IPランドスケープによる経営戦略支援のための共起語ネットワーク作成

2020032　平井遥斗　　情報基盤工学講座　指導教員 奥原浩之

要約

特許情報は過去の情報をアーカイブしたいわば発明の保管庫的なデータであり，それを活用することで経営戦略・技術的発展等広く社会に役立てることができる．

しかし，現状の特許プラットフォームは人手で少数の特許事例を調べるのには必要充分であるが，ビッグデータとして特許全体の分析を行いたい場合には整理されているとはいいがたい．

また，コロナ化をはじめ，サステナビリティやESGの推進など，昨今の社会変化に手対応していくためには多面的な視点から経営戦略を策定することが不可欠である．そこには攻めの知財情報活用であるIPランドスケープが有効である．

キーワード： 自然言語処理，知的財産戦略，IPランドスケープ，特許情報処理，テキストマイニング

1 はじめに

近年，コロナ禍をはじめグローバリゼーション，インターネット技術やAI，IoT等のデジタル化技術の進展や顧客のニーズの多様化や社会環境などの急速な変化などの様々な要因により将来の予測が困難な時代となっている．このような社会変化がすさまじく不確実性が高まる社会に対応していくためには，コア技術を高めることによって差別化を行い優位性を確立することが重要であり，多面的な視点から経営戦略を策定することが不可欠である．このような変化は，企業にとってチャレンジの機会となっており，IPランドスケープには，その役割の一つとして，新規事業に乗り出そうとする会社がその保有する要素技術を生かして参入することができ，競争力を高めるための新たな市場・用途・商品・サービス等を提案することが期待されている．

本研究では莫大な特許情報を整理し，可視化を行うことで，新たな市場・用途・商品・サービスの探索・提案を行う手法の提案を行う．

2 知的財産戦略と特許情報処理

2.1 知的財産戦略とIPランドスケープ

知的財産戦略（知財戦略）とは，企業が自身の知的財産を活用し，経営戦略に組み込むためのアプローチである．事業環境が大きく変化する時代において，企業の重要な資産の一つである知的財産を活用することで，事業を成功に導き，企業価値を高めることを目的としている．

知財戦略と経営戦略とは，企業の持続的な発展に向けて密接に関連しており，知財戦略は，経営戦略の一部として位置付けられるが，経営戦略において各機能別戦略の方向性を決定する重要な役割を果たす．

これらの知財戦略は日本においては特許などの「Intellectual Property（知財）」と景観や風景を意味する「Landscape」を組み合わせた造語で「IPランドスケープ」と呼ぶことが多い．

2021年に特許庁が公表している「経営戦略に資する知財情報分析・活用に関する調査研究」によるとIPランドスケープが必要であると回答した者は約8割であった．しかし，IPランドスケープを十分に実施できていると回答したものは約1割であった[1]．現在，必要性は理解しているがまだ実施に至っている企業が少ないという状態である．

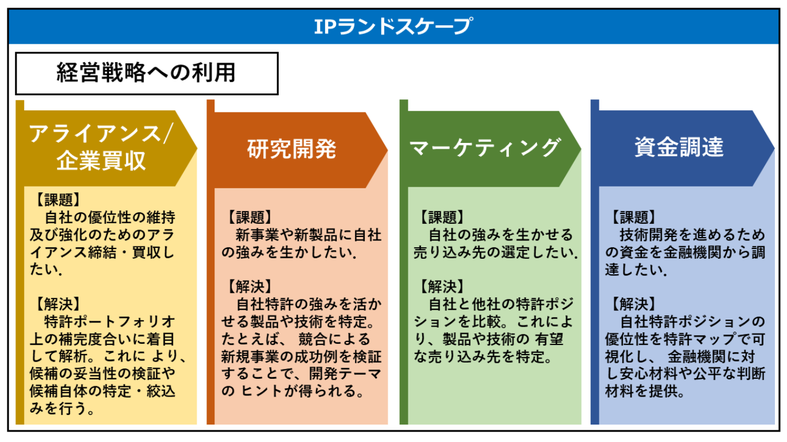

図1 IPランドスケープの活用

2.2 特許情報処理

特許とは知的財産の一部であり発明の保護を目的として，ある発明に対して独占・排他的に実施されるために付与された権利であり，特許権とも呼ばれる．日本では約34万件の特許が出願され，多様な分野の発明が蓄積されている[2]．

