機械学習を用いた効率的な特許調査方法
文書のベクトル化方法と文書分類への応用

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
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

最近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。最近の Japio YEAR BOOK でも AI 関連の情報は非常に多くなっている。昨年の時点で人工知能を搭載した商用の特許情報調査・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。企業においては事前情報収集中、導入に向けて検証中、既に導入済で調整中、スケールアップ中、実務で活用中といろいろなステージにあると思われる。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。

筆者が活動しているアジア特許情報研究会 [2)] は昨年設立 10 周年を迎え記念講演会を開催した。研究会では研究成果を外部に発表すること重視しており過去の活動の資料も研究会のホームページで公開している。研究会は中国、韓国等の東アジアチーム、アセアン諸国を中心にした新興国チーム、地域の枠を越えた観点からの知財情報解析チームで活動している。知財情報解析チームのメンバーはテキストマイニング、機械学習、AI の動向等に興味を持って各自の研究テーマを設定しつつメンバー間で積極的に情報交換を行いながら研究を進めている [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。研究会の中には自作 AI を業務に活用しているメンバーも少数ながら複数存在している。

本稿では事前情報収集、検証実験、実務で活用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 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 1 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 2 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。

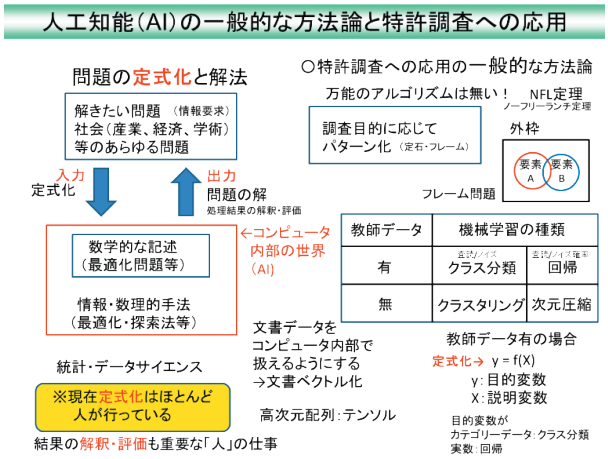


図１　人工知能の一般的な方法論と特許調査への応用

情報要求の４レベル
R. S. Taylor の 1968 年の論文（Question-negotiation and information seeking in libraries）において、人間の情報要求（information need）が４つのレベルに分類されている。

1．直感的要求（visceral need）
現状に満足していないことは認識しているが、それを具体的に言語化してうまく説明できない状態

2．意識された要求（conscious need）
頭の中では問題を整理できるが、あいまいな表現やまとまりのない表現でしか言語化できない状態

3．形式化された要求（formalized need）
問題を具体的な言語表現で言語化できる状態

4．調整済みの要求（compromised need）
問題を解決するために必要な情報の情報源が同定できるくらい問題が具体化された状態

図２　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。

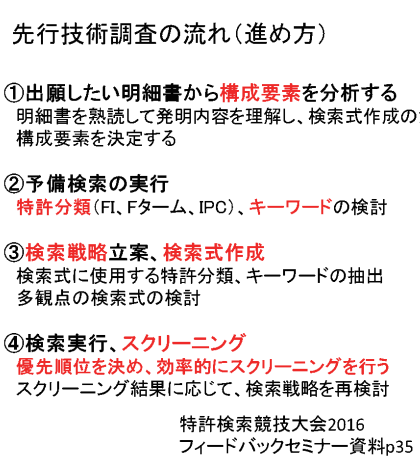


図３　先行技術調査の流れ

理想的には図３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。

3 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４に示す。

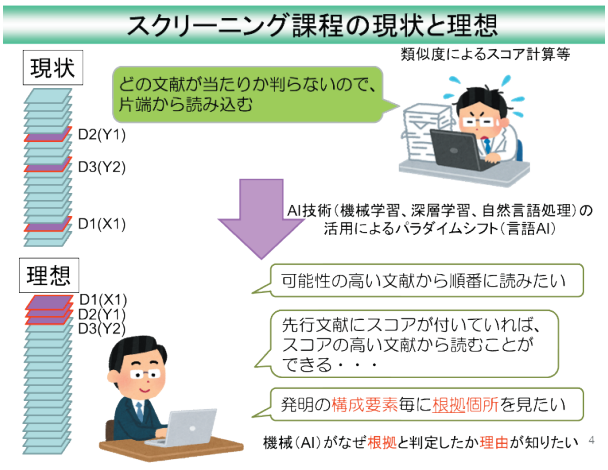


図４　スクリーニング課程の現状と理想

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)]。

図５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。

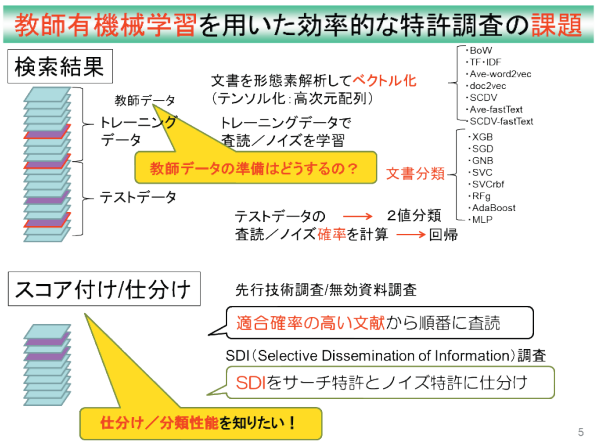


図５　教師有機械学習を用いた特許調査の課題

4 文書のベクトル化と文書分類方法

文書のベクトル化と文書分類の概要を図６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（図１）、７種類の文書のベクトル化方法を検討した。図６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[6)]のCountVectorizerを使用した。② TF・IDF

