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―

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

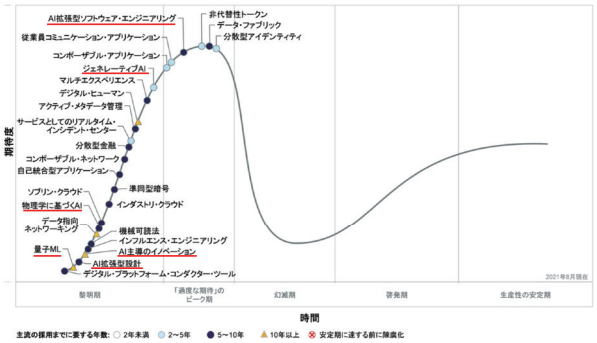
1985 年現花王株式会社入社、研究開発に従事
1999 年研究所の特許調査担当（新規プロジェクト）、2009 年知的財産部、2021 年より現職
2011 年よりアジア特許情報研究会所属
2020 年　特許情報普及活動功労者表彰 日本特許情報機構理事長賞「技術研究功労者」受賞
情報科学技術協会、人工知能学会、データサイエンティスト協会　各会員

ando.t@kao.com

1 はじめに

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 2021 年版を図 1 に示す。新しい AI 応用技術が現れている。

先進テクノロジのハイプ・サイクル：2021年



Gartner、「先進テクノロジのハイプ・サイクル：2021年」を発表
https://www.gartner.co.jp/ja/newsroom/press-releases/pr-20210824

図 1　先進テクノロジーのハイプサイクル 2021 年

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。AI の利用を謳う商用の特許調査・分析ツールは 10 システムを超えている [1)]。既に事前情報収集の段階は通り過ぎて実際に導入している会社も相当数存在していると思われる。ただ上手くいっている会社だけでなく期待通りの結果が得られず困惑している方々も多いのではないかと思われる。実際にエンドユーザーと話をすると AI に過度な期待を抱いている人や、従来の特許調査システムとの違いに苦労されている人も見受けられる。

本稿では人工知能と人間知能（HI：Human Intelligence）の役割分担を踏まえて、特に人間知能主導による事前の情報収集、検証実験、トライアル、実務で活用等の各工程で必要な留意点と実際に自分の手を動かして、試して効果を実感できる特許調査の効率化手法を検討した。

2 知財分野における AI（人工知能）の整理

知財分野における「AI」の性能を客観的に評価するには下記課題があり何をどう評価したらよいか検討対象をまず整理した。

「AI」の性能を客観的に評価するにあたっての課題

①学術的にも定まった「AI：人工知能」の定義が無い [2)]

②ベンダーが提供している AI 利用ツールの「AI」についても各社各様であり定まった定義が無い

③マーケティング目的で「AI」の定義を拡大解釈したものもある

④エンドユーザーが「空想の AI」を念頭に極端な汎用 AI（強い AI）のイメージを抱き過大な期待を抱いて

いる

一言で AI と言っても何をイメージしているかは人により様々である。本稿では便宜的にまず表 1 のように仕分けした。

表 1　便宜的に分けた AI の種類

No.	AIの種類
①	稼働中のAI
②	研究中のAI （自然言語処理、特許情報分野を注目）
③	空想のAI（漫画、SF等） 例：鉄腕アトム、ドラえもん
④	名前だけAI 仮想例：AI審査官、AIサーチャー、AIデータサイエンティスト

知財分野における「稼働中の AI」ツールの出来ることと、出来ないことを理解して、人間知能で行うべきこと人工知能（機械）に行わせることを見極める必要がある。種々の検討を行うにあたり漠然と「AI」を対象としても焦点が合い辛いので以降は AI の中心技術である「機械学習」を中心に検討する。上位概念順に、AI、機械学習、ディープラーニングの関係にある。

「研究中の AI」に関しては自然言語処理、特許情報分野に影響しそうなものを後述する。「名前だけ AI」は端的に偽物の AI と呼ぶ専門家もいる。

3　特許調査 AI の現状に関する問題提起

AI 利用の特許調査・分析ツールを既に導入している会社もベンダーサイドのウェビナーを聴講すると増えているようである。ただ学会発表や筆者が参加しているアジア特許情報研究会関連の話を聞いても華々しい成功事例はあまり聞かない。粛々と使いこなしている人はいるようである。

図 2 に身近な AI の成功事例として「将棋 AI」と「特許調査 AI」との比較を示す。「将棋 AI」と「特許調査 AI」を比較することで AI の使いこなしに関して参考になる。将棋 AI と特許調査 AI それぞれ従来型とディープラーニング型の製品と簡単な特長、課題を図 2 に示す。

将棋 AI は既に数年前よりトッププロ棋士を超えている。それに対して残念ながら、特許調査 AI はプロサーチャーと直接対決を行うレベルに達していない。将棋 AI の目的は将棋に勝つと目的は明確である。それに対して特許調査 AI は特許調査自体が何種類かあり、特許調査のタスク自体の定義が明確とは言い難い。将棋 AI

図 2　「将棋 AI」と「特許調査 AI」との比較

の従来型の製品として「水匠」、ディープラーニング型の製品として「dlshogi」を示した。8 月 15 日に水匠対 dlshogi の計 3 局の対局が行われ、dlshogi の 2 勝 1 敗としたようである。プロ棋士が将棋 AI を利用するときは将棋そのものに関しては十分に分かっており各種の研究用として使用するものと思われる。一般人が将棋 AI の評価値を見る場合は将棋の勝負の優勢の参考にしたり、AI 推奨の最善手を見つつプロ棋士の勝負の行方を楽しむためと思われる。プロ棋士にしても一般人にしても将棋 AI の中身はブラックボックスでも出力は十分に有用である。対して特許調査 AI はどうか？というのが本稿の問題提起であり主題である。

