多視点クラスタ間の正規化自己相互情報量変化に基づく特許のホワイトスペースの発見

Discovering Patent White Space Based on Changes in Normalized Pointwise Mutual Information Between Multi-View Clusters

宮澤 脩一*・藤井 健輔**
Shuichi Miyazawa, and Kensuke Fujii

要旨

特許の技術ランドスケープにおける未探索領域（ホワイトスペース）を自動的に検出する手法BLANCを提案する。本手法は、まず特許文書を用途・新規性・進歩性の3つの意味的視点で独立にクラスタリングする。次に、視点間のクラスタペアの統計的関連度を正規化自己相互情報量（NPMI）で定量化する。そして、ユーザー指定キーワードによるフィルタリング前後のNPMI差分を指標として未探索領域を検出する。本手法をフロートガラス/結晶化ガラス分野の特許302件に適用した結果、キーワード「フッ素」により検出されたホワイトスペースが、知財専門家が手動分析で事前に特定していた技術機会と一致し、専門家の知見を事前情報なしに再発見できることを実証した。

Abstract

We propose BLANC, a method for automatically detecting unexplored regions (white space) in patent technology landscapes. The method first clusters patent documents independently along three semantic perspectives: application/use, novelty, and inventive step. It then quantifies the statistical association of cluster pairs across these perspectives using Normalized Pointwise Mutual Information (NPMI). Finally, it detects white space based on changes in NPMI before and after filtering the corpus by a user-specified keyword. Applying the method to 302 patents in the field of float glass and glass-ceramics, we found that the white space detected using the keyword “fluorine” coincided with a technology opportunity that intellectual-property experts had previously identified through manual analysis. This result demonstrates that the method can rediscover experts’ insights without prior information.

*AGC株式会社デジタル・イノベーション統括部（shuichi.miyazawa@agc.com）
**AGC株式会社知的財産部（kensuke.fujii@agc.com）

1. 緒言

産業R&Dおよび知的財産（IP）部門において、新規特許出願テーマのブレインストーミングは重要な活動である。しかし、その準備作業として大量の関連特許を読み込むプロセスは、以下のような非効率性を抱えている。

(1) 明確な目的なく特許を読むことは多大な労力を要し、動機づけが難しい。
(2) ブレインストーミングの質は個人のドメイン知識に大きく依存し、参加者間で一貫性のない結果につながる。
(3) 従来の特許マップは作成に時間がかかり、グループ作業を要し、空白領域が技術的に意味のあるものか判断が難しい。
(4) AIベースの分類アプローチは、組織固有の技術知識を反映せず、一般的な視点のみを提供する。

これらの問題に対し、本研究では特許ランドスケープにおける「あってもよいはずなのにないもの」を自動的に検出する手法を提案する。本手法は、広範な特許母集団では十分に確立されている技術テーマの組合せが、ユーザーの関心キーワードでフィルタリングした際に結合が弱まるクラスタペアを発見するものであり、これらの弱まった結合は新規特許出願の候補テーマとなるホワイトスペースを表す。

提案手法は以下の3つのフェーズで構成される。

フェーズ1　多視点クラスタリング：特許文書の用途、新規性、進歩性といった各テキストを独立の意味的次元として扱い、各次元でニューラルトピックモデリング手法BERTopic[1] により独立にクラスタリングする。

フェーズ2　共起定量化：異なる次元のクラスタ間の統計的関連度をNPMIで定量化する。NPMIは正規化自己相互情報量すなわちNormalized Pointwise Mutual Informationの略であり、共起（1つの文書の中に2つの要素が同時に現れること）の強さを−1から+1の範囲で表す指標である。

フェーズ3　ホワイトスペース検出：ユーザー指定のキーワードで特許コーパス[*1]をフィルタリングし、フィルタ前後のNPMIの差分（ΔNPMI）が大きいクラスタペアをホワイトスペースとして検出する。

本論文の貢献は以下の3点である。(1) 多視点BERTopicクラスタリングとNPMI共起分析を統合した特許ホワイトスペース検出パイプラインの提案、(2) ΔNPMI指標の導入、(3) 実際の特許コーパスへの適用と検証である。

