論　説

テキストマイニング技術の活用に関する研究

情報検索委員会
第３小委員会*

抄　録　企業の知財部門が，マクロ分析の手段として用いる，テキストマイニング技術の活用について検討した。

テキストマイニングの結果を活用する上での課題として，特定の観点における特徴的なキーワードがキーワードマップ上に描画されないという点に着目し，その原因の一つとして，抽出されたキーワードリスト中に様々な観点のキーワードが混在していることがあると考え，観点の絞り込み方法を検討した。この結果，第１に分析対象の選択，第２に選択した分析対象と他の箇所とのキーワードの使用頻度差に基づく特徴語の抽出やストップワードの設定が有効であることを確認したので報告する。また最後に，本研究を通して見出された，テキストマイニングツールを提供するベンダーへの要望を述べる。

目　次

1．はじめに

特許や技術論文に代表される，知財情報の有用性に対する理解が広がるにつれ，企業の知財部門に対して，経営層に向けた知財情報分析に基づく提言を求められる場面が増えている。知財部門では，かかる要求に応えるため，迅速かつ質の高い知財情報分析を進める必要がある。

このような知財情報分析のステップの例としては，マクロ分析によるポジション把握，セミマクロ～セミミクロ分析によるテーマ設定，ミクロ分析によるテーマ深掘り等を経ての将来予測が提案されている[1)]。

他方，知財情報のマクロ分析を迅速に行う手法として，テキストマイニング技術を用いた分析（テキストマイニング）が広く知られている。また，テキストマイニングを行い，その結果を描画するツール（テキストマイニングツール）

* 2018年度　The Third Subcommittee, Information Search Committee

も広く知られている。

そこで本稿では，テキストマイニングによるマクロ分析の結果を，迅速かつ質の高い知財情報分析に繋げる上での課題を検証し，テキストマイニングの質を高めうるテキストマイニングツールの活用方法を検討したので報告する。

また最後に，検討を通してメンバーが実感した市販のテキストマイニングツールに対するベンダーへの要望を述べる。

2．課題の検証

本章では，テキストマイニングツールによる分析結果と，特許庁が公開している特許出願技術動向調査等報告書（技術動向調査報告）に示された概要とを比較し，知財情報をテキストマイニングする上での課題を検証した。

なお，技術動向調査報告を比較対象とした理由は，第１に，かかる報告が「企業の研究開発戦略において大変有用な情報[2)]」であるとされており，「企業のグローバル活動に伴う，世界規模での特許出願動向の基礎資料として，各国・機関における特許出願動向調査 -マクロ調査[2)]」が実施され，提言の形にまとめられていることから，実施されている知財情報分析の目的が，我々のそれに近いと考えられるためである。また第２に，技術動向調査報告の概要中には簡潔な俯瞰図が示されており，テキストマイニングで抽出されたキーワードが描画されたマップ（キーワードマップ）との比較が容易なためである。

2．1　課題の検証方法

検証対象とする技術動向調査報告のテーマとしては，活用分野が広く，メンバーの専門分野にかかわらず比較的評価しやすいと考えられた，平成25年度報告の「ロボット」を選定した。

母集団は，かかる技術動向調査報告に記載された検索式に基づいて抽出した日本出願5,847件とした。この母集団について，テキストマイニングを実施し，キーワードマップを作成し，技術動向調査報告の概要中に示された俯瞰図と比較し，テキストマイニングによるマクロ分析の結果を，迅速かつ質の高い知財情報分析に繋げる上での課題を検証した。

検証に用いたテキストマイニングツールは，Derwent Innovation（ThemeScape），Biz Cruncher，KH Coder，Orbit Intelligence，TechRader，Text Mining Studio，CyberPatent Desk テキストマイニング（旧「TRUE TELLER パテントポートフォリオ」（以下，「CPD」））の７つである。

2．2　課題の検証結果

参照した技術動向調査報告の概要[3)]には，技術俯瞰図が示されており，観点毎にキーワードがまとめられている。

すなわち，まず技術全体を応用技術と要素技術に分け，応用技術を，「産業用ロボット分野」と「サービスロボット分野」に分け，さらに「サービスロボット分野」を「特殊環境用ロボット」と「サービスロボット」に分けた上で，それぞれの具体的な用途が列挙されている。

また，要素技術は，「全体構造技術」，「安全技術」，「制御技術」，「知能化技術」，「認識・コミュニケーション技術」の５つに分けた上で，それぞれの具体的な技術が列挙されている。

一方，テキストマイニングツールを用いて作成したキーワードマップについて，説明する。Biz Cruncherで描いた図１では，左側にアーム関連，右下に歩行関連，右上に制御関連のキーワードが配置されているが，それらのキーワードの観点は統一されていない。

すなわち，技術動向調査報告に示された俯瞰図が，用途，技術などの観点毎にキーワードを整理しているのに対して，キーワードマップは，様々な観点のキーワードが混在している。

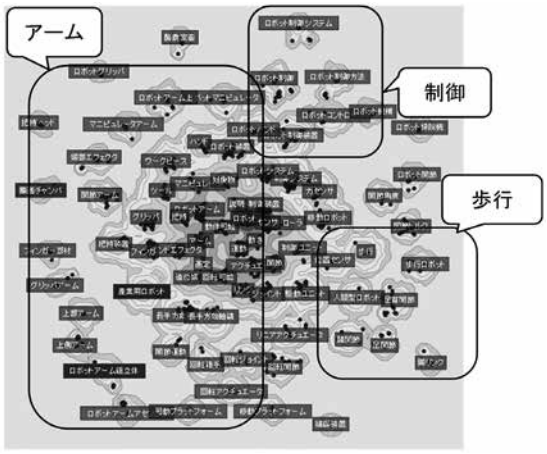

図１　Biz Cruncherによるキーワードマップの例

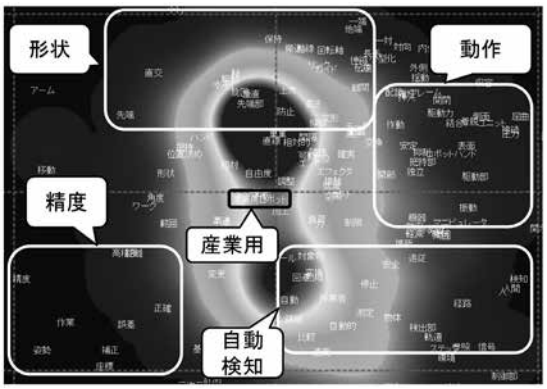

図２　CPDによるキーワードマップの例

我々は，技術動向調査報告のように，観点毎にキーワードを整理して俯瞰することが，我々の目的において求められると仮定し，テキストマイニングツールを用いて作成したキーワードマップ中のキーワードを観点毎に整理しようと考えた。しかしながら，様々な観点のキーワードが描画されたキーワードマップでは，特定の観点における特徴的なキーワードが，描画されていない場合があることに気づいた。