特許調査 AI の従来型として Web 上にマニュアル類等の技術情報が広く公開されている Patentfield[3] を使用した。ディープラーニング型の製品としてアイ･ピー･ファイン社の Deskbee5[4] を試用させていただいた。CNN:Convolutional Neural Network（畳み込みニューラルネットワーク）は画像認識で使われる手法である。基本的な畳み込みネットワークの構造は入力層―畳み込み層―全結合層―出力層となっている。畳み込み層では画像の小領域（例 3 × 3 = 9 ピクセル）を用いてフィルターとして画像の特徴を抽出する。フィルターを順番にスキャンして画像とフィルターの値を掛け合わせる。掛け合わせた数値の総和を求め新たな 2 次元データを取得する。Deskbee5 の 1 次元 CNN は画像認識に優れた性能を示す CNN を自然言語処理向けに 1 次元で使用するように工夫されたものである。図 3 にカーネル法の概念図、図 4 に 1 次元 CNN の概念図を示す。

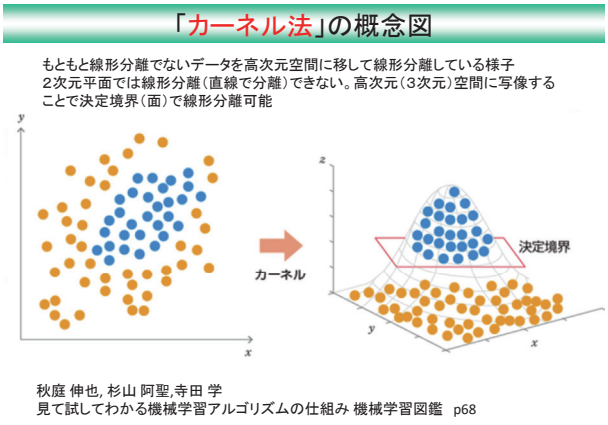


図３　カーネル法の概念図

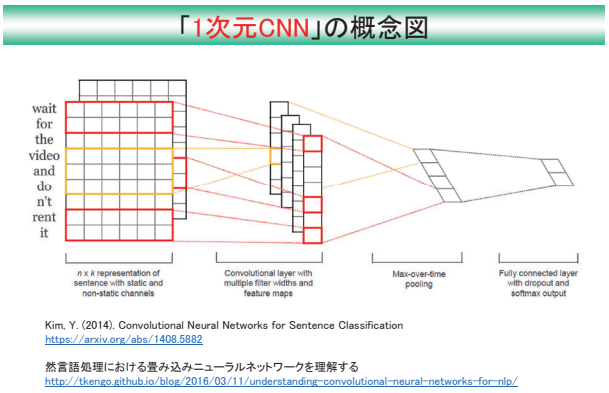


図４　次元 CNN の概念図

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５に示す。

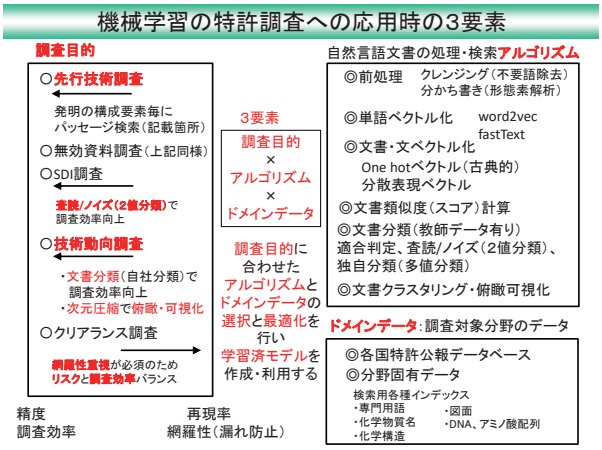


図５　機械学習の特許調査への応用時の３要素

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。

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。

（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。

（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。

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。

図５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、DNA、アミノ酸配列検索等である。この辺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。

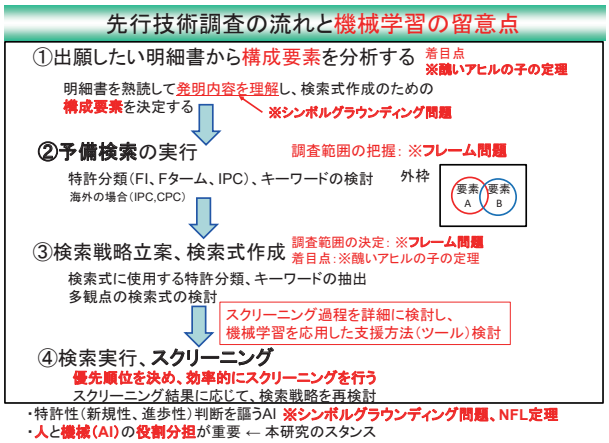


図６　先行技術調査の流れと機械学習の留意点

図６に特許検索競技大会推奨の先行技術調査の流れとその右側に機械学習の留意点を示す。理想的には図６の各行程に適合したアルゴリズムの選択及びチューニングを行い一気通貫に結果が得られることが望ましいが、商用の特許調査 AI ツールの場合は、内部はブラックボックスの場合が普通であり各工程別の性能の評価は困難である。また全工程を備えている訳ではなく一部工程は人手に任せているツールも存在する。ツールの出力結果の評価方法に関しては後述する。

