機械学習を用いた効率的な特許調査方法
― AI 調査ツール活用のためのデータサイエンスベースの特許調査―

Effective patent search methods using Machine Learning

花王株式会社　研究開発部門　研究戦略・企画部／アジア特許情報研究会
安藤　俊幸

1985 年現花王株式会社入社、研究開発に従事
1999 年研究所の特許調査担当（新規プロジェクト）、2009 年知的財産部、2021 年より現職
2011 年よりアジア特許情報研究会所属
2020 年特許情報普及活動功労者表彰 日本特許情報機構理事長賞「技術研究功労者」受賞
情報科学技術協会、人工知能学会、データサイエンティスト協会　各会員

ando.t@kao.com

1　はじめに

最近では知財情報業務への人工知能（AI:Artificial Intelligence）の適用も身近な存在になってきている。商用の AI を利用した特許調査ツールも複数登場[1]している。ただ、これら商用の AI 調査ツールをユーザーが使いこなす上で押さえておくべき基本事項や限界・課題も多いのも現実である。本稿では、特許調査おける AI 活用のためにデータサイエンスベースの特許調査について、調査システムのユーザーの立場として述べる。

研究開発中の AI 全般の動向としてガートナーの先進テクノロジのハイプ・サイクル見ると「人工知能」は 2018 年には「過度な期待度のピーク期」を越え、2019 年に、『人工知能』は、幻滅期に位置付けられている。ここで「ピーク期とは最も良い状態」あるいは「幻滅期は悪い状態」という文字通りの意味ではない。ピーク期は「過度な期待」によって理想と現実にギャップがある状態のことである。幻滅期は「冷静な判断」を行う時期で、「本物と偽物の区別」が行われるのもこの時期とされている。2020 年版では「人工知能」関連技術が 11 技術に細分化されている。ハイプ・サイクルの 2022 年版を図 1 に示す。AI 自動化の加速をサポートするテクノロジとして、「オートノミック・システム、コーザル AI、ファウンデーション・モデル、ジェネレーティブ・デザイン AI、機械学習コード生成」が挙げられている。

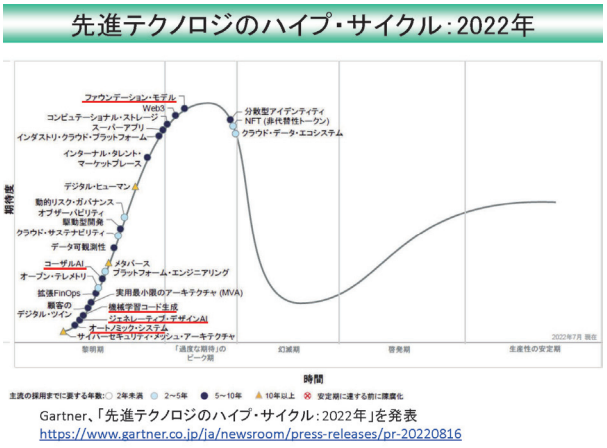


図 1　先進テクノロジーのハイプサイクル 2022 年

2017 年の「情報の科学と技術」誌の特集：特許情報と人工知能（AI）[2]に桐山勉、安藤俊幸で総論を執筆した。そこから結論部分を一部抜粋して、2022 年 4 月時点の現状と比較して 5 年前の予測を振り返り現状と更に今後の進むべき道を考察[3]した。5 年前の予測は、『人工知能の特許情報への活用は、特許情報の深層学習機械翻訳と深層学習特許検索と深層学習特許分類付与の 3 点が主要なものと考える。開発研究には費用と投資が先行することも現実である。そのため、他の業界分野に比べると特許情報への活用は、資金力と人材力が乏しく、AI の活用・応用研究が遅れている。』であった。

人工知能の特許情報への活用については、2022 年時点で、特許情報の深層学習機械翻訳は、訳抜けの問題等一部課題は残るが実用化されており、外国特許のスクリーニング等には大変便利である。深層学習を応用した

特許検索と特許文書の分類機能を備える商用のシステムも複数登場しているが、使用する上で様々な課題も散見され、所謂 AI 検索に関して後程、詳述する。AI 分類に関しては 2021 年の本寄稿論文[4] で SDI 調査における 2 値分類に関して具体的に紹介した。

2 データサイエンスベースの特許調査

本稿のメインテーマである、AI 調査ツール活用のためのデータサイエンスベースの特許調査のプロセスを図 2 に示す。

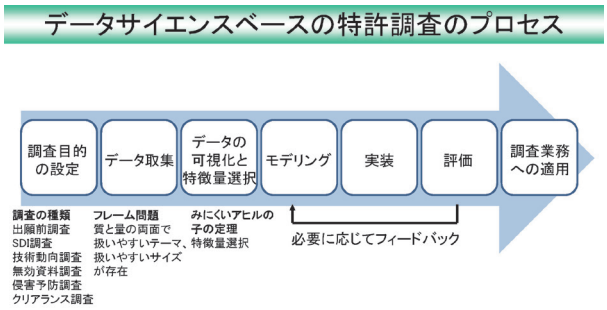


図 2　データサイエンスベースの特許調査のプロセス

各プロセス毎の留意点を述べる。

調査目的の設定

調査目的の設定にはなるべく具体的に整理された情報要求が重要である。情報要求とは、直面する問題を解決するために必要な情報を入手したいとする欲求のことである。調査の種類によって押えるべきポイントがことなる。

データ収集／整形時の注意点について

機械学習を用いた学習では、ある程度の量がまとまったデータが必要となる。「データ」といっても目的によって収集すべきデータはさまざまである。調査目的によって必要なデータを見極め、それらをどのようにして集めるかを理解し実行することが重要である。

以下にデータ収集／整形時の注意点について述べる。有効なデータの見極め方下記項目に留意する。

①データの信憑性

示されたデータが、信じるに値する科学的根拠となり得るかどうかを判定する。信頼できる機関によって取得・提供されたものかどうか、集計・分析済のものであればその方法は正しいかどうかなどを確認する必要がある。日本特許庁提供の書誌、特許公報、経過情報等のデータの品質は非常に高いレベルにあるが完璧ではない。また日本特許庁のデータもデジタル化公報（1993 年）より前の公報データには OCR の誤りや欠損値がある。残念ながら US、EP、CN 等の主要国のデータ品質は日本のレベルに達していないものも多い。新興国のデータ品質は更に各種問題を含んでいることに注意する必要がある。データ品質は調査目的にも依存する。

②データの量

収集されたデータを使って、全体の傾向を把握したり、示唆を得る上で、データ量が十分かどうかを判断する必要がある。データは多ければ多いほど良いというわけではないが、あまりにデータが少なすぎると、母集団に対する推定に誤差が生じ、解釈やその後の意思決定が事実にそぐわない危険性がある。

