寄稿１

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[1]における研究活動紹介

花王株式会社 知的財産部　　安藤 俊幸

抄録

最近ではAIの中心技術である各種機械学習のオープンソースライブラリが容易に入手可能である。特許調査担当者の実務的な観点から機械学習を用いた効率的な特許調査の可能性について述べる。先行技術調査ではdoc2vecによる公報文書単位のスコアで査読の優先順位を付け、文単位で発明の要素毎の類似文抽出検討を行い、13種類の教師あり分類アルゴリズムで適合判定を検討した。文単位の類似文抽出で記載の根拠箇所特定の可能性が示せた。動向調査では教師あり機械学習の１次元CNNによる文書分類と教師なしの次元圧縮による文書の可視化検討を行った。文書分類はSDI調査の効率化を目指している。調査目的に応じたアルゴリズムと特徴量の選択が重要である。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。

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[2]。特許庁の「特許行政年次報告書2018年版」[3]でも「AIと特許」としてコラムで取り上げられている。また「特許出願技術動向調査等報告」の平成26年度電気・電子分野で「人工知能技術」[4]としてまとめられている。

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[5][6]。「知財管理」誌でも2018年８月号で「ミニ特集：第４次産業革命（その１）AIに関する知財動向とビジネス上の留意点」が組まれている[7][8]。

筆者が所属しているアジア特許情報研究会は2008年に設立し中国、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。特許庁の特許情報室の皆様とアジア特許情報研究会ではここ数年定期的に情報交換をさせていただいている。研究会の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。この分野は情報が陳腐化するのが速いこともあり研究成果は学会発表、論文投稿等で旬なうちに発表を

1）アジア特許情報研究会　http://www.geocities.jp/patentsearch2006/asia-research.html
2）「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
3）特許庁「特許行政年次報告書2018年版」第２部 第１章、127ページ（2018年）
4）特許庁「特許出願技術動向調査等報告」電気・電子、平成26年度、「人工知能技術」（2014年）https://www.jpo.go.jp/shiryou/pdf/gidou-houkoku/26_21.pdf
5）「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/
6）「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
7）ソフトウェア委員会第２小委員会、「AIにおける知財戦略に関する調査・研究－世界動向と法改正の方向を踏まえた，AIに係る各プレイヤーの留意点－」、知財管理、p1019-1052
8）ソフトウェア委員会第２小委員会、「AI発明のビジネス上の留意点に関する研究」知財管理、p1053-1065

目指して活動している。

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化してきており最新の種機械学習ライブラリが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。AIや機械学習関係の書籍も毎月新しい本が店頭に並び自分のPCで手軽に試してみることができる。ただし自分の業務で使いこなして有用な結果を得るまでには「人」の側で習得すべき事項も多い。本稿では企業内の特許調査担当者の実務的な観点から機械学習を用いた効率的な特許調査の可能性について述べる。まだ検討途中のことも多く筆者の私見であることに予めご留意頂きたい。

2. 機械学習を用いた効率的な特許調査の概要

図1は最近筆者が考えている特許調査と機械学習の精度（調査効率）向上への取り組みの全体像の概要である。一文に要約すると「調査目的に合わせたアルゴリズムとドメインデータの選択と最適化を行い学習済モデルを作成・利用する。」となる。ここで調査目的とは特許調査の種類を指している。

先行技術調査は第1ステップとして新規性について考慮し進歩性に関しては次の段階で考えることとする。先行技術調査では発明の構成要素毎にパッセージ検索を行い該当箇所の抽出を目指す。パッセージ検索におけるパッセージとは、文書中でユーザーのクエリの内容と強く関連する内容の記載箇所のことである。無効資料調査は先行技術調査と検索・機械学習の観点からは概ね同様と捉えている。

SDI（Selective Dissemination of Information）調査の調査範囲は既に確定（検索式有）済として、人が判定した査読/ノイズの教師データはそれなりの数があるものとして機械学習での2値分類（査読/ノイズ）での調査効率向上を最初の目的としている。

技術動向調査は機械学習による文書分類（自社分類）で調査効率向上を目指す。さらに教師データ無しの機械学習による次元圧縮で各公報間の関係を俯瞰・可視化する。

クリアランス調査（侵害防止調査）は網羅性重視が必須のためリスクと調査効率バランスが重要である。調査の観点からは調査を実行する調査担当者側の開発製品、技術、リスク評価の正しい理解・把握、クリアランス対象製品の市場動向（規模）、製品がライバル関係にある競合会社の動向等のビジネス情報等の非特許情報も含めた調査対象の確定とそれに対応した特許調査範囲の確定が重要である。海外のクリアランス調査はさらに考慮すべき要因が増えるので必要な場合は専門家への相談も考えると良い。機械学習の観点からはリスクをどのように見積もるかが査読の優先順位を付ける上で重要である。クリアランス調査は本稿ではこれ以上詳細には扱わない。

