テキストマイニング（データマイニング）技術紹介

NRIサイバーパテント株式会社　企画営業部　　中居　隆

近年、厳しい経済環境の下で、企業においては、研究開発投資の効率性、収益性が厳しく問われている。そこでは、他者との競争の中で、自らのポジション、強み・弱みを的確に把握し、知的財産のポートフォリオの最適化を図るという視点が求められる。

こうした中で、大量の特許や論文等のデータに対して、テキストマイニング技術を用いて解析することで、自社・他社の知財ポートフォリオをマクロ、ミクロな視点から詳細に分析・評価する手法の活用が広がっている。本稿では、テキストマイニング技術を用いた情報の分析について、分析のプロセス、特徴、企業・大学・研究機関における活用状況、今後の展望を示すとともに、特許庁の審査関連業務への適用可能性について考察する。

1. テキストマイニング技術の発展

はじめに、日本国内におけるテキストマイニング技術の発展と拡がりの経緯を紹介する。

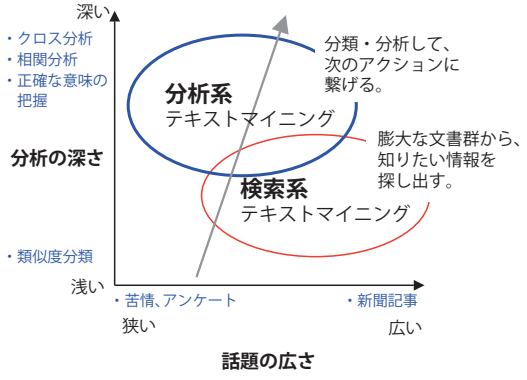
1）分析系はマーケティング分野を中心に発展

テキスト情報を扱う「自然言語処理」の技術は、検索系と分析系の2つの流れで発展し、活用が進んできた（図表1）。検索系は、インターネットの普及とともに、飛躍的に発展し、利用が広がってきた。特に、特許や文献などの技術情報の検索サービスでは、大量のテキストデータを対象に、指定された条件に基づく、漏れのないデータ抽出が求められ、日進月歩で高速化、利便性の向上が図られている。一方、分析系のテキストマイニングは、これまで日本では、消費者向けのサービス事業者や、消費財メーカのマーケティング部門など、CRM（Customer Relationship Management）をキーワードとして、主にマーケティング分野を中心に活用されてきた。

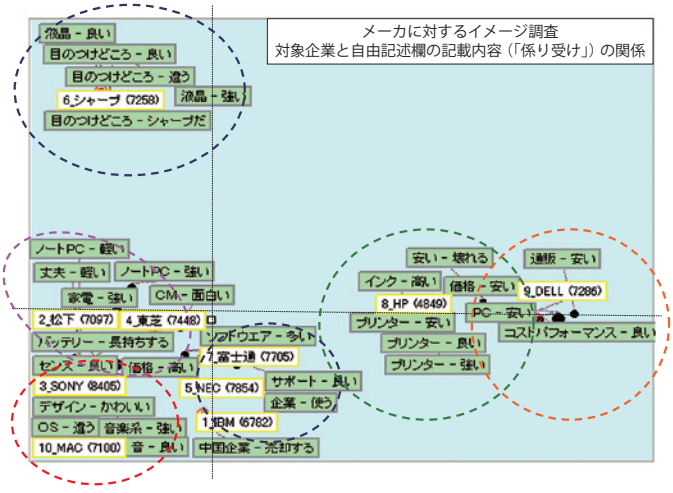
ひとつは、アンケート調査結果の分析である。アンケート調査では、回答者の隠れた潜在意識や、想定外の回答まで引き出すために、選択式の設問と、自由記述式の設問が使い分けられるが、従来、後者については、人手で1件1件を読み込み、仕分け・集計することに多大な労力・コストが掛かった。こうしたデータを電子化し、テキストマイニングによって解析することで、回答者の年齢、性別、居住地といった属性と、自由記述に出現する言葉の関係を見たり、設問間での回答の相関を見ることが可能となる（図表2）。

もうひとつは、お客様からの苦情・要望に対応する「コールセンター」に寄せられる声の分析である。例えば、「先週、発売した新製品Aに対して、10代女性からは『蓋が固くて空けにくい』といった不満の声が増えている」といった分析結果から、製品・サービスそのものの改善や、プロモーション戦略の見直しにつなげるといった具合だ。顧客の声に対応するという、コストセンターとしての取り組みが、プロフィットに繋がるということもあり、急速な拡がりを遂げてきた。さらに、先進的な企業では、こうした取り組みをシステム化することで、コールセンターに寄せられた声が、翌日には分析レポートとして、担当役員、社員に社内のポータルサイトを通じて、自動配信されるといったことが行われている。

こうした分析系のテキストマイニングでは、深く文章の意味やニュアンスを捉え、分析者の求める観点や問題意識に応じて簡単に分析・検証できたり、分析者



図表1　検索系と分析系のテキストマイニング



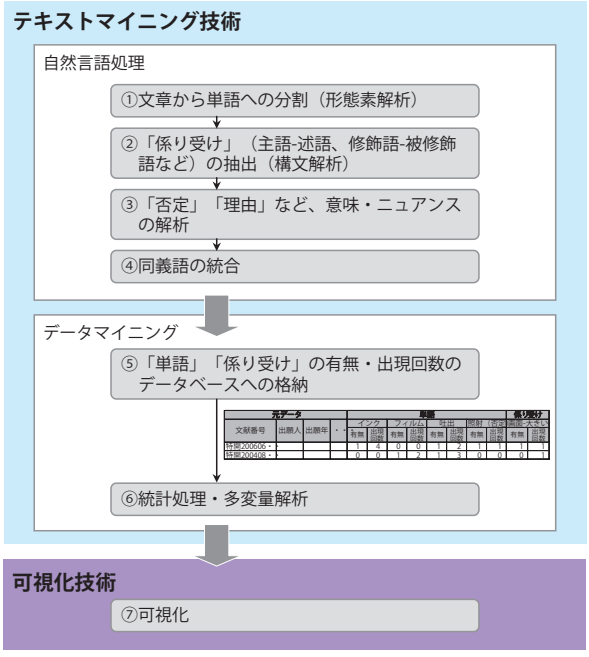
図表2　マーケティング分野におけるテキストマイニングの活用例

が気付かない項目間の関係を発掘するといった、検索系のテキストマイニングとは異なる性能・機能が求められる。

以上のように、マーケティング分野においては、すでにテキストマイニングが、企業の定常業務を支える当たり前のツールになりつつある。これに対して、技術情報に対する日本語でのテキストマイニングが、大手企業を中心に活用され始めたのは、2004年前後であり、まだまだ、発展途上の段階にあると言える。

2）テキストマイニングによるポートフォリオ分析　3つの技術要素

テキストマイニングによるポートフォリオ分析は、大きく、テキストマイニング技術と可視化技術から構成される。さらにテキストマイニング技術は、自然言語処理の技術とデータマイニングの技術に分けることができる（図表3）。

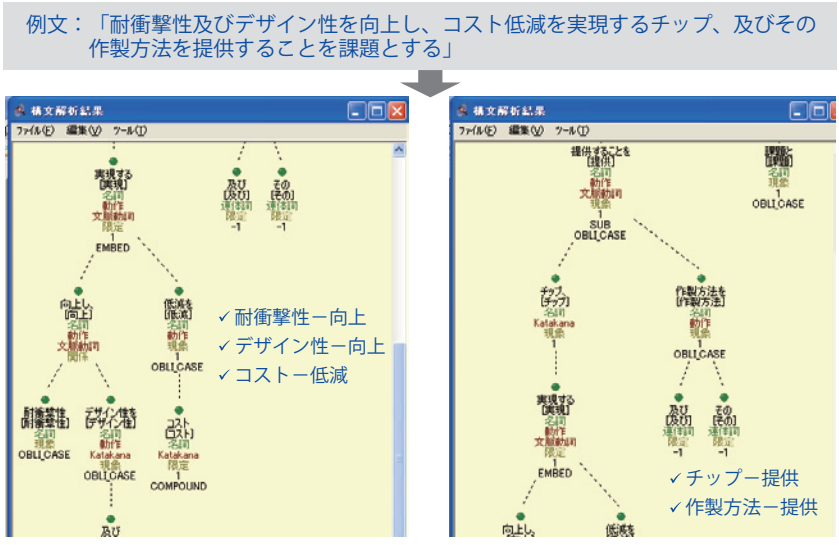


図表3　テキストマイニングによるポートフォリオ分析技術構成

（1）テキストマイニング技術　～自然言語処理～

「自然言語処理」は、大量のテキスト情報について、自然文（文章）から、その構成要素や構造、意味を抽出し、解析するものである。初めに、①句読点や接続詞と、②テキストマイニングツールが保有する言葉のデータベース（「辞書データ」）を手掛かりに、「形態素解析」と呼ばれる手法により、文章から単語の切り出しを行う。

