STI Horizon 2020 Vol.6 No.2
https://doi.org/10.15108/stih.00210

レポート

特許文書情報を対象としたコンテンツ分析の手法と出願人タイプ別特性比較

第 2 調査研究グループ 上席研究官 小柴 等
第 1 研究グループ 客員研究官 池内 健太、元橋 一之

概　要

本研究では文書を数値表現する手法を用い、特許を数値表現することで新たな分析の軸を導入し、さらに、それらを用いた分析を試みた。これにより、内容ベースでの分野区別が可能となったほか、本稿で紹介したように、特許間の距離が定義できたことで出願人タイプごとの異なる特徴を明らかにすることができた。提案手法は既存手法と補完的に用いることが可能であるため、本稿及び参考文献1）で述べた以外にも、様々な組合せの分析が想定できる。また、提案手法は特許や言語に限らず提要が可能であるため、今後は論文との関係性の分析や、英語で記載された特許の分析等に発展をさせていく予定である。

キーワード：特許，分散表現，高次元ベクトル近傍探索

はじめに

本稿では、人工知能関連技術（自然言語処理）を用いて特許文書（公開特許公報）の記載内容を数値表現する手法と、その結果に基づき、大学、企業、個人など出願人の種別ごとに特許の特徴を調査した結果について述べる。

具体的には、分散表現（単語埋め込み；Word Embedding）と言われる手法で特許文書を数値表現し、高次元ベクトル近傍探索と言われる手法と組み合わせて使用する。これにより、既存手法が有する技術面・分析面での課題を緩和した。結果として、従来の分析を補完する新しい分析軸を開発した。

この特許の数値表現という新しい分析軸を用いることで、任意の特許間の距離（類似度）を容易かつ定量的に得ることができる。分析の結果からは、個人の申請する特許は相対的に他の特許との類似度が低いこと、企業の申請する特許は内容的に類似する特許が多いこと、などが分かった。また、大学や産学連携特許は両者の中間に位置すること、などが分かった。

以下ではこれら手法及び分析結果の概要について紹介する。なお詳細は参考文献1）を参照されたい。

背景

既存手法等

科学技術・学術活動の結果として生み出される具体的でわかりやすい成果物としては論文や特許を挙げることができ、科学技術・学術政策に資する資料の作成等を目的として、これら論文・特許データを用いた様々な計量書誌学的分析がなされている[2~13]。

計量書誌学的分析の手法には、単に（分野別、期間別、機関別などの何らかの種別やキーワードの有無、それらの組合せなど）当該文献に静的に紐付く何らかの属性を軸として数を数えるなどの分析を行うもの（単純集計分析）、引用・被引用など他の文献との関係によって動的に変化する属性を軸として分析を行うもの、などが挙げられる[14]。被引用数の多さは当該文献の影響力と読み替えることもでき、文献の質の高さを表す指標として用いられることもある。

「他の文献との関係」については、既に述べた引用関係を用いるものと、内容の類似度を用いるものに大別できる。論文・特許においては引用文献を明記する風習があるため、（表記ゆれなどの課題はあるものの）引用は関係性を明確に規定できる点でメリットがあり、これを用いた分析も多数なされている。内容の類似度を用いるものは「類似度」の設定方法などに難が

あるほか、類似度を計算するコストの点、関係性は飽くまで「類似度」によって推定されるものである点、などにも課題を有する。その一方、引用関係はないが極めて類似する文献を見つけることができたり、「分野」などのあらかじめ決まった分類をまたいだ関係性を見いだすことができたり、引用関係とは異なる関係性を見いだせる可能性があり、内容の類似度に基づく分析も種々なされている[9、10)、15、16)]。

これらの分析は特に対立するものではなく、目的やそれぞれの手法の特徴に照らしてうまく組み合わせることで、有用な知見を得ることができる。

類似度を用いた分析の課題

ただし、内容の類似度を用いるものは既に述べた種々の課題も有している。例えば類似度の定義について、Jaccard 係数を用いるもの[10)]、cos 類似度を用いるもの[9)]などがある。さらに、先に挙げた参考文献 9）では、米国特許約 530 万件特許間の距離を計算するためにクラウド環境を用いて数週間の期間をかけ、計算結果は約 300TB のデータになったとしている。