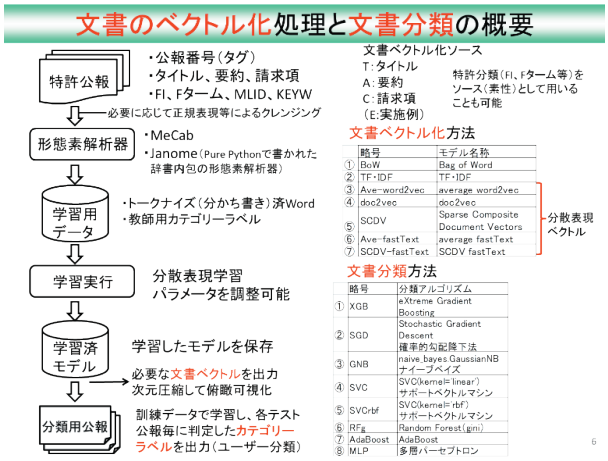


図６　文書のベクトル化処理と文書分類の概要

モデル作成には scikit-learn の TfidfVectorizer を使用した。図６の③～⑦の分散表現ベクトル作成には gensim[7] を使用した。SCDV は Sparse Composite Document Vectors using soft clustering over distributional representations[8] の略である。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。

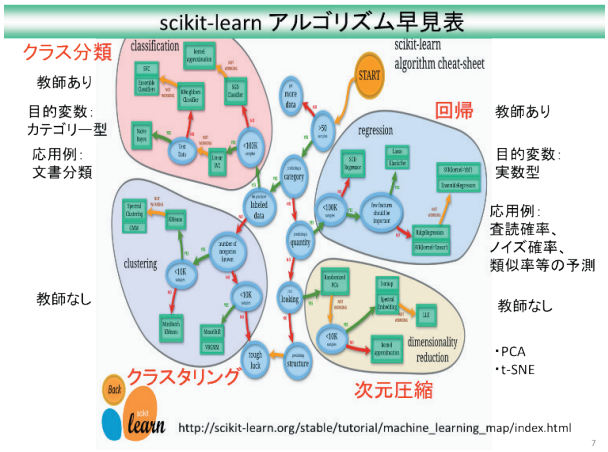


図７　scikit-learn のアルゴリズム早見表

図７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７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」[9] 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。

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[10]。

5　文書のベクトル化と文書分類（多値分類）検討

インクジェットインク関係の特許 3098 件を 12 カテゴリーにクラス分類（多値分類）する検討を行った[11]。

図８に文書ベクトルと文書分類（一部抜粋）の組み合わせによる正解率比較のグラフを示す。横軸は正解率、縦軸は各文書ベクトルである。上のグラフでは、文書分類方法が① XGB で、文書ベクトル② TF・IDF、⑤ SCDV、⑦ SCDV-fastText が良い。中央のグラフでは、文書分類方法が③ GNB（ナイーブベイズ）で、文書ベクトル① BoW、② TF・IDF が良い。ただし各グラフで横軸（正解率）が統一されていないことに注目して、

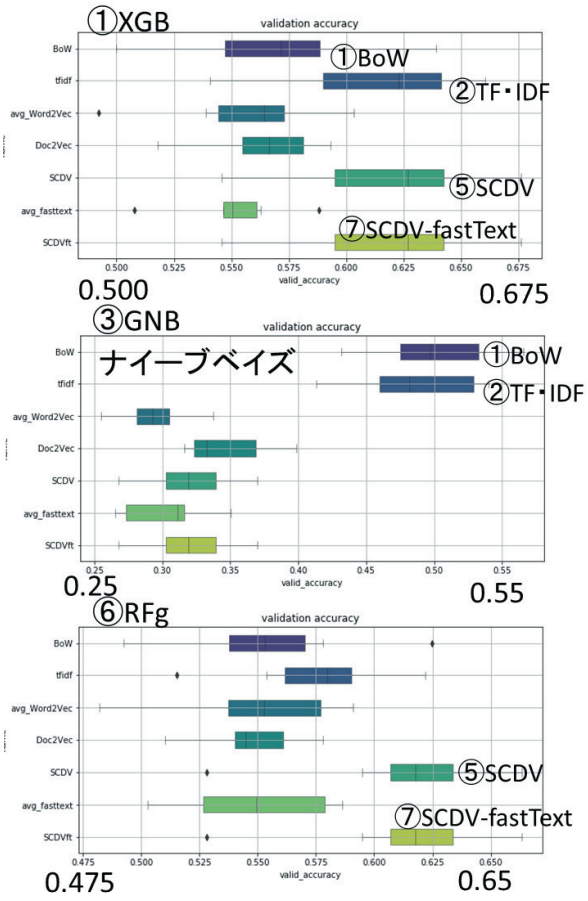


図８　文書ベクトル×文書分類（一部抜粋）の正解率比較

絶対値で文書分類方法を比べると、ナイーブベイズは従来のテキストデータのベクトルでは有効であるが最近の分散表現ベクトルには不向きであることが解る。

下のグラフは、文書分類方法は⑥RFgで、⑤SCDV、⑦SCDV-fastTextベクトルが良い。8分割交差検証を行い正解率の範囲をプロットしている。交差検証については後述する。

図9に文書ベクトルと文書分類の組み合わせによる正解率比較をまとめて示す。横軸は文書分類方法、縦軸は正解率である。折れ線の種類は文書ベクトル方法である。図9上部のグラフは訓練時、下部はテスト時の結果である。テスト時の結果が重要である。訓練とテストを入試に例えると訓練は問題集や模擬試験の成績である。テストは入試本番の点数に相当する。訓練時に使用した公報は、原則テスト時には使用禁止である。SVC（サポートベクトルマシン）とRFg（ランダムフォレスト）は訓練時には100パーセントの正解率を示している。人の学習に例えると正解を丸暗記した状態である。テスト時に訓練時に使用した公報を入力すると訓練時同様に高い正解率を示すが重要なのは未知のテストデータに対する分類（識別）能力である（汎化能力）。このデータセットのテストの分類性能は文書のベクトル化⑤SCDV×文書分類①XGBの組み合わせが良かった。