例えば，CPDで描いたキーワードマップを示す図２においては，用途における特徴的なキーワードとしては，「産業用」が描画されているのみである。結果として，かかるキーワードマップから，ロボット分野における用途を俯瞰することは困難である。

なお本稿では，「用途における特徴的なキーワード」とは，「ロボット自体の用途に関するキーワード（例えば「産業用」）」を意味し，ロボットを用途とする材料や部材等に関するキーワード（例えば「ロボットアーム」）は，含まないものとする。

2．3　設定した課題

以上のことから本研究では，特定の観点における特徴的なキーワードがキーワードマップ上に描画されないという点を課題として設定した。

次に，設定した課題の原因を検証した。図２に示したキーワードマップの元になる，テキストマイニングで抽出されたキーワードリストの上位300件を確認したところ，表１に示すように，150位以降にキーワードマップ中に描画されなかった用途における特徴的なキーワードが複数見つかった。

表１　キーワードリスト中の用途における特徴的なキーワード

順位	単語
31	産業用
173	搬送ロボット
196	歩行補助ロボット
255	移動ロボット
262	医療用

このことから，用途における特徴的なキーワードの多くがキーワードマップ中に描画されない理由は，キーワードとして抽出されていないわけではなく，様々な観点のキーワードが混在して抽出されるため，全体としては使用頻度が低いキーワードと判断され，描画できるキーワードの数（または現実的に確認可能なキーワードの数）の上限によって描画対象から外れているためと考えた。

3．検討結果

そこで，本研究では，テキストマイニングによるマクロ分析の質を向上させるべく，様々な観点のキーワードの混在が少なく，注目する観点における特徴的なキーワードが描画されたキーワードマップを得る方法を検討した。

また，検討した方法が，より多くの会員企業に実用的なものとなることを目指した。具体的には，特定のテキストマイニングツール独自の機能によることなく活用できることも考慮し，フリーツール（KH Coder）を含む複数のテキストマイニングツールで同様の考え方に基づいた分析を実施した。

この結果，第１に分析対象を適切に選択すること（3．1（2）項，3．2（1）項，3．3（1）項参照），第２に選択した分析対象と他の箇所とのキーワードの使用頻度差に基づく特徴語やストップワードを設定すること（3．1(3)～(4)項，3．2（2）項，3．3（2）項，3．4節参照）が有効であることを確認したので，以下，事例を用いて報告する。

なお，母集団としては，上記した課題の検証に用いたロボットに関する日本出願5,847件を用いた。

また，分析対象としてはDWPI抄録と明細書を選択した。

3．1　DWPI抄録の分析事例

本節では，DWPI抄録を分析した事例を示す。

(1) DWPI抄録とThemeScape

DWPI抄録（DWPI）は，Clarivate Analytics社が提供する，世界50以上の特許発行機関が発行した特許文献を対象に，手作業で付加価値情報を付与したデータである。

DWPIは様々な観点ごとにまとめられている。このうち，新規性の観点に関するテキストを収録した「DWPI－新規性」，用途の観点に関するテキストを収録した「DWPI－用途」，発明の優位性の観点に関するテキストを収録した「DWPI－優位性」は，すべての発明に対してデータが存在しており，特に有用である。

DWPIを基礎的データの１つとした，研究開発活動の調査と分析のための情報ソリューションとして，Derwent Innovation（DI）がある。

DIには，ThemeScapeと呼ばれるテキストマイニングツールが存在する。ThemeScapeを用いて，DWPIに含まれるテキストデータを分析し，その類似性に基づいて特許文献を配置したキーワードマップが得られる。またかかるキーワードマップでは，件数が集中する領域には，共通するキーワードが描画される。

このようなキーワードマップを用いることで，分析対象において，例えば，どのような技術が主流かといった情報や，どの特許文献が類似しているかといった情報を容易に把握することが可能となる。また，ThemeScapeは，分析対象とするDWPIの観点を選択できる。さらに，分析対象としたくないキーワードをストップワードに設定することも可能である。

(2)「DWPI－用途」の分析

図３に前記母集団の「DWPI－用途」を分析対象として，ThemeScapeで作成したキーワードマップを示す。

図３に示したキーワードマップには，用途における特徴的なキーワードとして，例えば，「Elderly people」，「Motor vehicle」といったキーワードが描画されている。一方で，「Robot hand」や「Joint」といった，用途との関連性が低いキーワードも同時に描画されている。

用途との関連性が低いキーワードがキーワードマップ中に描画された影響を調べるために，「DWPI－用途」に「hand」というキーワードが含まれている特許文献がどのように描画され

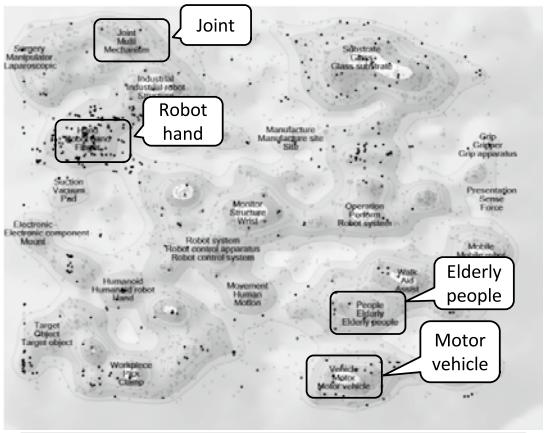

図３　「DWPI－用途」のキーワードマップ

ているかの確認を行った。その結果，これらの特許文献が，特にキーワード「Robot hand」の周辺に多く配置されていることが分かった（図３中に濃い点として表示）。さらに，これら，「Robot hand」の周辺に配置された特許文献の「DWPI－用途」の記載を確認したところ，用途における特徴的なキーワードを含むものがあることを確認できた。

そこで，これら「Robot hand」の周辺に配置された特許文献を，用途における特徴的なキーワードの周辺に再配置できるように，キーワードマップを調整する方法について検討を行った。