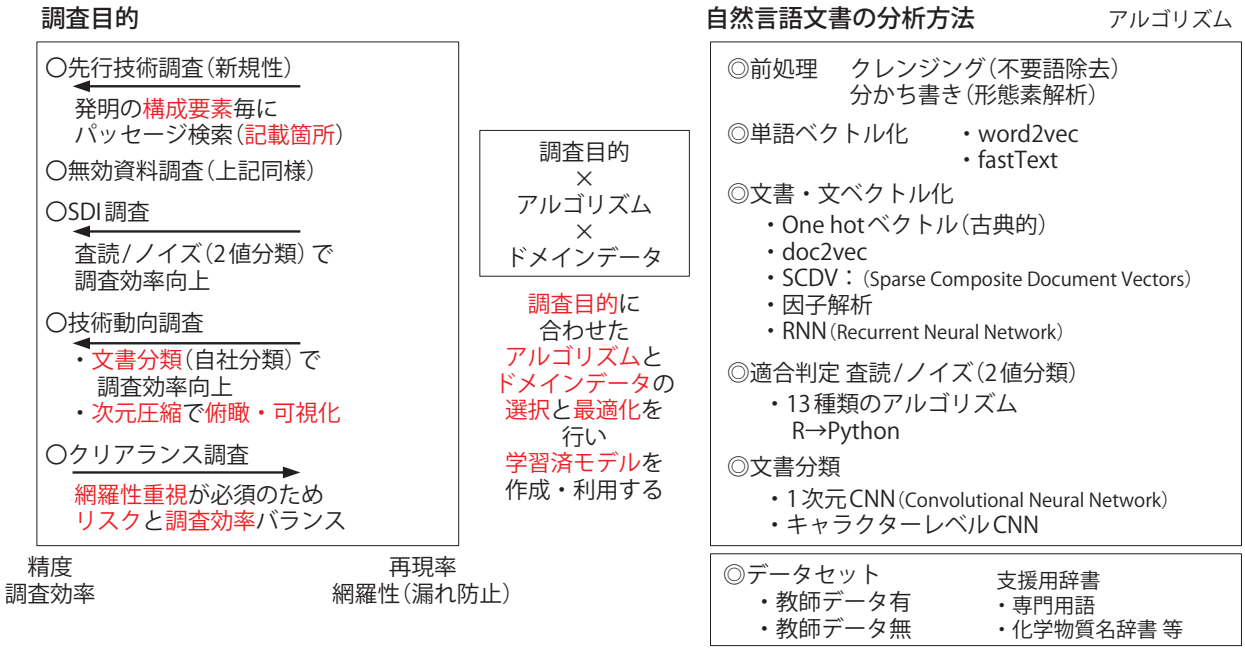


図1　特許調査と機械学習の概要

図1の各調査の種類の下の矢印の向きはそれぞれの調査が精度（precision）重視か再現率（recall）重視か筆者の考える一般的な方向性を示している。個々の具体的な調査ではケースバイケースでこれと異なる場合も当然あり得る。

図1右側の自然言語文書の分析方法の部分は上から下に機械学習を行うための特許文書の大まかな処理とその検討対象アルゴリズムを示している。この部分は後程詳細に検討する。

図1右下のデータセット部分は調査対象のドメインデータであり調査対象母集団とその教師データのデータセットである。図では小さな矩形領域で示されているがこの部分のデータと教師データの質と量で行える機械学習の種類や学習済みモデルの性能が決まる重要な部分である。支援用辞書は教師データ作成支援を念頭に置いているが各種調査のキーワード関連の支援ツールとしても使用可能である。後程具体例を示す。

3. 特許調査における機械学習使用時の留意点

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機械学習には原理的な限界が存在することが理解できる。機械学習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人工知能には、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。特許調査において調査目的に応じてどこまで調べるか調査範囲を決めることは非常に重要である。特許分類（IPC,FI,F ターム等）を有効利用することは重要ポイントである。

(2) ノーフリーランチ定理（NFL 定理）[9]

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。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。特許調査に当てはめると調査目的に合った適切な機械学習のアルゴリズムを選択することが重要である。

(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。「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。より具体的には醜いアヒ

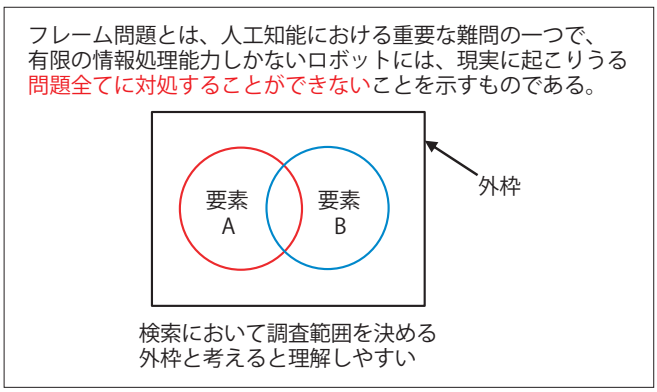
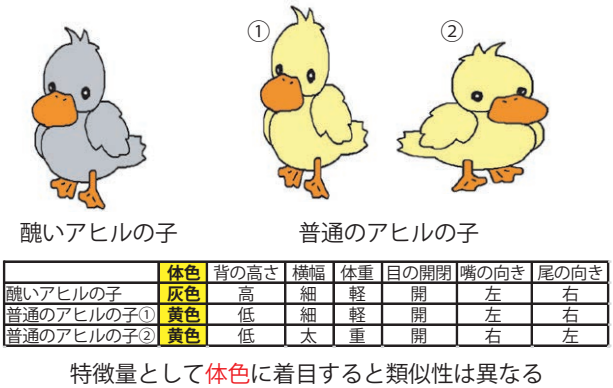


図2　フレーム問題



	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3　醜いアヒルの子の定理

9）ノーフリーランチ定理　https://ja.wikipedia.org/wiki/ノーフリーランチ定理

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。

ディープラーニング登場前の機械学習では「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行われていた。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する。

(4) 過学習[10)]

過学習の概念は機械学習において重要である。通常、学習アルゴリズムは一連の訓練データを使って訓練される。つまり、典型的な入力データとその際の既知の出力結果を与える。学習者はそれによって、訓練データでは示されなかった他の例についても正しい出力を返すことができるようになると期待される。しかし、学習期間が長すぎたり、訓練データが典型的なものでなかった場合、学習者は訓練データの特定のランダムな（本来学習させたい特徴とは無関係な）特徴にまで適合してしまう。このような過剰適合の過程では、訓練データについての性能は向上するが、それ以外のデータでは逆に結果が悪くなる。

4. 特許公報（文書）の類似度による先行技術調査

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4に特許検索競技大会2016の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。ここでは特許公報（文書）の各種の類似度を使用してスクリーニング過程を詳細に検討した。

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」[11)]、NRIサイバーパテントデスク（株）の「CyberPatent Desk」[12)]、を使い検索競技大会の問題文（図5）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

