

AI for Materials Science(AI4MS)

知識多様体について



TOHOKU
UNIVERSITY

東北大学大学院工学研究科 航空宇宙工学専攻 教授
副理事（産学連携担当）・リサーチプロフェッサー
グリーンクロステック研究センター長
研究代表者 岡部朋永

知識多様体に関する研究

<知識多様体に関する研究> (2026年6月3日up)

知識多様体とは、たくさんの文書を「意味の位置関係」として並べた空間のことです。たとえば、文書として学術論文を例にとると各論文の本文をそのまま読む代わりに、まず文書を埋め込みベクトルやTF-IDFベクトルに変換します。これは、文書に含まれる語句や表現の特徴を数値の列で表したものです。内容が似ている文書同士はベクトルも近くなり、異なる内容の文書は遠くなります。この高次元の関係を2次元に写すと、論文同士の意味的な地図ができます。地図上の点は論文を表し、近い点ほど研究内容が近いと考えます。さらに、点と点の間を補間すれば、既存論文の中間にある「仮想的な知識」も推定できます。測地線は、その知識空間のゆがみを考慮した最短経路で、単なる直線ではなく、意味が自然に変化する道筋を表します。例えば、ChatGPT PROにご自身で準備した20本の論文PDFと仕様書 ([2026_06_03_知識多様体解析仕様書.pdf](#)) を渡し、実行を依頼すれば、解析が出来ます。もしも、これを簡単に試してみたいときには、ご自身で準備した20本の論文PDFと仕様書 ([2026_06_03_プロンプト仕様書.pdf](#)) に記されているプロンプトを順に打ち込んでください。簡易な解析が行えます。

<Research on Knowledge Manifolds> (June 3rd, 2026 up)

A knowledge manifold is a space in which many documents are arranged according to their “semantic positional relationships.” Instead of reading the full text of each paper directly, each document is first converted into an embedding vector or a TF-IDF vector. This is a numerical representation of the features of the words and expressions contained in the document. Documents with similar content have vectors that are close to each other, while documents with different content are farther apart. By projecting these high-dimensional relationships into two dimensions, a semantic map of the papers can be created. Each point on the map represents a paper, and points that are closer together are considered to have more similar research content. Furthermore, by interpolating between points, it is possible to estimate “virtual knowledge” that lies between existing papers. A geodesic represents the shortest path that accounts for the distortion of this knowledge space; it is not merely a straight line, but a path along which meaning changes naturally. For example, if you provide ChatGPT Pro with 20 paper PDFs that you have prepared, together with the specification document “[2026_06_03_Knowledge Manifold Analysis Specification.pdf](#),” and ask it to perform the analysis, the knowledge manifold analysis can be carried out. If you would like to try a simpler version, prepare 20 paper PDFs and enter the prompts in sequence as described in “[2026_06_03 Prompt Specification Document EN.pdf](#)” This will allow you to conduct a simplified analysis of the knowledge manifold. Please note that some adjustments may be needed when using them in languages other than Japanese. We recommend working with an LLM to tune them appropriately.

Knowledge Manifold Analysis Specification (2D, with Uniform-Placement Constraint) - Revised Version v2.0