③データの偏り（バイアス）

データを収集する際は、収集の対象となる母集団について、標本の大きさや標本の抽出方法を選択する必要がある。これらを考慮せずにデータを収集し機械学習を行った場合、不正確な予測結果を招いてしまう原因となる。取得方法や対象が恣意的に選定されていないかなどの確認が必要である。

④欠損データ

何らかの理由により取得ができなかったデータが存在する場合を指す。データは常に全ての項目が揃っているとは限らない。精度の高い学習・予測を行うためにはデータが揃っていることが望ましいが、機械学習のアルゴリズムでは、こうした欠損値に対して補完をする手法も存在する。

データの可視化と特徴量選択

データの可視化

データの可視化は目的により各種存在する。特許調査データベースに組み込まれているものもある。またパテントマップソフトやテキストマイニングツールには各種の可視化手法があらかじめ用意されている。

特徴量選択

特徴量選択（feature selection）はデータサイエンスにおいて非常に重要なプロセスである。特徴量選択とは、機械学習のモデルを使用する際に有効な特徴量の組

み合わせを探索するプロセスのことを表している。

モデリング

モデリングとは、収集したデータをコンピュータが処理できる形式でコンピュータに入力し、機械学習のプログラムを実行して結果を出力するプロセスを指す。「モデルを作る」「モデリングを実行する」などとも呼ばれている。

教師あり機械学習のモデリング

教師あり機械学習におけるモデリングの成果は、機械が発見した法則やパターンから識別や予測を行う数理アルゴリズムである。これに新たに未知のデータを投入することにより、その性能を検証することができる。「予測モデル」「識別モデル」など、機械学習の目的に応じて呼びわけられることもある。

教師なし機械学習におけるモデリングの成果は、実行結果そのものである。データを似ているもの者同士にクラスタリングした分割数や、分かれ度合いなどを人間が解釈する。

実装

実装に関しては別途詳細に述べる。

評価

モデルができ上がったら、そのモデルが調査業務に適用できそうなものかどうか確認するために評価を行う。具体的には収集したデータを２つに分割、一方のデータ群でモデリングを行い、もう一方のデータ群でモデルの評価を行う。データ量が不十分な場合はホールドアウト法やｋ分割法を用いてモデリングと評価を行う。ホールドアウト法とは、データを「学習用データ」と「テストデータ」に分割して、モデルの精度を確かめていく手法のことである。

評価とは簡単に言えば「作ったモデルが実際の調査業務へ適用可能なものであるか判断すること」である。いかなる手法においてもモデルの評価をすると、その結果の数値が出力される。評価結果は自動的に出力されるが、その数値を見て適用できるかどうかを判断するのは人間である。

どの手法においても予測と実際の差が少なかったり、判別した結果が合っている方が適合しているモデルということになるが、手法や取り組んでいる課題によって適用できるかどうかの閾値は異なってくる。

3 特許調査用学習済モデルの作成と評価

特許調査用学習済モデルの作成とその評価方法を図３に示す。上側が特許調査用学習済モデルの作成の概要である。下側が学習済モデルの利用と評価の概要である。

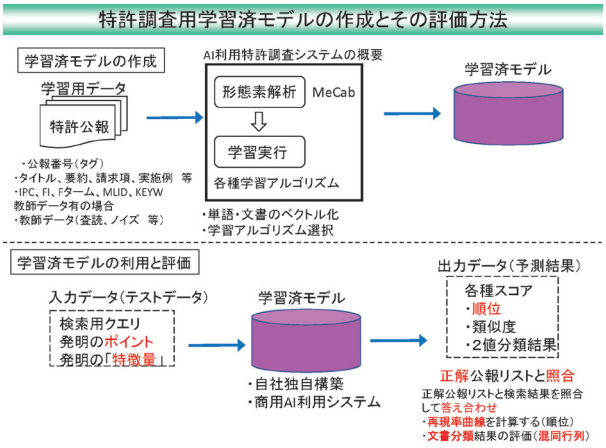


図３　特許調査用学習済モデルの作成とその評価方法

文書分類結果の性能評価に有効な混同行列を図４に示す。

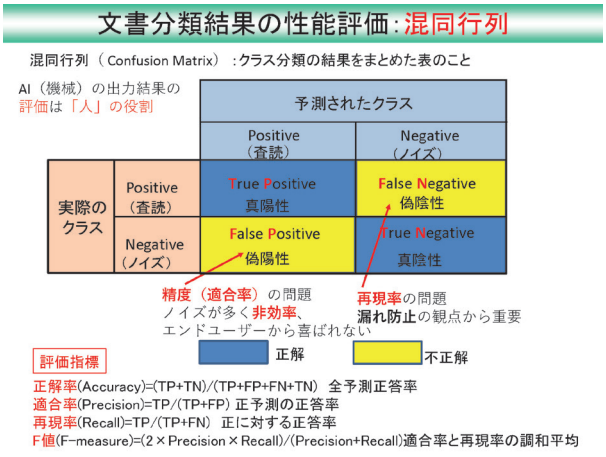


図４　文書分類結果の性能評価：混同行列

混同行列（Confusion Matrix）とは２値分類問題で出力されたクラス分類の結果をまとめたマトリックス（行列）のことで、２値分類機械学習モデルの性能を測る指標として使われる。混同行列から図４の４種類の評価指標を計算できる。

偽陽性は精度（適合率）の問題であり、ノイズが多く非効率、エンドユーザーから喜ばれない。

偽陰性は再現率の問題であり、漏れ防止の観点から重要である。

文書分類モデルの性能評価に使われる交差検証法を図

5 に示す。図は 8 分割の例である。

K- 分割交差検証 Cross Validation（クロスバリデーション法）

データ全体を K 個に分割した上で、そのうちのひとつをテストデータとし、残った K-1 個を訓練用データに分解していくテストデータと学習用データの入れ替えを行いながら繰り返し、すべてのケースがテスト事例となるまで検証を行っていく。つまり、K 個に分割されたデータは、K 回の検証が行われることになる。

図 5　K 分割交差検証

交差検証とは、統計学において標本データを分割し、その一部をまず解析して、残る部分でその解析のテストを行い、解析自身の妥当性の検証・確認に当てる手法を指す。データの解析（および導出された推定・統計的予測）がどれだけ本当に母集団に対処できるかを良い近似で検証・確認するための手法である。