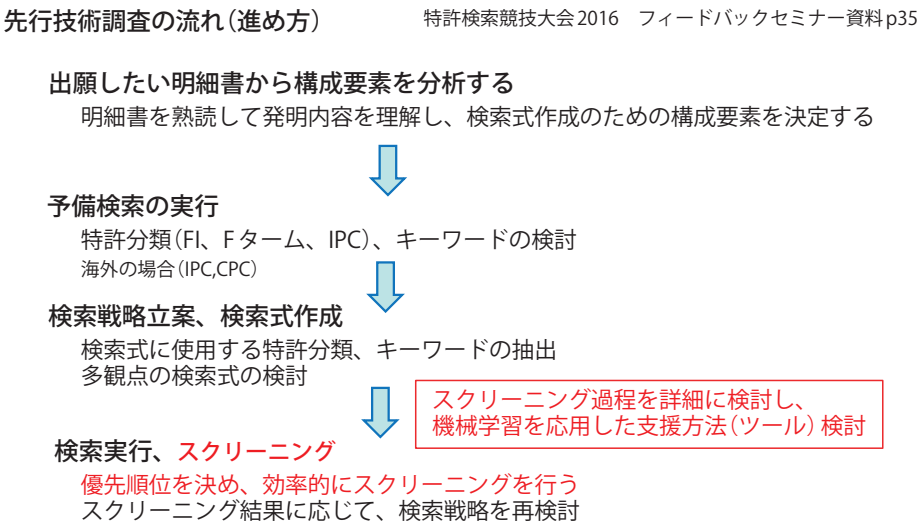


図4　先行技術調査の流れ

10）過剰適合　https://ja.wikipedia.org/wiki/過剰適合
11）日立 特許情報提供サービス「Shareresearch」　http://www.hitachi.co.jp/Prod/comp/app/tokkyo/sr/
12）NRI サイバーパテントデスク（株）提供「CyberPatent Desk」　https://s.patent.ne.jp/

49件の和集合746件をデータセットとした[13]。2017年作成のデータセットをそのまま使用した。

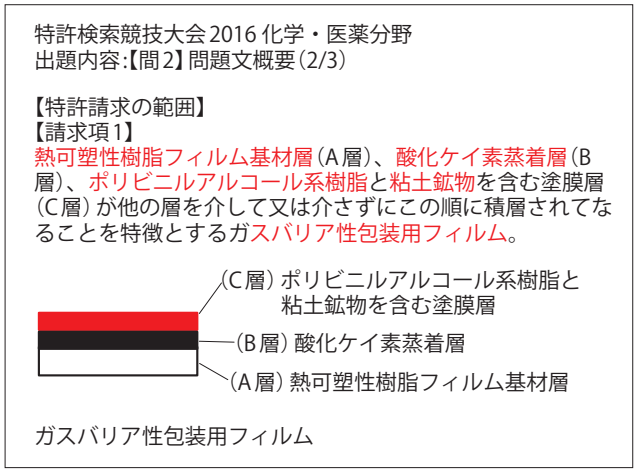


図5　特許検索競技大会2016の化学・医薬分野の問2

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

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[14]。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[15]。

図6に形態素解析(MeCab)による分かち書き、図7に専門用語による分かち書き例を示す。専門用語は名詞の連接部分を専門用語として抽出している。

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6　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7　専門用語による分かち書き

図8に分かち書きと重み付けの再現率への影響を示す。横軸は査読時のスコア上位の公報から内容を確認していく場合の確認数である。縦軸は再現率である。再現率は確認数における正解公報(この例ではトータル49報中の)出現割合である。精度(調査効率)重視の観点からは再現率曲線の立ち上がりが急峻な方が良い。正解公報が全て確認数の上位に並ぶ理想的な場合を「理想」として図中にプロットしている。確認数が少ない再現率曲線の出だしはSR(Shareresearch)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局的に見ると差は意外に少ない結果となった。SRはShareresearchの略で概念検索(類

13) 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
14) 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
15) 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

分かち書き（形態素、専門用語）と重み付け（TF 、 TF・IDF）の再現率への影響

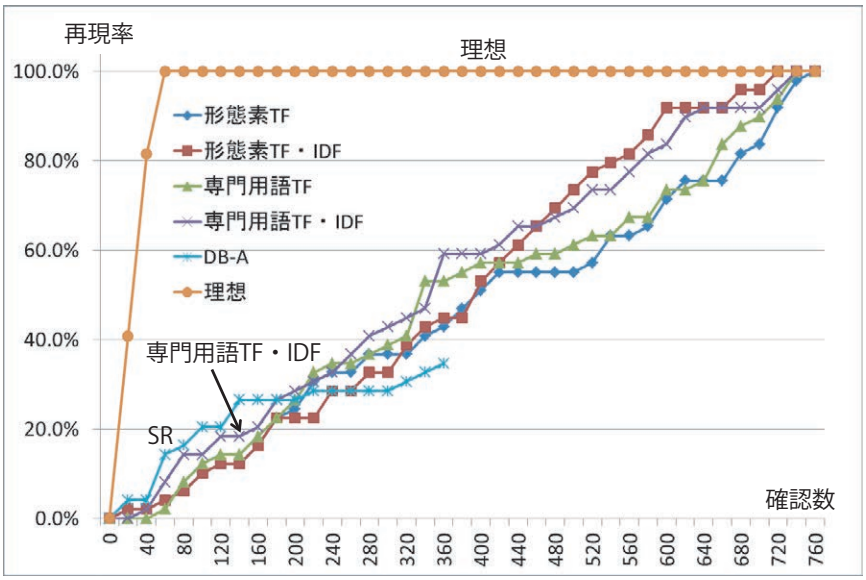


図8　分かち書きと重み付けの再現率への影響

似検索）結果をベースラインの比較対象として同時にプロットした（以下同様）。

5. 単語の分散表現用いた文書のベクトル化検討

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（50～300）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[16-19]。