Purpose of This Revision

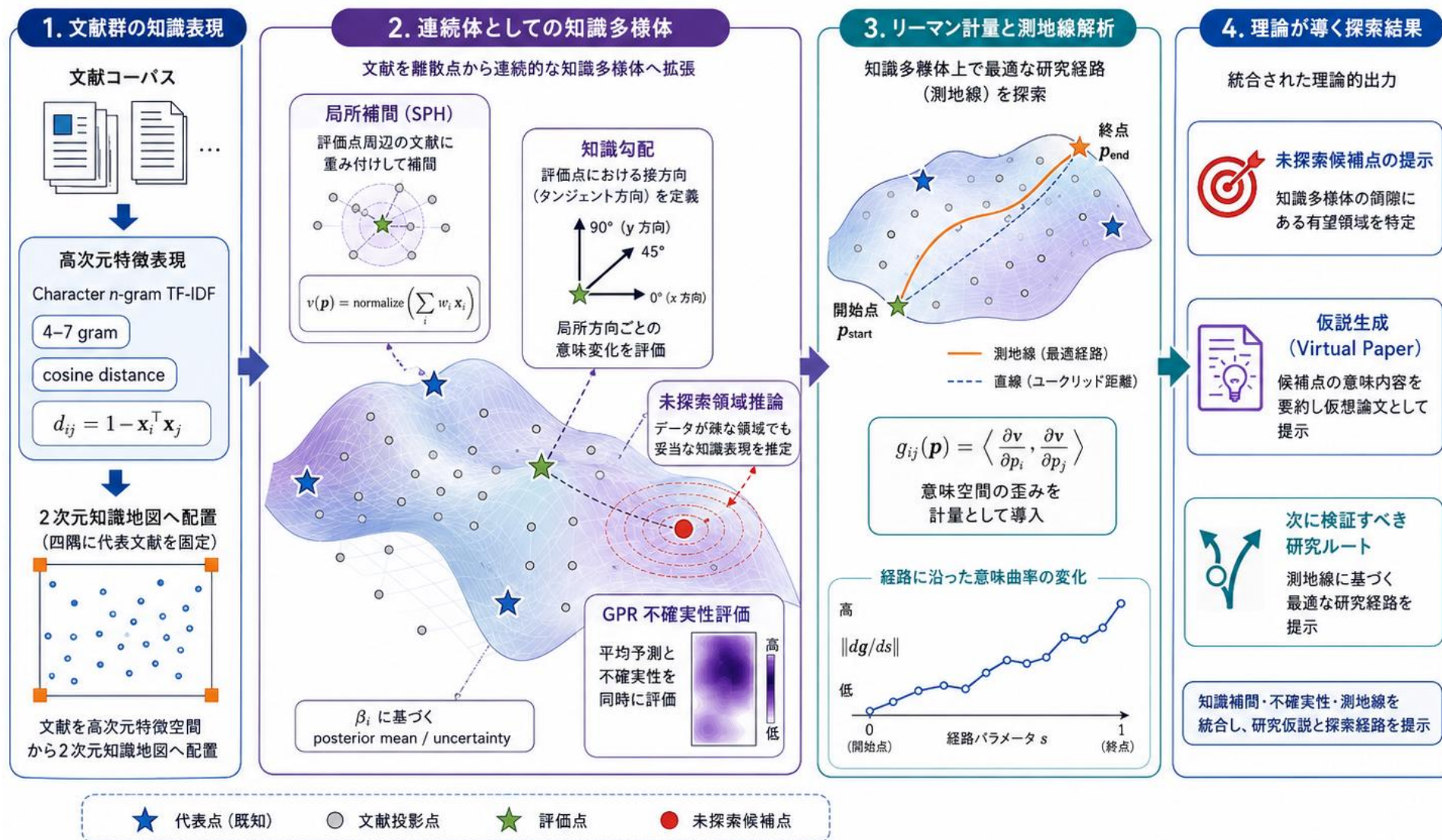
This revised version prevents the issue in which geodesic optimization could be skipped by a runtime fallback and the straight initial path could be output as-is. The specification is designed so that the geodesic is reliably output from the beginning as an optimized path.

In this specification, the primary objective of the 2D knowledge map is to preserve semantic distance relationships based on cosine distance, while the secondary objective for visualization is to spread points as uniformly as possible within the region $-1 < x < 1$, $-1 < y < 1$.

Methods not described in this specification must not be changed arbitrarily. Any steps where results may vary due to differences in the execution environment or library versions must be recorded in `manifest.json` and `analysis_report.md` under “Sources of possible variation”.

