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―

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
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

1　はじめに

第３次 AI（Artificial Intelligence）ブームと騒がれ始めてから数年が経過し、新聞、雑誌、Web 等において AI の話題を見かけない日はないぐらい AI 関係の情報で溢れている[1, 2]。ただ AI が導入されると何でもできると過大な期待を抱いたり、逆に自分の仕事・職業が無くなってしまうのではないかと不安を感じている人もいるようである。一口に AI と言っても単独の AI 技術が存在する訳ではなく、また定まった AI の定義がある訳でもない[3]。最近では AI の中心技術である各種機械学習のツールがコモディティ（普通の存在）化してきており最新のライブラリが Web 上でマニュアルと共にフリーで公開されることが増えている。その気になれば誰でも入手可能である。ただし自分の業務で使いこなして有用な結果得るまでには「人」の側で習得すべき事項も多い。

特許情報の分野においても「情報の科学と技術」誌で昨年に続き 2018 年 7 月号（68 巻 7 号）でも「特集：特許情報と人工知能（AI）-II」の特集が組まれている[4, 5]。

筆者は AI のコア技術である機械学習を用いて研究員の特許調査を効率化して生まれるゆとり時間を研究員本来のクリエイティブな活動に充てて新しい価値創出支援を目的に機械学習を試行している。その経験からまず現状の「AI」に何ができて、何ができないのかの理解が必須であり、人が行うべき事、「AI」に任せた方が良い事の役割分担を意識する重要性を実感している。現在「AI」と呼ばれている複数の技術にはそれぞれ各種原理的な限界も存在しこれらを理解して「AI」を正しく使い分ける必要がある。更に人も「AI」もそれぞれの特性に合わせて育成する観点が重要なことに日々ことあることに痛感させられている。

本稿では特許調査・解析の実務に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。前半では昨年の Japio YEAR BOOK 2017 の論文と同じ例題で特許検索競技大会 2016 の化学・医薬分野の問 2（ガスバリア性包装用フィルム）を使用し機械学習の先行技術調査への適用を念頭に検討した[6, 7]。後半では技術動向調査を念頭にインクジェットインクのサンプル集合で検討した。

2　特許調査への機械学習適応時の留意点

現在、様々なベンダーからいろいろな「AI」製品が提供されており謳い文句も様々である。いろいろな「AI」製品のパンフレット等には良いことしか書かれていない場合が往々にしてあるが人工知能の分野には昔からいろいろな難問が存在している。これらの難問を知ることで現状の「AI」には原理的な限界が存在することが理解できる。AI 活用における留意点として以下に重要なものを述べる。

（1）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。

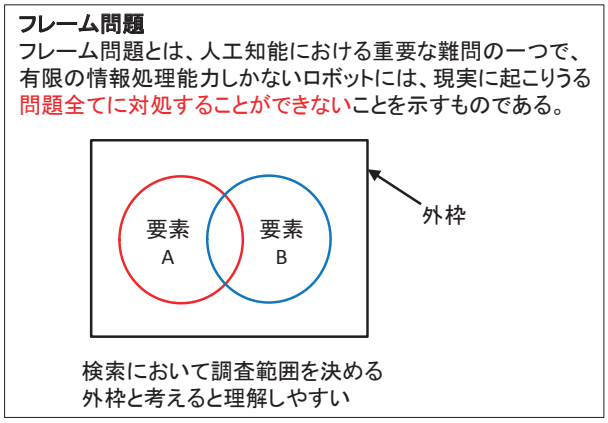


図 1　フレーム問題

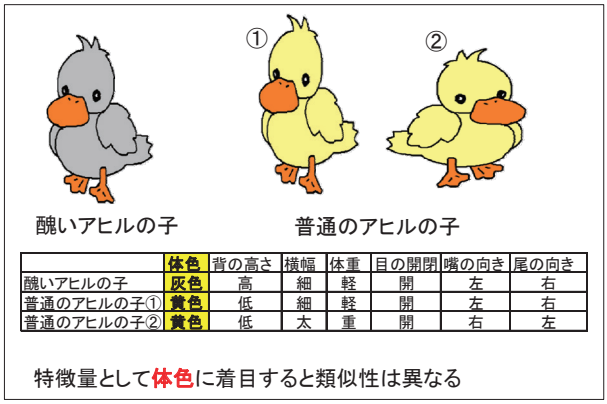
他社権利の侵害防止を目的に行うクリアランス調査でも調査範囲の決定は重要である。先行技術調査とは異なる観点からリスクと調査効率のバランスを考慮して調査範囲を決定する必要がある。

(2) ノーフリーランチ定理（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。この定理は数学的に証明されており解こうとする最適化問題に対する学習アルゴリズムに万能なものはないので問題にあったアルゴリズムを選択したり設計することの重要性を説いている。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) 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。



	体色	背の高さ	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

特徴量として体色に着目すると類似性は異なる

図 2　醜いアヒルの子の定理

「醜いアヒルの子を含む n 匹のアヒルがいるとする。このとき醜いアヒルの子と普通のアヒルの子の類似性は任意の二匹の普通のアヒルの子の間の類似性と同じになる」という定理。各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ある。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、例えば向いている方向、背の高さ、体重、目を開いている／閉じている等々の特徴量を増やすと類似性で区別できなくなる。

人は子供でも例えばリンゴとバナナの識別は容易である。たとえ黄色いリンゴとバナナの識別も容易にできる。これがディープラーニング登場前の「AI」では意外に難しい。「特徴量エンジニアリング」と呼ばれる特徴量の選択手法を用いて専門家が注意深くチューニングした機械学習が従来より行われおり有効な方法である。ディープラーニング（深層学習）では従来の専門家による特徴

量抽出が自動的に行われる。ただしディープラーニングには大量の学習データと計算能力が必要となる。計算能力は GPU（Graphics Processing Unit）の使用で大幅に改善される。

（4）シンボルグラウンディング問題

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。

3 先行技術調査への機械学習の応用

先行技術調査への機械学習の応用例として特許検索競技大会の問題を例題にして検討を行った。図３にフィードバックセミナー資料より先行技術調査の流れを示す。機械学習の先行技術調査過程への適用例として調査範囲の確定、検索キー（特許分類、検索キーワード）の抽出、スクリーニング支援（要査読かノイズの仕分け等２値分類、査読の優先順位をレコメンドするスコアリング）等が考えられる。各調査プロセスの性質に応じた機械学習の個別のアルゴリズムの応用が考えられるが本稿では発明の構成要素に注目したスクリーニングを検討した。構成要素の分析（抽出）は醜いアヒルの子の定理を念頭に「人」が行うことを前提にしている。予備検索を実行して調査範囲を決定するのはフレーム問題で述べたように調査効率（精度）と検索漏れ（再現率）とのバランスを考えて決める必要がある。