次に、ひとつの文章の中で、言葉間の繋がり＝「係り受け」（主語－述語、述語－目的語、修飾語－被修飾



図表4　構文解析による「係り受け」の抽出

語など）を捉える。さらに、動詞の活用や文末の表現からより詳細に意味、ニュアンスを捉える（「構文解析」）。さらに、シソーラスを活用することで、同義語の統合まで行うことができる（図表4）。

（2）テキストマイニング技術　～データマイニング～

「データマイニング」とは、統計処理や人工知能等の手法を用いて、大量のデータから、データ項目間の相関性を見つけ出すデータ分析の技術である。例えば、米国のウォルマートは、POSデータをデータマイニングで分析することにより、金曜日の夕方は、ビールとオムツを一緒に購入する男性客が多いことを発見したというもので、一見、無関係そうで、人の経験や勘では気づかない発見を見つけ出す（掘り起こす＝マイニング）ものである。

「テキストマイニング」は、「自然言語処理」により自然文をデータ処理しやすい「単語」「係り受け」にデータ化した上で、属性項目を含めた項目間の関係を「データマイニング」により解析するものである。特許データの場合には、①出願人や出願年、特許分類コードなどの書誌的事項、審査経過情報と、②公報の文章から切り出した単語や係り受けの関係性を、データマイニングにより解析するということになる。

（3）可視化技術

データマイニングを用いた分析では、多くのデータ項目間で二軸、三軸による分析を簡単に行うことができるため、逆に、いかに分析結果を分かりやすく、シンプルに可視化できるかが鍵となる。特に、知財のポートフォリオ分析では、知財の専門家ではない、経営者、現場の研究者と、知財部門のスタッフが、同じ土俵の上で理解・議論できるアウトプットが求められる。単に、クロス集計の結果をグラフや表形式で示すだけでなく、テキストマイニングならではの言葉をベースにしたマッピング手法や、組織や人の繋がりをネットワーク図として可視化するなど、多様な可視化手法が提案されている。

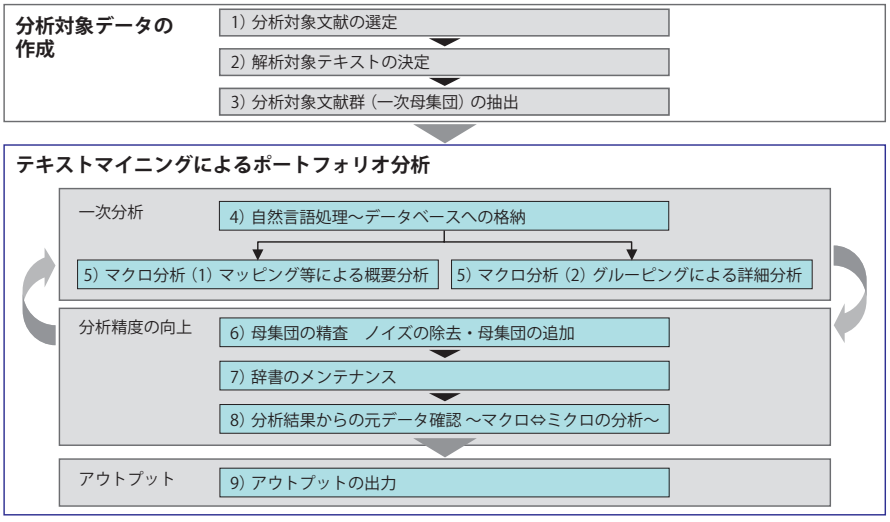
2. テキストマイニングによる分析プロセス

本章では、現在、技術情報のテキストマイニングとして活用が拡がっている、企業の事業戦略、研究開発戦略、知財戦略の策定、検証を目的とした知財のポートフォリオ分析を例として、そのプロセスと特徴を説明する（図表5）。

1）分析対象文献の選定

テキストマイニングによる技術情報の分析では、分析の目的に応じて、技術に関わるさまざまなデータソースを使い分けたり、統合して解析することができる。

例えば、大学や研究機関における基礎研究を含めた



図表5　テキストマイニングによる知財ポートフォリオ分析のプロセス

最先端の技術動向を広く捉えるためには、公開された特許情報に加えて、論文情報を見る必要がある。その際、特許情報と論文情報では、特許分類のような共通のコード等がないため、言葉をベースとしたテキストマイニング分析が有効な手段となる。

その他にも、自社の最近の技術開発への取り組みについて、関連する他社の取り組みとの関係・ポジションを把握するために、自社の公開前の出願情報を、公開された特許情報と重ねて分析することも行われている。

2）解析対象テキストの決定

特許情報をテキストマイニングを用いてマクロに分析する場合、解析対象とするテキストは、発明の内容・特徴が簡潔に記述され、関連する重要な言葉が網羅的に含まれている一方、発明の本質とは直接関係しない言葉は極力含まれないものが望ましい。

例えば、公開公報の「要約」は、発明の対象、特徴、課題、解決手段が、簡潔に記載されているケースが多い反面、出願人・代理人によって書き方や記載内容のレベルにばらつきがある。「特許請求の範囲（請求項）」は、権利範囲が明確に記述されているが、請求項によって、発明の本質との関係性、重要性が異なるため扱いが難しい。そこで、第一請求項、あるいは、独立請求項のみを自動抽出し解析対象とする場合もある。そのほか、発明の用途を分析したい場合には、「実施例」や「産業上の利用可能性」が有効となる。特許データのXML（Extensible Markup Language）化により、明細書の中から、必要な項目のテキストデータを切り出すことが容易になり始めており、分析の目的に応じた解析対象テキストの使い分けが可能な環境が整いつつある。

また、例えば、論文データの場合には、商用のデータベース事業者が「抄録」や「キーワード」を付与しているケースがある。テキストマイニング分析には、極めて有効なデータソースと言える。

3）分析対象文献群（一次母集団）の抽出

分析の対象とする文献群は、従来の特許分類やキーワードによる検索や、いわゆる「概念検索」などを用いて、特許データベースから抽出する。

筆者は、テキストマイニング分析が有効な母集団の件数について問合せを受けることが多いが、これまでの経験から、母集団が広くなれば、関連技術を含めた動向を幅広く把握できる反面、母集団全体に対する一次分析の観点、粒度は荒くなると言える。一方、母集団が絞り込まれれば、より詳細な観点、粒度による分類・分析ができるが、周辺技術まで把握することはできない。

テキストマイニングのソフトウェアによっては、母集団に出現した単語・係り受けリストから、想定外のノイズを見つけ除外することもできるため（後述）、このようなツールでは、ある程度のノイズを許容するつ

もりで、少し広めに一次母集団を抽出し、分析に着手することも有効である。

4）自然言語処理 ～データベースへの格納

一般に、テキストマイニングのシステム・ソフトウェアでは、母集団のデータを取り込む際に、1章で述べた、「自然言語処理」を一括して行う。さらに、抽出された単語や「係り受け」を、文献データの属性情報（特許の場合、出願番号、出願年、出願人、発明者などの書誌的事項、法的ステータスを表す審査経過情報など）と紐付けた形で、データベースに格納する。