図9にDoc2Vecよるベクトル化処理の概要を示す。

Doc2Vec[21]は、Word2Vec[20]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

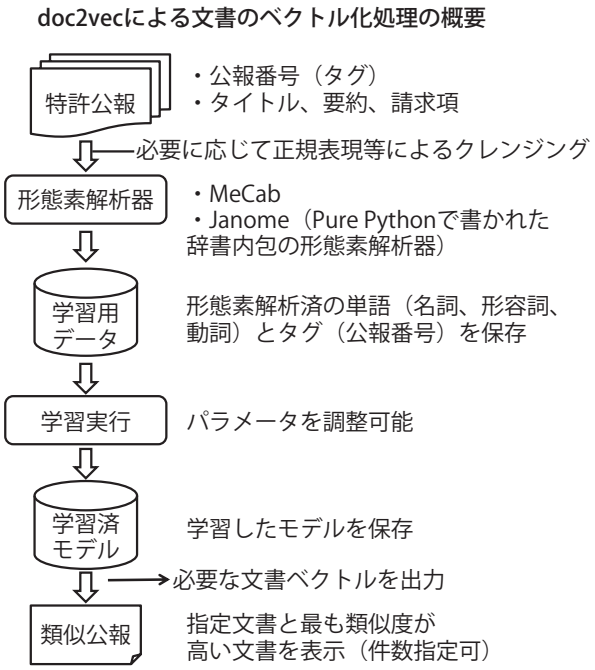


図9　Doc2vwcによる文書のベクトル化処理の概要

16) 岩波データサイエンスvol.2 [特集] 統計的自然言語処理ーことばを扱う機械　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

17) 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）

18) 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

19) 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

20) 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.

21) Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In International Conference on Machine Learning, Vol. 14, pp. 1188-1196, 2014.

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[22]（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[23]を使用した。Janomeは形態素解析用の辞書としてMeCab[24]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10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SRの再現率曲線を圧倒している。もちろんSR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。

6. 分散表現を用いた「文」単位の類似度計算検討

6章では前章の公報文書のベクトル化に対して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

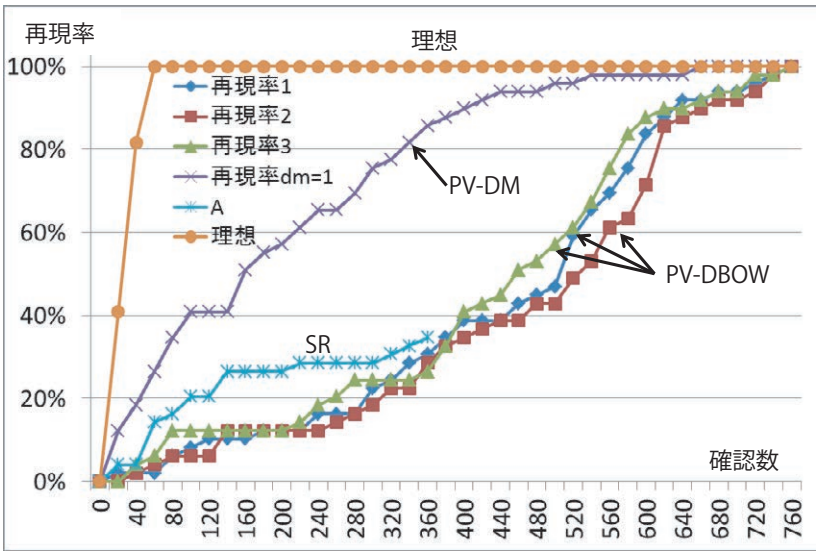


図10　文書の分散表現ベクトルの学習モデルと再現率

22) gensim　https://radimrehurek.com/gensim/

23) Janome　http://mocobeta.github.io/janome/

24) MeCab　http://taku910.github.io/mecab/

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11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。

図12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13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。

