１Ｃ１７
言語モデルを用いた特許抄録による研究開発トレンドの調査方法と開発技術の定量評価の検討

○黒田夢子，鈴木潤，隅蔵康一（政策研究大学院大学）
doc25051@grips.ac.jp

1. はじめに

文章をベクトルに変換する Transformer 型の言語モデルは，近年軽量で高精度な埋め込み専用モデルが利用可能となっている．そこで，比較的狭い技術分野の特許を対象として，技術内容を定量的に調査する方法を検討した．特許抄録をベクトル化し，k-means 法でクラスタリングしたところ，クラスターの重心に近いキーワードから各クラスターの技術的特徴を把握できた．さらに，クラスターの重心ベクトル間のコサイン類似度を用いて位置関係を整理すると，①要素技術，②具体的な課題を解決する技術，③製品に直接かかわる技術，の三種類に分類できた．出願件数の推移のピークを比較すると①は 2019 年から②では 2020 年，③では 2021 年に初めて現れ，応用的なクラスターほどピークが遅れる傾向が認められた．個別の特許については，文章ベクトル間のコサイン類似度を利用することで，既存特許や論文との技術的類似性を分析できることを確認した．

2. 研究の背景・先行研究

特許は，規格や表現内容が揃った文章であるために，自然言語処理の進展とともに，特許審査や産業調査を目的とした自動化のタスクが研究対象となってきた．近年の大規模言語モデルの登場に伴い，特許文献の文章の分類で，IPC サブクラスに対する分類記号の予測精度の向上が報告されるようになった．特許の分類タスクは，申請された特許の新規性の可否と分類記号の付与に分けられる．特許審査官が行う先行技術調査ために，言語モデルにより与えられる特許同士の類似度が利用されている．関連特許の抽出を行うタスクでは，言語モデルの選定や調整方法で様々な手法が検討されている[1][2]．言語モデルは，特許の類似性をある程度正しく捉えていると考えられており，産業調査においては，研究開発内容の可視化が試みられている[3]．本章では，特許データを対象とした，ディープラーニングにより進展した大規模言語モデルの分類タスクやモデルの評価方法を，精度に着目して紹介する．

AI が自動化できるタスクは問題設定に制約があり，正解のあるデータセットにより学習を行う場合が多い．特許文献は，抄録や請求項と分類記号で構造化されたデータセットであり，それらが発明の内容を補完的に示している．発明者，出願人や特許と学術論文の引用情報も利用可能であり，分類記号や引用関係などを正解として，様々なデータ項目の組み合わせが検討されている．近年，単語の並び順や文のセクションを取り込んだベクトル化を行う Transformer 型の言語モデルが実用化された．ベクトル化を埋め込みと呼び，文章から変換したベクトルは，人間が受け取る意味が近くなるよう配置される．この特性は，特許の申請時に近い特許が既に出願されていないかを調べる先行技術調査に利用できると考えられており，精度の向上が研究のスコープである．

特許の分類は，特許が採択されるかを予測するための新規性の可否や，IPC や CPC 分類記号を予測する．これらのタスクでは，利用する言語モデルの埋め込みが特許の意味を必要な形で与えられるかが分類精度向上の鍵である．類似する特許を取得するタスクでは，正規化された二つの文章ベクトルの内積で求めるコサイン類似度を利用する[4]．これは二つのベクトルのなす角のコサインで文章の類似性を示す．人間が内容を読んだり単語の検索を行うことなく，類似する特許を見つけることが可能だが，言語モデルの精度に依存する．Ascione(2024)[2]では，事前学習済みモデル SBERT に対し，最も詳細な技術分類である CPC 分類記号を用いたファインチューニングを行い，精度を向上させたと言う．CPC 分類記号は，国別に異なる分類記号を橋渡しするために EPO と USPTO が提案した特許分類であり，一番詳細な技術分類である．この研究では，審査官引用の特許の組み合わせと，調整済み言語モデルのコサイン類似度が最も大きい特許の組み合わせが 60%一致した．

