

AIを用いた特許調査における 業務効率化に関する研究

——教師データの作り方の検討を中心に——

情報検索委員会
第2小委員会*

抄 録 昨今、AI搭載を謳った特許調査ツール（以下、AIツール）が数多くリリースされており、これを活用した特許調査等、各種の知財情報調査業務の効率化が期待されている。しかしひとくちに特許調査といっても、クリアランス調査、技術動向調査、先行技術調査、無効資料調査とその内容は多岐にわたり、各々の調査結果に期待される精度も様々である。そして調査業務毎の要望に応じたAIツールの個別調整方法、例えば教師データの構築方法に関する報告等は、これまでほとんど目にする機会が無かったものと思われる。そこで本研究では、期待される調査結果を出力するための教師データの作成方法を検討した。具体的には、教師データを構成する各項目の在り方と、その調整方法、ならびに好ましい教師データの構成について検討を行った。併せて、AIツールの出力結果（判定結果の分類・スコアリングデータ等）を業務効率化に役立てるための利用方法について検討した。

目 次

1. はじめに
2. 課題とAIツールの現状
 - 2.1 背景
 - 2.2 AIツールの特徴
 - 2.3 特許調査業務へのAIツール活用
3. 検討方法について
 - 3.1 検討方法概要
 - 3.2 仮説1／サーチ教師とノイズ教師の技術的距離
 - 3.3 仮説2／教師データの学習項目数
 - 3.4 仮説3／教師データのサンプル数
 - 3.5 仮説4／サーチ教師の拡張
 - 3.6 仮説5／教師データ文字数の統一
4. 結果
 - 4.1 まとめ
 - 4.2 業務への適用可能性
5. 結論、提言
6. おわりに

1. はじめに

現在、第4次産業革命が進行中であると言われていている。第4次産業革命とは、「IoT及びビッグデータ」と「AI」をコアとする技術革新を指し、AIやロボットの技術革新により人間の労働の補助や代替などが可能になるとして、実用化が期待されている¹⁾。

ビッグデータやAIにおいて共通するのは、活用すべきデータの蓄積を前提とすることであり、AIは、蓄積されたデータから新たなデータを生み出す技術とされている。ここで、AIの利活用モデルとしては、①AIの学習に用いる教師データの作成プロセス、②教師データを用いたAIの学習プロセス、③学習済みAIを用いたAIの利用プロセス、が想定されている²⁾。

* 2019年度 The Second Subcommittee, Information Search Committee

いずれのモデルにおいても、蓄積された大量の情報を活用することが重要である。

そして特許情報はまさに蓄積された技術情報の宝庫であり、統一された形式で記載されていることからデータとしての取り扱いの利便性も高い。したがって知財業務に携わるものであれば、特許情報の整理から活用まで、AIを活用できないか模索するのはある意味当然と言える。

我々情報検索委員会第2小委員会第2ワーキンググループでは、知財業務へのAIの活用における課題と現状の調査として、まずは論文³⁾等による情報収集の上、AIツールを提供するベンダーとの意見交換を行った。その結果、現在のAIツールとしては、①スクリーニング効率化・レイティング、②クラスタリング・分類展開を目的とするものが多数であることが確認された。そしてこれは知財業務へのAIの活用において、上記目的への要求が多いことを反映しているものと推測されたため、本研究では、AIツールを使用したスクリーニング効率化を検討することとした。本研究が調査業務における効率化の一端を担えればと考える。

なお本研究は、2019年度情報検索委員会第2小委員会第2ワーキンググループの和田学（ワーキンググループリーダー、凸版印刷）、篠崎直樹（東ソー）、瀬口洋平（日本ガイシ）、武田真弓（ソシオネクスト）、橋本幸一（豊田自動織機）、山本清太郎（三菱ケミカル）によるものである。

2. 課題とAIツールの現状

2.1 背景

現在の知財情報調査業務、とりわけ特許調査業務における情報入手業務に関しては、インターネットを介した機械検索および結果出力を行うことが一般的である。そして近年では、キーワードや分類を用いた検索式構築を要せずに、

調査対象の技術内容を文章入力するのみで（程度の差こそあれ）一定の調査結果を検索・出力されるまでに進化しており、特許庁へ出向いて公報を“手めくり”していた頃とは比べものにならないほど業務効率は飛躍的に向上した。

一方近年、社会における権利意識、法令順守意識の高まりと共に、侵害訴訟等の権利行使の回避等、法的リスク低減の重要性がますます増加している。そこで例えば、他者の権利を尊重して企業活動を実施していることの証左として事前の調査結果を整えておく等、上記活動の根拠となる証拠確保の点から、当該調査結果に対する更なる安全性確保が求められており、その正確性、即ち「漏れの無い調査結果」が一層、望まれている。そのためクリアランス調査においては、調査対象が少数、かつ特定の特許分類について他者特許を確認すれば足りる事案を除き、多くの場合、機械検索にて調査対象の母集合を作成する際の調査漏れを極力低減するために調査範囲を広げた、いわゆる“広め”の検索を行い、調査結果から膨大な数のノイズを除去する一次スクリーニング作業など、非効率的な業務を余儀なくされているものと推察される。

そこで近年発展が目覚ましいAI技術に対して、これを用いた調査業務の効率化に期待が寄せられている。即ち、「漏れの低減（皆無化）」と「ノイズ割合の低減」という両面を、AIに担保して欲しいという期待が高まっている。

2.2 AIツールの特徴

先述の通り、現在、AIを搭載した各種の調査・分析ツールが市販されており、スクリーニング効率化やレイティングを目的としたものがある。本研究では、図1に概要を示したような、予めサーチデータとノイズデータに分類している教師データ（特許テキストデータ）を学習させることで、分類したい判定データ（特許テキストデータ）を分類し、スコアリングするもの