5 特許調査における基礎的な性能評価指標

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７に再現率（網羅性）と適合率（精度）の求め方を示す。

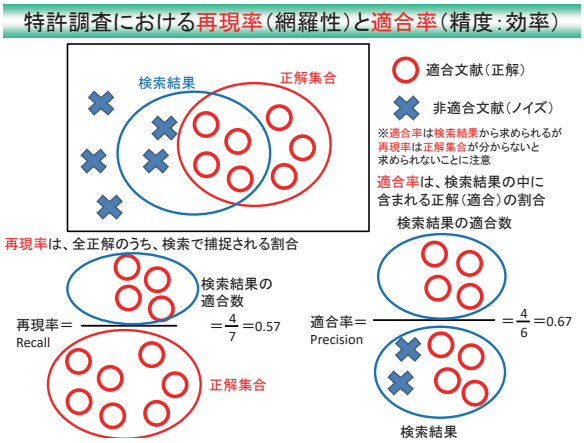


図７　特許調査における再現率（網羅性）と適合率（精度）

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。

特許調査における教師有機械学習の課題を図８に示す。

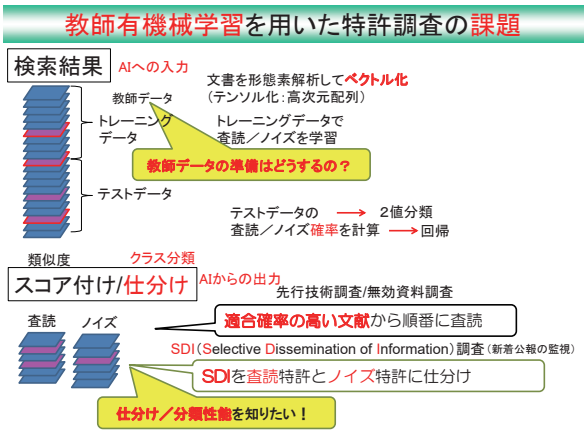


図８　特許調査における教師有機械学習の課題

大きな課題として教師データの準備とクラス分類の性能評価が挙げられる。教師データの準備に関しては別途具体的事例を基に説明する。教師有機械学習の応用例としては SDI 調査（特定分野に関する特許情報を定期的に入手して査読特許を抽出する調査）が検討しやすい。

混同行列（confusion matrix）は、機械学習の分野においてアルゴリズムの性能を可視化するための特有の表配置である。図９に査読／ノイズの２値分類の混同行列を示す。

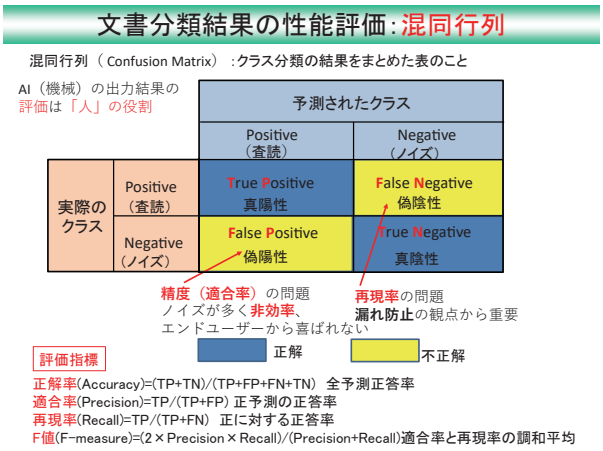


図９　文書分類結果の性能評価：混同行列

混同行列を用いて評価指標として正解率（Accuracy）、適合率（Precision）、再現率（Recall）、F 値（F-measure）を計算できる。筆者は混同行列を集計して上記評価指標を計算する Excel VBA を作成して重宝している。

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[5]。Word2Vec[6] 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[7]（sparse composite document vectors）等がある。Paragraph Vector[8] を実装した Doc2Vec[9] や、SWEM[10] 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[11] を用いたテキスト間の類似度尺度 BERTScore[12] も提案されている。

筆者も文書単位[13]、文単位[14] 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[13]、[14]。

6　商用の AI 利用特許調査・分析ツールの評価方法

一般化した特許調査システムとその評価方法を図 10 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。

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。文書分類の SDI 調査の実例は後述する。

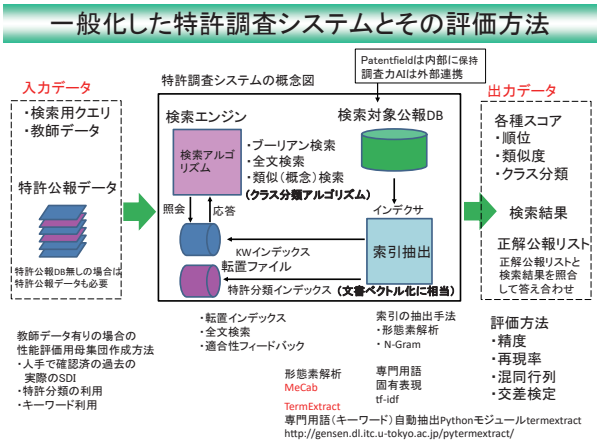


図 10　一般化した特許調査システムとその評価方法

7 SDI 調査の事例検討

SDI 調査への適用を念頭に教師有機械学習検討用のお薦めの母集団作成方法を図 11 にまとめた。

教師有機械学習検討用のお薦めの母集団作成方法

母集団作成方法	人手で確認した過去のSDI実務データ	特定の特許分類を利用	特定のキーワードに着目して作成
例	・特定研究所過去5年分 ・特定テーマ20年分	インクジェットの顔料/染料インク	インクジェットの顔料/染料の類似語を抽出して利用
メリット	実務に直結	使える特許分類があれば便利	キーワードを使用するので汎用性あり
デメリット	社外に論文などで公開しづらい（社内利用では影響無し）	使える特許分類があるとは限らない	キーワードを抽出して辞書化する手間が必要
備考	複数評価者で評価した場合や長期にわたってSDIを継続している場合等、判定基準を揃える必要がある	・特定の特許分類があれば分類付与するExcelマクロを用意しておくと便利	・KW、類似語等をダウンロード可能なDBあり ・特定のキーワードがあれば分類付与するExcelマクロを用意しておくと便利