大規模言語モデルを用いた分類記号を推定するタスクでは，最高性能が更新されている．Roudsari(2022)[5]では，IPC サブクラスの分類で XL NET が Precision:0.82，LRAP:0.807 となり，それまでの単語の埋め込みを SVM により分類する方法の性能を上回った.Chikkamath(2023)[6]では,USPTO の特許データで事前学習を行なった BERT-for-Patents で，CPC および IPC のサブクラスの分類を行い micro-F170%を達成した．Bakamiri2024 は，BERT-for-Patents を利用した拡張データセットで追加学習した埋め込み用モデルの PateteBERTa を用いて CPC サブクラスのマルチラベル分類で accuracy=54%，F1>66%となり，その時点での SOTA(State of the art)を塗り替えたと言う[7]．言語モデルの進展は急速であり，新しい開発モデルが利用できる．

先行技術調査は，特許審査官のみならず，企業の研究開発担当者にもニーズがある．必要な類似特許の検索には，専門知識に精通した担当者が必要な技術用語を適切な検索式で検索することが必要である[8]．類似する特許の検索もれは，特許を申請してから類似特許が見つかり拒絶を受ける理由となりうる．本研究では，k-means 法により特許クラスタリングを行なった．文章ベクトルにより意味の近い文章をグループ化し，より広い検索範囲の特許を文書ベクトルのクラスタリングにより分割することは，先行技術調査の効率化に寄与する．また，既存の分類より詳細な技術概要に沿ったクラスターが得られるため，その分野の研究成果を深く知ることができる[9]．

特許と論文の研究開発の近接関係を分析する研究では，TF-IDF を用いた単語ベクトルの埋め込み方法により[10]特許と論文を一まとめにして分析されるが，言語モデルによる埋め込みでは，特許と論文で文章表現の違い[11]が文章ベクトルに現れるため別々に扱う．TF-IDF では，汎用的な表現に用いられる単語は値が小さくなり，主に技術や知識の内容を代表する単語が相対的に高い値を持つベクトルとなるため，特許と論文を一律に比較できると判断されている．言語モデルによる埋め込みでは，単語の前後関係や文脈を取り込むのでベクトルの比較は文章の意味の比較となる．本研究では，既存特許・論文と既存の文献との近接性が，企業か大学かの出願セクターで異なるかをコサイン類似度を用いて検証した．

3. 研究対象

詳細な分析に対する調査範囲として，Simultaneous Localization and Mapping(SLAM)を選定した．SLAM は，自己位置推定と地図作成を同時に行うアルゴリズム群であり，モビリティの環境認識技術として一般的な解決方法である[12]．センサから物体までの距離情報を取得する LiDER や，デプスカメラ等のセンサの開発とともに，技術が蓄積されてきた経緯がある．ロボットやドローンの自立走行や，自動運転における位置や周辺環境の「認識」の実現に対して汎用的な要素技術であり，高度な専門知識が要求される．また，産業上の必要性から学術研究と企業の技術開発が拮抗していることが予想される研究分野である．研究開発方法に一様さがあり開発組織の規模が同程度で，プロジェクト単位の研究開発のマネジメント方法が同等と考えられる技術範囲であり，後の定量的な評価に対しフィットする．

4. 手法

(1) 特許論文の書誌データの取得とデータセットの作成

“SLAM”と “Simultaneous Localization and Mapping”をキーワードとして検索を行い，コンピュータサイエンス分野のプレプリントを公開する arXiv，欧州特許庁 (European Patent Office) の特許書誌データベースの PATSTAT で,それぞれ論文と特許を取得した.抽出方法を以下に示す．論文は arXive の公開する API で抄録に“SLAM”か “Simultaneous localization and Mapping”を含むと言う条件でリクエストし，2000 件がヒットした．タイトル，抄録，出版年，著者を取得した．

特許は,PATSTATonline で検索を行い，“SLAM”を抄録に含む特許と言う条件で 17923 件が得られた．“SLAM”という略語は 4 文字の単純な文字列であり，別分野の略語や他の単語の一部として含まれる．Simultaneous Localization and Mapping を意図した特許だけに絞り込むため，一つ一つの特許の記載事項を検討する必要がある[13]．そこで，抄録に① “Simultaneous Localization and Mapping”が含まれれば採択した．“Simultaneous Localization and Mapping”を示す単語を抄録の文章から特定し，②単語の中に “SLAM”と言う文字列を含む 1 語(86 語)，“SLAM” を含む 2 語 (178 語) が含まれる特許を採択した．③ “SLAM”を含む特許から，別分野に関係する単語(30 語)が含まれるものを除外した．抄録の言語(appln_abstract_lg)が英語(en)の 2470 件を抽出した．