として、AIツールAとツールBを使用した。

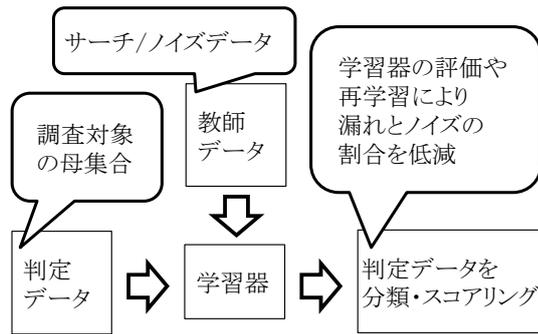


図1 AIツール概要

AIツールAの学習器は、ディープラーニング又はランダムフォレストを用いており、教師データを形態素解析し、明細書用のシソーラス辞書を使用してAI処理する。明細書用のシソーラス辞書は、約30年分の特許公開公報の単語を、同義語、上位概念語、下位概念語、関連語として辞書を作成し、この辞書を用いて文書ベクトルを生成し、機械学習させたものである。また、学習器の評価をすることもできる。

AIツールBの学習器は、畳み込みニューラルネットワークを用いており、教師データを形態素解析で名詞抽出を行い、データの実数ベクトルにより、AI処理する。判定データのサーチ確率、ノイズ確率を散布図で見える化し、重要キーワードが含まれているポイントを可視化することができる。また、その重要キーワードが含まれている集合を抽出し再学習させることができる。

なお、本研究の検証では、AIツールA、AIツールBが備える機能の一部のみを使用し検証しているため、各AIツールの持つ最大の精度がでないものではない。また、本研究実施以降も、ここに紹介していない新たな機能が追加され、精度の向上が図られているようである。

2.3 特許調査業務へのAIツール活用

AIツールを用いて、各調査業務において業

務の効率化が期待されている。例えば、①クリアランス調査、②技術動向調査、③先行技術調査、④無効資料調査である。しかし、どのように教師データを作成し学習させると、期待する判定データが抽出されるかは未知数である。また各調査業務で期待されている精度も異なるため、AIツールから出力された判定結果の分類・スコアリング結果をどのように業務の効率化に役立てられるかを検討する。

3. 検討方法について

3.1 検討方法概要

(1) 評価する技術分野について

AIの教師データの作り方によって、判定結果がどのように変化するかを体系的に評価するにあたり、技術分野によって教師データの作り方の影響が異なることが想定された。そこで本研究では、技術分野を1つに固定し、その分野の範囲内で教師データのパラメータを振ることによって、体系的な知見を得ることとした。

なお対象技術分野は、各委員の所属企業によらず技術的に把握がしやすいものとして、「釣り具分野」を選定した。

(2) 評価方法について

評価は「釣り具分野」の中から、教師データ、判定データともランダムに抽出した。教師データには、抽出したい技術のサンプル群を「サーチ教師」と表現し、逆に抽出不要な技術のサンプル群を「ノイズ教師」と定義した。

ここで「抽出したい／抽出不要」の判別、ならびに、AIが判定した評価の正誤は、いずれも、特定の特許分類を有しているか／有していないかの判断材料にて、客観的に評価する方法を採用した。

(3) 評価する項目について

評価すべき項目については、下記の様に系統的に調べることにした。まずここでは全体像を示し、次節以降で詳細に記載する。

まず、「正解とノイズとの技術的距離が、判定の精度に及ぼす影響」を調べた。これは、以降の検証の難易度を左右する一番大きなファクターであるためである。

事前の想定としては、「正解とノイズとの技術的距離が近ければ、学習した結果に基づいて、判定データが正解かノイズかをAIが判断する難易度は高くなる」と推定した。また、どの程度まで近い技術を区別することができるのかを確認するために実施した。

次に、上記で得た知見をベースとして、「学習する項目（要約、請求項、名称など）」の影響度の検証を行った。この観点も、以降に検証を進めるうえで、キーになると考えたためである。

そのうえで、「教師数が、判定精度に及ぼす影響」について調べた。事前の想定としては、「サンプル数を増やせば判定精度は高くなる」と考えたが、具体的にどのような依存性を示すかについて定量的な知見を得ることを目的とした。

以上の知見を得たうえで、下記2つの観点を最適化することで精度アップが期待できるのではないかという推論を元に、検証を行った。まず、「教師データの正解とノイズの境界」については、「再現率を高めるには、領域の情報を強化することが有効に作用するのでは」という仮説に基づいて検証を行った。

更には、教師データの各データ量は、「揃っていた方が良いのか」という仮説に基づいて各教師データの文字数を強制的に調整して検証を実施した。

以上が教師データ作成プロセスに必要な知見を得るために行った検討項目である。

(4) サーチ教師数とノイズ教師数のバランスについて

今回検討した範囲では、サーチ教師数とノイズ教師数は、基本的に1：1の比率で行った。いずれかの件数比率が高くなると、多い側に判定が引きずられるという傾向があることはベンダーから情報を聞いていたので、今回は中立条件で固定して行った。

(5) 評価指標に関して

本検証は、表1に記載のTP、FP、FN、TNの4つの表記方法で判定結果の属性を区分した。なお、表1において、「真の結果」は人手による判定の結果、「予測結果」はAIによる判定の結果であり、それぞれサーチを「正」、ノイズを「負」としている。上記4つの表記方法は、この正(P: Positive)及び負(N: Negative)と、AIによる判定が人手による判定と一致していた場合(T: True)及び不一致であった場合(F: False)との組み合わせからなる。

また判定精度の評価指標は、表2に定義した通り、サーチを適切に抽出できるかを表す適合率「 $TP/(TP+FP)$ 」と再現率「 $TP/(TP+FN)$ 」だけでなく、ノイズを適切に除外できるかも加味した正解率「 $(TP+TN)/(TP+FP+TN+FN)$ 」及び、クリアランス調査で重要となるサーチ漏れ率「 $FN/(TP+FN)$ 」を使用した。