以上までのプロセスが、いわば、分析の「前処理」ということになる。

5）マクロ分析

（1）マッピング等による概要分析

①言葉をベースにしたマッピング

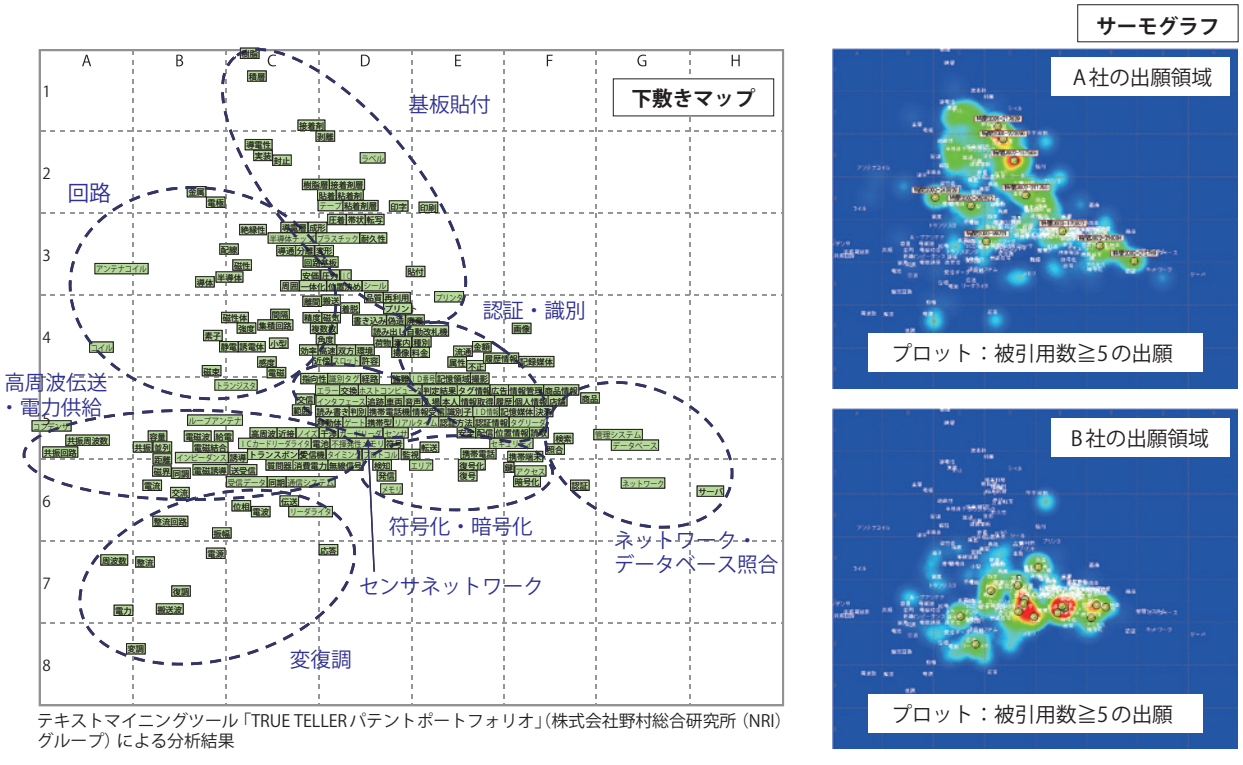
母集団全体の技術領域と、出願人、出願年別の出願状況と推移の概要を把握する方法として、テキストマイニングならではの、母集団の出現単語・係り受けの頻度分布に基づくマッピング分析が拡がりを見せている。

図表6のサーモグラフでは、はじめに、母集団の文中に頻出する技術用語どうしが同時に出現する公報の件数が多いほど、マップ上で単語間の距離が近くなるように単語を自動配置し、分析のベースとなる「単語マップ」を作成する。こうして単語の分布によってできた技術的な領域を、共通の「下敷き」として、そこに出願人や発明者別の文献の分布密度を温度表示（「サーモグラフ」）として表している。

図表7のコレスポンデンスマップでは、出願年と単語群の相関、出願人と単語群の相関を、1枚のマップに表現することで、各年代、各企業における特徴的な取り組みを浮き彫りにしている。

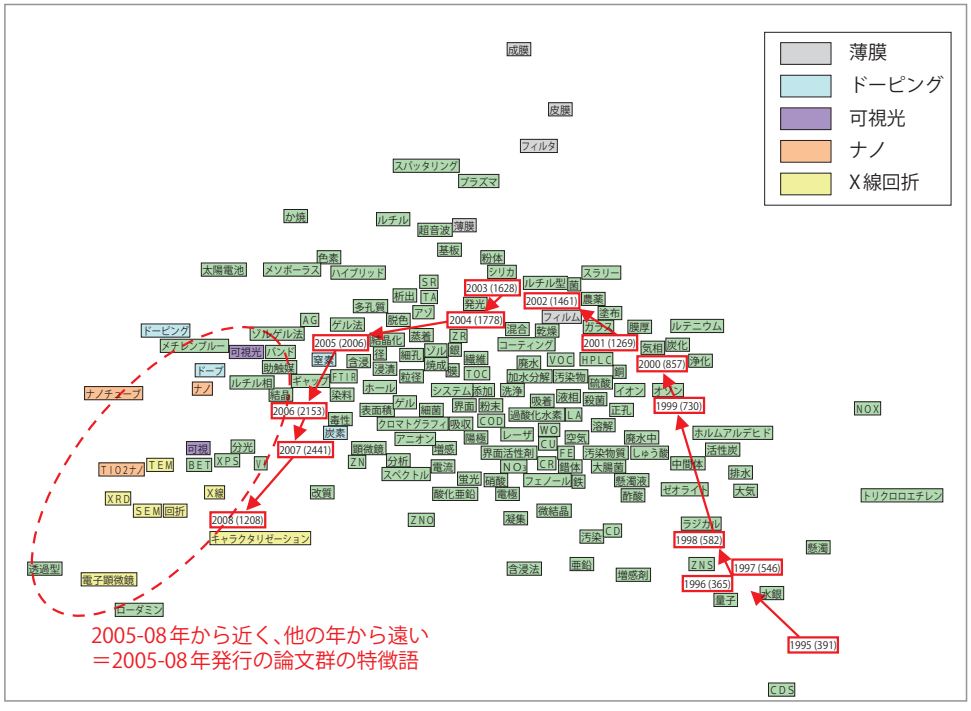
②企業間や発明者間の繋がり（共同出願・共著）を可視化するネットワーク分析

図表8は、ある研究機関における特許・論文データから、発明者間、および、発明者と企業との共同発明の繋がりをネットワーク図として可視化したもので、テキストマイニング技術と、ネットワーク分析の技術を



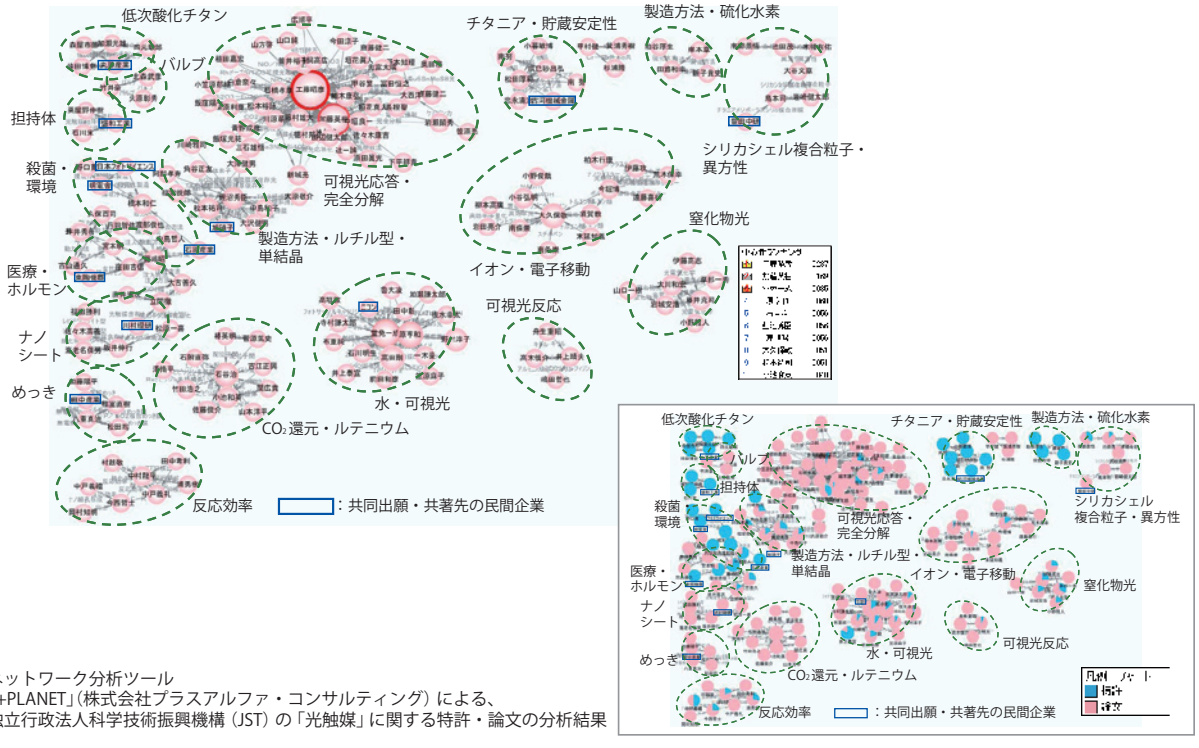
テキストマイニングツール「TRUE TELLER パテントポートフォリオ」（株式会社野村総合研究所（NRI）グループ）による分析結果

