機械学習を用いた効率的な特許調査
ニューラルネットワークの特許調査への応用

○安藤俊幸[1)], 桐山　勉[2)]
花王株式会社[1)], はやぶさ国際特許事務所[2)]
〒131-8501　東京都墨田区文花 2-1-3
Tel: 03-5630-9538　FAX: 03-5630-9712
E-mail: ando.t@kao.co.jp

Effective patent search methods using Machine Learning:
Application of neural network to patent search

ANDO Toshiyuki [1)], KIRIYAMA Tsutomu[2)]
Kao Corporation [1)], HAYABUSA INTERNATIONAL PATENT OFFICE[2)]
2-1-3, Bunka, Sumida-ku, Tokyo 131-8501 Japan
Phone: +81-3-5630-9538　Fax: +81-3-5630-9712
E-mail: ando.t@kao.co.jp

【発表概要】

ニューラルネットワークを利用した機械学習を用いて効率的な特許調査方法を検討した。特に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。

後半はニューラルネットワークの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。

本報で検討した分散表現ベクトル（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。特許調査の精度を上げるには前処理の形態素解析による「分かち書き」が重要になる。

【キーワード】

ニューラルネットワーク, 機械学習, 分散表現, doc2vec, word2vec, 類似度, 特許調査, 先行技術調査, 特許情報解析, 可視化

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・プランが公表されており各種の実証試験が試行されている。

本報では特許調査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. 目的

機械学習の特許調査への応用の目的として下記2種類の特許調査をベースに目的を設定した。

①先行技術調査

機械学習の観点では教師データが少なくても効率的に学習して再現率と精度を両立可能な調査手法。特許検索の観点では検索漏れを少なくするように網羅性を重視した検索母集団を作成し精度を重視したスクリーニングを行い調査目的に適合したスコア付けを行う調査手法を目的とする。更に適合した部分を例えば段落単位で提示する。

②技術動向調査

膨大な特許情報から技術動向を効率的に把握する。全体像が直感的に把握できて関心がある特許公報にインタラクティブ(対話的)にアクセスできるような俯瞰・可視化とインタラクティブ操作ができる手法が理想的である。日本語、英語、中国語で解析可能であること。

3. 検討方法

図 1 に機械学習の特許調査への適用の基礎検討概要を示す。単語の One hot ベクトル表現とは文書に出現するすべての単語に固有の「その単語の有無」を表すベクトルを割り当てて表現する。単語の出現(種類)数の次元を要する。単語の出現数が増えると数万次元におよぶこともある。「単語」の分かち書き方法は形態素、専門用語、N グラム等がある。

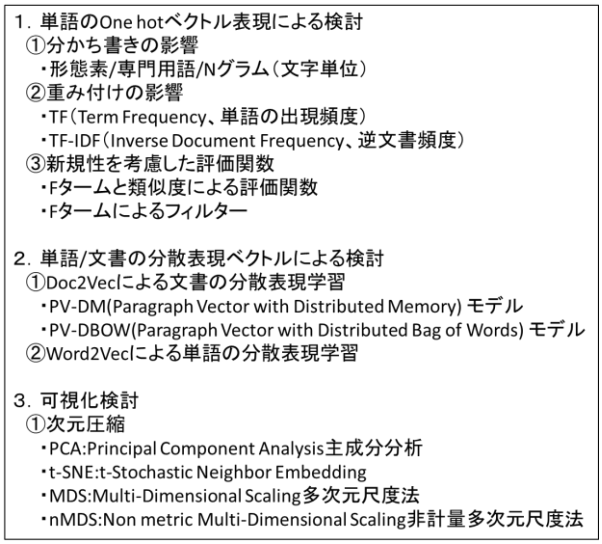

図1. 機械学習の基礎検討の概要

下記①～③に本研究で使用したデータベースとツール類を記す。

①商用特許データベースの類似検索とデータセット作成

類似(概念)検索の類似度(スコア)検討のため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、発明通信社 HYPAT-i2、NRIサイバーパテントデスク2 を使用した。データセット作成にはNRIサイバーパテントデスク2のタイトル、要約、請求項を csv 形式でダウンロードして使用した。

②機械学習

機械学習は Python3.6 で機械学習ライブラリ(scikit-learn[3)]と gensim[4)])を使用した。python 環境構築は Anaconda を使用して行った。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[5)]。