上記式に基づくと、適合率はAIが正解と判断した判断が「どれだけ正しかったか」、また、再現率はAIが抽出すべき真の正解特許を「どれだけ抽出できたか」を示す指標である。すなわち、再現率と適合率は本質的に相反する性質

表1 真の結果と予測結果の表記方法

		真の結果	
		正	負
予測結果	正	TP	FP
	負	FN	TN

がありこの2値は、重要なパラメータとなる。

表2 正解率などの定義

正解率	$(TP + TN) / (TP + FP + TN + FN)$
適合率	$TP / (TP + FP)$
再現率	$TP / (TP + FN)$
サーチ漏れ率	$FN / (TP + FN)$

3.2 仮説1 / サーチ教師とノイズ教師の技術的距離

仮説1は、判定結果がサーチ教師とノイズ教師の技術的距離に依存することである。サーチ教師とノイズ教師の技術が近い（教師データの範囲が狭い）程、判定難易度が高くなり、反対にサーチ教師とノイズ教師の技術が遠い（教師

データの範囲が広い）程、判定難易度が低くなると予想した。そこで、サーチ教師とノイズ教師の技術的な距離を3段階に分け、AIツールを用いて仮説1を検証した。特許分類の階層の深さであるメイングループ、サブグループ、分冊識別記号を技術的な距離に適用し、3段階にレベル分けを行った（表3及び図2参照）。

レベル1は、サーチ教師とノイズ教師の技術的距離が遠くなるよう、特許分類の階層をメイングループとし、教師データを作成した。サーチ教師は「釣竿（A01K 87/00）」、ノイズ教師は「疑似餌（A01K 85/00）」、「リール（A01K 89/00）」とした。サーチ教師である「釣竿」と、ノイズ教師である「疑似餌」、「リール」では、技術が大きく異なることから、レベル1の教師データによる判定難易度は「低」と予想した。

表3 条件一覧表

レベル	技術的距離	特許分類の階層	サーチ教師	ノイズ教師	難易度
1	遠い	メイングループ (IPC)	A01K 87/00 (釣竿)	A01K 85/00 (漁撈用の疑似餌) A01K 89/00 (漁撈用のリール)	低
2	中間	サブグループ (IPC)	A01K 87/08 (握柄)	A01K 87/02 (竿の接続具) A01K 87/04 (竿上の釣糸案内) A01K 87/06 (竿上にリールを固定する用具)	中
3	近い	分冊識別記号 (FI)	A01K 87/08, B (リールシートと関連を有するもの)	A01K 87/08, A (構造に特徴を有するもの) A01K 87/08, C (素材または製造方法に特徴を有するもの) A01K 87/08, Z (その他のもの)	高

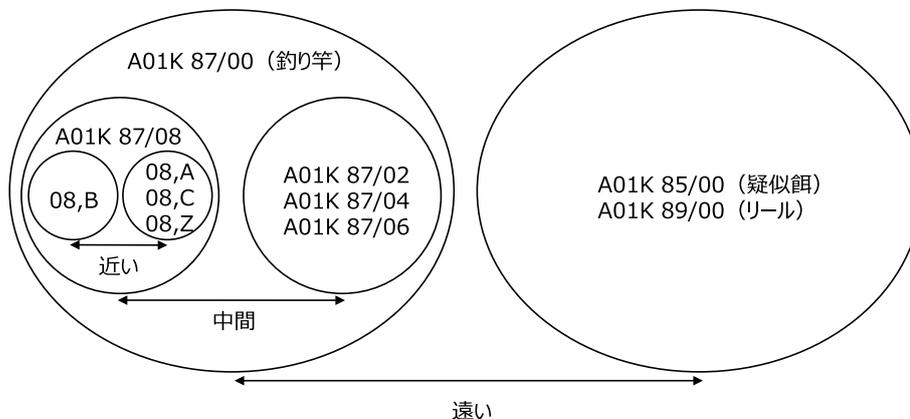


図2 階層イメージ

レベル2は、サーチ教師とノイズ教師の技術的距離が中間となるよう、特許分類の階層をサブグループとし、教師データを作成した。サーチ教師は「握柄 (A01K 87/08)」, ノイズ教師は「竿の接続具 (A01K 87/02)」, 「竿上の釣糸案内 (A01K 87/04)」, 「竿上にリールを固定する用具 (A01K 87/06)」とした。レベル2の教師データは、「釣竿」の技術分類の中でサーチとノイズを区切っており、判定難易度は「中」と予想した。レベル3は、サーチ教師とノイズ教師の技術的距離が近くなるよう、特許分類の階層を分冊識別記号とし、教師データを作成した。サーチ教師は「リールシート (A01K 87/08, B)」, ノイズ教師は「構造 (A01K 87/08, A)」, 「素材または製造方法 (A01K 87/08, C)」, 「その他 (A01K 87/08, Z)」とした。レベル3の教師データも、レベル2と同様に「釣竿」の技術分類の中でサーチとノイズを区切っているものの、より細分化された技術でサーチとノイズが区切られており、判定難易度は「高」と予想した。

サーチ教師とノイズ教師は、表3に記載の特許分類を母集合とし、ランダムに各々100件抽出して作成した。判定データも教師データと同様にサーチとノイズを各々100件抽出して作成した。ここで、教師データと判定データは重複しないものとした。

判定結果を表4に示す。

表4 判定結果

	レベル1 (遠い)	レベル2 (中間)	レベル3 (近い)
正解率	96%	86%	85%
適合率	95%	89%	84%
再現率	96%	82%	86%
サーチ漏れ率	4%	18%	14%

サーチ教師とノイズ教師の技術的距離が遠いレベル1では、正解率、適合率、再現率がいず

れも高く、信頼性の高い結果となった。一方、レベル2 (技術的距離：中間) 及びレベル3 (技術的距離：近い) では、レベル1と比較し、正解率、適合率、再現率のいずれも低くなった。