図表6　言葉によるマッピング分析例1 「サーモグラフ」



テキストマイニングツール「TRUE TELLERパテントポートフォリオ」(株式会社野村総合研究所（NRI)グループ）による、
「光触媒」に関する分析結果

図表7　言葉によるマッピング分析例２　「コレスポンデンスマップ」



ネットワーク分析ツール
「+PLANET」(株式会社プラスアルファ・コンサルティング）による、
独立行政法人科学技術振興機構（JST）の「光触媒」に関する特許・論文の分析結果

図表8　発明者・企業間ネットワーク

融合したものである。共同発明を特徴付けるキーワードが接続線の上に示されており、時間の経過とともに、人が繋がり、技術開発への取り組み、融合が進んでいった様子を可視化している。特許が中心のグループは、産業・事業に近い取り組み、論文が中心のグループは基礎研究に近い取り組みと見ることができる。

製品が複雑化し、多様な技術の融合が進む中で、自社の人材・技術をどう活かして、次のイノベーションを生み出すか、あるいは、同業他社の取り組みを把握する上で有効な手法と言える。

(2) グルーピングによる詳細分析

テキストマイニングでは、文中に出現した言葉の頻度順リストから、人手で言葉を選択することで、文献群を意味的に仕分けするルールを簡単かつ精緻に設定し、意味的なグループごとの出願の傾向・動向を定量的に分析することができる。

また、言葉の選択、仕分けルールを人手で設定する方法のほかに、すでに分類コードが付与されている文献データを「教師データ」として、未分類の文献群に対して簡易的に分類コードを付与したり、母集団全体の言葉の出現傾向から、完全自動で分類を行う方法がある。以下では、それぞれの方法について、詳細と特性を説明する。

①マニュアル分類

文中に出現した言葉の頻度順リストから、分析者が言葉を選択し、それらの「and条件」「or条件」「not条件」やその組み合わせにより、文献群を意味的に仕分けするものである。その際、例えば、「画面が大きい」というように、「画面」「大きい」のいずれか1単語では意味が特定されず、「画面ー大きい」という「係り受け」になって初めて意味が特定される場合には、「係り受け」を仕分けの条件として活用することで、より精緻な分類が可能となる。

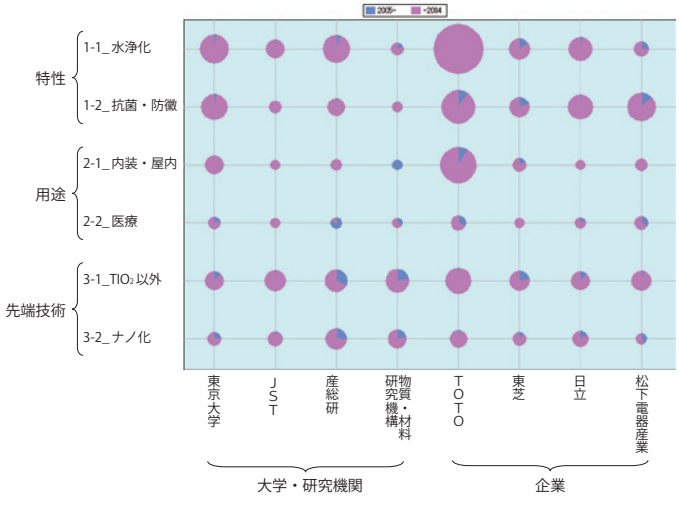
図表9では、光触媒について、特性、用途、先端技術という観点から、分類・分析を行っており、企業間の取り組みの違いを詳細に分析している。

このように、分析者の問題意識・仮説に基づき、さまざまな観点（課題・要素技術・機能・用途・製造方法など）から文献群を仕分けすることで、詳細なポートフォリオ分析が可能となる。これまで、公報に付与

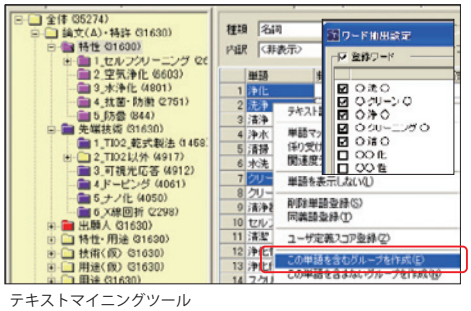
グルーピング条件の設定（例）

グループ名		単語・係り受け（○-○）
1. 特性	1-1_水浄化	廃水、懸濁液、水質、排水、浄水、COD、液（部分一致）-浄（部分一致）・・
	1-2_抗菌	菌、ウィルス、バクテリア、微生物、ウイルス、カビ・・
2. 用途	2-1_内装・屋内	タイル、風呂、浴室、浴槽、キッチン・・
	2-2_医療	手術、病院、医療、病気、腫瘍・・
3. 先端技術	3-1_TIO_2以外の金属	亜鉛、ZNO、タングステン、WO、酸化ニオブ、タンタル・・
	3-2_ナノ化	ナノ粒子、ナノファイバ、ナノチューブ、超微粒子・・

グループ×出願人×出願年によるクロス分析



グルーピング条件の設定画面（例）



テキストマイニングツール
「TRUE TELLERパテントポートフォリオ」
（株式会社野村総合研究所（NRI）グループ）による、
「光触媒」関連の分析例

図表9　グルーピング分析

されたIPC（International Patent Classification）や、審査のために付与されたFI（File Index）などの特許分類によるマクロ分析が、企業の意思決定の材料として馴染みにくいとの指摘があったことは、分析者の抱える問題意識や仮説を、分類の視点・切り口として反映しにくいことがひとつの原因と言える。

文献群の記載内容を読み込み、仕分けして分析することは、企業によっては、これまでも「めくり」という言葉に表されるように、多くの人手と時間を掛けて行われてきた。しかし、文献数に比例した時間、労力、コストがかかるため、現実には、分析可能な件数に限界があった。

もうひとつ、こうした人手による仕分けの問題として、先行技術調査の際など、分析担当者の当該技術に対する理解、知識が十分でない中でも、予め、仕分けのルール＝「箱」を決めなければならないといったことが挙げられる。例えば、一旦「箱」を決め、途中まで文献群を読み進めたところで、新たな「箱」の必要性、あるいは、読み飛ばしてきた言葉の重要性に気付けば、再度「箱」の体系を見直し、そうした視点を持って、1件目から改めて目を通していかなければならなかった。

テキストマイニングによる仕分けでは、一旦、母集団データを前処理し、頻出単語を頻度順に見ながら、仕分けの「箱」を決め、同時に仕分けのルールを決めていくことに特徴がある。また、一つの出願が、複数の機能、課題、技術要素から構成されている場合もあり、重複を厭わずに仕分けのルールを決めていくことができる。さらに、最終的にどこの「箱」にも入らなかった文献群についても、再度、その中の頻出単語を確認することで、すでに作成した「箱」に言葉を加えたり、新たな「箱」を作るといったことが簡単にできるのも、テキストマイニングならではの手法と言える。

またこのようにして、一旦、作成した仕分けのルールは、企業にとって有形のノウハウ、蓄積になっていく。例えば、研究開発を次のステップに進めるか否かの意思決定を行う際には、研究開発の開始時に行った分析に、その後の新着情報を追加で取り込むことで、自動的に仕分けを行い、差分を確認することができる。また、新着情報に初めて出現した単語を自動抽出し、仕分けの条件を更新していけば、ノウハウ・蓄積そのものが成長し、継続的に技術動向をウォッチしていくための「仕組み」「インフラ」にすることができる。

なお、マニュアルでの言葉の発見、仕分けを支援するために、例えば、「特許請求の範囲」では、「・・を特徴とする○○」といった表現に続いて、発明の対象を表す言葉がきたり、「○○用」「○○に用いられる」といった表現から、発明の用途を表す言葉がくることが多く、こうした表現を手掛かりとして、言葉の抽出することができる。

②「教師データ」に基づく「半自動」分類

企業によっては、過去の自社の出願、保有特許について、自社の事業や組織、プロジェクト等の切り口で、独自の社内分類コードを付与し、データの検索や管理、分析・評価のツールとして活用しているケースが多い。このように、あらかじめ、分類が付与された自社の出願、保有特許を「教師データ」として、他社の出願、特許に対して、類似する自社の分類コードを機械的に付与することができる。

具体的には、（ア）各コードが付与された自社の出願、特許データから、各コードに偏って出現する「特徴語」と、特徴度を表す指数（スコア）を自動算出し、これをもとに、（イ）他社の特許1件1件についてもっとも近いコード、あるいは、近い順に上位n位までのコードを付与するといったプロセスになる。これは、各コードに対して、いわゆる「概念検索」「類似検索」を一括して行い、仕分けを行うといったことである。