4　特許調査への機械学習適応時の留意点

筆者が考える、機械学習の特許調査への応用時の 3 要素を図 6 に示す。

3 要素を統合して「調査目的に合わせたアルゴリズムとドメインデータの選択と最適化を行い学習済モデルを作成・利用する」のは人間知能が主導して行うべきと考える。調査目的の各調査の矢印の向きは精度と再現率のどちらを指向しているかを定性的に示したもので実際の個々の調査ではケースバイケースである。アルゴリズムは AI 調査ツールの寄与が大きい。ドメインデータは公報データベースの寄与が大きい。図 6 の「ドメインデータ：調査対象分野のデータ」は図では小さな面積しか占めておらず目立たないが特許調査では非常に重要である。いわゆる特許公報データベースでありこのデータの内容・品質が特許調査に直接影響する。例えば収録されていない国の調査はできない。日本の特許データは日本特許庁より非常に綺麗に整備された状態で提供されるのでデータベースベンダーの違いはあまり感じないが海外のデータに関しては注意深く調べるとかなり差がある。また分野固有のデータを独自に検索できるようにしているシステムもある。例えば化学構造検索や化学物質名[5,6]、DNA、アミノ酸配列検索等である。これらを使いこなそうとすると現状では経験を積んだプロのサーチャーや研究者のスキルが必要となる。

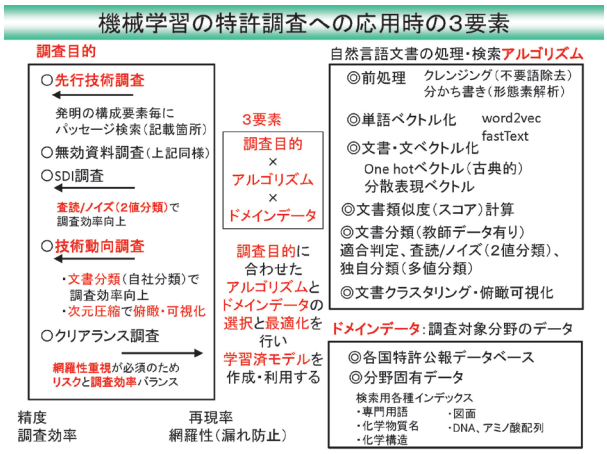


図 6　機械学習の特許調査への応用時の 3 要素

特許調査への人工知能適用時の留意点として人工知能分野の原理的な難問から実務上の留意点まで簡単に列記する。

（1）シンボルグランディング（記号接地）問題

シンボルグラウンディング問題とは、記号システム内のシンボルがどのようにして実世界の意味と結びつけられるかという問題。記号接地問題とも言う。現在の「AI」は人間と同じように自然言語を理解しているわけではないことに注意する必要がある。

（2）ノーフリーランチ（NFL）定理

最適化問題であらゆる問題に適用できる性能の良い万能のアルゴリズムは無いという意味である。ある特定の問題に焦点を合わせた専用アルゴリズムの方が性能が良いということである。現状は汎用の AI（強い AI）は無く、特定の問題に強い専用の AI（弱い AI）が多いことと関係している。この定理の名前の由来は「無料の昼食は無い」というところからきている。酒場の広告で「ドリンク注文で昼食無料」というのがあったが実際は「ドリンクに昼食料金が含まれている」ということでハインライ

ンの SF 小説『月は無慈悲な夜の女王』（1966 年）で有名になった格言に由来している。この定理の数学的な意味も重要であるが名前の由来になった格言の意味も実際の AI 製品の広告やパンフレットを吟味する場合重要である。特に「AI を導入するとなんでも／簡単にできる」という意味のフレーズには要注意である。「なんでもできる＝万能のアルゴリズム」は無い。「簡単にできる＝無料の昼食」は本当に無料なのか、特に教師あり機械学習において教師データを用意したり、機械学習の出力結果を判定／検証するコストを考慮しているのか要チェックである。

（3）フレーム問題

フレーム問題とは、人工知能における重要な難問の一つで、有限の情報処理能力しかないロボットには、現実に起こりうる問題全てに対処することができないことを示すものである。特許調査や学術文献調査等の検索においてどこまで調査するのか調査範囲を決める外枠と考えると理解しやすい。特許調査においては調査目的に応じてどこまで調べるか調査範囲を決めておくとフレーム問題を回避あるいは軽減できる可能性がある。もう少し具体的には発明を特許出願する前に行う先行技術調査では発明に新規性、進歩性があるか調査するがその発明が属する技術範囲を適切に決めると調査が効率的に行える。調査対象国により IPC、CPC、FI 等を適切に使い分ける、あるいは併用すると良い。日本特許の場合は FI、F タームを利用すると調査精度を高めることができる。

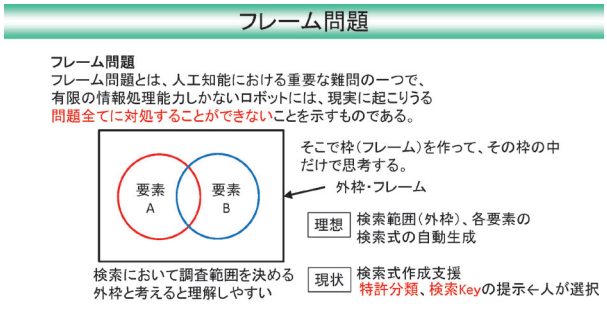


図 7　フレーム問題

（4）過学習（汎化性能）

過学習（overtraining）とは、機械学習において、訓練データに対して学習されているが、未知データ（テストデータ）に対しては適合できていない、汎化できていない状態を指す。汎化能力の不足に起因する。その原因の一つとして、統計モデルへの適合の媒介変数が多すぎる等、訓練データの個数に比べて、モデルが複雑で自由度が高すぎることがある。不合理で誤ったモデルは、入手可能なデータに比較して複雑すぎる場合、完全に適合することがある。過学習は機械学習の実務上、細心の注意が必要である。

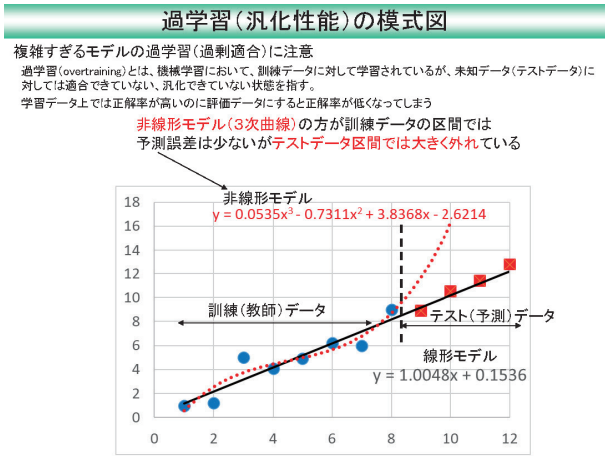
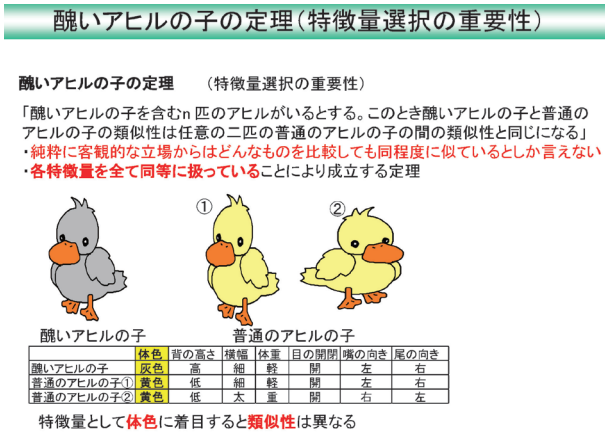