PATSTATonline は，各国特許をカバーし収録数が多いため，多数のデータテーブルに分けられ ID で結合する正規化されたデータベースである．出願人のセクターを特定するため，tls206_person，

tls207_pers_appln の psn_name，person_name_orig_lg を調査し，企業(I)，大学(U)，その他に分類した．また，一つの特許 ID に複数の出願人が含まれ，二つのセクターの出願人を持つ特許は，産学協同出願(UI)と分類した．出願セクターの特定は，Grid(2010)[14]を参考とした．

(2) 文章ベクトルのクラスタリングを利用した技術トレンド調査

“intfloat/multilingual-e5-large-instruct”により，取得した論文 2000 件，特許 2470 件をベクトルに変換した．初めにモデルの解釈性能を確かめるため，k-means 法で２つに分類し，特許か論文かを正解とした混同行列を図 2 に示す．言語モデルは，特許と論文を異なる文章群と見分けた．一つのクラスターが 100 件程度のグループになるよう，論文と特許を別々に 20 個のクラスターに分類した．特許のクラスタリングの結果を紹介する．クラスターの技術概要を推定するため，単語の TF-IDF で重み付けした文章ベクトルの総和を求め，各クラスターの重心に近い 30 語を選定した．30 個のキーワードから chatGPT5.0 によりクラスターの技術概要を推定した．

表 2 特許と論文を合わせた k-means の結果(k=2)

	クラスタ 0	クラスタ 1	合計
特許	2448	22	2470
論文	0	2000	2000

20 個のクラスターの重心同士のコサイン類似度を算出し，重心の最も近しいクラスターが隣り合うよう並べることで，SLAM 特許の技術要素を分析した．

(3) コサイン類似度による特許・論文抄録の近接性の分析

2470 件の特許と 2470 件の特許の総当たりと，2470 件の特許と 2000 件の論文の全ての組み合わせに対し，文章ベクトル同士のコサイン類似度を求め，既出の特許と論文の中でコサイン類似度の最も大きいものを一番近い特許，一番近い論文とした．最新の言語モデルの埋め込みでは，ベクトル同士の近さが，おおよそ文章の意味の近さを表している．求められた一番近い特許，一番近い論文とのコサイン類似度は，出願時に既に発表された特許，論文との記載内容の違いの程度を示している．初めに各特許とコサイン類似度が近い 20 件の特許，論文を選び，出願年，発表年が 1 年以上古いものうち，コサイン類似度が一番大きいものを選出した．20 件以内に既出の特許，論文がない特許は，近傍 20 件の中で最初にその技術や知識を発表した特許となる．今回はそれらに対しては，分析を行っていない．

企業，大学，産学協同の出願人のセクターに分け，コサイン類似度に違いがあるか分析を行なった．順に，Wilcoxon 順位和検定，一番近い特許・一番近い論文とのコサイン類似度とその差の経験累積確率分布関数の作成，分布の違いを調べるコルモゴロフ・スミルノフ検定である．出願セクターとは無関係に，特許，論文の出願年や発表年が離れているほど，時間の経過により技術が進展している可能性がある．コサイン類似度と出願年差の関係をスピアマンの相関検定で確認した．

5. 結果と考察

(1) 研究クラスタリング

2470 件の特許を 20 クラスターに分類し，キーワードから技術概要の推定を行なった（図 1）．

図 1　SLAM 特許のクラスターのタイトルと出願数が最大となる出版年

クラスターの出願数がピークの年にそのクラスターのタイトルを配置した．技術内容が近いクラスターほど，クラスターの重心同士のコサイン類似度が大きくなるため，各クラスターでコサイン類似度が一番近くなるクラスターを求め，隣接関係を確かめた．コサイン類似度の大きいクラスターが近くなるように調整すると，図 2 に示すように，①SLAM の要素技術，②実装上の応用に対応した SLAM，③SLAM が実装された製品の特許にまとめることができた．①では 2019 年，②では 2020 年，③では 2021 年に最初のクラスターが現れ，応用度が高いほど出願ピークが後になる傾向が認められた．