(3) 用途に関するキーワードの自動判定

先にも述べたが，ThemeScapeはストップワードを設定可能である。そこで，「Robot hand」や「Joint」といった用途との関連性が低いキーワードをストップワードに設定することで，埋没していた用途における特徴的なキーワードが浮かび上がってくると考えた。

DWPIは観点ごとに分けられているが，共通して用いられるキーワードが存在する。例えば，「Robot hand」は，「DWPI－新規性」では「The robot hand has …」のように構成の説明に用いられる一方，「DWPI－用途」では「Robot hand for …」のように利用シーンの説明に用いられる。このように観点を縦断して現れ，出現頻度も観点ごとに差が少ないキーワードは，特定の観点（例えば，新規性や用途）における特徴的なキーワードではない可能性が高いと考えた。

すなわち，「『DWPI－新規性』や『DWPI－優位性』よりも『DWPI－用途』に多く記載されているキーワードは，用途における特徴的なキーワードである可能性が高い」という仮説を立て，この仮説に基づき，DWPIに記載されたキーワードが用途における特徴的なキーワードである可能性を示すスコアを以下のように定義した。

$$\frac{DF(\text{用途})}{\max\{DF(\text{新規性}),\ DF(\text{優位性})\}+1}$$

上記定義において，DF(X)は，対象とするキーワードを含むXの観点のDWPI「DWPI－X」が付与された特許文献の数を表す。このスコアは「そのキーワードが『DWPI－新規性』及び『DWPI－優位性』と比較して，どの程度『DWPI－用途』に多く現れるか」の指標となる。例えば，「DWPI－用途」と比べて「DWPI－新規性」または「DWPI－優位性」に数多く現れるキーワードは，本スコアが１より小さくなり，「DWPI－新規性」および「DWPI－優位性」のいずれよりも「DWPI－用途」に多く現れるキーワードではスコアが１以上となる。本スコアがしきい値以下となるキーワードをストップワードとして設定することで，用途における特徴的なキーワードを効率よく抽出し，描画できると考えた。

しかしながら，すべてのキーワードに対して，手作業で上記スコアを算出することは，現実的ではない。例えば，用いた母集団中に現れるキーワード（単語１～３つの組み合わせ）は，38,374通り存在する。そこで，１行１公報で列

ごとに項目が分けられている形式の公報データから，自動的に上記スコアを算出するプログラムを作成した。本稿の付録[4]に本プログラムとその解説を記載する。

(4) ストップワードを設定したキーワードマップ

図４に，上述の方法で算出したスコアが，２未満のキーワードをすべてストップワードに設定して描画したキーワードマップを示す。

「DWPI－用途」に「hand」というキーワードが含まれている特許文献（濃い点で表示）が，図３と比較して分散して配置されていることが分かる。

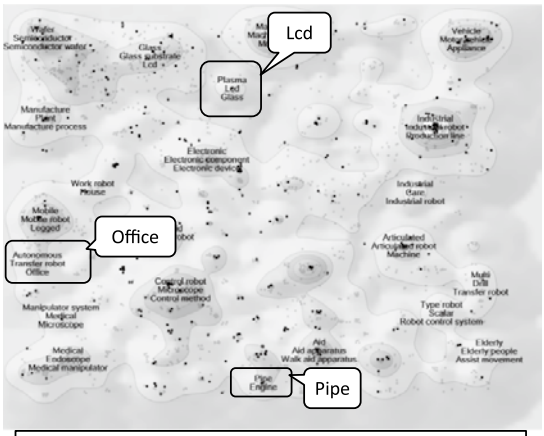

図４　「DWPI－用途」のキーワードマップ（ストップワードあり）

一方，図４のキーワードマップに描画されたキーワードに着目すると，「Robot hand」や「Joint」が描画されなくなった一方で，「Office」，「Pipe」，「Lcd」等の図３では描画されていなかった用途における特徴的なキーワードが新たに発掘された。

また，図３において「Robot hand」の周辺に配置された，いくつかの特許文献について，表２に，「DWPI－用途」の記載（表中「DWPI USE」としているカラム）と，図４においてどのキーワードの周辺に配置されたか（表中「図４での配置」としているカラム）を示す。このように，「DWPI－用途」に，記載された用途における特徴的なキーワードに基づいて，図４で再配置されていることが確認できた。

このように，上記スコアを用いてストップワードを設定することで，用途を俯瞰しやすいキーワードマップを作成できることが分かった。

表２　「Robot hand」の周辺に配置された案件の「DWPI－用途」の記載と図４での配置領域

DWPI USE	図4での配置
Robot having hand for cellular manufacturing system.	Manufacture
Robot hand of carrier robot, for gripping workpieces.	Carrier robot
Robot hand used for carrier robot (claimed).	Carrier robot
Multi-fingered robot hand for an industrial robot and humanoid robot.	Humanoid
Hand structure for use in a robot (claimed) i.e. triaxial drive-type robot, for gripping a cylindrical article i.e. pipe.	Pipe

3. 2　明細書の分析事例１

本節以降は，前記母集団について，明細書を分析した結果を示す。

本節では，KH Coderを用いて分析した事例を示す。明細書における「要約」あるいは「技術分野」をテキストマイニングによる分析対象とし，テキストデータはインターネット特許検索サービスであるCKSWebを用いて準備した。キーワードマップは，関連性の高いキーワード同士を距離と実線のネットワークとして描画できる共起ネットワーク図を選択した。

(1) 分析箇所検討事例

まず,「要約」を分析対象にして，分析，描画を実行した例を示す（図５）。

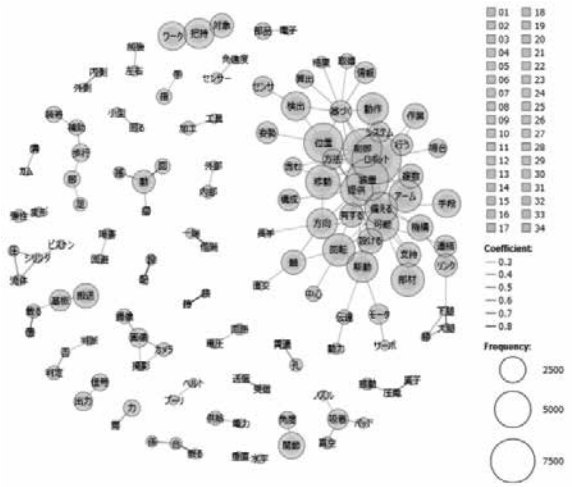

図５ 「要約」の共起ネットワーク図

当該図５の右側には比較的広大なネットワークを確認できるが，具体的な課題までは捉えられず,「要約」中に同じく含まれる解決手段と紐付いた状態として認識することもできない。

