，2015年までの特許情報と2016年以降の非特許情報を母集団に用いたが，これを実務に適用する場合は，例えば「現時点で公開されている（１年半前までの）特許情報とそれ以降（直近１年半）の非特許情報」を母集団として同様の分析を試みるといったことを検討されたい。

4. 2　新用途探索

続いて対象業務として「新用途探索」を検討した結果を紹介する。具体的には，「多孔体」の市場展開に向けたターゲットアイテムの探索を想定し，KH Coderにて分析を行った。

(1) 分析手法

1）母集団作成

今回，「自社製品である多孔体の持つ特徴が，どのような分野で要求されうるか」という観点で分析を行うため，「自社の多孔体に関する特許[17]」を母集団として用いることとした。

2）IPC分類の学習による自動分類付与

技術動向調査と同様に，KH Coderを用いて，IPC分類リストを学習させた結果から，1）の母集団への自動分類付与を実施した。なお今回はIPC分類の階層をサブクラスまでのみとした[18]。

3）共起ネットワークによるキーワード分布

KH Coderを用い，「ツール」→「抽出語」→「共起ネットワーク」にて，1）の母集団中の特許に含まれるキーワードの共起ネットワークを描画した。なお，2）で付与済のIPC分類を外部変数として併用することで，各キーワードに紐付くIPC分類も表示させた。

