

テキストマイニングの活用事例

～新機能を用いた分析手法の紹介～

柴田 洋輔 (NRI サイバーパテント株式会社 IP コンサルティング室)

要旨：

NRI サイバーパテントは 2004 年より提供してきた知的財産分析サービス「TRUE TELLER パテントポートフォリオ (以下、TTL)」の後続サービスとして「CyberPatent Desk テキストマイニング (以下、CPD-TM)」を 2018 年 10 月 1 日にリリースした。本論では、その新サービスである CPD-TM の特徴に触れつつ、バージョンアップにより新たに実装された分析手法の活用方法について、具体例を挙げて紹介する。

はじめに ～テキストマイニングとは～

事例分析に入る前の基礎知識として、まず CPD-TM の概要及び特徴について記載する。CPD-TM は言語解析を行うためのサービスであり、特許データや論文データなどのテキスト情報を「テキストマイニング」という手法により分析する。

ここで、テキストマイニングとは、文章を単語や文節の単位で区切り、それらの出現傾向や頻度などの情報から、特徴的なキーワードは何であるか、何に主眼が置かれている文章なのか、その傾向を予測するシステム手法になる。

CPD-TM の特徴

従来の TTL で実装されていた代表的な分析手法である統計分析、グルーピング分析、マッピング分析は残しつつ、CPD-TM では新たにホットスポット分析、特徴マップ分析等が追加されているため、本論はこれら新たな分析手法について主に記載する。従来からある分析手法についての説明は割愛をするが、グルーピング分析やマッピング分析による事例紹介は、前回の「NRI-CP 通信 2017 Vol.1」に記載がある

ので、そちらも参照いただきたい。

CPD-TM は、従来の TTL と同様に NRI サイバーパテントが提供する知的財産情報サービス「CyberPatent Desk」(以下、CPD) と連係をしており、CPD の検索結果をダイレクトに CPD-TM にインポートできる。現時点では、引き続き日本と米国のみがインポート対象国であるが、今後は対象国を拡大していく予定である。

また、分析対象は CPD の特許検索の結果に限らず、国内外の論文データや新聞等の外部データを取り込んで分析することも可能であり、網羅的なデータの統合分析を 1 つのプロジェクトで処理できる点に特徴がある。

なお、CPD-TM で分析することができる言語は日本語と英語であるが、まもなく中国語の解析も可能になる。

本論では、CPD-TM の特徴に触れながら、具体的に光学計測機器をモデルとしてホットスポット分析、特徴マップ分析を用いた分析を行い、どのような知見が今回の CPD-TM から新たに得られるようになったかについて紹介していく。

母集団、グルーピング軸の定義

本論では、公開公報と登録公報を対象に国際特許分類 IPC のみの条件(G01B 11/)で「光学計測機器」に関する母集団を定義する。

まず、ホットスポット分析の前段階として、各グループを定義するための軸を作成するが、縦軸の項目を課題の観点で「小型化」、「感光性」、「自動化」、「光強度」、「測定精度」と定義し、横軸の項目を要素技術の観点で、「入射」、「反射」、「屈折」、「波長」、「光軸」として定義し、図1のような特許マップを作成する。具体的には CPD-TM のクロス分析の結果をマトリクス形式で描画した。図1の各交点の円の中にある数字は、縦軸と横軸の両方の条件を同時に満たしている対象文献の件数になる。

ホットスポット分析

ホットスポットとは、グループ内において、対象とする属性値の出現数がどの程度偏っているのかをスコアリングする手法である。

ポイントは、あくまで件数の大小には関係がなく、そのグループに偏って出現しているほどスコアが高く計算されていることであり、閾値を超える項目は赤色、閾値は超えないがプラスの項目は黄色でハイライトされる。なお、閾値は変更することができるため、どれだけの偏りがある場合にハイライトをするのかはユーザー自身で設定できる。特定のグループに偏って出現しているほど高いスコアなるため、グラフ中で注目すべき「特徴」を見つけ出すのに有用である。