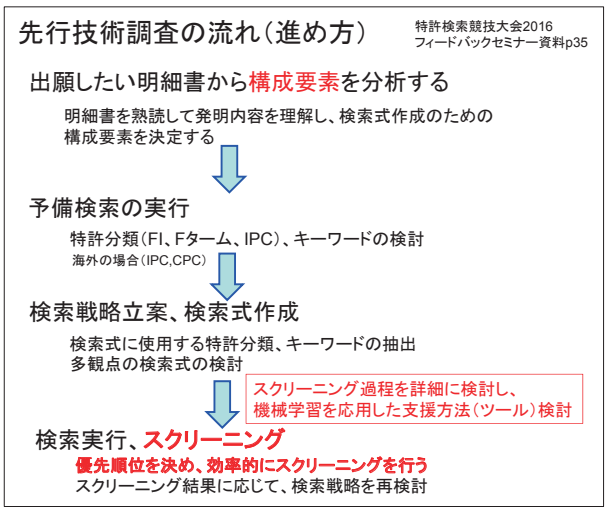


図３　先行技術調査の流れ

特許検索競技大会 2016 の化学・医薬分野の問２（ガスバリア性包装用フィルム）を例題として選択し各種の検討を行いやすいデータセットを作成した。商用特許データベースとして日立の特許情報提供サービス「Shareresearch」[8]、NRI サイバーパテントデスク２[9]、を使い検索競技大会の問題文（図４）の請求項１を入力して概念（類似）検索を行い各々上位 376 件と正解公報 49 件の和集合 746 件をデータセットとした[6]。2017 年作成のデータセットをそのまま使用した。

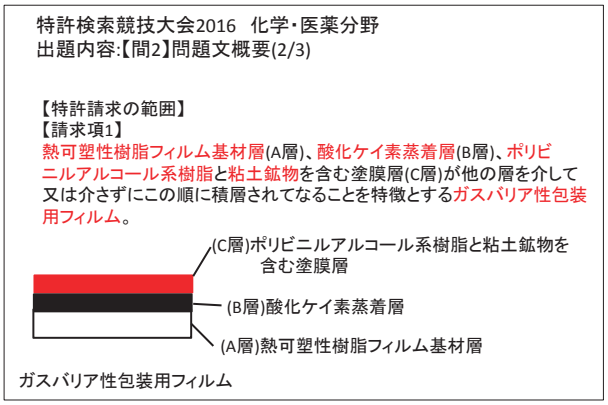


図４　特許検索競技大会 2016 の化学・医薬分野の問２

近年、単語のベクトル化手法として word2vec[10]、文書のベクトル化手法として doc2vec[11] が考案されこれらの発展手法が各種登場してきている。

図５に doc2vec による文書のベクトル化処理の概要を示す。2017 年の検討では doc2vec を使用して文書単位にベクトル化して類似文書のスクリーニング検討を行った。

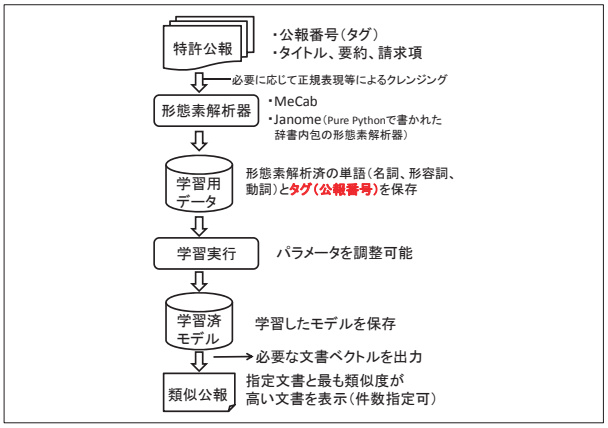


図５　doc2vec による文書のベクトル化処理の概要

本稿では改良ポイントとして下記①～③の検討を行った。

改良ポイント

①公報を文単位に分解してタグ付け

②実施例追加

③クエリ：請求項単位、構成要素単位等

①は公報を文書（documents）単位から文（sentence）単位でタグ付けしてベクトル化した。②はタイトル、要約、請求項に実施例を追加した。③はクエリとして請求項単位、発明の構成要素単位等任意のクエリを入力できるようにした。タグ付けの詳細は下記のように公報番号に記載部分の通し：文番号とした。

タグ付け詳細

公報番号_記載部分：文番号

例：P2001-123456_C6

記載部分略号は下記のように決めた。

記載部分略号

T：タイトル

A：要約

C：請求項

E：実施例

上記のように公報を文単位に分解してタグ付けしクエリも任意の文あるいは句単位で入力可能なようにすることで発明の構成要素単位で根拠個所の抽出が期待できる。

図６に検索競技大会の模範解答の発明の構成要素分節例を示す。

熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、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。

正解例と解説:【問2】(1)構成要素分析

(1)調査依頼された請求項1に対して、検索すべき技術の構成要素(概念)を記述しなさい。

記号	構成要素(概念)	重み1	重み2
a	熱可塑性樹脂フィルム基材層	10%	5%
b	酸化ケイ素蒸着層	20%	30%
c	ポリビニルアルコール系樹脂を含む塗膜層	10%	10%
d	塗膜層に粘土鉱物を含む	30%	30%
e	他の層を介してまたは介さずにこの順に積層	5%	1%
f	ガスバリア性	15%	19%
g	包装用フイルム	10%	5%

※構成要素の分け方は本例に限定しない　　同じ重みだと1/7=14.3%

図６　構成要素分析（検索競技大会の模範解答例）

図７に分布仮説に基づいた文脈中の単語の重み学習のword2vecの模式図を示す。doc2vecはword2vecを拡張してタグ付き文書を入力する。固定長ベクトルは単語（文書）間の距離（類似度）計算や次元圧縮による可視化、別のネットワークの入力に利用できる。

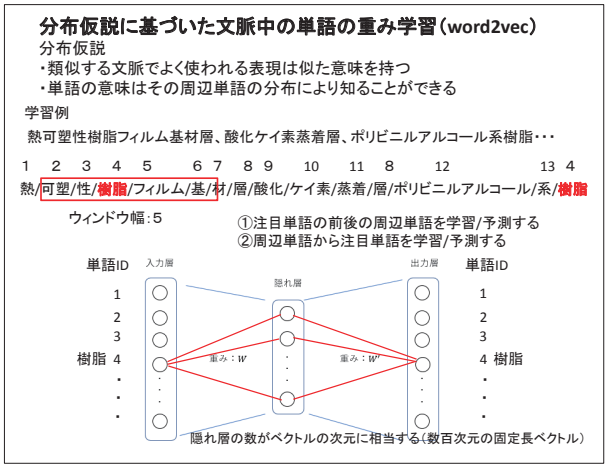


図７　分布仮説に基づいた文脈中の単語の重み学習

図８に「文」単位での類似度計算による再現率曲線を示す。確認数が少ない立ち上がり部では「文単位要素 a-g の平均値」が最も良い再現率を示している。確認数の全体を通して「文書」単位の類似度計算結果も良い結果を得ているが「文」単位の類似度計算は発明の構成要素毎に根拠個所を特定したりあるいは適合判別の可能性が考えられる。SR は日立の特許情報提供サービス「Shareresearch」の類似検索結果をベースラインとして比較している（以下同様）。