(2) 分析結果

前項の手順により描画した多孔体関連特許のキーワードとIPC分類サブクラス（外部変数）の共起ネットワークを図3に示す。

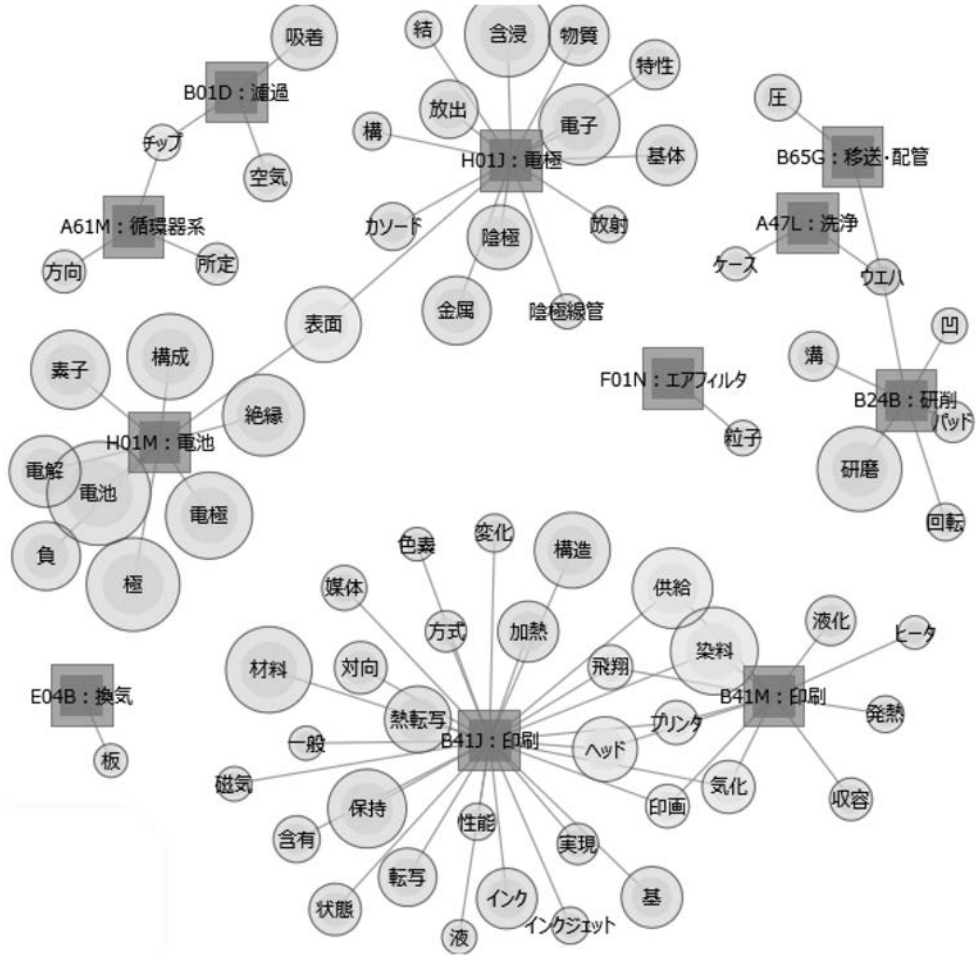

図3　多孔体関連特許のキーワードとIPC分類サブクラスの共起ネットワーク[19]

ここに表示されたIPC分類から，母集団中の特許に該当する多孔体がどのような分野に適用可能かを想定することができる。

もちろん自動付与されたIPC分類の多くは，母集団中の特許自体に元々付与されていたものと重複する可能性が高いが，中には特許自体に付与されていなかったIPC分類が，自動分類によって初めて付与される場合がある。

今回の結果では，図３中のA61M（循環器系），B01D（濾過），B24B（研削），B41JおよびB41M（印刷），B65G（移送・配管），H01J（電極），H01M（電池）は母集団中の特許自体にも付与されていたが，F01N（エアフィルタ），E04B（換気），A47L（洗浄）は自動付与された分類にのみ存在していた。これらは単にノイズかも知れないが，当該多孔体を適用できる可能性がありながら，出願時点は想定していなかった，潜在的な新用途を示唆している可能性もある。

今回，上記用途への実際の適用可能性の検証までは行っていないが，本手法は簡単な操作により短時間[20]で行うことができるため，新用途探索において検討の価値はあると期待する。

4. 3　分析における課題と対策

４章１節と４章２節では，本研究の分析手法について，妥当性のある結果を容易に得られる手法として紹介してきたが，実際には例示した結果を得るまでには幾つかの課題があり，その対策としての工夫を行ってきたため，以下ではこの点について述べる。

(1) 課　題

１）同義語の表記統一

特許情報か非特許情報かを問わず，発信者が違えば（時には同じ発信者の同じ文献内でも）同じ対象が異なる用語で表記されている場合が多々存在する。テキストマイニングでは文中の語句の出現頻度や位置関係を分析することから，同義語の異表記を合わせてカウントするか別々にカウントするかによって結果が変わりうる。当然，分析の趣旨からすれば同義語は合わせてカウントすべきであることから，これらを同じものとみなして処理するための工夫が必要となる。

２）一般的用語による技術的特徴語の埋没

特に有価証券報告書やプレスリリースなど，技術者以外の読者層も想定される非特許情報においては，技術的特徴語よりも一般的用語での記載が多くなる傾向がある。このため，技術的特徴語の出現頻度が相対的に低下[21]して，分析結果上には一般的用語ばかり表示されてしまい，目的の情報が得られない可能性がある。そこで，技術的特徴語を抽出して分析するための工夫が必要となる。

実際に上記１）や２）を考慮せず，取得した特許情報と非特許情報のテキストデータをただ統合したのみの母集団を，ある市販のテキストマイニングツール[22]にそのままインポートし，ヒートマップを描画の上，データソースを表示させたところ，特許情報由来のキーワード群と非特許情報由来のキーワード群とがマップ上で明確に分離して表示されてしまい，かつ非特許情報由来のキーワード群には一般的用語が多く，技術的特徴語が埋没してしまっていたことから，技術の推移が把握できなかった。

(2) 対策検討

１）同義語の表記統一

テキストマイニングに限らずパテントマップ等においても，従来から表記統一の課題は存在しており，特定の基準でキーワードを分類するいわゆる「コーディング」が行われてきたため，基本的にはこれに準じた対応で良いが，本研究では特にAIの活用による効率化手法として，インパテックが提供している「パテントマップEXZ[23]」の新機能「Wiki Dict.」を検討した。