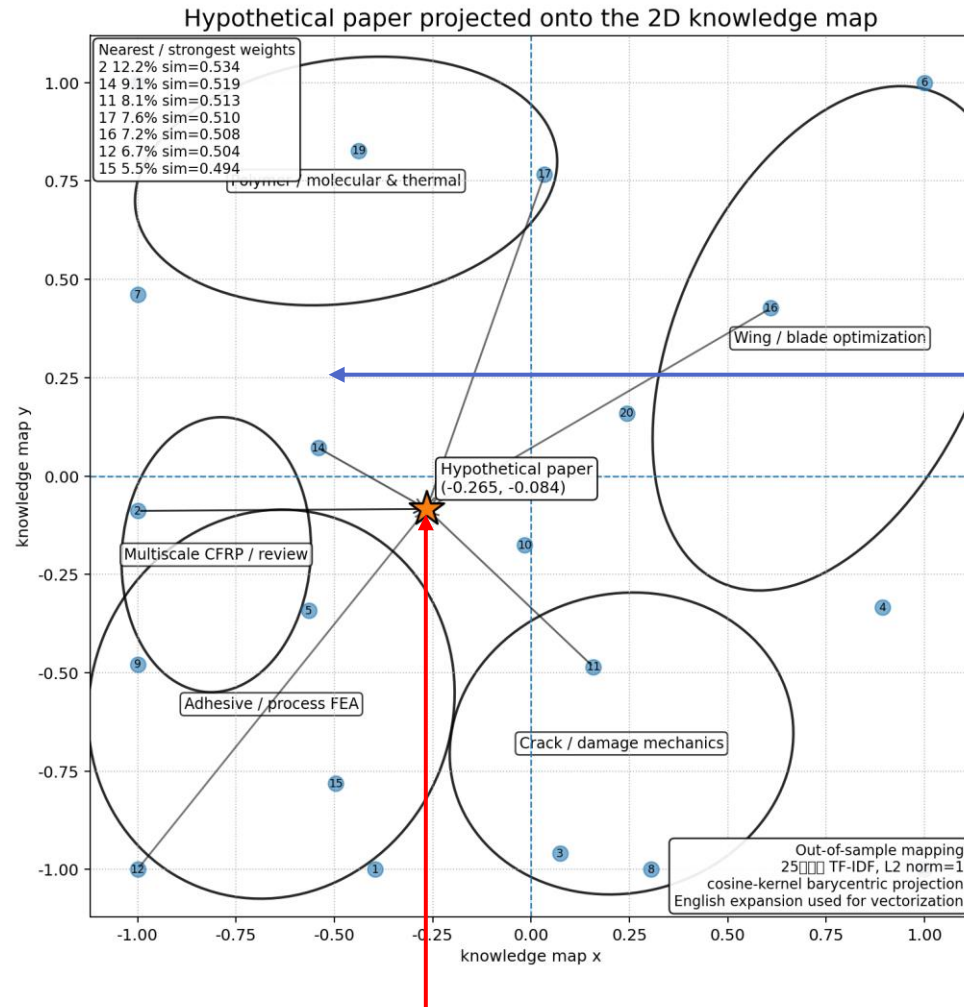
1. Common Conditions

1. Extract PDF body text in page order. Figure/table captions and references must be included. Do not use OCR. Record page numbers only for pages from which text cannot be extracted.
2. Text preprocessing is limited to Unicode NFKC normalization, lowercasing of alphabetic characters, and replacement of consecutive whitespace with a single space. Do not perform stop-word removal, stemming, translation, or summarization.
3. TF-IDF must use an implementation equivalent to scikit-learn's `TfidfVectorizer`, with the following settings fixed:
 - `analyzer="char_wb"`
 - `ngram_range=(4,7)`
 - `sublinear_tf=True`
 - `norm="l2"`
 - `max_features=250000`
 - `min_df=1`
4. If the number of features exceeds 250,000, fix the TF-IDF vocabulary by following the library's standard vocabulary ordering and feature-limit handling, not by frequency order.
5. Use `random_state=0` for every procedure that uses randomness.
6. L2-normalize every document vector. All subsequent similarities are cosine similarities.



知識多様体における知識多様体(簡単な例__2次元)

3



(推定論文)

本研究では、反応誘起相分離を伴う熱硬化性樹脂をマトリックスとするCFRPを対象に、分子構造から複合材レベルの残留変形・力学特性までを一貫して予測するマルチスケール解析手法を提案する。まず、樹脂組成や架橋構造が硬化収縮、熱膨張、粘弾性特性に与える影響を分子シミュレーションおよび熱機械試験により評価する。次に、**得られた樹脂物性を繊維/樹脂スケールの有限要素モデルへ導入し、界面近傍の応力集中、微視的残留応力、有効弾性率を算出する。**さらに、均質化されたラミナ物性を積層板解析へ受け渡し、成形後の反り変形、長期変形、損傷発生傾向を予測する。実験的に得た変形量および微細構造観察結果と比較することで、樹脂設計がCFRP構造の寸法安定性と耐久性に及ぼす影響を明らかにする。