これより、AIツールを用いた判定結果は、サーチ教師とノイズ教師の技術的距離に依存する傾向が見られた。

なお、レベル1の教師データを用いた場合、正解率、適合率、再現率がすでに高いことから、次節以降の仮説2～5では、レベル2または3相当の教師データを用いて、判定結果を向上させるための検証を行った。

3.3 仮説2 / 教師データの学習項目数

仮説2は、教師データの学習項目数が判定に依存することである。教師データは、特許テキストデータの学習項目が多いほど、情報量が多くなるため、判定精度が高くなると予想し、以下の条件1と条件2で検証を行った。まず条件1の学習項目を表5に示す。

表5 条件1の学習項目

条件	学習項目
教師1	請求項1のみ
教師2	全請求項
教師3	全請求項+要約
教師4	全請求項+要約+発明の名称

サーチ教師とノイズ教師は筆頭IPCで分類し、教師データと判定データを各200件準備した。抽出条件を表6に示す。サーチ分類は、各検索において、抽出したい内容が含まれる分類であり、ノイズ分類は、抽出したい内容以外の分類としている。教師データは、サーチ分類(1分類)とノイズ分類(4分類)を各々100件抽出して作成したが、判定データは、サーチ分類(1分類)とノイズ分類(4分類)を分類ごとに各々40件抽出して作成した。準備したデータについて、教師1～教師4までの項目ごとに検

証を行った。なお判定データの項目は、各教師データの項目と同じとした。例えば教師1では判定データも請求項1のみとした。

表6 教師データ及び判定データ件数

サーチ分類	教師データ	判定データ
A01K 87/06	100件	40件
ノイズ分類	教師データ	判定データ
A01K 87/00	40件	40件
A01K 87/02	20件	40件
A01K 87/04	20件	40件
A01K 87/08	20件	40件
総計	200件	200件

検証結果を図3に示す。

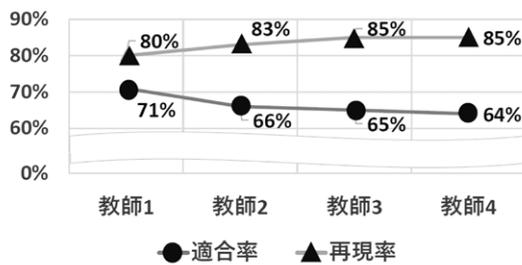


図3 検証結果 (再現率・適合率推移)

検証結果からは、教師1・教師2では、適合率が高く、再現率が低い結果となり、教師4・教師3では、再現率が高く、適合率が低くなった。

次に条件2として表7の学習項目を検証した。

表7 条件2の学習項目

条件	学習項目
教師1	発明の名称+要約+請求項(独立項)
教師2	発明の名称+要約+全請求項
教師3	発明の名称+要約+全請求項+発明が解決しようとする課題+発明の効果
教師4	発明の名称+要約+全請求項+詳細な説明(全文)

サーチ教師とノイズ教師は筆頭IPCで分類し、教師データと判定データを準備した。抽出条件を表8に示す。教師データは、サーチ分類(1分類)とノイズ分類(1分類)を各々100件抽出して作成したが、判定データは、サーチ分類(1分類)を100件、ノイズ分類(1分類)を300件抽出して作成した。準備したデータについて、教師1~教師4までの項目ごとに検証を行った。なお判定データの項目は、各教師について教師1~4と同様の4通りのもの(判定1~4)を準備した。例えば教師1に対し、判定1~4まで検証し、その平均値を結果とした。

表8 教師データ及び判定データ件数

サーチ分類	教師データ	判定データ
A01K 87/08B含む	100件	100件
ノイズ分類	教師データ	判定データ
A01K 87/08B含まない	100件	300件
総計	200件	400件

検証結果を図4に示す。

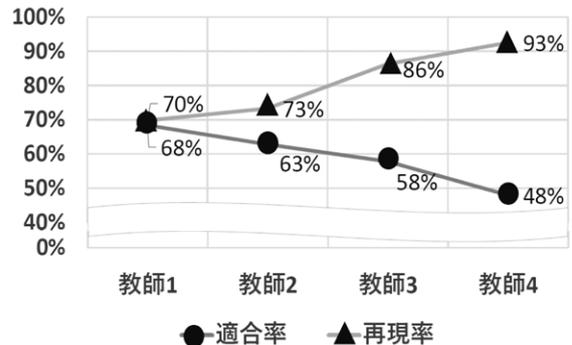


図4 検証結果 (再現率・適合率推移)

検証結果からは、条件1と同様の傾向がみられ、教師1・教師2では、適合率が高く、再現率が低くなり、教師4・教師3では、再現率が高く、適合率が低い結果となった。

再現率は、表1、表2で示されていた評価指標の通り、人手による判定結果がサーチのうち、AI判定結果がサーチだった割合である。教師

4が高いことから、検索案件の漏れを少なくする必要がある調査には、学習項目を増やす方がよいことがわかる。適合率は、表1、表2で示されていた評価指標の通り、AI判定結果がサーチのうち、人手による判定結果がサーチだった割合である。教師1が高いため、ノイズを減らす目的ならば、学習項目は少なくともよいことがわかる。

条件1と条件2の検証結果から、情報量が多くなれば、再現率がよくなる一方、適合率は落ちていくことが証明された。

3. 4 仮説3／教師データのサンプル数

教師データのサンプル数（以下、N数）と判定精度の相関の検証は以下の様に行った。

(1) 母集合データの作成

IPCでA01K 87/08（全IPC）を有する国内特許を抽出した。

(2) サーチ教師／ノイズ教師の作成

上記の母集合から、「FI=A01K87/08B」を有する案件をサーチ、それ以外をノイズとしてデータを選別した。その中から、学習用／判定用にデータを振り分けて使用した。