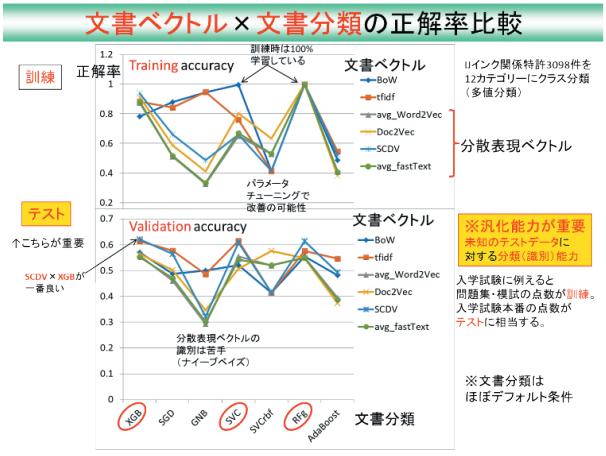


図9　文書ベクトルと文書分類の組み合わせ比較

処理時間（計算コスト）

ベクトル化手法	次元	ベクトル化	t-SNE	XGB	SGD	GNB	SVC	SVCrbf	RFg	AdaBoost	MLP	合計（秒）	（時間）
BoW	16632	0.41	226.80	4090.27	28.72	53.71	2090.64	2262.17	47.84	610.81	メモリエラー	9411.4	2.6
tfidf	16632	0.41	226.58	4190.11	27.54	51.37	1975.358[	1917.530[	44.92	607.59	メモリエラー	5148.5	1.4
Word2Vec average	300	12.39	23.83	227.99	0.59	0.86	31.67	36.42	21.04	60.75	メモリエラー	415.5	0.1
Doc2Vec	300	21.99	33.29	233.18	0.60	0.81	42.60	36.70	20.86	61.43	メモリエラー	451.5	0.1
SCDV	18000	155.56	246.55	5838.03	31.86	54.32	1954.29	2060.65	64.83	1115.41	メモリエラー	11521.5	3.2
fasttext average	300	34.11	25.08	227.23	0.63	0.83	31.909[	36.12	21.16	60.89	メモリエラー	406.0	0.1
SCDV-fasttext	18000	255.05	246.63	5817.18	31.91	54.59	1945.49	2064.68	65.13	1113.92	メモリエラー	11594.6	3.2
処理時間合計（秒）		479.92	1028.76	20623.99	121.84	216.48	8064.68	8496.74	285.77	3630.79		38949.0	10.8
（時間）		0.13	0.29	5.73	0.03	0.06	1.68	1.80	0.08	1.01		10.8	

使用CPU Intel Core i7-8700K 3.70GHz コア数：6　論理プロセッサ数:12
1論理プロセッサを使用して計測
タスクマネージャーによる負荷計測：CPU 4.35GHz　約11%

ベクトル化共通パラメータ（一部）
features_num = 300
min_word_count = 1
context = 5
downsampling = 1e-3
epoch_num = 10

分類精度重視の観点からはXGBがお奨め、
ただし処理時間はかかる

分類精度、使いやすさ、処理時間、
マルチプロセッサ対応の観点からは
Random Forestがお奨め
ただしTF・IDFは精度が悪い

図10　文書ベクトルと文書分類の組み合わせの処理時間

文書ベクトルと文書分類の組み合わせの処理時間を図10に示す。分類精度重視の観点からはXGBが良かった。ただし処理時間がかかる。目的と使用データにより文書ベクトル方法と文書分類の組み合わせを選択するのが良い。

6 文書のベクトル化と文書分類（2値分類）－事例1

分かりやすい事例としてインフレータの査読（サーチ）／ノイズの2値分類例を示す。インフレータとはエアバッグの一部の構成部品で衝撃でガスを発生させる。インフレータについて主に記載された公報を査読、それ以外をノイズと人が判定して教師データとした。母集団の全文書数は635件、査読52件、ノイズ613件である。7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11、図12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。濃い紺色が査読で黄色がノイズである。③Avg. word2vecと⑥Avg. fastTextによる文書ベクトルが紺色の査読文献のまとまりが良い。

t-SNEによる3次元への次元圧縮例を図13に示す。単語による文書ベクトル②TF・IDFを1314次元から3次元に次元圧縮したデフォルト状態（図13左）では青のサーチが裏側に隠れている。回転させて表に出すと右側のようになる。人間は4次元以上の高次元の関係をイメージとして思い描けないが図13の3次元の各公報の相互関係を2次元にマッピングした一例が図11の②TF・IDFの散布図である。3次元画像を視点を変えて見ると同じ3次元の散布図データでも大幅に異なって見える。3次元の地球儀と2次元の世界地図の関係に例えることができる。高次元データを扱うときも同様の観点から注意が必要である。

