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―

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

近年ニューラルネットワークを用いた機械学習が特に画像認識において成功をおさめディープラーニングへと発展し様々な分野で応用研究がなされている[1)]。特許情報の分野においても「情報の科学と技術」2017 年 7 月号（67 巻 7 号）で、特許情報と人工知能（AI）の特集が組まれている[2)]。日本特許庁においても人工知能（AI）技術の活用に向けたアクション・プラン[3)]が公表されており各種の実証試験が試行されている。

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。例題として特許検索競技大会 2016 の化学・医薬分野の問 2（ガスバリア性包装用フィルム）を選択し機械学習の先行技術調査への適用可能性を検討した。

2　特許調査への機械学習適用の留意点

機械学習で実際のデータを使用して学習モデル構築を始める前に押えておいた方がよいと思われる機械学習使用に当たっての基本的な留意点を述べる。

（1）ノーフリーランチ定理[1]（NFL 定理）[4)]

この定理は「あらゆる問題で性能の良い汎用最適化戦略は理論上不可能であり、ある戦略が他の戦略より性能がよいのは、現に解こうとしている特定の問題に対して特殊化（専門化）されている場合のみである」ということを立証している（Ho and Pepyne、2002 年）。

工学者や最適化の専門家にとって、この定理は、問題領域の知識を可能な限り使用して最適化すべきだということを示しており、領域を限定して特殊な最適化ルーチンを作成すべきであることを示している。

高度に最適化された特殊アルゴリズム（赤）と汎用ア

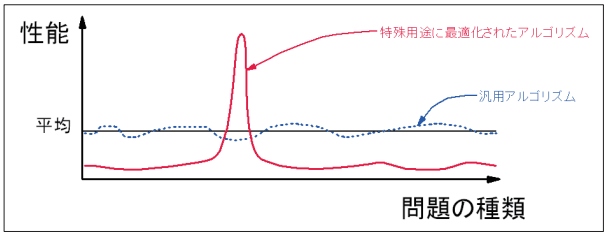


図 1　ノーフリーランチ定理の概念図

1　この定理の名称は、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の There ain't no such thing as a free lunch. に由来する。かつて酒場で「飲みに来た客には昼食を無料で振る舞う」という宣伝が行われたが、「無料の昼食」の代金は酒代に含まれていて実際には「無料の昼食」なんてものは有る訳がない、という意味。TANSTAAFL（タンスターフルと発音）というアクロニム（acronym：連なったアルファベットを通常の単語と同じように発音して読むもの）も同作品で広まった。同作品には自意識を持つ今でいうクラウドコンピュータが登場しストリー展開に重要な役割を果たす。50 年以上過去に現在でいう仮想現実、CG、音声応用、データサイエンス等が描かれており欧米人の AI 観やコンピュータ史の観点からも興味深い。

ルゴリズム（青）。どちらも平均すれば同程度の性能となることに注意。

NFL 定理に関係して AI 分野におけるアルゴリズムの重要性は「最強囲碁 AI アルファ碁 解体新書 深層学習、モンテカルロ木探索、強化学習から見たその仕組み」[5]ではアルゴリズムの特徴を生かして組み合わせておりその考え方が参考になる。

(2) 醜いアヒルの子の定理　機械学習の「朱鷺の杜 Wiki」

醜いアヒルの子の定理（ugly duckling theorem）「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」

これは，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。この定理は，特徴選択や特徴抽出が識別やパターン認識にとって本質的であることを示唆している。

醜いアヒルの子（白鳥の雛）と普通のアヒルの雛の本質的な特徴量の種類を選択するか重み付けの重要さを示している。例えば白鳥の雛は灰色、アヒルの雛は白という特徴量だけを選択すると類似性は異なり容易に識別できるが白鳥の雛の識別とは関係がない特徴量を加えれば加えるほど識別は難しくなる。

(3) 機械学習における過学習[6]

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

3 機械学習検討用データセット作成

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。機械学習適用のメインターゲットとしてスクリーニング支援用に査読の優先順位を推薦するスコアリングを想定した。筆者のこれまでの検討で調査対象文書と調査対象集合の各特許公報の各種類似度（スコア）を求めても審査官が実際に新規性で拒絶理由に採用した文献の類似度を比べると乖離が大きいことが課題であった。そこで実際の審査過程を考慮して問題が作成され「正解」公報とその先行技術調査プロセスの模範解答が示される特許検索競技大会[7]に着目した。図 2 に特許検索競技大会 2016 フィードバックセミナー資料より推奨されている先行技術調査の流れを示す。

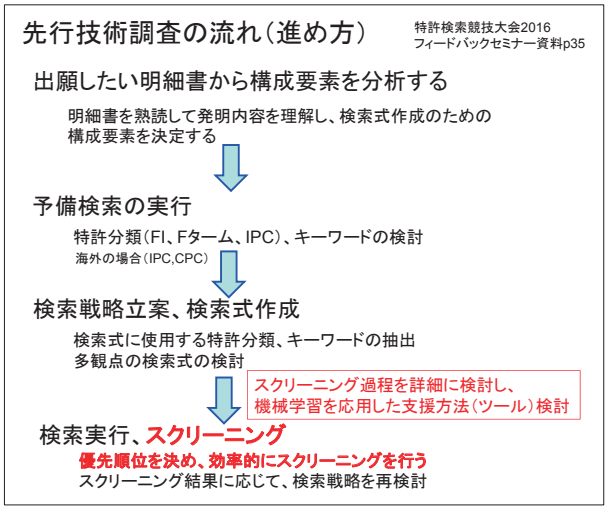


図 2　先行技術調査の流れ

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）[7]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。