Wiki Dict.とは，パテントマップEXZに入力されたキーワードについて，ウィキペディア[24]を参照し，自動的に表記統一（名寄せ）および分類・階層化する機能である。図４に例示するように，４章１節の自動運転車の分析に用いた特許情報に対してWiki Dict.の処理を実施し「ブレーキ」の項目を確認すると，「ブレーキ」自体の他に，「制動装置」，「リターダ」等の語が含まれることが見て取れる。

図４　Wiki Dict.による自動分類の例

こうして統合された各階層の項目はそのままマップ描画に用いることも可能だが，過不足等があれば，さらに人手でデータの追加等を行い仮想項目やリストの作成に用いることもできる。

従来，このようなキーワードの統合は手作業で抽出・編集を行っていたため，多大な工数を要していた。また自動化を図る場合も，初めに手作業でコーディングルールを設定する必要があり，特に新事業探索や新用途探索等馴染みのない分野の調査を行う際は，都度コーディングルールを新たに設定する必要があることに加え，当該分野の技術用語の知識が十分でない場合，コーディングルールの不備によりキーワードの適切な抽出や分類ができない懸念があった。

そこでこのような機能を活用することにより，馴染みのない分野を初めて調査する場合でも，キーワードの抽出や同義語の表記統一が，簡便かつ適切に行えることで，効率と精度の向上に繋げられることが期待される。

なお，Wiki Dict.を用いて表記を統一すべき同義語を把握した上で，前述のKH Coderでの分析に適用する方法としては，把握した同義語を検索条件としてテキストファイルに記述した，「コーディングルール・ファイル」を作成し，KH Coderでこれを読み込む方法[25]がある。

２）一般的用語による技術的特徴語の埋没

上記方法に対し，KH Coderに用いる母集団として初めから同義語を表記統一したテキストデータを用いる方法もあるため，以下紹介する。

この方法では，Wiki Dict.で表記を統一したキーワードを軸に用いてパテントマップEXZにて「マトリクスチャート」を作成し，それを文献ごとの技術的特徴語のリストに変換[26]し，

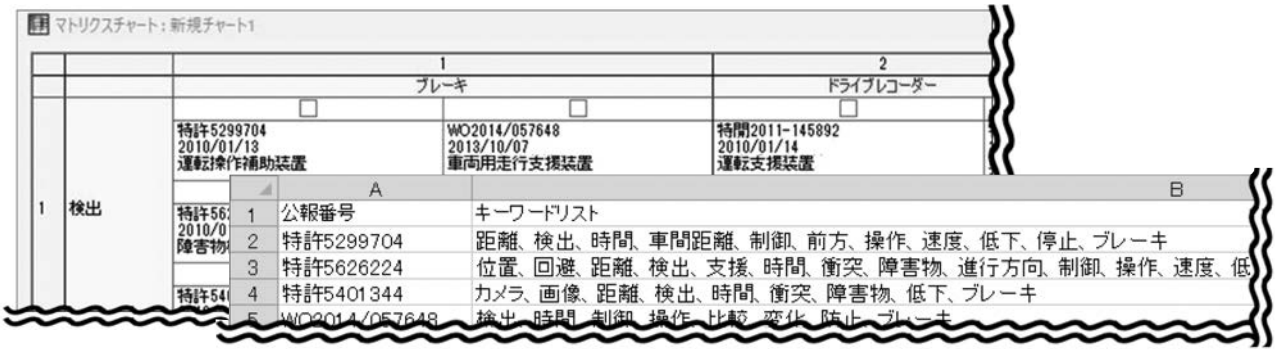

図５　自動運転車の特許情報のマトリクスチャート（左上）と技術的特徴語リスト（右下）

KH Coderの母集団として入力する。例として図５に，自動運転車の特許情報データを用いたマトリクスチャート（左上側）および，それを変換した技術的特徴語リスト（右下側）を示す。

