卒業論文

IP ランドスケープによる
経営戦略支援ための
共起語ネットワーク作成

Data Fusion through Web-GIS Visualization
Using Open Data for Evidence-Based Policy Making

富山県立大学 工学部 情報システム工学科
2020032 平井遥斗
指導教員 奥原 浩之 教授
提出年月: 2024 年 2 月

目 次

図一覧

表一覧

記号一覧

以下に本論文において用いられる用語と記号の対応表を示す.

用語	記号
MultiHead Attention における単語ベクトルに W^Q を掛けたもの	Q
MultiHead Attention における単語ベクトルに W^K を掛けたもの	K
MultiHead Attention における単語ベクトルに W V を掛けたもの	V
MultiHead Attention における次元数	v^T
MultiHead Attention における訓練される重み行列	W^O
Positional Encoding における位置エンベディングの次元数	i
Positional Encoding における埋め込みベクトルの次元数	d_{model}

第1章

はじめに

§ 1.1　本研究の背景

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 [1]

特許庁の調査によれば，IP ランドスケープが必要であると回答した者は約 8 割であったのに対し，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 [2]．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．

§ 1.2　本研究の目的

§ 1.3　本論文の概要

本論文は次のように構成される．

第 1 章 本研究の背景と目的について説明する．

第 2 章

第 3 章

第 4 章

第 5 章

第 6 章 本論文における前章までの内容をまとめつつ，本研究で実現できたことと今後の展望について述べる．

第2章

知的財産戦略と特許情報

§ 2.1　知的財産戦略

知的財産戦略とは企業が保有する知的財産を経営戦略の一環として取り入れ，企業の競争力を高め，事業目標を達成することを目的とする戦略である．知的財産には，特許，商標，意匠，著作権，ノウハウなど，さまざまな種類があり，これらの知的財産をどのように活用すれば，企業の価値を最大化できるのかを考えることが重要である [3].

知的財産戦略は，経営戦略と密接に関係しており，企業全体の戦略において各部門や機能の方向性を決定する重要な役割を果たしている．日本において，知的財産戦略は特許などの知的財産（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」と呼ばれることが多い．

知的財産戦略の目的は，以下の 3 つにまとめることができる [4].

（1）オープンイノベーション創出に貢献する知的財産戦略

- オープンイノベーションによる事業創出に貢献する知的財産戦略

  オープンイノベーションによる事業創出とは，近年の変化が激しい事業環境下において，従来のような社内のみで行う研究開発では，新規事業の創出に限界があることを踏まえ，競合企業やスタートアップ，大学等などの外部からの技術やアイディアを自社に取り組むこと等を通じて新たな価値を創造し，事業を創出しようとするもの．

- プラットフォーム戦略の推薦による事業創出に貢献する知的財産戦略

  プラットフォーム戦略の推進による事業創出とは，顧客や事業など，様々な主体を同一のプラットフォーム上に集めることで，事業のエコシステムを創出するビジネスモデルであるプラットフォーム戦略の推進により事業を創出しようとするものである．

- プソリューションビジネスの事業創出に貢献する知的財産戦略ソリューションビジネスとは，従来のモノ売りのビジネスから脱却し，顧客の課題を解決するコト売りへと進化したビジネスである．すなわちソリューションを創出するビジネスである．従来は知財部門が顧客の課題解決に直接関与することは少なかったが，近年は知財部門が積極的に関与し，新たなソリューションのコアを早期に特定し，これを適切に保護する知財ポートフォリオを構築している企業が増えている．

（2）事業競争力の強化に貢献する知的財産戦略

- コアインピーダンス強化に貢献する知的財産戦略コアコンピタンスとは，競合他社との差別化につながる競争優位性をもたらす自社の強みであり，これを技術として支えるのがコア技術である．コアコンピタンスを現状からさらに磨き，深化させることは，競争優位性を維持・強化するために重要である．

- グローバル事業展開に貢献する知駅財産戦略

  グローバル事業展開の形態として，輸出，ライセンシング，戦略的提携，買収及び現地子会社の新設等がある．

- M&A による事業ポートフォリオの拡大に貢献する知的財産戦略 M&A による事業ポートフォリオ拡大とは，社外に存在する事業を M&A を実施して買収することで，自社の事業ポートフォリオを拡大することである．M&A は，既存の事業の規模拡大の経済効果や，新規事業への参入新たな技術やノウハウの獲得など，様々な目的で実施される．