類似度（スコア）検討のため最初に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」[9]、発明通信社　HYPAT-i2[10]、NRI サイバーパテントデ

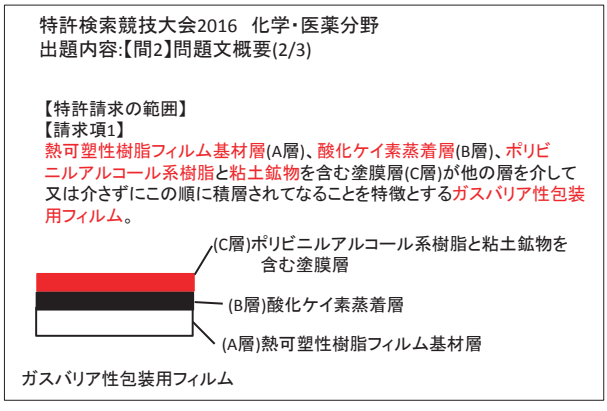


図 3　特許検索競技大会 2016 の化学・医薬分野の問 2

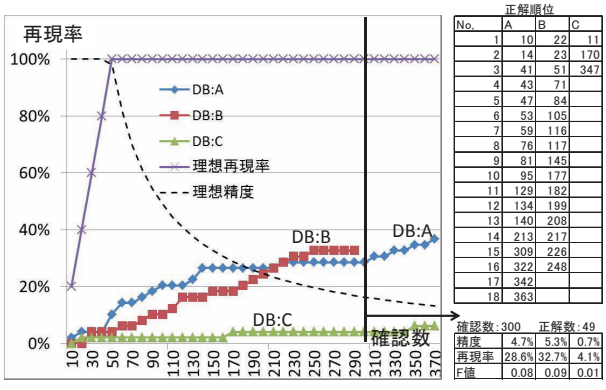
スク 2[11]、を使い検索競技大会の問題文（図 3）の請求項 1 を入力して概念（類似）検索を行い検索競技大会の正解公報 49 件の出現順位を調べた。概念検索は各システムのデフォルト条件で行った。

図 4 に各商用データベースの概念（類似）検索の再現率比較のグラフ示す。横軸は概念検索結果をスコアの高い順に確認した場合の確認数である。確認数 300 の時の精度、再現率、F 値を図の右下に示す。F 値は精度と再現率の調和平均である。正解公報が理想的に確認できた場合の理想再現率と理想精度（破線）を示す。次節以降の検討結果はグラフの見やすさの点から再現率でプロットしているが精度（調査効率）重視の観点からはグラフの立ち上がりが急峻な方が良い。再現率（網羅性）重視の観点からはなるべく早く 100%に近づく方が良い。再現率の理想曲線に対して精度を計算してプロット（破線）すると再現率が 100%未満の間は精度 100%で理想的な場合は 100%で交差して再現率が 100%に到達以降は新公報は全てノイズとなるので精度は横軸の確認数とともに単調に減少する。次節以降の検討では理想再現率と DB：A の再現率を比較のベースラインとしてプロットする。

普通、概念（類似）検索だけで調査を行うことは少なくブーリアン検索等と補完的に組み合わせて使用する。概念検索を予備検索として検索キー（特許分類、検索キーワード）の抽出用に使用する場合は精度が高いことが望ましい。

特許調査の実務上どこまで確認するか、言い換えるとどこで調査（確認）を打ち切るかは重要なポイントである。調査目的や調査の重要性（どのくらいコストを掛けられるか）によっても異なる。

性格の異なるデータベース A と C の概念検索各々上



正解順位

No.	A	B	C
1	10	22	11
2	14	23	170
3	41	51	347
4	43	71	
5	47	84	
6	53	105	
7	59	116	
8	76	117	
9	81	145	
10	95	177	
11	129	182	
12	134	199	
13	140	208	
14	213	217	
15	309	226	
16	322	248	
17	342		
18	363		

確認数：300　正解数：49

	A	B	C
精度	4.7%	5.3%	0.7%
再現率	28.6%	32.7%	4.1%
F値	0.08	0.09	0.01

図 4　商用データベースの概念（類似）検索の再現率比較

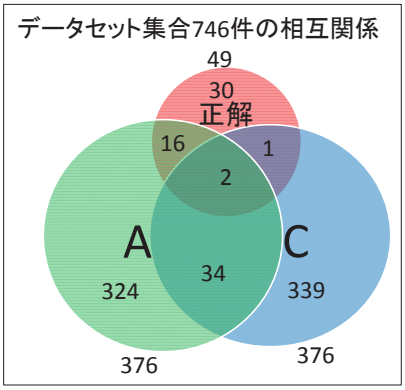


図 5　データセット集合 746 件の相互関係

位 376 件と正解 49 件の和集合 746 件を各種検討用のデータセットとした。C は上位 10000 件確認し正解 3 件であった。

4　類似検索シミュレーション検討

前節で作成したデータセットを用いて類似度計算に影響する要素（アルゴリズムや各種パラメータ等）を実験的に検討した。類似度計算に影響する要素として各文書のタイトル、要約、請求項を下記ベクトル化（特許公報に含まれる単語を基に複数の数値で表す）する手法を検討した。類似度計算方法の評価方法はデータセットにおけるクエリ文書：本願 P0 に対する各公報の類似度（スコア）を計算して降順にソートし正解公報の順位を求め横軸に公報確認数、縦軸に再現率をプロットして評価した。