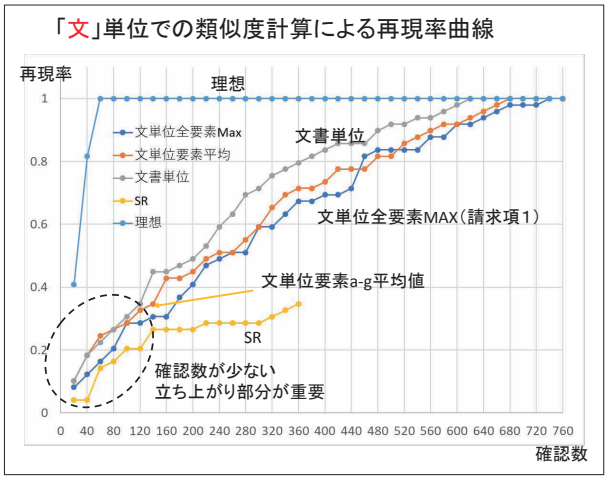


図８　「文」単位での類似度計算による再現率曲線

図９に発明の構成要素の重み付け検討結果を示す。重み付けは図６の重み１、重み２を使用した。確認数の後半で再現率への効果が大きいが重要な確認数の前半で再現率を若干悪化させる。i100 は重み付けを変えていない曲線である。i100 の意味は doc2vec のハイパーパラメータの一つである学習回数である。

図１０、図１１に文の分節とクエリ拡張の影響を示す。クエリに主要な構成要素を含む実施例を入力した場合が最も良い結果になっている。

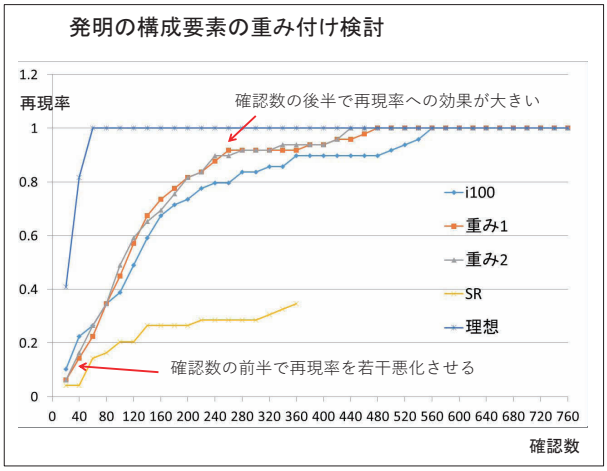


図９　発明の構成要素の重み付け検討

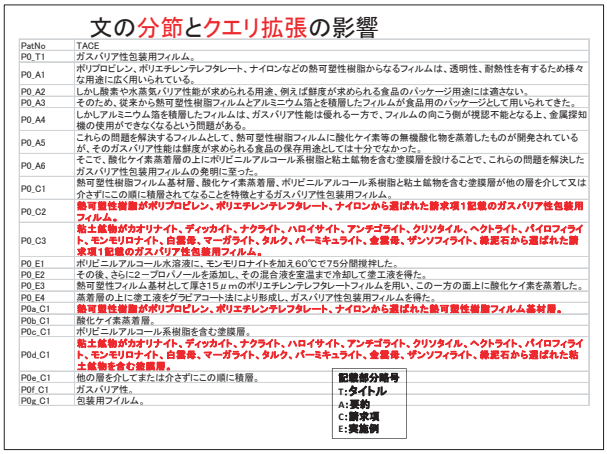


図 10　文の分節とクエリ拡張の影響（クエリ）

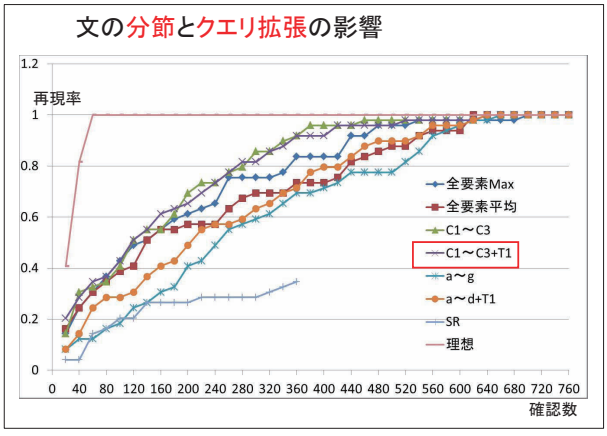


図 11　文の分節とクエリ拡張の影響（結果）

図 12 に発明の構成要素毎の根拠個所（文）抽出結果を示す。結果の解釈に関して下記の注意を要する。

①構成要素 a, f, g の順に寄与が大きいが発明の特徴量としてはあまりふさわしくない→構成要素の重み付けである程度の改善が見込める

②適合は人（筆者）が判定している→教師あり学習で改善が見込める（教師あり学習による適合判定については現在検討中である。）

各構成要素の最大類似度「文」の平均値で順位2位P1998-076325　正解公報

構成要素	記載部	類似度	該当文	適合
a	E94	0.728	さらに、これらの熱可塑性樹脂基材は、透明であることが好ましい。	○
b	E99	0.595	金属及び／または金属酸化物は特に限定されないが、アルミニウム、ケイ素、亜鉛、マグネシウムなどの金属及び／または金属酸化物であることが好ましい。	○
c	E55	0.523	さらに、本発明では塗膜中に架橋剤を含んでいてもよい。	×
d	E125	0.489	塗膜の構成成分を含んだ塗剤は、溶媒に無機板状粒子が均一に分散もしくは膨潤しかつ水溶性または水分散性ポリマーが均一に溶解もしくは分散した溶液が好ましい。	○
e	E140	0.511	フィルム走行装置を具備した真空蒸着装置内にフィルムをセットし、冷却ドラムを介して走行させる。	×
f	E217	0.714	ガスバリア性に特に優れるフィルムが得られた。	○
g	T1	0.633	ガスバリアフィルム及び包装材料	○

構成要素　平均値：0.599
a:熱可塑性樹脂フィルム基材層
b:酸化ケイ素蒸着層
c:ポリビニルアルコール系樹脂を含む塗膜層
d:塗膜層に粘土鉱物を含む
e:他の層を介してまたは介さずにこの順に積層
f:ガスバリア性
g:包装用フィルム