（3）組織・基盤の強化に貢献する知的財産戦略

- ブランド価値向上に貢献する知的財産戦略

  ブランド価値の向上は，顧客からの信頼や好感を高め，他社に対しての競争優位性を構築するだけでなく，資金調達や人事確保の容易化など，企業の組織・基盤の強化にもつながる．ブランド価値は，高い経営理念に基づいた企業活動によって向上させることができる．

- デジタルトランスフォーメーション（DX) 等による事業基盤の強化に貢献する知的財産戦略

  デジタルトランスフォーメーションによる事業基盤の強化とは，IT やデータ等のデジタル技術をを活用して，自社の事業基盤の強化を図るものである．近年，知財情報等を自社の事業基盤を強化するために利用する取り組みが注目を集めており，DX において，知財部門が貢献できることは少なくない．

- SDGs への貢献に関わる知的財産戦略 SDGs（持続可能な開発目標）の取り組みは，国際社会から企業への信頼を高め，グローバルな投資家から高い評価を得るために重要である．また，企業の持続的発展のためにも欠かせないものとなりつつある．

経営戦略

「経営戦略」とは，企業が競争環境の中で自らの経営目的・経営目標を達成するための方針や計画全般を意味する．どれほど巨大な企業であっても，保有する経営資源（ヒト，モノ，カネ）は有限だ．企業が掲げる目標や目的に応じて，選択し分配していく必要がある．そうした企業活動の基軸となる指針や指標，また方策を実現するための体制づくりなども「経営戦略」には含まれる．企業はすべてのリソースを有しているわけではない．当然ながら，強みや弱み，特性がある．それらを経営者は理解・把握した上で，組織改革や事業の方向性を決定していかなければならない．やるべきこと，やらなければいけないことは数

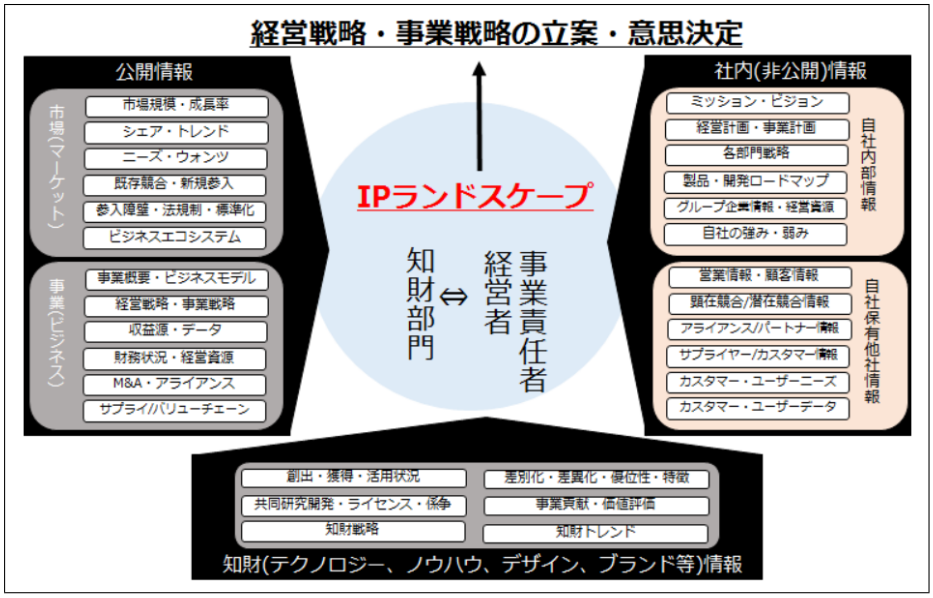

図 2.1: IP ランドスケープの概要 [5]

多くある．それらにどのような優先順位を付けて実行していくかを明確に打ち出していくためにも，戦い方の根幹がどうしても必要になってくる．それが,「経営戦略」を策定する目的と言って良いだろう．

グローバル化の進展や IT・AI の普及，ニーズの多様化，競争環境の激化など，現代は変化のスピードがますます増している．こうした時代において，企業は 10 年，20 年先の生き残りに向けてどのような成長シナリオを描いていくかが，一段と問われるようになってきている．そのためにも，経営者は自社の強みや特性を把握・理解し，組織改革や事業の方向性をスピーディ，かつダイナミックに決定していかなければならない.「経営戦略」の必要性が高まっているのもそのためである．

経営戦略のフレームワーク

経営戦略の策定では自社を取り巻く外部の環境要因に打ついて分析する外部環境分析や，自社内の環境を分析する内部環境分析を踏まえ自社の強みや弱み，機械や脅威を把握することで，戦略オプションを立案して最適な戦略を選択することが大切である．それらを行うために役立つ代表的なフレームワークとして，PEST 分析ファイブフォース分析，3C 分析，VRIO 分析，SWOT 分析，STP 分析，4P 分析などが挙げられる [6].