文書のベクトル化手法検討（類似度計算用）

- BoW（Bag-of-words）モデル：単語の出現頻度、出願順序を考慮しない
- TF・IDF モデル：TF（Term Frequency、単語の出現頻度）と IDF（Inverse Document Frequency、逆文書頻度）の積
- 単語として形態素あるいは専門用語（複合語）を使用
- N-グラム（単語単位ではなく文字単位で分解）

形態素解析器は MeCab、専門用語は自作の PatAnalyzer を使用して抽出した[12]。類似度は自作の類似度計算プログラム SimCalc1 を使用して計算した[13]。

図 6 に分かち書きと重み付けの再現率への影響を示す。精度重視の観点から重要な確認数が少ない最初の方は DB：A が一番よかった。次は専門用語で分かち書きした TF・IDF である。詳細に見ると差は出ているが大

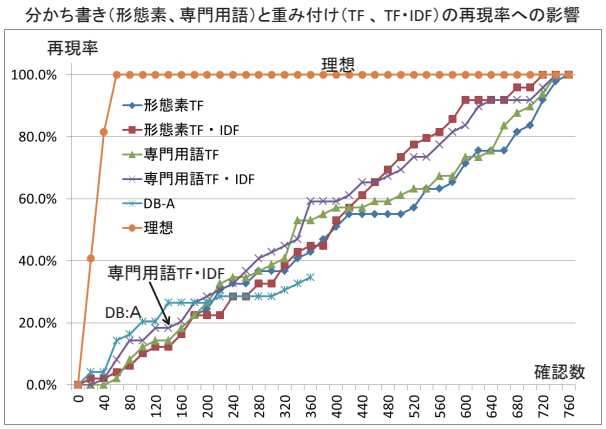


図６　分かち書きと重み付けの再現率への影響

局的に見ると差は意外に少ない結果となった。

図３の検索競技大会の問題である請求項１の分かち書きを形態素と専門用語で行った場合の結果を図７、図８に示す。類似度計算は記号を除いて行っている。専門用語は名詞の連接部分を抽出している。

熱	名詞,一般,*,*,*,*,熱,ネツ,ネツ
可塑	名詞,一般,*,*,*,*,可塑,カソ,カソ
性	名詞,接尾,一般,*,*,*,性,セイ,セイ
樹脂	名詞,一般,*,*,*,*,樹脂,ジュシ,ジュシ
フィルム	名詞,一般,*,*,*,*,フィルム,フィルム,フィルム
基	名詞,一般,*,*,*,*,基,モト,モト
材	名詞,接尾,一般,*,*,*,材,ザイ,ザイ
層	名詞,接尾,一般,*,*,*,層,ソウ,ソー
、	記号,読点,*,*,*,*,、,、,、

図７　形態素解析（MeCab）による分かち書き（一部）

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
粘土鉱物
塗膜層
他
層
積層
特徴
ガスバリア性包装用フィルム

図８　専門用語による分かち書き

図９に N-グラムの文字数 N と重み付けの影響を示す。ここでの N-グラムは文字の N グラムである。形態素あるいは単語の N-グラムではない。重みを２値（0 か 1）にした場合より TF（単語頻度）にした場合の方が良い結果を示している。文字数 N のピークは３か４にありそうだが N＝３は確認していない。

類似度計算方法として下記を検討した。

・Cosine 係数、Dice 係数、Jaccard 係数
・新規性を考慮した評価関数を定義

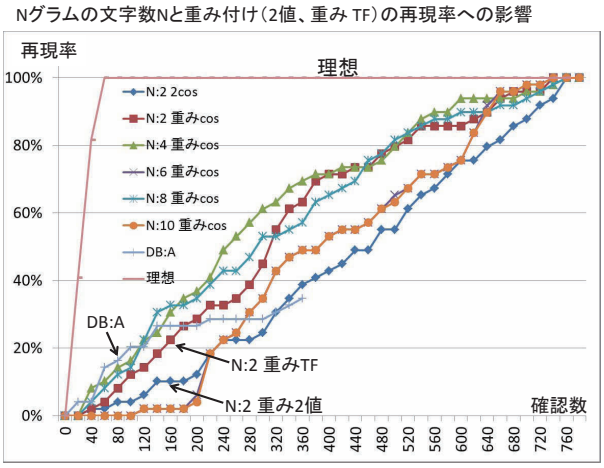


図９　N-グラムの文字数 N と重み付けの影響

Cosine 係数、Dice 係数、Jaccard 係数の再現率への影響は小さかった。

図 10 の構成要素分析模範解答例と図８の専門用語による分かち書きを比べると形態素解析と専門用語では専門用語の方がマッチングした場合は類似度が高くなると考えられる。形態素の場合はマッチングの確率が高くなると考えられる。

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)
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	包装用フイルム

※構成要素の分け方は本例に限定しない

図 10　構成要素分析（検索競技大会の模範解答例）

新規性を考慮した評価関数として図 10 の構成要素分析例を参考に図 11 の F タームと TF による類似度を加算した評価関数を設計した。図 10 の構成要素に該当する F タームがマッチングした時に重み 1 を加算し更に形態素の TF による類似度を加算した単純な合成関数である。公報確認数を横軸に評価関数を縦軸にプロットしたものが図 11 のグラフである。