(3) 教師データの振り分け条件

教師データは表9及び表10に示す条件1～4を作成した。

表9 教師データの振り分け

条件	サーチ教師 (FI=A01K87/08B を有する)	ノイズ教師 (FI=A01K87/08B を有さない)
条件1	10件	10件
条件2	25件	25件
条件3	50件	50件
条件4	100件	100件

表10 判定データについて

構成（正解数）	件数
サーチデータ (FI=A01K87/08Bを有する)	120件
ノイズデータ (FI=A01K87/08Bを有さない)	180件
合計数	300件

(4) 結果について

図5から、N数と適合率、正解率は強い相関が見られた。

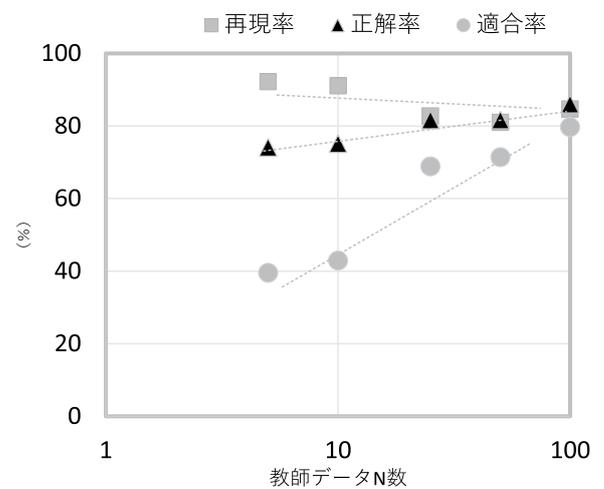


図5 教師データN数と精度の関係

今回の結果からは教師データのN数が100までしか評価できていないがこの範囲においては、N数は多いほど適合率と正解率が高くなることが確認できた。

3. 5 仮説4／サーチ教師の拡張

仮説4では、サーチ教師の対象を拡張することで、サーチ領域を拡張できるかを検証した。サーチ領域が拡張できれば、再現率の改善が期待できる。

まず、AIツールBを例に、サーチ領域を補足説明する。AIツールBのAIの出力では、1つの判定データに対し、サーチである確率（サーチ確率）、ノイズである確率（ノイズ確率）の、2つが出力値として得られる。判定データは、

得られた2つの出力値を座標と見なすことで、図6に示す座標系に配置される。これが前述の散布図である。AIツールBでは、図6の座標系で、左下より右上に延びる境界線を設定し、境界線より左上にある判定データをサーチ、右下にある判定データをノイズと判定する。境界線の左上の領域が、サーチ判定の対象となるサーチ領域である。AIツールAにおいては、このようなサーチ領域の可視化は行われませんが、サーチ確率又はノイズ確率に相当する値は、スコアとして出力される。

図6では、サーチ・ノイズが分かっている評価用データの判定結果を、例示している。境界線が実線の位置の場合、再現率73%、適合率80%である。仮に、サーチ領域を右下側の隣接領域まで拡張し、境界線を一点鎖線の位置に移動させることができると、再現率100%、適合率73%と変化する。このように、サーチ領域の拡張により、再現率は改善される。

AIツールBでは、境界線の位置を直接変更することができるが、変更できないツールも多い。AIツールAもできない。そこで、汎用性の高い方法として、教師データにより、サーチ領域を拡張する方法を検討した。

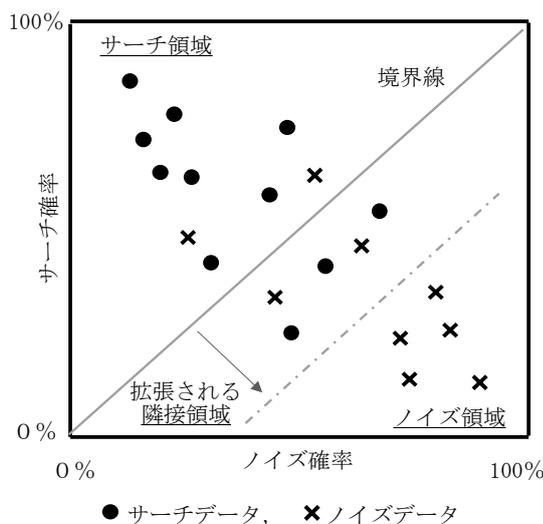


図6 サーチ・ノイズ領域及び境界線

原理的に、隣接領域に対応するノイズ教師を、サーチ教師に加え学習させることができれば、境界線の位置は移動する。ただし、課題が2つある。1つ目の課題は、ノイズ教師の特定である。背反として適合率が低下するので、過度の適合率の低下を避ける為、境界線に近い隣接領域にあるノイズ教師を特定する必要がある。2つ目の課題は、結果の確認である。教師データを変更すると、学習結果が変わり、判定データに対し出力されるサーチ確率・ノイズ確率も変わる。つまり、図6の座標系によっても、結果の適否の確認は容易でない。そこで、1つ目の課題については、仮説1の検証結果を参考とした。仮説1の検証結果にて、特許分類に基づく技術的距離と判定の精度とに相関が見られたことより、特許分類に基づく技術的距離と、図6の座標系における距離とに相関があると仮定した。本検証では、ノイズ教師中より、サーチ教師に付与された特許分類の上位分類、または同時付与分類(同時に付与されることが多い分類)が付与されたノイズ教師を、サーチ領域に近い、隣接領域に位置するノイズ教師と仮定し、各々検証を行った。2つ目の課題については、想定された精度の変化、再現率改善・適合率低下、の有無により、サーチ領域に意図する拡張が生じたか、判断することとした。

具体的な検証方法を記す。母集合を「釣竿(A01K87)」, 基となるサーチ領域を「構造に特徴-リールシートと関連(A01K87/08B, 以下08Bとも記載)」とする。隣接領域は、上位分類「構造に特徴(A01K87/08A, 以下08Aとも記載)」, 又は、同時付与分類「スクリー-ロック-リールシート(A01K87/06B, 以下06Bとも記載)」, とした。A01K87/08Bの公報群に08Aの公報群を追加することでサーチ教師1を、06Bの公報群を追加することでサーチ教師2をそれぞれ作成し検証を行った。検証条件の概要は表11のとおりである。