なお、ソフトウェアが自動抽出した特徴語に対して、ノイズの原因になると思われる一般語や技術的に意味を持たない言葉（例：課題、装置など）を、人手を介して除外するプロセスを挟むと、仕分けの精度が向上することから、筆者はこれを「半自動分類」と呼んでいる。

③完全自動分類「クラスタリング」

テキストマイニングでは、単語どうしが同時に出現する傾向や、シソーラスを材料にして、例えば、「全体を5つのグループに分類する」と指定するだけで、文献群を完全自動で仕分けする「クラスタリング」と呼ばれる手法がある。極めて短時間で手軽に仕分けができる反面、ロジック・プロセスがブラックボックス化し、明解な説明・検証が難しくなるため、企業における意思決定の材料としては、活用場面が限定される。マーケティングの分野では、例えば、分析担当者レベルで、

図表10　テキストマイニングによるグルーピング手法の比較

	グルーピングの方法	グルーピングロジックの明確さ・分かりやすさ	手間・工数
1. 単語・係り受けの選択によるマニュアル分類	・頻出単語・係り受けから、ユーザの視点で、意味的に仕分け。	◎ ・選択した単語・係り受けの有無という、明確で分かりやすい条件でグループ化。	△ ・単語・係り受けの選択に、一定の工数と当該技術に対する知識・知見が不可欠。
2. 既存分類に基づく自動分類 (Ex.) ・A社特許群と似た他社特許の抽出 ・自社分類に基づく他社特許の分類	・既存分類ごとに、「特徴語」と、特徴度を表す「スコア」を算出。 ・「特徴単語」と「スコア」に基づき、類似文献を自動分類。	△ ・グルーピング条件の確認・調整は 可能。ただし、複雑。 ・たまたま既存分類に、特徴的な表現があると、影響を受ける可能性。	○ ・既存分類に基づき、類似文献の自動抽出～グルーピングが可能。 ・分類精度を確保するためには、人手による閾値等の調整が必須。
3. 自動分類＝「クラスタリング」	・単語の出現傾向等から、類似したテキストを自動的にグループ化。 ・自動作成されたグループ毎の特徴語に基づき、グループ名を自動付与。	× ・グルーピング条件の確認・調整は 不可　＝「ブラックボックス化」 ・1文献を1グループに振り分け。	◎ ・ソフトが自動的にグルーピング。

「1.単語・係り受けの選択によるマニュアル分類」の特長	従来の手法・他の手法との比較
1. 想定外の表現を含めて、言葉を網羅的に拾える。	・いわゆる「キーワード検索」では、言葉が思い浮かばないと、関連文献を拾えない。
2. 頻出順に言葉を見ていくことで、グループを体系的、かつ、網羅的に構築できる。	・「めくり」「目視」による分類では、予め、手探りで、分類枠を想定しなければならない。 ・途中で分類枠を変えれば、大幅な、手戻りが発生する。
3. グルーピングのロジックが明確で説明・検証できる。	・自動分類型（ブラックボックス型）のツールでは、「○○ツールによる分析結果は……」としか言えない。
4. 分析者の観点（課題／技術／製造方法……）、仮説に応じて、グループを設定できる。	・課題、技術といったさまざまな観点から、「めくり」「目視」により、特許を分類するには、高いスキルと経験が求められる。 ・自動分類型のツールでは、ユーザの求める観点・仮説が 入り込む余地がない。
5. 作成したグルーピング条件は、企業としての、形のある蓄積・ノウハウになっていく。（新着特許を追加インポートすれば、作成した　グループに、自動的に振り分けられる）	・「めくり」による分類では、いつまでも、ノウハウが属人化。 ・新着特許の分類にも、いつまでも、件数比例の手間・工数 がかかる。

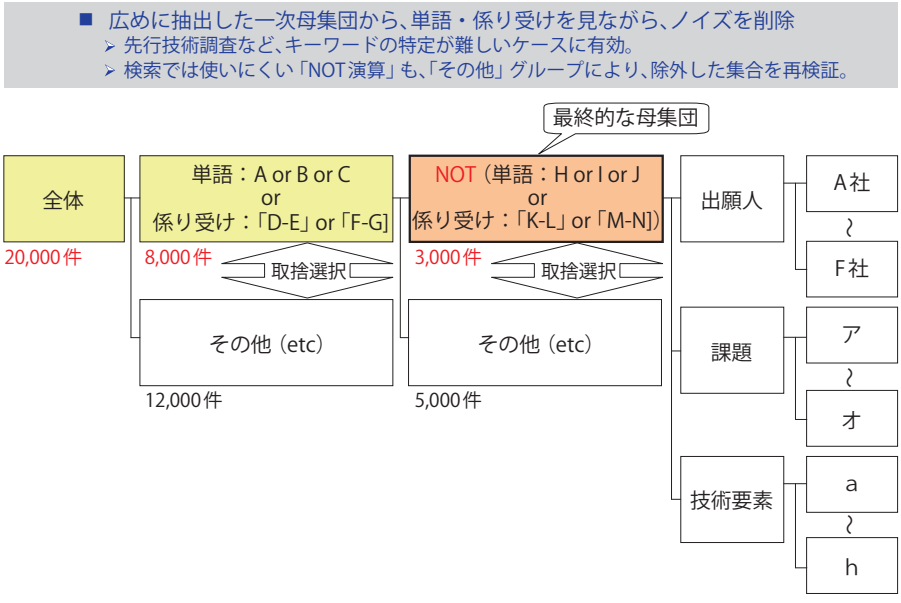
まず全体の概要を把握したい場合などに使われることがある。

以上のように、テキストマイニングによる仕分けの手法は、「分析に掛けられる時間、労力」と「分析の精度、ロジック・プロセスの明確さ・分かりやすさ」がトレードオフの関係にあり、分析の目的、場面に応じた使い分けが求められる（図表10）。誤解を恐れずに言えば、テキストマイニングによる知財ポートフォリオ分析は、一つボタンを押せば、すぐに有用な分析結果と示唆が得られるといったものではない。言葉をもとに分析を行うことから、分析の各プロセスにおいて、分析担当者の知識・知見が反映されることで、初めて、真に役立つ分析が可能となる。

6）母集団の精査　ノイズの除去・母集団の追加

テキストマイニングによる分析では、分析のプロセスを通じて、発見した言葉をもとに、想定外のノイズを見つけ、除外するための条件を簡単に設定することができる。また逆に、気付いた関連用語を、そもそもの母集団の抽出条件にフィードバックし、追加データとして取り込むといったプロセスを通じて、母集団を精査していくことができる（図表11）。

これまで、例えば、企業にとって新規分野への参入可能性を調査する場合など、当該技術に対する予備知識が十分でない中で、手探りで検索式を立てる必要があり、極めて属人的な経験・ノウハウ頼みの部分があった。さらに、文献群を抽出できても、人が目を通せる件数には限界があること、「めくり」による分類作業を外部の調査会社に委託する場合には、件数比例のコストがかかることから、「1,000件以内に絞り込むために、検索条件をどう設定するか」といった本末転倒とも言える思考を行わざるを得なかった。テキストマイニングによる分析では、このような件数による拘束が小さ



図表11　テキストマイニングを活用したノイズの除去

く、最初は、少し広めに母集団を設定し、分析のプロセスを通じて、当該技術・用語に対する理解を深めながら、徐々に母集団を精査していくことができる。

さらに今後、検索ツール・システムとテキストマイニング技術の融合が進めば、こうした分析対象データの絞り込み、追加といった処理の簡素化、自動化が可能になっていくだろう。

7）辞書のメンテナンス

テキストマイニングによる分析では、言葉のデータベース（「辞書データ」）が分析のベースとなる。しかしながら、極めて専門性の高いテクニカルタームや、新出の技術用語をカバーしていくことには限界があり、これを補う形で、ユーザ固有の辞書を簡単に追加・メンテナンスできることが求められる。辞書の追加・メンテナンスを支援する機能として、例えば、「母集団の文中に何回も出てきた、『辞書データ』にはない文字列」を言葉の候補として自動抽出し、これを分析者の判断で登録することにより、ユーザ独自の辞書を育てていくことができる。これによって、精度が高く、企業の意思決定に資する分析が可能となる。

なお、こうした技術用語の辞書データについては、テキストマイニングを支える「インフラ」として、公的機関等による整備・提供・定常的なメンテナンスが期待されるところである。

8）分析結果からの元データ確認 〜マクロ⇔ミクロの分析〜

知財ポートフォリオの分析では、マクロな分析結果から、逆に、ミクロに個別文献を確認・検証しながら、分析を深堀りしていくプロセスが不可欠である。分析ツールには、マクロ⇔ミクロな分析・確認を円滑に行うことができる機能性・操作性が求められる。さらに、検索システム・サービスの公報データにリンクすることで、関連図面や最新の審査経過情報にアクセスできると望ましい。