図 12 の評価関数を用いた再現率への影響の実験結果は確認数の大きい後半では DB：A を上回るが前半ではあまり差は無い。評価関数を用いて更に正解公報を上位に持ってくるには構成要素によって重みを変える、F タームと形態素 TF 類似度の寄与率を変える等々いろいろ考えられる。実験的あるいは重回帰分析のような手

Fターム利用評価関数　　各要素のFタームの重み：1+TFによる類似度

要素	b1 酸化ケイ素	b2 蒸着	c PVA	d 粘土鉱物	f ガスバリア	g 包装用フィルム
FI	B32B9/00@A		B32B27/30,102			
Fターム	4F100AA20	4F100EH66	4F100AK21 4F100AK69	4F100AC03 4F100AD01	4F100JD02	4F100GB15

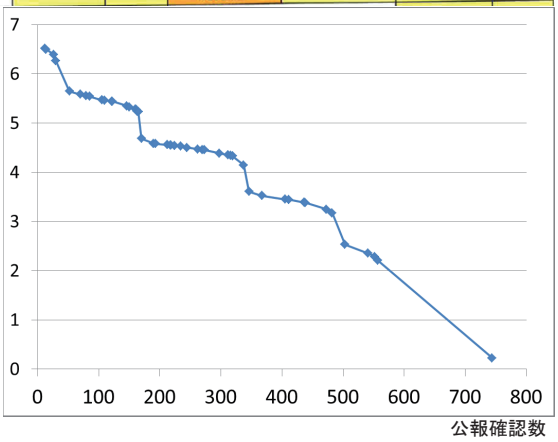


図 11　F タームと形態素 TF 類似度による評価関数

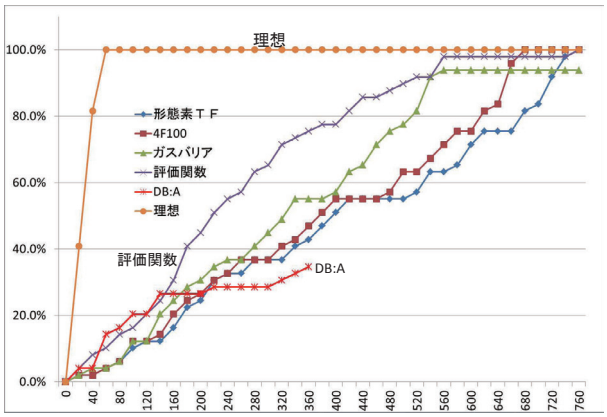


図 12　評価関数とフィルターの影響

法で重み付けを調整することでチューニングの余地はあると考える。評価関数をどこまでチューニングできるか興味深い。形態素 TF がベースラインで 4F100 は F テーマコードでフィルターしたものであり、ガスバリアのラインは要素 f のガスバリアに該当する F ターム 4F100JD02 でフィルターしたものである。フィルターとはメールのスパムフィルターのように該当 F タームが付与されていない公報を除いている。フィルターでは公報に構成要素の F タームが付与されていないと除かれて検索漏れが発生する。実際にガスバリアの再現率曲線は 100%に達しておらず検索漏れが発生している。

本稿では先行技術調査のスクリーニングを主な対象としているが評価関数はクリアランス（侵害防止）調査向けに設計することも可能である。ただしクリアランス調査向けの評価関数としては生死情報やリーガルステータスを利用したフィルターあるいはスクリーニングの優先順位を下げるマイナスのスコアリング等の考え方をクリアランス調査の要求特性に合わせて設計する必要がある。

5　分散表現学習モデル検討

単語の分散表現：Distributed Representation あるいは単語埋め込み：word embedding と呼ばれる手法を用いて単語を比較的低次元（50～500）の実数ベクトル化して利用する研究は様々な分野で行われている[14-17]。分散表現に関しては岡崎による一般向け解説[14]と、専門的な解説[15-16]がある。スタンフォード大学の Pennington らが GloVe（Global Vectors for Word Representation）[21]と呼ばれるモデルを提案し公開されている。Facebook の発表した fastText[22]は Google の word2vec[20]の延長線上にあるもので、より精度が高い表現を、高速に学習できるとされている。word2vec による類義語抽出は 7 節で後述する。

doc2vec[19]は、word2vec[18]の拡張であり、（単語ではなく）任意の長さの文書を数百次元の固定長ベクトルとして表現する手法である。doc2vec と呼ばれているが内部的には 2 つの学習方法が実装されている。word2vec と同様に CBOW モデルを拡張した PV-DM（Paragraph Vector with Distributed Memory）モデルと Skip-gram モデルを拡張した PV-DBOW（Paragraph Vector with Distributed Bag of Words）モデルの 2 種類のニューラルネットワーク構造が組み込まれている。PV-DBOW は単語の順序を考慮しないシンプルなモデルで計算効率が良く、PV-DM は単語の出現頻度と出現順序を考慮したモデルで PV-DBOW と比べると少し複雑でより多くのパラメータが必要になる。doc2vec の実行には gensim[23]（Python 用のトピックモデルライブラリ）を使用した。形態素解析器はインストールと Python からの利用が容易な Janome[24]を使用した。Janome は形態素解析用の辞書として MeCab[25]と同じ IPA 辞書を使っている。形態素解析速度は MeCab の方が一桁高速である。

図 13 に doc2vec よるベクトル化処理の概要を示す。

3 節で作成したデータセットで再現率曲線を求めて学習モデルと学習パラメータの評価を行った。Core i5、メモリ 6GB、Windows7（32bit）の普通のノート PC でこのデータセットでは形態素解析処理：170 秒、学習時間 2 秒程度、本願 P0 に対する 746 件の類似