図 8　過学習

（5）特徴量選択（醜いアヒルの子の定理）

醜いアヒルの子の定理とは、純粋に客観的な立場からはどんなものを比較しても同程度に似ているとしか言えない、という定理である。特徴量を全て同等に扱っていることにより成立する定理で特徴量選択の重要性を示している。もう少し具体的には醜いアヒルの子（白鳥の雛で灰色）、普通のアヒルの子（黄色）の特徴量（灰色、黄色）に着目すれば識別可能だが識別に無関係の特徴量を増やすと区別できなくなる。



	体色	首の高さ	横幅	体重	目の開閉	嘴の向き	尾の向き
醜いアヒルの子	灰色	高	細	軽	開	左	右
普通のアヒルの子①	黄色	低	細	軽	開	左	右
普通のアヒルの子②	黄色	低	太	重	開	右	左

図 9　醜いアヒルの子の定理

商用の AI 調査ツールを導入する場合は上記五つの留意点を考慮されているかの観点も検討すると良い。上記

五つの留意点を踏まえて特許調査のプロセスに適合したアルゴリズムを選択して、組み合わせて、実務を想定した各種データで実験し、チューニングすることにより、より良い出力（予測結果）を期待できる。

AI特許調査ツール利用時の注意点まとめ

AIツール利用時の注意点	具体的な対応方法
(1)シンボルグランディング（記号接地）問題 記号システム内のシンボルがどのようにして実世界の意味と結びつけられるかという問題	・現在の「AI」は人間と同じように自然言語を理解しているわけではない ・ユーザー側でドラえもん（過大評価）を期待していないか→適切な評価が必要
(2)ノーフリーランチ（NFL）定理 ①万能のアルゴリズムはない ②無料の昼食はない→常にコストがかかる	・AIのアルゴリズムに注意が必要 →それぞれ特徴がある ・常に様々な観点からのコスト意識が必要
(3)フレーム問題 有限の情報処理能力しかないAIにどこまで処理させるか、処理対象範囲を限定する必要	・調査目的に応じて調査範囲を限定 ・DBとしての収録範囲の確認　等々
(4)過学習（汎化性能） 訓練データに対して学習されているが、未知データ（テストデータ）に対しては適合できていない、汎化できていない	・教師有機械学習使用時には常について回る→都度、留意が必要
(5)特徴量選択（醜いアヒルの子の定理） 特徴量選択の重要性を示している	・システムが提示する類似度と人が感じる類似の違いに注意

図 10　AI 特許調査ツール利用時の留意点まとめ

教師有り機械学習を使用する上で 2 つのトレードオフの関係に注意が必要である。一つ目は適合率（精度）と再現率のトレードオフである。二つ目はバイアス（偏り）とバリアンス（分散）のトレードオフである。

バイアスとバリアンスのトレードオフ（Bias-Variance Tradeoff）とは、機械学習モデルによる予測において汎化誤差を最小化させるために、偏り誤差を小さくするとバラツキ（分散）誤差が大きくなり、逆にバラツキ誤差を小さくすると偏り誤差が大きくなるという、両者のトレードオフの関係性を示す。

機械学習（や統計学）のモデルによる予測においてバイアス（偏り：Bias）とは、予測値と真の値（正解値）とのズレ（つまり「偏り誤差：Bias error」）を指す。この予測誤差は、モデルの仮定に誤りがあることから生じる。またバリアンス（分散：Variance）とは、予測値の広がり（つまり「ばらつき誤差:Varianceerror」）を指す。この予測誤差は、訓練データの揺らぎから生じる。

モデルによる予測においてバイアス（偏り誤差）が大きすぎる場合、そのモデルは入力と出力の関係性を正確に表現できていない（訓練データでさえも正確に予測できない）といえる。いわゆる「学習不足（過少適合：under-fitting)」の状態である。

またモデルの予測においてバリアンス(ばらつき誤差)が大きすぎる場合、そのモデルは訓練データのノイズまで学習してしまっている（テストデータなど未知のデータでは正確に予測できない)。いわゆる「過学習（過剰適合：over-fitting)」の状態である。ターゲットの真ん中を正解としたときのバイアスとバリアンスのイメージを図 11 に示す。

機械学習に置けるバイアスとバリアンスのイメージ図

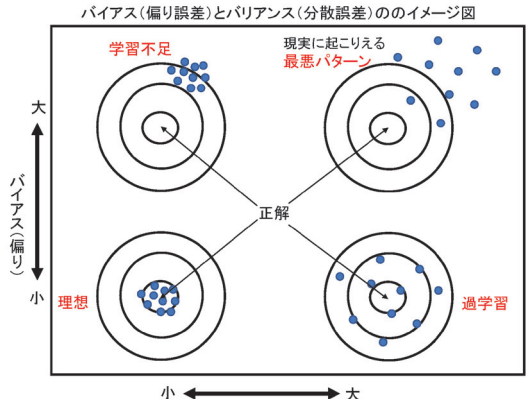


図 11　バイアスとバリアンスのイメージ図

特許調査との関係でバイアスとバリアンスを考えることも重要である。例えば特許調査範囲を表す特許分類の IPC や FI の選定を誤るとバイアスが大きくなると考えられる。発明の特定に重要なキーワードの類義語を不必要に多く検索式に含めるとバリアンスが大きくなると考えられる。

5　特許調査システムの検証と評価方法

特許調査システムの検証と評価方法として、主な使用目的と齟齬が無ければ再現率と適合率での評価からスタートすると良い。特許調査における基本的な検索性能の評価指標として図 12 に再現率（網羅性）と適合率（精度）の求め方を示す。

