# 機械学習を用いた効率的な特許調査方法
## ―単語・文書のベクトル化方法と特許調査への応用―

Effective patent search methods using Machine Learning

花王株式会社　知的財産部／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・サイクルを見ると「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。ハイプ・サイクルの 2020 年版を図 1 に示す。「人工知能」関連技術が 11 技術に細分化されている。

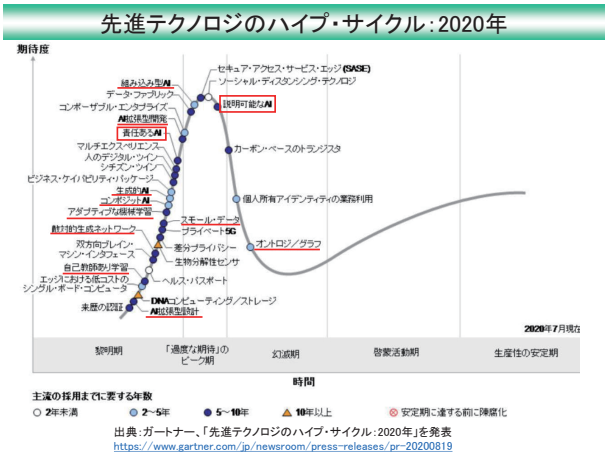



図 1　先進テクノロジーのハイプサイクル 2020 年

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、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<br>（自然言語処理、特許情報分野を注目） |
| ③ | 空想のAI（漫画、SF等）<br>例：鉄腕アトム、ドラえもん |
| ④ | 名前だけAI<br>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 「完全一致」と「最良一致」検索モデル比較

検索モデルとは、情報検索をコンピュータで実現するための仕組みである。既存の検索手法は「完全一致（exact match）」と「最良一致（best match）」に大別される [3)]。完全一致モデルでは、検索クエリで指定された条件に完全に一致する文書集合とそれ以外の文書集合を区別することが目的であり、検索された文書に順位を付けることが目的ではない。ただし検索結果を、文書番号、出願日、IPC、出願人等でソートすることはできる。キーワードや特許分類記号を組み合わせた論理式で検索クエリを構成するブーリアンモデル（Boolean model）がある。

最良一致モデルでは、検索クエリは文、あるいは一つ以上のキーワードであり、論理演算子などでキーワード間の関係は明示しない。検索クエリから抽出された検索語をベクトル化して各文書ベクトルとの類似度を計算してスコアとして、スコアの降順に文書を表示する。ベクトル空間モデル（vector space model）、確率的言語モデル（probabilistic language model）等があり商用の特許検索システムに概念検索、類似検索として従来より搭載されている。表２に「完全一致」と「最良一致」検索モデルの比較を示す。

表２　「完全一致」⇔「最良一致」検索モデルの比較

| 検索モデル | 「完全一致」 | 「最良一致」 |
|---|---|---|
| クエリ入力 | 特許分類（IPC,FI,Fターム等）、キーワード、出願人、公報番号等 | 発明の特徴を表す文、あるいは一つ以上のキーワード |
| 演算子 | AND、OR、NOT、隣接、近接 | 特に指定しない |
| 公報の抽出方法 | キーワードや特許分類記号を組み合わせた論理式に「完全に」一致する特許文書を抽出する | 入力された文またはキーワードに応じて並び替える |
| メリット | 各文書が検索された理由が明確 | ・ユーザーは文、あるいは一つ以上のキーワードを入力するだけでよい<br>・一覧の上位から閲覧すれば所望の文書を効率よく見つけられる<br>（検索された各文書は一定の基準に基づいて順位付けされる） |
| デメリット | ・キーワードや分類記号を使うには、調査対象分野や特許分類の体系に詳しくなければならない<br>・検索された公報すべてを閲覧する必要がある（検索結果に順位がつかないため） | ・順位付けの基準がユーザーにはわかりにくい<br>・何件まで査読すれば良いのか明白ではない |
| 主なユーザー | 専門家が好む傾向 | 一般ユーザーが好む傾向 |

「完全一致」と「最良一致」検索モデルの特徴やメリットとデメリットを正しく理解して使い分けることが望ましい。各種 AI 関連ツールを選んだり、使いこなす上でも非常に重要である。このことはユーザー側だけでなくツールのベンダー側にとっても重要である。ツールの間違った使い方を勧められても、逆に特性上難しいことをツール（検索システム）に求めてもお互いに不幸になるだけである。

## 4 特許調査への機械学習適応時の留意点

現在の大部分の人工知能を考える上で押さえておくべきポイントとして問題の定式化がある。問題の定式化とは解きたい問題をコンピュータが扱えるようにすることである。

この問題の定式化と特許調査への応用の概要を図２に示す。解きたい問題の把握は非常に重要である。情報検索の世界では昔から情報要求として知られている。情報検索リテラシーの入門としても必須と考える。情報要求の詳細は図３を参照されたい。特許調査に置いても何を調査したいのか明確になっていないと特許調査そのものが失敗する可能性が高まる。人工知能をこの情報要求を明確化する工程に、例えば質問応答システムとして組み込まれるとその後の検索精度の向上が期待できるが、この部分は現状では調査対象分野の経験を積んだサーチャーのレベルに達するのは次世代の言語 AI に期待すべきと考える。現時点では情報要求を踏まえて解きたい問題の定式化を行うのは人の重要な役割である。AI からの出力である処理結果の解釈・評価も重要な人の役割

である。商用の特許情報調査・分析ツールの性能評価も人の役割として重要である。現状の特許調査関連の AI ツールは残念ながら、「誰でも」、「何も考えずに」、使える魔法のような万能の AI ツールは無いとみるべきである。その根拠として最適化の分野において「万能のアルゴリズムは無い」というノーフリーランチ（NFL）定理がある。NFL 定理については後述する。