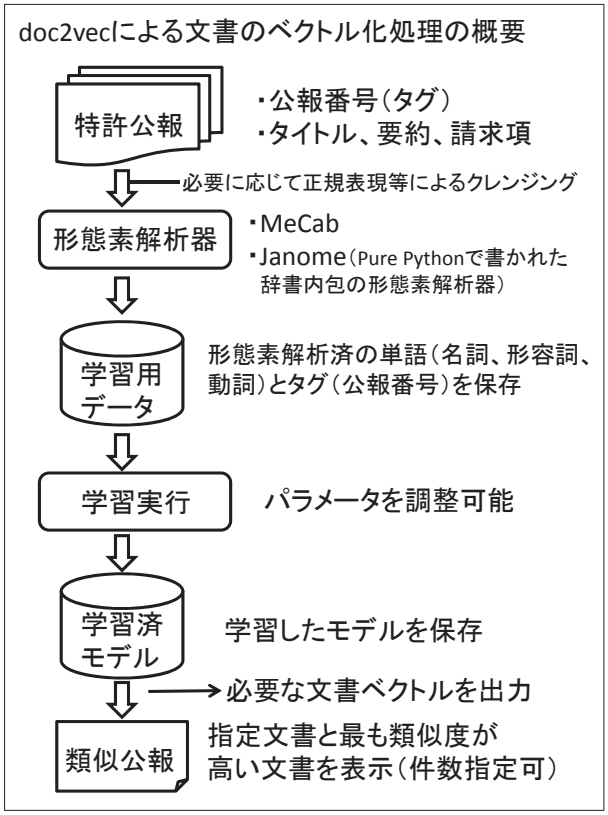


図 13　doc2vec による文書のベクトル化処理の概要

度計算 1 秒以下であった。

図 14 に文書の分散表現ベクトルの学習モデルと再現率を示す。単語の出現頻度と出現順序を考慮したモデル PV-DM はリファレンスとしてきた DB：A の再現率曲線を圧倒している。もちろん DB：A は DB 全体、本検討では非常にスモールサイズのデータセットであり直接比較の対象ではない。本検討はデータベースの検索は適切に行った後のスクリーニング過程を念頭においている。PV-DBOW では同じデータで 3 回学習を行いそれぞれ再現率曲線を求めた。再現率 1～再現率 3 である。学習のつど結果は異なっている。doc2vec の学習パラ

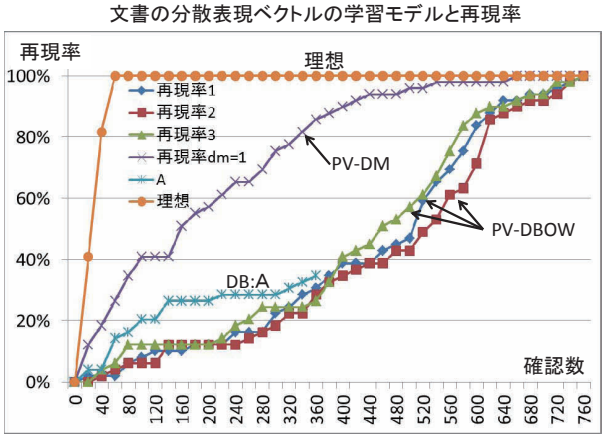


図 14　文書の分散表現ベクトルの学習モデルと再現率

メータ dm＝0 が PV-DBOW であり、dm＝1 を指定すると学習モデルが PV-DM になる。他の学習パラメータはデフォルトで行った。

図 15 に PV-DM モデルの分散表現ベクトルの次元数（Size）の影響を示す。精度重視の立ち上がりでは 200 次元が良い結果を示す。他の学習パラメータの検討も行う予定である。

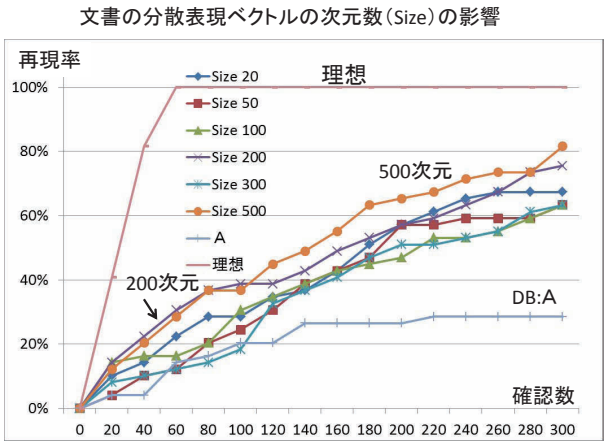


図 15　分散表現ベクトルの次元数（Size）の影響

6　可視化検討

図 16 は R 言語の非計量多次元尺度法の関数 isoMDS を用いて各公報間の距離（距離＝1－類似度）からデータセットの各公報の 3 次元座標を求め rgl パッケージを使用して OpenGL（3D グラフィックスライブラリ）で 3D 可視化した。各公報間の相互距離は TF・IDF 法による単語の重み付ベクトルを用いたコサイン類似度から求めた。類似の公報を近くに配置するように類似度か