記載部分略号
T:タイトル
A:要約
C:請求項
E:実施例

図 12　発明の構成要素毎の根拠個所（文）抽出結果

発明の構成要素 b：「酸化ケイ素蒸着層」の該当文は「金属及び／または金属酸化物は特に限定されないが、アルミニウム、ケイ素、亜鉛、マグネシウムなどの金属及び／または金属酸化物であることが好ましい。」であり、直接「酸化ケイ素」の記載はないが「ケイ素の金属酸化物」が該当する。同様に発明の構成要素 d：「塗膜層に粘土鉱物を含む」の該当文は「塗膜の構成成分を含んだ塗剤は、溶媒に無機板状粒子が均一に分散もしくは膨潤しかつ水溶性または水分散性ポリマーが均一に溶解もしくは分散した溶液が好ましい。」であり「塗膜」と「無機板状粒子」が該当する。

doc2vec では直接的な記載がなくても文脈中の単語の並びを反映した学習を行い類似の文を提示しており非常に興味深い結果が得られた。

4　ニューラルネットワークによる学習の基礎検討

本格的なディープラーニングの特許調査への応用を検討する前にネットワーク層が少ないパーセプトロン（Perceptron）、多層パーセプトロン（Multilayer perceptron、略称：MLP）を使った基礎的な検討を行った。MLP は線形分離可能ではないデータを識別できる。

パーセプトロンは、人工ニューロンやニューラルネットワークの一種である。心理学者・計算機科学者のフランク・ローゼンブラットが 1957 年に考案した。

視覚と脳の機能をモデル化したものであり、パターン認識を行う。シンプルなネットワークでありながら学習能力を持つ。1960 年代に爆発的なニューラルネットブームを巻き起こしたが、1969 年に人工知能学者マー

ビン・ミンスキーらによって線形分離可能なものしか学習できないことが指摘されたことによって下火となった。他の研究者によってさまざまな変種が考案されており、ニューロン階層を多層化し入出力が二値から実数になったボルツマンマシン（1985 年）やバックプロパゲーション（1986 年）などによって再び注目を集めた。2009 年現在でも広く使われている機械学習アルゴリズムの基礎となっている[12]。

図 13 に Chainer（チェイナー）[13] の多層パーセプトロンによる階段関数の学習を示す。ニューラルネットワークの各エッジの重みを少しずつ調整し、正解ラベル（教師データ）との誤差を小さくする事を、「学習」と呼ぶ。階段関数をほぼ学習するのに 10 万回の「学習」を要している。Chainer は Preferred Networks で開発されているニューラルネットワークの計算および学習を行うためのオープンソースソフトウェアライブラリである。

図 14 に Iris（アヤメ）の品種分類の Iris data set を示す。Iris data set は scikit-learn[14] に含まれている。scikit-learn は Python のオープンソース機械学習ライブラリである。

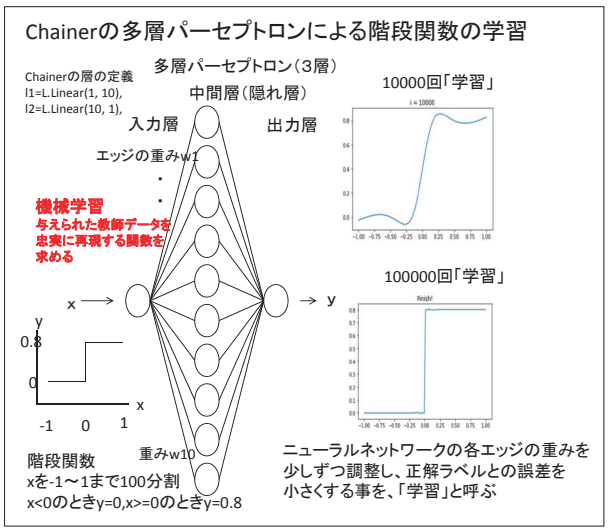


図 13　多層パーセプトロンによる階段関数の学習

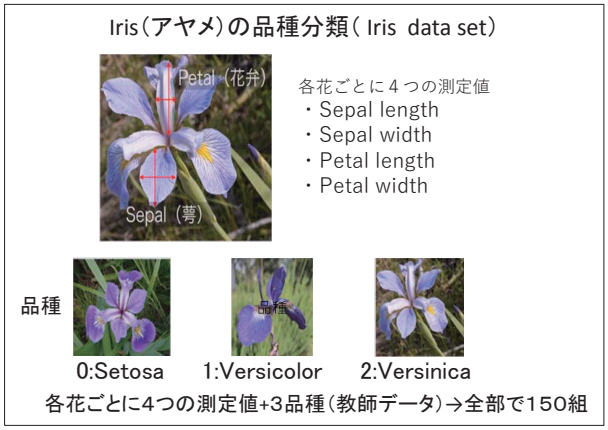


図 14　Iris（アヤメ）の品種分類（ Iris data set）

図 15 に Chainer によるニューラルネットワークの訓練と訓練済モデルの品種分類実行例を示す[15]。Iris data set の半分の 75 件でニューラルネットワークを訓練して残り 75 件を品種分類した結果 Acuuracy：96.0 %（正答率）であった。ニューラルネットワークの中間層が 2 層以上ある時は（狭義の）ディープラーニング（深層学習）と呼ばれる。

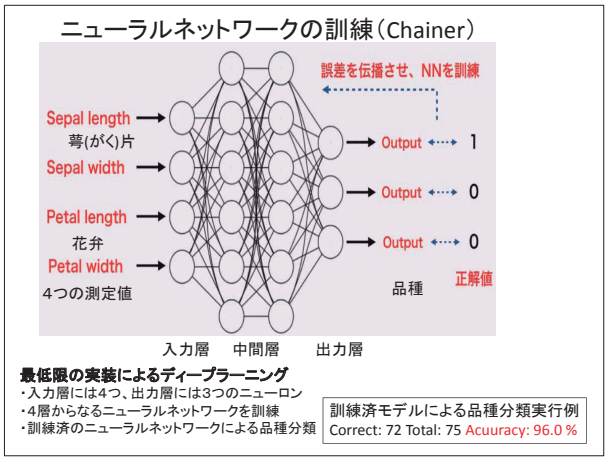


図 15　Chainer によるニューラルネットワークの訓練

5 ディープラーニングによる先行技術調査の予備検討

NTT データ数理システムの Visual Mining Studio[16] 8.4 の Deep Learning アドオン（Deep Learner）[17] を使用して先行技術調査の予備検討を行った。Deep Learner の機能は多層ニューラルネットワークによる教師あり学習・教師なし学習を行う機能である。教師あり学習では、カテゴリ値の予測については判別モデル、数値の予測に対しては回帰モデルを構築する。教師なし学習では、データを次元圧縮し低次元化された表現を得ることができる。 入力するデータは、1 行 1 件のデータである通常のテーブル形式に加え、可変長の時系列データや同社のテキストマイニングツール Text Mining Studio[18] で分かち書きされたテキストデータも扱うことができる。図 16 にデータタイプ別の教師あり、なしの学習の処理内容と特色を示す。画像データは同社の AutoDL で分類モデルを構築できる。