文書分類方法①XGBによる7種類の文書ベクトルを用いた査読／ノイズの8分割交差検証結果を図14上

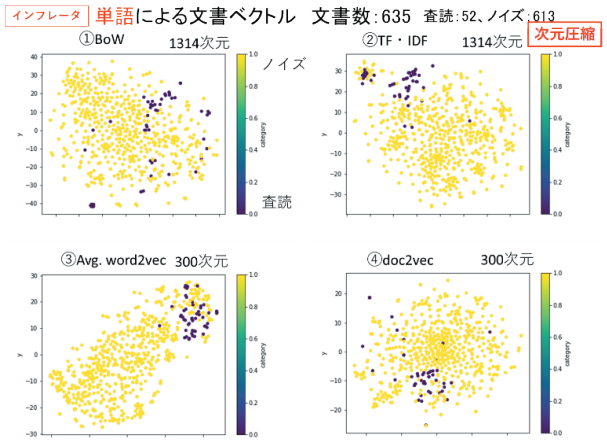


図 11　単語による文書ベクトルの次元圧縮・可視化

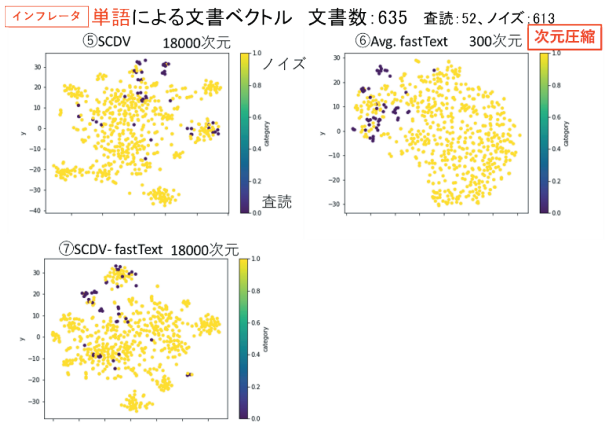


図 12　単語による文書ベクトルの次元圧縮・可視化

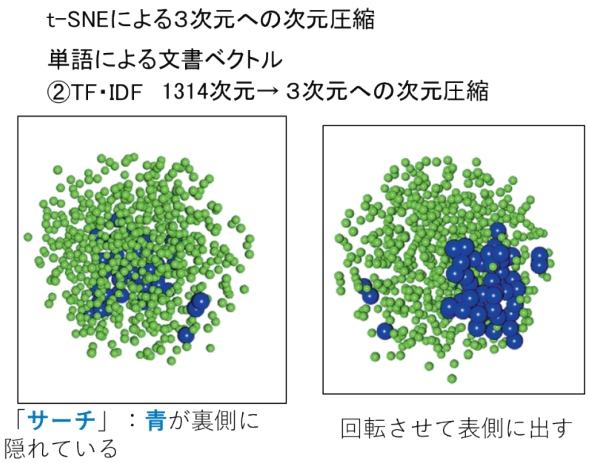


図 13　t-SNE による 3 次元への次元圧縮

に示す。平均値では③ Avg. word2vec の正解率が最も良い。① BoW も良い。⑥ Avg. fastText は最大値は良いが最小値は下方にあり性能のばらつきが大きい。図 14 下に文書分類⑥ RFg の 7 種類の文書ベクトルの分割交差検証結果を示す。全体的に① XGB より若干正解率が低く（スケールに注意）、似た傾向である。図 11、図 12 の次元圧縮による査読／ノイズ公報の位置の分布と図 14 の文書分類結果をよく見比べると、文書分類は直接文書ベクトルの各文書の 2 次元に圧縮された位置（座標）情報を用いているわけではないが相関はありそうである。特に査読／ノイズの各公報からの教師データのサンプリングにより分類結果に影響を与えそうであることが示唆される。

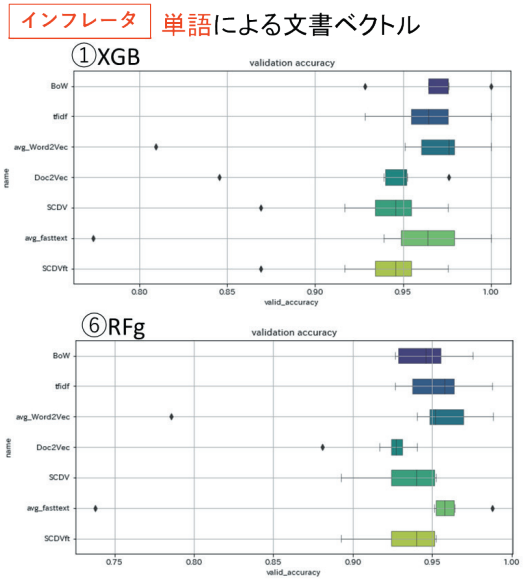


図 14　① XGB、⑥ RFg の 8 分割交差検証

機械学習で交差検証（Cross-validation）とは、母集団データを分割し、その一部をまず訓練データとして、分類器を学習させる。残りの部分で分類器のテストを行い、分類器の妥当性の検証・確認に当てる手法を指す。データの分類（出されたカテゴリーの予測）がどれだけ本当に母集団に対処できるかを良い近似で検証・確認するための手法である。

8 分割交差検証法を図 15 に示す。8 分割交差検証では、8 個あるデータブロックのうち、1 個を検証用（テスト）データ、その他を訓練データとして使う。データブロック 1 個ずつ分類器に対して訓練を行い、それを k-1 ＝ 7 ブロック分行う。訓練の終わりには必ず別にしておいた検証用（テスト）データを用いて検証する。これを繰り返し、必ず分割されたデータブロックが 1 回ずつテスト用データとして使われるようにする。

交差検証において訓練データ、検証データ、テストデータと大きく 3 種類のデータに分割する場合もある。

交差検証における訓練データ、検証データ、テストデータの使い方を入試に例えると、訓練データ＝過去問、検

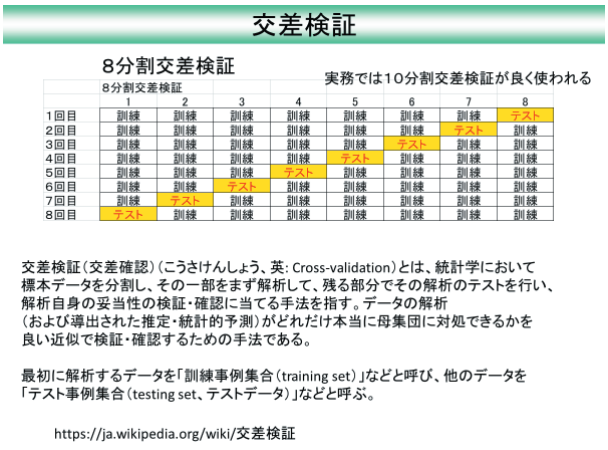
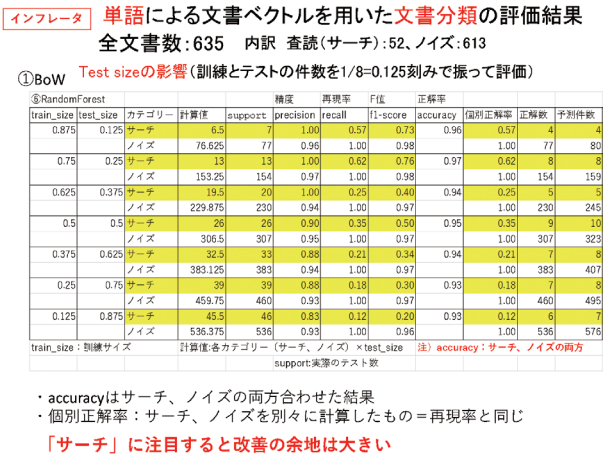