内閣府や特許庁による IP ランドスケープの積極的な推進に代表されるように，IP ランドスケープは研究機関においても積極的に検討されるべき対象であると考えられる．また，その具体的な取り組みの多くに ICT を活用した取り組みが多数行われていることから，IP ランドスケープの効率的な実施には ICT の活用が不可欠であり，情報工学との親和性が高いものと思われる．これらのことから，本研究は情報技術を用いた IP ランドスケープの支援を目的とする．

§ 2.2　特許情報処理

特許情報とは，特許・実用新案・意匠・商標の出願や権利化に伴って生み出される情報である．この情報は，研究開発の重複防止，既存技術の活用，無用な紛争の回避などに役立つ．

特許情報は，研究開発の策定から商品化，更には他人の権利調査に至るまでの様々な事業活動において活用されている．

具体的な活用例は，以下のとおりである．[8].

特許情報の活用例

- 技術動向調査

  将来性を見据えた研究テーマの選定や過去になされた研究との重複回避のために，特許情報を利用して技術動向調査が行われる．特定の技術分野における特許出願の動向や出願件数の推移を調査することにより，過去にどのような技術が存在したか，また，今後開発すべき技術分野の把握の参考になる．

- 出願前の先行技術調査

  研究成果として発明がなされたとき，権利化するか否かの判断が必要となる．特許出願をする際に関連する分野の先行技術について調査することにより，権利として認められる見込みのない無駄な出願を未然に防止することができる．

- 権利調査

  開発製品が他人の産業財産権を侵害すると，製造・販売の中止や製造品の廃棄，あるいは権利者への損害賠償にまで発展する恐れがある．これらを未然に防止するために，設計から製造前段階にかけて，他人の権利範囲の調査を行う．

- 公知例調査

  他の権利者から警告を受けた場合などの対抗手段として，自社の発明・考案を事業化する際に障害となる他人の特許権・実用新案権を無効にするため，その特許・実用新案登録の出願前の公知例を調査する．

- 公知例調査

  事業を営む上で多くの場合には競合他社が存在している．その競合他社がどのような戦略で事業を行っているか調査する上で，特許情報は貴重な情報源となる．競合他社の過去から現在に至るまでの出願動向を把握することにより，研究開発動向等を読み取ることが可能である．また，競合他社の出願動向を継続的に監視し，自社にとって障害となる出願等の早期発見に努めることも重要である．

特許番号

§ 2.3　テキストマイニングと自然言語処理

テキストマイニングとは，定型化されていない文章から情報を抽出する技術です．SNS やアンケート，コールセンターの応対など，さまざまな場面で活用されている．テキストマイニングは，AI 技術の進展により，より高度な分析が可能になった．また，テキストデータの量も増加しており，テキストマイニングツールの種類も増えている．[9]

テキストマイニングを行うことで，企業は，顧客のニーズや市場動向を把握したり，新製品の開発やマーケティングの戦略を策定したりすることができる．

このことによって，単一のデータの可視化のみでは表面化してこなかった課題をくみ取ることや，逆に，課題に対する解決策を一見関係のなさそうな分野から発見するといったことが可能となる．

形態素解析

形態素解析は，

分かち書き

自然言語処理（NLP）において，分かち書きは，テキストを単語や句などの意味的な単位に分割する処理である．分かち書きは，テキストの意味理解や解析の基礎となる重要な処理であり，多くの NLP タスクで必要となる．

分かち書きの目的は，テキストの意味を正確に理解するために，テキストを単語や句などの意味的な単位に分割することである．例えば，文の意味を理解するためには，文を主語，述語，目的語などの句に分割する必要がある．また，単語の意味を理解するためには，単語を品詞や語義などの単位に分割する必要がある．

分かち書きの処理方法は，大きく分けて以下の 2 つに分けられる [7].

