分散表現学習を利用した効率的な特許調査
文書のベクトル化方法と文書分類への応用

○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 method using Distributed representations
Document Vectorization Method and Application to Document Classification

ANDO Toshiyuki [1)]
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.com

【発表概要】

ニューラルネットワークを利用した単語・文書の分散表現学習を用いて効率的な特許調査方法を検討した。特に SDI 調査を念頭に約 3000 件のインクジェット関連特許を人手で分類付与した実験用データセットを作成して文書のベクトル化方法とその用途として次元圧縮による文書の俯瞰可視化、文書分類への応用、類義語の抽出支援を検討した。

文書のベクトル化手法として OneHot ベクトルの Bag of Word(BoW)モデル、TF・IDF モデル、分散表現ベクトルのモデルとして Ave-word2vec、doc2vec、SCDV(Sparse Composite Document Vectors)[4)]、Ave-fastText、fastText-SCDV を検討した。

機械学習による文書分類の手法としてはBoostingとRandom Forestsを組み合わせて集団学習させる Python 用 XGBoost(eXtreme Gradient Boosting)パッケージを利用した。XGBoost の他に 7 種類の文書分類アルゴリズムを検討した。

各モデルを交差検証した結果 SCDV による文書ベクトルを用いて XGBoost による文書分類モデルが一番良かった。これは調査目的や調査の活用シーンに合わせて使えば十分特許調査実務に応用可能である。機械学習を用いて公報を文書分類する場合、教師データ（作成）を考慮した分類体系の設計が重要である。

【キーワード】

分散表現，doc2vec，word2vec，fastText，機械学習，文書分類，次元圧縮，特許調査，先行技術調査，特許情報解析，可視化

1. はじめに

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検討してきた[1)]。近年、word2vec のような単語の分散表現手法やそれを文書のベクトル化に拡張した doc2vec 等の有用性が注目されている。

本報では文書のベクトル化方法とそのベクトルを用いた機械学習による文書分類と特許調査への応用を検討した。文書分類の検討にはインクジェットインク特許約 3000 件に人手でカテゴリーを付与した教師データセットを作成して検討した。

2. 目的

機械学習の特許調査への応用の目的として下記の三つの目的を設定した。

①SDI 調査

予め人手で付与した社内分類等を教師データとして学習させておき定期的に発生する新規公報に対してどの程度の精度で分類できるか確認する。

②技術動向調査

文書・単語ベクトルを次元圧縮して全体像を直感的に把握し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・可視化マップを検討する。

③類義語の抽出支援ツール

注目語の類義語の抽出を支援するツールとして使用できるか検討する。日本語、英語、中国語で使用可能であることが望ましい。

3. 検討方法

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。

下記①～③に本研究で使用したデータベースと関連ツール類を記す。

① 商用特許データベース

Questel 社 Orbit.com を日本語、英語、中国語による原語検索、ファミリーデータ、英語化学物質名 ID(MLID)、英語コンセプト（テクニカルターム：KEYW）等各種データをダウンロードして使用した。

NRIサイバーパテントデスク社 CyberPatent Desk を日本特許のタイトル、要約、請求項、FI、F タームのデータソースとして csv 形式でダウンロードして使用した。

② 機械学習

機械学習のオープンソースライブラリとして scikit-learn 0.20.3[2)]、gensim3.4.0 技術[3)]、XGBoost を使用した。Python3.7 環境構築は Anaconda を使用して行った。

商用の単語の分散表現作成ツールとして NTT データ数理システムの Text Mining Studio 類義語アドオンツール 5) を試用した。

③ パテントマップ作製・解析ツール

商用のパテントマップ作製ツールとしてインパテック社のパテントマップ EXZ、特許情報の解析ツールとして Questel 社 Orbit.com のオプションの分析モジュールを使用した。