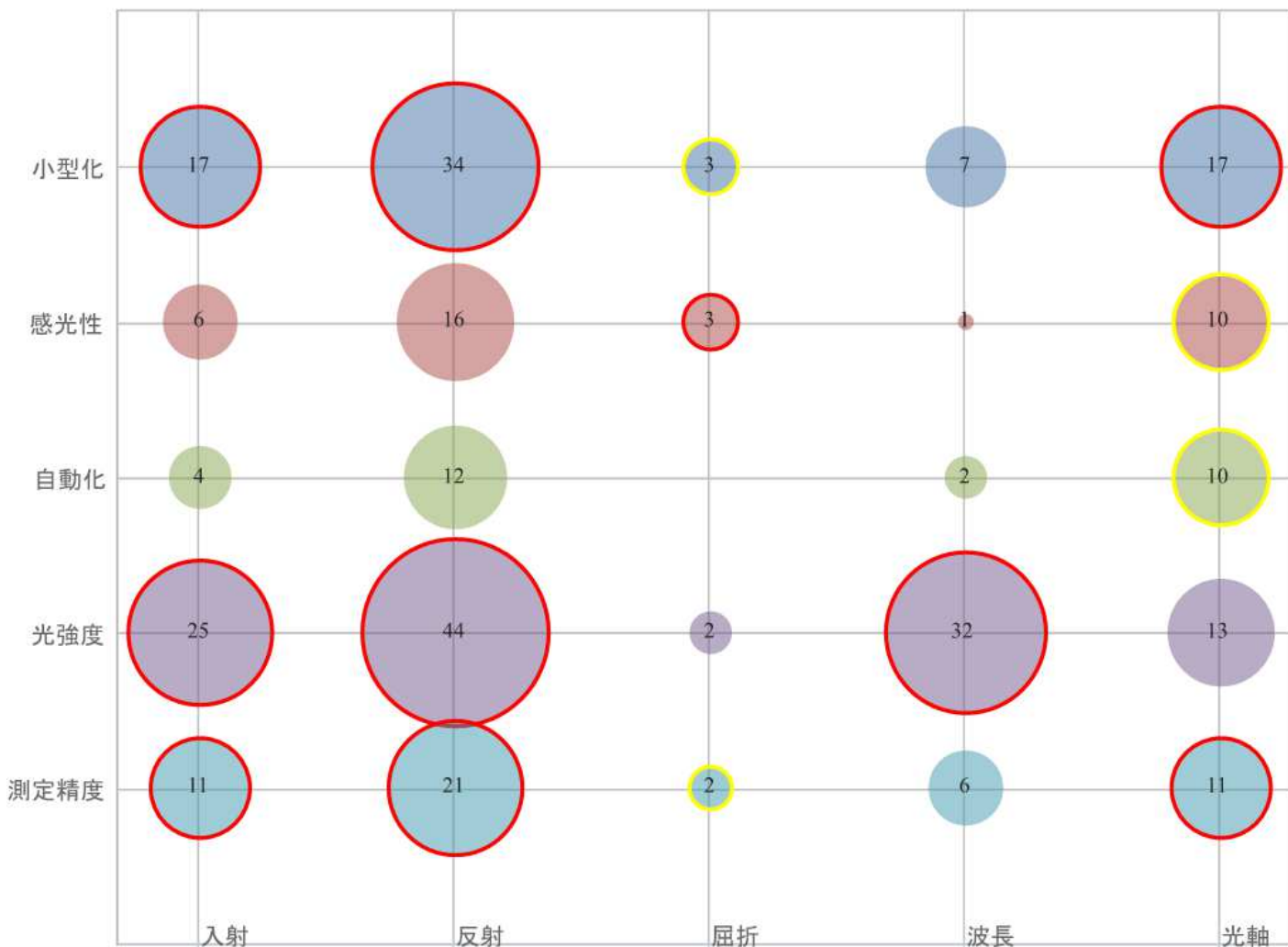


図1 ホットスポット分析による特許マップ

図1は、光学機器業界のある企業の特許マップになる。「自動化」の観点での出願件数が少ない上に、「自動化」の行においてはハイライトされている箇所が「光軸」との1箇所しかないことから、その企業の特徴と言える技術分野ではないと考えられる。

