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，現在に至るまでの発明を保存したデーターベースのようなものであり，様々な発明が保管されている．これらを活用することで，経営戦略や技術の進歩など多岐にわたり，社会に貢献することができる．しかし，現状の特許プラットフォームでは，少数の特許を手作業で調査するには十分であるが，特許全体をビッグデータとして分析を行いたい場合には適しているとは言い難い．また，新型コロナウィルスの影響や，持続可能性やESGの推進など，最近の社会変化に対応していくためには多面的な視点から経営戦略を立てることが必要である．その中で，積極的に知的財産情報を活用する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，コロナウィルスの影響やグローバル化，インターネット技術やAI，IoT等のデジタル技術の進展，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，さまざまな要素が絡みあうことにより，将来を予測することが難しくなっている．急激な変化と不確実性が高まる社会に対応するためには，企業が保持しているコア技術を強化して差別化を行い，優位性を確立することが重要である．また多角的な視点から経営戦略を策定することが不可欠である．

これらの変化は，企業にとってチャレンジへの機会となっている．IPランドスケープでは，新しいビジネスを始めようとする企業が自社の保有する技術を活用して新しい市場に参入することを支援することが可能である．これにより，企業は競争力を高め，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，それらを経営戦略の一部として取り入れることを指す．事業環境が急速に変化する現代において，企業の大切な資産である知的財産をうまく活用することで，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，企業の持続的な成長を目指すうえで密接に関連している．知財戦略は経営戦略の一部として位置付けられ，企業の全体的な経営戦略において各部門や機能の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：IP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[?]．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．

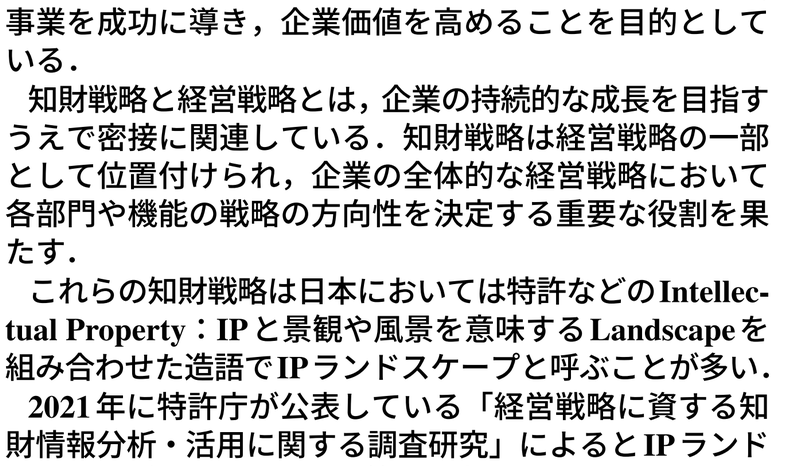

図1 IPランドスケープの活用

2.2 特許情報処理

特許とは知的財産の一部であり，一般的には20年間の期限が設けられている．その期間自身の発明を他者に盗用さることを防いだり，他者が発明を利用する際に，使用料金などを受け取ることが許可される．

特許番号は，特許文章の表紙に記載されており，特定の特許を識別，検索するのに役立っている．特許番号には主に出願番号，公開番号，登録番号の3つが存在する．出願番号および，公開番号は，出願または公開された西暦と通し番号で構成されており，登録番号は通し番号のみで構成されている．

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[?]．

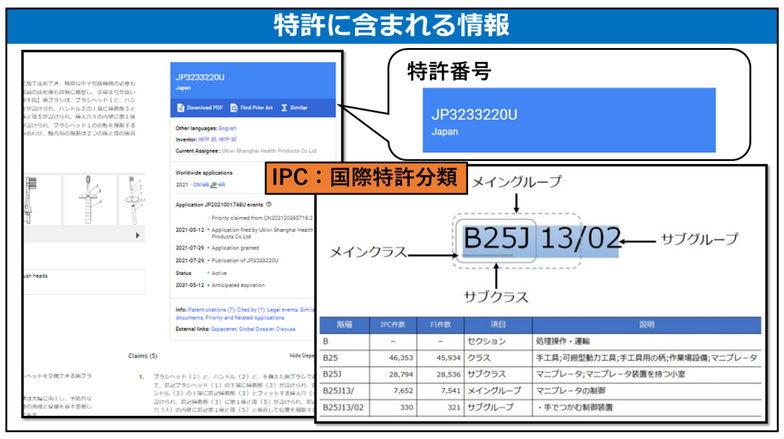

図2 特許情報処理

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が理解し，処理するための技術である．これは，人工知能の研究分野で重要な部分でありコンピュータが人間のような思考や学習を行うための基礎となっている．自然言語処理技術は大きく分けて，言語理解と言語生成の2つに分けることができる．それらの中でも言語理解は人間が書いた文章をコンピュータが処理し，文章を理解する技術のことである．これにより，コンピュータは与えられた文章の内容を把握し，適切な対応を行うことができる．具体的な応用例としてはメールの自動分類やウェブ検索などがある．

テキストデータは，人間の言葉や意見などの定性的なデータを含んでおり，これらのデータから有益な情報を抽出することがテキストマイニングの目的である．人間は自然言語からアイディアを生み出すことが一般的であるため，インターネット上のテキストデータを自然言語処理することは非常に重要である．現代社会では，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．これらインターネット上の情報を収集し，分析することでIPランドスケープに生かすことができると考える．

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，全体を俯瞰して見渡せるように，これらの情報を可視化する必要があると考える．

ベクトル化にはSentence-Bidirectional Encoder Representation from Transform：Sentence-BERTを用いる[?]．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る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．