3. テキストマイニング分析の「精度」

筆者は、テキストマイニング分析の精度について問合せを受けることが多いが、「自然言語処理」「データマイニング」「可視化」の各プロセスを区別して議論する必要がある。

まず、自然言語処理については、現在、検索システムやツールで用いられる技術と基本的に同様のものであり、エンジンによって特性の違いはあるものの、大きな差異はない。実際に、数百件の公報群を対象に、国内の主要な特許検索サービスでキーワード検索を行った場合

と、テキストマイニング専用ソフトウェアを使って言葉の抽出を行った場合で結果を比較したところ、100％に近い精度で同様の集計結果が得られている。ただし、このわずかな差異が重要な場合もあり、テキストマイニングでは、文章からの単語の切り出し＝「形態素解析」のベースとなる辞書データの精度、網羅性が重要となる。

次に、データマイニングについては、多くの解析手法が数学的に裏付けされた統計手法に基づくものであり、こちらも、閾値・パラメータの設定によるところはあるものの、分析結果が大きくぶれることは少ない。むしろ、前述のグルーピング分析などでは、分析者の求める観点、問題意識、仮説が明確であること、当該技術に対する一定の知見・知識といったバックグラウンドが重要となる。また、分析の精度・粒度と分析に掛けられる時間はトレードオフの関係にあり、マクロな概況把握を行いたいのか、仕分けした結果から、最後は個別の公報まで目を通したいのかといった目的によって決まるものである。

最後に、可視化技術については、さまざまな手法が提案されており、それぞれ、特性を理解して使い分ける必要がある。例えば、言葉をベースにしたマッピング手法については、当該技術に対する知見・理解がまだ十分でない状態において、極めて短時間で母集団の概略を把握したり、代表的な文献群を抽出することに優れている。一方、分析の細かさには限界があり、また、分析者の求める観点、仮説、問題意識が入り込む余地は少ない。一方、グルーピング結果のグラフ化は、極めて詳細に、定量分析ができる反面、前段のグルーピングの条件設定によっては、大きなノイズや漏れが発生する可能性がある。特に、定量分析の結果は、数字が一人歩きするリスクも伴い、慎重な条件設定・確認が求められる。

なお、テキストマイニングによる分析ツールでは、分析の「精度」を上げるために、以下に挙げるような補助的な機能が提供されている。

・選択した単語に対する、「前方一致」「部分一致」によるグルーピング条件の設定（例：「量産（部分）」により、「量産」「量産化」「量産性」を同時に選択）
・選択した単語に対する、同義語候補の自動抽出（広義語、狭義語、一部文字列一致など）
・長音、ハイフンの自動抽出・修正支援
・人名に含まれるスペース等の統一

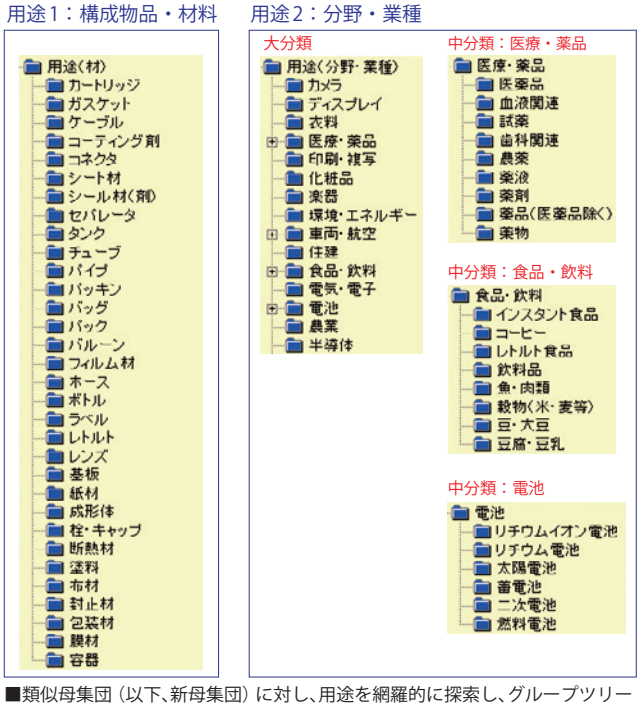
4. 活用目的・場面の拡がり

テキストマイニングによる知財ポートフォリオ分析は、さまざまな目的、場面に拡がっている。以下に、最近の活用例をいくつか紹介する。

1）自社技術の新たな用途、提供・提携先の発掘

自社が保有する要素技術（シーズ）に対して、特許情報をヒントに、これまで想定していなかった新たな用途（ニーズ）や、新たな提携・提供先の候補を発掘するといったアプローチが行われている。たとえば、自社が保有する素材の特性（例：「耐熱性」など）を検索条件として、技術分野の異なる特許まで含めて幅広く抽出し、「実施例」「産業上の利用可能性」などのテキストを解析することで、従来とは異なる分野への適用可能性を探ったり、異業種の企業による出願を発掘し、新たな提携・提供先のヒントを得るといったものである。

多くの企業が、次の新事業の創出に苦慮する中で、現在、保有する要素技術の強みをもとに、新たな事業機会を模索するものである（図表12）。



■類似母集団（以下、新母集団）に対し、用途を網羅的に探索し、グループツリーを作成

図表12　用途別のグルーピング

2）M&Aの戦略策定・デューディリジェンス

製造業界におけるM&Aにおいて、技術的な融合・シナジーがどれだけ期待できるかは、極めて重要な意思決定の要素である。しかしながら、対象企業の財務状況、有形資産については検討の初期段階から、詳細に調査（デューディリジェンス）するのに対して、技術を客観的に評価し、経営者を含めた関係者の共通認識を形成することはなかなか難しい。当然、ある程度、検討が進んだ段階では、詳細な評価を行うことになるが、検討の初期段階では、担当者の主観的な評価に委ねられることが少なくない。そこで、自社と対象企業の知財ポートフォリオを、テキストマイニングにより詳細に比較・分析することで、M&Aによる技術的な補完・囲い込みの可能性を評価、シミュレーションするといったことが行われている。

3）研究機関の成果評価・研究開発のターゲット選定

公的な研究機関における研究開発や、いわゆる「ナショナルプロジェクト」としての研究開発など、国としての特定の技術分野への資金投入の成果について、数年後にどれだけの知的財産を生み出し、波及効果を与えているかといった観点から特許データや論文データの解析を行うものである。

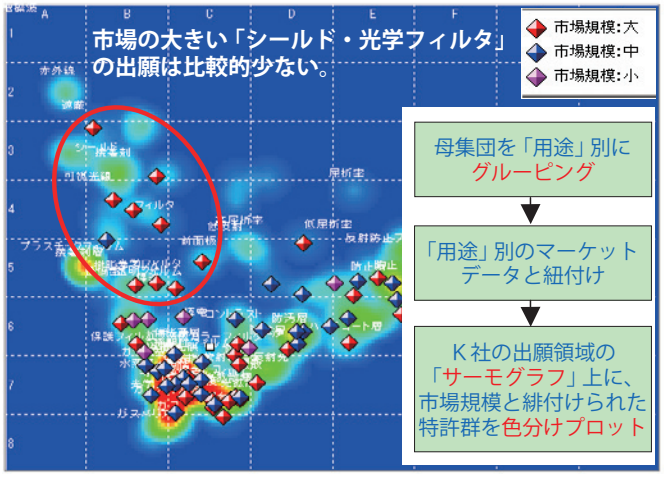
さらに、国内外の特許・論文データを幅広く分析することで、他国に対して優位性の高い技術領域を客観的に見極め、今後、国として予算投入すべき技術開発や、戦略的に標準化を狙う技術のターゲットを決めるといったアプローチも行われ始めている。

4）自社の事業・製品と知的財産の紐付け～財務情報・マーケット情報との統合分析

主に製造業において、企業が事業に関する意思決定をする際には、市場の規模や将来予測、現在のシェアと競合状況、社内リソース（人材、設備投資）の投入状況といった関連情報に基づく総合的な判断が求められ、知財ポートフォリオの分析は、そのひとつに過ぎない。本来、知的財産は、事業や製品・サービスと結びついて初めて価値を生み出すものである。

すなわち、自社・他社の特許群を、自社の事業・製品と紐付けることができれば、事業に関わるさまざまな情報と知財情報を重ね合わせた分析が可能となり、より高度な、事業戦略、研究開発戦略の判断材料となる。例えば、人や設備投資といったリソースの投入状況と重ね合わせて見れば、自社の事業戦略と、研究開発戦略・知財戦略とのギャップを評価することができる。通常、自社の出願・特許については、研究開発の初期段階から、事業・製品と紐づけた管理が行われている場合も多い