2. 関連研究

本手法は、主に3つの研究領域の知見を統合するものである。

2.1. 特許の空白分析・技術ギャップ検出

特許ランドスケープにおける未探索領域の特定する研究は、Yoonらの形態学マトリクス手法に始まる[2]。これは特許から抽出したキーワードを表形式で整理する手法である。たとえば縦軸に技術の観点、横軸にその観点における選択肢を並べ、特許が存在する組合せのセルを埋めていくと、空白のセルは未探索の技術組合せを表すことになる。Leeらはキーワードベクトルをを PCAで削減した2D特許マップ上で特許密度の低い空洞を機会として特定した[3]。Sonらはこれを発展させ、確率的生成トポグラフィックマッピング（GTM）による空白検出手法を提案した[4]。Kimらはテキストデータと引用ネットワークを用いて、クラスタ間の接続性を分析する手法を提案した。この手法では正規化されていない共起指標を用いて、接続が弱いが収束ポテンシャルの高いクラスタペアを機会としてフラグ付けする手法を提案した[5]。近年では、次節で紹介するBERTopicを用いた特許分析が進み、Jeonらは特許ドメイン特化埋め込みPatentSBERTaと組み合わせてデジタル治療分野の技術機会を発見し[6]、Yangらは特許ライフサイクル分析を報告している[7]。

2.2. ニューラルトピックモデリング（BERTopic）

BERTopic[1] は、大規模言語モデルの事前学習済み知識を活用したニューラルトピックモデリング手法である。そのパイプラインは以下の4つのステップから構成される。第1ステップでは、事前学習済み文埋め込みモデルSBERT[8] を用いて各文書を密なベクトル表現に変換する。SBERTはSentence-BERTの略であり、文を意味的に類似したベクトル表現に変換するモデルである。生成された高次元の意味的文埋め込みにより、文書間の意味的な類似度をコサイン類似度で定量評価できる。第2ステップでは、次元削減手法UMAP（Uniform Manifold Approximation and Projection）[9] によって、SBERTによる高次元埋め込みを低次元空間に圧縮する。UMAPは局所的な構造を保存しつつ高次元データを低次元に射影する代表的な手法である。第3ステップでは、密度ベースのクラスタリング手法HDBSCAN（Hierarchical Density-Based Spatial Clustering of Applications with Noise）[10] によりクラスタを検出する。HDBSCANはデータの密度が高い領域を自動的にクラスタとして検出する手法であり、クラスタ数を事前指定する必要がない点と、どのクラスタにも属さないノイズデータを自動的に検出できる点が特徴である。第4ステップでは、クラスベースTF-IDF（Term Frequency-Inverse Document Frequency）すなわちc-TF-IDFにより各クラスタの代表的なキーワードを抽出する。TF-IDFは、ある文書内での語の出現頻度（TF）と、その語がコーパス全体のどれだけ多くの文書に現れるかの逆数（IDF）を掛け合わせ、その文書に特徴的な語を高く重みづける古典的な重み付け手法である。c-TF-IDFは、各クラスタの全文書を1つの仮想文書とみなしてTF-IDFを計算する手法である。BERTopic

＊1　文書ごとに複数のテキストフィールドを持つ構造化されたテキストの集合

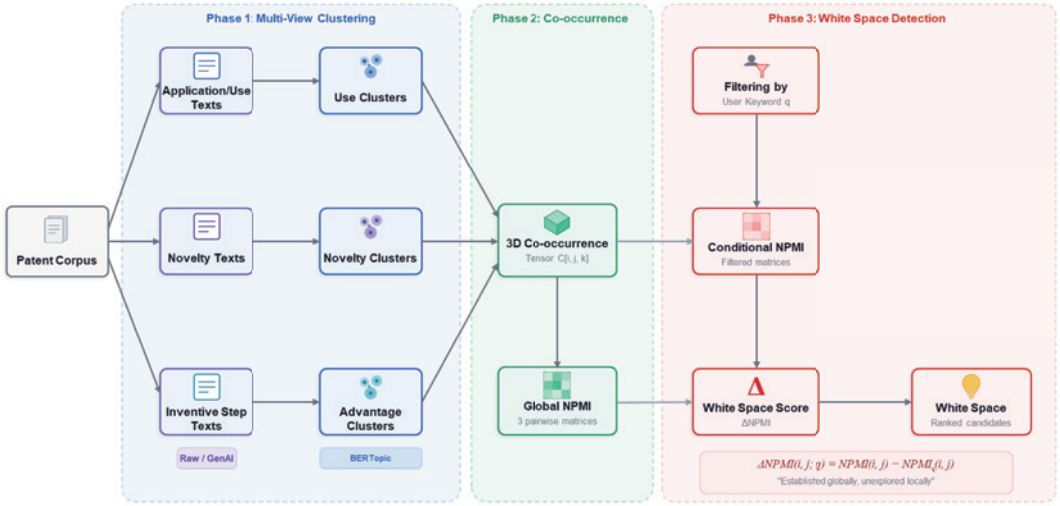


Fig. 1 System Architecture and Data Flow of the BLANC Pipeline. Dimension-specific texts are obtained from raw structured fields or via generative AI.