図 17 に Deep Learning アドオン（Deep Learner）の設定画面を示す。最初に Iris data set を csv ファ

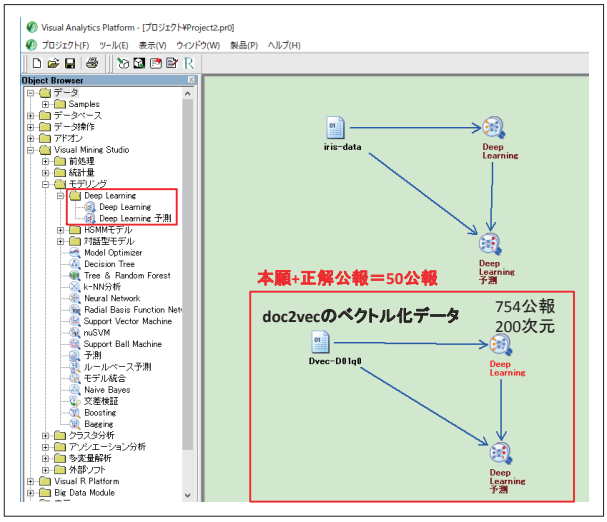


図 16　Deep Learning アドオン (Deep Learner) の設定画面

データタイプ・学習別処理内容と特色

	教師あり学習	教師なし学習
テーブル	分類分析・回帰分析	次元圧縮
時系列	系列を考慮した 分類分析・回帰分析 例）時系列センサーデータ等	可変長の系列データから 固定長の次元圧縮表現を獲得
テキスト	テキストの分類分析	
特色	目的変数は数値、カテゴリを 問わず複数指定可能	次元圧縮により得た表現を VMSの他のアイコンで使用可能 例）クラスタリング、可視化等

図 17　Deep Learner のデータタイプ・学習別処理内容と特色

イルより読込み動作確認を行った。次に doc2vec により文書単位でベクトル化したデータを csv ファイルで読込みモデル選択の用途を「予測」としてデータ形式を「テーブル」、目的変数を「正解ラベル（整数）」、説明変数を doc2vec の 200 次元ベクトル（実数）を設定した。ニューラルネットワークのモデルデザインは入力層、中間層を全結合層 1（出力次元数 300）、全結合層 2（出力次元数 2）、出力層とデザインした。活性化関数を ReLU、Dropout Ratio を 0.0 とした。

Deep Learning アドオンに本願＋正解公報 49 件の計 50 件を含むトータル 754 件のラベル付き（教師）データ（doc2vec による 200 次元の固定長ベクトル）を入力して学習させた。予備検討として全データを Deep Learning 予測を行った。

表 1　多層ニューラルネットワークによる教師あり学習

	正答数	誤答数	正答率	精度	再現率
正解公報	30	20	60.0%	100.0%	60.0%
ノイズ公報	704	0	100.0%		
	734	20	97.3%		

上記結果は doc2vec も Deep Learning アドオンのニューラルネットワークもハイパーパラメータのチューニングを行っていないが予備検討として精度 100% という興味深い結果が出ている。ただし検索漏れ防止（再現率）の観点からはパラメータチューニングや他の手法との併用等の検討課題も示している。

6　ディープラーニングによる技術動向調査の予備検討

近年世界の特許・実用新案出願の約 6 割を中国が占める状況にある。中国特許調査も念頭に技術動向調査としてインクジェットインク分野を対象に予備的に検討した。Questel 社のグローバル特許データベース Orbit.com[19] のファミリー単位の FamPat データベースで下記検索式 2501 件（2018.08.06 時点）を母集団とした。

（4J039GA24）/FTM AND（CN）/PN 2501 件

（CN に限定しない 4J039GA24/FTM　18818 件）

4J039GA24 は F タームのインクジェットインクである。JP 特許の F タームと発行国 PN に CN を指定しているので JP と CN にファミリーがある 2051 件が抽出されている。図 18 は Orbit. の分析モジュールの教師なし学習であるクラスタリングを使用したフォームツリーマップである。

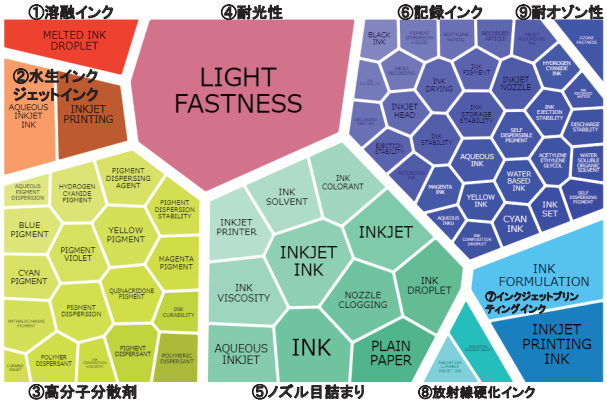


図 18　Orbit のフォームツリーマップ

①～⑨はグループ表示を ON にした時のグループ名称の日本語表記である。実際には英語キーワードで表示される。説明の便宜上日本語で表記する。このチャートではシステム内部で自動的に 9 クラスターが指定されクラスタリング計算が行われる。新しい公報が収録されたり検索式を変更したりして母集団が多少変わるとクラ

スタリング状況が変化する。例えば母集団の 27 件の増加で③高分子分散剤が「青顔料」、⑤ノズル目詰まりが「普通紙」、④耐光性の表示が無くなり「非水性インク」、「硬化性」が新しいクラスターとして表れている。クラスタリングアルゴリズムは教師なし機械学習なのである意味当然ではあるがユーザー視点としては特許件数が多少増減しても同じグループ表示が欲しい人も多い。またクラスターのカテゴリーとしてインクの種類のカテゴリーと効果·性能のカテゴリーには多少違和感がある。ユーザーが定義あるいは公報を仕分けしたクラス分類（文書分類）が使えると便利である。

図 19 に公報データをベクトル化して次元圧縮して 2 次元で表わしたランドスケープマップを示す。各公報が色つきのドットで表示されクラスター毎に色分けされている。マウスで領域を指定するとその集合がリスト表示される。表示されたリストを見て内容を確認して自分で選択した表示領域にラベル（コメント）を付けることができる。ランドスケープマップのクラスター分けは内部で自動的に行われユーザー側では関与できない。図 18 のフォームツリーマップのクラスタリングの指摘と同様にランドスケープマップ上でユーザーが定義あるいは公報を仕分けしたクラス分類（文書分類）が使えると便利である。7 章で次元圧縮とクラスタリング応用のマップ利用の注意点を述べ 8 章で教師データありの文書分類との併用を検討する。

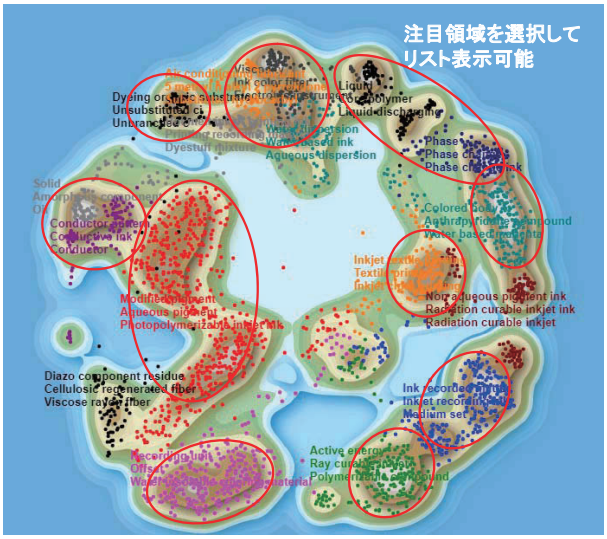


図 19　Orbit のランドスケープマップ

7 機械学習（AI）の出力利用を考える上での注意点

2 章では「特許調査への機械学習適応時の留意点」として人工知能分野で昔から難問とされてきた原理的な限界の観点から留意点を述べた。本章では次元圧縮やクラスタリング応用マップを利用する上でのユーザーの観点からの注意点を述べる。