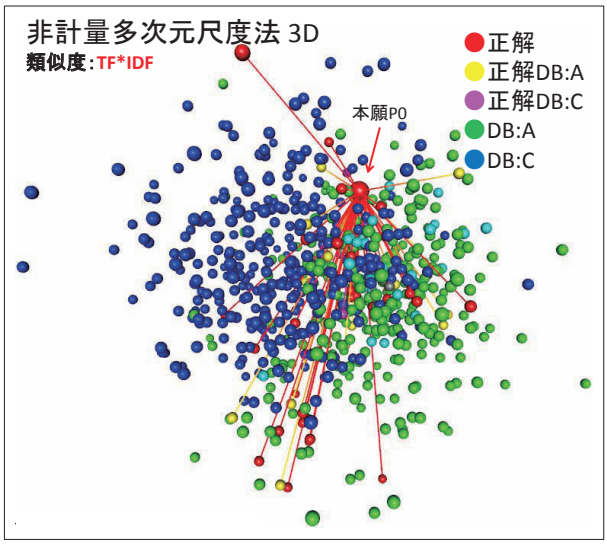


図 16　非計量多次元尺度法による各公報の可視化

ら各公報間の距離を求めて isoMDS で実際の 2 次元あるいは 3 次元の座標を求めている。マップの精度を上げるには精度のよい類似度算出法が求められる。各公報を示す球のカラーマッピングは図 4 データセット集合の相互関係の色に対応させている。緑球はデータベース A 由来の非正解公報、黄色球は DB：A 由来の正解公報、同様に青球は C 由来の非正解、ピンク（マゼンタ）球は C 由来の正解、赤球はデータベース A と C の両方の類似検索集合に入らなかった正解公報である。本願 P0 と矢印で示した大きい赤球が検索対象の本願である。本願 P0 と正解公報をラインで結んでいる。R 言語の 3D 表示はマウスでインタラクティブに回転させることで視点を変えて見ることが可能である。データベース A、C の類似検索結果の性質の違いを表現できている。球のカラーマッピングは自分で出願人や特定の特許分類、特定の専門用語に対して必要に応じて割り当てが可能である。

筆者は公報の位置関係を類似度から求め、引用／被引用関係をネットワーク表示させる手法 [12] を検討したことがある。そのときは近くに配置されるべき引用公報が遠くに配置され類似度計算方法の正確性が課題であった。

図 17 に doc2vec の分散表現ベクトルによる類似度を用いて公報間の位置関係を可視化したマップを示す。doc2vec の学習パラメータ dm＝1（PV-DM）、Size＝200（次元）で行った。処理時間は前処理の形態素解析 170 秒、学習 2 秒、747 件の相互類似度計算 14 秒であった。特徴的なのは本願 P0 が他の公報群から離れた位置にマッピングされていることである。これは本願の入力データが請求項 1 のみであり他の公報と比べると非常に短く、他の公報との共通する情報量が少ないことが原因と考えられる。その意味では直感的なイメージと合っているマップと言える。

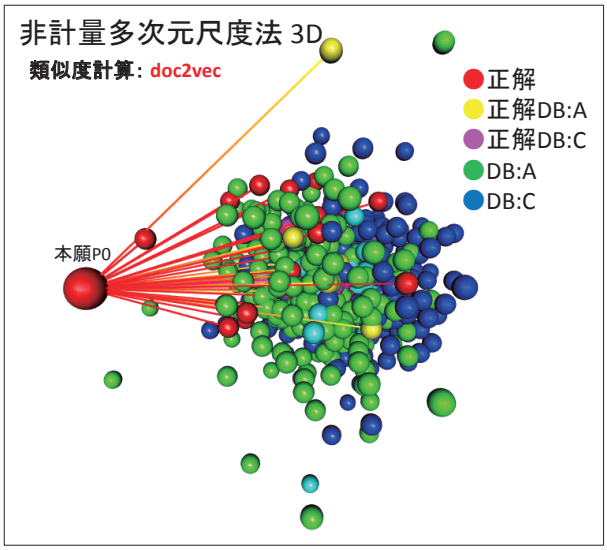


図 17　doc2vec の類似度による各公報の可視化

7　word2vec による類似語抽出

word2vec を使用すると単語の分散表現ベクトルが得られる。歴史的には word2vec の方が doc2vec より先に開発されている。単語の分散表現ベクトルから指定した単語の類似語が簡単に得られる。word2vec は gensim を用いて Python より実行した。3 節のデータセットに図 13 の doc2vwc による文書のベクトル化処理の概要に準じて Janome により形態素解析して名詞、形容詞、動詞を抽出し分かち書き後 word2vec による学習後「粘土」の類似語、類似度の上位 20 位を求め関連しそうなワードを別途確認した。確認として MeCab による形態素解析、PatAnalyzer による専門用語抽出を行い形態素、専門用語の頻度ランキングの出現順位と頻度を求めた結果を図 18 に示す。図 20 の Wikipedia の主な粘土鉱物と比較すると教師なし機械学習でデータセットを学習しただけで「粘土」に対する具体的な粘土鉱物を「類似」と学習している。word2vec による「粘土」の類似語の順位 1 位のスメクタイトを詳細に見ると形態素としてはデータセット中で頻度 26 で順位は 555 である。図 18 と図 19 の専門用語の各スメクタイトを合計すると 26 になる。Wikipedia の粘土鉱物の項に無い順位 4 のサポナイトとは粘土鉱物の一種モンモリロナイト族の中で Mg に富む種である。ただカオリナイトは形態素解析でカオリ（固有名詞）とナイト（名詞）に分解されているが粘土と近い類似度になっている。同じことがセリサイトでもいえる。word2vec により「粘土」を含まない文からサポナイト、ヘクトライト、スチーブンサイト、カオリナイト、マイカ、タルク、セリサイト（の一部）を粘土の類似語として学習できたことになる。ただ形態素解析の分かち書き精度が word2vec のニューラルネットワークのニューロンの「単語埋め込み」精度に影響している。例えば図 18 の「粘土」の類似語の黄色のセルは形態素解析で正しく形態素として解析できなかったものである。図 18 から