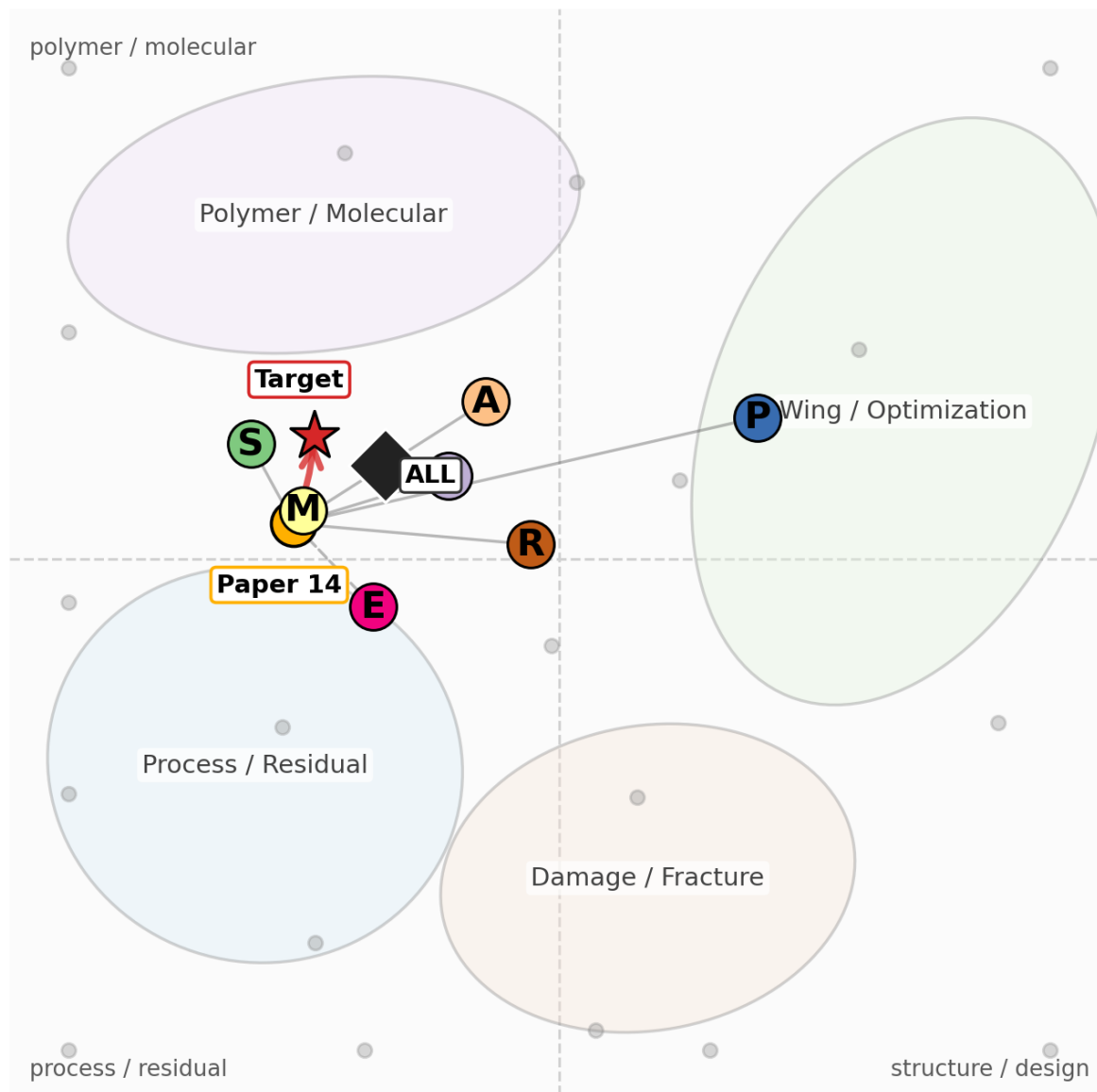
(仮想論文)

量子化学・分子シミュレーションで界面特性を同定し、それを複合材料全体の有効物性へ落とし込み、**放射光計測で妥当性を検証し、さらに航空機構造重量への影響まで推算する、**分子-材料-構造-システムを接続するマルチスケール設計論文

SCAMPERにおける研究戦略(簡単な例__2次元)

4

Computed SCAMPER positions on knowledge map



SCAMPER labels

Placed by TF-IDF projection of each strategy text

S Replace empirical inputs with molecular predictions.

経験則からMD

C Add synchrotron evidence to validate morphology.

A Apply microstructure design to phase-separated CFRP.

相分離構造を反映

M Expand targets to stiffness, Tg, toughness and shrinkage.

各種物性を加える

P Use improved properties for airframe weight impact.

E Reduce process-only focus; keep transferable physics.

R Reframe the story as molecule to Method structure to aircraft.