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11　構成要素分析（検索競技大会の模範解答例）

分布仮説に基づいた文脈中の単語の重み学習（word2vec）

分布仮説
・類似する文脈でよく使われる表現は似た意味を持つ
・単語の意味はその周辺単語の分布により知ることができる

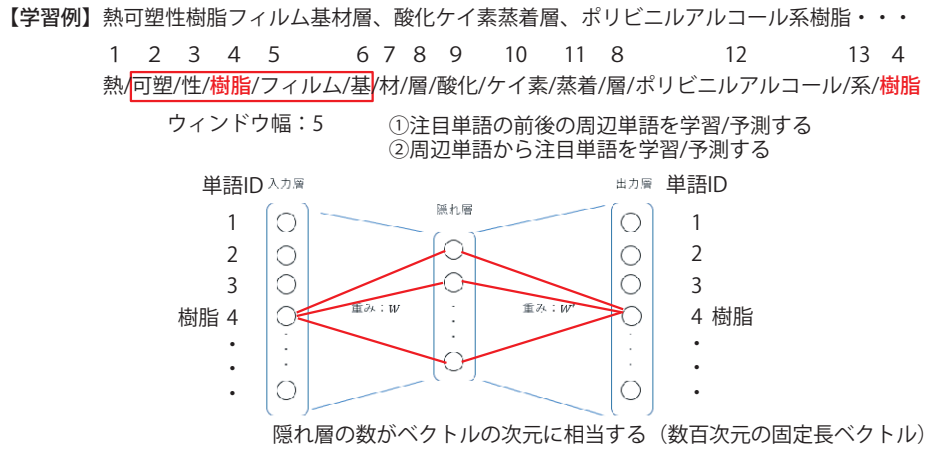


図12　分布仮説に基づいた文脈中の単語の重み学習

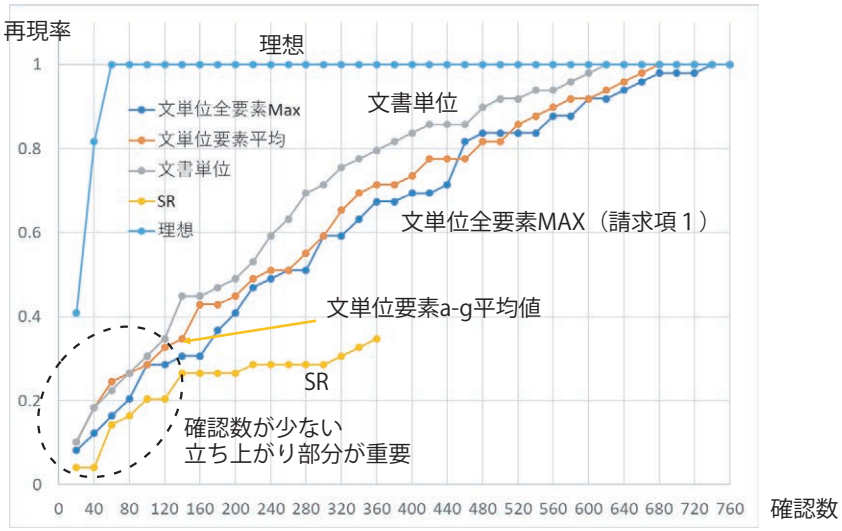


図13　「文」単位での類似度計算による再現率曲線

図14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11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15、図16に文の分節とクエリ拡張の影響を示す。クエリに請求項1（P0_C1）、主要な構成要素を

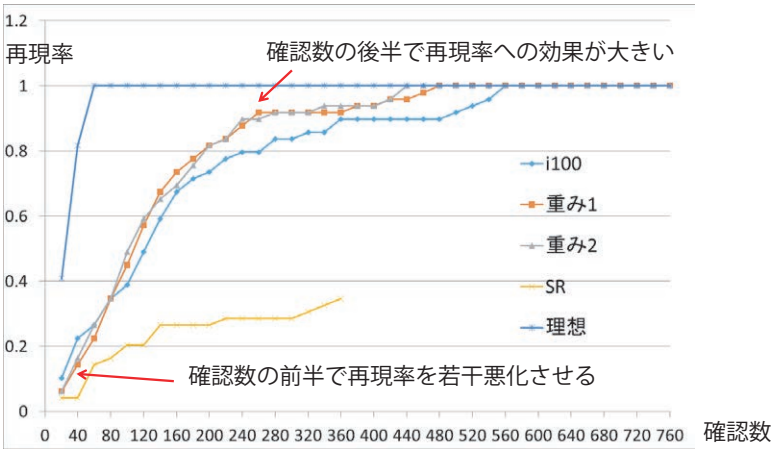


図14　発明の構成要素の重み付け検討

PatNo	TACE
P0_T1	ガスバリア性包装用フィルム。
P0_A1	ポリプロピレン、ポリエチレンテレフタレート、ナイロンなどの熱可塑性樹脂からなるフィルムは、透明性、耐熱性を有するため様々な用途に広く用いられている。
P0_A2	しかし酸素や水蒸気バリア性能が求められる用途、例えば鮮度が求められる食品のパッケージ用途には適さない。
P0_A3	そのため、従来から熱可塑性樹脂フィルムとアルミニウム箔とを積層したフィルムが食品用のパッケージとして用いられてきた。
P0_A4	しかしアルミニウム箔を積層したフィルムは、ガスバリア性能は優れる一方で、フィルムの向こう側が視認不能となる上、金属探知機の使用ができなくなるという問題がある。
P0_A5	これらの問題を解決するフィルムとして、熱可塑性樹脂フィルムに酸化ケイ素等の無機酸化物を蒸着したものが開発されているが、そのガスバリア性能は鮮度が求められる食品の保存用途としては十分でなかった。
P0_A6	そこで、酸化ケイ素蒸着層の上にポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層を設けることで、これらの問題を解決したガスバリア性包装用フィルムの発明に至った。
P0_C1	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。
P0_C2	熱可塑性樹脂がポリプロピレン、ポリエチレンテレフタレート、ナイロンから選ばれた請求項1記載のガスバリア性包装用フィルム。
P0_C3	粘土鉱物がカオリナイト、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、パーミキュライト、金雲母、ザンソフィライト、緑泥石から選ばれた請求項1記載のガスバリア性包装用フィルム。
P0_E1	ポリビニルアルコール水溶液に、モンモリロナイトを加え60℃で75分間撹拌した。
P0_E2	その後、さらに2-プロパノールを添加し、その混合液を室温まで冷却して塗工液を得た。
P0_E3	熱可塑性フィルム基材として厚さ15μmのポリエチレンテレフタレートフィルムを用い、この一方の面上に酸化ケイ素を蒸着した。
P0_E4	蒸着層の上に塗工液をグラビアコート法により形成し、ガスバリア性包装用フィルムを得た。
P0a_C1	熱可塑性樹脂がポリプロピレン、ポリエチレンテレフタレート、ナイロンから選ばれた熱可塑性樹脂フィルム基材層。
P0b_C1	酸化ケイ素蒸着層。
P0c_C1	ポリビニルアルコール系樹脂を含む塗膜層。
P0d_C1	粘土鉱物がカオリナイト、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、パーミキュライト、金雲母、ザンソフィライト、緑泥石から選ばれた粘土鉱物を含む塗膜層。
P0e_C1	他の層を介してまたは介さずにこの順に積層。
P0f_C1	ガスバリア性。
P0g_C1	包装用フイルム。

記載部分略号
T：タイトル
A：要約
C：請求項
E：実施例

図15　分の文節とクエリ拡張の影響（クエリ）

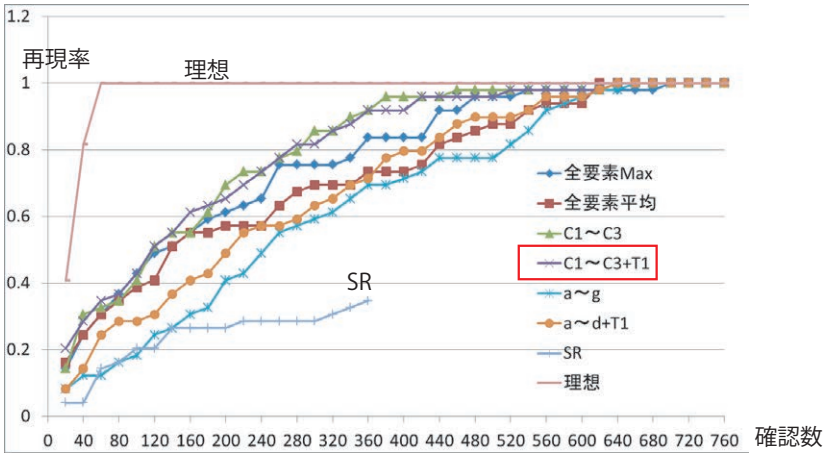


図16　文の文節とクエリ拡張の影響（結果）

含む実施例（P0_C2, P0_C3）、タイトル（P0_T1）を入力した場合が最も良い結果になっている。

図17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次章で検討する。）

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。

7. 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[25] 8.4のDeep Learningアドオン（Deep Learner）[26]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