特許ポートフォリオと市場規模のギャップ分析



※1　公報の文中に含まれる単語・係り受けから、特許をフィルムの用途に分類。用途別の市場規模データ（※2）とマッチングした。
※2　「2006液晶関連市場の現状と将来展望」（株式会社富士キメラ総研）

図表13　マーケット情報と特許情報の統合分析

が、他社の出願・特許を紐付けることは容易ではない。そこで、テキストマイニングにより、言葉をもとに、他社の特許群を、自社の事業・製品の切り口で仕分けすることで、効率的に、自社・他社の強み・弱みを関連情報と統合して分析することが可能となる（図表13）。

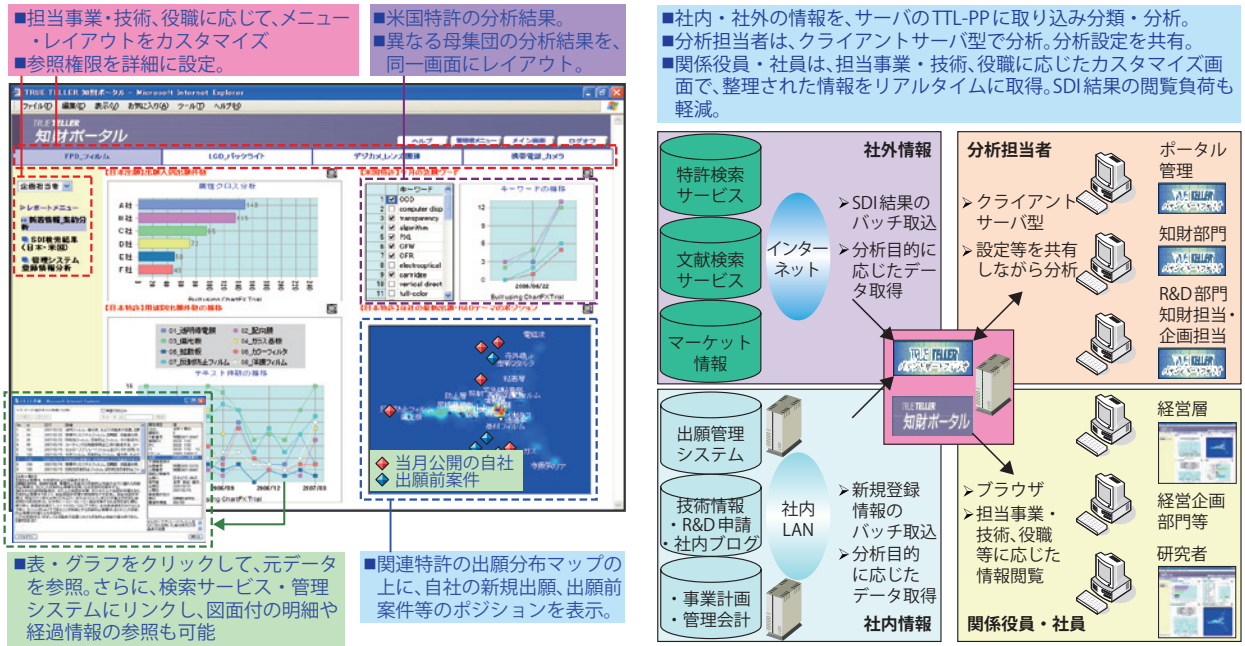
5）新出／急増技術用語のアラート配信、技術マーケティング

メーカーのマーケティング部門において、毎週、大量に発行される特許・論文データから、テキストマイニングにより、新出の技術用語、急増している技術用語をいち早く自動抽出することで、他社の動向を察知し、自社にとっての事業のヒントを見い出そうとする取り組みが行われている。例えば、BtoBで素材や部品を提供する企業においては、セットメーカの出願情報は、新しいマーケット・潜在顧客を察知するための有効な情報ソースである。さらに、特許情報から、顧客企業のキーマンを探索し、効果的な営業に繋げるといったことも行われている。

6. 分析ツールから業務インフラへ「知財ポータル」

企業の知財担当者や、研究開発の企画担当者は、それぞれの問題意識、仮説に基づき、自ら手を動かして、必要なポートフォリオ分析を行う立場にあるが、多くの技術者・研究者は、担当する事業、業務、役職に応じて、必要な情報を、整理された形で分かりやすく提供してほしいというニーズを持つ。例えば、毎週発行される国内・海外の公報に関するSDI（Selected Dissemination Information）検索の結果を確認することは、技術者・研究者にとって必要な業務ではあるものの、未加工の情報を、未整理の状態で読み込んでいくことの負担感は決して小さくない。

テキストマイニングにより、担当者が求める切り口に沿って、あらかじめ仕分けされ、可視化された結果を、担当する事業、業務、役職に応じてカスタマイズされた画面によって、社内のイントラネットで配信することができる。これにより、担当者は、それぞれの自分の関心の高い切り口の情報から確認していくことが可



「TRUE TELLER知財ポータル」（株式会社野村総合研究所（NRI）グループ）の画面より

図表14 「知財ポータル」画面イメージと概念図

能となる。情報ソースについても、国内、海外の新着特許公報や、新着論文情報、社内で管理されている公開前の出願情報や技術系の社内ブログなど、技術関連のテキスト情報を言葉をベースに整理、統合して、必要な人に、必要な形で配信するものである（図表14）。

なお、すでに、マーケティングの分野では、「顧客の声」を社内のポータルサイトを通じて、発信するといったことが多くの企業で行われ始めている。「顧客の声」の場合には、情報の速報性が重要となるが、技術情報の場合には、①最新の情報を、過去の大量の情報との関係の中で、整理して配信すること、②最新の情報から、過去の関連情報を簡単に辿れることなどが求められる。

7. 審査関連業務への適用の可能性

本章では、ここまで述べてきたテキストマイニング技術について、特許庁をはじめとする関連機関における、審査関連業務への適用の可能性について考察する。

1）審査関連業務へのテキストマイニングの適用案

（1）審査業務の高度化・効率化

テキストマイニングの適用は、審査業務のプロセスの効率化に寄与するものと考える。

①担当審査官への振り分け業務の効率化

新たな出願群に対して、テキストマイニング処理を行い、予め作成したグルーピングのロジックに従って、一次的に分類を付与したり、関連しそうな分類に対して類似性・関連性を指数化して示すことで、これを担当審査官への振り分けの参考情報として活用することが考えられる。

②審査官向け参考情報の提供

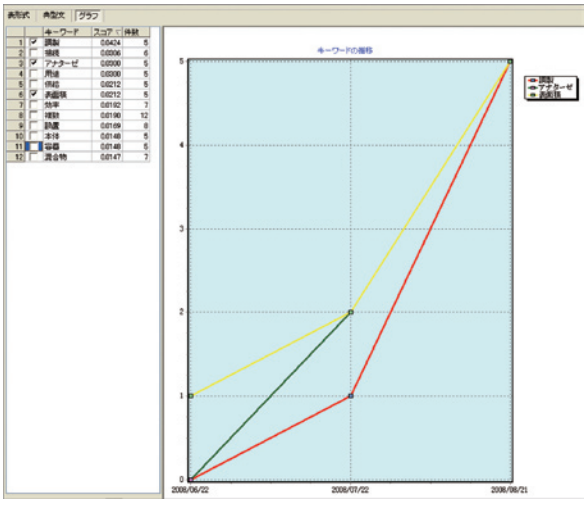
審査対象の出願について、①で付与した特許分類「候補」や、関連性の高そうな過去の代表的な出願群を参考情報として提供したり、さらに関連特許群の文中に出現するキーワードリストやマッピング結果から、関連特許群を絞り込んだり、辿ったりすることが簡単にできる仕組みが提供されれば、審査業務の効率化に寄与するものと考える。

また、ひとつの文献に同時に付与されることの多い、分野が異なるFI、Fタームの繋がりや、その文献群に出現するキーワードの関係を抽出しておけば、さらなる業務の効率化や、考慮漏れの軽減に繋がる。ひいては、審査の質の向上に資するものと考える。

（2）新出／急増技術用語のアラート配信

審査官に対して、担当する技術分野における新出用語や急増用語のアラートを定常的に配信することで、審査官の関連技術の最新動向に対するアンテナを高め、関連知識の効率的、効果的な習得を支援することができる（図表15）。

