文埋め込みに対するグラフベースクラスタリング：適応的局所スケーリングと共有近傍の統合

新妻巧朗 [1]　中分遥 [2]　田森秀明 [1]　吉田光男 [3]
[1] 朝日新聞社メディア研究開発センター　[2] 北陸先端科学技術大学院大学 先端技術研究科
[3] 筑波大学ビジネスサイエンス系
{niitsuma-t,tamori-h}@asahi.com　nakawake@jaist.ac.jp
mitsuo@gssm.otsuka.tsukuba.ac.jp

概要

文埋め込みのクラスタリングでは，高次元空間におけるハブ現象や文脈の連続性が構造抽出の課題となる．本研究では，バイオインフォマティクスに着想を得て，適応的局所スケーリングと共有近傍（SNN）を統合したグラフベース手法を提案する．本手法は，近似近傍探索と GPU 実装により，数千万件規模のデータに対してもスケーラブルに動作する．実験の結果，既存の密度ベース手法や Mini Batch K-means と比較して，連続的な意味構造を持ちうるデータセットにおいて高い品質を達成し，大規模実務データにおいても有効性を実証した．

1　はじめに

大規模なテキストデータセットから教師なしに意味的な構造を抽出するトピック分析は，計算社会科学，実務的なデータ分析において不可欠な技術である．近年では，Transformer ベースの事前学習済みモデルによる高品質な文埋め込み（sentence embedding）[1] が利用可能となり，これらをクラスタリングするアプローチが標準化している．代表的なフレームワークである BERTopic [2] では，UMAP [3] による次元削減と HDBSCAN [4] による密度ベースクラスタリングが採用されている．

しかし，これらの既存手法は，トピックが離散的な「島」として存在することを暗黙に仮定しており，実際のデータ構造と乖離する場合がある．現実のテキストの文脈はグラデーションのように連続的に移り変わることも多く，球状クラスタを仮定する K-means や，密度差による明確な分離を前提とする HDBSCAN を適用すると，本来は連続している文脈が過度に分断されたり，話題が移り変わる領域がノイズとして大量に棄却されたりする問題が生じる．加えて，高次元ベクトル空間では特定のデータ点が多数の近傍となるハブ現象 [5] が生じやすく，距離に基づく単純な構造把握を困難にしている．

この「高次元空間における連続的な構造の抽出」という課題に対し，本研究ではバイオインフォマティクス分野の知見に着目する．同分野のシングルセル解析 [6] では，細胞の分化過程のような連続的な多様体構造を捉えるため，次元削減後の空間で密度クラスタリングを行うのではなく，高次元データから直接構築した近傍グラフに対してコミュニティ検出を行うアプローチが標準的である．本研究はこの「グラフ構築とコミュニティ検出に基づく枠組み」をテキスト分析へ導入し，適応的局所スケーリングと共有近傍（Shared Nearest Neighbor, SNN）を統合することで，密度の不均一性とハブ現象にロバストなトピック抽出を実現する．

本研究では，公開ベンチマーク（MTEB）および数千万件規模の実データ（Amazon の商品説明）による実験を行った．その結果，提案手法は，連続性の高いデータセットにおいて K-means や BERTopic アプローチよりも高いクラスタ品質を達成し，かつ大規模データに対しても実用的なスケーラビリティを持つことを実証した．

2　関連研究

テキストのトピック抽出は，LDA 等の確率モデルから，高品質な文埋め込み [1] を活用したクラスタリングへと移行している．BERTopic はその代表格であり，次元圧縮と密度クラスタリングを組み合わせることで高いコヒーレンスを実現した．しかし，密度ベース手法は空間密度の不均一性やハブ現象に脆弱であり，連続的な文脈を持つデータではクラス