人間と機械学習（AI）の役割分担を良く考えることが大前提である。人間の役割を列記する。

人間の役割

- ゴール（目的）を決める
- ゴールへのルートを決める / 選択する
- 出力を利用して判断する / アクションする
- 出力（マップ）の信頼性、精度、特徴を理解して使用する
- 出力を得るアルゴリズムの特徴を理解する
- 気付き（セレンディピティ）を得る
- マイニングするものを決める
- 正解（教師データ）を準備する
- 間違いに気付く / 修正する

図 20 に地図の図法とその特徴を示す。地球は 3 次元の球状であるが 2 次元平面上の地図に次元圧縮すると各種図法に特有の歪みが生ずる。各図法の特徴を理解して使用することが重要である。ランドスケープマップも各公報をベクトル化して高次元空間上の配置を次元圧縮して 2 次元平面上にマッピングしている。商用のツールでは詳細は非開示のブラックボックスのことが多いが次元圧縮やクラスタリングアルゴリズムは各種存在する。自分で行う場合は特に使用目的や用途に合わせて特徴を理解して利用するのが重要である。

図 21 に scikit-learn のアルゴリズム早見表を示す。

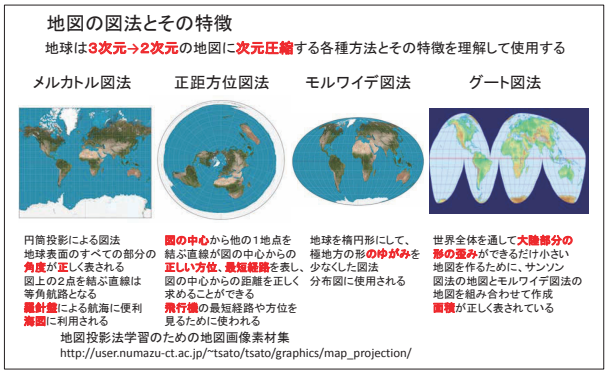


図 20　地図の図法とその特徴

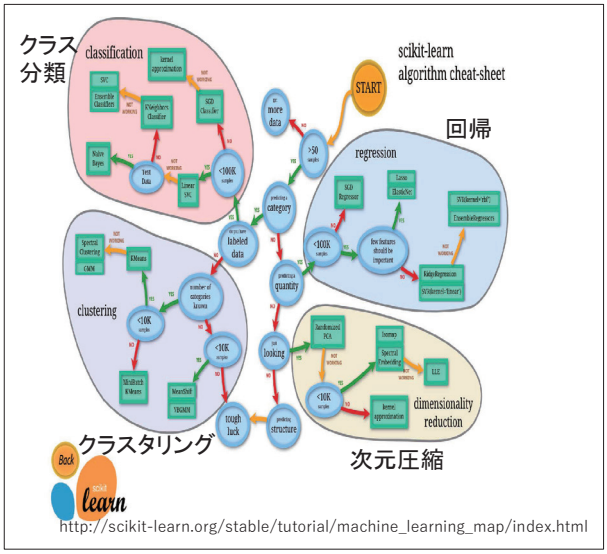


図 21　scikit-learn のアルゴリズム早見表

scikit-learn algorithm cheat-sheet[20] のページから各アルゴリズムの更に詳細な情報にアクセスできる。

NTT データ数理システムの「Visual Mining Studio チートシート」[21] も公開されている。

8 教師データありの文書分類と次元圧縮による可視化

Apache MXNet[22] というディープラーニングフレームワークを使用して教師データありの文書分類を検討した[23]。MXNet は Python を始め R、Scala、Julia、Perl、C++ 等多くののプログラミング言語に対応している。予備検討として 15 カテゴリーの記事が収録されている「Wikipedia 日英京都関連文書対訳コーパス（Version 2.01）」[24] を使用して文書分類の検討を行った。「Wikipedia 日英京都関連文書対訳コーパス」は、高性能な多言語翻訳、情報抽出システム等の構築を支援することを目的に作成された日英対訳コーパスである。国立研究開発法人情報通信研究機構が Wikipedia の日本語記事（京都関連）を英語に翻訳し、作成したものである。文書分類のアルゴリズムはディープラーニングの一種である 1 次元 CNN（Convolutional Neural Network）[25] を使用した。トレーニング文書で教師データ（15 カテゴリー：学校、鉄道（交通関連）、旧家、建造物、神道、人名、地名、伝統文化、道路、仏教、文学、役職・称号、歴史、神社仏閣、天皇）を学習させ、テスト文書を各カテゴリーに分類して正解数をカウントする。トレーニング文書：9877 記事、テスト文書：4234 記事で行った。テスト文書の分類結果は Accuracy＝0.799953 であり約 80%正解率であった。

文書のベクトル化、次元圧縮による可視化として下記 3 種類を検討した。テスト文書を各カテゴリーに分類したクラス毎にベクトルの標準偏差を計算して全体の標準偏差で割り score を計算した。クラス毎のベクトルのばらつきが全体のばらつきに対して小さいので score は小さい方がクラス毎に良くまとまっていることを示す。また 2 次元に次元圧縮して公報の散布図を作成することで可視化できる。次元圧縮は scikit-learn の主成分分析を使用した。

①SCDV:Sparse Composite Document Vectors[26] による文書のベクトル化 score＝0.756449

全ての単語に対する単語ベクトル辞書を作成する（FastText）。全ての単語ベクトルを MinBatchKMeans によってクラスタリングする。各クラスターに属する単語のベクトルを加算して合成して文章ベクトルを生成する。fastText は Facebook が開発した単語のベクトル化とテキスト分類をサポートした機械学習ライブラリである。FastText は Gensim[27]（Python 用の自然言語処理ライブラリ）から実行した。カテゴリー毎に色付けしている。