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。

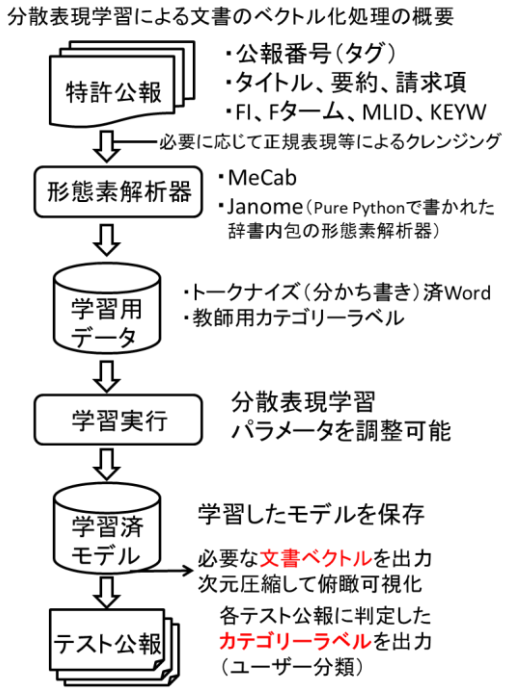

図1. 分散表現学習によるベクトル化

図1に分散表現学習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. 予備検討（目的・課題抽出）

SDI 調査、技術動向調査を念頭に Orbit でファミリー単位のデータベース FAMPAT を使用して下記検索式の検索結果 2584 ファミリーを母集団として現状の一般的な特許情報の解析手法やパテントマップ作成時の課題等を検討・抽出した。

検索母集団：(4J039GA24)/FTM AND (CN)/PN。ここで 4J039GA24 はインクジェットインクの F ターム、(CN)/PN は発行国として中国が含まれるファミリーである。結果的に F タームを使用していることで日本と中国のファミリーがある集合 2584 ファミリーが得られる。このファミリーから日本公報 3098 件を抽出し機械学習の検討用母集団とした。

Orbit の分析モジュールを使用して解析手法に対する現状の課題と自分で機械学習を利用して解析する場合の改善ポイント、目的を抽出した。図2にテクニカルドメインによる技術概要を示す。各ヘキサゴン（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。インクジェットインク関連特許は Basic materials chemistry に 2575 ファミリー、Textile and paper machines に 2210 ファミリーが一部重複して属している。全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。

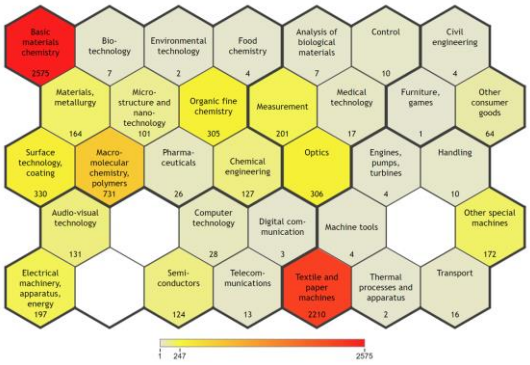

図2. テクニカルドメインによる技術概要

図3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。こ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、耐光性、耐オゾン性のような「効果」であったりして一定しないことである。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である。

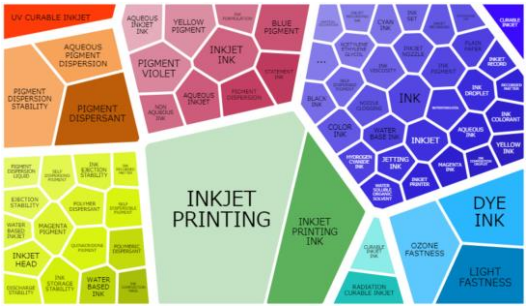

図3. コンセプトクラスター

また公報が複数のクラスターに属することを許すソフトクラスタリングになっており重複がかなり多いが表面上は分かり辛い。このツールのメリットとしては気付き(セレンディピティ)が得られる。注目したクラスターの公報リストを表示するのはインタラクティブ性に優れ非常に良い等々メリットも多い。

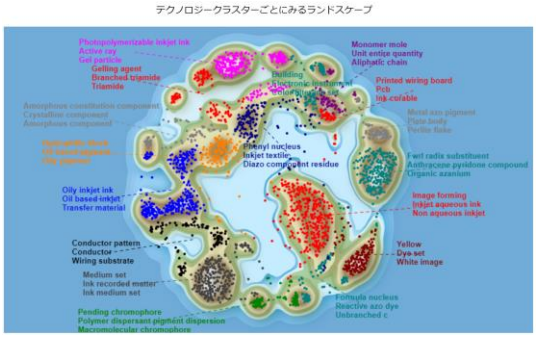

図4. テクノロジークラスター

図4にテクノロジークラスター(ランドスケープマップ)を示す。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3のコンセプトクラスターと同様のデメリットを有している。
本検討では次元圧縮による公報の俯瞰可視化と文書分類を組み合わせてクラスタリングのデメリットを解消できないか検討する。