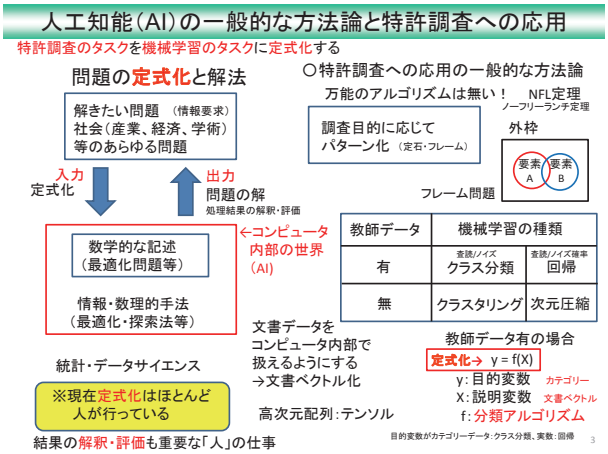



図２　人工知能の一般的な方法論と特許調査への応用

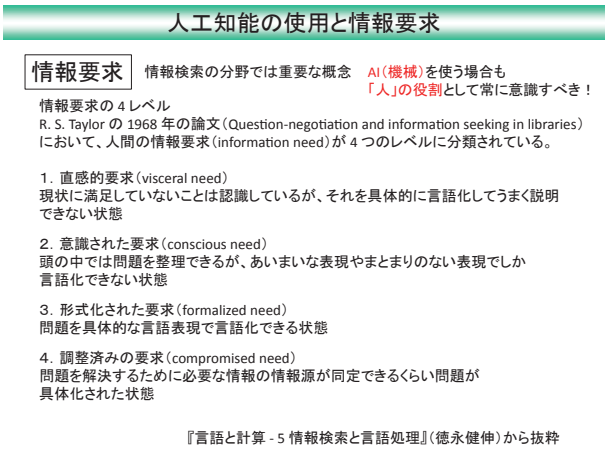



図３　Taylor の情報要求の 4 レベル

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。

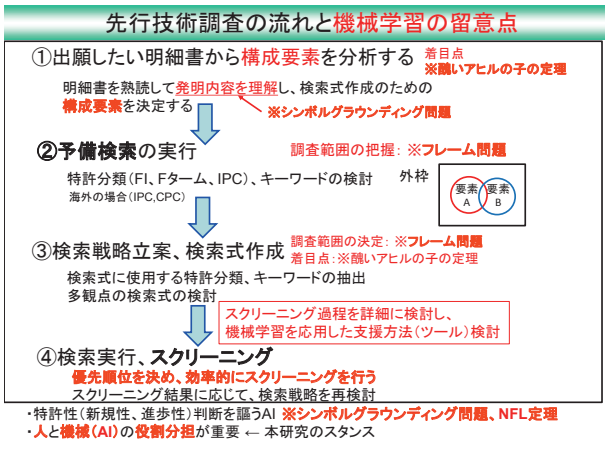



図４　先行技術調査の流れと機械学習の留意点

理想的には図４の全行程に適合したアルゴリズムの実装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、コストや開発期間を考えると実験や検証段階では注目している工程に絞って検討するアジャイル開発も選択肢として有効と思われる。

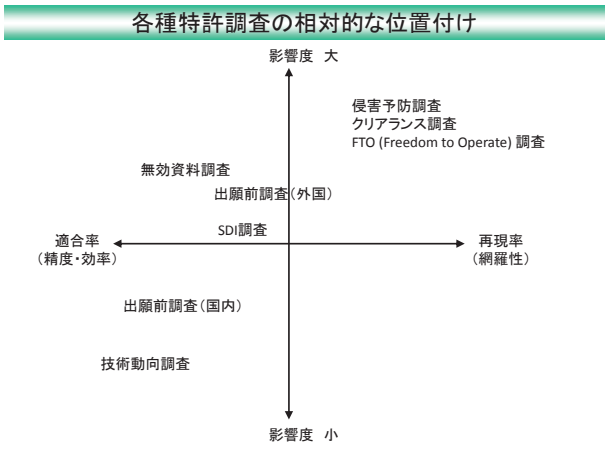



図５　各種特許調査の相対的な位置付け

図５は横軸に適合率と再現率をとり、縦軸に特許調査の失敗を想定した時の影響度とした場合の各種特許調査の相対的な位置付けを示したものである。実際の調査案件によりケースバイケースで位置付けは異なると思われるが侵害予防調査系が難易度が高いと考えられる。筆者が特許調査担当になり始めて受けた日本知的財産協会（JIPA）の特許調査の研修で、１件の特許を見逃したのが原因で、既に 50 億円の投資を行った事業から撤退せざるを得なくなった事例の新聞記事を教材に講習を受けたことを鮮明に覚えている。影響度により、侵害予防調査の対象、いわゆる「イ号製品」をどこまで想定して検索クエリを作成するかを、考える必要がある。再現率（検索漏れ防止）でむやみに調査件数を増やせばよいというものでもない。あまりにも件数が多いとスクリーニング時に見逃す可能性が高まる。海外で事業展開する場合は外国語公報を調べる必要がある。SDI 調査は侵害予防重視か技術動向重視かでポジションは異なる。

## 5　特許調査における現状の課題抽出

特許調査における現状の課題としてスクリーニング課程に関してまとめたものを図６に示す。

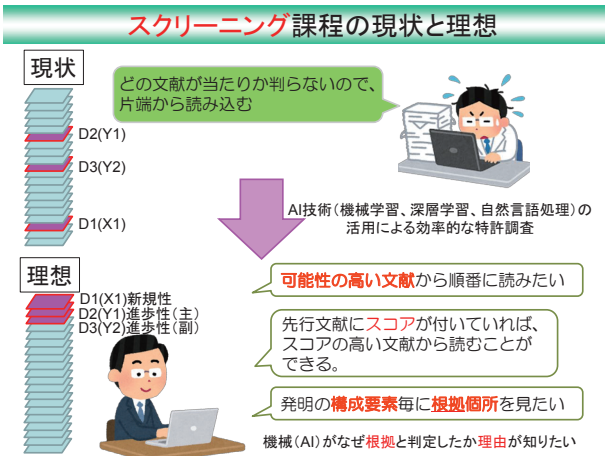



図６　スクリーニング課程の現状と理想

特許調査においてブーリアン演算により作成した集合をどの文献が当たりかわからないので片端から読み込む場合も特に初心者の場合は多いのではないかと思われる。もちろん上級者は構成要件毎の検索集合で優先順位を付けて査読したり、ブーリアン演算の集合と類似（概念）検索と組み合わせて類似度の高い順に読み込んだりと工夫している人も多数いると思われる。筆者も文書単位[4]、文単位[5]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[4),5)]。

侵害予防調査においては「類似」の順番ではなくリスクの高い順番にスクリーニングするのが合理的であるのでリスクを予測してソートすることも課題である。

図７に教師データありの機械学習を用いた特許調査の課題の一部をまとめている。特に教師データありの機械学習を特許調査へ適用することはなじみが薄いと思われる。この場合の最初の課題は「教師データの準備をどうするか？」とか「トレーニング（訓練）データとテスト（評価）データをどう分けるか？」とか機械学習によりスコア付け（回帰）あるいはカテゴリー分け（クラス分類）された出力結果を「どのように使うか？」、出力結果の性能評価を「どのようにするのか？」等と思われる。これらの課題を本稿で明らかにしていきたいと考えている。

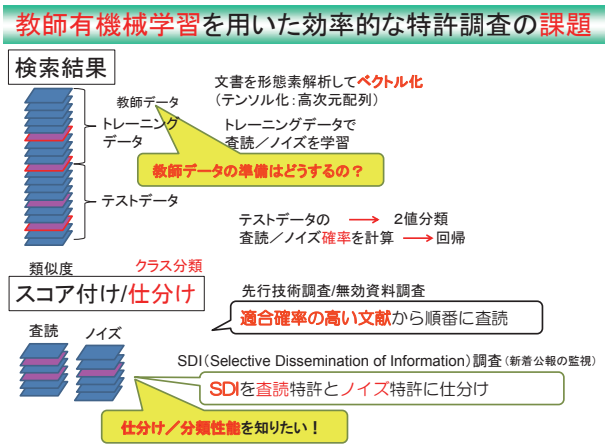



図７　教師有機械学習を用いた特許調査の課題

## 6 商用の AI 利用特許調査・分析ツールの動向

AI 利用特許調査・分析ツールの導入に際しては導入目的に照らして「出来ること／出来ないこと」を明確にして、性能、コスト等を考慮して決めると良いと思われる。AI ということで戸惑いがあるかもしれないが身近な高額商品、例えば車を購入するプロセスを想定して評価・確認項目のチェックリストを作成すると良い。表３に簡単な例を示す。性能評価の詳細に関しては後述する。