はLDA（Latent Dirichlet Allocation）[11] やNMF（Non-negative Matrix Factorization）[12] といった従来のトピックモデルを、一貫性・多様性の両指標で上回ることが報告されている。

特許ドメインでは専用の埋め込みモデルも開発されている。BERTを特許データでファインチューニングしたCPC分類モデルとしてPatentBERT[13] が、特許特化の文埋め込みモデルPatentSBERTa[14] が提案されている。いずれも、特許テキストの意味的類似度を捕捉する埋め込みの品質が向上させるものである。

2.3. 正規化自己相互情報量（NPMI）と共起分析

相互情報量（Pointwise Mutual Information; PMI）は、2つの事象の共起が独立の場合と比べてどの程度強いかを測る情報理論的指標であり、Churchらにより計算言語学の実用的指標として導入された[15]。その正規化版であるNPMIは、BoumaによりPMI値を−1から+1の範囲にスケーリングする指標として提案された[16]。ここで−1は共起が全くないこと、0は独立、+1は完全な共起を意味する。この有界性により、異なる頻度帯のクラスタペア間で関連度を公平に比較できる。

特許分析における共起指標の応用として、Breschiらは特許の技術分野を示す国際特許分類IPCコードの共分類を通じた知識関連性の測定を行い[17]、CurranらはIPC共分類分析による技術収束のモニタリングを提案した[18]。いずれも、同一特許に付与された分類コードの共起が偶然以上に関連しているかを検定するロジックであり、本研究のNPMIベースのクラスタ共起分析の理論的基盤となるものである。なお、Jaffeは特許分類ベクトルを用いた技術的近接性の定量化を先駆的に示しており[19]、共起指標で技術関連性を測定するアプローチの経済学的正当化を与えている。

2.4. 本研究の位置づけ

既存研究のうち最も近いものは、Kimら[5] のクラスタ間接続性分析と、Jeonら[6] によるBERTopicを用いた特許機会分析の2件である。前者は単一のテキスト表現から得たクラスタ間の接続性を分析するもので、複数の意味的次元を独立に扱う視点は持たない。同様に後者も単一観点でのクラスタリングのみで、統計的共起指標による空白検出は実装されていない。これに対し、本研究は、多視点クラスタリング、NPMI共起分析、条件付きホワイトスペース検出の3つを統合した初めてのパイプラインである。特に、複数の意味的次元で独立にクラスタリングした結果をNPMIでクロス集計する多視点アプローチは先行研究に存在しない。さらに、全体コーパスと条件付きコーパスのNPMI差分をホワイトスペースの指標として用いるΔNPMIは、本研究が初めて導入するものである。こうした多視点クラスタリングは、観点別に分離された特許テキストを大規模に得ることが従来は困難であったため実現されてこなかった。

3. 提案手法

3.1. 手法全体像

本手法のフローをFig. 1に示す。特許コーパスを入力とし、前処理、多視点クラスタリング、キーワード抽出、共起分析、ユーザー入力、条件付き共起計算、ホワイトスペース検出の順に処理を行う。テキストフィールドとしては、用途記述、新規性記述すなわち解決手段の記述、進歩性記述すなわち課題の記述の3次元を設定する。構造化されたテキストフィールドが特許文書内に直接利用可能な場合、たとえば抄録・請求項・要約などが利用できる場合は、それらをそのまま各次元にマッピングする。構造化フィールドが利用できない場合は、生成AIモデルを用いて全文テキストから各次元に対応する情報を抽出することも可能である。

3.2. 前処理とDTM構築

特許テキストから自然言語処理技術を用いて複合語

フレーズを抽出し、行列形式の文書・用語行列（Document-Term Matrix：DTM）を構築する。DTMは行が文書、列が語彙語、セル値が出現回数を表す行列であり、通常多くのセルが0の値をとる疎行列となる。ここで語彙語とは単語またはフレーズを指す。全テキストフィールドを統合したマスタDTMと、各次元別のDTMをそれぞれ構築する。マスタDTMは後述のキーワードフィルタリングに使用し、次元別DTMは各次元でのフィルタリングに使用する。