1. ルールベース型
   ルールベース型分かち書きは，あらかじめ定義された対象となる言語の文法ルールに基づいて分かち書きを行う方法である．ルールベース型分かち書きは，人手でルールを定義するため，単純な分かち書きを行う場合は比較的容易に実装でき，調整も可能であるが，複雑な分かち書きを行う場合は，ルールを複雑にする必要があり，高度な専門知識が必要となり，誤りが生じやすくなる．
2. 統計学習ベース型
   統計学習ベース型分かち書きは，機械学習によって導き出されたルールに基づいて分かち書きを行う方法である．統計学習ベース型分かち書きは，複雑な分かち書きを行う場合でも，比較的正確に分かち書きを行うことができる．また，機械学習に大量のテキストデータが必要であり，計算量が大きいという問題もコンピュータの高速化と低価格化により解決に向かっている．

分かち書きの精度は，分かち書きの目的や，分かち書きを行うテキストの種類によって異なる．例えば，新聞記事などのフォーマルなテキストであれば，ルールベース型分かち書きでも比較的高い精度で分かち書きを行うことができる．一方，SNS の投稿などの非フォー

マルなテキストであれば，統計学習ベース分かち書きの方が高い精度で分かち書きを行うことができる．

近年，NLP 技術の進展により，分かち書きの精度も向上している．また，クラウドサービスやオープンソースソフトウェアの普及により，分かち書きの利用が容易になってきている．

また，現状の分かち書きには，以下の課題がある．

- 日本語の曖昧さ

  ルールベース型分かち書きは，あらかじめ定義されたルールに基づいて分かち書きを行う方法である．ルールベース型分かち書きは，単純日本語は，英語と比べて曖昧な表現が多い言語である．例えば,「私は，彼に会いました.」という文は，文法的には「私は，彼に会いに行きました.」という意味にも解釈できる．このような曖昧な表現を正確に分かち書きすることは，困難である．

- 新語や流行語

  常に新しい言葉や表現が生まれてくるため，分かち書きのルールや統計モデルを常に更新する必要がある．

- 誤ったデータの影響分かち書きの精度は，分かち書きの対象となるデータの品質に大きく影響を受ける．誤ったデータが含まれていると，分かち書きの精度が低下する．

今回用いる特許の本文には，専門的な用語や複合語が多数含まれているため，それらを正しく抽出する必要がある．そのため，python のモジュール termextract を用いて専門用語や複合語の抽出を行い，それらを分かち書きの辞書に登録する．

tarmextract

tarmsextract は，東京大学情報理工学系の松本研究室によって開発されたテキストデータから専門用語を抽出するための Python モジュールである．

今後，テキストマイニングと自然言語処理は，AI や ML 技術を活用することで，より高度な分析が可能となり，より幅広い分野で活用されるようになると考えられる．また，テキストデータの量の増加に対応するため，テキストマイニングと自然言語処理の自動化や，テキストデータの検索・分析・活用を効率化する技術の開発が進んでいくと考えられる．

第3章

特許情報の可視化

§ 3.1　特許情報のベクトル化

特許情報は，日々蓄積され，今では莫大な量となっており，それらの分析は困難を極める．そこで，特許情報を効率的に分析するためには，各特許をベクトル化して整理をおこない，全体を俯瞰できるように可視化する必要があると考える．本研究では，特許本文の文章を対象にベクトル化を行う．特許本文には，特許技術の内容が詳細に記載されているため，これらの情報をベクトル化することで，特許の技術分野や技術トレンドなどを把握することができると考える．

具体的には，特許本文を Sentence-Bidirectional Encoder Representation from Transform（Sentence-BERT）を用いることで文章全体を単位にベクトル化を行う [11]．Sentence-BERT は，Bidirectional Encoder Representations from Transformers（BERT）をベースに開発されており，文章の単語の順序を考慮して，文章の意味を表現するベクトルを生成する．

sentence-BERT は，文章の意味を理解する能力に優れているため，自然言語処理の様々なタスクに活用されている．

Transformer

近年，翻訳などの入力文章を別の文章で出力するというモデルは，Attention を用いたエンコーダー，デコーダ形式の RNN や CNN が主流であった．しかし，Transformer は，RNN や CNN を用いず Attention のみを用いたモデルである．Transformer は，再帰も畳み込みも一切行わないので並列化が容易であり，他のタスクにも汎用性が高いという特徴がある．Transformer においては Attention を多数並列に配置した Multi-Head Attention が用いられ，一般的に以下の式（3.2）のように定式化される [10].