表４に国内で提供されている商用の AI 利用特許調査・分析ツールを示す。網羅的に調べたわけではないのでこの表にには含まれていないものもあると思われる。概要はベンダーの Web ページより抜粋したものである。

最近は各種解説記事や論文も多く出ている。次に一例

表３　AI 利用特許調査・分析ツールの導入ポイント例

| 評価項目 | 自動車 | AI調査ツール |
|---|---|---|
| 使用目的1 | 自家用、社有車 | 調査の種類（先行技術、動向・・） |
| 使用目的2 | レジャー、通勤 | 独自分類付与、可視化 |
| 使用者 | 本人限定、家族 | 研究員、知財部員、サーチャー |
| 購入方法 | 追加購入、買い替え | 追加契約、置き換え契約 |
| 事前評価 | 試乗 | （有償、無償）トライアル |
| 性能に関する重視ポイント | 走行性能、燃費 | 適合率、再現率 |
| エンジン | ガソリン、ディーゼル | 検索モデル、検索エンジン |
| 外観 | 色、デザイン | ユーザーインターフェイス |
| 信頼性 | 故障率 | DBの収録率、データ精度 |
| コスト | 初期、ランニング | 初期、ランニング |
| サポート | アフターサービス | ユーザー教育 |
| 営業 | 自社製品をユーザー視点で説明できるか | |
| 将来性 | 各種開発力 | 開発スピード |

を示す。AI SAMURAI[6)]、Deskbee[7)]、xlpat[8)]、サーチャーの視点からの解説[9)]も「人」の役割に関して特に参考になる。

## 7 商用の AI 利用特許調査・分析ツールの評価方法

特許調査システムの概念図とその評価方法を図８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、「完全一致」の検索モデルは入力（検索用クエリ）と出力（検索結果）の関係は理解しやすい。「最良一致」の検索モデルではユーザーが出来ることは限られている。入力に関しては「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。類似度によるスコアと文書分類の実例は後述する。文書の「類似度」もコサイン類似度，Jaccard 係数，Dice 係数，Simpson 係数、単語の分散表現を用いたテキスト間の類似度等、各種の算出方法がありそれぞれ特徴を有している[10)]。

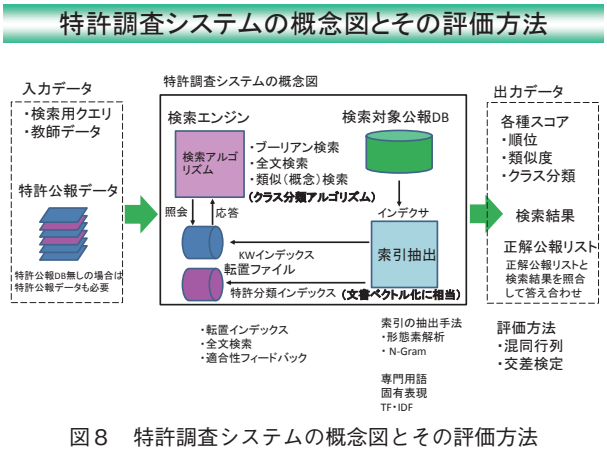



図８　特許調査システムの概念図とその評価方法

表４　商用の AI 利用特許調査・分析ツール

| No | 製品名/AI関連機能 | 概要 | ベンダー | URL |
|---|---|---|---|---|
| ① | AI SAMURAI | 日米中3カ国対応型 類似文献評価システム | AI SAMURAI | https://aisamurai.co.jp/ |
| | 先行技術調査 | 「発明内容」を文章で入力すると、AIによる類似文献評価を行います。「発明内容」から高速で国際特許分類（IPC）を認定、最も類似する文献5件を抽出し「発明内容」の類似度の高さをA～Dの４段階で評価。５つの類似文献を並べたクレームチャートを自動で生成します。 | | |
| | クリアランス調査 | 「発明内容」を入力すると、AIによる検索によって類似度が高い順に抽出された500件の特許文献のリストを約1分で生成しCSVデータをダウンロードできます。 | | |
| | 無効資料調査 | 無効にしたい特許公報の登録番号もしくは、「発明内容」と調査したい基準日を入力して検索すると類似特許の抽出を自動的に行い、AIにより無効化可能性の評価をし5つの類似文献を並べたクレームチャートを生成します。 | | |
| | AIコラボ検索 | ・AIが「発明の内容」からキーワード抽出や類義語展開、重み付けを実施して検索式を自動的に生成します。<br>・人が検索式を調整 | | |
| ② | amplified | 1億2000万件を超える世界中の特許文献をディープラーニングで学習し、全特許間における類似性を把握した独自開発のAIが、ユーザーの発明と類似する特許を数秒で発見し提示します。 | amplified ai | https://www.amplified.ai/ja/home |
| ③ | Deep Learner | Deep Learner は、ディープラーニング (Deep Learning, 深層学習) のモデルを対話的に設計し、実行するためのモジュールです。テキストに紐づく属性情報が存在している場合、その情報もモデルの学習に組み入れることができます。<br>Document Embedding アイコン<br>・Word Embedding アイコンで算出した 単語の分散表現を利用して、Simple Word Embedding-based Model（SWEM）を用いて文書のベクトル表現を作成する<br>・テキストデータ（分かち書き結果）から、頻度ベースの手法（Bag of Words,tf-idf）を利用して、文書のベクトル表現を作成する。 | NTTデータ数理システム | https://www.msi.co.jp/deeplearner/ |
| | Text Mining Studio | 簡単な操作で本格的なテキストマイニングが行えるツールです。 | | https://www.msi.co.jp/tmstudio/aboutTMS.html |
| | Text Mining Studio 類似抽出アドオン | 文章と文章の類似度を算出し、類似/非類似の分類を行うことができる、ラベリング支援ツールです。 | | https://www.msi.co.jp/tmstudio/TMSSimirarTextTool.pdf |
| | Visual Mining Studio | 簡単な操作で本格的なデータマイニングが行えるツールです。 | | https://www.msi.co.jp/vmstudio/functions.html |
| ④ | Derwent Data Analyzer | 機械学習による分類の自動化 | クラリベイト | https://clarivate.jp/products/derwent-data-analyzer/ |
| ⑤ | Deskbee | Deskbeeは、特許調査の短時間化を目的とし、 独自にAI技術を組み合わせて最も手間のかかるノイズ除去の簡略化を実現しています。 | アイ・ピー・ファイン | http://www.ipfine.com/deskbee/ |
| ⑥ | Innovation Q Plus | 世界中の主要国の特許情報を包括的にカバーしているデータベースです。IEEE が発行する文献情報など非特許コンテンツも同時に検索することが可能です。検索はセマンティックサーチを搭載しています。 | IP.com | https://ip.com/products/innovationq/ |
| ⑦ | Nomolytics | Nomolyticsとは、Narrative Orchestration Modeling Analyticsの略で、従来のテキストマイニングにクラスタリング技術のPLSA（確率的潜在意味解析）とモデリング技術のベイジアンネットワークという2つの人工知能技術を組み合わせたテキストデータの新しい分析技術です（特許登録済：特許第6085888号）。 | アナリティクスデザインラボ | http://www.analyticsdlab.co.jp/technology/nomolytics.html |
| ⑧ | Patent Explorer | 人工知能「KIBIT（キビット）」を活用し、発明の新規性・進歩性を否定する根拠となる可能性がある特許文献を、迅速に発見・抽出することによって、特許調査を効率化する特許調査・分析システムです。 | FRONTEO | https://kibit.fronteo.com/products/patent-explorer/ |
| | KIBIT | 「KIBIT」は人工知能関連技術のLandscapingと行動情報科学を組み合わせ、FRONTEOが独自開発した日本発の人工知能エンジンです。 | | |
| | Concept Encoder | 「単語と文書のベクトル化」により、解析の対象となる自然文からより多くの情報量を抽出できます。 | | https://www.fronteo.com/products/conceptencoder/ |
| ⑨ | Patentfield | AI特許総合検索・分析プラットフォーム | Patentfield | https://patentfield.com/ |
| | AIセマンティック検索 | AIによって関連する技術分野ごとに独自にクラスタリングされている | | |
| | AI分類予測 | 教師データを使用した2値分類、多値分類、多ラベル分類 | | |
| ⑩ | Patent Noise Filter | AI を使った特許自動分類サービス | アイ・アール・ディー | |
| ⑪ | Shareresearch | 特許情報提供サービス | 日立製作所 | https://www.hitachi.co.jp/New/cnews/month/2019/06/0613.html |
| | AI読解支援オプション | 課題を自動で抽出し、特許の内容把握を効率化する | | |
| | 自動分類付与オプション | 膨大な特許情報を高精度に分類する | | |
| | 技術マップオプション | 特許出願技術の動向を可視化する | | |
| ⑫ | xlpat | 特許技術とビジネスデータにAIと機械学習を適用します | xlpat labs | https://en.xlpat.com/ja/ |
| | ノベルティチェッカー&アイディエーションツール | 新規性を数分で確認します | | |
| | 無効資料調査 | 先行技術を数分で迅速に取得 | | |
| | パットディガー | ポートフォリオ内の数千件の特許をインテリジェントに評価します。 | | |
| | IPランドスケイパー | テクノロジー・ドメインにおけるIPと市場のトレンドを直感的に視覚化します。 | | |