word2vec「粘土」の類似語　　形態素　　専門用語抽出

順位	類似語	類似度	順位	頻度	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

図 18　Word2vec による「粘土」の類似語抽出

専門用語抽出（続き）

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

図 19　専門用語抽出（続き）

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）

https://ja.wikipedia.org/wiki/粘土鉱物

図 20　主な粘土鉱物

は化学の常識として除いているが 6 位に「硫酸」が現れる。これは粘土鉱物と列記された中に「硫酸バリウム」がありそれが形態素解析で「硫酸」と「バリウム」に分かち書きされたためと考えられる。形態素解析としては正しいが化学の観点からは「硫酸バリウム」が望ましい場合が多いと考えられる。また形態素解析で「ガス」と「ガスバリア」になる場合があり word2vec の類似語は「ガス」と「ガスバリア」でそれぞれ異なり別物である。

従来から単語の共起に基づいて共起ネットワークを描いたり、形態素の名詞の隣接情報により専門用語抽出は可能である [12]。これらと word2vec を組み合わせるとクエリ拡張の支援ツールとして更に有効だと考えられる。また word2vec では有名な例で「王－男＋女＝女王」、「パリ－フランス＋日本＝東京」のような演算ができ単語の意味関係に基づいて関係を理解する一助にできると報告されている。単語間の類似度（非距離）から単語の相互関係を多次元尺度法等で可視化することも可能である。

8　ニューラルネットワークの応用事例

ニューラルネットワーク [26] は、教師信号（正解）の入力によって問題に最適化されていく教師あり学習と、教師信号を必要としない教師なし学習に分けられる。明確な解答が用意される場合には教師あり学習が、データ・クラスタリングには教師なし学習が用いられる。中間層が 2 層以上ある深層学習（ディープラーニング）[27] においては、出力に近い最後の中間層を教師あり学習で、それよりも入力に近い側を教師なし学習で学習する方法がジェフリー・ヒントンらにより提案されている。結果としていずれも次元削減されるため、画像や統計など多次元量のデータで線形分離不困難な問題に対して、比較的小さい計算量で良好な解を得られることが多い。 現在では、特徴量に基づく画像認識、市場における顧客データに基づく購入物の類推など、パターン認識、データマイニング等に応用されている。

自己組織化マップ [28-29]

自己組織化マップ（self-organizing map：SOM）は教師なし学習を行うニューラルネットワークの 1 種である。SOM は n 次元のベクトルを平面（二次元配列）に写像する方法である。

自己組織化マップは NTT データ数理システムの Visual Mining Studio（VMS）に実装されている。

図 21　Visual Mining Studio（VMS）の自己組織化マップ

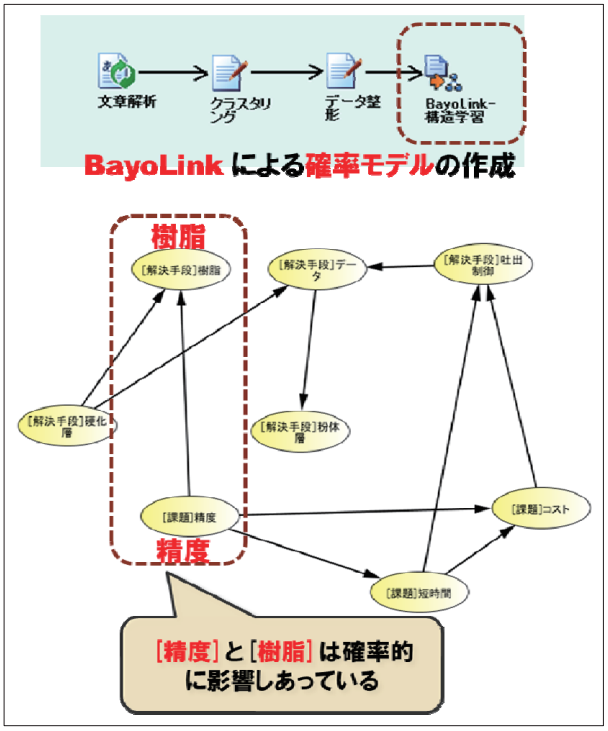


図 22　BayoLink によるベイジアンネットワーク

筆者は特許公報の発明のカテゴリーを自己組織化マップを使用して文書分類できないか試したことがありカテゴリーが同じ公報が SOM 上の近くのセルに集まる傾向が確認できた。[30]

ベイジアンネットワーク [31]

ベイジアンネットワーク（Bayesian network）は、因果関係を確率により記述するグラフィカルモデルの 1 つで、複雑な因果関係の推論を有向非巡回グラフ構造により表すとともに、個々の変数の関係を条件つき確率で表す確率推論のモデルである。ジューディア・パールが 1985 年に命名した。ジューディア・パールはこの研究の功績によりチューリング賞を受賞した。人工知能の分野では、ベイジアンネットワークを確率推論アルゴリズムとして 1980 年頃から研究が進められ、既に長い研究と実用化の歴史がある。