一方、赤色や黄色でハイライトされた箇所に注目すると、課題の観点では「小型化」「光強度」「測定精度」等に多くのハイライトがあり、自社の特徴・強さを発揮できている分野と予測できる。また、要素技術の観点では「光軸」の列には4箇所のハイライトがあり、この企業にとって特徴的な技術分野になっていることが分かる。

特徴マップ分析

特徴マップ分析とは、先程のホットスポット分析と同様、カイ2乗値を利用したクラメールの連関係数により、確率分布の中央値(期待値)からのずれ(偏差)を計算してビジュアル的な描画に置き換える手法である。

グループ内で偏って出現している属性値を母集団から網羅的に抽出し、関連性の高いものをグラフィカルに表現できるため、それぞれの項目間の特徴・関連性を視覚的に一目で捉えることができる点で有用である。

図2は、特徴マップを用いて、光学計測機器のコア技術である「光軸」「撮像」「測定」「精度」「光源」をグループとして、それらに関連

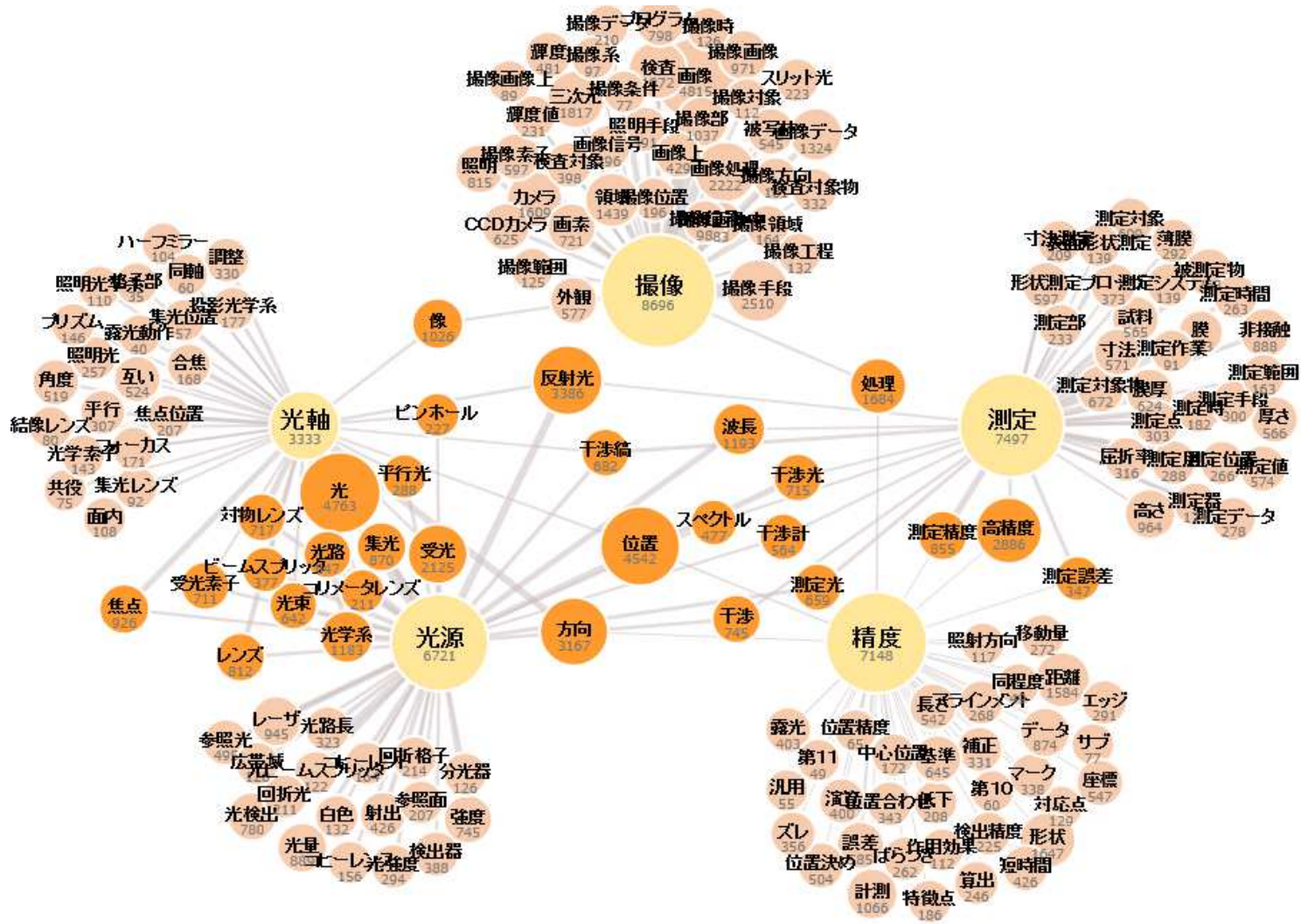


図2 特徴マップによる技術網羅性の俯瞰図

するキーワードの関係性を描いた技術網羅性を示す俯瞰図になる。これらの5つのコア技術の複数個に共通して関連するキーワードが図の中央に配置され、いずれか1つだけのコア技術に関連するキーワードが図の外側に配置されている。

例えば、「光軸」と「光源」の間には多くの共通関連キーワードが存在しており、特に密接なコア技術として存在していることが分かる。また、画面中央に描画された「位置」が最も多くのコア技術と関連していることより、技術俯瞰図の中核を担うキーワードであることが分かる。なお、マップの「光軸」「位置」等のバブルをクリックすることによりCPDに収録されている特許公報にアクセスができるため、具体的な内容の詳細も容易に確認できる。