<Multi-Head Attention>

$$MultiHeadAttention(Q, K, V) = Concat(head_1, head_2, \cdots, head_h)W_o \tag{3.1}$$

$$\text{where } head_i = ScaledDotProductAttention(QW_i^Q, KW_i^K, VW_i^V) \tag{3.2}$$

ここで，Scaled　Dot-Product Attentioin では，内積を利用したベクトル間の類似性に基づく変換を行う．

<Scaled Dot-Product Attention>

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.3}$$

3.2 では，学習パラメータを持っていない Scaled Dot-Product Attention の表現力を広げるために，入力の直前に学習パラメータを持つ Linear 層の追加を行っている．これにより，入力されるベクトルの特徴空間に依存しない注意表現を学習することができる．Linear 層の追加を行った Scaled Dot-Product Attention を一般に Single-Head Attention を呼ぶ．

Scale Dot-Product Attention は，ある単語に対して，その単語が文章に含まれる単語とどれだけ類似しているのかを計算し，それらを確率的に表現したものである．Transformer における Attention の入力には主に以下の 2 種類の入力方法が用いられている．

1. Self-Attention（softmax に与える Query，Key，Value を同じ値にする）
2. SourceTarget-Attention（Key，Value を同じ値にし，Query を異なる値にする）

Single-Head Attention では多種多様な意味や文法をもつ単語に対しても単一の注意表現が生成される．そこで，Single-Head Attention を多数並列に配置して Multi-Head にすることで，複数の特徴部分空間における注意表現の獲得をすることができる．

以上のことから，文章を行列で表せることが分かった．しかし，文章というのは，文字を読む方向が重要であり，行列として表され，かつ，一括で処理する場合，文字の順番の概念がなくなってしまう．これは，文章を正しく扱えなくなる可能性がある．そのため，Embedding 層からの行列に位置情報を含んだ行列を足し合わせることで，文字の順番の概念を扱えるようにする必要がある．これを可能にするのが Positional Encoding である．

<Positional Encoding>

$$PE_{(pos,\,2i)} = sin\left(\frac{pos}{10000^{\frac{2_i}{d_{model}}}}\right) \tag{3.4}$$

$$PE_{(pos,\,2i+1)} = cos\left(\frac{pos}{10000^{\frac{2_i}{d_{model}}}}\right) \tag{3.5}$$

入力文章の単語数が 50 個まで扱えて，Embedding 層の埋め込み次元数が 128 次元の場合，Positional Encoding が生成する行列は 128 次元の行ベクトルが縦に 50 個並んだ行列になる．この行列は，各行のベクトルが絶対に同じものにならいため，この行列から単語の位置情報を表すことができる．具体的には，行ベクトルの各次元は，単語の位置情報に応じて，異なる値が割り当てられている．例えば，最初の行ベクトルの最初の次元は，単語の位置が 0 であることを示し，最後の行のベクトルは，単語の位置が 49 であることを示す．このように，Positional Encoding は，単語の位置情報を行ベクトルに埋め込むことで，Transformer モデルが単語の順序情報を利用でいるようにしている．

BERT

自然言語処理タスクにおいて，精度向上には言語モデルによる事前学習が有効である．この言語モデルによる事前学習には「特徴量ベース」と「ファインチューニング」の 2 つの方法がある．まず,「特徴量ベース」とは 事前学習で得られた表現ベクトルを特徴量の 1 つとして用いるもの で，タスクごとにアーキテクチャを定義する．ELMo[Peters, (2018)] がこ

の例である．また,「ファインチューニング」は 事前学習によって得られたパラメータを重みの初期値として学習させるもの で，タスクごとでパラメータを変える必要があまりない．例として OpenAI GPT[Radford, (2018)] がある．ただし，いずれもある問題がある．それは事前学習に用いる言語モデルの方向が 1 方向だけということだ

Sentence-BERT

BERT では 2 つの文章を入力し，それらの類似度を測ることができる．しかし，複数の文章を入力する場合は BERT では容易ではない．そこで本研究では Sentence-BERT を用いる．BERT で求められた埋め込み表現を pooling し，それらを Softmax 関数を用いて，分類を行う．