表11 検証条件

構成		分類	件数
母集合		A01K87/	-
教師 1	サーチ教師 1	A01K87/08B+08A	100件
	ノイズ教師 1	サーチ教師 1 以外	100件
教師 2	サーチ教師 2	A01K87/08B+06B	100件
	ノイズ教師 2	サーチ教師 2 以外	100件
判定	サーチ判定	A01K87/08B	100件
	ノイズ判定	サーチ判定以外	300件

上記の条件にて、サーチ教師とノイズ教師の件数割合を変更して検証した結果を図7に示す。

教師1では、サーチ教師における08Aの割合を0%より次第に増加するように変化させたところ、再現率が改善し、適合率は低下した。一方、教師2にてサーチ教師の割合を同様に変化させたところ、適合率の低下は無く、再現率が改善したが、その変化幅は小さかった。

教師1は、想定通りの変化より、サーチ領域が拡張された、と考えてもよさそうである。適合率は低下するが、変化幅は比較的小さく、再現率を優先したい場合には、選択肢となりえる。ただし、留意すべき点も確認された。他の公報群にて検証を行ったケース（図示無）では、同分類・同条件にて、サーチ教師に含める08Aの割合が10~20%の範囲のみ再現率の改善効果が見られ、それ以外では改善が見られなかった。効果が得られる範囲は、少なからずデータ（公報）の内容にも依存するようである。

教師2については、意図する結果は得られなかった。再現率・適合率に変化はあったが、数値の挙動より、サーチ領域の拡張とは異なると判断される。

上記のように、サーチ領域を拡張することで、再現率を改善できることが確認できた。また、教師データにてサーチ領域を拡張する手法の一例として、上位分類が使えることが確認できた。

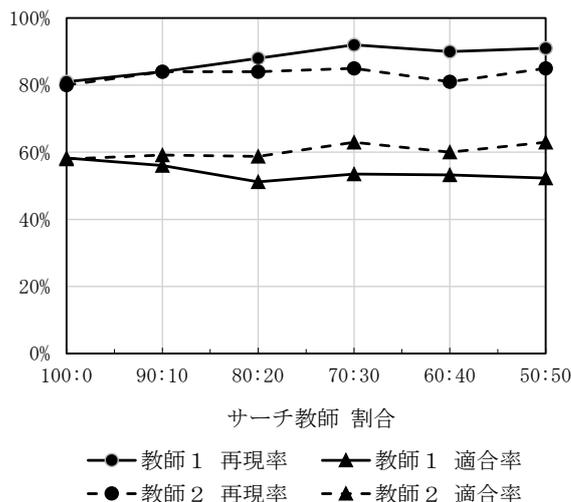


図7 検証結果 (再現率・適合率推移)

3.6 仮説5 / 教師データ文字数の統一

仮説5として、教師データの全請求項の文字数を揃えた場合に精度が向上する可能性に関して検証した。

特許公報における請求項のデータを教師として用いAI処理する場合、以下の問題が生じる可能性がある。

- ・技術分野によっては公報の数が少ない
- ・出願時期が古い公報だと請求項の文字数が極端に少ない

このような、公報の数を十分に揃えられずに、文字数が少ない公報も教師データとして用いる必要がある場合、AIを活用しても解析対象から十分に特徴をとらえることができず、精度が上がらないものと考えられる。

そこで、このようなデータ不足を解消する手法である「データ拡張」⁴⁾の一環として、文字数を増やす「水増し」処理を行うことで教師データを拡張することを検討した。「データ拡張」処理には、様々な方法があるが、本稿ではシンプルに、「ある一定の文字数まで請求項の文章を繰り返しコピーする」処理を行った。ここで、「ある一定の文字数」を「最も文字数の多い公報」基準にすると、文字数の少ない公報における繰

り返しが過剰となり、その公報がAI処理に与える影響が過大となる可能性がある。そこで、基準とする文字数は、教師集合における中央値に留めることにした。また、一般に特許の請求項に関しては、上位の請求項に記載される内容は発明の主要な特徴である一方、下位の従属項で記載される内容は、相対的に低い重要度の特徴であると考えられる。中央値を超える文字数の部分に記載される内容は相対的に低い重要度の記載であり、解析上、ノイズとなる可能性がある。そのため、中央値を超える文章は削除する、「カット」処理も併せて検討した。

(1) データの構成

母集合：FIでA01K 87/08（下位分類を含む）の集合を作成

期間： 全期間（～2020年1月）

サーチ教師：FI=A01K87/08 Bを有する案件
271件

ノイズ教師：FI=A01K87/08 Bを有さない案件
395件

(2) データの統計情報

A01K 87/08集合に対し、請求項の文字数の統計処理を行った結果を以下に記す。図8は集合の各公報を全請求項の文字数50文字毎に区間毎に分け、各区間における公報数を頻度として