3.3. 多視点クラスタリング

特許明細書は標準的に、用途・新規性・進歩性という異なる観点のテキストセクションを含んでいる。新規性のセクションは解決手段の記述に、進歩性のセクションは課題の記述に相当する。本研究の鍵となる考え方は、これらの異なるテキストフィールドをそれぞれ独立の意味的次元として扱い、機械学習における多視点学習と同様に独立にクラスタリングすることである。単一のクラスタリングでは捉えられない次元間の技術組合せのパターンを、NPMI共起分析によって定量化できる点が本手法の利点である。本研究では次元数K=3とし、個々の次元を添字k（1≤k≤K）で表す。第1の次元は用途テキストであり、その特許がどこに使われるかを表す。第2の次元は新規性テキストであり、何が新しいかを表す。第3の次元は進歩性テキストであり、どんな利点があるかを表す。

各次元kについて、第2.2節で述べたBERTopicパイプラインを独立に実行する。すなわちSBERT埋め込み、UMAP次元削減、HDBSCANクラスタリング、キーワード抽出の4ステップである。結果として、各文書dにクラスタの三つ組d →（c_1, c_2, c_3）が割り当てられる。ここでc_k（k=1,2,3）は文書dが第k次元（それぞれ用途・新規性・進歩性）で属するクラスタのIDを表す。すなわち同一の文書が、用途視点ではクラスタc_1、新規性視点ではクラスタc_2、進歩性視点ではクラスタc_3に同時に属するという、次元ごとのクラスタ所属の組で表現される。この三つ組が後述の次元間共起集計の単位となる。なお、HDBSCANによりどのクラスタにも属さないと判定された文書はノイズとして扱い、クラスタIDを−1とする。HDBSCANで十分なクラスタ数が得られない場合は、k-meansによるクラスタ数指定を併用することも可能である。

3.4. PMIベースキーワード抽出

BERTopic標準のc-TF-IDFによるキーワード抽出は、特許テキストのような専門的な文書では高頻度語に偏る傾向がある。そこで、クラスタ内での出現頻度とコーパス全体に対する特異性をバランスよく評価するため、以下の複合スコアを導入した。

$$score(w,i) = \alpha \cdot NPMI(w,i) + (1-\alpha)\cdot P(w \mid i) \quad (1)$$

ここで、$P(w \mid i)$はクラスタiに属する文書のうち語wを含む文書の割合、すなわちクラスタ内頻度を表す。NPMI(w, i)は語wとクラスタiの正規化自己相互情報量であり、コーパス全体に対する特異性を表す。αは両者のバランスを制御する混合重みであり、デフォルトは0.5とする。高い$P(w \mid i)$はクラスタ内での網羅性を、高いNPMIはコーパス全体に対する弁別性を保証する。上位n語をクラスタキーワードとし、nのデフォルト値は10とする。なお、キーワード抽出におけるNPMI計算では、確率推定の数値安定性のため$\varepsilon = 10^{-12}$を使用する。

3.5. 3D共起テンソルとNPMI計算

3つの次元のクラスタ数をそれぞれN_1、N_2、N_3とすると、各次元のクラスタに同時に属する文書数を数える3D共起テンソルを構築する。

$$C[i,J,k] = |\, D_i^{(1)} \cap D_j^{(2)} \cap D_k^{(3)} \,| \quad (2)$$

このテンソルを各次元で周辺化し、ペアワイズの2D共起行列を得る。たとえば、用途×新規性の行列は進歩性次元を合計することで得られる。各ペアにNPMIを計算する。

$$NPMI(x_i, y_j) = \frac{PMI(x_i, y_j)}{-log_2 P(x_i, y_j)} = \frac{log_2\left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}\right)}{-log_2 P(x_i, y_j)} \quad (3)$$

3.6. 条件付きホワイトスペース検出

ホワイトスペース検出の核心的なアイデアは、「全体コーパスで高い関連度を持つクラスタペアが、特定のキーワードを含む文書の部分集合（サブコーパス）に限定すると関連度が低下する」という現象を検出することである。ユーザーはキーワードqを指定し、DTMを用いてキーワードを含む文書集合D_qを抽出する。この縮小されたコーパスで条件付き共起テンソルと条件付きNPMIを計算する。

ホワイトスペース指標ΔNPMIは以下で定義される。

$$\Delta NPMI(i,j;q) = NPMI(i,j) - NPMI_q(i,j) \quad (4)$$

ここで、NPMI(i, j)は全体コーパスでのクラスタペアの関連度を表し、NPMI$q(i, j)$はキーワードqでフィルタリング後の関連度を表す。ΔNPMIの正の値は、フィルタリングにより関連度が低下したことを示す。すなわち、全体では確立された技術組合せが、ユーザーの関心領域では未探索であることを意味する。さらに以下の3条件を満たすペアのみを保持する。第1に、条件付き共起が存在すること。第2に、ΔNPMI>0であること。第3に、元のNPMIが最低閾値$\theta_{\min}$を超えること。ここで、$\theta_{\min}$のデフォルト値は0.3とする。この第3の条件は本手法の設計上重要である。元のNPMIが低いクラスタペアでは、関心領域と無関係なキーワードを与えるだけでΔNPMIが形式的に正となりうるが、これらは本来検出したい「未探索の有望領域」ではない。そこで元のNPMIがθ minを超える