特許調査における再現率（網羅性）と適合率（精度：効率）

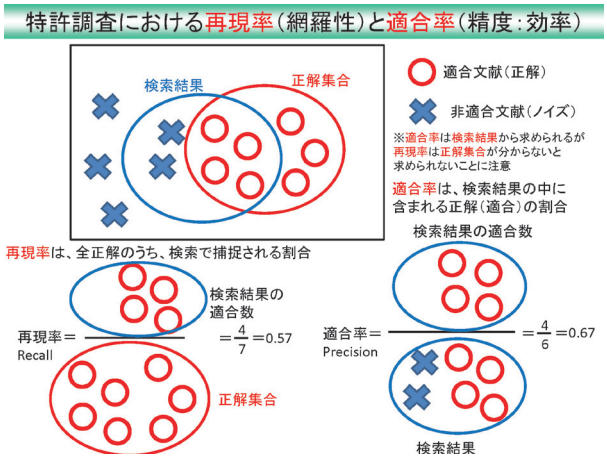


図 12　特許調査における再現率と適合率

適合率（精度）は、検索結果の中に含まれる正解（適合）の割合である。再現率は、全正解のうち、検索で捕捉される割合である。適合率は検索結果から求められるが再現率は正解集合が分からないと求められないことに注意する必要がある。

一般化した特許調査システムとその評価方法を図 13 に示す。中央の長方形内は特許調査システムの概念図である。一般的に内部はブラックボックスであるが利用しているアルゴリズムをマニュアルやウェビナー等で公開している場合や問い合わせるとある程度教えてくれるベンダーも存在する。検索モデルに関しては「完全一致」のブーリアン型は入力（検索用クエリ）と出力（検索結果）の関係は理解しやすい。何らかの類似度を使用する「最良一致」（例えば後述する Patentfield のセマンティック検索）の検索モデルではユーザーが検索結果の理由を明示的に理解することは困難である。

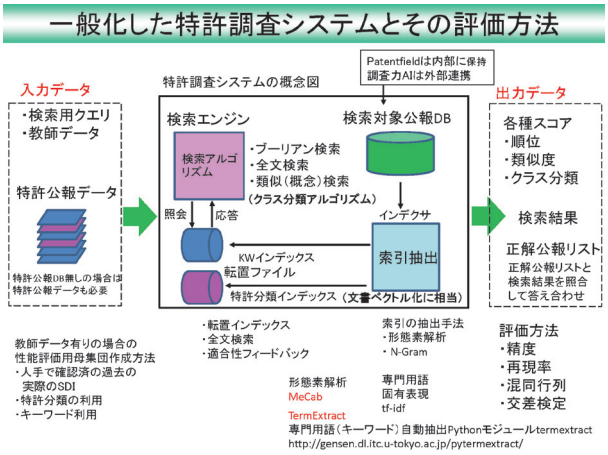


図 13　一般化した特許調査システムとその評価方法

入力に関しては何らかの類似度を用いた検索の場合は「発明の特徴を表す文章、あるいは一つ以上のキーワード」をクエリとして入力するのが基本である。入力が教師データ有りの場合、出力はクラス分類結果である。入力に対して類似の公報を求める場合の出力はスコア（主に類似度）による順位付きの文書リストである。クラス分類の評価方法としては混同行列が用いられる。

クエリ文と類似の順番にスコア付けしてソートするために特許調査ツールの内部では様々な「類似度」が使用されている。スコアの名称も類似度ではなくツール固有の命名がされている場合もある。一口に「類似度」と言ってもテキストに含まれる単語を用いて計算する時に、単語の集合間の類似度を計算する、Jaccard 係数、Dice 係数、Simpson 係数がある。テキストに含まれる単語の重要度を tf-idf を用いて重み付けして計算するコサイン類似度がある [7]。Word2Vec[8] による単語の分散表現を用いたテキスト間の類似度計算方法が多数提案されている。テキストに含まれる単語の Word2Vec による単語の分散表現ベクトルの平均を求める Ave.-Word2Vec、SCDV[9]（sparse composite document vectors）等がある。Paragraph Vector[10] を実装した Doc2Vec[11] や、SWEM[12] という，単純にテキスト中の全単語のベクトルを平均したり、ベクトルの各要素の最大値のみ抽出したりするといった複数の手法が提案されている。SWEM は非常に高速に動作し、比較的良い結果が得られるのでよく使われている。BERT[13] を用いたテキスト間の類似度尺度 BERTScore[14] も提案されている。

筆者も文書単位 [15]、文単位 [16] の類似度計算を用いた先行技術調査への応用を検討した。2017 年、2018 年の Japio YEAR BOOK で紹介している [15),16)]。

6　特許調査への機械学習の導入

図 14 に、オープンソースソフトウェア（Open Source Software）の使用を前提にした「自分でできる」OSS ツールによる構築と「誰でも使える」既成商用ツールの導入の比較を示す。それぞれ特徴をよく把握して使い分けると良い。既成商用ツールの導入は OSS ツールによる構築に比べて使用の障壁は低いとは言え、ツールを使いこなすにはそれなりの経験と練習が必要となる。またシステムにより提示された公報を読んで自分が求めているものかどうかの適合判定はユーザーが行う必要がある。

特許等の調査業務にオープンソースを用いたプログラミングを活用することで調査効率（精度）や網羅性（再現率）の向上が期待できる。特許調査にも先行技術調査、動向調査等各種あり、調査のプロセスも細分化される。どのプロセスのどのようなタスクをプログラミングによって効率化したいのかによりそのタスク処理が得意なプログラミング言語が異なる。

R は統計解析用の言語として広く使われており開発／実行環境は OSS として提供されている。最近では書籍も多数出版され R 言語そのものやテキストマイニング

等の応用事例に関する情報も容易に入手可能である。

Python は汎用のプログラミング言語であり、コードがシンプルで扱いやすく設計されている。データサイエンス、AI の中心技術である機械学習分野でもよく使われている。図 15 に R と Python の基本機能と代表的なライブラリを示す。

「自分でできる」OSSツールと「誰でも使える」既成商用ツールの比較

	「自分でできる」(構築)	「誰でも使える」(導入)
前提条件	OSSツール使用	既成商用ツール使用
内部処理の透明性	高い	通常ブラックボックス
プログラミングスキル	必要	不要
自由度	高い	低い
使用の障壁	高い	低い
初期コスト	低い	高い
ランニングコスト	低い	高い
特徴	・習得に時間がかかる ・使い慣れると非常に強力かつ便利	用意された機能は簡単に使用できる

図 14　OSS ツールと既成商用ツールの比較

RとPythonの基本機能とライブラリ

	R	Python
ベクトル・行列計算	標準(base)	NumPy, SciPy
データフレーム	標準(base)	pandas
基本的な統計機能	標準(stats)	StatsModels
基本的なグラフィックス	標準(graphics)	Matplotlib
拡張グラフィックス	ggplot2	Seaborn
3Dグラフィックス	rglパッケージ	plotly
ネットワーク分析	igraph	NetworkX
機械学習	Caret (Classification And Regression Training)	Scikit-Learn PyCaret