また、マップ上にプロットされた属性間の線は、線が太ければ太いほど関連性が高いことを示している。図2では、「光軸」と「光源」に共通のキーワードとして「焦点」があるが、光軸-焦点の線が光源-焦点の線よりも太いことから、「焦点」は「光軸」寄りのキーワードであると言える。

特徴マップの別の用途としては、例えばコア技術の代わりに、課題として図1で偏りが大きかった「光強度」と「測定精度」をグループとして定義して、その両方の課題に共通する要素技術のキーワードを模索すると「光源」と「反射光」などが抽出でき、課題解決の新たな観点を得られる(図3を参照)。

あくまで特徴マップの使い方の一例にすぎないが、他社が保有する特許技術の理解についてキーワードの漏れは無い、課題解決に際して見落としている観点は無い、業界の技術俯瞰図の中核となっているキーワードとは何かを浮かび上がらせることにより、新たな創造へとつなげることができる。

CPD-TM ～AI搭載型の特許解析エンジン～

従来のTTLでは、図1のような特許マップを作成するために必要な縦横のグルーピング条件は、全て人間が手作業で調査して作成する必要があったため、多大な手間を要していた。しかし、CPD-TMでは新たに機械学習理論が実装されているため、システムが自分で案件を調査してグルーピングの軸とするべき項目の候補を抽出してくれる。

具体的には、教師データ(システムが手本とする特許情報)を必要としない「k平均法」と「トピックモデル」、教師データを必要とする「ナイーブベイズ法」の3種類のアルゴリズムが実装されている。「k平均法」と「トピックモデル」は、それぞれ単語が登場する傾向に着目する分析手法、単語の持つトピック(話題性)に着目する分析手法という点で異なる。また、ナイーブベイズ法はシステムが教師データに基づいて各グループの単語の出現頻度を学習し、新たな特許情報が登録された場合にどのグループに分類をするのが最も適切かをシステムに予測させる分析手法になる。