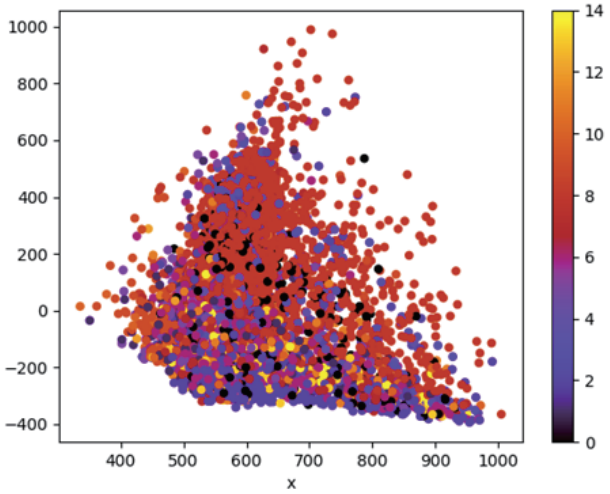


図 22　SCDV と次元圧縮による文書の散布図：

②因子解析によるによる文書のベクトル化 score＝0.838752

文書内に含まれているすべての単語ベクトルから因子成分を作成しその因子を文書の意味合いを表すベクトルデータとする。図 23 に散布図を示す。

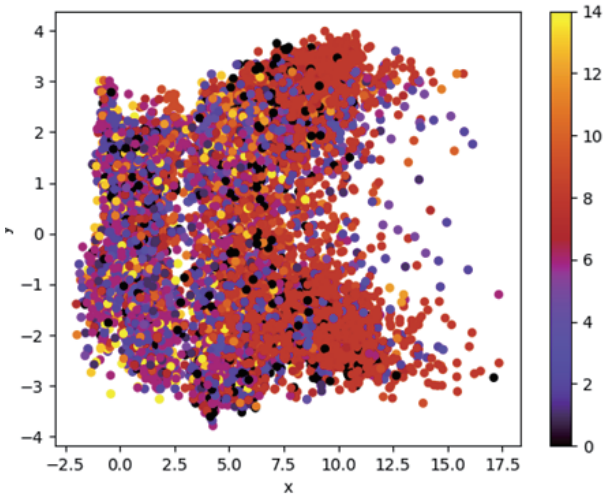


図 23　因子解析による文書のベクトル化散布

③ RNN による文書のベクトル化 score＝0.935703

Recurrent Neural Network（RNN）により直接文書データをベクトル化した。RNN は、時系列データやテキストデータを扱うことのできるニューラルネットワークの 1 つである。RNN は過去に計算した中間の状態を記憶しておけるため時系列の処理ができる。ただし計算に非常に時間がかかるため学習回数（エポック数）を 5 として実行した。計算時間は 4 コア CPU（GPU を使用しない）で 1 時間以上であった。学習回数を増やすと score は改善する。

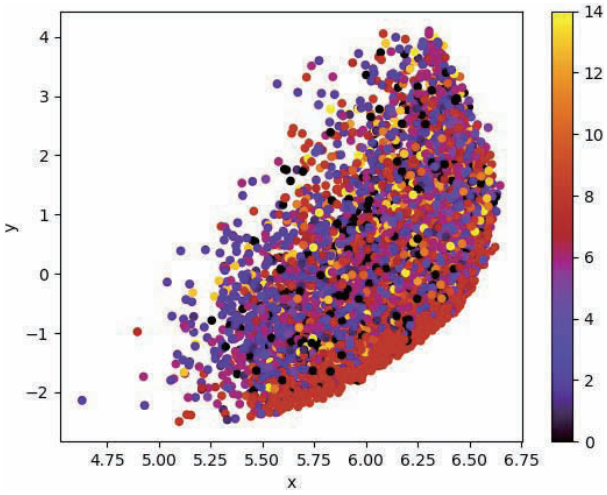


図 24　RNN による文書のベクトル化散布図

9　特許調査における教師データの利用について

5 章の表 1「多層ニューラルネットワークによる教師あり学習」では本願を含めて 50 公報を正解公報の教師データとして利用できたがこれは非常にレアケースで通常の先行技術調査や無効資料調査では正解公報の用意が困難である。あるいは逆に数件の正解公報が用意できたとすると先行技術調査や無効資料調査は結論が実質的に明らかとなりそれ以上の調査は不要と考えられる。特に特許出願はまず新規性の確保を目指して明細書が準備されるので文書単位でそれなりの数の正解公報を用意すること自体が特許出願の性格上難しいと言える。これを発明の構成要素単位あるいは文（センテンス）単位で用意することは文書単位に比べれば格段に対応の選択肢が増える。とはいうものの特許特有の長くて難解な特許文書中から発明の構成要素の記載箇所を見つけて教師データとする必要がある。doc2vec の文単位の類似度計算は正解候補の抽出に利用できる。他にも上手く使えば教師データの準備に有効と思われる手法を紹介する。

Orbit.com のダウンロードデータ活用事例を図 25 に示す。図 25 は特定の公報に付与された KEYW：コンセプトと MLID：化学物質名の ID である。

①KEYW：コンセプト（テキストマイニング手法で抽出した専門用語）括弧内の数字は（重要度、頻度 TF）

②MLID：化学物質名の ID

上手く使うと教師データとして興味深い使い方が可能となる。

Orbit.comダウンロードデータ活用事例

①KEYW：コンセプト（テキストマイニング手法で抽出した専門用語）括弧内の数字は（重要度、頻度TF）
例
INKJET(100,25)|MAGENTA INK(100,56)|INKJET RECORDING(58,21)|INKJET MAGENTA INK(54,4)|MAGENTA INKJET RECORDING(54,2)|INK PRINTING TROUBLE(45,1)|AZO PIGMENT(42,29)|PIGMENT DISPERSING RESIN(35,23)|ALKYLAMINO(34,1)|INSOLUBLE ABSORBENT SUBSTRATE(34,18)|　・・・・・

②MLID：化学物質名のID
例
(JP2018115325)|256(1,2-PROPANEDIOL|PROPYLENE GLYCOL);|1792(1,3-BUTANEDIOL);|1280(STEARYL ACRYLATE);|1793(DIETHYLENE GLYCOL MONOPROPYL ETHER);|58625(ETHYLENE GLYCOL METHYL ETHYL ETHER);|773(POTASSIUM HYDROXIDE);|1285(LAURYL METHACRYLATE);|　・・・・・

図 25　Orbit.com のダウンロードデータ活用事例

4 章で検討した（4J039GA24）/FTM AND（CN）/PN 2501 件の中国特許に付与された KEYW：コンセプトの上位 20 を表 2 に示す。

各公報のレコードに付与されている KEYW：コンセプトには頻度 TF の数値があるので表 2 のように公報単位で集計した公報件数 DF を使うと簡単に TF・IDF を計算することができる。この母集団のトータル付与 KEYW：コンセプトは 246943 種類であった。

MLID：化学物質名の ID も各公報のレコードに付与されている。ただし付与は完全ではなく付与されて

表 2　KEYW：コンセプトの上位 20 位

No.	KEYW Ranking	Google翻訳	公報件数DF
1	ink	インク	1674
2	methylpyrrolidone	メチルピロリドン	1025
3	inkjet printing	インクジェット印刷	965
4	inkjet recording	インクジェット記録	954
5	inkjet ink	インクジェットインク	888
6	hydrogen atom	水素原子	861
7	nonionic surfactant	非イオン性界面活性剤	808
8	colorant	着色剤	783
9	additive	添加剤	752
10	carboxyl group	カルボキシル基	749
11	ink viscosity	インク粘度	749
12	inkjet	インクジェット	722
13	inkjet printer	インクジェットプリンター	682
14	anionic surfactant	アニオン性界面活性剤	658
15	acetylene ethylene glycol	アセチレンエチレングリコ	629
16	plain paper	普通紙	623
17	viscosity modifier	粘度調整剤	602
18	organic solvent	有機溶剤	588
19	surfactant	界面活性剤	584
20	ball mill	ボールミル	562