1. Vectorize each strategy in 250k TF-IDF space
2. Project by cosine-kernel barycenter
3. Plot computed coordinates

Combined point: (-0.35, +0.19)

Target alignment: 0.71

TRIZ戦略

論文14を評価点 (-0.5, 0.25) に移すための矛盾マトリックス

残留変形中心の研究を、分子・界面・計測・材料設計寄りへ再構成する



重点3原理

5 結合

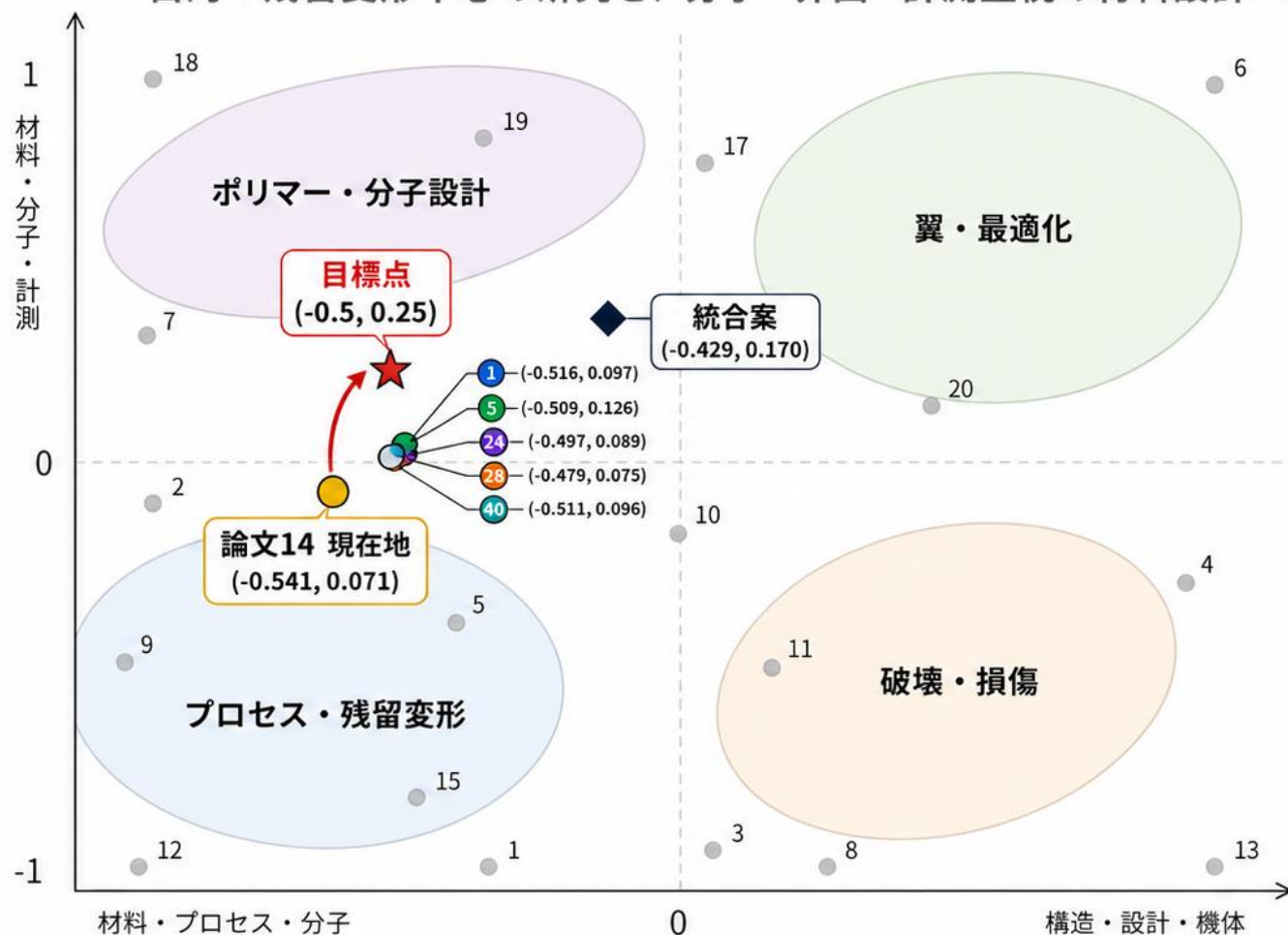
24 仲介

1 分割

計測・界面・階層化を軸に、材料設計型の論文へ再構成

論文14を評価点 $(-0.5, 0.25)$ に近づけるTRIZ戦略

目的：残留変形中心の研究を、分子・界面・計測重視の材料設計へ再構成する



番号位置は、TRIZ戦略文を25万次元文字n-gram TF-IDF空間に写像し、既存20論文との類似度から2D投影して計算

TRIZ戦略 (計算配置)