4－2. 文書のベクトル化検討

文書のベクトル化手法として表1の7種類を検討した。

	文書ベクトル化方法	
	略号	モデル名称
①	BoW	Bag of Word
②	TF･IDF	TF･IDF
③	Ave-word2vec	average word2vec
④	doc2vec	doc2vec
⑤	SCDV	Sparse Composite Document Vectors
⑥	Ave-fastText	average fastText
⑦	SCDV-fastText	SCDV fastText

表1. 文書ベクトル化方法

BoWモデルは古典的な非常にシンプルなモデルで出現単語にIDを付け文書の各単語の有無だけを集計する。単語の出現順や頻度は考慮しないOne hotベクトルである。TF･IDFモデルは単語頻度と単語が出現する文書頻度を考慮して重み付けする。Ave-word2vecモデルは文書に含まれる単語の分散表現ベクトルの平均値を使う。doc2vecモデルはword2vecを文書に拡張したものである。SCDVはword2vecの単語ベクトルを基にidf値を計算し、GMMクラスタリングする。一つ一つの単語ベクトルが各クラスタに属する予測確率を単語ベクトルにかけて、単語ベクトル数×クラスタ数に次元を広げて、文書ベクトルとする。クラスタ数K=60とした。Ave-fastTextは、word2vecの代わりにfastTextを使用した。表1の③～⑦が分散表現による文書ベクトルモデルである。word2vec、doc2vec fastText、のベクトルの次元数(サイズ)は300、分かち書きした単語を取り込むWindow幅は5、取り込み最小単語数は1とした。doc2vecの取り込みモデルを選択するパラメータdm=1で単語の語順を考慮するモデルである。
SCDVはword2vecの300次元かけるクラスタ数K=60なので300×60=18000次元になる。公報文書の分散表現ベクトルのデータソースとしてはタ

イトル、要約、請求項とした。One hot ベクトルによる文書ベクトルとして Orbit の英語化学物質名 ID(MLID)、英語コンセプト(テクニカルターム:KEYW)、CyberPatent Desk の FI、F タームによる文書ベクトルも補助的に検討した。各文書ベクトルを用いて文書分類精度への影響、次元圧縮による各文書の俯瞰可視化マップも検討した。

4－3. 文書分類検討

機械学習による文書分類の手法として表 2 の 8 種類の分類アルゴリズムを検討した。

	略号	分類アルゴリズム
①	XGB	eXtreme Gradient Boosting
②	SGD	Stochastic Gradient Descent 確率的勾配降下法
③	GNB	naive_bayes.GaussianNB ナイーブベイズ
④	SVC	SVC(kernel='linear') サポートベクトルマシン
⑤	SVCrbf	SVC(kernel='rbf') サポートベクトルマシン
⑥	RFg	Random Forest(gini)
⑦	AdaBoost	AdaBoost
⑧	MLP	多層パーセプトロン

表 2. 文書分類手法

XGBoost は Boosting と Random Forests を組み合わせて集団学習させるもので Python 用 XGBoost パッケージを使用した。他は scikit-learn の実装を利用した。文書分類精度は XGBoost が良かった。

文書分類検討にあたり下記3種類の分類の粒度での検討を計画した。
①発明の主題レベル(筆頭 FI)
②発明の構成要素レベル(F ターム)
③明細書の文言記載レベル
以下①発明の主題について述べる。一番抽象的と考えられる大きな粒度で大分類を想定している。

表3に日本公報 3098 件の筆頭 FI ランキング上位 10 位を示す。

筆頭FI	内容	件数
C09D 11/00	インク	923
B41M 5/00 A	記録方法	175
C09D 11/30	インクジェットインク	150
C09D 11/322	顔料インク	99
C09D 11/38	非高分子添加剤	76
C09D 17/00	顔料ペースト	64
C09D 11/326	顔料分散剤	51
C09D 11/328	染料	45
G02B 5/20 101	カラーフィルター	39
C09D 11/34	ホットメルト	38