ペアに限定することで、「全体コーパスでは強く結びついているにもかかわらず、関心領域に限定するとその結びつきが消える」という意外性のあるペアのみを抽出する。最終的に、上位kペアを候補テーマとして提示する。ここで、kのデフォルト値は10とする。ここで、キーワードqは自動的に選定されるものではなく、特許分析者であるユーザーが自身の関心領域に基づいて指定する。

3.7. マイクロレベルフレーズ分析

上述のマクロレベルのクラスタペア分析に加え、ユーザーが選択したクラスタ組合せ内のフレーズレベルの共起分析も提供する。クラスタ制約に合致する文書を抽出し、高頻度フレーズから共起行列を構築する。ここで対象とするフレーズは最大100語までとする。構築した行列からNPMIを計算し、ヒートマップとして表示する。フレーズはユーザーのキーワードとのNPMIでランク付けされ、ホワイトスペースの具体的な技術要素を見つける手がかりを提供する。

4. 実験

4.1. データセット

フロートガラスおよび結晶化ガラスすなわちガラスセラミックに関連する特許302件を収集した。対象期間は2019年～2023年、検索条件は「フロートガラス×結晶化」の技術分野である。各文書には、一行要約、技術背景、課題、解決手段、用途、追加情報の構造化フィールドが含まれる。BLANCの3次元マッピングは以下の通りとした：用途記述を用途次元、解決手段記述を新規性次元、課題記述を進歩性次元に割り当てた。**Table 1**にデータセット統計を示す。

Table 1 Summary of Patent Dataset

項目	値
文書数	302件
対象期間	2019–2023年
用途（平均文字数）	78
解決手段（平均文字数）	304
課題（平均文字数）	79
マスタDTM語彙数	3,200
トークナイザ	fugashi（MeCab）

4.2. クラスタリング結果

3つの意味的次元について、多言語Sentence-BERTモデル（paraphrase-multilingual-MiniLM-L12-v2）で埋め込みを生成し、UMAP（n_neighbors=10, n_components=5）で次元削減後、クラスタリングを実行した。用途次元および進歩性次元ではHDBSCAN（min_cluster_size=8, min_samples=3）を使用した。新規性（解決手段）次元ではHDBSCANが2クラスタにしか分割しなかったため、k-means（k=10）によるクラスタ数指定で10トピックへの分割を行った。k=10の選択は、用途次元（15クラスタ）および進歩性次元（11クラスタ）のクラスタ数を参考にしたドメイン知識に基づく設定であり、10前後のトピック数が本技術分野の主要テーマを網羅するのに適切と判断した。ただし、一部の次元でクラスタ数を手動で指定することは、検証の客観性を一定程度損なう可能性がある。本研究では原則としてクラスタ数を自動推定するHDBSCANを使用し、それが有意な分割を生成しない次元に限って補完的にKMeansを用いた。クラスタ数の選択がホワイトスペース検出結果に与える感度の体系的な検証は、第5.3節で述べるとおり今後の課題である。

結果、用途次元で15クラスタ（ノイズ率12.6 %）、新規性次元で10クラスタ（ノイズ率0 %）、進歩性次元で11クラスタ（ノイズ率15.9 %）が得られた。ここでノイズ率とは、HDBSCANによりどのクラスタにも分類されなかった文書の割合を指す。用途次元はコーティング、合わせガラス、太陽電池、フロート強化、磁気ディスク等の技術テーマに分かれた。新規性次元は組成規定、強化/反り抑制、結晶化LAS、バリア層等を含む。進歩性次元は基板平滑化、反り抑制、核生成、CTE制御、調理器等のテーマを形成した。各クラスタの詳細は**Table 2**を参照されたい。なお、各クラスタには上位10個の代表キーワード（複合語フレーズを含む）がスコア順に付与される。本稿中のクラスタ名称（例：「強化/反り抑制」「結晶化LAS」等）は、これらの代表キーワード群を著者が技術的に解釈して命名したラベルであり、提案手法による直接出力は単語やフレーズの順位付きリストである。