## 8 先行技術調査の事例検討

先行技術調査の予備検索工程への応用を目的に答えが分かっており比較検証が可能な特許検索競技大会の過去問[11]を使用して事例検討を行った。特許検索競技大会の模範解答は「完全一致」型検索モデルに基づいている。検討に使用した問題を図９に示す。

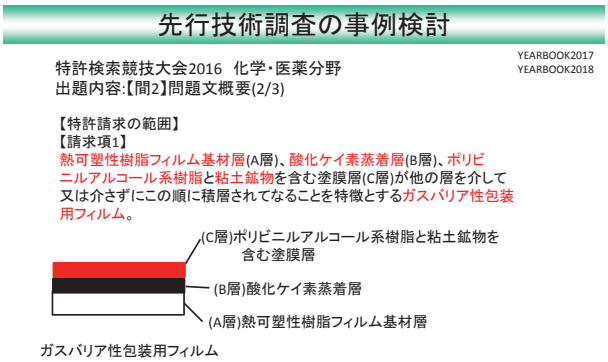



図９　特許検索競技大会 2016 の化学・医薬分野の問２

予備検索工程は検索キーである、発明の該当分野の特許分類（FI、F ターム、IPC）と特徴キーワードを求めるために行われる。ここでは、商用特許データベースとしてサイバーパテント株式会社 CyberPatent Desk と日立 Shareresearch の概念検索を「最良一致」型検索モデルとして使用した。CyberPatentDesk の概念検索画面を図 10 に示す。特徴は IPC のセクション（例：B）とメインクラス（例：B32）と２段階で概念検索の対象分野を限定できることである。この場合の概念検索の範囲は要約となる。IPC による分野別に限定しない全分野の場合、要約と請求の範囲が指定できる。図４の機械学習の留意点の「フレーム問題」に限定的ながら対応できる。この効果は後で示す。


図 10　CyberPatentDesk の概念検索画面

図 11 に日立 Shareresearch の概念検索画面を示す。特徴は分かち書き結果の特徴タームが重みと共に表示されユーザーが重みを 1 ～ 1000 の範囲で調整できる。Shareresearch の概念検索の対象は、全文、要約、請求項から選択できる。


図 11　日立 Shareresearch の概念検索画面

図 12 に CyberPatentDesk の概念検索結果と２種類の検索クエリを示す。クエリ①請求項１とクエリ②明細書記載部である。検討に用いた特許検索競技大会の過去問は 49 の正解公報が分かっている。検索結果の表の「含む正解」は概念検索の上位 1000 件に含まれる正解件数である。「含む正解」の件数はクエリ②が良い傾向である。

CyberPatentDeskの概念検索結果

| 検索集合 | 検索入力 | 検索対象 | 分野（IPC）指定 | | |
|---|---|---|---|---|---|
| | | | セクション | メインクラス | 含む正解 |
| G1 | クエリ① | 請求項 | 全分野 | | 12 |
| G1b | クエリ① | 要約 | 全分野 | | 20 |
| G2 | クエリ① | 要約 | B | B全体 | 26 |
| G3 | クエリ① | 要約 | B | B32 | 27 |
| G10 | クエリ② | 請求項 | 全分野 | | 29 |
| G11 | クエリ② | 要約 | B | B全体 | 34 |
| G12 | クエリ② | 要約 | B | B32 | 33 |

クエリ①【請求項1】
熱可塑性樹脂フィルム基材層(A層)、酸化ケイ素蒸着層(B層)、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層(C層)が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。

クエリ②　明細書記載部
(A層)熱可塑性樹脂フィルム基材層
熱可塑性樹脂フィルムであれば特に限定はなく、ポリプロピレン、ポリエチレンテレフタレート、ナイロン等のフィルムを使用することができる。また、基材層自体が熱可塑性樹脂フィルムを何層か積層したフィルムであってもよい。
(B層)酸化ケイ素蒸着層
熱可塑性樹脂フィルム基材上に酸化ケイ素蒸着層を形成したものである。酸化ケイ素の蒸着膜は透明性とガスバリア性を兼ね備える。蒸着層の厚さを調整することで、用途に応じてフィルムの透明度とガスバリア性能のバランスを取ることが可能である。
(C層)ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層
図1は(C層)のガスバリア性能の発現機構を示す図である。ポリビニルアルコール系樹脂／水／粘土鉱物を含む塗工液は、粘土鉱物の層間がへき開して分散し、塗膜を形成するとへき開した粘土鉱物が面状に配向し、高いガスバリア性を発現する。ポリビニルアルコール系樹脂は、ポリビニルアルコールの他、エチレン－ビニルアルコール共重合樹脂等のビニルアルコールを含む共重合体でもよい。
粘土鉱物としては、カオリナイト、ディッカイト、ナクライト、ハロイサイト、アンチゴライト、クリソタイル、ヘクトライト、パイロフィライト、モンモリロナイト、白雲母、マーガライト、タルク、バーミキュライト、金雲母、ザンソフィライト、緑泥石等が挙げられる。人工の粘土鉱物でもよい。