前者の方法で，ソースから取得したテキストデータをそのままKH Coderに取り込んだ上で，コーディングにより同義語を表記統一する場合，当該キーワードの文章内での出現位置の情報（文脈）を反映した解析が可能となるが，一方，後者の方法であらかじめ技術的特徴語リストに変換（一般的用語を除外）したテキストデータをKH Coderに取り込む場合，一般的用語への埋没を抑制した解析が可能となるため，これらを適宜使い分けることを推奨する。

(3) その他ツールの検討

本研究において，なるべく多くの会員企業で適用できるように，汎用的な手法の提案を意識してはいたが，JDreamⅢやパテントマップEXZ（Wiki Dict.）のような有料ツールの他，コーディングやスクレイピングなどの専門的な知識やスキルが要求される手法も含まれていた。

そこで以下では，上述した以外の無料ツールを用いた簡便な分析手法として，「AI自動要約による特許要約文から，AIテキストマイニングにより特徴語を抽出する手法」を検討した結果を紹介する。

１）AI自動要約

図研プリサイトから，ウェブサイト上で利用可能な自動要約ツール[27]が提供されている。

当該サイトでは，入力フォームに最大20MBまでのテキストデータを入力し，「要約文生成」ボタンをクリックすることで，１～100行までの指定行数の要約文を生成することができる。

今回，自動運転車の分析対象に用いた各特許文献から「発明が解決しようとする課題」と，「発明の効果」を抽出し，これを入力データに用いて，特許の要約書と同等の400文字程度となるよう行数を調整し要約文を自動生成した。なお，上記にて「課題」と「効果」を抽出した理由は，

図６　AI自動要約の生成画面の例

最終的には要約文を「新用途探索」のための分析（結果は割愛）に適用する想定で，当該範囲には用途に関する記載が含まれている可能性が高いと推測したためである。

当該サイトでの自動要約生成画面の例を図６に示す。

２）AIテキストマイニング

UserLocalから，ウェブサイト上で利用可能なAIテキストマイニングツール[28]が提供されている。

当該サイトでは，テキストデータを入力することでワードクラウドや共起ネットワークなど，テキストマイニングで一般的に用いられる各種の分析結果を表示させることが可能である。

また，１つの母集団を解析するのみでなく，２つの母集団を比較して，どちらにどのようなキーワードが多く出現するか，またはどちらか一方にしか出現しないキーワードがあるか等の結果を表示する機能も備えている。

なお，これらの機能は無料で利用することが可能だが，入力できるテキストの文字数として10,000文字まで（ファイルをアップロードする場合は10MBまで）に制限されている。そこで，なるべく多くの文章を分析対象に含めるには，上記１）の自動要約などの併用も有用と考える。

今回，「２つの文章を比較」の機能を用いて，自動運転車の各特許にもともと付されていた「公報記載の要約」と，上記１）によって生成した「明細書（課題・効果）の自動要約」とを比較[29]した結果を図７および図８に示す。