図 15　8 分割交差検証



⑥RandomForest					精度	再現率	F値	正解率			
train_size	test_size	カテゴリー	計算値	support	precision	recall	f1-score	accuracy	個別正解率	正解数	予測件数
0.875	0.125	サーチ	6.5	7	1.00	0.57	0.73	0.96	0.57	4	4
		ノイズ	76.625	77	0.96	1.00	0.98		1.00	77	80
0.75	0.25	サーチ	13	13	1.00	0.62	0.76	0.97	0.62	8	8
		ノイズ	153.25	154	0.97	1.00	0.98		1.00	154	159
0.625	0.375	サーチ	19.5	20	1.00	0.25	0.40	0.94	0.25	5	5
		ノイズ	229.875	230	0.94	1.00	0.97		1.00	230	245
0.5	0.5	サーチ	26	26	0.90	0.35	0.50	0.95	0.35	9	10
		ノイズ	306.5	307	0.95	1.00	0.97		1.00	307	323
0.375	0.625	サーチ	32.5	33	0.88	0.21	0.34	0.94	0.21	7	8
		ノイズ	383.125	383	0.94	1.00	0.97		1.00	383	407
0.25	0.75	サーチ	39	39	0.88	0.18	0.30	0.93	0.18	7	8
		ノイズ	459.75	460	0.93	1.00	0.97		1.00	460	495
0.125	0.875	サーチ	45.5	46	0.83	0.12	0.20	0.93	0.12	6	7
		ノイズ	536.375	536	0.93	1.00	0.96		1.00	536	576

train_size：訓練サイズ　計算値:各カテゴリー（サーチ、ノイズ）×test_size　注）accuracy：サーチ、ノイズの両方
support:実際のテスト数

図 16　単語による文書ベクトルを用いた文書分類結果

証データ＝模試、テストデータ＝本番入試というのが近い。

・訓練データ

訓練データは分類器を訓練する場合に使う。学習データという場合もある。重みやバイアスといったパラメータの学習に利用する。

・検証データ

訓練済みの分類器をこれを使って分類器の性能テストに利用する。これで分類器のハイパーパラメータを調整する場合もある。

・テストデータ

テスト用である。汎化性能をチェックするためには訓練、検証に使用していない、分類器にとって未知のデータである必要がる。

単語による文書ベクトルを用いた文書分類結果を図 16 に示す。文書ベクトル化① BoW ×文書分類⑥ RFg の組み合わせである。テストサイズの影響を、訓練とテストの件数を 1/8=0.125 刻みに振って評価している。サーチに注目して個別正解率＝再現率をみると訓練サイズの大きい方が良い傾向である。accuracy はサーチ、ノイズの両方合わせた結果（正解率）である。個別正解率はサーチ、ノイズを別々に計算したもの＝再現率と同じである。support は実際にテストした件数である。予測件数は分類器が予測した件数、正解数は予測した件数のうちの正解数である。

冗長であるが理解を助けるために一番上の実際の行データを左から説明すると train_size（訓練サイズ）:0.875 は比率である。test_size（テストサイズ）:0.125 は同様に比率である。train_size+test_size = 1 である。カテゴリーは各公報に人が付与したサーチ／ノイズの２値分類結果であり目的変数として使用する。計算値は各カテゴリーの公報数（サーチ=52、ノイズ=613）× train_size で、support は実際にテストした件数で（サーチ=7、ノイズ=77）である。precision は精度、適合率と呼ばれることもある。recall は再現率である。F1-score は F 値で精度と再現率の調和平均である。accuracy はサーチ／ノイズ両方合わせた正解率である。個別正解率はサーチ／ノイズを別々に計算した正解率で再現率と同じになる。正解数はサーチ :4、ノイズ :77 で、それぞれの正解件数である。予測件数は分類器が実際にサーチ :4、ノイズ :80 と予測した件数である。サーチについては予測件数 4 で正解数 4 であるので精度＝正解数 4/ 予測件数 4=1.00 である。サーチの再現率は正解数 4 ／ support7=0.57 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 17 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 [12] 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 ター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。

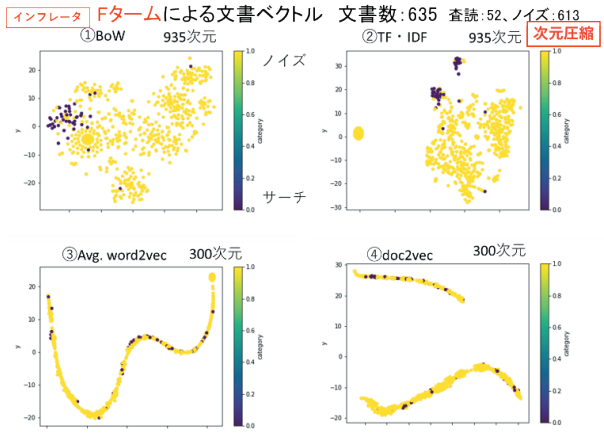


図 17　F タームによる文書ベクトルの次元圧縮・可視化

③ Avg.word2vec や④ doc2vec 等の分散表現ベクトルには不向きなデータ構造と考えられる。実際に③ Avg.word2vec、④ doc2vec の次元圧縮結果は特異なパターンを示し分類結果も良くない。