図 12　CyberPatentDesk の概念検索結果

図 13 に CyberPatentDesk の概念検索結果　再現率曲線を示す。クエリ②（明細書記載）の方がクエリ①（請求項１）より顕著に良い結果である。IPC で検索分野を絞ると良い傾向である。

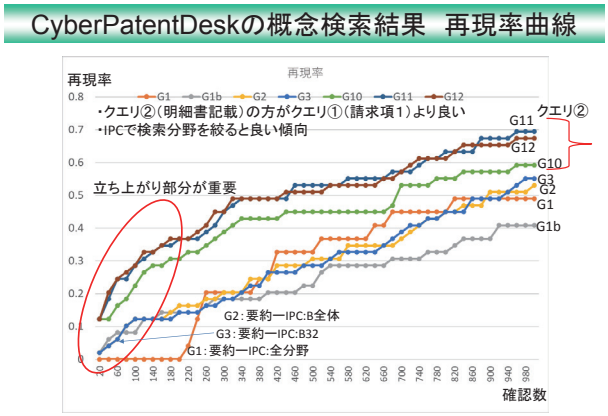
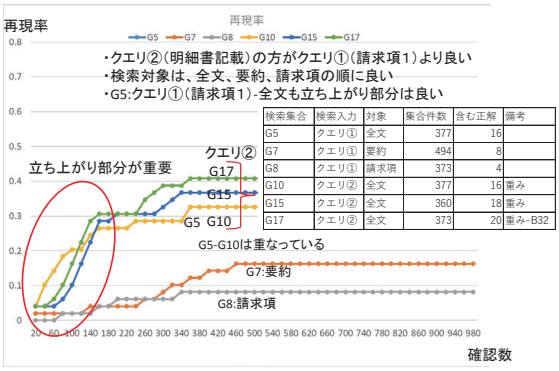



図 13　CyberPatentDesk の概念検索結果　再現率曲線

図 14 に Shareresearch の概念検索結果　再現率曲線を示す。クエリ②（明細書記載）の方がクエリ①（請求項１）より良い結果を示している。検索対象は、全文、要約、請求項の順に良い。G5：クエリ①（請求項１）-全文も立ち上がり部分は良い結果を示している。




| 検索集合 | 検索入力 | 対象 | 集合件数 | 含む正解 | 備考 |
|---|---|---|---|---|---|
| G5 | クエリ① | 全文 | 377 | 16 | |
| G7 | クエリ① | 要約 | 494 | 8 | |
| G8 | クエリ① | 請求項 | 373 | 4 | |
| G10 | クエリ② | 全文 | 377 | 16 | 重み |
| G15 | クエリ② | 全文 | 360 | 18 | 重み |
| G17 | クエリ② | 全文 | 373 | 20 | 重み-B32 |

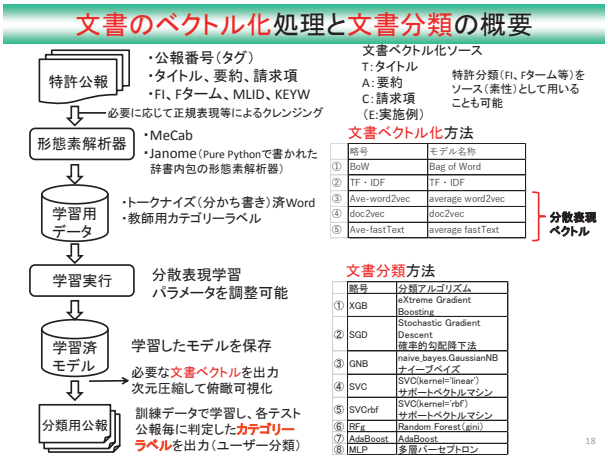
図 14　Shareresearch の概念検索結果　再現率曲線

## 9 文書のベクトル化と文書分類方法

文書のベクトル化処理と文書分類の概要を図 15 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。図 15 に文書のベクトル化処理と文書分類の概要を示す。① BoW モデル作成には scikit-learn[12] 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 15 の③～⑤の分散表現ベクトル作成には gensim[13] を使用した。文書のベクトル化手法として図 15 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。また文書ベクトルのデータソースとして F タームによる文書ベクトルも検討した。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。

表 1 の CyberPatent の概念検索結果の集合 G2+G3+G11+G12 の合計 1064 件（正解公報 38 件）を母集団として文書ベクトル化、文書分類を検討した。

文書分類方法として８種類の分類アルゴリズムを検討した。




| | 略号 | モデル名称 |
|---|---|---|
| ① | BoW | Bag of Word |
| ② | TF・IDF | TF・IDF |
| ③ | Ave-word2vec | average word2vec |
| ④ | doc2vec | doc2vec |
| ⑤ | Ave-fastText | average fastText |

| | 略号 | 分類アルゴリズム |
|---|---|---|
| ① | XGB | eXtreme Gradient Boosting |
| ② | SGD | Stochastic Gradient Descent 確率的勾配降下法 |
| ③ | GNB | naive_bayes.GaussianNB ナイーブベイズ |
| ④ | SVC | SVC(kernel='linear') サポートベクトルマシン |
| ⑤ | SVCrbf | SVC(kernel='rbf') サポートベクトルマシン |
| ⑥ | RFg | Random Forest（gini） |
| ⑦ | AdaBoost | AdaBoost |
| ⑧ | MLP | 多層パーセプトロン |

図 15　文書のベクトル化処理と文書分類の概要

図 16 に scikit-learn のアルゴリズム早見表を示す。上部のクラス分類と回帰は教師データありの機械学習アルゴリズムで、下部のクラスタリングと次元圧縮は教師データなしの機械学習アルゴリズムである。機械学習アルゴリズムは種類も多く図 16 は代表的なものである。scikit-learn にはクラス分類のアルゴリズムだけで 40 種類が実装されている。またアルゴリズムの中身も複雑

で何をしているのか分かりにくい。「見て試してわかる機械学習アルゴリズムの仕組み 機械学習図鑑」[14] はわかりやすい視覚イメージと実際に試すことで理解が進む参考文献である。