<Siamese Network>

$$O = softmax(W_t(u, v, |u - v|))W_t \in R^{3_n \times k} \tag{3.6}$$

事前学習モデルは Hugging Face や GitHub などのサイト公開されている．また東京大学や京都大学なども独自のモデルを公開している．本研究では Hugging Face に登録されている ”sonoisa/sentence-bert-base-ja-mean-tokens ”を用いる．

§ 3.2　次元圧縮手法とクラスタリング手法

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ため，次元圧縮を行う．次元圧縮手法には線形次元圧縮手法と，非線形圧縮手法がある．線形次元圧縮手法は，計算がようであるが，データの非線形的な構造を表現することが難しい．一方で，非線形次元圧縮手法は，データの非線形的な構造を表現することができるが，計算が複雑で，処理に時間がかかる．今回行う次元圧縮では，ベクトル同士の近さを保持する必要である．ベクトル同士の近さを保持するためには，非線形次元圧縮を用いる必要がある．そこで本研究での次元圧縮手法には UniformManifold Approximation and Projection of DimensionReduction（UMAP）を用いる．

UMAP

UMAP は 2018 年に，Leland Mclnnes，John Hecht，James Melville によって提案された手法である [12]．従来の非線形次元圧縮手法である t-SNE よりも実行時間が高速であり，圧縮後の情報保持力が高いという特徴がある．また，4 次元以上にも圧縮することが可能である．UMAP は高次元空間での近いデータ同士を低次元空間でも近く，異なる点同士を遠ざけるという処理を行うこと，で高次元でのデータ同士の近さや遠さを低次元でも表現できるようにしている．

高次元空間におけるデータ同士の近さは以下のように定義され，3.9 によって対称化される [13].

< 重み付き k 近傍（高次元）>

$$v_{j|i} = exp(\frac{-(r_{ij} - \rho_i)}{\sigma}) \tag{3.7}$$

$$\rho_i = min_{j \in K}\{r_{ij}\} \tag{3.8}$$

< 対称化 >

$$v_{ij} = (v_{j|i} + v_{i|j}) - v_{j|i}v_{i|j} \tag{3.9}$$

低次元空間におけるデータ同士の近さは以下のように定義され，3.11 によって対称化される．

< 重み付き k 近傍（低次元）>

$$w_{ij} = exp(-max\{0, d_{ij} - \rho'\}) = \tilde{w}_{ij} \tag{3.10}$$

< 対称化 >

$$w_{ij} = \frac{1}{1 + a \cdot d_{ij}^{2 \cdot b}} \tag{3.11}$$

トポロジカル表現の最適化によって高次元空間での近さと低次元空間での近さができるだけ同じになるように最適化を行う．

< トポロジカル表現の最適化 >

$$L = \sum [v_{ij} log \frac{v_{ij}}{w_{ij}} + (1 - v_{ij}) log \frac{1 - v_{ij}}{1 - w_{ij}}] \tag{3.12}$$

K-means

次元圧縮を行ったデータに対してクラスタリングを行い，潜在的なグループ化を行う．クラスターの重心を求める必要があるためクラスタリングには k-means を用いる [?].

n 個の個体を $\vec{x_i} = (x_{i1}, \ldots, x_{iD}), i = 1, \ldots, n$ で表し，この個体の集合を X とする．データを K 個の重なりの無いクラス $X_k, k = 1, \ldots, K$ に分類するため，次の目的関数を用いる．

$$J = \min_{\{\vec{c_k},\, k=1,\cdots,\, K\}} \sum_{i=1}^{n} \sum_{\vec{x} \in X_k} \|\vec{x_i} - \vec{c_k}\|^2 \tag{3.13}$$

ここで，$\|\vec{x_i} - \vec{c_k}\|^2 = \sum_{d=1}^{D}(x_{id} - c_{kd})^2$ とし，クラスター中心は $\vec{c_k} = (c_{k1}, \ldots, c_{kD}$ である．式について最小化を行うため，K-Menas アルゴリズムと呼ばれる以下の反復アルゴリズムを使用する．