公開番号とは，個々の公開特許公報に付与される番号である．出願番号と同様に公開年が識別できるような書式で付与される．ただし，2000年以降は，「特開2006-123456」の書式．同じく，公表特許公報には「特表平10-123456」「特表2006-123456」のような公報番号が付与されるが再公表特許公報には「WO2006/123456」のような国際公開番号がそのまま公報番号として付与される．

国際特許分類（International Patent Classification：IPC）は，特許文献の国際的な利用の円滑化を目的に作成された世界共通の特許分類である．2023年11月現在，IPC第8版（2006年1月発効）が最新の分類となっており，技術の進展に柔軟に対応するため，適宜改正が行われている．

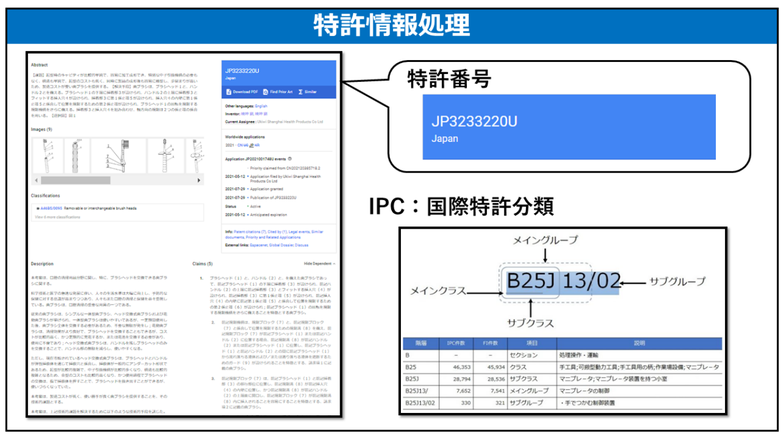

図1 多目的最適化の定式化

2.3 自然言語処理とテキストマイニング

自然言語処理とは，人が書いたり話したりする言葉をコンピュータで処理する技術である．人工知能の研究分野で中核を成す要素技術の一つといえる．自然言語処理技術は「言語理解」と「言語生成」に大きく二つに分けることができる.「言語理解」は人が書いた文章に対してなんらかの処理をする技術であり，メールの自動分類，ウェブ検索などが典型的な応用になる.「言語生成」は，コンピュータに文章を生成させる技術で，文章の要約や機械翻訳などを含む．

テキストデータは,「定性データ」の代表的なもので，この「定性データ」から付加価値の高い情報を収集することがテキストマイニングの目的である．アイデア発想において人間は自然言語から思考して発想することが一般的である．そこでサイバー空間にあるテキストデータを自然言語処理することを考える．現代社会においてインターネット上の情報量は莫大になっており，今後も増え続けることが予想される．このインターネット上の情報を収集して分析することでIPランドスケープに生かせると考える．

3 共起語ネットワークの作成

3.1 特許情報のベクトル化

莫大な情報を特許情報の分析を行うためにそれぞれの特許をベクトル化し整理したうえで可視化を行う必要があると考える．

ベクトル化にはSentence-BERT(Sentence-Bidirectional Encoder Representation from Transform)を用いる[3]．Sentence-BERTは、2018年10月11日にGoogleが発表した自然言語処理モデルである「BERT」を改善したモデルである．BERTは2つの文章を比較することにはたけているが，複数の文章を比較すルには精度がいまいちである．そこで「Saimese Network」という手法を用いて複数文章をインプットすることができるようになったものが「Sentence-BERT」である．本研究では，事前学習に「Hugging Face」に登録されている日本語用のSentence-BERTの事前学習モデルを使用した．