Table 2 Clustering Results for Three Semantic Dimensions

次元	クラスタ数	手法	ノイズ率	最大クラスタ
用途	15	HDBSCAN	12.6 %	27件
新規性（解決手段）	10	k-means	0 %	50件
進歩性（課題）	11	HDBSCAN	15.9 %	64件

4.3. NPMI共起分析の結果

Table 3 Top NPMI Cluster Pairs Across Dimension Combinations

次元ペア（A×B）	クラスタA	クラスタB	NPMI	共起
新規性×進歩性	強化/反り抑制	強化/反り抑制	0.624	21
用途×進歩性	薪ストーブ	結晶化ガラス	0.443	3
用途×進歩性	合わせガラス	結晶化ガラス	0.379	5
用途×進歩性	セラミック材料	調理器	0.363	3
用途×進歩性	コーティング	調理器	0.360	6
用途×新規性	暖炉	結晶化LAS	0.332	3
用途×新規性	コーティング	バリア層	0.327	9
用途×新規性	フラットガラス	強化/反り抑制	0.296	8

3つの次元ペアのNPMI行列を計算し、**Table 3**にまとめた。最も注目すべき結果は、新規性×進歩性の次元ペアにおいて「強化/反り抑制」クラスタ同士が示したNPMI=0.624という非常に強い結合である。これは、反り抑制の課題と強化手法の新規性が本技術分野で確立された組合せであることを定量的に裏付けている。その他の高NPMIペアとして、暖炉クラスタと結晶化LASクラスタの組合せ、コーティングクラスタとバリア層クラスタの組合せ、薪ストーブクラスタと

結晶化ガラスクラスタの組合せ、コーティングクラスタと調理器クラスタの組合せなどが挙げられる。Fig. 2に用途×進歩性のNPMIヒートマップを示す。

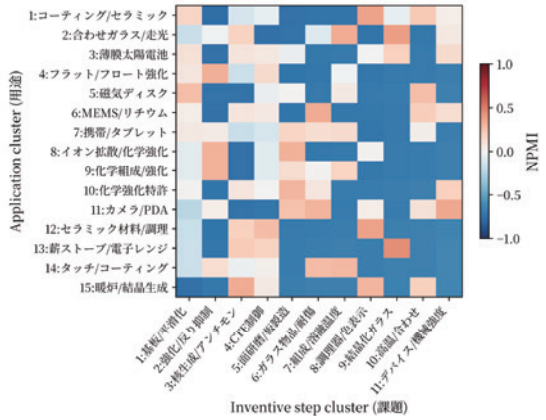


Fig. 2 NPMI Heatmap for Application × Inventive Step Dimension Pair

4.4. ホワイトスペース検出のケーススタディ

ホワイトスペース検出のケーススタディとして、「フッ素」をキーワードとした分析を行った。NPMI閾値はθ_{min}=0.2を使用した。302件の小規模コーパスではθ_{min}=0.3ではフィルタリング後の候補が極端に少なくなるため、閾値を緩和した値である。

■キーワード「フッ素」(14件、4.6 %)

用途×新規性の次元ペアにおいて、フラット/フロート強化クラスタと強化/反り抑制クラスタの組合せがΔNPMI=0.464で最大候補として検出された。結果の詳細はTable 4に示す。用途×進歩性でも同一テーマのペアがΔNPMI=0.480で1位となった。なお、Table 4のNPMI（g）は全体コーパスでのNPMI（global、すなわち式（3）のNPMI（i, j））を、NPMI（c）はキーワードでフィルタリングした条件付きコーパスでのNPMI（conditional、すなわちNPMIq（i, j））を表し、ΔNPMI = NPMI（g）−NPMI（c）である。3つの次元ペアすべてにおいて、「強化/反り抑制」クラスタがフッ素フィルタにより最大のΔNPMIを示した。この結果は、知財部の専門家が手動の特許分析で事前に特定していた「フッ素表面処理による反り抑制」というホワイトスペースと一致する。

Table 4 White Space Candidates for Keyword 'Fluorine'

次元ペア	クラスタA	クラスタB	NPMI (g)	NPMI (c)	ΔNPMI
用途×課題	フラット/フロート強化	強化/反り抑制	0.316	-0.163	0.480
用途×課題	化学組成/強化	強化/反り抑制	0.311	0.116	0.195
用途×新規性	フラット/フロート強化	強化/反り抑制	0.296	-0.168	0.464
用途×新規性	タッチ/コーティング	強化/反り抑制	0.245	0.102	0.143