これは,「要約」には課題と解決手段という複数の観点からの文章が含まれているにもかかわらず，テキストマイニングツールではそれらを区別したり紐付けたりする機能を持たないことが原因であると推測される。

また，技術的な内容を表すものではない一般的な動詞も多く見られ，それにより用途における特徴的なキーワードの一部が埋没して，描画されていない可能性が考えられる。

次に,「技術分野」を分析対象にして，分析，描画を実行した例を示す（図６）。

この例では，特定の用途及び構造を示すと考えられるネットワークを確認することができる。

これらは,「要約」を分析対象とした場合には確認されなかったもので,「技術分野」を分析対象として選択したことにより，単一の観点から分析が行われたためであると考えられる。

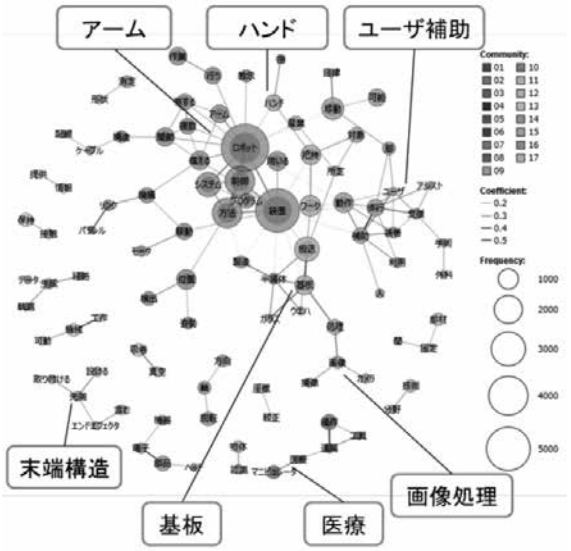

図６ 「技術分野」の共起ネットワーク図

このように，当事例においては，集合に含まれる主な技術内容を俯瞰するような大まかな分析の場合には,「技術分野」が分析対象として適していることが示唆される結果となった。

なお,「要約」及び「技術分野」の共起ネットワーク図の作成に際して，描画結果の見やすさを最大化した上で両者のネットワーク分布の粗密を均一にすべく，設定に差異を設けた。具体的には,「要約」に対しては，ストップワードとして,「図」,「少なくとも」,「それぞれ」,「前記」，及び「発明」を設定し，かつ最低出現数が170以上のキーワードを使用した共起のうち，上位120を表示する設定とした。「技術分野」に対しては，ストップワードとして,「特許」,「出願」,「米国」,「本明」,「細書」,「参照」,「開示」,「実施」,「形態」,「発明」,「特に」，及び「係る」を設定し，かつ最低出現数が100以上のキーワードを使用した共起のうち，上位130を表示する設定とした。最低出現数は，任意回数以上出現したキーワードのみを描画させるための足切り設定である。

(2) ストップワード設定方法と事例

「技術分野」を分析対象として，用途におけ

る特徴的なキーワードをより積極的に描画するため，ストップワード自動判定を検討した。

ストップワード自動判定には，各キーワードに以下のように定義するスコアを用いた。

$$\frac{DF(「技術分野」)}{\max\{DF(「請求項」), DF(「課題」), DF(「効果」)\}+1}$$

上記定義においては，DF(X)は，対象とするキーワードが，記載箇所Xに記載された特許文献の数を表す。

本定義では，キーワードのスコアが1.0以上であるとき，そのキーワードが他の記載箇所（請求項，課題，効果）と比べて技術分野により多く出現することを意味する。一方スコアが小さいほど，技術分野以外の記載箇所により多く出現することを意味する。

ストップワードを設定しなかった場合（図７），スコアが0.4以下のキーワードをストップワードに設定した場合（図８），及びスコアが1.0以下のキーワードをストップワードに設定した場合（図９）の共起ネットワーク図を示す。

それぞれの共起ネットワーク図は，「技術分野」における出現数が多い順に上位200語のキーワードを描画した。

共通設定として，特許明細書の定型文として登場する「出願」,「米国」,「本明」,「細書」,「参照」,「開示」,「実施形態」,「発明」,「特に」,「係る」を予めストップワードとして設定した。

スコアが0.4以下のキーワードをストップワードとした場合（図８）では,「自動車」や「コミュニケーション」など用途における特徴的なキーワードのネットワークがストップワードを設定しなかった場合（図７）と比較して新たに出現した。

一方，スコアが1.0以下のキーワードをストップワードとした場合（図９）では,「外科手術」など用途における特徴的なキーワードの一部がストップワードに含まれて消失し，特許明細書