類似度の定義に関する課題

Jaccard 係数は類似度の指標として代表的なもののひとつである。Jaccard 係数は文章中の単語をリストアップした上で 2 つの文章中で重複する単語の割合を計算することで求められる。割合であるため重複が全くなければ 0、全て一致すると 1 となり、類似度の指標として好ましい。ただし、単語の重複率であるので頻度を考慮していない点には課題がある。例えば、文書 1 には、単語 A が 100 回、B、C が 1 回出現していたとする。一方、文書 2 には A が 1 回、B が 100 回、C が 1000 回出現していたとする。この場合、Jaccard 係数では両文書ともに A、B、C の単語で構成されるので類似度は 100% となる。

他方、Jaccard 係数と並ぶ代表的な手法として cos 類似度も存在する。Jaccard 係数と比較すると、cos 類似度は単語の重複だけでなく頻度も考慮するような指標と言える。ここでは、単語それぞれを次元、出現頻度を大きさと捉えて文書をベクトルで表現する。その上でベクトルの内積（cos）を類似度とする。単語の出現頻度は 0 を含む自然数であるためベクトルの内積は単語の重複が一切なければ 0、単語の頻度パタンも含めて全て一致すると 1 となり、類似度の指標として好ましい。

単語の定義に関する課題

ただし、Jaccard 係数、cos 類似度とも「単語」の関係については考慮していない点に課題もある。例えば「みかん」と「ミカン」は表記が異なるため別の単語と見なされるが、その意味するところは極めて近いと想定される。「みかん」と「オレンジ」「柑橘（かんきつ）類」なども近そうに思える。一方で「みかん」と「遊星歯車列」「再生核ヒルベルト空間」などは関係が無さそう（遠そう）に思える。しかしながら、上記の手法ではそれらは一切考慮されない。その結果、人間にとっては似たような意味内容を有するが単語の重複がないような文章、例えば「細君のバースデーにケーキを購入して帰宅した」「妻の誕生日に “いちごショート” を買って帰った」という 2 つの文章の類似度は 0 となる。

こうした課題に対応するためには、例えば「みかん」に対して「ミカン」は類似度 98%、「オレンジ」は 96%、「柑橘（かんきつ）類」は 80%など、単語同士の類似度を定義する必要がある。そのために、シソーラス（類似語辞書）やオントロジーを用いる方法も考えられるが、昨今では「分散表現」（単語埋め込み ; Word Embedding）という方法も開発・利用が進んでいる。

分散表現にも幾つか手法があるが、大まかなイメージとしては大量の文章を与え、ある単語とセットで出てくる確率の高い単語は関係が強い、と、学習させるようなものと言える。技術的には深層学習やそのベースとなるニューラルネットワークを用い、特定の単語を任意の次元のベクトルとして出力する。この際、関連が大きいものほど座標値が近くなる。ここで cos 類似度ではベクトルを取り扱っていた。したがって、分散表現を用いることで先ほどの「みかん」「ミカン」問題を回避した、類似度の算出が可能となる。

こうした背景から、単語の分散表現をベースに文章の分散表現を獲得する手法[17)]や、文章の分散表現間で cos 類似度を求めて分析した結果なども報告がなされている[18)]。

計算コストに関する課題

ここまでで、「みかん」「ミカン」問題を回避し、単語の持つ意味も多少考慮したような形で類似度の算出が行えることになった。しかしながら、類似度を用いた分析には他にも計算コストの問題を有している。類似度は 2 つの文書について計算するものである。当然ながら計算するまで類似度は分からないため、基本的には全ての文書の組合せについて計算する必要があり、計算量は $0(n^2)$ となる。例えば、10 件の文書があったとき、その組合せは $(10\times9)/2=45$ 件、100 件では $(100\times99)/2=4950$ 件、1000 件では $(1000\times999)/2=499500$ 件、となる。つまり、対象がたかだか 1000 件であってもその類似度は約 50 万の組合せを計算することになり計算量は膨大である。また、

計算結果を保持する必要もあるため大量のストレージも必要となる。結果、例えば参考文献 9）では、米国特許約 530 万件特許間の距離を計算するためにクラウド環境を用いて数週間の期間を要し、計算結果は約 300TB のデータになったとしている。