4.5. マイクロフレーズ分析の例

フッ素キーワードで検出されたホワイトスペース候補、すなわちフラット/フロート強化と強化/反り抑制の組み合わせに対し、マイクロレベル分析を適用した。このクラスタペアに属する文書群から頻出フレーズを抽出し、フレーズ間のNPMI行列を計算してWard法[20]による階層的クラスタリングを行った。結果として、反り制御に関連するフレーズ群とフッ素処理に関連するフレーズ群が別々のテーマグループを形成し、両者の結合が弱いことが確認された。前者には化学強化、イオン交換、応力プロファイルといったフレーズが含まれ、後者には表面処理、フッ化処理、濃度分布といったフレーズが含まれた。これにより、フッ素表面処理技術と反り抑制課題の組み合わせが、フレーズレベルでもホワイトスペースとして支持されることが示された。この知見は、具体的な出願テーマの検討材料として特許ブレインストーミングに活用された。

なお、本節の302件のケーススタディに加え、手法の汎用性を検証するため、公開特許データを用いた別ドメイン（機械学習/AI分野5,417件、ガラス組成分野1,982件）でも検証実験を実施した。これは、既知のクラスタペアの結合を人工的に弱めた（ホワイトスペースを注入した）コーパスを作成し、本手法がその注入位置を再び検出できるかを定量評価するものである。詳細な定量結果は§5.2で触れるとおり別稿で報告予定であるが、人工注入したホワイトスペースが多視点分解とNPMI正規化により高い精度で回収されることを確認している。

5. 考察

5.1. 結果の解釈と実務的含意

本手法はキーワード起点で運用されることを想定している。すなわち、研究者はまず自身の専門領域に関連するキーワードqを起点として設定し、複数のキーワードを試行的に入力する。その中で、全体コーパスでは強く結びついていたクラスタペアの結びつきがそのqによって低下し、高いΔNPMIを示すペアを見いだした場合、そこを出願機会の候補として深掘りする。このワークフローにおいて、本手法は、特許ブレインストーミングの準備作業において、3つの実務的価値を提供する。第1に、関連度の高いクラスタペアとその変化を定量的に示すことで、ブレインストーミングの質が個人のドメイン知識に依存し、参加者間で一貫性を欠くという課題（緒言の課題（2））に対応する。第2に、「あるべきってもよいはずなのにないもの」を定量的に特定することで、明確な目的なく特許を読む労力と動機づけの難しさという課題（課題（1））に対応する。第3に、インタラクティブUIによる3ステップワークフローが、作成に時間とグループ作業を要し空白領域の技術的意義の判断が難しい従来の特許マップの課題（課題（3））に対応し、大幅な時間節約を実現する。

実際に、本手法の分析結果を用いた特許ブレインストーミングにおいて、「フッ素表面処理×反り抑制」のホワイトスペース候補が議論の出発点となった。この候補は、知財部の専門家が独立に手動分析で特定し

ていたものであり、BLANCが事前情報なしに同じ結論を導いたことで、手法の実用性が確認された。手動分析では数週間を要する作業が、BLANCでは数分で完了する点も実務上の大きな利点である。

5.2. 多視点アプローチの有効性

用途・新規性・進歩性の3つの次元で独立にクラスタリングし、次元間のNPMIでクロス集計することの利点は、単一観点では検出できない技術組合せを発見できる点にある。たとえば、「用途Aに対して新規性Bのアプローチは全体では確立されているが、特定のキーワード領域では未探索である」といった発見は、全テキストを一括でクラスタリングする単一観点アプローチでは可視化できない。

本手法の汎用性については、公開特許データを用いた機械学習/AI分野（5,417件）およびガラス組成分野（1,982件）の2ドメインで人工的なホワイトスペース注入実験を実施し、多視点分解とNPMI正規化の両方が不可欠であることを定量的に実証した。さらに、特許請求項からの法的定型句除去や要約からの図面説明除去といったテキスト前処理がクラスタリング品質に決定的な影響を与えることも明らかにした。これらの定量評価の詳細は別稿で報告予定である。

5.3. 制限事項

本手法にはいくつかの制限がある。第1に、ホワイトスペース検出は本質的に「存在しないもの」を見つけるタスクであり、客観的な正解データは存在しない。本研究で用いる専門家評価は最善の代替手段であるが、評価者の主観やドメイン知識に依存するという限界がある。第2に、フィルタリング後のコーパスが小さい場合、NPMI推定が統計的に不安定になる。平滑化や最低出現回数フィルタの導入が推奨される。第3に、ホワイトスペース候補の技術的意義の判断には専門家の知識が依然として必要である。本手法が検出するホワイトスペースは、それ自体が有望な出願機会を保証するものではなく、特許ブレインストーミングの出発点を効率的に提示し、その後の議論を早めることに寄与するものである。したがって本手法は専門家の判断を置き換えるものではなく、その判断を支援するツールである。第4に、クラスタリングの結果はHDBSCANのパラメータや埋め込みモデルの選択に依存するため、パラメータ感度の検証が重要である。