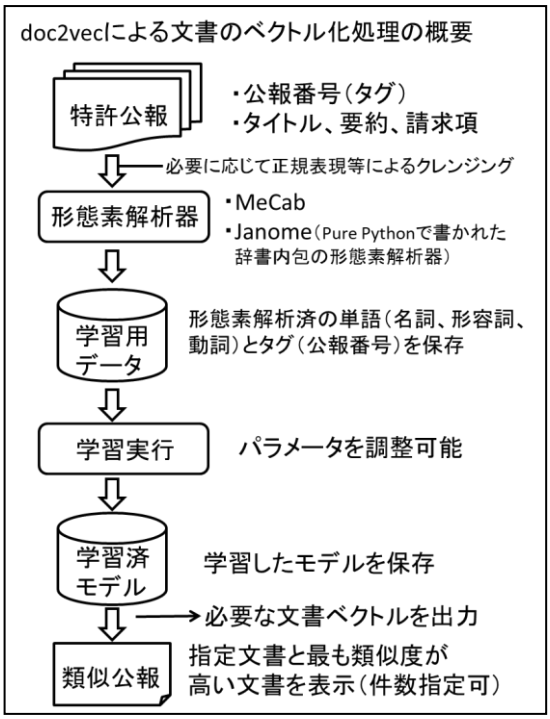

図2. doc2vec によるベクトル化処理

図2に doc2vec による文書のベクトル化処理の概要を示す。word2vec による単語の分散表現学習も同様に行った。

4. 検討・分析結果

4－1.　One hot ベクトル表現検討

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。機械学習適応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に着目した。

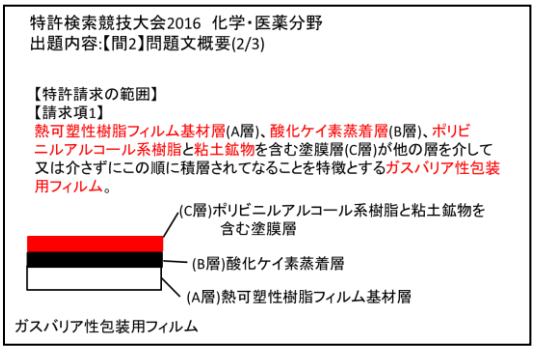

図3. 特許検索競技大会の問題

図3に特許検索競技大会 2016 の化学・医薬分野の問2を示す。請求項1を使用して商用データベースの類似検索を行い再現率で比較したグラフを図4に示す。再現率=正解数/全正解数である。

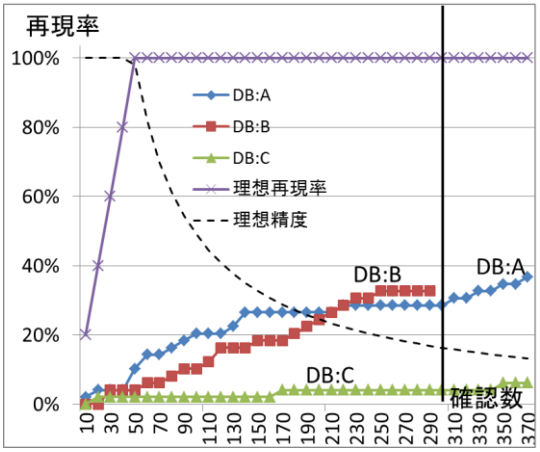

図4. 類似(概念)検索の再現率比較

確認数:300　　全正解数:49

	DB:A	DB:B	DB: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

計算例
←2/49

表1. 確認数 300 の精度、再現率、F 値

図4の横軸は類似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 時点の精度、再現率、F 値を表1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。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。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。

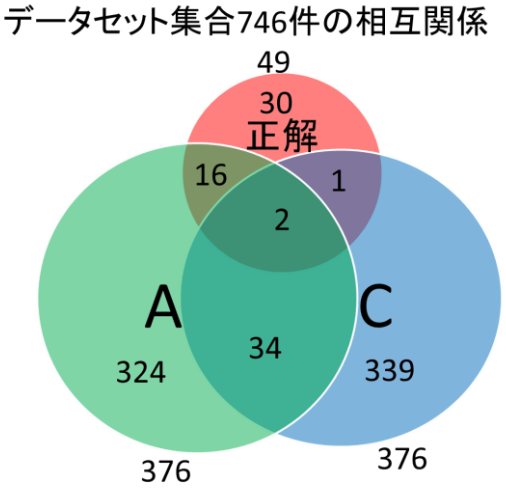

図5. データセット集合の相互関係

性格の異なるデータベース DB:A と DB:C の概念検索各々上位 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。図5にデータセット集合 746 件の相互関係を示す。