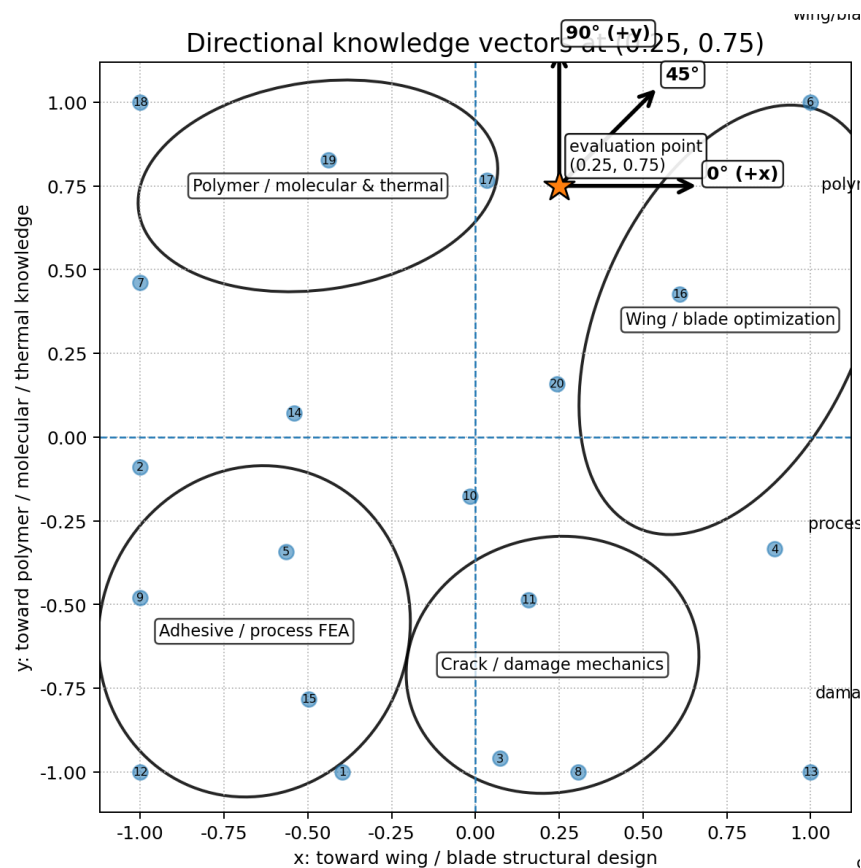
- 1 分割
分子→界面→複合材→構造
- 5 結合
放射光計測+残留変形解析
- 24 仲介
界面特性で橋渡し
- 28 機械置換
X線CT・散乱で直接可視化
- 40 複合材料
相分離樹脂・界面設計



要点：残留変形中心の研究を、分子・界面・計測重視の材料設計へ再構成する。

知識多様体における知識勾配(簡単な例__2次元)