6. 総括

本研究では、特許ランドスケープにおけるホワイトスペースを自動検出する手法を提案した。本手法の主な貢献は以下の3点である。

第1に、多視点BERTopicクラスタリングとNPMI共起分析を統合した特許ホワイトスペース検出パイプラインを提案した。特許文書を用途・新規性・進歩性の3つの意味的次元で独立にクラスタリングし、次元間をNPMIでクロス集計するこの3コンポーネントの統合は先行研究に存在しないものである。

第2に、条件付きNPMI差分による新しいホワイトスペース指標ΔNPMIを導入した。ΔNPMIは全体コーパスとキーワードフィルタリング後のコーパスのNPMI差分を計算する指標である。これにより、「一般的には確立されているが特定領域では未探索」な技術組合せを定量的に特定できる。

第3に、フロートガラス/結晶化ガラス分野の302件の特許コーパスに本手法を適用した。キーワード「フッ素」で分析した結果、「フロート強化ガラス×反り抑制」のホワイトスペースをΔNPMI=0.48で検出した。この候補は知財部の専門家が独立に手動分析で特定していたものと一致し、BLANCの実務的有用性を実証した。

今後の展望として、3つの方向性が挙げられる。第1に、大規模言語モデルを用いたホワイトスペース候補の自動説明生成である。すなわち、なぜこの組合せが未探索なのかを言語化する機能の実装である。第2に、ブートストラップ等による統計的有意性検定の導入である。第3に、出願年によるコーパス分割を用いたホワイトスペースの時間的変化の追跡である。

付記

本論文に記載された手法の一部について特許出願を行っている。出願番号は特開2025-148721号である。

―参考文献―

(1) M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv:2203.05794 (2022).

(2) B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: Keyword-based morphology analysis, Technol. Forecast. Soc. Change 72 (2005): 145–160.

(3) S. Lee, B. Yoon, Y. Park, An approach to discovering new technology opportunities: Keyword-based patent map approach, Technovation 29 (2009): 481–497.

(4) C. Son, Y. Suh, J. Jeon, Y. Park, Development of a GTM-based patent map for identifying patent vacuums, Expert Syst. Appl. 39 (2012): 2489–2500.

(5) B. Kim, G. Gazzola, J.-M. Lee, D. Kim, K. Kim, M. K. Jeong, Inter-cluster connectivity analysis for technology opportunity discovery, Scientometrics 98 (2014): 1811–1825.

(6) E. Jeon, N. Yoon, S. Y. Sohn, Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa, Technol. Forecast. Soc. Change 186 (2023): 122130.

(7) H. Yang, S. Chen, Industry 5.0: Life-cycle mapping of sustainable technologies using BERTopic-driven patent analytics, World Pat. Inf. 80 (2025): 102273.

(8) N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proc. EMNLP-IJCNLP 2019 (2019): 3982–3992.

(9) L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426 (2018).
(10) R. J. G. B. Campello, D. Moulavi, J. Sander, Density-Based Clustering Based on Hierarchical Density Estimates, Proc. PAKDD 2013, LNCS 7819 (2013): 160–172.
(11) D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, J. Mach. Learn. Res. 3 (2003): 993–1022.
(12) D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999): 788–791.
(13) J.-S. Lee, J. Hsiang, Patent classification by fine-tuning BERT language model. World Pat. Inf. 61 (2020): 101965.
(14) H. Bekamiri, D. S. Hain, R. Jurowetzki, PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT, Technol. Forecast. Soc. Change 206 (2024): 123536.
(15) K. W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Comput. Linguist. 16 (1990): 22–29.
(16) G. Bouma, Normalized (pointwise) mutual information in collocation extraction, Proc. GSCL 2009 (2009): 31–40.
(17) S. Breschi, F. Lissoni, F. Malerba, Knowledge-relatedness in firm technological diversification, Res. Policy 32 (2003): 69–87.
(18) C.-S. Curran, J. Leker, Patent indicators for monitoring convergence – Examples from NFF and ICT, Technol. Forecast. Soc. Change 78 (2011): 256–273.
(19) A.B. Jaffe, Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. American Economic Review 76 (1986): 984–1001.
(20) J. H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963): 236–244.