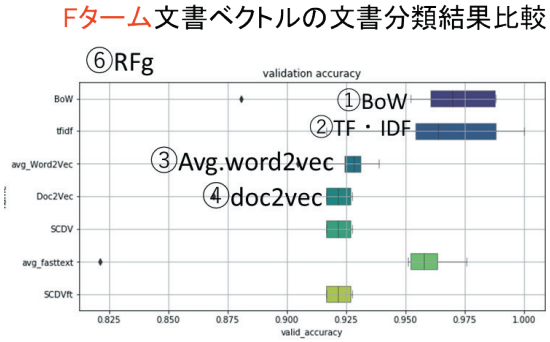


図 18　F タームよる文書ベクトルを用いた文書分類結果

文書分類方法⑥ RFg の① ~ ④の文書ベクトルの８分割交差検証結果を図 18 に示す。

インフレータ Fタームによる文書ベクトルを用いた文書分類の評価結果
全文書数:635　内訳 査読(サーチ):52、ノイズ:613

①BoW Test sizeの影響(訓練とテストの件数を1/8=0.125刻みで振って評価)

⑥RandomForest					精度	再現率	F値	正解率			
train_size	test_size	カテゴリー	計算値	support	precision	recall	f1-score	accuracy	個別正解率	正解数	予測件数
0.875	0.125	サーチ	6.5	7	1.00	0.86	0.92	0.99	0.86	6	6
		ノイズ	76.625	77	0.99	1.00	0.99		1.00	77	78
0.75	0.25	サーチ	13	13	0.79	0.85	0.81	0.97	0.85	11	14
		ノイズ	153.25	154	0.99	0.98	0.98		0.98	151	152
0.625	0.375	サーチ	19.5	20	0.86	0.60	0.71	0.96	0.60	12	14
		ノイズ	229.875	230	0.97	0.99	0.98		0.99	228	235
0.5	0.5	サーチ	26	26	0.78	0.54	0.64	0.95	0.54	14	18
		ノイズ	306.5	307	0.96	0.99	0.97		0.99	304	317
0.375	0.625	サーチ	32.5	33	0.90	0.55	0.68	0.96	0.55	18	20
		ノイズ	383.125	383	0.96	0.99	0.98		0.99	379	395
0.25	0.75	サーチ	39	39	0.90	0.46	0.61	0.95	0.46	18	20
		ノイズ	459.75	460	0.96	1.00	0.98		1.00	460	479
0.125	0.875	サーチ	45.5	46	0.94	0.35	0.51	0.95	0.35	16	17
		ノイズ	536.375	536	0.95	1.00	0.97		1.00	536	564
train_size：訓練サイズ			計算値:各カテゴリー（サーチ、ノイズ）×test_size					注）accuracy：サーチ、ノイズの両方			
			support:実際のテスト数								

・accuracyはサーチ、ノイズの両方合わせた結果
・個別正解率：サーチ、ノイズを別々に計算したもの＝再現率と同じ

訓練サイズが大きい方が「サーチ」の精度、再現率ともに良い

図 19　F タームによる文書ベクトルを用いた文書分類

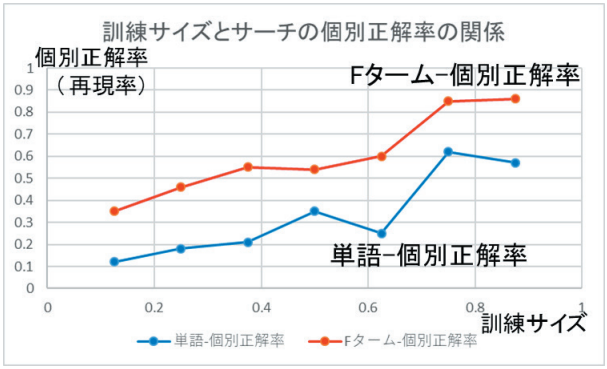


図 20　訓練サイズとサーチの個別正解率の関係

F タームよる文書ベクトルを用いた文書分類結果のまとめを図 19 に示す。

訓練サイズとサーチの個別正解率の関係を図 19 に示す。訓練サイズの大きい方がサーチの個別正解率（再現率）が高くなる。インフレータの母集団では F タームによる文書ベクトルの方が単語による文書ベクトルより全般に分類（識別）性能が良かった。

7 商用特許データベースにおける機械学習の利用事例

商用特許データベースにおける機械学習の利用事例を Questel 社 Orbit.com の分析モジュールを用いて紹介する。現状の特許検索から解析における一連の過程で機械学習はいろいろな箇所に既に組み込まれておりあまり意識しないで使用している場合もある。母集団は化粧料分野のテーマコード 4C083 の F タームを使用して次の検索式の 2779 件とした。

検索式（4C083CC23 or 4C083CC38）/FTM AND（JPA）/PN AND 2015-01-01:2019-06-30

F ターム：4C083CC23・・・洗浄剤（洗顔料，ボディ洗浄剤）、4C083CC38・・シャンプーである。洗浄剤という観点では技術内容は近いが、CC23 は洗顔あるいはボディ洗浄剤であり、CC38 は毛髪用のシャンプーであり使用部位が異なる。

（1）教師あり－クラス分類

図 21 に教師ありのクラス分類を使用したテクニカルドメインによる技術概要を示す。ヘキサゴン（６角形）は IPC で定義された技術領域である。特許全分野を５×７＝ 35 個の６角形で表している。化粧品の洗浄剤、シャンプー関連特許は Organic fine chemistry

に 2714 ファミリー、Basic materials chemistry に 947 ファミリーが一部重複して属している。IPC で約 6 万分類以上ある全分野の特許が予め定義された 35 分野に振り分けられるので技術分野の粒度が大き過ぎるのが課題である。また自分で定義したユーザー分類が使えると良い。