図 11　教師有機械学習検討用のお薦めの母集団作成方法

人手で確認した過去の SDI 実務データを利用できると実務に直結して各種ツールや使用方法での想定される性能等をある程度予測できる。ツールの社内への導入検討には向いている。ただ社外に論文等で公開しづらい。社内のみで利用する分には問題ない。特許分類を利用する方法は適切な特許分類があり適度な公報数が利用できると便利な方法である。本稿では FI を利用したインクジェットインクを顔料インク（査読）と染料インク（ノイズ）に仕分けする例を紹介する。特定のキーワードに着目してラベル（教師データ）付きの母集団を作成することも可能である。特許分類の利用、特定のキーワードに着目する、どちらの場合も Excel VBA マクロ作成しておくと便利である。

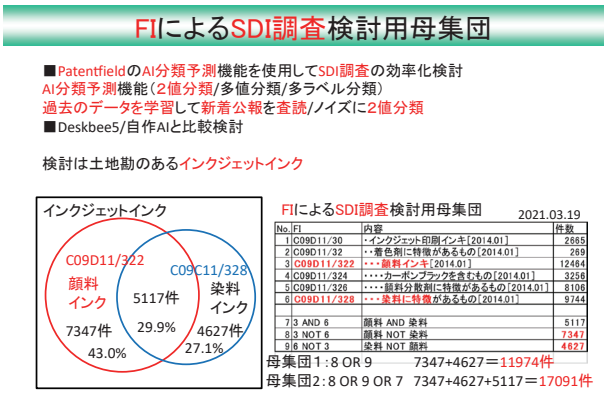


図 12　FI による SDI 調査検討用母集団

図 12 に FI による SDI 調査検討用母集団と関連 FI の件数を示す。使用データベースは Patentfield である。

母集団 1 は染料と顔料の重複する公報を含まないシンプルな母集団である。母集団 2 は重複する公報を含む母集団である。

商用の AI 利用特許調査システムとしては Patentfield（Patentfield ㈱）[3] と AI 判定によるノイズ除去を特徴とする Deskbee 5（アイ・ピー・ファイン㈱）[4] を使用した。

Patentfield には 3 種類の AI 分類予測機能が実装されている。3 種類の AI 分類予測機能の概要を図 13 に示す。

PatentfieldのAI分類予測機能

機能名称	2値分類	多値分類	多ラベル分類
概要	関係する／関係しないを「1 or -1」で分類予測	複数のラベルを用意して、最も近いラベルを1つ付与	複数のラベルを用意して、近いラベルを複数付与
テストデータ項目名	AI予測ラベル	AI予測ラベル	AI予測ラベル
使用例	・ノイズ除去 ・教師データとの関連順にソート（AI予測スコア）	・社内分類（ラベル）付与 ・観点（課題/効果等）付与	・社内分類（ラベル）複数付与 ・観点（課題/効果等）複数付与
備考	AI予測スコアで関係性の程度を「-1～1」の実数表現		

図 13　Patentfield の AI 分類予測機能

2 値分類は教師データを学習し、関係する／関係しないを「1 or -1」で分類予測する。教師データは公報単位で関係するものに 1、関係しないものに -1 を割り振り学習させる。任意の検索集合に対して教師データで学習した、学習済モデルを使用して AI で各公報が関係する／関係しないを予測する。テストデータの分類予測された項目名は 3 種類の予測機能全て「AI 予測ラベル」である。関係性の程度を -1 から 1 の範囲の実数で予測するのが AI 予測スコアである。多値分類、多ラベル分類はユーザー側で予測させたいラベルとその教師データを公報単位で用意する。例えば「顔料」、「染料」、「顔料分散剤」・・・である。多値分類は複数のラベルから最も近いラベルを一つ予測する。特許公報で例えると公報に一つ付与される筆頭 IPC のようなラベルである。多ラベル分類は公報に対して複数付与される F タームのようなラベルである。

AI 分類予測では、教師データで指定された公報から抽出された特徴キーワードと、予測対象公報の特徴キーワードを特徴量として、教師データとの類似度を計算する。AI 予測スコアが高いことは教師データとの特徴量が近いことを意味する。公報に付与された特許分類（IPC,FI,F タームなど）は、分類予測では考慮されない。

Patentfield では AI 分類予測の対象特徴量を「対象特徴量選択ボックス」で各種の分類対象の特徴量を選択可能である。代表的な特徴量を図 14 に示す。

PatentfieldのAI分類予測の対象特徴量

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

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

※Deskbee5の内部処理処理は基本入力テキスト→形態素解析→名詞の固有ID化→1次元CNN →サーチ/ノイズ確率計算→ 2値分類

図 14　Patentfield の AI 分類予測の対象特徴量

セマンティック付記の項目は、単語、文書を機械学習させた概念をもとにした文書特徴量（分散表現ベクトル）である。セマンティック系特徴量では抽出された特徴キーワードの類似キーワードを含めて類似判定を行う。