タが不安定化する課題がある．

これに対し，テキストをグラフ構造と捉え，コミュニティ検出を適用する研究が注目されている．hSBM [7] や ComTM [8] は単語共起グラフを用いるが，近年では文埋め込み間の類似度グラフを用いる手法 [9, 10] も提案されている．これらは意味的な一貫性を捉える上で有効だが，全ペア計算を前提とするものが多く，計算量が $O(N^2)$ となるため数万件以上のデータへの適用が困難であった．本研究では，近似近傍探索（NN-Descent [11]）により計算量を約 $O(N^{1.14})$ に抑えつつ，適応的局所スケーリングと SNN を導入することで，数千万件規模のデータに対してもロバストかつ高速な構造抽出を実現する．

3 提案手法

本手法は，(1) 次元圧縮と近似近傍探索，(2) 適応的局所スケーリングと共有近傍（SNN）を統合したグラフ構築，(3) コミュニティ検出，の 3 段階で構成される．

3.1 次元圧縮と近似近傍探索

大規模データの処理においてメモリ効率と計算速度を両立するため，まず Incremental PCA [12] を用いて文埋め込みを低次元へ射影する．続いて，近似近傍探索手法 NN-Descent [11] を用い，各データ点 x_i の k 近傍集合 $\mathcal{N}_k(i)$ を高速に特定する．距離尺度にはコサイン距離を用いる．

3.2 適応的重み付けによるグラフ構築

密度の不均一性とハブ現象に対処するため，局所的な密度情報と大域的な構造情報を統合してエッジ重みを決定する．

適応的局所スケーリング　各データ点 x_i 周辺の密度が異なれば，同じ距離でも意味的な親密さは異なる．そこで，Self-tuning spectral clustering [13] のアプローチを拡張し，密度適応的な接続確率 $p_{j|i}$ を定義する．本手法では，外れ値へのロバスト性を高めるため，単純な k 近傍距離ではなく，最近傍点との距離ギャップの分位点（Quantile）に基づいて局所スケール σ_i を決定する．具体的には，近傍点 $x_k \in \mathcal{N}_k(i)$ との距離ギャップ $\Delta_{ik} = d(x_i, x_k) - \rho_i$（ただし $\rho_i = \min_{j \neq i} d(x_i, x_j)$）を計算し，$\sigma_i$ をその分布の q-分位点とする．

$$\sigma_i = \mathrm{Quantile}_q(\{\Delta_{ik} \mid x_k \in \mathcal{N}_k(i)\}) \tag{1}$$

本手法では，$q = 0.3$ に固定して実験をしている．この操作は，各データ点周辺の密度分布に基づいて距離を局所的に標準化することに相当し，密度の異なる領域間での接続強度を公平に扱うことを可能にする．この σ_i を用いて，接続確率は以下のように定義される．

$$p_{j|i} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \tag{2}$$

これにより，局所的な疎密に適応しつつ，極端に遠い近傍点の影響を排除した重み付けが可能となる．さらに，UMAP [3] の定式化に倣い，以下のファジー集合和により，双方からの接続確率を統合しつつエッジを対称化する．

$$w_{ij}^{fuzzy} = p_{j|i} + p_{i|j} - p_{j|i}p_{i|j} \tag{3}$$

共有近傍（SNN）と指標の統合　高次元空間特有のハブ現象を抑制するため，構造的類似度として SNN を導入する．具体的には，近傍集合間の Jaccard 係数を算出する．

$$w_{ij}^{snn} = \frac{|\mathcal{N}_k(i) \cap \mathcal{N}_k(j)|}{|\mathcal{N}_k(i) \cup \mathcal{N}_k(j)|} \tag{4}$$

最終的なエッジ重み w_{ij}^{final} は，密度情報 w_{ij}^{fuzzy} と構造情報 w_{ij}^{snn} を，ハイパーパラメータ α $(0 \leq \alpha \leq 1)$ を用いて統合する．

$$w_{ij}^{final} = \alpha \cdot w_{ij}^{fuzzy} + (1 - \alpha) \cdot w_{ij}^{snn} \tag{5}$$