ここでいうネットワークとはグラフ理論 [32] の重み付けグラフのことである。グラフ理論（graph theory）は、ノード（節点・頂点）の集合とエッジ（枝・辺）の集合で構成されるグラフに関する数学の理論である。ベイジアンネットワークは NTT データ数理システムの BayoLink に実装されている。BayoLink ではベイジアンネットワークの確率モデルを作成することができそれぞれの技術がどのように影響しあうのかを可視化し、様々な仮定の下での推論を行うことができる。

9　まとめ

本稿では先行技術調査を念頭に特許検索競技大会 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データセットを作成して前半ではスクリーニング過程の再現率曲線に影響を与える要因を実験的に検討した。踏み込んだ検討はしていないが特徴量選択は本質的に重要である。後半は教師なし機械学習を用いて単語の分散表現で文書の固定長ベクトルが得られる doc2vec の学習モデルを使用して公報の類似度を計算する手法を検討した。その結果単語の出現頻度と出現順序を考慮したモデル PV-DM を使用すると非常によい類似度計算ができることがわかった。公報の類似度計算精度が向上すると特許調査において効率的なスクリーニングが可能となる。

本稿で検討した分散表現ベクトルを更に教師データ有りの機械学習の入力データとすることも可能である。教師データ有りの機械学習と組み合わせることで更なる精度向上が期待できる。また doc2vec の出力ベクトルを使用して各特許公報間の関係の可視化もできるので精度

の高い動向調査に応用可能である。gensim のような機械学習のフリーライブラリを用いると単語の分散表現学習は非常に簡単であるが特許調査の精度を上げるには前処理の形態素解析が重要になる。知財分野では新語の発生頻度も高く形態素解析用辞書の更新や専門用語辞書の活用も重要である。

10 終わりに

筆者は 2008 年頃より断続的にテキストマイニングによる効率的な特許調査手法を研究してきた [12-13]。本稿の前半部分はその結果のまとめに相当する。後半の doc2vec の出力ベクトルの検討はようやく始めたばかりだが素性の良さを実感している。今後の検討が楽しみである。

本報告は 2017 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）「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/
3）特許庁における人工知能（AI）技術の活用に向けたアクション・プランの公表について
https://www.jpo.go.jp/torikumi/t_torikumi/ai_action_plan.htm
4）https://ja.wikipedia.org/wiki/ノーフリーランチ定理
5）大槻知史．最強囲碁 AI アルファ碁 解体新書 深層学習、モンテカルロ木探索、強化学習から見たその仕組み，翔泳社，2017
6）過剰適合（過学習）
https://ja.wikipedia.org/wiki/過剰適合
7）特許検索競技大会
https://www.ipcc.or.jp/contest/
8）特許検索競技大会 2016 フィードバックセミナー，アドバンスコース
9）日立　特許情報提供サービス「Shareresearch」
http://www.hitachi.co.jp/Prod/comp/app/tokkyo/sr/
10）発明通信社　HYPAT-i2
https://www.hatsumei.co.jp/hypat_i2/index.html
11）NRI サイバーパテントデスク 2
https://www.patent.ne.jp/service/patent/
12）安藤俊幸．テキストマイニングを用いた効率的な特許調査方法
http://www.japio.or.jp/00yearbook/files/2015book/15_2_12.pdf
13）安藤俊幸．テキストマイニングと統計解析言語 R による特許情報の可視化　情報管理 Vol. 52 (2009) P 20-31
https://www.jstage.jst.go.jp/article/johokanri/52/1/52_1_20/_article/-char/ja/
14）岩波データサイエンス vol.2[特集] 統計的自然言語処理―ことばを扱う機械
岡崎直観，単語の意味をコンピュータに教える、

https://sites.google.com/site/iwanamidatascience/vol2/word-embedding
15）岡崎直観．言語処理における分散表現学習のフロンティア人工知能 Vol.31No.2p189-201（2016）
16）岡崎直観．単語の分散表現と構成性の計算モデルの発展
https://www.slideshare.net/naoakiokazaki/20150530-jsai2015
17）中村雄太ら．分散表現空間解析モデルに基づく研究トレンドに関する考察
http://db-event.jpn.org/deim2017/papers/305.pdf
18）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.
19）Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In International Conference on Machine Learning, Vol. 14, pp. 1188-1196, 2014.
20）word2vec
https://code.google.com/archive/p/word2vec/
21）GloVe: Global Vectors for Word Representation
https://nlp.stanford.edu/projects/glove/
22）fastText Japanese Tutorial
https://github.com/icoxfog417/fastTextJapaneseTutorial
23）gensim
https://radimrehurek.com/gensim/
24）Janome
http://mocobeta.github.io/janome/
25）MeCab
http://taku910.github.io/mecab/
26）ニューラルネットワーク
https://ja.wikipedia.org/wiki/ニューラルネットワーク
27）ディープラーニング
https://ja.wikipedia.org/wiki/ディープラーニング
28）Teuvo Kohonen, “自己組織化マップ”, 改訂版, シュプリンガーフェアラーク東京（2005）
29）https://ja.wikipedia.org/wiki/自己組織化写像
30）安藤俊幸ら，アジア特許情報のテキストマイニングによる解析―自動テキスト分類への挑戦―
https://www.jstage.jst.go.jp/article/infopro/2011/0/2011_0_13/_article/-char/ja/
31）ベイジアンネットワーク
https://ja.wikipedia.org/wiki/ベイジアンネットワーク
32）グラフ理論
https://ja.wikipedia.org/wiki/グラフ理論

上記 URL はいずれも 2017 年 8 月 25 日に確認したものである。