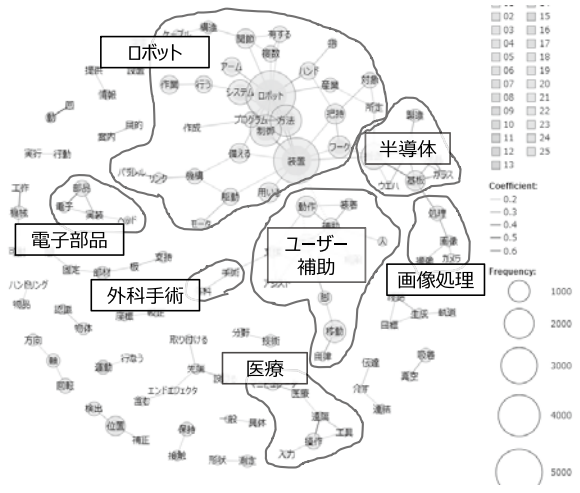

図７　ストップワードを設定していない「技術分野」共起ネットワーク図

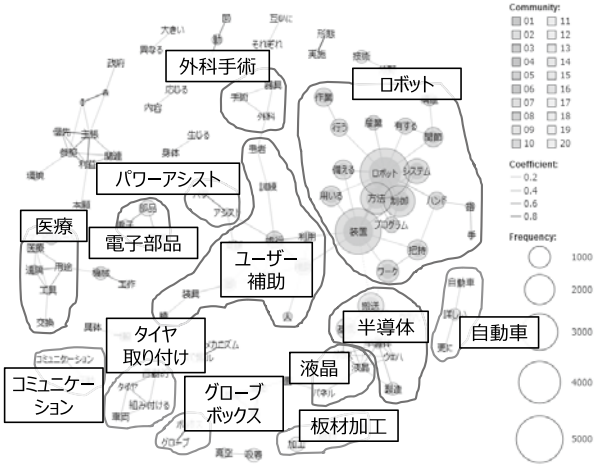

図８　スコア0.4以下のキーワードをストップワードとした「技術分野」共起ネットワーク図

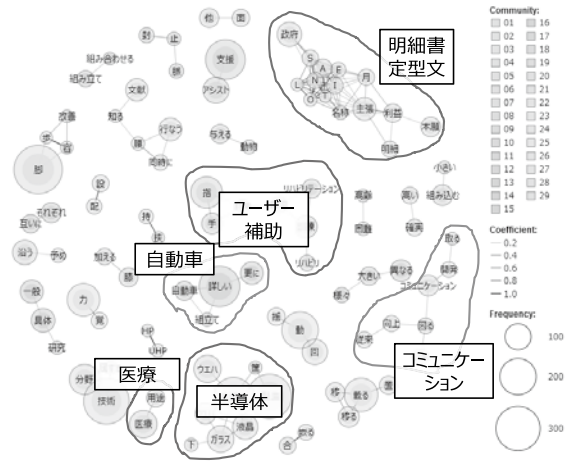

図９　スコア1.0以下のキーワードをストップワードとした「技術分野」共起ネットワーク図

の定型文である「本願」や「明細」などのノイズワードが新たに出現した。

このように，自動判定によってストップワードを設定して，描画することで，図８のように用途を俯瞰しやすいキーワードマップを作成できることを確認した。

一方で，ストップワードに設定するスコアのしきい値の設定には注意が必要である。参考までに，スコアのしきい値とキーワードの出現状況との関連を表３に示す。表中，それぞれのスコアのしきい値において出現したキーワードについて「○」で示す。

表３　ストップワードに設定したスコアと，キーワードの出現

ストップワードに設定したキーワードのスコア	画像処理	ロボット全般	半導体	ユーザ補助	外科手術	液晶	パワーアシスト	電子部品	自動車	医療	コミュニケーション	板材加工	タイヤ取り付け	グローブボックス
標準条件（図7）	○	○	○	○	○			○		○				
スコア0.3以下		○	○	○	○	○		○	○	○	○	○	○	
スコア0.4以下（図8）		○	○	○	○	○	○	○	○	○	○	○	○	○
スコア0.5以下		○	○	○	○	○	○	○	○	○	○	○	○	
スコア0.6以下		○	○	○	○	○	○	○	○	○	○	○	○	
スコア0.75以下		○	○	○		○		○	○	○	○		○	
スコア1.0以下（図9）			○	○					○	○	○			

3. 3　明細書の分析事例２

本節では，Biz Cruncherを用いて分析した事例を示す。

このテキストマイニングツールでは，各キーワードの希少性と多様性という２つの尺度に基づいて非公開の方法で算出される，重要度という独自のスコアを用いて，その公報内での特徴を示すキーワードが抽出される。

希少性とは，そのキーワードが対象の公報群内にどれだけ出現するかの尺度であり，出現数が少なければ希少性が高くなる。一方，多様性とは，公報内におけるそのキーワードの使われ方のバリエーションの尺度であり，バリエーションが多ければ多様性が高くなる。

(1) 記載箇所による観点の絞り込み

前述の重要度は，明細書内の記載箇所毎に算出することもできるので，記載箇所毎に特徴的なキーワードを抽出できると考えられる。

例えば，「請求項」であれば構成，「技術分野」であれば用途や技術における特徴的なキーワードを抽出できると予測した（表４）。

表４　記載箇所毎に含まれうると予測した観点の一覧

記載箇所	含まれうる観点
発明の名称	構成／用途
要約	構成／課題
請求項	構成
技術分野	用途／技術
背景技術	技術
課題	課題／用途／技術
発明の効果	効果／用途
利用可能性	用途

この予測に基づいて分析対象とする記載箇所の選択を行い，複数の記載箇所を組み合わせて分析する方法を検討した。

この方法によって，複数の記載箇所に記載されている注目する観点における特徴的なキーワードの重要度が上がり，それ以外のキーワードの重要度が相対的に下がることで，結果的にキーワードマップ上に，注目する観点における特徴的なキーワードが多く描画されると考えた。

ここでは，用途における特徴的なキーワードが含まれそうな記載箇所として，「技術分野」，「課題」，「発明の効果」，「利用可能性」を分析対象とした。

Biz Cruncherにおいて初期設定されている「発明の名称」，「要約」，「請求項」を分析対象として，重要度上位80語を描画してキーワード

マップ化すると，用途における特徴的なキーワードが13語描画された（表５及び図10）のに対し，分析対象を変更し，「技術分野」，「課題」，「発明の効果」，「利用可能性」とした結果では，「歩行補助装置」や「人型ロボット」等，新たなキーワードが抽出され，用途における特徴的なキーワードが26語に増加した（表６及び図11）。

しかし依然として，用途以外の観点のキーワードも多い上，単一の記載箇所のみに記載されているキーワードの中には，用途における特徴的なキーワードでありながら，記載箇所を組み合わせることで逆に埋没したものもあった。

(2) 複数記載箇所の減算組み合わせ

次に，複数の記載箇所を引き算の形で組み合わせる方法を検討した事例を報告する。

「技術分野」のみを分析対象として，同様に重要度上位80語を描画してキーワードマップ化すると，用途における特徴的なキーワードが23語描画されたものの，逆に用途との関連性が低い「ロボットアーム」や「マニピュレータ」といった，技術に関するキーワードも多く含まれていた（表７及び図12）。

そこで，技術に関するキーワードがより多く含まれそうな「背景技術」を分析対象として算出した重要度を，「技術分野」の分析で算出された各キーワードの重要度から減算し，重要度の再集計を行った（図13）。

つまり，「技術分野」で抽出されたキーワードのうち，「背景技術」でも抽出されたキーワードは，再集計された重要度が下がることになる。

この処理の後に作成したキーワードマップ（図14）では，「荷搬送ロボット」などの埋没していたキーワードが描画され，重要度上位80語中49語が用途における特徴的なキーワードとなった。これらのキーワードには，「ワーク搬送装置」，「ワーク搬送方法」，「ワーク把持方法」等，