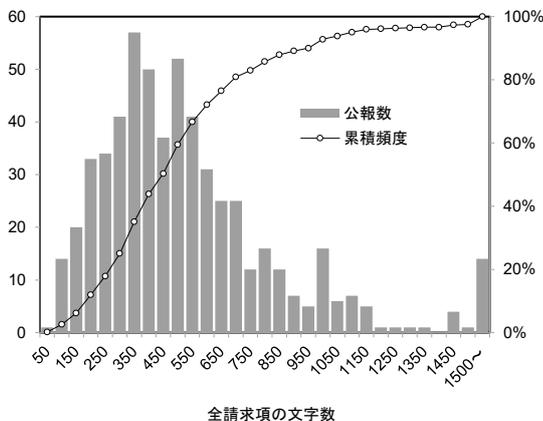


図8 文字数のヒストグラム

表したヒストグラムである。

平均値 565文字
中央値 449文字
最小 34文字
最大 14,213文字

文字数を揃える基準として、中央値の近傍で切りの良い450文字を選定した。

(3) サーチ教師／ノイズ教師の作成

サーチ教師（A01K 87/08B）271件から、ランダムに教師データ用に100件、判定データ用に100件を選択した。

同様に、ノイズ教師（A01K 87/08B以外）395件から、ランダムに教師データ用に100件、判定データ用に100件を選択した。

(4) 教師データの前処理

解析対象項目は「請求項」のみとし、教師公報200件に対し、以下の3条件の前処理を行い、3水準を設定した。

教師1：未処理

教師2：全請求項文字数450字以上の公報全てに対して、450字で文字数を「カット」した（図9の①）。

教師3：教師2に対して、さらに全請求項文字数450字未満の公報に対して、450字となるように文字数を「水増し」した（図9の②：表計算ソフトの機能を使用し請求項を単純にコピーし、450字を超えた部分をLEFT関数により削除した）。

各処理の概念を表したものを図9に示す。グラフは、全請求項の文字数の小さい方から順番に各公報を並べ、各公報の文字数を散布図として表したものである。①のカット処理では、ノイズとなる可能性のある、各公報における中央

値を超えた文章を削除したイメージである。一方、②の水増し処理では、中央値に満たない公報において文章を繰り返しコピーし、データ拡張を行ったイメージである。

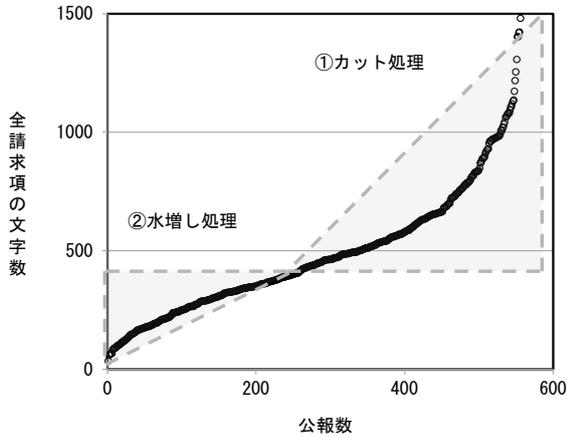


図9 各処理の概念図

(5) 解析結果

AIツールAでの結果を表12に示す。

表12 検証結果（正解率・適合率等比較）

	教師1 未処理	教師2 カット	教師3 カット+ 水増し
正解率	80%	80%	81%
適合率	77%	80%	80%
再現率	83%	80%	81%
サーチ漏れ率	17%	20%	18%

教師1と教師2の結果を比較すると、文字数カット処理により、適合率の改善と、再現率の低下とが確認された。これは、ノイズ公報を減らすことができたが、一方で、サーチ公報の漏れも増えたことを意味する。即ち、教師データに文字数をカットする前処理を行うことは、スクリーニングの際、サーチ公報/ノイズ公報の両方を落とすことになる。

教師2と教師3の結果を比較すると、文字数水増し処理により、適合率の維持と、再現率の

改善が確認された。観点を変えて、教師1と教師3の結果を比較すると、再現率を維持しつつ、適合率を改善できたことになる。これは、サーチ公報を漏らさずに、選択的にノイズ公報を落とすことができたことを意味する。

今回の検証によれば、教師データの前処理において、全請求項の文字数を減らす「カット」処理と、文字数を増やす「水増し」処理の両方を行うことで、スクリーニングの精度を向上させる可能性が示唆された。

4. 結果

4.1 まとめ

仮説1～5の検証結果を属性とし、精度（正解率、再現率、適合率）への影響の観点で、次ページの表13にまとめた。

今回の検証結果の活用については、一例を次節で触れる。

4.2 業務への適用可能性

仮説1～5の検証にて確認された属性より、AIツール利用時の作業フローの検討を行い、次ページの図10に示した。まず、「3. 教師数」の目安は、理想としては200件である。「3. 教師数」がそれ以下の場合、「1. 技術的距離」、または「5. 教師文字数統一」を確認し、適合率や再現率が低くなる傾向にあるかを確認するとよい。教師データを設定したら、検証を行う。検証前の確認結果、又は検証結果にて、精度が不十分な場合には、まずは「2. 項目数」、「4. 教師周辺領域補完」、「5. 教師文字数統一」の観点で改善を試みる。適合率が不足する場合には、「2. 項目数」あるいは「5. 教師文字数統一」の操作により、改善される可能性がある。一方、再現率が不足する場合は、「2. 項目数」と「4. 教師周辺領域補完」、「5. 教師文字数統一」を検討する。前記操作によっても、十分

表13 仮説1～5の調査精度への影響

属性	1. 技術的距離		2. 項目数		3. 教師数		4. 教師 周辺領域 補完	5. 教師 文字数 統一
	遠	近	少	多	少	多		
正解率	↗	↘	↗	↘	↘	↗	↘	↗
再現率	↗	↘	↘	↗	↘	↗	↗	↗
適合率	↗	↘	↗	↘	↘	↗	↘	↗
備考					本テーマでは200件 以上		改善レン ジには幅	文字数の 水増し可

な精度が得られない場合には、「3. 教師数」を見直し、教師データの追加を検討する。

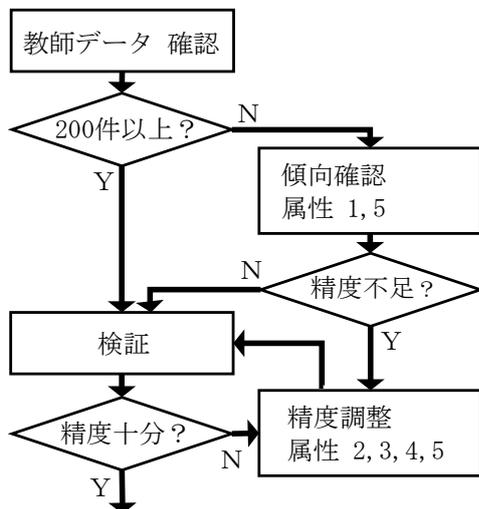


図10 AIツール利用時の作業フロー

従来、AIツールで十分な精度が得られないと、一足飛びに「3. 教師数」教師データ数の不足に目が向けられる傾向にあった。しかしながら、通常、教師データの追加は、新たな教師データの作成作業をとめない、作業者の負担が大きい。