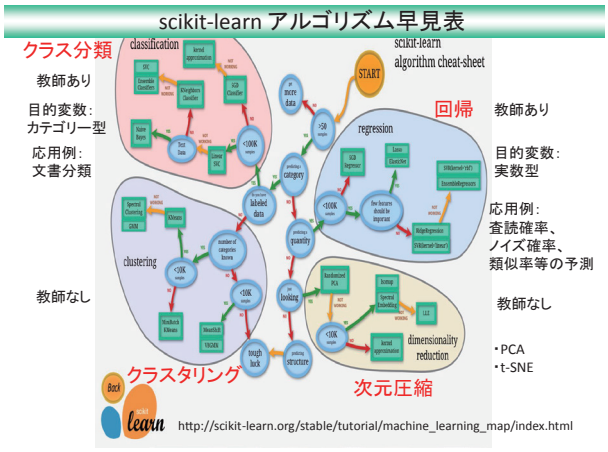



図 16　scikit-learn のアルゴリズム早見表

① XGB：XGBoost（eXtreme Gradient Boosting）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。Boosted trees は Gradient Boosting と Random Forest のアルゴリズムを組み合わせたアンサンブル学習を行う[15]。

CyberPatent の概念検索結果の集合 G2+G3+G11+G12 の合計 1064 件（正解公報 38 件）を母集団とする集合：CP1064 の５種類の文書ベクトルを t-SNE で２次元に次元圧縮した結果を図 17、図 18 に示す。t-SNE（t-distributed Stochastic Neighbor Embedding：t 分布型確率的近傍埋め込み）は、高次元データの可視化に適している次元圧縮アルゴリズムである。濃い紺色が正解公報で黄色がノイズである。どのベクトル化方法の文書の俯瞰マップも紺色の正解公報はある程度まとまっているが２～３件の孤立した公報が存在する。図 18 の破線の下の① BoW、② TF・IDF の文書ベクトルはデータソースの素性を F タームとしたものである。

F タームによる文書ベクトルの次元圧縮・可視化を図 18 下に示す。F タームによる文書ベクトルとは各特許公報に付与されている F タームを使用して特許公報文書をベクトル化したものである。単語による文書ベクトルとの違いは単語は通常公報文書に複数回現れ頻度情

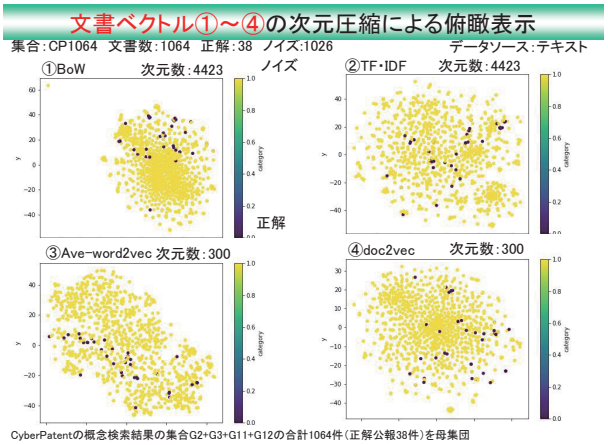



図 17　文書ベクトル①～④の次元圧縮による俯瞰表示

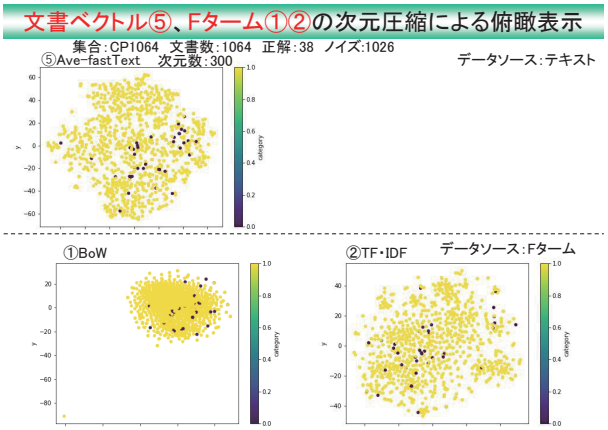



図 18　文書ベクトル⑤、F タム①②の次元圧縮による俯瞰表示

報が得られる。F タームは公報に付与される場合は各 F タム種類ごとに１個である。また F タームは多観点で付与され観点と階層による分類体系が決まっている。

F タームが公報に付与されているか否かの２値ではなく、F タームの付与のされやすさや重み付けを考慮した研究[17] がある。本稿の F タム文書ベクトル① BoW は F タームの有無の２値である。② TF・IDF は F タムが付与されている場合 TF=1 であり、IDF 項は定義通りに計算される。

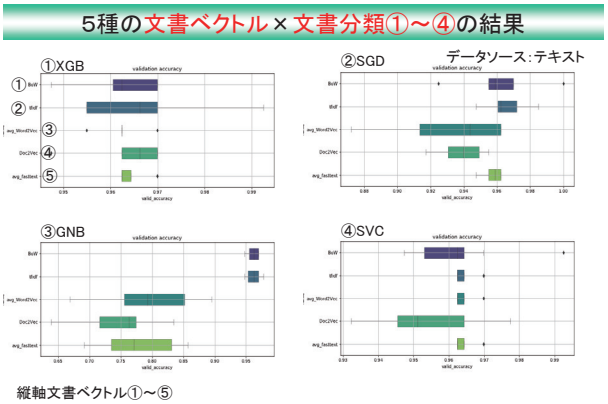



図 19　５種文書ベクトル×文書分類①～④の結果

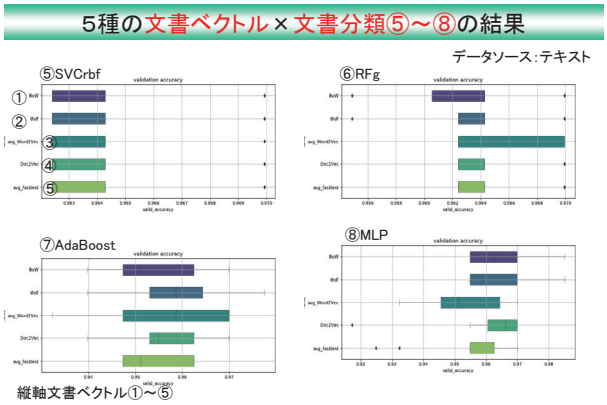



図 20　5種文書ベクトル×文書分類⑤～⑧の結果

図 19、図 20 にデータソースをテキストとした文書ベクトル①～⑤（縦軸）の文書分類方法①～⑧の８分割交差検証結果を示す。

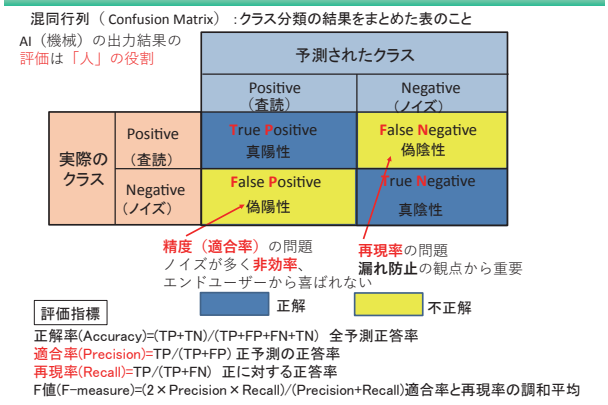



図 21　混同行列による文書分類結果の性能評価方法

図 21 に混同行列による文書分類結果の性能評価方法を示す。

## 10 分散表現を利用した言語モデルの活用

分散表現の概要を word2vec を例にして図 22 に示す。

分布仮説に基づいた文脈中の単語の重み学習（word2vec）を図 23 に示す。

図 24 に Word2vec による「粘土」の類似語抽出結果を形態素、専門用語抽出の結果と比較して示す。一番左の列がが Word2vec による「粘土」の類似語の順位である。次の類似語が実際の抽出された類似語である。類似度の降順に抽出されたリストより人手で目視により抜粋している。形態素、専門用語は Excel のシート上で Word2vec による類似語を基にサーチ機能で順位を確認している。黄色セルは形態素解析による分かち書き