(2) 一番近い特許・一番近い論文とのコサイン類似度とその差

分析対象の出願セクターごとの件数を表 3 に示す．コサイン類似度が高い 20 件の特許の中にその特許より古い特許があるものは 2262 件であり，残りの 208 件は近傍の 20 件の中では最も先に出願した特許であると言える．論文では，2193 件が見つかり，近傍の 20 件の論文がすべてその特許より後の出版の論文であるケースが 207 件であった．この分析では，コサイン類似度を用いるため，一番近い特許とのコサイン類似度 2262 件，論文とのコサイン類似度 2193 件を分析した ．

表 3 SLAM 特許の出願セクターごとの件数(N=2470)

大学	産学共同出願	企業	その他	データなし
1206	80	1086	74	27

一番近い特許，一番近い論文とのコサイン類似度に対しパーセントランクを算出して，Wilcoxon 順位和検定を行なった．表 4 に中央値を示す．出願人が大学の特許と出願人が企業の特許は，中央値=0.5 が棄却された．論文でも特許でも，大学は高い方のランキングに偏り，企業は低い方のランキングに偏っていることが有意に確かめられた．

表 4 パーセントランクの中央値

	一番近い特許		一番近い論文	
	パーセントランクの中央値	p value	パーセントランクの中央値	p value
大学	0.576736	2.10E-09	0.6236314	2.20E-16
産学協同出願	0.4692614	0.9433	0.5050182	0.7436
企業	0.4343211	7.51E-09	0.3845803	2.20E-16

コサイン類似度の分布を比較するため，出願年が 1 年以上前の特許の中で一番近いものとのコサイン類似度の経験累積確率分布関数を作成した（図２）．

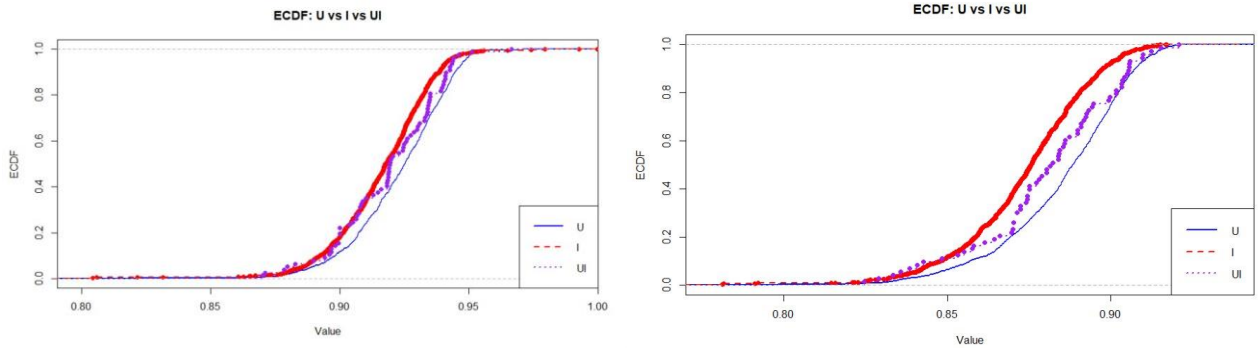

図 2　一番近い特許（左）一番近い論文（右）とのコサイン類似度の累積確率分布関数

大学，企業，共同出願を比較すると，企業の特許が最もコサイン類似度が低い．このことには二つの可能性がある、一つ目として，審査への対策として独自性や新規性が認められるよう，表現の工夫がされている，その程度が確実な成果を求める企業の方が高いことが考えられる．二つ目は，実際の技術内容も，より審査を通るよう独自性が高い発明に至るような努力がされている可能性である．また，大学