今回の検証結果を業務に適用することで、「2. 項目数」の調整や、「4. 教師周辺領域補完」、「5. 教師文字数統一」など、既存の教師データの加工といった、より負担の少ない作業で必要な精度が得られるケースが期待できる。

5. 結論、提言

上述の通り、AIツールによるスクリーニングでは、教師データを調整することで判定精度が向上することが証明できたものの、サーチ漏れが生じるという課題は解決していない。このため、本研究においては調査業務全般の一次スクリーニングを完全に人からAIツールへ置き換えることは時期尚早との結論に至った。しかし、技術動向調査や先行技術調査等、多少のサーチ漏れを許容できる調査業務においては、現時点でも業務への活用が十分に可能であると考えられる。

次ページの表14に調査タイプ別のAIツールの活用推奨度を記した。調査タイプは、先行技術調査、クリアランス調査、無効資料調査、技術動向調査の4つに分類し、AIツールを活用するうえでの2つの観点を軸に活用推奨度を決定した。

1つ目の観点は、調査に漏れないことが必須か否かである。先にも述べたように、技術動向調査や先行技術調査は多少のサーチ漏れは許容される傾向にあるが、クリアランス調査や訴訟リスクの面から確実に無効化しなければならない案件の無効資料調査では、少しのサーチ漏れが致命的な調査ミスにつながる可能性がある。以上の観点から、技術動向調査や先行技術調査を○評価とした。

表14 調査タイプ別AIツール活用推奨度

調査タイプ	活用観点① 漏れないことが最優先	活用観点② 確認箇所	活用推奨度
先行技術調査	○	△：全文	○～△
クリアランス調査	△	○：クレーム	○～△
無効資料調査	△	△：全文	△
技術動向調査	○	○：クレーム	○

一方で、現在のAIツールには、サーチ／ノイズ判定だけでなく、判定の確度を示すスコアリング機能が付いている。スコアの高低で読み込む公報に優先順位を付けてノイズの内容も確認することで、たとえサーチ漏れがあった場合でも、効率的に対象公報を発見することが可能となりうる。これらを踏まえ、漏れないことが最優先のクリアランス調査と無効資料調査には、△評価を付した。

また、調査タイプに応じて特許の読み込み箇所が異なるため、2つ目の観点では、読み込みが必要な記載箇所について評価している。技術動向調査とクリアランス調査は、最低限独立クレームを読み込むことで、サーチとノイズ判定が可能な調査タイプである。しかし、先行技術調査と無効資料調査は、クレームだけでは不十分なケースが多く、全文を読み込まなければならないケースが多々ある。これに対し本研究で検討したAIツールでは、データ処理や判定精度の都合上、教師データ／判定データともに、全文での判定は難しい傾向にある。したがって、先行技術調査と無効資料調査を△評価とした。

以上、2つの活用観点を元に、現状のAIツールに最も適した調査タイプは、技術動向調査であり、次点に先行技術調査とクリアランス調査、無効資料調査はその他タイプと比べて活用推奨度は低いと評価した。

また、(技術分野にも依るが、)今回検証したAIツールの判定精度は極めて高いことも事実である。「人の目で一次スクリーニングを実施

した際にノイズと判定した文献をAIツールで改めて判定し、サーチ判定の漏れがないか確認する」または「コアとなる調査範囲は人の目で精査し、本来調査すべきか迷うような周辺領域をAIツールで判定する」といった人とAIツールとの協業により、調査業務の効率化や正確性を高めることが可能と思われる。

6. おわりに

本研究では、特許調査業務の中で大きな作業負荷となっている一次スクリーニングを人からAIツールへ置き換えることができれば、特許調査業務の飛躍的な効率化に繋がるとの思いから研究テーマを選定し、本稿に掲載した教師データの取り扱いについて検討を進めた。

各仮説検証により、教師データの取り扱いを工夫することで、AIツールによるスクリーニングの判定結果が向上することが確認できた。しかし、サーチ漏れの課題を解決するための糸口はつかめていないため、今後の更なる研究による精度向上を期待したい。

本稿では紙幅の都合上、検討した全ての仮説を紹介することはできなかったが、例示した教師データの各種取り扱い手法や各種調査とAIツールとの考え方が、会員企業各社における調査業務の効率化の参考になれば幸いである。

注 記

- 1) 内閣府、「日本経済2016-2017」, 第2章第1節(2017)

本文の複製、転載、改変、再配布を禁止します。

<https://www5.cao.go.jp/keizai3/2016/0117nk/index.html>

- 2) 多田幸司, 「第四次産業革命下におけるIoTに関する現状認識－特許の観点から－」, 特許懇286号, pp.47-53 (2017)

<http://www.tokugikon.jp/gikonshi/286/286tokusyul-6.pdf>

- 3) 野崎篤志, 「特許情報をめぐる最新のトレンドー人工知能, IPランドスケープおよび特許検索データベースの進化ー」, Japio YEAR BOOK 2018, pp.60-67 (2018)

http://www.japio.or.jp/00yearbook/files/2018book/18_a_08.pdf

- 4) 大林弘明, 「データ拡張を用いた固有表現抽出の精度向上」, 言語処理学会 第25回年次大会, 発表論文集 (2019)

https://www.anlp.jp/proceedings/annual_meeting/2019/pdf_dir/P7-3.pdf (参照日: 2020.10.2)

(URL参照日は4)を除き2020年8月25日)

(原稿受領日 2020年9月1日)