Patentfieldの「顔料」「染料」の学習済み類似語

	顔料	染料
1	有機顔料	アントラキノン系染料
2	無機顔料	建染
3	着色剤	水溶性染料
4	体質顔料	酸性染料
5	キナクリドン系顔料	直接染料
6	有機着色顔料	アントラキノン染料
7	有機系顔料	塩基性染料
8	顔料分散剤	反応性染料
9	顔料組成物	分散染料
10	カーボンブラック顔料	着色剤
11	キナクリドン顔料	油溶性染料
12	白色顔料	アゾ系染料
13	無機系顔料	アゾ染料
14	染料	キサンテン染料
15	カラー顔料	フタロシアニン系染料
16	メタリック顔料	顔料
17	ピグメント	金属錯塩染料
18	ジケトピロロピロール	色染
19	ジケトピロロピロール系顔料	ナフトール染料
20	顔料誘導体	トリアリールメタン染料
21	顔料微粒子	染顔料
22	ピグメントバイオレット	バット染料
23	アニリンブラック	ジアゾ染料
24	着色材	硫化染料
25	ブラック顔料	インジゴ染料
26	ジオキサジンバイオレット	有機顔料
27	無機着色顔料	染着
28	ピグメントイエロ	イエロー染料
29	マゼンタ顔料	アゾイック染料
30	染顔料	色素

31	フタロシアニン系顔料	反応染料
32	顔料含有量	ブラック染料
33	ピグメントグリーン	アジン染料
34	モノアゾ顔料	キサンテン系染料
35	銅フタロシアニン顔料	シアン染料
36	アゾ系顔料	カチオン性染料
37	顔料分散体	メチン系染料
38	キナクリドン系	マゼンタ染料
39	ピグメントブル	アニオン染料
40	アゾ顔料	無機顔料
41	フタロシアニンブル	媒染染料
42	無機顔料粒子	アニオン性染料
43	二酸化チタン顔料	染着性
44	顔料分散液	トリアリールメタン系染
45	イソインドリン顔料	蛍光染料
46	酸化鉄顔料	天然染料
47	パール顔料	フタロシアニン染料
48	黒色着色剤	キノフタロン染料
49	白色着色剤	含金
50	着色材料	シアニン系染料
51	隠蔽力	ジスアゾ系染料
52	モノアゾ	建染染料
53	顔料ペースト	アントラキノン系
54	アルミニウム顔料	カチオン染料
55	白色系顔料	染料組成物
56	アルミニウムフレーク	ニトロ系染料
57	フタロシアニングリーン	有機系顔料
58	キナクリドン	ジスアゾ染料
59	ピグメントレッド	アゾ色素
60	水溶性染料	水不溶性染料

図 15　Patentfield の「顔料」「染料」の学習済み類似語

図 15 に Patentfield の「顔料」および「染料」の類似キーワード上位 60 を示す。セマンティック検索では類似キーワードを含めて検索している。

黄色のハイライト表示で示したように顔料と染料はお互いに類似キーワードの関係にある。「着色剤」、「染顔料」のような顔料と染料の上位概念のキーワードは顔料インクと染料インクをクラス分類（仕分け）しようとする場合に誤分類の要因になる。同様に顔料の類似語として「水溶性染料」や、染料の類似語として「有機顔料」も誤分類の要因になる。

Patentfieldによる母集団1のAI分類予測の性能評価

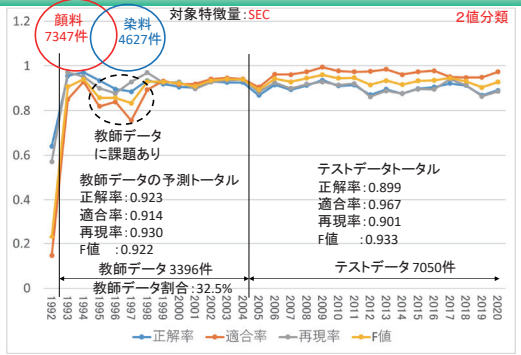


図 16　Patentfield の AI 分類予測（SEC：母集団 1）

Patentfieldによる母集団 1 のAI分類予測（2値分類）結果を混同行列を用いて公開年毎に集計し、評価指標として正解率、適合率、再現率、F 値を求めてプロットした結果を図 16 に示す。この AI 分類予測の対象特徴量は図 14 の SEC である。

教師データとしてデジタル化公報以降の公開年 1993 年から 2004 年の 3396 件を使用して学習させ、2005 年以降の公報を SDI の新着公報に見立てて

評価した。

正解率、適合率、再現率のいずれも 0.8 を上回っている。1992 年以前の結果が振るわないのは、公報データに欠損、OCR の誤字、脱字等のデータ自身の問題があるためで、以降の検討からは、1992 年以前のデータは除いた。1997 年頃に AI 分類予測結果が落ち込んでいるのは教師データに質的な課題があるためと考えられる。図 16 の教師データの割合は 32.5% であり教師データの量としては十分と考えられる。

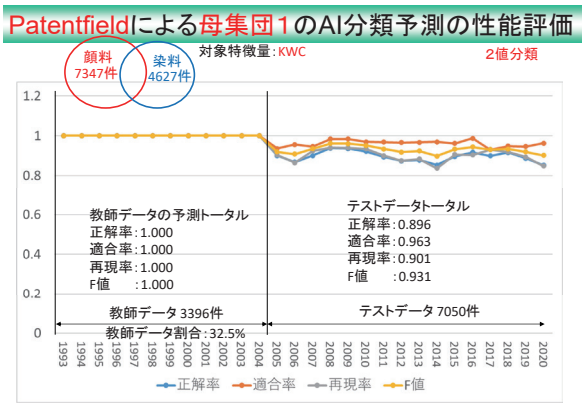


図 17　Patentfield の AI 分類予測（KWC：母集団 1）

図 17 に「名称 / 要約 / 請求の範囲」のキーワードを対象特徴量とする分類予測（KWC）の性能評価結果を示す。教師データの予測がすべて 1 であり、過学習が懸念されるがセマンティックの特徴量（SEC）と同程度の性能を示している。