特許が相対的にコサイン類似度が大きいことは，既出の技術や知識から演繹的に発明を行っている，技術の実現可能性を精緻に確認していくため既出の特許や論文と近い発明となっていると考える事ができる．いずれとしても，その特許が求める請求の範囲が，コサイン類似度が小さい方が広いと考えられる．また，特許が論文を引用し，その論文との間に特許の新規性の境界が現れる場合があるが，論文の引用は稀である[15]．今回求めた一番近い論文は，おおむね引用された論文ではない．一番近い論文とのコサイン類似度が小さいことは，学術研究の成果を取り込んでいることを示していると考えられる．

一番近い論文，一番近い特許ともに，企業の特許が大学特許よりコサイン類似度が低い傾向が見られた．違いが統計的に有意であるかを，コルモゴロフ・スミルノフ検定により確かめた（表 4）．大学と企業では差が有意である．共同出願の特許は一番近い特許では企業と差が小さく，論文では大学との差が小さい．ただし，共同出願特許のサンプル数は限られており，p 値は参考値として解釈する必要がある．

表 4 一番近い特許と一番近い論文とのコサイン類似度のセクターごとの分布の差の検定結果

	一番近い特許		一番近い論文	
	Max Distance	p value	Max Distance	p value
U-I	0.16708	3.81E-13	0.27699	2.20E-16
UI-U	0.13233	0.1636	0.14886	0.09633
UI-I	0.11554	0.2915	0.17281	0.03489

特許と論文では，文章表現の違いが顕著なため，特許×特許と特許×論文のコサイン類似度は，すべての特許で後者の方が低い．そこで，それぞれの特許で一番近い論文と特許とのコサイン類似度の差をとり，出願セクター別に比較を行った．図 3 の企業，共同出願，大学の中央値は，0.042，0.039，0.038 である．企業は高く，大学は低く，産学共同出願はその間であることが分かる．一番近い論文との差が小さく，一番近い特許との差が大きい時，この差は小さくなる．値が小さい特許は，論文で発表される学術研究の内容に近く，一番近い特許からは遠い．そのため，値が小さい特許の中には，学術からの知見を吸収した革新的な特許が存在すると考えられる．

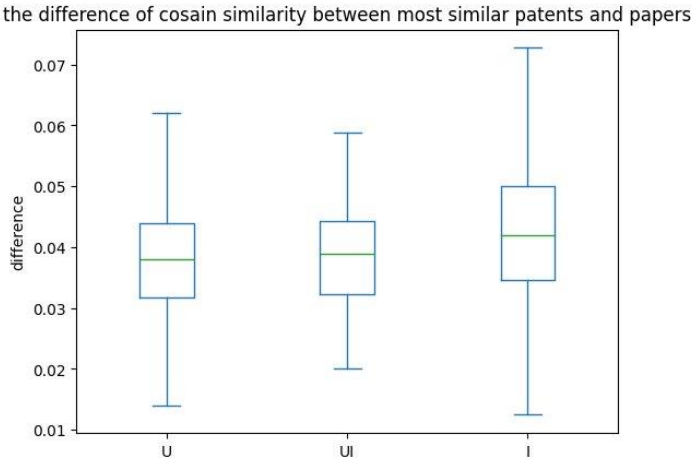

図 3 出願セクターごとの一番近い論文・特許とのコサイン類似度の差（外れ値を除く）

(3) 出願年の差の検定

図 2 では，産学協同出願のグラフが大きく蛇行している．これは，産学協同出願の特許数が少ないため，他のステイタスが出揃っておらず，部分的に件数が少ない値の区間ができてしまうためと考えられる．論文や特許とのコサイン類似度に影響を与える可能性があるデータとして，一番近い特許・論文との出願年の差がある．ピアソンの積率相関検定で，一番近い特許とのコサイン類似度と出版年の差には，弱い相関が認められた．一方で，論文では，相関は認められなかった．

特許となる研究開発は，企業間の交流があったり，似たニーズによる研究開発であったりするために，時間の経過とともにどの研究開発も同様に技術の更新が進んでいくと考えられる．つまり，1 年前に一番近い特許が存在する場合より，3 年前に一番近い特許が存在する場合の方が研究開発が進んだ分だけ