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。

図6に形態素と専門用語による分かち書きと TF、TF･IDF による重み付けの再現率への影響を示す。確認数が少ない立ち上がりでは形態素 TF･IDF が良くその後は専門用語 TF･TDF が良いが DB:A には及ばない。

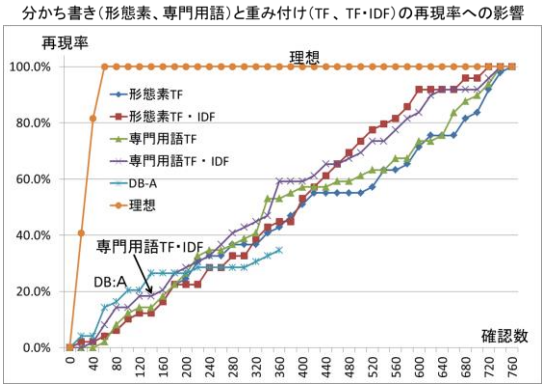

図6. 分かち書きと重み付けの影響

新規性を考慮した評価関数として検索競技大会の模範解答の構成要素分析例を参考に F タームと類似度による評価関数を設計した。図7上部の表部分は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を示している。構成要素 a（熱可塑性樹脂フィルム基材層）、要素 e（他の層を介してまたは介さずにこの順に積層）は該当する F タームが存在しない。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7のグラフである。

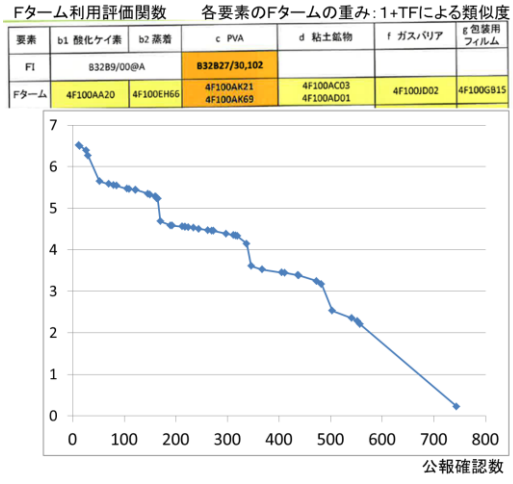

図7. F タームと類似度による評価関数

図7の評価関数を用いた再現率への影響を図8に示す。シュミレーション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。

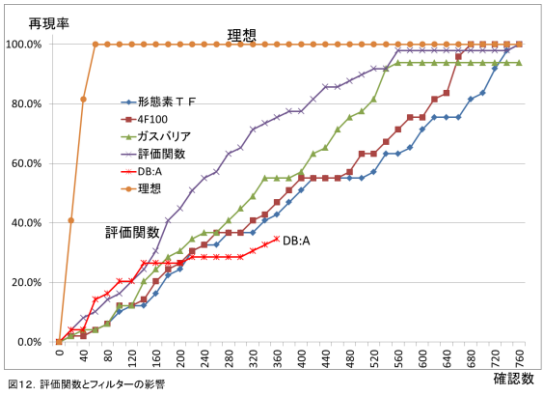

図12. 評価関数とフィルターの影響

図8. 評価関数とフィルターの影響

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検索漏れが発生している。

4－2. 分散表現によるベクトル化検討

図9に文書の分散表現ベクトルの学習モデルの再現率を示す。

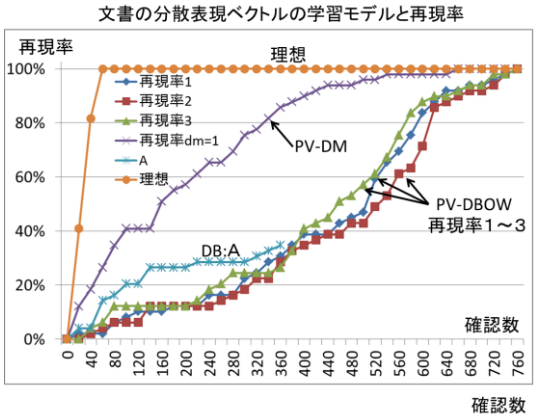

図9. 分散表現ベクトルによる再現率

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 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い。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。