※両者の機能には差異があり、一対一で対応するものではない

図 15　R と Python の基本機能とライブラリ

R の機械学習のパッケージとして Caret（Classification And Regression Training）がある。Caret は構築できる数理モデル（アルゴリズム）が 200 種類を超え、線形回帰モデルから、決定木系、ニューラルネット系と幅広く扱える。Caret には、機械学習に必要な機能が、前処理、データ分割（学習データとテストデータ）、特徴量選択、モデル学習、モデル評価、パラメータチューニング、予測等、一式揃っている。

最近では、Python 版の Caret である PyCaret も開発されている。PyCaret は、機械学習のデータ前処理や可視化、モデル開発等の一連の作業を数行のコードで自動化してくれるオープンソースの機械学習パッケージである。いくつかの主要な機械学習ライブラリ（scikit-learn,XGBoost,LightGBM 等）を Python でラッパーしたものである。分類や回帰、クラスタリング、異常検知、自然言語処理等が扱える。

OSS を用いた特許文書解析として「AI 系基盤技術と、オープンソースを用いた機械学習による特許文書解析」[17] が参考になる。「特許調査のためのプログラム事例紹介」[18] では、図 16 の Doc2Vec を用いて調査対象の特許文書集合を学習し、文書・単語ベクトルを求め、指定文書・単語（クエリ）に対し類似度の高い文書・単語を表示する Python のソースプログラムを紹介している。

Doc2Vecによる文書のベクトル化処理の概要

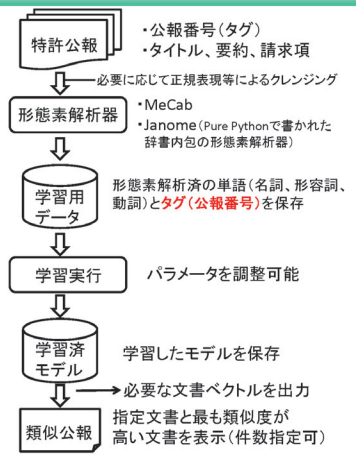


図 16　Doc2Vec による文書のベクトル化処理の概要

図 16 の Doc2Vec による文書のベクトル化処理と指定文書・単語に対し類似度の高い文書・単語を出力する機能を Windows サーバーに Python 実行環境を構築して、所謂 PoC として実装した。PoC（Proof ofConcept：概念実証）とは、新たなアイデアやコンセプトの実現可能性やそれによって得られる効果などについて検証することである。どの部分を自分で実装して、どの部分を他に任せるかもポイントである。サーバー管理やセキュリティ等も含めて全て自分で面倒を見るのはかなりの負担である。自社の専門部署や他社の専門の会社に管理を任せるのも選択の範囲である。

現在は複数の AI 利用ツールが商用サービスとして提供されており選択肢は増えている。多くのツールで期間限定の無償のトライアルも可能であり実際の実務データを使用してのツールの性能評価の敷居も下がっている。

7 知財分野における第４世代 AI の基礎検討

最近、人工知能（AI:Artificial intelligence）の使用を謳っている特許調査システムが商用ベースで複数提供されておりコモディティ化している。ただいろいろと課題も多いのが現状である。また現在の深層学習（第３世代 AI）の限界も指摘され、第４世代 AI が提案[19]されており、特許庁よりニーズ即応型技術動向調査の概要[20]が公表されているが、知財分野における応用はこれからである。

第４世代 AI を念頭に、特許調査と機械学習の補完・融合を目的に基礎検討を行った[21]。

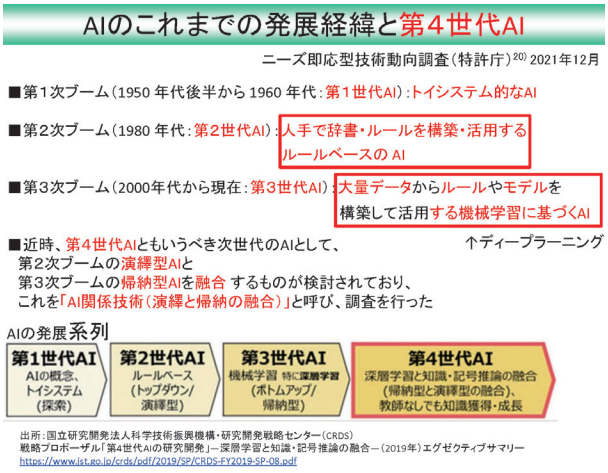


図 17　AI の発展経緯と第四世代 AI

図 17 に AI のこれまでの発展経緯と簡単な特徴、第四世代 AI を示す。第２世代 AI は、人手で辞書・ルールを構築・活用するルールベースの AI である。第３世代 AI は大量データからルールやモデルを構築して活用する機械学習に基づく AI である。

現在の深層学習（第３世代 AI）の限界として下記指摘がされている。

・学習に大量の教師データや計算資源が必要であること
・学習範囲外の状況に弱く、実世界状況への臨機応変な対応ができないこと
・パターン処理は強いが、意味理解・説明等の高次処理はできていないこと等

第４世代 AI は深層学習と知識・記号推論の融合（帰納型と演繹型の融合）、教師なしでも知識獲得・成長が期待できる。

検証対象として特許調査のプロセスと正解公報が詳細に解説されている特許検索競技大会 2021 年過去問[22]の化学・医薬分野の問題（徐放性マイクロニードル）図 18 を選択した。検索結果の性能の指標として正解公報９件（図 19）を使用して、正解率、精度、再現率、正解公報のランキング順位を使用した。

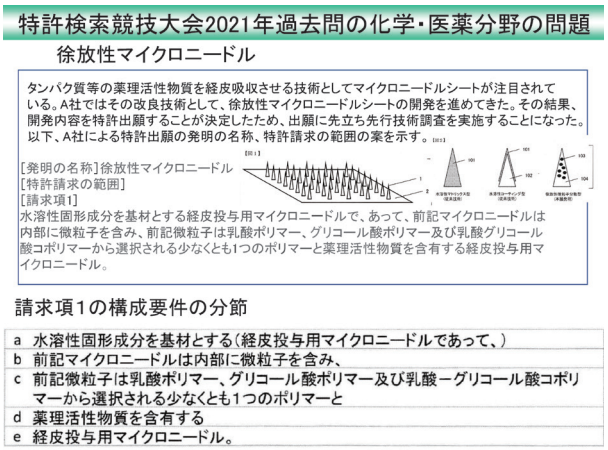


図 18　特許検索競技大会 2021 年の過去問

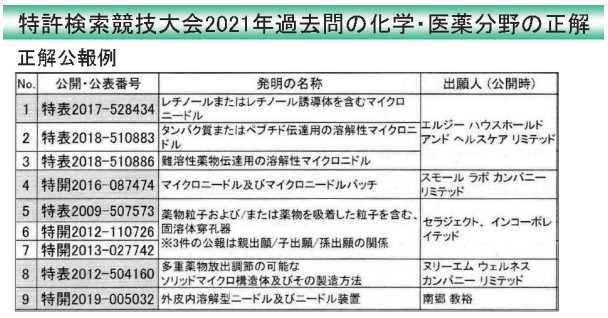


図 19　特許検索競技大会 2021 年の正解公報例

図20に完全一致、最良一致検索モデルの比較をしめす。

図 20 を基に、特許調査時の検索モデルの類型化を試みた。その結果、特許分類、キーワード等をブーリアン演算で使用する「完全一致型」、発明の特徴を表す文を入力する「最良一致型」、最良一致型をさらに従来の概念検索型と AI 検索型に分けた。