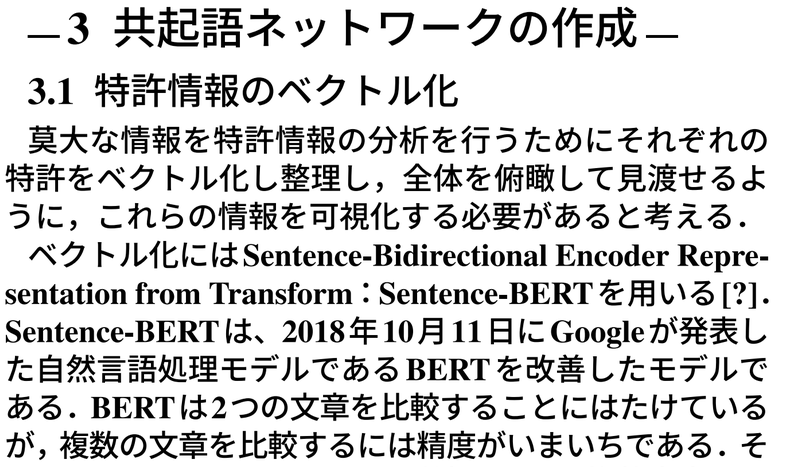

図3 BERT

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Uniform Manifold Approximation and Projection of Dimension Reduction：UMAPを用いる[?]．

本研究では，k-means法を用いてクラスタリングを行う[?]．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リングを行う手法である．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．

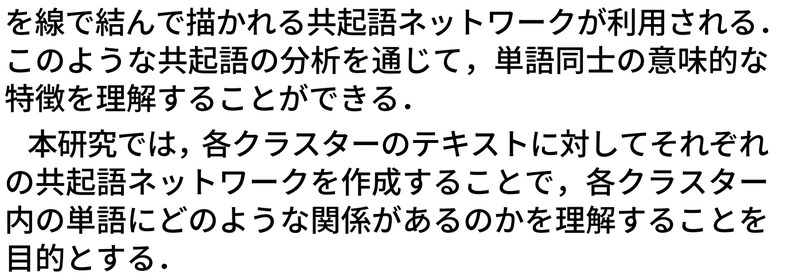

図4 UMAPによる次元圧縮

3.3 共起関係と共起語ネットワーク

共起するとは，ある単語と別の単語が同時に文章中に出現することを指す．関連性の高い単語は，一緒に出現することが多いため，それらの単語の共起関係を調べることで，単語間の関係性を理解することができる．共起分析では単語同士のJaccard係数とい指標を用いて単語同士の共起度合いを比較し，共起関係にある単語と単語を線で結んで描かれる共起語ネットワークが利用される．このような共起語の分析を通じて，単語同士の意味的な特徴を理解することができる．

本研究では，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ことで，各クラスター内の単語にどのような関係があるのかを理解することを目的とする．

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，Sentence-BERTを用いてベクトル化する．さらに，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．最後に，共起度合いを算出したのち，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．

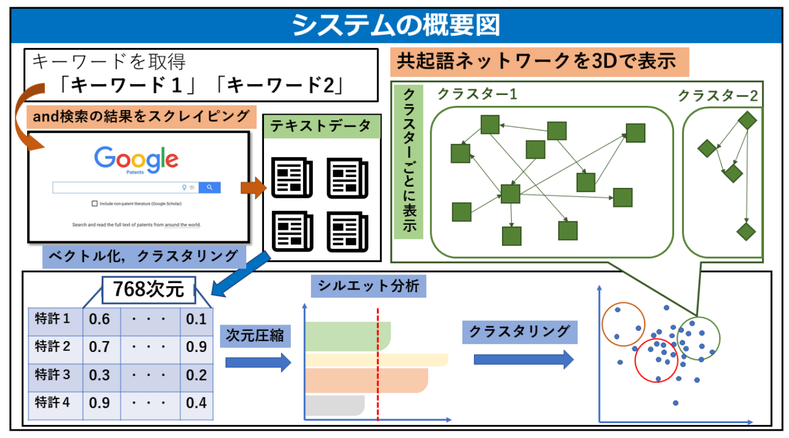

図5 提案手法

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，特許全体を俯瞰してみることが可能となった．そのことにより，従来よりも広い分野探索や研究課題の発見が可能になったと考える．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．

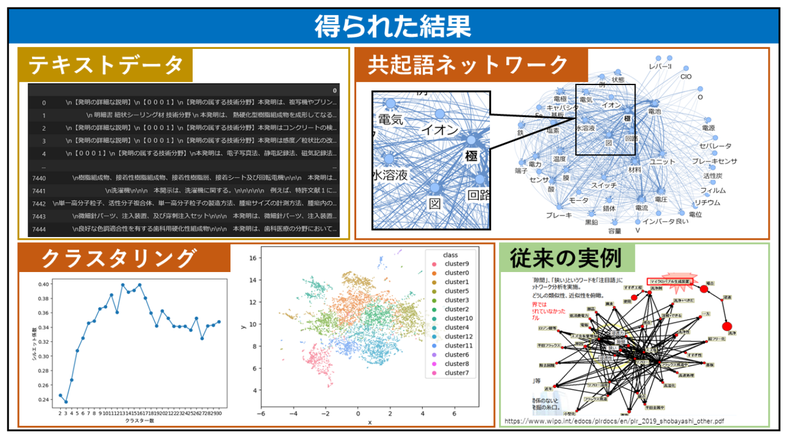

図6 実験結果

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] リサーチ・ナビ，”日本の特許を特許分類から調べる” ，https://rnavi.ndl.go.jp/jp/patents/post_398.html，閲覧日 2023．11．07

[3] Nils Reimers，Iryna Gurevych．”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．