ところで、分散表現を用いると各文書を任意の次元のベクトルとして表現できるのであった。ある程度の次元数のベクトルの場合、「高次元ベクトル近傍探索」[19),20)]と言った手法を用いることで、近傍のベクトルのみを取得することができる。正規化されたベクトルの場合、ベクトルの距離は類似度に反比例するため、近傍のベクトルのみ取得できれば、もともと類似度が 0 に近いようなものを計算する手間を省くことができる。そのため、大幅に計算コストを削減することが可能となる。また、「高次元ベクトル近傍探索」の手法によっては数百次元、数十万件規模のベクトルの中から数ミリ秒の単位で取得できるため、必要に応じて近傍ベクトルを探索し、都度類似度を計算すれば良い。したがって、ストレージのコストについても大幅に低減できる。ただし、「高次元ベクトル近傍探索」は近似手法のため 100% の精度を得られない点に注意が必要である。

このほか、各文書をベクトルで表現した場合、多次元尺度法などの次元圧縮を行うことで、2 次元空間での可視化なども実現し、分析結果の解釈などが容易になる。実際にこれらの手法を用いた分析もなされている[18)]。

これらを勘案して、今回は公開特許公報を対象に、分散表現を用いた内容の分析を行うことにした。

分析の手続

分析の大まかな手順・手続をまとめると以下の通りとなる。

1. 特許データから、タイトル及び概要文を抽出する
2. 形態素解析器にかけ、名詞句のみ抽出する
3. 上記、2. のデータに基づき単語の分散表現を獲得する
4. 獲得した単語分散表現を用い、各特許の分散表現を獲得する
5. 場合により、特許分散表現をもとにクラスタリングを行う
6. 場合により、特許分散表現をもとに次元圧縮を行い 2 次元で可視化する
7. 特許分散表現に対して高次元ベクトル近傍探索用のグラフインデクスを作成しておくことで、任意の特許データに類似する特許データを高速に取得する

提案手法に基づく分析結果

公開特許公報など、特許文書に対して類似度若しくは距離を定義できることの利点は既に述べたとおり、引用関係にない文書との関係を推定できること、また、その関係性の強さを推定できることにある。これにより、引用関係等の既存の分析に加えて、新たな分析の分類を実施することが可能になる。

そこで本試行では、単語分散表現や文書分散表現（特許分散表現）の作成と妥当性の確認、既存の分野割りである IPC 分類と分類結果の対応の調査等を行った上で、まず距離の分布について分析し、どの程度の距離があれば似ていると言えそうかを明らかにした。その上で、大学、企業など組織ごとに距離の分布にどのような特徴があるのかを調べた。

それらの結果について以下で紹介する。

データの詳細

本実験では、特許庁が提供・公開している「公開特許公報」のデータを用いた。

データの期間は公開日ベースで 2005 年 1 月から 2019 年 4 月末日まで、種別としては「A」公開特許公報、公表特許公報及び「S」再公表特許に分類されているものを対象とした。

結果、対象となる公開特許公報の件数は 4,069,503 件となっている。

次に、これらの公開特許公報データ（以後、単に特許データという）の概要文について「【課題】」や「【解決手段】」などのラベル文字をルールベースで削除した上で、特許データのタイトル、概要文を分析対象に設定した。

分散表現の獲得結果

まず、特許分散表現についてクラスタリング手法（K-Means++）[21)]で 16 分類した上で、次元圧縮手法（UMAP）[22)]を用いて 2 次元に可視化したものを図表 1 に示す。

図表 1 の点のひとつひとつが特許に対応し、点の色が 16 分類に対応する。また図上の文字は 16 分類ごとの頻出単語に基づいて、著者らが設定した 16 分野それぞれに対応するラベルである。詳細については参考文献 1）を参照されたい。

図表 1 を見ると、「車両」や「印刷」「金属」のように複数のエリアにスプリットしているクラスタもあるものの、多くは 2 次元に圧縮した状態でも近くに配置されている。また、「情報」の近くに「画像」や「電子回路」が、「化合物」の近くに「バイオ」や「樹脂・膜」など関連が強いと思われるものが近くに

図表 1　特許分散表現の 2 次元表現

配置されている。他にも例えば複数にスプリットしている車両についても、制御系に関しては「電子回路」と、燃料制御や空力特性などは「流体」と、駆動系は「モータ」と関連が近いと考えられ、全体としてある程度妥当と思われる結果が得られている。