7



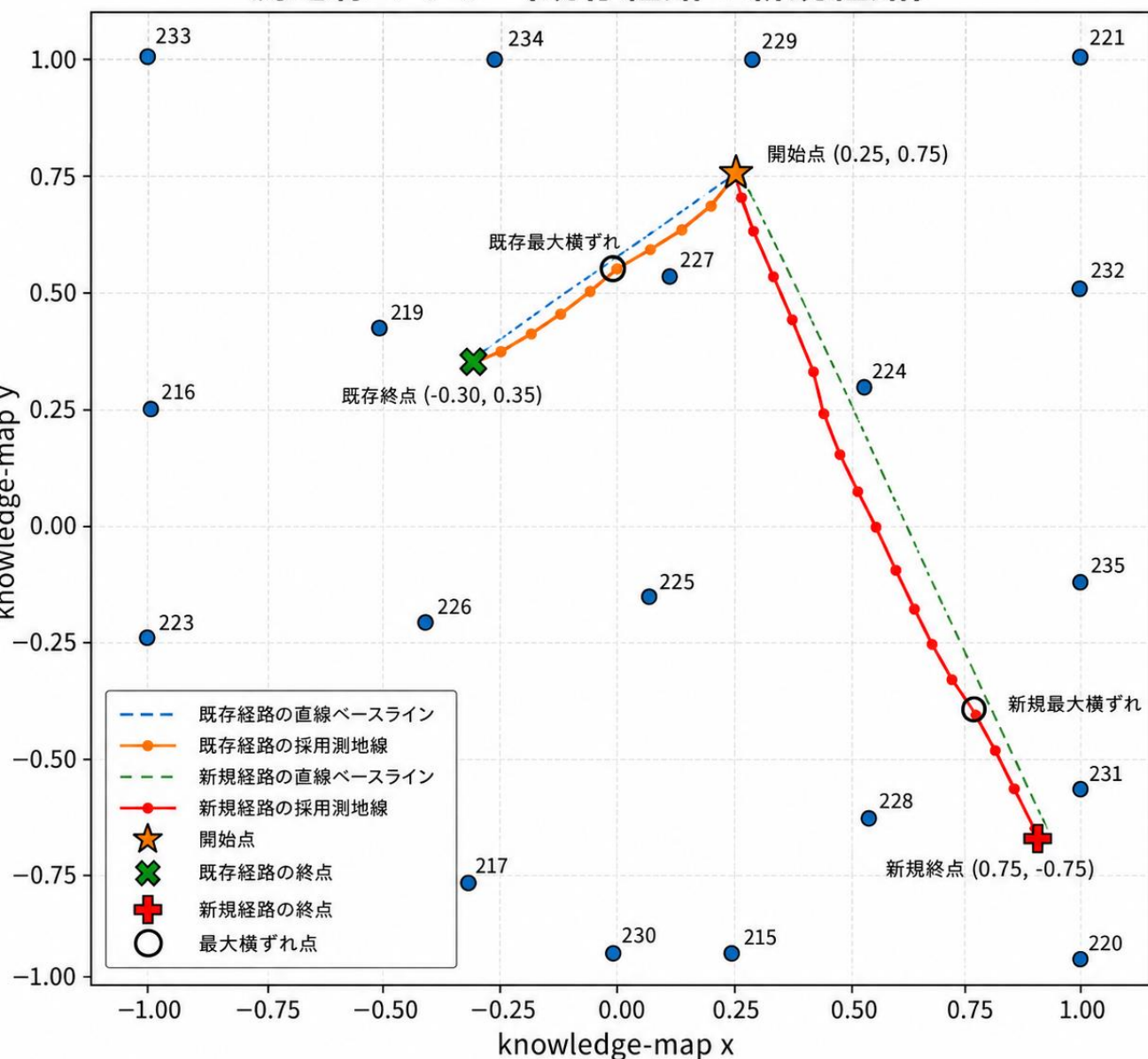
方向	数学的定義	日本語での言語化
0度 (x方向)	$+\partial V/\partial x$	樹脂・プロセス・分子/材料側から離れ、航空機翼・空力構造連成・構造最適化へ進む方向。 表をコピーする
45度	$(\partial V/\partial x + \partial V/\partial y)/\sqrt{2}$	損傷・残留応力・破壊力学の詳細から離れ、材料情報を保持しつつ、翼・ブレードの設計最適化へ統合する方向。
90度 (y方向)	$+\partial V/\partial y$	破壊・損傷・繊維/マトリックス局所力学から離れ、ポリマー・樹脂・分子/熱特性と高次構造設計が結合する方向。
方向間の類似度は以下でした。		
比較	cosine類似度	解釈
0度 vs 45度	0.766	45度は0度の構造設計方向を強く含む
45度 vs 90度	0.716	45度は90度の材料・樹脂方向も含む
0度 vs 90度	0.101	0度と90度はほぼ別の知識軸

要するに、0度は構造最適化軸、90度は材料・樹脂/分子設計軸、45度は両者を接続する統合軸です。

知識多様体における最適経路探索(簡単な例__2次元)

8

測地線マップ (既存経路 + 新規経路)