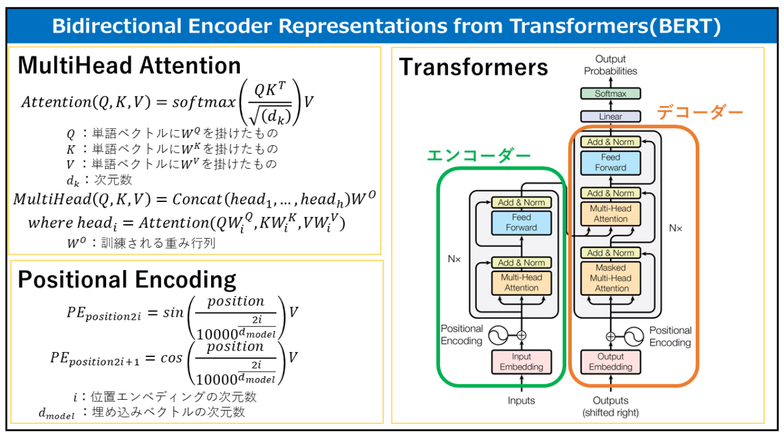

図2 BERT

3.2 次元圧縮とクラスタリング

今回扱うデータは768次元と高次元であるためクラスタリングを行う際に「次元の呪い」が発生することが考えられる．この「次元の呪い」を回避するために，次元圧縮を行う．次元圧縮手法にはUMAP(Uniform Manifold Approximation and Projection of Dimension Reduction)を用いる[4]．

本研究では，k-means法を用いてクラスタリングを行う[5]．k-means法は，初めに指定したクラスタの数だけ「重心」をランダムに指定して，その重心をもとにクラスタをグルーピングしていくという手法である．k-means法を活用すれば，データ間の距離を計算する必要がなくなるというメリットがある．ここで，k-meansでは事前にクラスターの数を与える必要があためクラスター数を決定するためにクラスター分析を行う．クラスター分析にはシルエット分析を用いる．シルエット分析におけるシルエット係数は，クラスタリングの品質を評価する指標であり，クラスター内のサンプルがどれだけ密集，分離しているのかを示す指標である．本研究ではシルエット係数が一番大きくなるクラスター数でクラスタリングを行う．

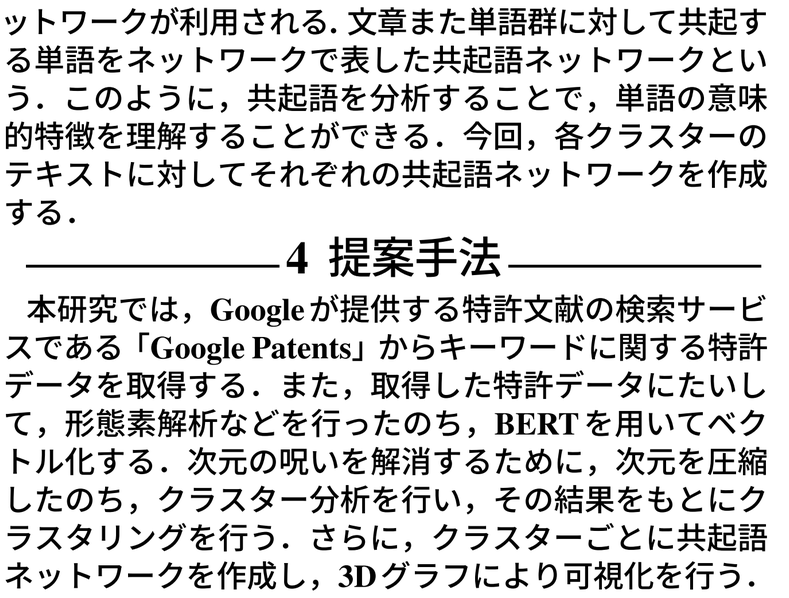

図3 UMAP

3.3 共起関係と共起語ネットワーク

ある単語とある単語が同時に出現することを共起するといい，文章において関係深い単語は共起することが多い．共起分析では単語同士のJaccard係数を比較したり，共起関係を持つ単語と単語を線で結んで描かれる共起ネットワークが利用される．文章また単語群に対して共起する単語をネットワークで表した共起語ネットワークという．このように，共起語を分析することで，単語の意味的特徴を理解することができる．今回，各クラスターのテキストに対してそれぞれの共起語ネットワークを作成する．

4 提案手法

本研究では，Googleが提供する特許文献の検索サービスである「Google Patents」からキーワードに関する特許データを取得する．また，取得した特許データにたいして，形態素解析などを行ったのち，BERTを用いてベクトル化する．次元の呪いを解消するために，次元を圧縮したのち，クラスター分析を行い，その結果をもとにクラスタリングを行う．さらに，クラスターごとに共起語ネットワークを作成し，3Dグラフにより可視化を行う．