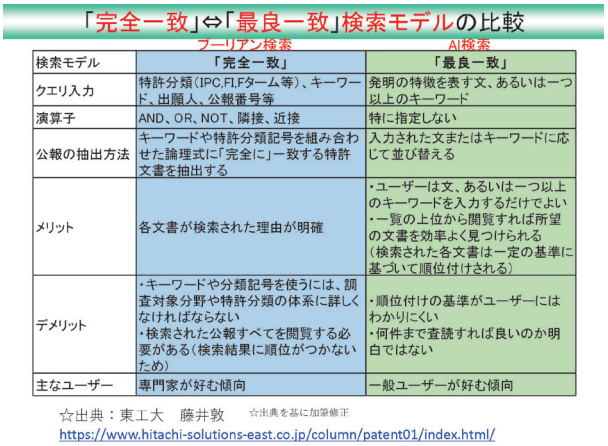


図 20　完全一致、最良一致検索モデルの比較

概念検索型、AI 検索型どちらも内部で使用しているアルゴリズムが分かればさらに細分化することが可能である。概念検索型はベクトル空間法、確率的言語モデルに分けられる。AI 検索型は商用の実際のツールに依存する。

図 21 に特許検索競技大会の模範」解答に記載の IPCC（工業所有権協力センター）推奨の検索方法と機械学習の留意点を右側に記入したものを示す。

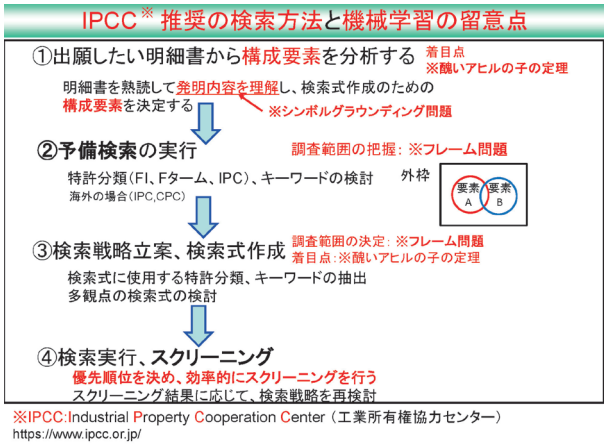


図 21　IPCC 推奨の検索方法と機械学習の留意点

ルールベースの AI と AI 検索の組み合わせの基礎的な検討を図 22 のように 2 段階に分けて計画した。第 1 ステップで人間知能（Human Intelligence）主導で重要特許分類、重要キーワードを抽出し HI 検索 +AI 検索をシュミレーションして検索性能が向上するか確認する。第 2 ステップでルールベースの AI による重要特許分類、重要キーワード抽出ができるか検討する。

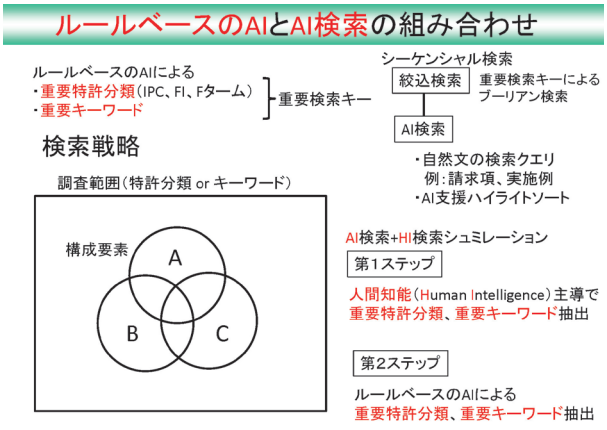


図 22　ルールベースの AI と AI 検索の組み合わせ

Patentfield[23] の検索種別と対象特徴量（対象範囲）を図 23 に一部抜粋して示す。

Patentfield のセマンティック検索の正式名称は AI セマンティック類似検索であり AI 検索に分類する。同じセマンティック検索という名称でも商用データベースによっては概念検索の場合もあり注意を要する。

Patentfieldの検索種別と対象特徴量（対象範囲）

セマンティック検索 SE

セマンティック系特徴量　（分散表現ベクトル）fastText	コマンド
セマンティック検索(名称/要約/請求の範囲/明細書/審査官フリーワード)	SE
セマンティック検索(名称/要約/請求の範囲)	SEC
セマンティック検索(請求の範囲)	SEAC
セマンティック検索(請求の範囲 トップクレーム)	SETC

赤枠はデフォルト設定

全文検索 KW

キーワード系特徴量　（重み付(BM25)KWベクトル）	コマンド
名称/要約/請求の範囲/明細書/審査官フリーワード	KW
名称/要約/請求の範囲	KWC
請求の範囲（出願/付与）	CL
請求の範囲 トップクレーム（出願/付与）	TCL

一部抜粋

図 23　Patentfield の対象特徴量

Patentfield のセマンティック検索は内部では fastText の分散表現ベクトルを使用しており収録されている全分野にわたって類似語を学習している。Patentfield の全文検索は明細書の検索対象を選択してキーワードにより検索できる。セマンティックスコアを使用することで、探したい任意の自然文書または特定の文献から類似度が高い順にスコアリングされ、教師データを用意せずに探したい技術内容に近い特許文献から調査することができる。セマンティックスコアはブーリアン演算によるコマンド検索の母集団に組み合わせることも可能で、任意のキーワード、文書、特許番号などを指定することにより検索結果の母集団を変更することなく、容易に近い内容の技術から効率的に調査を行うことができる。