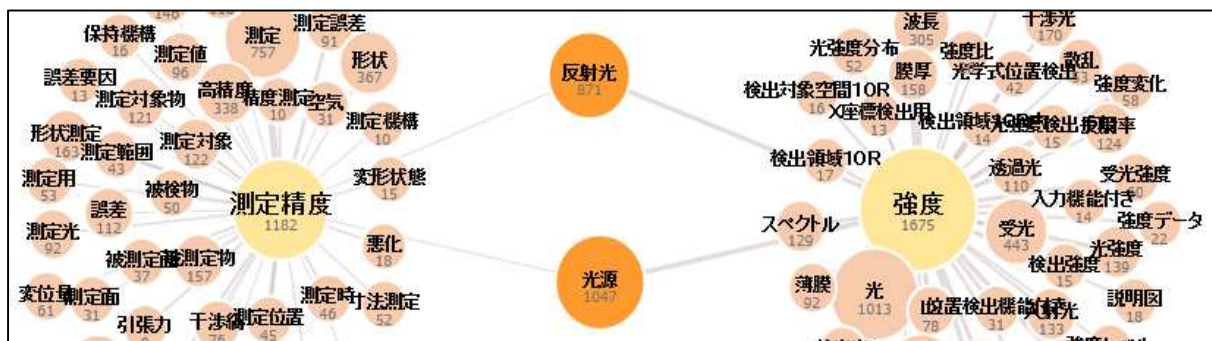


図3 特徴マップによる課題解決技術の俯瞰図(一部抜粋)

「k 平均法」「トピックモデル」は、教師データを与える必要が無い場合初動の負担が少なく、時間をかけずに大雑把な動向を把握したいような短期的な調査の場面で有用なツールである。一方、「ナイーブベイズ法」は、教師データをある程度与える必要があるため初動に負担がかかるものの、継続的に精度の高い分析を大量に行いたいような調査の場面で有用なツールである。

なお、どちらの分析手法を使うべきかは調査の目的や場面により判断すべきであり、お互いが補完関係にある分析手法と言える。例えば、調査の初期段階では「トピックモデル」により大雑把な動向を確認して継続的にウォッチしていくべきプロジェクトの選定を行い、それらの中から厳選された少数のプロジェクトに対しては「ナイーブベイズ法」により教師データを与え中長期的な視点で分類精度を上げていくなどの使い分けが想定される。

分析のまとめ

このように、CPD-TM に新たに実装された分析手法によって、分析対象の特徴や課題をより視覚的に整理することができるようになってきた。この分析結果から得られた新たな着眼点は、例えば、競合他社の技術動向の把握、自社技術の強み・弱みの評価、R&D テーマの選定時の新たな気づきとして、活用することが期待できる。

また、本論では紹介しなかったが、CPD-TM は新機能が追加されただけでなく、解析エンジンを刷新することにより、プロジェクトの処理パフォーマンス、データ処理の応答性、アウトプットの視認性、インタフェース円滑化なども同時に改善がされているので、そちらもご期待いただきたい。

おわりに ～受託調査の紹介～

弊社では、新規に CPD-TM を導入間もないお客様や、高度な分析をされるお客様に対して、お客様に代わって特許調査・分析を行うサービスも提供している。

お客様より調査の内容を伺った上で、課題把握・仮説立案・分析軸の構築作業を弊社にて実施し、調査・分析結果（調査報告書、CPD-TM のアクセス権）を納品する。標準的な調査期間は 2～3 ヶ月である。

このサービスの特徴は、特許情報に限らずに、論文や新聞などの様々な情報も統合し、1 つのプロジェクトとして分析できる点にある。また、CPD-TM を導入していただいているお客様には分析結果のプロジェクトをご契約の環境に直接納品するため、調査が終了した後も母集団の特許データを常に最新の状態にしてグルーピング分析などを継続的に使用していただける点にある。

参考文献

1. “その"技術開発"本当にそのまま続けますか？” 日刊工業新聞社 (2011)
2. “知的財産イノベーション研究の展望” 白桃書房 (2014)
3. “テキストマイニングによる解析サービス「TRUE TELLER パテントポートフォリオ」の特徴”， 情報の科学と技術, Vol.66 No.5, pp.243-246 (2016)
4. 「TRUE TELLER パテントポートフォリオの活用事例 ～光学機器をテーマとした分析例～」 NRI-CP 通信 2017 Vol.1 (https://www.patent.ne.jp/company/letter/NRI_CP_LTR_2017-01.pdf)

以上