表５　初期設定で分析対象となる記載箇所と観点

記載箇所	含まれうる観点
発明の名称	構成／用途
要約	構成／課題
請求項	構成

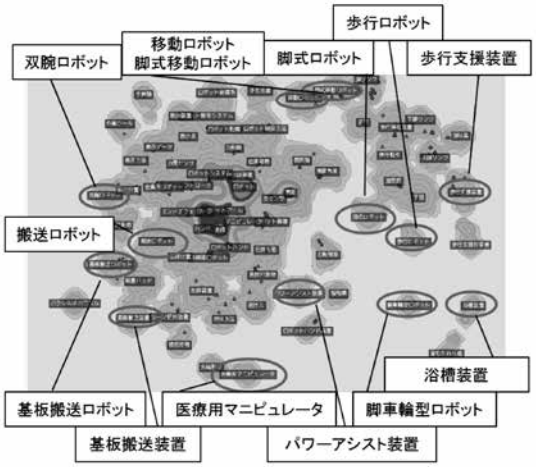

図10　初期設定での分析

表６　変更後の分析対象（記載箇所）と観点

記載箇所	含まれうる観点
技術分野	用途／技術
課題	課題／用途／技術
発明の効果	効果／用途
利用可能性	用途

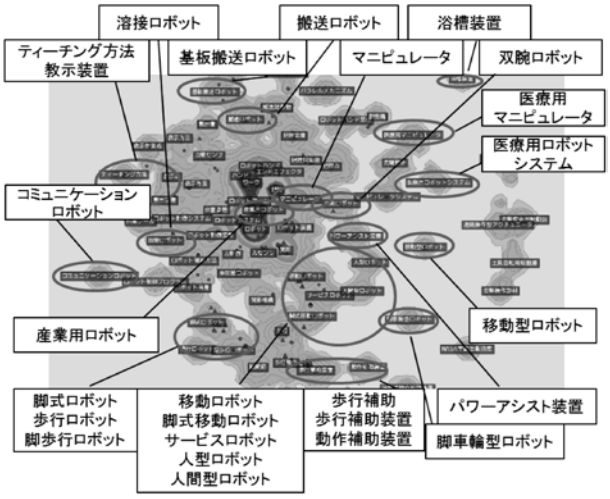

図11　分析対象を変更

表7　技術分野に含まれると考えられる観点

記載箇所	含まれうる観点
技術分野	用途／技術

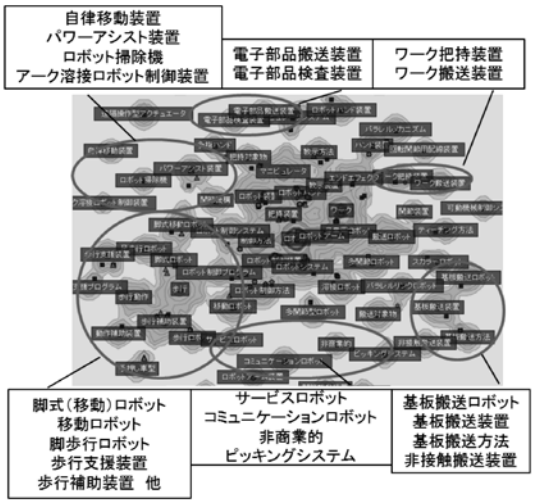

図12　技術分野のみの分析

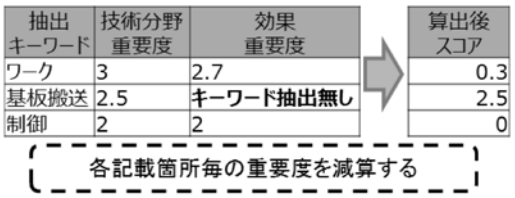

図13　重要度の再集計に関する概念図

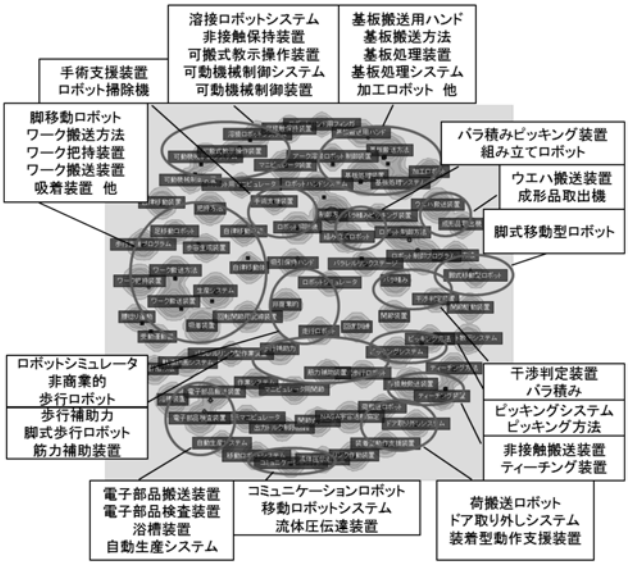

図14　技術分野から背景技術を減じた分析

より細かい用途における特徴的なキーワードも含まれていた。

これらの方法は，重要度というBiz Cruncher独自のスコアを用いて算出しているが，例えば記載箇所間のキーワードランキングの違いを用いるなど，同様の考え方を使って，他のテキストマイニングツールにも応用できるのではないかと考えている。

3. 4　明細書の分析事例３

本節では,CPDを用いて分析した事例を示す。

まず分析対象となる記載箇所を選択し，次いで，選択した記載箇所と他の記載箇所とのキーワードの比較に基づいて特徴語を抽出した。

テキストデータはCKSWebを用いて準備した。キーワードの抽出は名詞のみから行った。また特徴語はCPDの標準機能で，χ^2値を利用したクラメールの連関係数の,確率分布中央(期待値）からのずれに基づいて算出した。

(1) 特徴語抽出における比較対象の影響

分析対象として，明細書中の「発明の効果」を選択し，キーワードを抽出したが，効果における特徴的なキーワードは殆どランキング上位に入っていなかった。

そこで，別の記載箇所と比較して，特徴語を算出した。比較対象とする記載箇所として,「請求項」,「技術分野」,「課題」を用いて，それぞれ特徴語を算出した。

表８に,「発明の効果」から抽出した上位15位までのキーワード（「KWランキング」として示したカラム）と，比較対象毎に算出した上位15位までの特徴語（「特徴語ランキング（比較対象別)」として示した３つのカラム）を示した。白抜きのキーワードは効果とは異なる観点のキーワードであり，黒字のキーワードが効果に関するキーワードである。