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	○

平均値：0.599

構成要素
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

記載部分略号
T：タイトル
A：要約
C：請求項
E：実施例

図17　発明の構成要素毎の根拠個所（文）抽出結果

25) Visual Mining Studio　https://www.msi.co.jp/vmstudio/
26) Deep Learning アドオン（Deep Learner）　https://www.msi.co.jp/vmstudio/deepLearning.html

のテキストマイニングツールText Mining Studio[27]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18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。画像データの分類処理は別製品AutoDL[28]である。

NTTデータ数理システムDeep Learnerの
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18　データタイプ・学習別処理内容と特色

図19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。

Deep Learningアドオンではリアルタイムに誤差が減っていく様子を確認できる。

図21の結果は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。

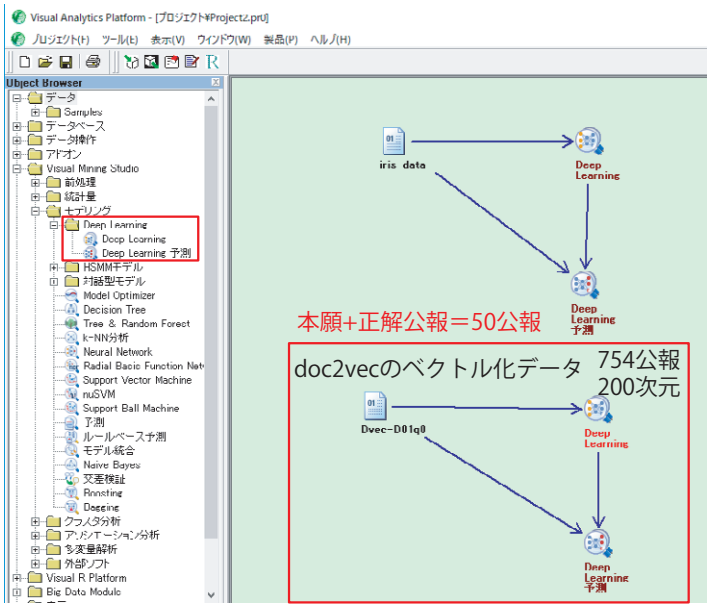


図19　Deep Learningアドオン（Deep Learner）

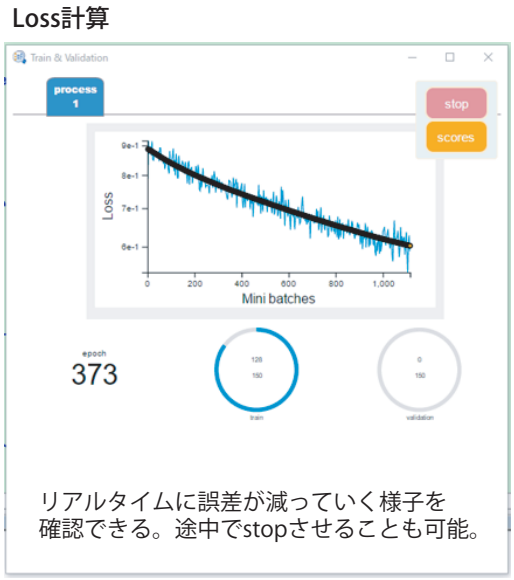


図20　Deep Learningアドオン

	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%		

図21　多層ニューラルネットワークによる教師あり学習

27) Text Mining Studio　https://www.msi.co.jp/tmstudio/
28) AutoDL　http://www.msi.co.jp/AutoDL/

8. 適合判定への応用検討（分類器アルゴリズム比較）

統計解析に強いR言語を使用して適合判定への応用を目的に13種類の分類器のアルゴリズムを予備的に比較検討した[29]。データセットはRパッケージkernlabの英文メール4601通（スパム1813、非スパム2788）のスパムメール識別のデータセットを使用した。このデータセットは57個の特徴項目を備えている。図22にスパムメール識別の正解率の箱ひげ図を示す。正解率が良かった分類器はエイダブースト、ランダムフォレストであった。DNNetはH2O社が開発したオープンソースのディープラーニングプラットホームH2Oで構築された隠れ層（250,500,100,50,20）5層の深層ネットワークである。隠れ層1層のニューラルネットワークaveNNetとあまり差が無い。スパムメールフィルターに使われているナイーブベイズがあまり奮わず意外であった。図17の発明の構成要素毎の根拠個所（文）抽出結果の適合判定に応用してみたいと考えている。

9. 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[30]というワシントン大学とカーネギーメロン大学によって開発された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[31]。MXNetはPythonを始めR、Scala、Julia、Perl、C＋＋等多く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（Version2.01）」[32]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一種である1次元CNN（Convolutional Neural Network）[33]を使用した。トレーニング文書で教師データ（15カテゴ