4－3. 可視化検討

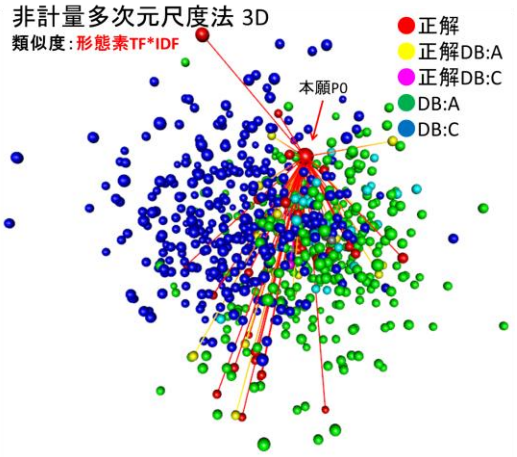

図10. One hot ベクトルによる可視化

図10に One hot ベクトルによる公報の可視化結果を示す。

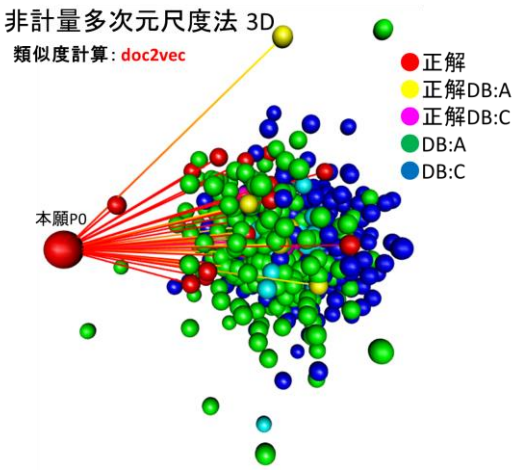

図11. 分散表現ベクトルによる可視化

図11に doc2vec を利用して各公報間の類似度から非計量多次元尺度法により可視化したマップを示す。

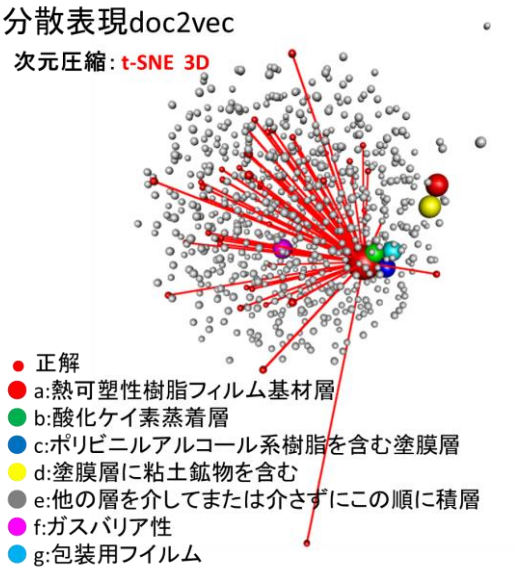

図12. 正解公報と構成要素の可視化

5. 今後の展望

本報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。更なる精度、再現率向上には教師データ有りの機械学習と組み合わせることが必須と考える。教師データ有りの機械学習としては評価関数を用いて構成要素によって重みを変える、F タームと形態素の類似度の寄与率を変える等々いろいろ考えられる。重み付けの調整や識別を利用することで改善の余地は大きいと考える。評価関数をどこまでチューニングできるか興味深い。特許調査の精度を上げるには前処理の形態素解析による「分かち書き」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。

6. 結論

単語の分散表現で文書の固定長ベクトルが得られる doc2vec の単語の出現頻度と出現順序を考慮した学習モデルを使用して公報の類似度を計算すると非常によい再現率が得られることがわかった。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。

公報の類似度計算精度向上は動向調査にも有効である。

7. おわりに

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[6)]。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。

「謝辞」

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す。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。

8. 参考文献

[1] IPA.AI 白書,KADOKAWA,2017
[2] 桐山勉, 安藤俊幸. 特許情報と人工知能(AI):総論. 情報の科学と技術. 2017, vol. 67, no. 7, p. 340-349.
[3] scikit-learn http://scikit-learn.org/stable/ accessed 2017.09.14
[4] gensim https://radimrehurek.com/gensim/ accessed 2017.09.14
[5] 岡崎直観.単語の意味をコンピュータに教える,岩波データサイエンス vol.2,p.47-61
[6] 安藤俊幸.Japio YEAR BOOK 2017 機械学習を用いた効率的な特許調査方法