比較対象として,「請求項」を用いた場合に

効果における特徴的なキーワードが最も効率的に抽出できることが分かった。

表８　キーワードランキングと，特徴語ランキングへの比較対象の影響

順位	KWランキング	特徴語ランキング（比較対象別）		
		比較対象		
		請求項	技術分野	課題
1	ロボット	精度	位置	方向
2	位置	高精度	方向	位置
3	ワーク	小型化	精度	距離
4	動作	向上	形状	回転
5	精度	効率	力	互い
6	方向	安全性	時間	長さ
7	姿勢	短時間	姿勢	形状
8	作業	負担	向上	回動
9	形状	信頼性	高精度	移動
10	力	軽量化	小型化	力
11	移動	生産性	距離	動作
12	アクチュエータ	コスト	作業者	角度
13	アーム	小型	効率	上方
14	制御	作業効率	長さ	一端
15	センサ	低コスト	安全性	態様

(2) 特徴語抽出における観点統一の影響

次に母集団をキーワードで絞って５つの集合を作成し，各集合と残りの４つの集合との，「発明の効果」で使われているキーワードを比較することで特徴語を算出した。特徴語は集合毎に上位10個ずつ描画した。

集合の絞り込みに用いるキーワードとして，まず前記した上位キーワードの「ロボット」，「位置」，「ワーク」，「動作」，「精度」を用いた結果が図15である。複数のキーワードに共通して特徴語とされたキーワードが多いことが分かる。

これに対して，比較対象を「請求項」とすることで抽出した効果における特徴的なキーワードである「精度」（「高精度」も類義語として扱い，あわせて集合を作成した），「小型化」，「効率」，「安全性」，「短時間」の５つを，集合の絞り込みに用いた結果が図16である。

この結果，複数のキーワードに共通して特徴語とされたキーワードがなく，より多くの特徴語を抽出できた。

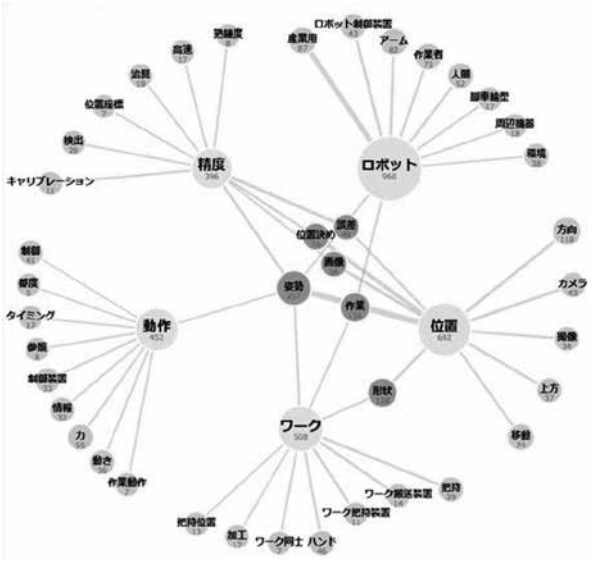

図15　上位キーワードを含む集合の特徴語

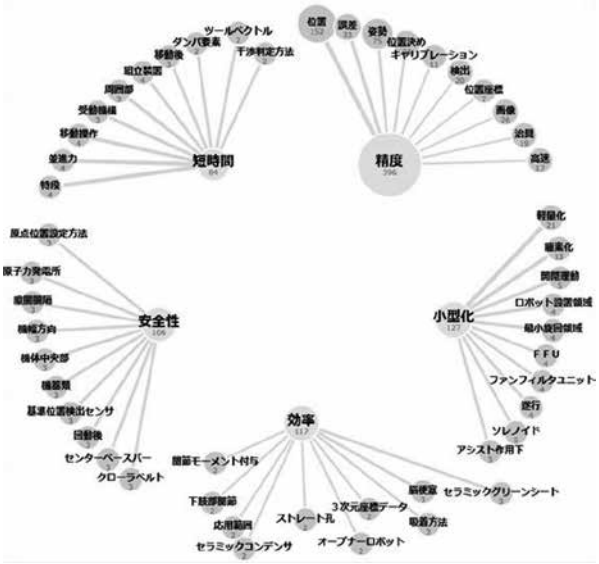

図16　効果における特徴的なキーワードを含む集合の特徴語

(3) 特許分類を用いた特徴語抽出

次に記載箇所を限定した上で，付与された特許分類で比較して特徴語を抽出した例を示す。

国際特許分類（IPC）のセクションＡ（生活必需品）が付与されている出願には，用途発明が多く，さらにそれら用途発明中の用途におけ

る特徴的なキーワードは，「発明の名称」に記載されているという仮定に基づき，母集団中のセクションＡが付与されている出願を抽出し，それらの「発明の名称」に使われているキーワードを抽出した場合の上位15位までのキーワードが表９の左のカラム（KWランキング）である。

表９　記載箇所とIPCを組み合わせたキーワード抽出

	セクションAが付与された出願群	セクションAが付与された出願群 vs それ以外の出願群
順位	KWランキング	特徴語ランキング
1	歩行補助装置	歩行補助装置
2	マニピュレータ	医療用
3	医療用	歩行支援装置
4	制御方法	浴槽装置
5	歩行支援装置	動作補助装置
6	システム	遠隔操作型
7	アクチュエータ	運動補助装置
8	遠隔操作型	装着式動作補助装置
9	動作補助装置	装着型
10	浴槽装置	歩行支援プログラム
11	プログラム	マニピュレータ
12	運動補助装置	他動運動機器
13	マニピュレータシステム	動作支援装置
14	制御装置	手術用
15	装着型	脚装具

白抜きのキーワードは用途とは異なる観点のキーワードであり，黒字のキーワードが用途における特徴的なキーワードである。予想通り，上位の多くを占めていることが分かる。

次に，かかるセクションＡが付与されている出願群と，セクションＡが付与されていない出願群との「発明の名称」に記載されているキーワードを比較して，セクションＡが付与されている出願群の特徴語を抽出したのが表９の右のカラム（特徴語ランキング）であり，より効率的に用途における特徴的なキーワードを抽出できることが分かった。