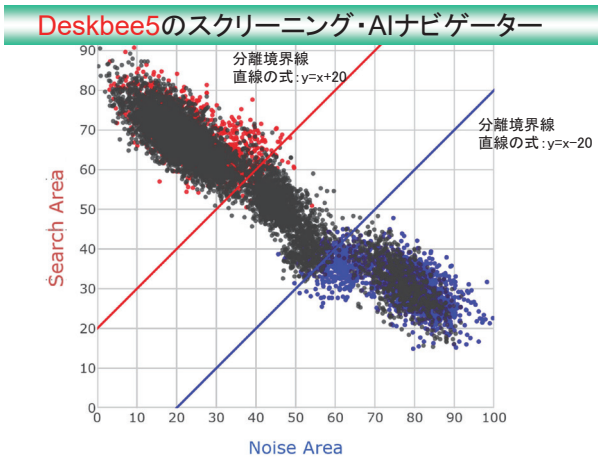


図 18　Deskbee5 のデフォルト画面

図 18 に母集団 1 の Deskbee5 のスクリーニング・AI ナビゲーターのデフォルト画面を示す。赤線と青線の 2 本の分離境界線を散布図の状態や予測結果を参考に 2 本の分離境界線を「適切」に設定するのが大切なポイントである。

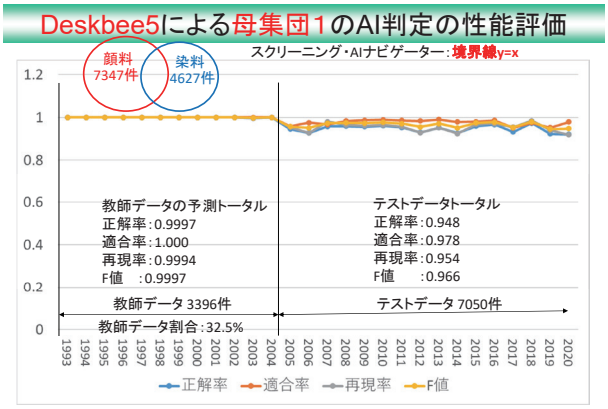


図 19　Deskbee5 の AI 判定（境界線 y = x：母集団 1）

スクリーニング・AI ナビゲーターの 2 本の分離境界線を原点を通る y = x の直線にした性能評価結果を図 19 に示す。Deskbee5 は教師データとテストデータ合計の最大件数が 1 万件なので同じ教師データでテストデータを 2005 ～ 2018 年、2019 年～ 2020 年の 2 回に分けて AI 判定を行った。

図 20 に Deskbee5 の母集団 2 の AI 判定結果（境界線 y = x）を示す。こちらは 3 回に分けて AI 判定を行った。教師データの予測結果は人間に例えるとほぼ丸暗記して満点に近い予測値になっている。