1. $\vec{c_k}^t$ が与えられたとき，それぞれの $\vec{x_i}$ に関して次式を計算する．

$$a = \underset{k}{argmin} \|\vec{x_i} - \vec{c_k}^{(t)}\|^2 \tag{3.14}$$

2. $X_k^{(t)}$ が与えられたとき，次式を計算する．

$$\vec{c_k}^{t+1} = \frac{1}{n_k^{(t)}} \sum_{\vec{x_i} \in X_k^{(t)}} \vec{x_i}, k = 1, \ldots, K \tag{3.15}$$

ここで，$n_k^{(t)}$ は $X_k^{(t)}$ に属する個体の数であり，$\vec{c_k}^{t+1}$ は $X_k^{(t)}$ の $(k+1)$ 回目のクラスター中心でありクラスターの代表点に相当する．
ある小さい定数を ϵ とし，すべての k について終了条件である $\|\vec{c_k}^{t+1}\vec{c_k}^{t}\| < \epsilon$ を満たすまで，(1) と (2) の手順を繰り返す．

シルエット分析

K-means を行うには初めにクラスター数を与える必要があるため，シルエット分析を用いて最適なクラスター数を決定する．シルエット分析はクラスタ内は密に凝集されているほど良い．異なるクラスタは花らているほど良い．この二つをもとに最適なクラスタ数を求める．各データサンプル $x^{(i)}$ に関して，以下のようにシルエット係数を求める [15]．

＜凝縮度＞

$$a^{(i)} = \frac{1}{|C_{in} - 1} \sum_{x^{(j)} \in C_{in}} \|x^{(i)} - x^{(j)}\| \tag{3.16}$$

＜乖離度＞

$$b^{(i)} = \frac{1}{|C_{near}|} \sum_{x^{(j)} \in C_{near}} \|x^{(i)} - x^{(j)}\| \tag{3.17}$$

＜シルエット係数＞

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{max(a^{(i)}, b^{(i)}} \tag{3.18}$$

§ 3.3　単語間のつながりと共起語ネットワーク

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．

Jaccard 係数

2 つの集合の類似度を測る指標で，以下のように定式化される．

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.19}$$

共起語ネットワークを 3D グラフと 2D グラフによって可視化を行う．3D グラフと 2D グラフにはそれぞれメリットデメリットが存在する．

3D グラフのメリット，デメリット

- 単語の共起関係を 3 次元で表現できるため 2D グラフに比べて表現できる情報量が多い．

- 情報量の多さや 3 次元空間であることから，視認性が 2D グラフに比べて悪い

これらのことを踏まえ，3D グラフと 2D グラフの両方で共起語ネットワークを表示できるようにした．3D グラフの作成には Three.js，2D グラフの作成には Graphvis を用いた．

Three.js

Three.js はウェブブラウザ上で 3 次元コンピュータグラフィックスを描画すための JavaScript ライブラリである．HTML5 の規格に従っており，プラグイン不要で利用することができる．また，WebGL という 3D グラフィックス API をラッピングしており，簡素なコードで 3DCG を描画することができる．3 次元コンピュータグラフィックスとは，3 次元の立体的な仮想物体を，コンピュータで演算することで平面上に奥行きや質感のある画像を表す手法である．従来は大型計算機が必要であったが，プロセッサの性能向上と GPU の一般化により，物性シミュレーションや 3D ゲームなど，さまざまな分野で利用されている [16]．
描画には”3D Force-Directed Graph”というモジュールを用いており，Json ファイルでデータを与えることで，有向グラフを作成することができる．

第4章

提案手法

§ 4.1 Google Patentsからの取得，分類，抽出

Google Patentsから特許情報を取得する．まず特許番号を取得し，それらの番号を用いて特許が表示されているページにアクセスし，そこから特許本文をテキストとして取得する．ユーザーが指定したキーワードのand検索をGoogle Patentsで行う．Google Patentsにおいてand検索を行うにはワードとワードの間にスペースを開ける必要がある．Google Patentsでは一度に1000件までしか表示することができない．そのため，それぞれが1000件を超えないように年代を1年ごと区切ってスクレピングを行う．特許の出願日がその年の1月1日から12月31日である特許を取得した．pythonのモジュールであるthreadsを使用してマルチスレッドでスクレイピングを行った．threadsでは，それぞれ1年ごとに分けられたスクレピングの処理を分割し，それらを平行に処理する．