表3. 筆頭 FI ランキング上位 10 位

表4にカテゴリー別の doc2vec ベクトルモデルの XGB による分類結果を示す。

カテゴリー	件数	正解率	精度	再現率	F値
インク	926	0.66	0.51	0.66	0.57
インクジェットインク	150	0.17	0.45	0.17	0.24
顔料インク	198	0.08	0.36	0.08	0.12
顔料分散剤	51	0.06	0.23	0.06	0.09
染料	45	0.15	0.58	0.15	0.23
着色剤	17	0.19	0.25	0.19	0.21
ホットメルトインク	38	0.23	0.50	0.23	0.30
非水性溶媒	38	0.03	0.06	0.03	0.04
多色インクセット	37	0.16	0.46	0.16	0.22
非高分子添加剤	76	0.00	0.00	0.00	0.00
記録方法	239	0.40	0.58	0.40	0.47
その他	1283	0.82	0.68	0.82	0.74

表4. カテゴリー別分類結果(8 分割交差検証)

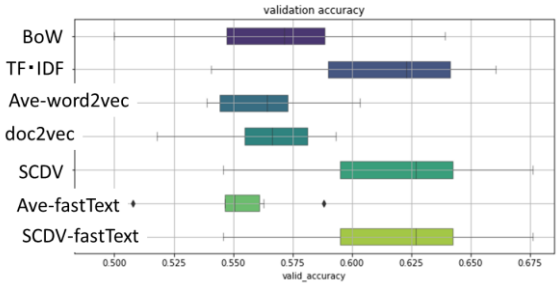

図5. 文書分類 XGB の 8 分割交差検証

図5に XGB で分類した 7 種類の文書ベクトル(縦軸)の8分割交差検証結果を示す。横軸は validation accuracy である。SCDV が良いが発明の主題に関してはあまりうまく文書分類されていない。

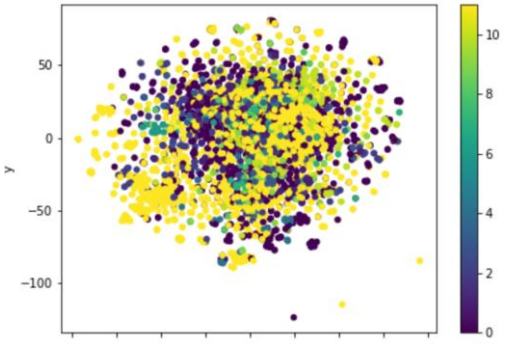

図6. BoW 文書ベクトルの次元圧縮

図6に BoW 文書ベクトルの次元圧縮結果を示す。次元圧縮は t-SNE で行った。カラーマッピングは教師データのカテゴリーを使用している。

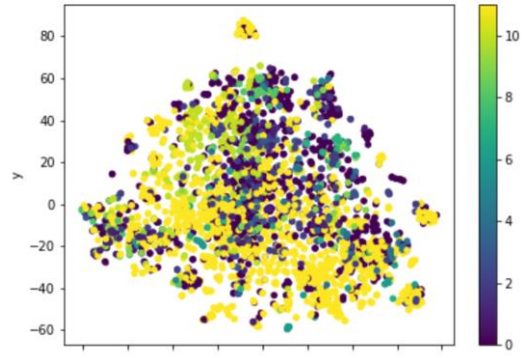

図7. SCDV 文書ベクトルの次元圧縮

図7に SCDV 文書ベクトルの次元圧縮結果を示す。BoW モデルと比べて同じカテゴリーの公報がまとまっている。
②発明の構成要素レベル（F ターム）、③明細書の文言記載レベルの文書分類については発表時に報告する。

5. 今後の展望

本報では文書の BoW、TF・IDF ベクトル、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して文書分類を検討した。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、BoW モデルに特許分類を入力等で改善の余地は大きいと考える。

6. 結論

文書の分散表現ベクトルと教師ありの文書分類を組み合わせることで SDI 調査や動向調査の効率化の可能性を示せた。文書分類に関してはパラメータチューニング、教師データの分類体系の設計等が必要である。

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。最近は機械学習を用いて効率的な特許調査に取り組んでいる。まだまだ改善の余地は大きいと考えている。今後の検討が楽しみである。

「謝辞」
本報告は2019 年の「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] 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.
[2] scikit-learn http://scikit-learn.org/stable/ accessed 2019.03.25
[3] gensim https://radimrehurek.com/gensim/ accessed 2019.03.25
[4] SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations https://arxiv.org/pdf/1612.06778.pdf
[5]Text Mining Studio 類義語アドオン https://www.msi.co.jp/tmstudio/TMSSynonymAddon.pdf