図８より，「明細書の自動要約」には，「公報記載の要約」では見られていなかった「逸脱」や「白線」といった単語が出現していたことが確認された。

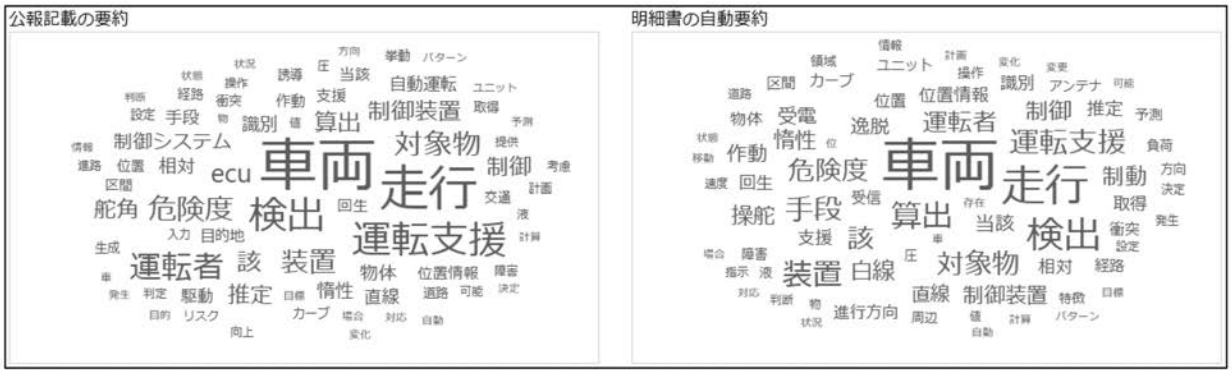

図７　公報記載の要約と，明細書の自動要約の比較（ワードクラウド）

公報記載の要約にだけ出現	公報記載の要約によく出る	両方によく出る	明細書の自動要約によく出る	明細書の自動要約にだけ出現
	可能	車両　走行　装置　検出 位置　制御　危険度 支援　設定　算出 運転支援　運転者　物 場合　対象物　車　推定 操作　直線　惰性　物体 対応　目標　相対　障害 判断　発生　当該 位置情報　識別	手段　情報　取得 ユニット　方向　作動 予測　決定	逸脱　白線

図８　公報記載の要約と，明細書の自動要約の比較（特徴語の出現頻度）

特許情報の分析において，全文を対象とした場合，処理に時間がかかることやノイズが増加する懸念があることから，要約書を対象とする場合もあると推測するが，分析の目的次第では今回のように対象を明細書の特定範囲としつつ，自動要約を用いることで，データ量を削減するとともに一般的用語を省略して，効率的な分析ができる可能性もあるものと思われる。

5．おわりに

以上のように本研究では，特許情報と非特許情報を組み合わせた分析として，非特許情報を母集団に含めたテキストマイニングを試行し，その過程で，非特許情報データの収集と分析における課題および対策を検討した。

その結果，AI（機械学習）によるIPC分類の自動付与を活用し，技術動向調査や新用途探索において，ある程度効果が期待されうる手法を見出すとともに，その過程で，EDINETからの有価証券報告書データ取得や，Wiki Dict.での同義語表記統一結果をKH Coderの外部変数に適用する際の，データ処理を効率化するためのツールも作成することができた。

一方で，PDFファイルからのテキスト抽出等については，これが容易にかつ適切に行えれば，解析に利用できる非特許情報が大幅に増加することが期待されるものの，具体的な手法の確立には至っていないため，今後の検討を期待する。

近年はPython等の普及により，機械学習やプログラミングも身近なものとなってきているため，会員企業各社の知財部員にも，それらを扱う素養を身に着けておくことで，自分自身の業務を効率化できる選択肢を広げることを提言したい。

とはいえ，やはり餅は餅屋ということもあるため，調査・分析ツールの各ベンダーにても，特許情報と非特許情報を組み合わせた母集団の分析，特に非特許情報データの収集も考慮したツールの開発を期待したい。

本研究が，IPランドスケープをはじめとする今後の会員企業各社における知財情報分析や，ベンダー各社におけるツール開発の検討の上で，多少でも参考になれば幸いである。

注　記

1) 日本知的財産協会2018年度情報検索委員会第５小委員会による，2019年７月度東西部会報告「IPランドスケープに関する実態調査と考察」における定義（「特許情報・非特許情報を活用した，経営に貢献するための知財分析活動」）を前提とする。http://www.jipa.or.jp/kaiin/katsudou/houkoku/bukaihoukoku/1907/05a_kensaku.pdf