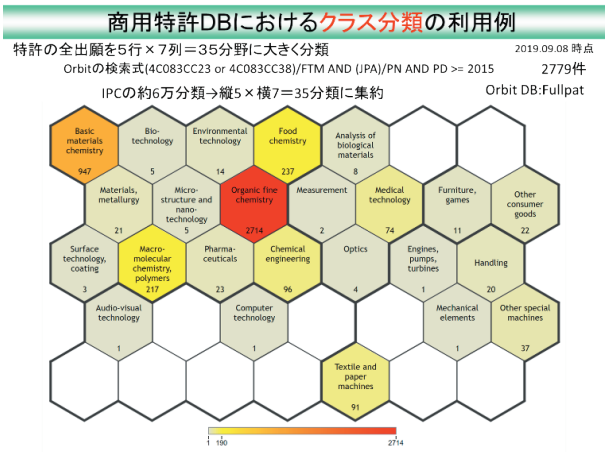


図 21　商用特許 DB におけるクラス分類の利用例

(2) 教師なし－クラスタリング

教師なしクラスタリングの事例として図 22 にコンセプトクラスターを示す。この図は英語のコンセプト（テクニカルターム）を用いて教師なし機械学習であるクラスタリングを行っている。あまりよく知らない分野における気付き（インサイト）やセレンディピティが期待される。このクラスタリングの課題は特許件数が増加あるいは減少するとクラスタリング結果が場合により大幅に異なる。また各多角形に表示されるラベルのカテゴリーが「物」であったり、「効果」であったり、例えば「シャンプー」と「ヘアシャンプー」が別のクラスターにあったりして一定しないことも課題である。また各多角形がクラスターになっておりクリックすると公報リストを表示するのだがラベルが適切に選ばれているとは言い難く中身のリストを見ないとクラスターが何を表しているか分からないことも課題である。注目したクラスターをマウスでクリックすると公報リストを表示するのはインタラクティブ性に優れクラスタリングの性質を理解して使う場合にはメリットも多い。

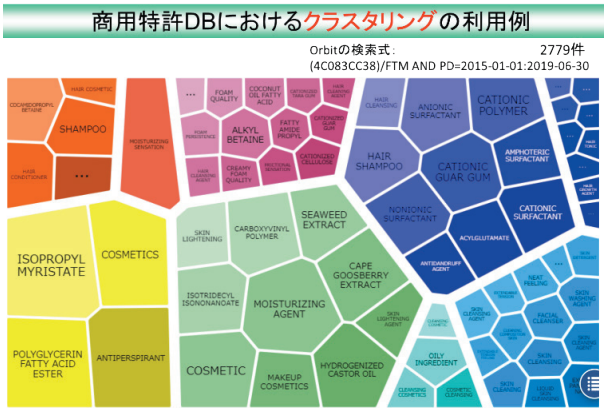


図 22　コンセプトクラスター

(3) 教師なし－次元圧縮

教師なし機械学習の次元圧縮とクラスタリングを組み合わせた事例を図 23 に示す。図はテクノロジークラスター（ランドスケープマップ）である。このマップは英語のコンセプトを用いて各公報をベクトル化してさらに次元圧縮して２次元にマッピングしている。各公報の色とラベルの色はクラスタリング結果を基にして決めていると推定される。クラスタリングアルゴリズムを使用しているため図 22 のコンセプトクラスターと同様の注意点を有している。

図 23　商用特許 DB における次元圧縮とクラスタリング

8 文書のベクトル化と文書分類（２値分類）－事例２

５章のインフレータの事例はいわゆる筋が良いデータで各文書ベクトルの次元圧縮の散布図や文書分類による査読／ノイズの２値分類結果の性能も比較的良く、解りやすい事例であった。本章ではより実務に即した SDI 調査と技術動向調査への機械学習の応用として化粧品の身体用洗浄剤とシャンプーの事例を取り上げる。下記検索式④の 2073 件を母集団とした。①②の両方が付与

されている①×②×③の公報数は 626 件である。

① 4C083CC23・・・洗浄剤（洗顔料，ボディ洗浄剤）

② 4C083CC38・・シャンプー

③公開・公表日：20150101:20190630

検索式④＝（①＋②）×③

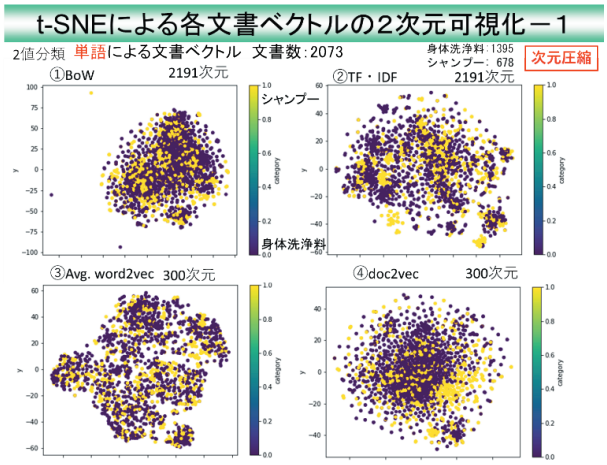


図 24　t-SNE による各文書ベクトルの２次元可視化－１

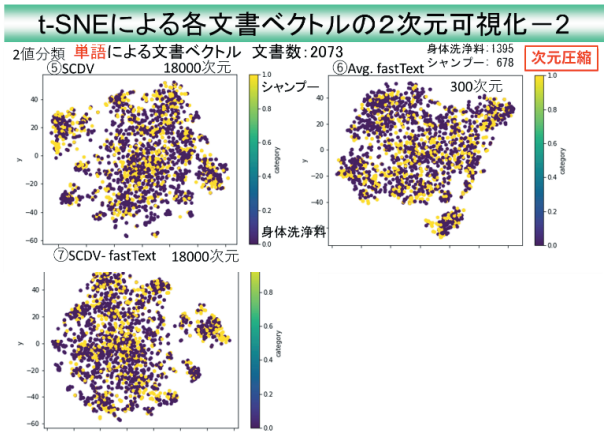


図 25　t-SNE による各文書ベクトルの２次元可視化－２

７種類の文書ベクトル化方法の t-SNE による次元圧縮結果を図 24、図 25 に示す。t-SNE では近いもの（類似）はより近くに、遠いもの（非類似）はより遠くに強調されて表示される。紺色が身体洗浄料、黄色がシャンプーである。身体洗浄料、シャンプーの両方に属する公報は身体洗浄料と見なして２値分類にカテゴリー設定している。カラーマッピングにはカテゴリーを使用している。カテゴリーはユーザー側で例えば出願人、特定の IPC、FI、F ターム、社内分類等を使用して任意に設定可能である。