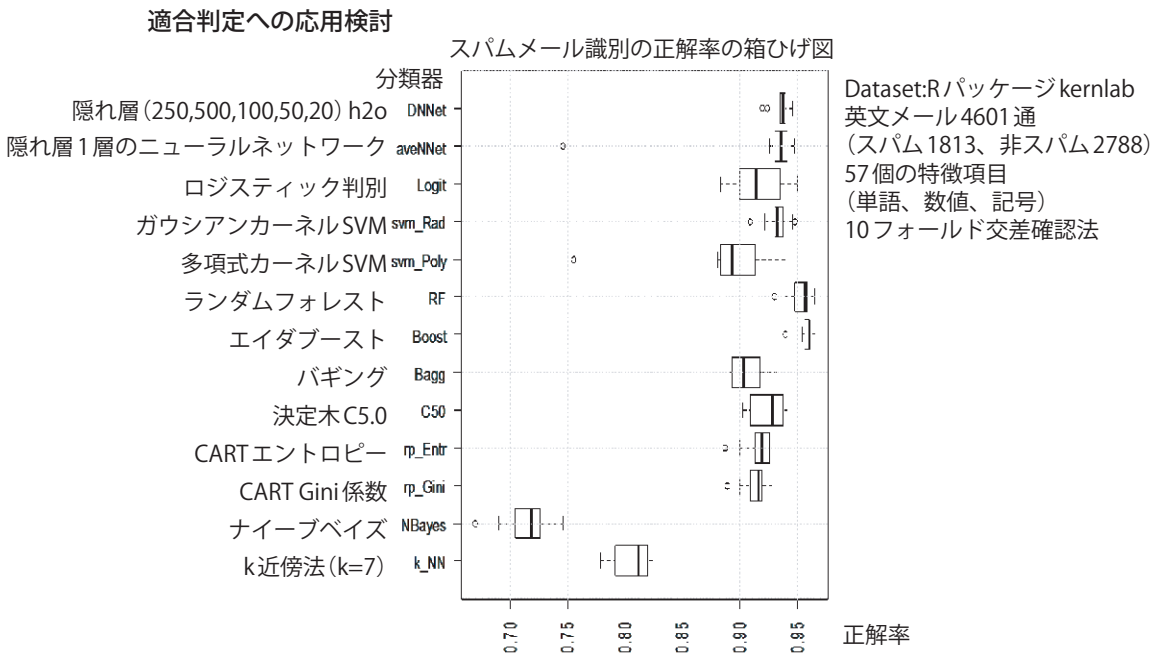


図22　スパムメール識別の正解率の箱ひげ図

29）金明哲.テキストアナリティクス,共立出版,2018, p.152-158.
30）Apache MXNet　https://mxnet.apache.org/
31）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
32）Wikipedia日英京都関連文書対訳コーパス　https://alaginrc.nict.go.jp/WikiCorpus/
33）Yoon Kim, Convolutional Neural Networks for Sentence Classification　https://arxiv.org/abs/1408.5882

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手法として下記①～③の3種類を検討した。自然言語で書かれた文書をコンピュータ処理しやすいように何らかの方法でベクトル化すると高次元空間上に類似の公報が距離が近くに配置される。高次元空間は人間には直接認識できないので何らかのアルゴリズムで2次元あるいは3次元に次元圧縮すると文書の相互関係を可視化できる。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[34]の潜在的意味解析（Latent semantic analysis: LSA）で使われる特異値分解（Singular value decomposition: SVD）を使用した。scikit-learnには各種の次元圧縮方法が実装されている。①～③の図23～図25のベクトル化散布図のX（横）軸、Y（縦）軸は距離（非類似度）に相当する。同じ色（カテゴリー）が近くにまとまっているほど人が分類したカテゴリーを上手く可視化していることを意味する。

①SCDV:Sparse Composite Document Vectors[35]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[36]。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[22]（Python用の自然言語処理ライブラリ）から実行した。15のカテゴリー毎に色付けしている。

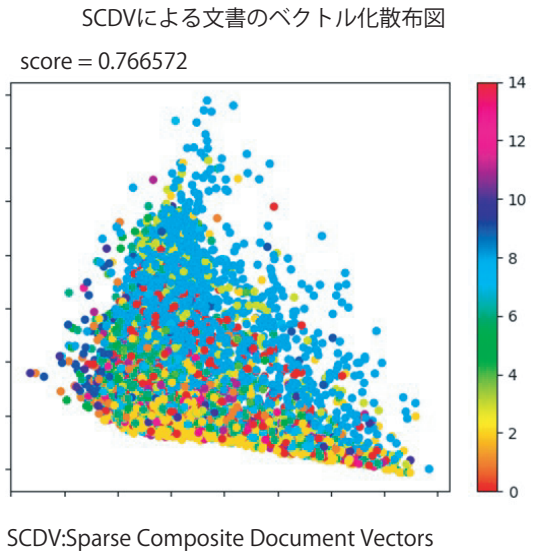


図23　SCDVと次元圧縮による文書の散布図

②因子解析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24に散布図を示す。

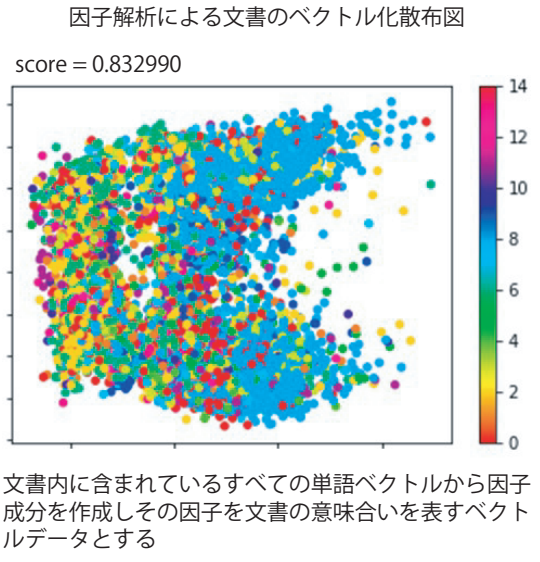


図24　因子解析による文書のベクトル化散布

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

35) SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations
https://dheeraj7596.github.io/SDV/　https://arxiv.org/abs/1612.06778