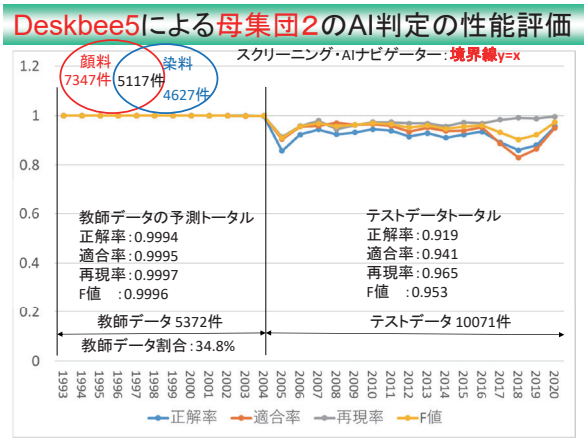


図 20　Deskbee5 の AI 判定（境界線 y = x：母集団 2）

図 21 に商用の AI 利用特許調査ツールの性能評価まとめを示す。

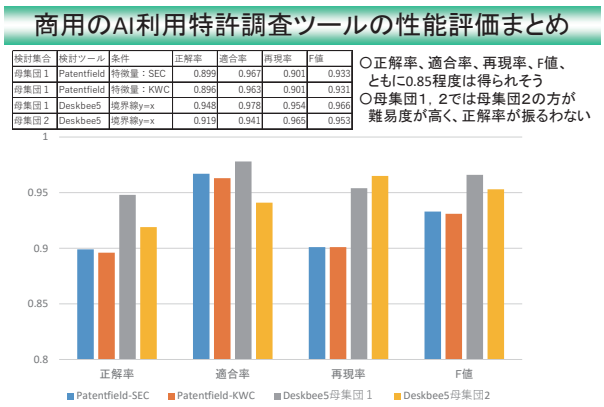


図 21　商用の AI 利用特許調査ツールの性能評価まとめ

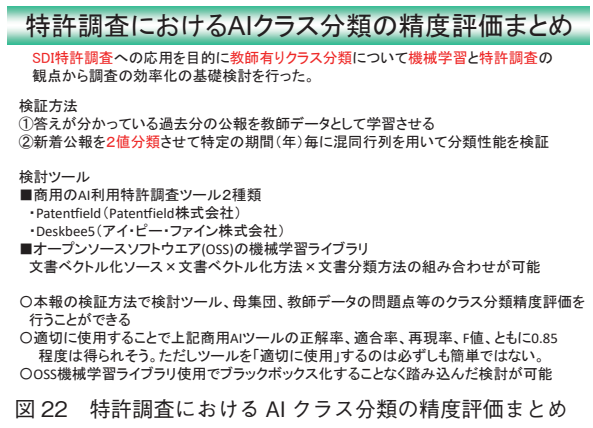


図 22　特許調査における AI クラス分類の精度評価まとめ

8　OSS による文書のベクトル化と文書分類

オープンソースソフトウェア（OSS）による文書のベクトル化処理と文書分類の概要を図 23 に示す。文書データをコンピュータ内部で各種機械学習により扱えるようにするため、5 種類の文書のベクトル化方法を検討した。① BoW モデル作成には scikit-learn[15] の CountVectorizer を使用した。② TF・IDF モデル作成には scikit-learn の TfidfVectorizer を使用した。図 23 の③～⑤の分散表現ベクトル作成には gensim[16] を使用した。文書のベクトル化手法として図 23 の表の 5 種類を検討した。BoW モデルは古典的な非常にシンプルなモデルで出現単語に ID を付け文書の各単語の有無だけを集計する。単語の出現順や頻度は考慮しない One hot ベクトルである。TF・IDF モデルは単語頻度と単語が出現する文書頻度を考慮して重み付けする。Ave-word2vec モデルは文書に含まれる単語の分散表現ベクトルの平均値を使う。doc2vec モデルは word2vec を文書に拡張したものである。Ave-fastText は、word2vec の代わりに fastText を使用した。文書ベクトル化方法の表の③～⑤が分散表現による文書ベクトルモデルである。word2vec、doc2vec fastText、のベクトルの次元数（サイズ）は 300、分かち書きした単語を取り込む Window 幅は 5、取り込み最小単語数は 1 とした。doc2vec の取り込みモデルを選択するパラメータ dm = 1 で単語の語順を考慮するモデルである。公報文書の分散表現ベクトルのデータソースとしてはタイトル、要約、請求項とした。各文書ベクトルを用いて文書分類精度への影響、次元圧縮による各文書の俯瞰可視化マップも検討した。

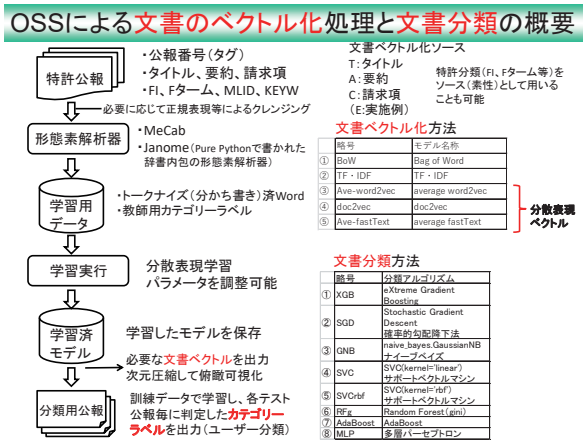


図 23　文書のベクトル化処理と文書分類の概要

① XGB:XGBoost(eXtreme Gradient Boosting)[17] は、勾配ブースティング木を使ったアルゴリズムをオープンソースで実装するソフトウェアである。性能が良いことでクラス分類や回帰で LightGBM と並びよく使われている手法である。

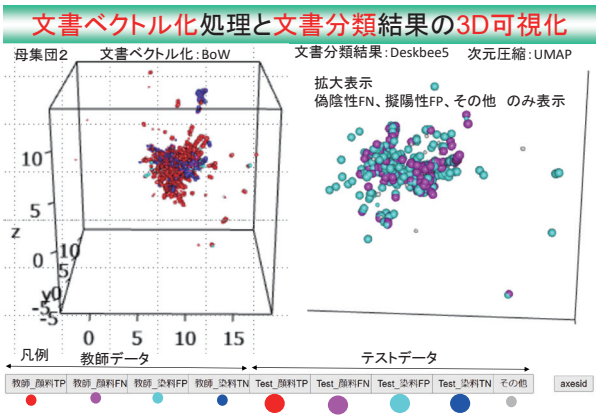


図 24　文書ベクトル化処理と文書分類結果の 3D 可視化

母集団２を BoW モデルで文書ベクトル化して、UMAP[18] で３次元に次元圧縮し、Deskbee5 の文書分類結果をカラーマッピングした。左側は約 15000 件の全体表示である。GPU ボード搭載のデスクトップ PC だとマウス操作ででリアルタイムで回転、拡大、縮小等の 3D 表示が可能である。画面下部の凡例に表示したボタンに割り当てた教師データ、テストデータ、混同行列の４つの結果（TP 真陽性、FN 偽陰性、FP 擬陽性、TN 真陰性）の表示を ON/OFF できる。右側は正しく判定された TP 真陽性、TN 真陰性の表示を OFF にして、誤った判定結果 FN 偽陰性、FP 擬陽性を表示させたものである。

9 AI による特許調査の責任の所在

最近、社外の研究会等で特許調査で AI に責任を取らすことはできないのかという趣旨の質問を何度か耳にする機会があり関連情報を調べてみた。直接特許調査とは関係がないが、「責任」という意味で、図 25 の自動運転レベル３に注目している。自動運転レベル３では「原則として自動運転システムが全ての操作（加速、操舵、制動）を行い、運転者は一切の操作をしない。ただし、自動運転プログラムの機能限界時などには、ドライバーに操作権限が移譲され、その場合には運転者が自ら運転操作を行うことが前提とされている。レベル３の自動運転車の事故における過失責任（の割合）については、場合分けをして考える必要があり、通常の場合、つまり「システムが操作権限を人間に委譲しなかった場合」は、原則としてドライバーの過失は認められなくなると考えられる。（権限委譲要請が無かった場合はドライバーに過失は無いと考えられるので、基本的には、ドライバー以外の者たちの中で、つまりシステムを開発・販売したメーカーや事故相手などの中での過失責任割合が裁判所で判断される、ということになる）。」

より直接的な事例がサーチャーの酒井美里氏の文献「AI 系調査ツールとの付き合い方」に関する視点の提案[17] で紹介されている。「検索結果の責任は誰が負うのか」を論じる中での結論は、「調査を依頼した、あるいは調査の手段を選択した側が責任を負う」という内容である。