公開日：直近3ヶ月で比較したときの急増ワード

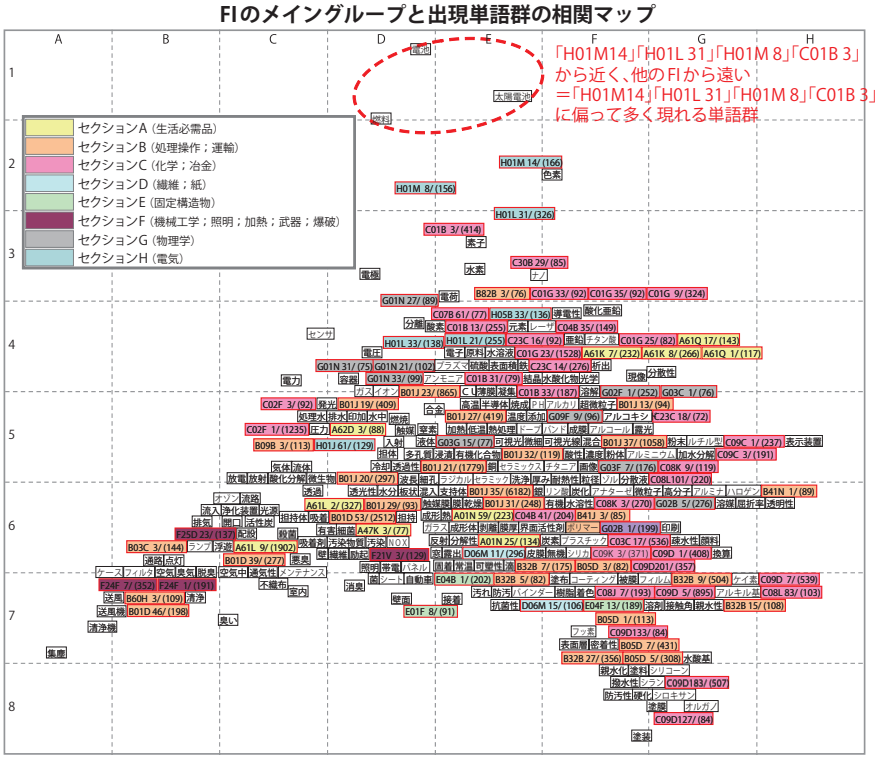


テキストマイニングツール「TRUE TELLERパテントポートフォリオ」（株式会社野村総合研究所（NRI)グループ）の画面より

図表15　急増ワードのアラート分析

（3）特許分類コードのメンテナンス・最適化

技術の細分化、融合が加速し、同時に、生み出された技術の短命化が進む中で、審査業務を支える特許分類コードについて、効率的にメンテナンスを行っていく仕組みが求められている。例えば、同じ出願に付与されることの多い特許分類コード（FI、Fターム）の関係を解析し、可視化することで、特許分類コードのメンテナンス、最適化に係る業務の効率化を図ることが考えられる。特定のFI、Fタームが急増している場合には、そこに出現する言葉に基づき、仕分けを行うことで、どのように



図表16　特許分類コードの相関分析

分類を細分化すれば、件数の均等化が図れるかといったシミュレーションを行うこともできる（図表16）。

2）審査関連業務への適用に向けた課題

以上に挙げたような、審査関連業務へテキストマイニングの適用に向けては、以下の課題が挙げられる。

（1）辞書データのメンテナンス・関連システム間での共有

テキストマイニングを支える、重要な要素のひとつが、辞書データベースである。科学技術用語については、関連機関における整備・更新が行われており、こうしたデータを統合し、システム間で共有・共用できることが求められる。

また、こうした辞書データの整備、メンテナンスのプロセスそのものにテキストマイニング技術を組み込むことで、新出用語、急増用語の自動抽出～人手による確認～辞書データへの登録といったプロセスが考えられる。さらに、審査官が審査業務を通じて、発見した重要な言葉が蓄積され、人手による確認を経て、辞書データに反映されるプロセスが組み込まれれば、より理想的だろう。

（2）FI、Fターム　定義文の更新／関連キーワード（単語、係り受け）のデータベース化

前述の特許分類「候補」の自動付与を実現するためには、テキストマイニングを活用した、特許分類ごとの仕分け条件の整備が必要となる。まずは、特許分類の定義文や解説書をテキストマイニング処理し、抽出されたキーワードから、一次的に仕分け条件を作成することが考えられる。さらに、FI、Fタームが付与された過去の出願、あるいは、被引用数の多い出願を「教師データ」として、キーワードを抽出すれば、同義語や、係り受けによる表現を含めた、より網羅性の高い仕分け条件が整備できると考えられる。

上に述べた、出願案件の担当審査官への振り分け、「観点」候補の自動付与を、テキストマイニングを用いて実現するためには、FI、Fタームごとの振り分けのロジックが必要となる。まずは、FI、Fタームの定義文、解説書からキーワードを抽出することが考えられるが、同義語、類義語への対応が不十分となる。そこで、すで

にFI、Fタームが付与されている代表的な特許群を「教師データ」として、各FI、Fタームに偏在するキーワードを自動抽出し、これを振り分けのロジックに反映させていくことが考えられる。

自動抽出された単語、係り受けに対して、目視による取捨選択、「and条件」「or条件」「not条件」の設定といったプロセスは不可避であり、一定の手間・工数が掛かることが想定されるが、一旦、構築された仕組みは、審査業務を支える強力なツールとなり、その後も精度は高まっていくだろう。出願が増えている技術分野から優先的に整備を進めるなど、全体効率を見据えた推進スキームが求められるだろう。

(3)「仕組み」を運用する体制の構築

(1)(2)で構築される辞書データや、システム上のロジックは、テキストマイニング技術を核としながら、技術用語、技術分類の体系に関する形式知(ナレッジ)を管理・共有する「仕組み」として捉えることができ、これをしっかり運用し、定常的にメンテナンスしていくための体制が求められる。

これに類似した例としては、企業において、営業情報や、技術に関する個人レベルの知見を共通の「知」として管理、共有する「ナレッジマネジメント」の仕組みが挙げられる。そこでは、単に、社員一人ひとりが入力したデータをそのまま残し、蓄積するのではなく「ナレッジ・エディタ」と呼ばれる役割の担当者が、インデックスの付与や、言葉の統一など、常にメンテナンスを行う「仕組み」を組み込むことで、生きた情報として活用できるようになると言われている。

審査関連業務を支えるインフラにおいても、これと同様に、テキストマイニングで抽出された言葉に対して、人手を介して加工を行う「仕組み」「体制」を組み込むことで、その有効性や精度を継続的に高めていくことができるだろう。

(4) 高い分析精度と情報処理速度の両立

コンピュータの情報処理能力、ネットワークの伝送能力の飛躍的な向上をひとつの契機として、分析系のテキストマイニングは、パソコンレベルでの活用が一気に広がってきた。しかしながら、審査業務を支えるインフラとしてテキストマイニングを活用するためには、大量の特許データに対して、高い分析精度を保ちながら、より高いレベルのレスポンスが求められることになるだろう。「ドッグイヤー」でのコンピュータの情報処理能力の進歩も期待されるが、分析精度を犠牲にせずに、情報処理速度を上げるための、ソフトウェア側の能力の向上が求められる。

さらに、審査官の業務を支えるインフラとして、シンプルで分かりやすいユーザインターフェイスも必要になるだろう。

(5) 審査官の経験・暗黙知と、システム化・データ化された形式知の融合

テキストマイニングは、言葉と属性項目の関係や、言葉そのものへの気付きを与えるものであるが、決して、完全自動でアウトプットを提示するものではない。テキストマイニングを審査業務に活かすためには、言葉を理解するための当該技術に対する知識・知見はもちろん、抽出された言葉群から重要な言葉に気づくための、審査官の経験・暗黙知が、いっそう重要となるだろう。

8. おわりに

わずか10年ほどのあいだに、インターネットによる特許検索が急速に普及・定着していったように、近い将来、テキストマイニングによる技術情報の調査・分析もまた、当たり前のものになっていくと確信している。企業の事業戦略、研究開発戦略、知財戦略を支える分析ツールとしてだけでなく、特許庁をはじめとする関連機関において、審査関連業務を支えるインフラとして、テキストマイニングが有効に活用されることで、業務がさらに効率化され、ひいては、知財立国の実現が加速されることを願っている。

profile

中居　隆(なかい　たかし)

1994年　東京大学工学部船舶海洋工学科修士課程修了。同年　株式会社野村総合研究所(NRI)入社。地理情報システム(GIS)、ナレッジマネジメントシステムの企画・開発、関連コンサルティング業務等に従事。2004年から、NRIグループのNRIサイバーパテント株式会社で、テキストマイニング技術を活かした特許ポートフォリオ分析ソフト「TRUE TELLERパテントポートフォリオ」の企画・設計・販売・サービス運営、関連調査・コンサルティングを担当。