Patentfield のブーリアン検索× AI セマンティック検索画面を図 24 に示す。

Patentfieldのブーリアン検索×AIセマンティック検索

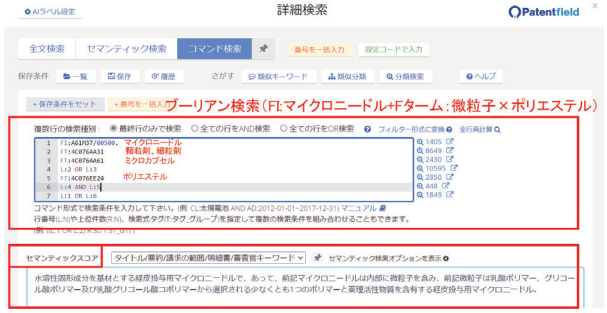


図 24　Patentfield のブーリアン検索× AI セマンティック検索

ブーリアン検索で FI: マイクロニードル +F ターム：

微粒子×ポリエステルの検索を行った。AI セマンティック検索には請求項 1 を使用した。

AI セマンティック検索単独と「ブーリアン検索× AI セマンティック検索」結果を図 25 に示す。「ブーリアン検索× AI セマンティック検索」と組み合わせることで大幅に出現順位が向上している。

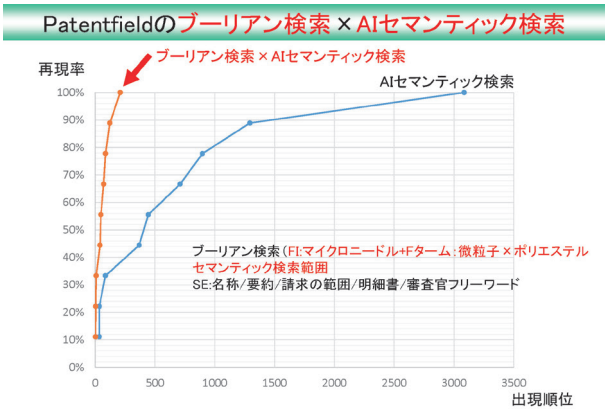


図 25　ブーリアン検索× AI セマンティック検索結果

8 まとめ

特許調査における機械学習利用時の留意点と実務テクニックを特許調査とデータサイエンスの観点から紹介した。

実務上の基本性能として調査目的に応じて下記項目に十分な注意が必要である。

①データ品質　収録（率）、正確性、機械翻訳精度（ベースとなるデータベースの性能）

②検索性能　再現率／適合率

③操作性　ユーザーインターフェイス、ハイライト、絞り込み機能等

AI に向き合う姿勢として、AI に丸投げではなく協調してうまく使いこなすために下記 2 点が重要である。

① AI に「過度な期待」はせず冷静な性能評価と使いこなしが必要である。

②人間知能 HI（Human Intelligence）と人工知能 AI（Artificial Intelligence）の特徴把握と役割分担が重要である。

現在は複数の AI 利用ツールが商用サービスとして提供されており選択肢は増えている。多くの場合、トライアルもでき、実際の実務データを使用してのツールの性能評価も可能である。調査目的に適合したツールを選択、あるいは構築することがポイントである。

9 終わりに

本報告は 2021 年度の「アジア特許情報研究会」のワーキングの一環として報告するものである。

研究会のメンバーの皆様には様々な協力をしていただきました。ここに改めて感謝申し上げます。

参考文献

1）野崎篤志，「特許情報をめぐる最新のトレンド」
http://www.japio.or.jp/00yearbook/files/2018book/18_a_08.pdf

2）桐山勉，安藤俊幸．特許情報と人工知能（AI）：総論
情報の科学と技術．2017，vol.67，no.7，p.340-349.
https://doi.org/10.18919/jkg.67.7_340

3）安藤俊幸．特許調査における AI 検索と概念検索の有効活用
情報の科学と技術．2022，vol.72，no.7，p.245-250
https://doi.org/10.18919/jkg.72.7_245

4）安藤俊幸．機械学習を用いた効率的な特許調査方法―「人間知能」主導による AI の特許調査への応用―
https://japio.or.jp/00yearbook/files/2021book/21_3_03.pdf

5）大島優香．全文系特許データベースにおける化学構造検索の事例研究
～索引系データベース Chemical Abstracts との比較～
情報の科学と技術．2022、vol.72、no.7、p.257-262
https://doi.org/10.18919/jkg.72.7_257

6）小島史照．Markush 構造検索および化合物の製造方法の特許調査における比較事例研究
情報の科学と技術．2022、vol.72、no.7、p.263-268
https://doi.org/10.18919/jkg.72.7_263

7）難波英嗣，「テキスト間の類似度の測定」
https://doi.org/10.18919/jkg.70.7_373

8）Word2Vec
Mikolov, T. et al. Distributed representations

of words and phrases and their compositionality. Proceedings of the 26th Neural Information Processing Systems, NIPS 2013, 2013,p.3111-3119.

9）SCDV
Mekala,D.et al. SCDV: sparse composite document vectors using soft clustering over distributional representations,Proceedings of EMNLP 2017, 2017, p.659-669.

10）Paragraph Vector
Mikolov, T.; Le, Q. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, ICML 2014, 2014,p.1188-1196.

11）Doc2vec:
https://radimrehurek.com/gensim/models/doc2vec.html

12）Shen, D. et al. Baseline needs more love: on simple wordembedding-based models and associated pooling mechanisms. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, p.440-450.

13）Devlin, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019.

14）BERTScore
Zhang, T. et al. BERTScore: Evaluating text generation with BERT. Proceedings of the 8th International Conference on Learning Representations, 2020.

15）安藤俊幸,「機械学習を用いた効率的な特許調査方法
ニューラルネットワークの特許調査への適用に関する基礎検討」
http://www.japio.or.jp/00yearbook/files/2017book/17_3_04.pdf

16）安藤俊幸,「機械学習を用いた効率的な特許調査方法
ディープラーニングの特許調査への適用に関する基礎検討」
http://www.japio.or.jp/00yearbook/files/2018book/18_3_05.pdf

17）西尾潤･安藤俊幸．AI 系基盤技術と、オープンソースを用いた機械学習による特許文書解析
http://www.tokugikon.jp/gikonshi/298/298tokusyu3.pdf

18）安藤俊幸．特許調査のためのプログラム事例紹介
情報の科学と技術．2020，vol.70，no.4，p.203-207
https://doi.org/10.18919/jkg.70.4_203

19）国立研究開発法人科学技術振興機構・研究開発戦略センター（CRDS）
戦略プロポーザル「第 4 世代 AI の研究開発」―深層学習と知識・記号推論の融合―
https://www.jst.go.jp/crds/pdf/2019/SP/CRDS-FY2019-SP-08.pdf

20）特許庁．ニーズ即応型技術動向調査
「AI 関係技術―演繹と帰納の融合―」
https://www.jpo.go.jp/resources/report/gidou-houkoku/tokkyo/document/index/needs_2021_ai.pdf

21）安藤俊幸．知財分野における第 4 世代 AI の基礎検討―機械学習と特許調査の融合
https://doi.org/10.11514/infopro.2022.0_49

22）特許検索競技大会　過去問 2021
https://japio.or.jp/service/service04_05.html

23）Patentfield
https://patentfield.com/