ただし、クラスタの名付けは主観的に行われており必ずしも正しく意味内容が反映・表現されているとは限らない。飽くまで印象の範囲にとどまっている点に注意を要する。

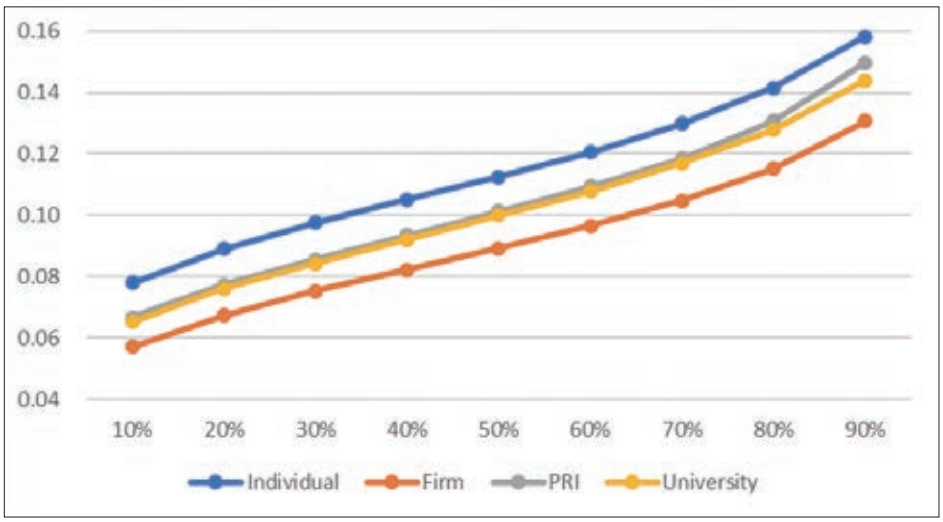
近傍 200 特許を用いた出願人タイプ別の分析結果

本章では、近傍 200 件の特許のデータを用いて、出願人タイプ別の特許の特徴について分析を行う。

図表 2 は出願人タイプ（個人出願人（Individual）、企業（Firm）、公的研究機関（PRI）及び大学（University））の違いによる 200 番目の特許との距離の分布を見たものである。全体として、個人出願人の距離が最も大きく、公的研究機関と大学がその次でほぼ同様の値、企業における距離が最も小さくなった。企業における出願特許は特定の技術スペースに集中しているのに対して、個人出願人はよりスパースな技術スペースに出願する傾向がある（公的研究機関・大学はその中間）ことを示している。

さらに、各特許を個人（IND）、企業（COM）、公

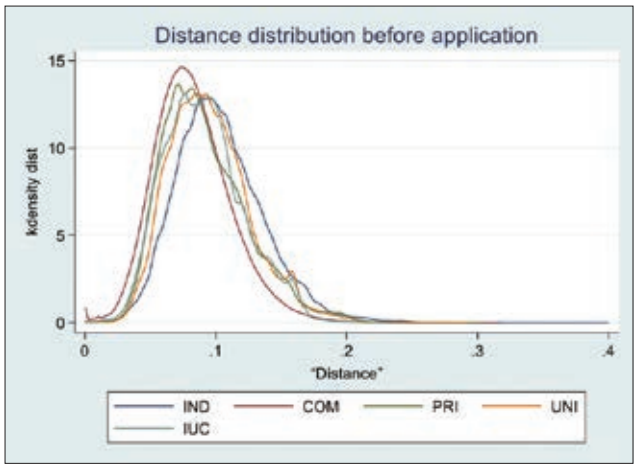
図表 2　出願人タイプ別の違い



的研究機関（PRI）、大学（UNI）に加えて産学連携（IUC）の５つのタイプに分類し、出願人のタイプ別に近傍特許との距離の分布を比較した。なお、各特許の出願日より前の５年以内に出願された特許との距離と出願日の後の５年以内に出願された特許との距離を比較する。そのため、比較の基準の特許は 2010 年に出願された特許に限定した。

図表３は出願人タイプ別の出願前の５年以内の近傍特許との距離の分布を示している。企業の特許の距離の分布は左側に寄っており、企業は出願時点で類似した特許が既に多く存在するスペースに特許を出願する傾向がある。一方、個人の特許の距離の分布は比較的右側に寄っており、個人発明家は類似した特許が比較的少ないスペースに特許を出願する傾向がある。他方、大学や公的研究機関、産学連携特許は企業と個人の中間に位置している。