特許の技術内容が離れ，出願年の差とコサイン類似度にマイナスの相関があると言える．しかし，一番近い論文が 1 年前でも数年前でも，特許との研究開発内容の差には変わりがない．学術論文で発表される研究も時間とともに発展していくが，その進展は特許で出願される技術には無条件では伝わらない．企業側の情報収集か，何らかの接触があった時に学術知識を取り入れた発明が行われる．すると，それが 1 年前の研究でも数年前の研究でも，特許と論文の内容の差の大きさは変わらないと考えられる．

表 5　一番近い特許・論文とのコサイン類似度と出版年の差の相関の検定結果

一番近い特許					一番近い論文				
相関係数	95%信頼区間		自由度	p value	相関係数	95%信頼区間		自由度	p value
-0.1516893	-0.1917047	-0.1111703	2260	4.09E-13	-0.0278111	-0.06958749	0.01406267	2191	0.193

6. 結論

精度の高い埋め込み用の言語モデルを用い，文章ベクトルにより研究開発内容のクラスタリングや定量的な比較を行い，研究開発動向調査に対して新しいエビデンスを提供することを試みた．言語モデルの解釈性能を証明することは難しく，関連業務の自動化タスク等，解決したい課題のはっきりとした問題にのみ使用されてきた言語モデルを分析に使用するための検討を行った．

企業が出願した特許と大学が出願した特許では，既出の一番近い特許・論文とのコサイン類似度の分布に統計的に有意に違いがあることが分かった．産学共同出願の特許では，大学と企業の中間の傾向があるが，一番近い論文では企業の特許との間に有意な差が確認され，一番近い特許では大学との差が大きい．一番近い特許のコサイン類似度はその特許の出願年との差が大きいと小さくなる傾向が統計的に有意であり，一番近い特許が出願されてから時間が経過した分，技術が進展していることが示唆された．

鈴木 2008[16]は， RIETI 発明者サーベイのデータと構造方程式モデリングにより，産学連携が特許のビジネス価値に対してマイナスの影響，技術的価値に対してプラスの影響を与えていることを明らかにしている．産学共同出願の特許が企業の特許と比較して，一番近い特許との差が小さくなっていることは前者に，一番近い論文との差が小さくなっていることが後者に対応していると考えられる．確かめられた傾向の背後で，実際に起こっていることは仮説の段階であり，追跡調査が必要と考えられる．特定の企業や大学の実証的調査や，データ数を拡充した別の分野での分析を行う．

参考文献

[1] Wang, Zihong, and Yufei Liu., Journal of Information Science, 50.4, 831-850(2024)
[2] Ascione, G. S., & Sterzi, V., arXiv preprint arXiv, 2403.16630(2024)
[3] Hain, Daniel S., et al., Technological forecasting and social change, 177, 121559(2022)
[4] 秋山 賢二，斎藤　隆文，情報処理学会論文誌デジタルプラクティス (DP)，4.3，58-68(2023)
[5] Haghighian Roudsari, Arousha, et al., Scientometrics, 127.1, 207-231(2022)
[6] Chikkamath, Renukswamy, et al., Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing, (2022)
[7] Bekamiri, Hamid, Daniel S. Hain, and Roman Jurowetzki, Technological Forecasting and Social Change, 206, 123536(2024)
[8] 下川公子，安藤敏，& 岡紀子，In 情報プロフェッショナルシンポジウム予稿集 第 7 回情報プロフェッショナルシンポジウム， 一般社団法人 情報科学技術協会，pp. 5-9(2010)
[9] 黒田　夢子，大庭　弘継，村上　祐子，情報処理学会第 219 回ソフトウェア工学研究会予稿集 (2025)
[10]Kazuyuki Motohashi, Hitoshi Koshiba, Kenta Ikeuchi, Scientometrics, 129, 2159-2179(2024)
[11]Xu, Shuo, et al., Scientometrics 126.9, 7445-7475(2021)
[12]NEDO, ロボット分野における研究開発と社会実装の大局的なアクションプラン，1-30(2023)
[13]Sandal, Nidhi, and Avinash Kumar., DESIDOC Journal of Library & Information Technology, 36.2, (2016)
[14]Thoma, G., Motohashi, K., & Jun, S., University Library of Munich, Germany, (2010)
[15]鈴木潤, RIETI Discussion Paper Series, 09-J-019, (2009)
[16]鈴木潤, RIETI Discussion Paper Series, 08-J-039, (2008)