4．ベンダーへの要望

本研究で得られた成果は，各種のテキストマイニングツールにおいて活用できる汎用性の高いものであると考えている。その一方で，今回の検証を通して，本研究の成果を実際に活用する上では，各種テキストマイニングツールには共通して望まれる事項のあることが明らかとなった。既に一部のテキストマイニングツールでは搭載済みの機能も含まれるが，検証を実施する上で有用であった事項を含め以下に記載する。

4．1　明細書の記載箇所毎に分離されたデータベースとの連携

本研究で着目したように，明細書の記載箇所毎に抽出されるキーワードの偏りを用いることで，特定の観点を浮き出させることが可能だと考えられる。例えば，「技術分野」には発明の用途における特徴的なキーワードが，「請求項」には発明の構成における特徴的なキーワードが，「発明の効果」には効果における特徴的なキーワードが，それぞれ多く含まれると考えられる。そこで，それぞれの記載箇所ごとに分析することで，用途や構成，効果といった特定の観点を俯瞰することが可能となる。

このような分析を行うには，データベース内の明細書のデータが，記載箇所毎に分離されている必要がある。テキストマイニングツールにおいても，この分離された記載箇所毎にキーワードを抽出できることが望ましい。

4．2　特許分析に特化した辞書，及び辞書機能の充実

特許を分析対象とするテキストマイニングの場合，どのような技術領域であっても，共通する分析に不要なキーワードが多く含まれている。また，実際には同義のキーワードであるにもかかわらず，キーワードの前後に特許特有の文字や符号が付与されていることによって，異義のキーワードとして扱われるケースも存在する。上記のようなキーワードをストップワード辞

書，あるいは同義語辞書に分析の都度登録することは大変手間がかかるため，迅速に分析を進めるためには，これらのキーワードについてあらかじめ処理されていることが望ましい。具体的には，ストップワードの辞書としては，「請求項」や「実施例」等のキーワードが確実に登録されていること，同義語としては，冒頭に「上記」や「前記」，あるいは末尾に数字が付されているキーワードを同義語として扱う処理がなされていることが望ましい。

また，キーワードマップにおいては，既に明らかとなっている観点がその他の観点とどのような位置関係にあるか等，描画したいキーワードがあらかじめ決まっているケースが想定される。そのようなケースのため，ユーザが描画したいキーワードを辞書に登録し，描画するキーワードの一部を指定することができる機能を有することが望ましい。

4．3　データ入出力（ツール間相互利用性）機能の充実

本研究では，複数のテキストマイニングツールで解析を行う中で，各テキストマイニングツールはそれぞれ独自の分析機能を有し，そのアウトプットから様々な切り口の情報を得ることが可能であることから，複数のテキストマイニングツールを用い，アウトプットを適宜組み合わせて総合的な分析を行うことが想定されると考えた。複数のテキストマイニングツールを用いて分析を行う場合，分析に使用するデータ，すなわち母集団データや，独自に設定した同義語，ストップワード等の辞書データは共通して用いることになる。その場合，例えばcsvファイルでの一括入力が可能である等，データの入力が迅速に行え，かつ簡易であることが望ましい。また出力も同様に行えることも同時に求められる。

4．4　分析ロジックの説明

本研究では，同じ母集団を用いて複数のテキストマイニングツールで解析を行ったが，それぞれから得られた分析結果（例えば抽出されたキーワード）は，予想以上に異なるものであった。これらの差異がなぜ生じるものなのか，本研究で追求することはできなかった。実際の業務においては，テキストマイニングを用いた分析結果に基づいて，分析結果が何を示唆するかについて論理的に説明することが求められる。しかしながら，用いたテキストマイニングツールの分析ロジックが非公開となっている場合は，分析結果の詳細についての言及が困難である。そのため，分析ロジックは可能な範囲で明らかになっていることが望ましい。

5．ま と め

当小委員会では，企業の知財部門が，例えば経営層に向けた提言を行う際に，マクロ分析の手段として用いる，テキストマイニング技術の活用について研究した。

技術動向調査報告に描かれた観点毎にまとめられた俯瞰図から，目指す提言において，観点毎に特徴的なキーワードを整理することが求められるとの仮定に基づき，テキストマイニングで得られたキーワードマップを確認したところ，特定の観点における特徴的なキーワードがキーワードマップ上に配置されていないという課題にたどり着いた。

この課題を解決するために，注目する観点の特徴的なキーワードを優先的に抽出する必要があると考え，検討した結果，第１に分析対象を選択すること，第２に選択した分析対象と他の箇所とのキーワードの使用頻度差に基づく特徴語の抽出やストップワードの設定が有効であることを確認した。

さらに，複数のツールを用いて上記を検証し，

会員企業の持つ各種のテキストマイニングツールにおいても活用できる，汎用性の高い方法であることが期待される結果を得た。

また，本研究で得られた成果を活用する上で，各テキストマイニングツールにおいては，①明細書の記載箇所毎に分離されたデータベースとの連携，②特許分析に特化した辞書，及び辞書機能の充実，③データ入出力（ツール間相互利用性）機能の充実，④分析ロジックを説明できること，などが実現されていることが望ましいと結論付けた。これらは，一部のテキストマイニングツールでは不十分な点もあり，今後のベンダー側での改善を期待したい。

6．おわりに

本研究に携わった2018年度情報検索委員会第３小委員会委員は，菊山茂樹（小委員長，クラレ），市川岳史（日清オイリオグループ），太田貴久（昭和電工），座古泰裕（シチズン時計），柴田潔子（住友電工知財テクノセンター），堤奈緒子（トヨタテクニカルディベロップメント），南宅崇人（ダイキン工業），吉武和志（ダイヘン），吉村裕子（大日本印刷）である。本稿が会員企業においてテキストマイニングツール活用の一助となれば幸いである。

注　記

1） Japio YEAR BOOK 2017, pp.198-205, 山内 明, IPランドスケープ実践に役立つ知財情報戦略－特許マーケティングを中心として－

2） 特許庁，資料・統計，刊行物・報告書，特許出願技術動向調査等報告書
https://www.jpo.go.jp/resources/report/gidou-houkoku/tokkyo/index.html
（参照日：2019年３月５日）

3） 平成25年度特許出願技術動向調査報告書概要（ロボット）
https://www.jpo.go.jp/resources/report/gidou-houkoku/tokkyo/document/index/25_robot.pdf
（参照日：2019年４月15日）

4） 「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）
http://www.jipa.or.jp/kaiin/kikansi/chizaikanri/furoku.html

（原稿受領日　2019年６月26日）