2) 日本知的財産協会2019年度情報検索委員会第２小委員会「特許調査におけるAI等の活用に関する研究」（知財管理，Vol.70，No.12，pp.1767～1782（2020）），同「AIを用いた特許調査における業務効率化に関する研究－教師データの作り方の検討を中心に－」（知財管理，Vol.71，No.1，pp.88～102（2021）），および平尾啓「知財AI活用研究会の研究事例紹介」（情報の科学と技術，Vol.70，No.7，pp.349～354（2020））等。

3) 日本知的財産協会2018年度ソフトウェア委員会第２小委員会「ソフトウェア・IoT関連業界におけるIPランドスケープの活用方法の調査・研究」（知財管理，Vol.69，No.8，pp.1094～1105（2019）），2019年度情報検索委員会第３小委員会「IPランドスケープに関する研究（その１）」（知財管理，Vol.71，No.2，pp.251～265（2021）），山内明「IPランドスケープ2.0」（Japio YEAR BOOK 2018，pp.200～209（2018））他多数。

4) 野崎篤志「IPランドスケープの底流―情報分析を組織に定着させるために」（IPジャーナル，Vol.9，pp.32～38（2019））においても，「確証バイアス」に留意すべき旨が言及されている。

5) 論文や書籍などでは技術分野ごとの分類が付与されている場合もあるが，必ずしも特許情報と共通かつ統一された様式で付与されているとは言えない。なお，ソースによっては発信者や時期なども不明な場合がある。

6) EDINETのサイトで提供されている「XBRLからCSVへの変換ツール」は，下記URLからダウ

ンロード可能である。
https://disclosure.edinet-fsa.go.jp/E01EW/BLMainController.jsp?uji.verb=W1E63070InitDisplay&uji.bean=ee.bean.W1E63070.EEW1E63070Bean&TID=W1E63070&PID=currentPage&SESSIONKEY=1623449481214&downloadFileName=&lgKbn=2&dflg=0&iflg=0

7） 本研究の中で作成した有価証券報告書からのデータ抽出ツールは下記URLの「知財管理」誌バックナンバー付録（JIPA会員専用ホームページ）より入手可能である。
http://www.jipa.or.jp/kaiin/kikansi/chizaikanri/furoku.html
なお，このツールはEDINET閲覧（提出）サイト（https://disclosure.edinet-fsa.go.jp/EKW0EZ0015.html）の「EDINET APIサンプルプログラム」をもとに当小委員会にて作成した。

8） テキストデータを前提とした本研究では検証していないが，JDreamⅢでは1981年以降の論文にIPC分類のメイングループも付与されている（https://jdream3.com/service/search/ipc/） ため，これを特許文献のIPC分類データと組み合わせて分析に用いることも期待される。なお，論文に付与された技術分類を特許情報と組み合わせて分析した事例としては，2018年７月13日に開催された第15回情報プロフェッショナルシンポジウムにて，3i研究会第５期発表として「特許と論文の複合解析による有望応用分野の予測－印刷技術を例に－」が報告されている。

9） Octopus Data Inc.,「Octoparse」
https://www.octoparse.jp/
有料ツールだが無料版でも基本的な機能は利用可能（登録可能URL数等に制限あり）。

10） 実際に図表を含むPDFファイルのテキスト抽出を試行したところ，図表のタイトルや図表内のテキスト等が，本文の途中（図表配置箇所に相当する部分）に挿入されてしまい，本文のテキストと分離しての抽出はできなかった。

11） ある程度複雑なレイアウトのPDFファイルにも対応できるとされているツールの例として，「いきなりPDF（https://www.sourcenext.com/product/pdf/）」や「PDFelement（https://pdf.wondershare.jp/）」等が挙げられる。

12） 科学研究費補助金および立命館大学研究推進プログラムによる助成を受けた研究成果の一部として提供されている，テキストマイニング用のフリーソフトである。
https://khcoder.net/