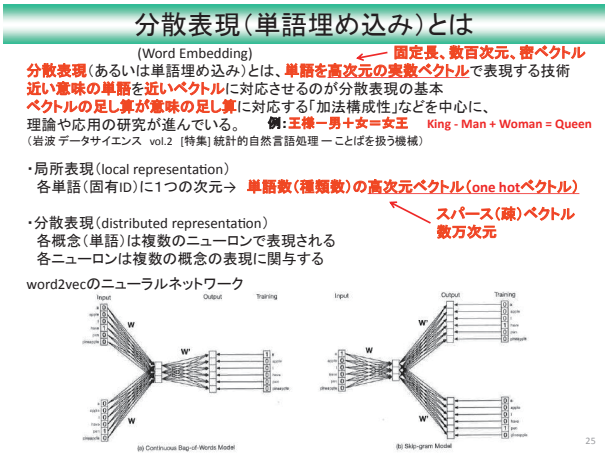



図 22　分散表現の概要

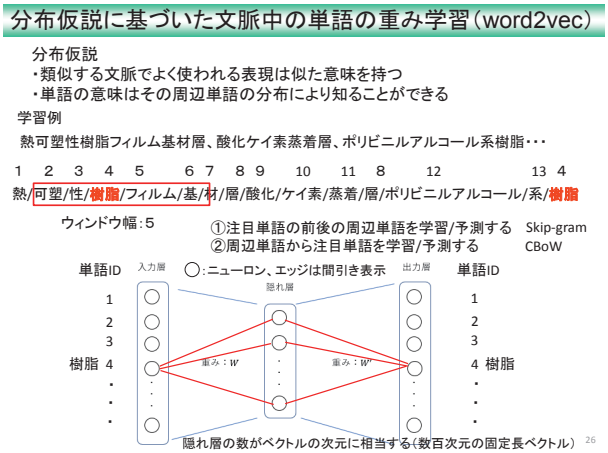



図 23　文脈中の単語の重み学習（word2vec）


| 順位 | 類似語 | 類似度 | 形態素 順位 | 形態素 頻度 | 専門用語抽出 専門用語 | 順位 | 頻度 |
|---|---|---|---|---|---|---|---|
| 1 | スメクタイト | 0.774 | 555 | 26 | スメクタイト | 1655 | 7 |
| 4 | サポナイト | 0.646 | 2101 | 4 | サポナイト | 4655 | 2 |
| 5 | ヘクト | 0.637 | 2099 | 2 | ヘクトライト | 4656 | 2 |
| 7 | スチーブン | 0.630 | 2100 | 2 | スチーブンサイト | 4703 | 2 |
| 8 | ナイト | 0.615 | 1448 | 4 | カオリナイト | 2669 | 4 |
| 9 | マイカ | 0.614 | 1449 | 4 | マイカ | 3441 | 3 |
| 11 | モンモリロナイト | 0.599 | 359 | 53 | モンモリロナイト | 246 | 52 |
| 12 | カオリ | 0.597 | 1635 | 3 | カオリナイト | 2669 | 4 |
| 14 | タルク | 0.587 | 1446 | 4 | タルク | 2691 | 4 |
| 16 | ゼオライト | 0.561 | 1175 | 7 | ゼオライト | 1652 | 7 |
| 17 | セリ | 0.554 | 2184 | 4 | セリサイト | 5112 | 2 |

| 専門用語 | 順位 | 頻度 |
|---|---|---|
| 水素型スメクタイト | 1657 | 7 |
| 合成スメクタイト | 1979 | 6 |
| スメクタイト族 | 3864 | 2 |
| スメクタイト群粘土鉱物 | 4002 | 2 |
| スメクタイト粘土鉱物 | 4740 | 2 |
| 合成マイカ | 7890 | 1 |
| カオリン | 7203 | 1 |

| 主な粘土鉱物(Wikipedia) |
|---|
| カオリナイト（高陵石） |
| スメクタイト |
| モンモリロン石（モンモリロナイト） |
| 絹雲母（セリサイト） |
| イライト |
| 海緑石（グローコナイト） |
| 緑泥石（クロライト） |
| 滑石（タルク） |
| 沸石（ゼオライト） |

図 24　Word2vec による「粘土」の類似語抽出

に失敗しているが類似語として上位に存在している。これは分布仮説に基づいた単語の重み学習の性質を良く表している。「完全一致」検索モデルと「最良一致」検索モデルの活用の仕方を考える上でのチェックポイントの一つである。