図表３　出願人タイプ別の出願前５年以内の近傍 200 特許との距離の分布



おわりに

本研究では文書を数値表現する手法を用い、特許を数値表現することで新たな分析の軸を導入し、さらに、それらを用いた分析を試みた。

これにより、内容ベースでの分野区別が可能となったほか、本稿で紹介したように、特許間の距離が定義できたことで出願人タイプごとの異なる特徴を明らかにすることができた。

提案手法は既存手法と補完的に用いることが可能であるため、本稿及び参考文献 1）で述べた以外にも、様々な組合せの分析が想定できる。また、提案手法は特許や言語に限らず提要が可能であるため、今後は論文との関係性の分析や、英語で記載された特許の分析等に発展をさせていく予定である。

参考文献・資料

1) 元橋一之, 小柴等, 池内健太. 特許文書情報を用いた発明内容の抽出と出願人タイプ別特性比較. Discussion Paper, No. 175, December 2019.
2) 富澤宏之, 林隆之, 山下泰弘, 近藤正幸. 有力特許に引用された科学論文の計量書誌学的分析. 情報管理, Vol. 49, No. 1, pp. 2–10, 2006.
3) 科学技術政策研究所科学技術動向研究センター . サイエンスマップ 2004. NISTEP REPORT, No. 100, March 2007.
4) 科学技術基盤調査研究室. サイエンスマップ 2006. NISTEP REPORT, No. 110, June 2008.
5) 科学技術基盤調査研究室. サイエンスマップ 2008. NISTEP REPORT, No. 139, May 2010.
6) 科学技術・学術基盤調査研究室. サイエンスマップ 2010&2012. NISTEP REPORT, No. 159, July 2014.
7) 科学技術・学術基盤調査研究室. サイエンスマップ 2014. NISTEP REPORT, No. 169, September 2016.
8) 文部科学省科学技術・学術政策研究所. サイエンスマップ 2016. NISTEP REPORT, No. 178, October 2018.
9) Kenneth A. Younge and Jeffrey M. Kuhn. Patent-to-patent similarity: A vector space model. SSRN, 07 2016.
10) Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. Text matching to measure patent similarity. Strategic Management Journal, Vol. 39, 08 2017.
11) 開本亮, 難波英嗣. 学術論文への国際特許分類 (IPC) 付与による産学連携の検討 : サブクラス分析とメイングループ 分析. 研究・イノベーション学会 年次学術大会講演要旨集, Vol. 32, pp. 336–337, 2017.
12) 開本亮, 難波英嗣. 学術論文への国際特許分類 (IPC) 付与による産学連携の検討 : IPC 分類と JST 分類の共用分析. 研究・イノベーション学会 年次学術大会講演要旨集, Vol. 33, pp. 177–180, 2018.
13) 元橋一之. AI におけるサイエンスとイノベーションの共起化: 米国における論文・特許データベースを用いた分析. DISCUSSION PAPER, No. 160, 2018.
14) 調ほか. 研究評価・科学論のための科学計量学入門, 丸善, 03 2004

15) 佐藤貢司, 安井基陽, 田中厚子, 中村昭博, 中田守. 被引用情報を用いた重要特許抽出方法の検証:. 情報プロフェッショナルシンポジウム予稿集, Vol. 2017, pp. 61 – 65, 2017.
16) 榑松理樹. 特許構成を考慮した文書類似度に基づく特許からの課題分類・手段分類推定システム . 人工知能学会全国 大会論文集, Vol. JSAI2014, pp. 1A32 – 1A32, 2014.
17) Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. arXiv preprint, 2018. arXiv:1805.09843.
18) 小柴等, 森川想. 議事録を用いた我が国における議会・行政の関係性分析手法. 人工知能学会論文誌, Vol. 34, No. 5, pp. E–J47_1 – 10, 2019.
19) 岩崎雅二郎. 商品画像検索へのグラフ構造型インデックスの適用. 画像電子学会誌, Vol. 42, No. 5, pp. 633–641, 2013.
20) Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki. On approximately searching for similar word embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.2265–2275, Berlin, Germany, August 2016. Association for Computational Linguistics.
21) David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, pp. 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.
22) Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv preprint, 2018. arXiv:1802.03426.