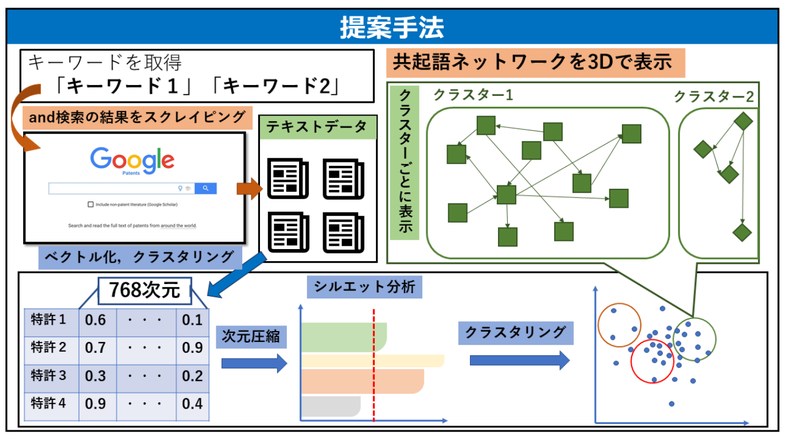

図4 提案手法

5 数値実験並びに考察

今回の数値実験では，従来の実例と同じくキーワードを「浸透」「隙間」「狭い」の3つを用いて共起語ネットワークを作成する．実験で扱う特許データは，2000年1月から2022年12月までに出願された特許を用いる．

従来の実例では共起語ネットワークを作成するだけであったが，提案手法では特許情報をベクトル化し，可視化を行うことで，莫大な情報を整理することができ広い分野の探索が可能になった．一方で，共起語ネットワークの作成にはまだ改善の余地が見られる．特許情報には複合語や専門用語などが多く含まれており，今回行った形態素解析ではそれらの抽出が行えておらず，別々の他の単語として抽出していたと考えられる．

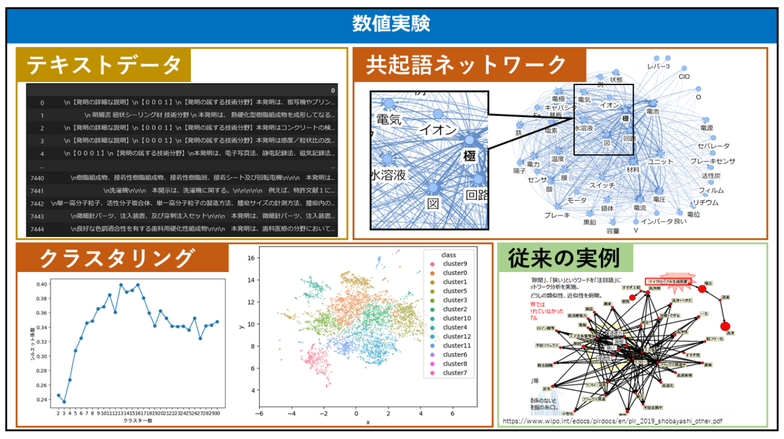

図5 実験結果

6 おわりに

本研究では，莫大な特許情報を整理し，可視化を行うことで，新たな市場・用途・商品・サービスの探索・提案を行う手法の提案を行った．今後の課題として，形態素解析を行う際の専門用語や複合語の抽出や，クラスターのタイトルをクラスター内の重要語などを用いて作成することでより，視覚的にわかりやすくなると考えられる．さらに，今回の実験ではスクレイピングから共起語ネットワーク作成までの時間が多くかかってしまう問題があり，実用化に向けて実行時間を短縮する必要がある．

参考文献

[1] 特許庁，”経営戦略に資する知財情報分析・活用に関する調査報告書”，https://www.jpo.go.jp/support/general/document/chizai-jobobunseki-report/chizai-jobobunseki-report.pdf，閲覧日 2023．11．07

[2] 藤井敦，山川英和，岩山真，難破英嗣，山本幹雄，山内将夫，”特許情報処理：言語処理的アプローチ”，コロナ社，2012．

[3] Jacob Devlin. et al．”Sentence-BERT：Sentence Embeddings using Siamese BERT-Networks”ArXiv e-prints，1908．10084，2019

[4] McInnes，L．，Healy，J．，& Melville，J：UMAP：Uniform Manifold Approximation and Projection for Dimension Reduction，ArXiv e-prints，1802．03426，2018

[5] Douglas Steinley，Michael J． Brusco，”Initioalizing k-menas Batch Clustering：A Critical Evaluation of Serveral Techniques”，Journal of Classification，Vol24，No．1，99-121，2007．