図 25 に word2vec による粘土の類似ワードの主成分分析（PCA）により次元圧縮して２次元可視化した

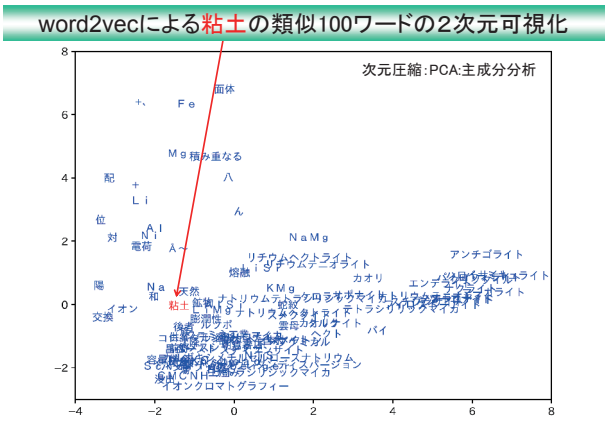



図 25　word2vec による粘土の類似ワードの２次元可視化

結果を示す。

## 11 最近の言語モデルの将来展望

表５　自然言語処理の代表的なモデル・アルゴリズム

| No. | モデル名 | モデル・アルゴリズム等 | 提案年 |
|---|---|---|---|
| ① | Word2Vec | CBOW（Continuous Bag-of-Words）モデル<br>Skip-gramモデル | 2013 |
| ② | Doc2Vec | PV-DM（Distributed Memory version of Paragraph Vector）<br>PV-DBOW（Distributed Bag of Words version of Paragraph Vector） | 2014 |
| ③ | Attention | 特定の部分に注意を向けるよう、学習させていく方法 | 2014 |
| ④ | GloVe | Global Vectors for Word Representation | 2014 |
| ⑤ | fastText | Word2Vecの提案者が作成した単語ベクトル獲得モデル | 2016 |
| ⑥ | SCDV | Sparse Composite Document Vectors | 2016 |
| ⑦ | Transformer | 「Attention is all you need」で提案されたモデル | 2017 |
| ⑧ | ELMo | Embeddings from Language Models | 2018 |
| ⑨ | GPT | Generative Pre-Training | 2018 |
| ⑩ | BERT | Bidirectional Encoder Representations from Transformers | 2018 |
| ⑪ | MT-DNN | Multi-Task Deep Neural Networks | 2019 |
| ⑫ | GPT-2 | GPTの後継 | 2019 |
| ⑬ | ERNIE | BERTを中国語に特化させたモデル | 2019 |
| ⑭ | XLNet | BERTの事前学習方法を改良したモデル | 2019 |
| ⑮ | RoBERTa | Robustly optimized BERT approach | 2019 |
| ⑯ | ALBERT | BERTを改良して軽量化したモデル | 2019 |
| ⑰ | T5 | Text-To-Text Transfer Transformer | 2019 |
| ⑱ | Reformer | Reformer: The Efficient Transformer | 2020 |
| ⑲ | GPT-3 | GPT2の後継 | 2020 |
| ⑳ | LaBSE | Language-agnostic BERT Sentence Embedding | 2020 |

表５に Word2vec 以降の自然言語処理の代表的なモデル・アルゴリズムを示す。BERT 以降の動向に注目している[18)-20)]。

表６に自然言語処理の代表的なモデル・アルゴリズムの提案論文を示す。

## 12 まとめ

本稿では「完全一致」と「最良一致」検索モデルの比較を示し、文書の BoW、TF・IDF ベクトル、分散表現ベクトル、F タームベクトルを作成し更に教師データ有りの機械学習の入力データとして文書分類を検討した。過去の独自分類結果の蓄積がある SDI 調査や技術動向

表６　自然言語処理の代表的なモデルの提案論文

| No. | モデル名 | 提案論文 |
|---|---|---|
| ① | Word2Vec | Efficient Estimation of Word Representations in Vector Space<br>https://arxiv.org/abs/1301.3781 |
| ② | Doc2Vec | Distributed Representations of Sentences and Documents<br>https://arxiv.org/abs/1405.4053 |
| ③ | Attention | Neural Machine Translation by Jointly Learning to Align and Translate<br>https://arxiv.org/abs/1409.0473 |
| ④ | GloVe | GloVe: Global Vectors for Word Representation<br>https://nlp.stanford.edu/projects/glove/ |
| ⑤ | fastText | Enriching Word Vectors with Subword Information<br>https://arxiv.org/abs/1607.04606 |
| ⑥ | SCDV | SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations<br>https://arxiv.org/abs/1612.06778 |
| ⑦ | Transformer | Attention Is All You Need<br>https://arxiv.org/abs/1706.03762 |
| ⑧ | ELMo | Deep contextualized word representations<br>https://arxiv.org/abs/1802.05365 |
| ⑨ | GPT | Improving Language Understanding by Generative Pre-Training<br>https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf |
| ⑩ | BERT | BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding<br>https://arxiv.org/abs/1810.04805 |
| ⑪ | MT-DNN | Multi-Task Deep Neural Networks for Natural Language Understanding<br>https://arxiv.org/abs/1901.11504 |
| ⑫ | GPT-2 | Language Models are Unsupervised Multitask Learners<br>https://openai.com/blog/better-language-models/ |
| ⑬ | ERNIE | ERNIE: Enhanced Representation through Knowledge Integration<br>https://arxiv.org/abs/1904.09223 |
| ⑭ | XLNet | XLNet: Generalized Autoregressive Pretraining for Language Understanding<br>https://arxiv.org/abs/1906.08237 |
| ⑮ | RoBERTa | RoBERTa: A Robustly Optimized BERT Pretraining Approach<br>https://arxiv.org/abs/1907.11692 |
| ⑯ | ALBERT | ALBERT: A Lite BERT for Self-supervised Learning of Language Representations<br>https://arxiv.org/abs/1909.11942 |
| ⑰ | T5 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer<br>https://arxiv.org/abs/1910.10683 |
| ⑱ | Reformer | Reformer: The Efficient Transformer<br>https://arxiv.org/abs/2001.04451 |
| ⑲ | GPT-3 | Language Models are Few-Shot Learners<br>https://arxiv.org/abs/2005.14165 |
| ⑳ | LaBSE | Language-agnostic BERT Sentence Embedding<br>https://arxiv.org/abs/2007.01852 |

調査へ応用すると効率的な特許調査に有用である。各学習モデルのパラメータチューニングはほとんど行っておらずデフォルト値を使用している。パラメータチューニング、教師データの分類体系の設計、特許分類を入力した BoW モデルと分散表現ベクトルのモデルの組み合わせ等で改善の余地は大きいと考える。今後の検討が楽しみである。

最近の言語モデル・アルゴリズムの進歩は速く様々な応用がなされている。Amplified は内部で Transformer 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。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。

## 13 終わりに

本報告は 2020 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）奥村学監修，「特許情報処理：言語処理的アプローチ，コロナ社，p23

4）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

5）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

6）三上　崇志ら，「特許検索タスクにおける AI システム導入の障壁―心理的障壁と組織的障壁―」
https://www.ipsj.or.jp/dp/contents/publication/43/S1103-S05.html

7）平尾　啓，「知財 AI 活用研究会の研究事例紹介」
https://doi.org/10.18919/jkg.70.7_349

8）坂元　徹「AI 技術を利用したグローバル特許調査・分析ツール「XIpat」の活用と可能性」
https://doi.org/10.18919/jkg.68.7_343

9）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

10）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

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

12）scikit-learn
http://scikit-learn.org/stable/

13）gensim
https://radimrehurek.com/gensim/ accessed 2019.03.25

14）秋庭伸也ら，「見て試してわかる機械学習アルゴリズムの仕組み 機械学習図鑑」，翔泳社，2019 年

15）XGBoost の主な特徴と理論の概要
https://qiita.com/yh0sh/items/1df89b12a8dcd15bd5aa

16）安藤俊幸，桐山勉，「分散表現学習を利用した効率的な特許調査」
https://www.jstage.jst.go.jp/article/infopro/2019/0/2019_31/_article/-char/ja
発表資料
https://sapi.kaisei1992.com/wp-content/uploads/2019/07/INFOPRO2019_A31.pdf

17）目黒光司ら，「F ターム概念ベクトルを用いた特許検索システムの改良」
http://www.lr.pi.titech.ac.jp/~meguro/NLP_2015_meguro.pdf

18）ついに読解力も人超え　「BERT 革命」の衝撃
https://xtech.nikkei.com/atcl/nxt/mag/nc/18/120400145/120400002/

19）2019 年大学入試センター試験英語筆記科目において AI が 185 点を獲得！
https://www.nii.ac.jp/news/release/2019/1118.html

20）革命かパンドラの箱か、新 AI ツール GPT-3 の波紋
https://www.itmedia.co.jp/business/articles/2007/29/news025.html