図 25 の下部に先行技術調査の技術要素をまとめた。最近の商用の AI 利用特許調査ツールの現状を念頭に強引に自動運転レベルのどこに該当するかを考えると大部分は運転支援のレベル１か、部分運転自動化のレベル２だと思われる。責任の所在は「人」で、特許調査の場合、AI の責任を云々するのは時期尚早と思われる。責任の所在は「人」と一口に言っても実際に誰が責任を負うかの考え方は各社各様と思われる。

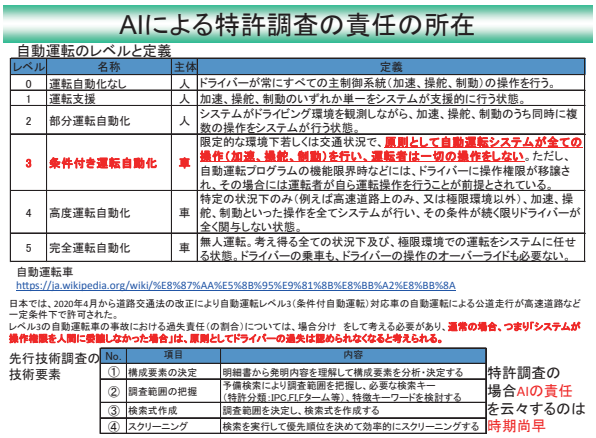


図 25　AI による特許調査の責任の所在

10 まとめ

本稿では「将棋 AI」と「特許調査 AI」の比較を通して「人間知能主導」による AI の特許調査への応用の可能性を示そうと試みた。特許調査 AI の従来型の代表として Patentfield、ディープラーニング型の代表としてアイ・ピー・ファイン社の Deskbee5 を選択し SDI 調査事例を題材に特許調査 AI の性能評価方法を検討した。

特許調査 AI の内部で使用されている TF・IDF ベクトル、分散表現ベクトル等の事例として「OSS による文書のベクトル化処理と文書分類の概要」を紹介した。

「特許調査 AI」はまだ発展途上であり各種の人工知能分野の難問の原理的な課題を抱えている。図 26 に AI 特許調査ツール利用時の注意点まとめを示す。

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 26　AI 特許調査ツール利用時の注意点まとめ

最近の言語モデル・アルゴリズムの進歩は速く様々な学習アルゴリズムが矢継ぎ早に登場している。なかでも Huggingface Transformers[20] は Hugging Face 社が提供している、自然言語処理のディープラーニングのフレームワークであり、BERT などの最先端のアルゴリズムを簡単に試すことができる。TensorFlow と PyTorch の両方に対応しており、テキスト分類や質問応答などの自然言語処理のタスクを実行可能である。ソースコードは全て GitHub 上で公開されており、誰でも無料で使うことができる。商用の特許調査 AI ツールでは、Amplified[21] は内部でディープラーニングの発展系である Transformer[22] を利用していると聞いている。ただしどの商用の AI 利用特許調査・分析ツールも魔法の箱ではない。「道具は使う人次第」というは AI 系のツールにも当てはまる。藤井聡太三冠の AI に対する下記姿勢 [23] は非常に参考になる。「

・AI が出した結論でさえも鵜呑みにせず検証していく

・AI の評価値を見るうえでまず気を付けること

どういう AI を使って、どれぐらい読んでいるのかというところ（基本情報）。

強い AI を使って、例えば十億手読んでいるものであれば、当然それなりにかなり信頼できるなどと考える。

複数の AI を使って、セカンドオピニオンのような形で見る場合もある。『そうした意味でも、AI の評価値が必ず正解なわけではありません。なので、参考にしつつも、自分でも局面を考えてみるという作業は、やはり常に必要になるのかなと思っています。』」

現状の特許調査 AI の利用においては、AI に検索を丸投げして人間は一切操作しないという状況からは程遠い。また誰しもが特許調査のプロというわけでもない。特許調査 AI に対する立場の違いによって求められる AI に対する姿勢も異なる。例えば特許調査 AI に対する立場として、下記役割では求められる AI に対する姿勢も大きく異なる。

・エンドユーザー（研究員 / 知財部員）

・企業への導入担当者

・ベンダー側の開発者

人間知能 HI（Human Intelligence）と AI の役割分担と使い分けが必須である。本稿がその一助となれば幸いである。

11 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）人工知能学会監修、「人工知能とは」．近代科学社

3）Patentfield
https://patentfield.com/

4）Deskbee5
http://www.ipfine.com/deskbee/

5）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

6）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.

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

8）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.
9）Doc2vec:
https://radimrehurek.com/gensim/models/doc2vec.html
10）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.
11）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.
12）BERTScore
Zhang, T. et al. BERTScore: Evaluating text generation with BERT. Proceedings of the 8th International Conference on Learning Representations, 2020.
13）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
14）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
15）scikit-learn
http://scikit-learn.org/stable/
16）gensim
https://radimrehurek.com/gensim/
17）XGBoost Documentation
https://xgboost.readthedocs.io/en/latest/index.html
18）UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
https://arxiv.org/abs/1802.03426
19）酒井 美里「「AI 系調査ツールとの付き合い方」に関する視点の提案」
https://doi.org/10.18919/jkg.70.7_355
20）Huggingface Transformers
transformers 4.10.1 documentation - Hugging Face
https://huggingface.co/transformers/
21）Amplified
https://www.amplified.ai/ja/
22）Attention Is All You Need
https://arxiv.org/abs/1706.03762
23）「第二の藤井聡太は生まれる？」「将棋 AI の評価値をどう考えている？」藤井聡太三冠が返答した“意外な持論”とは
『考えて、考えて、考える』より
https://bunshun.jp/articles/-/48475?page=2