13） 日本国特許庁「IPC分類表及び更新情報（日本語版）」
https://www.jpo.go.jp/system/patent/gaiyo/bunrui/ipc/ipc8wk.html

14） 実際の処理としては，２）により得られた，『IPC分類の各メイングループに含まれるキーワードの学習結果（例えばメイングループ「A01B」には「手作業具，鋤，ショベル，歯を有するもの，…」が含まれる）』をもとに，１）の『母集団の特許文献や非特許文献のテキストデータに含まれるキーワード』との対比から，母集団中の各文献がそれぞれどのIPC分類のメイングループに近いものであるかを判定させることで自動分類付与を行っている。

15） 母集団に用いた「技動」の特許情報の最先出願日または最先優先権主張日が2010年以降2015年以前であり，非特許情報の作成日または発行日が2016年以降2020年以前であるため，母集団全体には「2010年から2020年まで」の11年間の情報が含まれるが，図を見やすくするため便宜的に，2010年から2013年の４年間，2014年から2017年の４年間，2018年から2020年の３年間の３段階でラベルを区切った。

16） 図中でプロットされたIPC分類の幾つかには，吹き出しで「B60W40：衝突防止」等の説明を付しているが，これらは必ずしも日本国特許庁のIPC分類表に記載された説明をそのまま転記した訳ではなく，当該メイングループ内の各サブグループの説明等を踏まえ，便宜的に概略を記載したものである。

17） 具体的社名は伏せるが，ある企業の特許のうち，請求項に「多孔」や「ポーラス」を含むものを検索した結果を母集団とした。

18） IPC分類の階層の深さを，メインクラス，サブクラス，メイングループ等に変更した複数のリストをあらかじめ学習させておくことで，どの階層での分類を自動付与させるかを容易に変更することができる。今回は事前に各階層での分類付与を試行し，結果の見やすさ等を考慮しサブクラスでのリストを用いることとした。

19） 図中でIPC分類サブクラスに「H01J：電極」等の説明を付しているが，これらは図２と同様に

便宜的に概略を記載したものである。

20） 本手法では，あらかじめIPC分類リストを学習させておけば，母集団を変更した場合でも学習結果を用いた自動分類付与を，新たな母集団に対して容易に実施することが可能である。今回は事前に学習結果を準備済であったため，実際に図３の作成に要した時間は，母集団作成から描画までのトータルで30分程度であった。

21） 同義語の異表記が統一されていない場合，個々の表記の出現頻度はさらに低くなる。

22） 当該ツールのベンダーからは，条件付きで具体的内容開示の許可を得ているものの，特定ツールのネガティブな情報が独り歩きすることへの懸念を考慮して詳細は伏せている点，理解されたい。

23） インパテック株式会社，「パテントマップEXZ」
https://www.inpatec.co.jp/software/patentmap

24） フリー百科事典「ウィキペディア」
https://ja.wikipedia.org/wiki/メインページ

25） KH Coder「コーディングルール・ファイル」
https://khcoder.net/scr_docs_codfile.html

26） パテントマップEXZからマトリクスチャートをExcelにエクスポートすると，作成したマトリクスチャートの形式（各軸にキーワード，各セルに当該キーワードを含む文献の情報が入力された表）のままで出力される。本研究の過程でこれを図５右下のような，文献情報の列とキーワードを列挙した列のリスト形式に変換するためのExcelマクロを作成したため，これも「知財管理」の付録として，前掲注７）に記載のURLにて提供する。

27） 株式会社図研プリサイト，「ナレッジラボ①要約文の自動作成」
https://www.presight.co.jp/lp/detail/klab01.php

28） 株式会社ユーザーローカル，「AIテキストマイニング」
https://textmining.userlocal.jp/

29） 比較する２文章について，それぞれ最大10,000文字までという制限があることから，特許１件につき約400文字と想定して，25件の特許を無作為に抽出した集合を用いて比較した。

（URLの参照日は全て2021年７月12日）

（原稿受領日　2021年７月30日）