特許情報の量が少なすぎる又は多すぎる場合が存在するため6年単位で取得する年数を指定できるようにする．

本研究の提案手法は，大別すると以下のような4つの工程からなる．

1. 利用者の入力したワードにおけるGooglePatentsでの検索結果を取得する．
2. 取得した特許データの本文に対してSentence-BERTを用いてベクトル化を行う．
3. 出力されたベクトルに対して次元圧縮を行う．
4. 次元圧縮を行ったデータに対してシルエット分析を行いクラスタリングする．
5. それぞれのクラスターについて，K-meansで求めた重心から距離の近い10個のデータを用いて各クラスターのタイトルを作成する．
6. ユーザーが指定したクラスターに対して，Jaccard係数を用いて共起関係を導出する．
7. 求めたJaccard係数をもとに3Dグラフおよび2Dグラフを作成する．

§ 4.2　トピック推定からの3Dグラフによる可視化

§ 4.3　IPL(Intellectual Property Landscape)への活用

第5章

数値実験並びに考察

§ 5.1　数値実験の概要

§ 5.2　実験結果と考察

第6章

おわりに

謝辞

本研究を遂行するにあたり，多大なご指導と終始懇切丁寧なご鞭撻を賜った富山県立大学工学部電子・情報工学科情報基盤工学講座の奥原浩之教授，António Oliveira Nzinga René 講師に深甚な謝意を表します．最後になりましたが，多大な協力をしていただいた研究室の同輩諸氏に感謝致します．

2024 年 2 月

平井　遥斗

参考文献

[1] 特許庁, ”広報誌「とっきょ」”, 閲覧日 2024-02-04, https://www.jpo.go.jp/news/koho/kohoshi/.

[2] 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”, 閲覧日 2024-02-04, https://www.jpo.go.jp/support/general/document/chizaijobobunseki-report/chizai-jobobunseki-report.pdf.

[3] 東京知的財産総合センター, ”中小企業経営者のための知的財産戦略マニュアル”, 閲覧日 2024-02-04, https://www.tokyo-kosha.or.jp/chizai/manual/senryaku/rmepal000001vypy-att/senryaku_all_vol.9.pdf.

[4] 特許庁, ”経営戦略を成功に導く知財戦略”, 閲覧日 2024-02-04, https://www.jpo.go.jp/support/example/document/chizai_senryaku_2020/all.pdf.

[5] 特許庁, ”「経営戦略に資する知財情報分析・活用に関する調査研究」について”, 閲覧日 2024-02-04, https://www.jpo.go.jp/support/general/chizai-jobobunseki-report.html.

[6] 金融ナビ, ”経営戦略の策定に役立つフレームワーク 7 つ｜経営戦略の代表例も解説”, 閲覧日 2024-02-04, https://financenavi.jp/basic-knowledge/management_strategy_framework/#tag1.

[7] gikyo.jp, ”Perl による自然言語処理入門”, 閲覧日 2024-02-04, https://gihyo.jp/dev/serial/01/perl-hackers-hub/0031011.

[8] 特許庁, ”2019 年度_知的財産権制度入門”, 閲覧日 2024-02-04, https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/2019_syosinsya/1_3.pdf.

[9] 株式会社　日立ソリューションズ・クリエイト, ”テキストマイニングとは？ 手法や活用法を解説”, 閲覧日 2024-02-04, https://www.hitachi-solutions-create.co.jp/column/technology/text-mining.html.

[10] AGIRobots Blog, ”【Transformer の基礎】Multi-Head Attention の仕組み”, 閲覧日 2024-02-04, https://developers.agirobots.com/jp/multi-head-attention/.

[11] Nils Reimers，Iryna Gurevych．”Sentence-BERT：Sentence Embedding using Siamese BERT-Networks”, ArXiv e-prints, 1908. 10084, 2019

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

[13] Hatena Blog, ”UMAP の仕組み–低次元化の理屈を理解してみる”, 閲覧日 2024-02-04, https://kntty.hateblo.jp/entry/2020/12/14/070022.

[14] 倉橋　和子, ”分割・併合機能を有する K-Means アルゴリズムによるクラスタリング”, 奈良女子大学学位論文　2007

[15] Technical Note, ”シルエット分析”, 閲覧日 2024-02-04, https://hkawabata.github.io/technical-note/note/ML/Evaluation/silhouette-analysis.html.

[16] アンドエンジニア, ”Three.js とは？概要やできることを JavaScript 関連術を含めて解説”, 閲覧日 2024-02-04, https://and-engineer.com/articles/ZOWitBIAACMAFtEj.