36) fastText　https://fasttext.cc/

③RNNによる文書のベクトル化 score＝0.575706

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。ただし計算に時間がかかるためエポック数を10として実行した。エポック数とは、「一連の訓練データを何回繰り返して学習させるか」の数のことである。計算時間はCPU4コアを使用して計算し（GPUは使用しない）1時間以上を要した。ちなみにエポック数5でscore＝0.935703であり、エポック数10以降はscoreの減少は収束傾向にある。エポック数をむやみに増やしても計算時間、過学習の問題を生じる。RNNは学習モデルの作成に時間はかかるがエポック数10のRNNが検討した3種類の文書のベクトル化、次元圧縮による可視化方法のなかでscore（各カテゴリー毎のまとまり）が一番良かった。

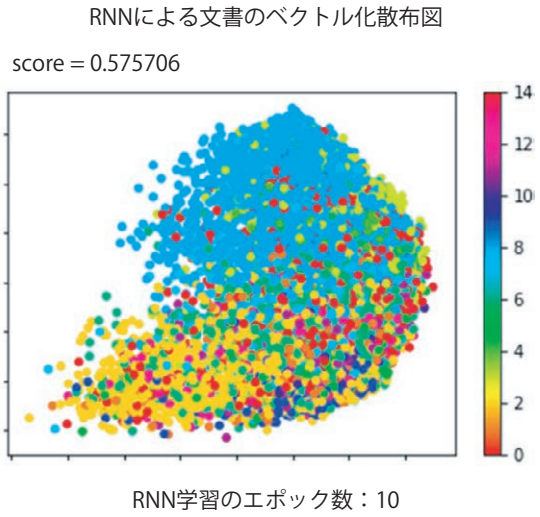
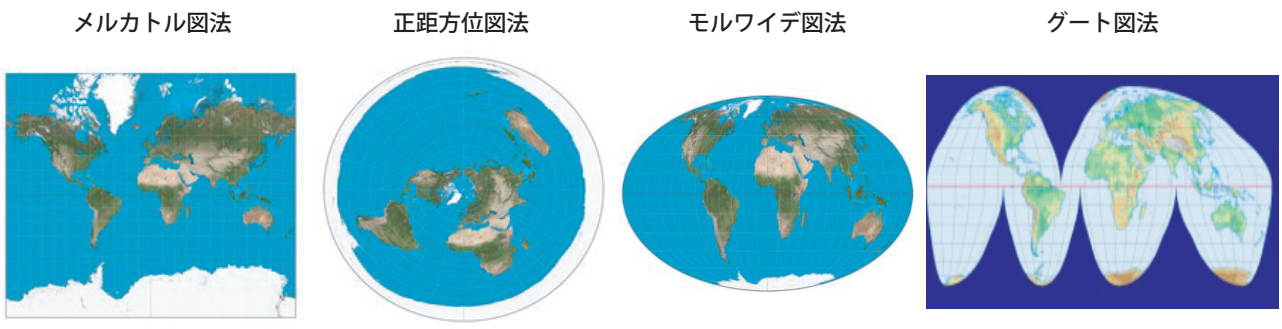


図25　RNNによる文書のベクトル化散布図

文書のベクトル化と次元圧縮の関係を直感的に理解できるように図26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。図23～図25も各文書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ことが重要である。

地球は3次元→2次元の地図に次元圧縮する各種方法とその特徴を理解して使用する



メルカトル図法

円筒投影による図法
地球表面のすべての部分の
角度が正しく表される
図上の2点を結ぶ直線は
等角航路となる
羅針盤による航海に便利
海図に利用される

正距方位図法

図の中心から他の1地点を
結ぶ直線が図の中心からの
正しい方位、最短経路を表し、
図の中心からの距離を正しく
求めることができる
飛行機の最短経路や方位を
見るために使われる

モルワイデ図法

地球を楕円形にして、
極地方の形のゆがみを
少なくした図法
分布図に使用される

グート図法

世界全体を通して大陸部分の
形の歪みができるだけ小さい
地図を作るために、サンソン
図法の地図とモルワイデ図法の
地図を組み合わせて作成
面積が正しく表されている

地図投影法学習のための地図画像素材集
http://user.numazu-ct.ac.jp/~tsato/tsato/graphics/map_projection/

図26　地図の図法とその特徴

10. 特許調査における教師データの利用について

特許調査における教師あり機械学習において教師データの利用についていくつかのヒントを示す。教師データとして利用可能なものは各種考えられる。公報文書に付与されている各種特許分類（IPC,FI,Fターム等）は教師データとして使用可能である。海外特許、例えば中国（CN）特許には限定的な例外[37]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。

特定の商用データベース、例えばCyberPatent DeskではPATOLIS（2014年1月31日にサービス終了）フリーキーワードを収録している。Questel社のグローバル特許データベースOrbit.com[38]ではKEYW：コンセプト（テキストマイニング手法で抽出した英語の専門用語）、MLID：化学物質名のIDが収録されている。

キーワードに関してデータベースに依存しない汎用的な方法として特許公報からの専門用語抽出、ワードクラウド、注目ワードに対する共起キーワード、共起キーワードのネットワーク分析による可視化[14]、word2vecによる類似キーワード抽出[13]等がある。

11. おわりに

本研究では機械学習を用いた効率的な特許調査についてある程度の可能性を示したがさらなる精度向上・実用化には様々な未検討課題がある。第一の課題はより効果的な分散表現の学習である。第二の課題は適合判定における特徴量選択である。第三の課題は分散表現学習と適合判定における特徴量選択を一気通貫に行った場合の分類アルゴリズムの選択、各種パラメータチューニングであり実用化に不可欠である。

本稿の前半では先行技術調査を念頭にdoc2vecによる文書／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どう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。本稿が効率的な特許調査の一助になれば幸いである。

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。

謝辞

特許庁総務部総務課特許情報室の皆様には情報交換の場で大変有益なアドバイスを頂きました。特許情報室の皆様に感謝申し上げます。

profile

安藤 俊幸（あんどう としゆき）

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

37）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
38）Questel社Orbit.com　https://www.orbit.com/