表 3　MLID 化学物質名 ID Ranking 上位 20 位

No.	MLID Ranking	名称	Google翻訳	公報件数DF
1	73	ester	エステル	1878
2	144	ethylene glycol	エチレングリコール	1863
3	152	amine	アミン	1679
4	34	acrylic acid	アクリル酸	1550
5	1184	diethylene glycol	ジエチレングリコール	1547
6	16	ether	エーテル	1540
7	186	glycerol	グリセロール	1481
8	256	propylene glycol	プロピレングリコール	1470
9	177	ethanol	エタノール	1398
10	225	polyethylene glycol	ポリエチレングリコール	1395
11	198	methanol	メタノール	1303
12	434	amide	アミド	1286
13	206	cyan	シアン	1255
14	28	ammonium	アンモニウム	1215
15	74	acrylate	アクリレート	1199
16	102	vinyl	ビニール	1181
17	1294	triethylene glycol	トリエチレングリコール	1168
18	695	ketones	ケトン	1165
19	26	carboxylic acid	カルボン酸	1141
20	232	polypropylene glycol	ポリプロピレングリコール	1129

いないレコードが 387/2501＝15.5％％存在した。上位 20 位までには表れていないが n-methyl-2-pyrrolidone とか 1，6-hexanediol のような表記も抽出されている。この母集団のトータル付与 MLID は 38174 種類であった。インクジェットインク組成物の例えば有機溶媒にも付与されており化学物質名称は異表記も多いが MLID として統制されており機械学習の教師データとして興味深い。

10 今後の展望

Orbit の日本語、中国語のダウンロードデータには目に見えない不可視の分割記号（16 進数表示で &H200B）が含まれている。日本語は形態素解析器 MeCab（和布蕪）[28] で分かち書きされているとのことである。この分割記号を使うと簡単に分かち書きできて doc2vec の入力や NTT データ数理システムの Deep Learning アドオンのようなディープラーニングへの入力に使用できる。もちろん日本語、中国語を自前で分かち書きを行っても良い。

海外特許、例えば中国（CN）特許には限定的な例外[29] を除いて F タームは付与されていないが日本（JP）にファミリー特許がある場合はファミリーの JP 特許の F タームを教師データとして利用することも可能である。KEYW：コンセプトのような専門用語辞書、MLID 化学物質名 ID のような化学物質名辞書、F タームあるいは独自構築の社内分類を教師データとして word2vec,fastText による単語の固定長ベクトルの分散表現、doc2vec による文・文書のベクトル化データと教師ありの機械学習を組み合わせると SDI（Selective Dissemination of Information）タイプの特許調査の有力な手段となりえる。

11 まとめ

前半では先行技術調査を念頭に doc2vec による文のベクトル化と発明の要素単位の類似文抽出検討を行い、後半で動向調査を念頭に教師あり機械学習の 1 次元 CNN による文書分類と次元圧縮による公報の可視化検討を行った。教師あり機械学習には良質な教師データの準備が重要である。ディープラーニングの機械学習には大量の教師データが準備できるかで学習済モデルの性能が決まる。調査目的に応じたアルゴリズムとデータの選択が重要である。

12 終わりに

本報告は 2018 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。

最後に大変有用な各種ツールに関して機械学習の初心者である筆者を様々な形でサポートしていただいた NTT データ数理システムの多くの皆様に感謝申し上げます。

参考文献

1）「AI 白書 2017～人工知能がもたらす技術の革新と社会の変貌～」，KADOKAWA, 2017
2）松尾豊，「人工知能は人間を超えるか」，KADOKAWA，2015 年

3）松尾豊 編著，「人工知能とは」，人工知能学会監修，近代科学社発行，2016 年
4）「情報の科学と技術」2017 年 7 月号(67 巻 7 号)．特集＝特許情報と人工知能（AI）
http://www.infosta.or.jp/journals/201707-ja/
5）「情報の科学と技術」2018 年 7 月号（68 巻 7 号)．特集＝特許情報と人工知能（AI)-II
https://www.jstage.jst.go.jp/browse/jkg/68/7/_contents/-char/ja
6）安藤俊幸，「機械学習を用いた効率的な特許調査方法」，Japio YEAR BOOK 2017，2017，p. 230-241.
http://www.japio.or.jp/00yearbook/files/2017book/17_3_04.pdf
7）安藤俊幸，桐山勉，「機械学習を用いた効率的な特許調査方法」，第 14 回情報プロフェッショナルシンポジウム予稿集，p. 83-88.
https://www.jstage.jst.go.jp/article/infopro/2017/0/2017_83/_article/-char/ja/
8）日立　特許情報提供サービス「Shareresearch」
http://www.hitachi.co.jp/Prod/comp/app/tokkyo/sr/
9）NRI サイバーパテントデスク 2
https://www.patent.ne.jp/service/patent/
10）Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words andphrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119, 2013.
11）Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In International Conference on Machine Learning, Vol. 14, pp. 1188-1196, 2014.
12）パーセプトロン
https://ja.wikipedia.org/wiki/ パーセプトロン
13）Chainer
https://chainer.org/
14）scikit-learn
http://scikit-learn.org/stable/
15）新納浩幸，「Chainer V2 による実践深層学習」，オーム社，2017 年
16）Visual Mining Studio
https://www.msi.co.jp/vmstudio/
17）Deep Learning アドオン（Deep Learner）
https://www.msi.co.jp/vmstudio/deepLearning.html
18）Text Mining Studio
https://www.msi.co.jp/tmstudio/
19）Questel 社 Orbit.com
https://www.orbit.com/
20）scikit-learn algorithm cheat-sheet
http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
21）Visual Mining Studio チートシート
https://www.msi.co.jp/vmstudio/tips/tips04_VMS_cheat_sheet.pdf
22）Apache MXNet
https://mxnet.apache.org/
23）坂本俊之，「MXNet で作る データ分析 AI プログラミング入門」，C&R 研究所
24）Wikipedia 日英京都関連文書対訳コーパス
https://alaginrc.nict.go.jp/WikiCorpus/
25）Yoon Kim, Convolutional Neural Networks for Sentence Classification
https://arxiv.org/abs/1408.5882
26）SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations
https://dheeraj7596.github.io/SDV/
https://arxiv.org/abs/1612.06778
27）Gensim
https://radimrehurek.com/gensim/
28）MeCab（和布蕪）
http://taku910.github.io/mecab/
29）中国特許文献の FI・F ターム付与データ提供について
https://www.jpo.go.jp/shiryou/s_sonota/china_patent.htm

上記 URL はいずれも 2018 年 8 月 30 日に確認したものである。