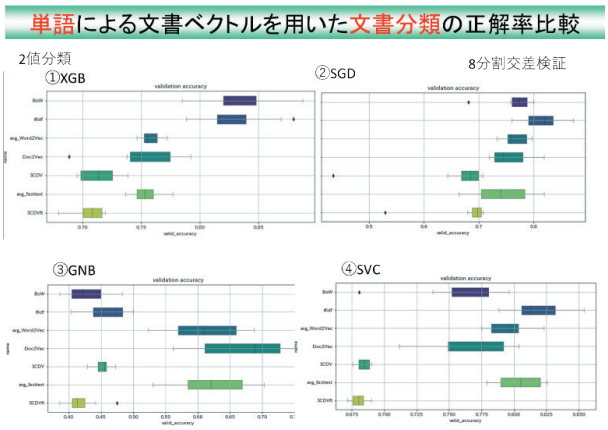


図 26　単語による文書ベクトルの８分割交差検証

①～④の４種類の文書分類方法の８分割交差検証方法による分類の正解率比較を図 26 に示す。７種類の文書ベクトルを使用して文書分類検討を行った。教師データは F タームを使用して身体洗浄料（4C083CC23）、シャンプー（4C083CC38）の２値分類とした。縦軸が７種類の文書ベクトル方法、横軸が正解率である。文書分類方法は① XGB が良かった。

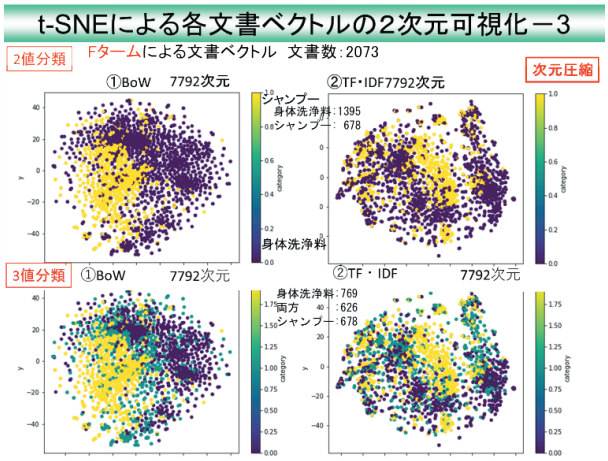


図 27　F タームによる文書ベクトルの次元圧縮・可視化

教師データとして F タームを使用して身体洗浄料（4C083CC23）、シャンプー（4C083CC38）の２値分類とした場合と、身体洗浄料、シャンプーの両方付与されているものを加えた３値分類の次元圧縮結果を図 27 に示す。F タームベクトルの２値分類と３値分類の次元圧縮結果を見比べると（それぞれ上下の図で比較）３値分類の両方の F タームが付与されている公報（黄緑色）が身体洗浄料（紺色）、シャンプー（黄色）の境界にプロットされている件数が多く興味深い。

F タームベクトルを説明変数として使用した① BoW

モデルの文書分類結果が２値分類、３値分類ともに①XGB、④SVC、⑦AdaBoostの分類モデルが100パーセントの正解率を示した。化粧品分野のようにFタームが体系的に付与されている分野では機械学習を用いた効率的な特許調査に、Fタームベクトルは有力な選択肢になると思われる。

9 まとめ

本稿では７種類の文書ベクトル化方法と８種類の文書分類方法の組み合わせを検討した。更にデータソースとして単語とFタームによる２種類の文書ベクトルを比較した。過去の独自分類結果の蓄積を教師データとして利用できる場合はSDI調査へ応用し新着公報の自動分類に有望である。教師あり－クラス分類と教師なし－次元圧縮を組み合わせた分類結果付きの俯瞰可視化マップは動向調査に有益である。

分散表現ベクトルは期待される性能をまだ十分に発揮していないように思われる。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（Fターム等）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。

10 終わりに

2018年の「BERT」[13]発表後の最近の言語モデルの発展[14)-16)]には目を見張るものがある。

今後もこの言語モデルの発展からしばらく目が離せない状況が続くと思われる。

本報告は2019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）アジア特許情報研究会
https://sapi.kaisei1992.com/
3）安藤 俊幸，「機械学習を用いた効率的な特許調査 アジア特許情報研究会における研究活動紹介」
http://www.tokugikon.jp/gikonshi/291/291kiko1.pdf
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）scikit-learn
http://scikit-learn.org/stable/
7）gensim
https://radimrehurek.com/gensim/
8）SCDV：Sparse Composite Document Vectors using soft clustering over distributional representations
https://arxiv.org/pdf/1612.06778.pdf
9）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
10）XGBoostの主な特徴と理論の概要
https://qiita.com/yh0sh/items/1df89b12a8dcd15bd5aa
11）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
12）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

13）BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/abs/1810.04805

14）ERNIE: Enhanced Representation through Knowledge Integration
https://arxiv.org/abs/1904.09223

15）XLNet: Generalized Autoregressive Pretraining for Language Understanding
https://arxiv.org/abs/1906.08237

16）RoBERTa: A Robustly Optimized BERT Pretraining Approach
https://arxiv.org/abs/1907.11692

上記 URL はいずれも 2019 年 8 月 30 日に確認したものである。