要素

数式

SPH重み

$$w_i = W_i / \sum_j W_j$$

補間知識ベクトル

$$\mathbf{v}(\mathbf{p}) = \text{normalize}(\sum_i w_i \mathbf{x}_i)$$

x方向勾配

$$\frac{\mathbf{v}(x+\delta, y) - \mathbf{v}(x-\delta, y)}{2\delta}$$

y方向勾配

$$\frac{\mathbf{v}(x, y+\delta) - \mathbf{v}(x, y-\delta)}{2\delta}$$

誘導計量

$$g_{ij} = \langle \partial_i \mathbf{v}, \partial_j \mathbf{v} \rangle$$

区間エネルギー

$$\Delta \mathbf{p}_k^T g(\mathbf{m}_k) \Delta \mathbf{p}_k$$

経路エネルギー

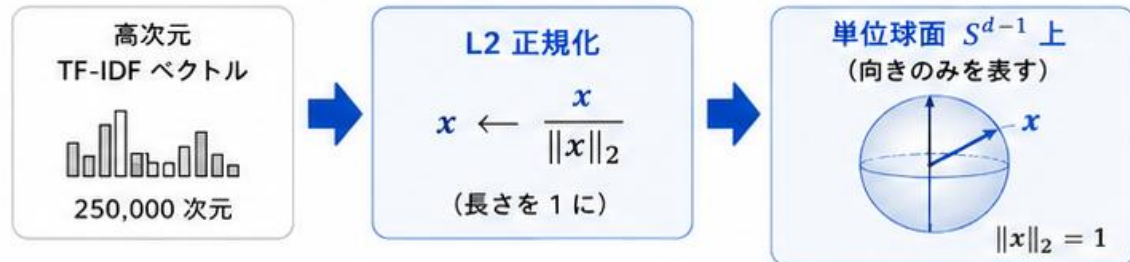
$$E = \sum_k \Delta \mathbf{p}_k^T g(\mathbf{m}_k) \Delta \mathbf{p}_k$$

リーマン長

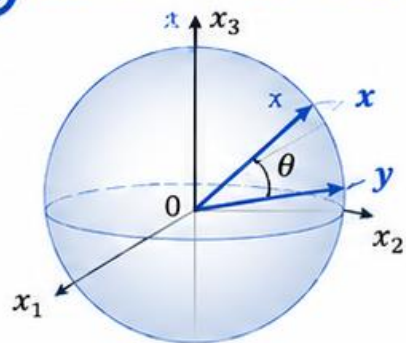
$$L = \sum_k \sqrt{\Delta \mathbf{p}_k^T g(\mathbf{m}_k) \Delta \mathbf{p}_k}$$

L2 正規化（ユークリッドノルムで1にする）

= ベクトルの「向き」を比較する幾何（単位球面上の幾何）



幾何のイメージ（例：3 次元の場合）



比較するもの

ベクトルの向き（角度）
= トピックの「方向性」

主な距離・類似度

- コサイン類似度
$$\cos \theta = \frac{x \cdot y}{\|x\|_2 \|y\|_2}$$
- ユークリッド距離（球面上）
- 角距離（ θ ）



解釈

「どの特徴（単語）の組合せの方向が似ているか」を見る
(絶対的な強さの違いには鈍感)

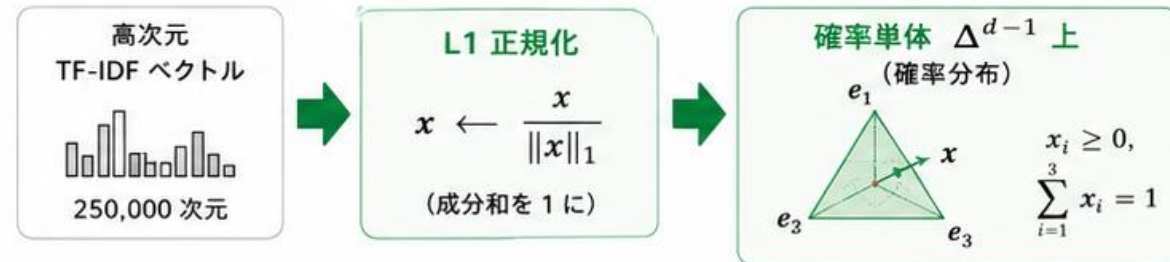


使い分けの目安

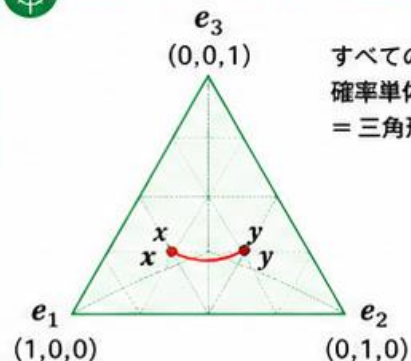
- トピックの方向性・類似トピック探索
- L2（コサイン）が主流

L1 正規化（要素の絶対値の和で1にする）

= 特徴量上の「確率分布」とみなす幾何（情報幾何学）



幾何のイメージ（例：3 次元の場合）



すべての点が
確率単体（シンプレックス）上にある
= 三角形の内部（3 次元の場合）

比較するもの

特徴（単語）の分布の違い
= 確率分布の「形の違い」

主な距離・類似度（情報幾何学）

- KL ダイバージェンス
$$D_{KL}(x \| y) = \sum_i x_i \log \frac{x_i}{y_i}$$
- Jensen-Shannon 距離
- Hellinger 距離
- Fisher-Rao 計量（リーマン計量）



解釈

「どの特徴の割合（確率）がどのように違うか」を見る
(全体の配分の変化に敏感)



使い分けの目安

- 特徴の配分・構成の違い・生成モデル的解釈
- L1（情報幾何学）が有効