この α により，密度の連続性を重視するか，トポロジーの堅牢性を重視するかのトレードオフを制御可能となる．

3.3 コミュニティ検出

構築されたグラフに対し，Leiden 法 [14] および Louvain 法 [15] を適用しクラスタを抽出する．

4 実験設定

4.1 データセット

本研究では，提案手法の有効性と汎化性能を検証するために，ドメインや規模の異なる 3 種類のデータセットを使用した．

Yahoo Answers Topics　MTEB ベンチマーク [16] に含まれる文書分類データセットである．前述の通り，本データセットのみを開発セット（Dev）と評価セット（Test）に 0.05 : 0.95 で分割し，開発セットを用いてハイパーパラメータの調整を行った．

表 1　使用したデータセットの統計情報

Dataset	# Dev	# Test	# Class
Yahoo Answers Topics	67,055	1,274,056	10
ArxivClusteringP2P	-	732,723	21
Amazon Products 2023	-	29,313,335	33

ArxivClusteringP2P　MTEB ベンチマークに含まれる，学術論文のタイトルとアブストラクトを用いたクラスタリングタスクである．本研究では MTEB により提供されるデータをそのまま使用し，全件を評価用データ（Test）として扱った．

Amazon Products 2023　大規模な Amazon の商品レビュー・メタデータセット [17][1)]である *Amazon Reviews 2023* から抽出したデータセットである．本研究では，商品（Item）のクラスタリングを行うため，メタデータに含まれる *Title* と *Description*，およびカテゴリラベルに欠損がないデータのみを抽出し，これを *Amazon Products 2023* とした．このフィルタリングの結果得られた 29,313,335 件のデータを，学習や調整には用いず，全て評価用データ（Test）として使用することで，大規模データに対するスケーラビリティと汎化性能を評価した．

4.2　テキスト埋め込みモデル

本研究は Sentence-Transformers [1] によってテキストのベクトル化をしており，文埋め込みモデルには `sentence-transformers/all-MiniLM-L6-v2` を使用した．

4.3　評価プロトコル

クラスタリング手法のハイパーパラメータ（クラスタ数 K の設定や密度パラメータ等）は，Yahoo Answers Topics の開発セット（Dev）において最適化された値を固定して使用し，ArxivClusteringP2P および Amazon Products 2023 に対してはパラメータ調整なし（Zero-shot 設定）で適用した．

4.4　評価指標

クラスタリングの品質評価には，調整相互情報量（AMI）[18] および調整ランド指数（ARI）[19] を採用した．AMI は，予測結果と正解クラスの間の相互情報量をエントロピーで正規化する正規化相互情報量（NMI) [20] から偶然による一致の影響で補正した指標である．ARI はペアの整合性を測る指標であり，共に $[-1, 1]$ の範囲を取り，正解と完全に一致した場合に 1 となる．参考値として NMI も併記する．

4.5　ハイパーパラメータの多目的最適化

本研究では，クラスタリングの品質と安定性を両立するため，Optuna [21] を用いた多目的最適化を行った．Yahoo Answers Topics の開発セットに対し，以下の 2 指標を最大化した．

AMI の期待値　データを割合 0.8 でサブサンプリングしたデータセットに対するクラスタリングの AMI（5 回のサブサンプリングに対する平均）．

安定性（Stability）　上記 5 試行のサブサンプリング結果間における，ペアワイズ ARI の平均値．

これにより，100 回の試行から得られたパレート最適解の中から，最も AMI が高い解を採用した．

5　実験結果

表 2 に各データセットにおける評価結果を示す．提案手法（Inc.PCA/Raw + Leiden/Louvain）と，比較手法について，AMI，ARI，NMI および検出されたクラスタ数，実行時間を比較した．Raw 設定は前処理なしを表し，またすべての手法に共通している NN-descent と適応的重み付けおよび SNN は表記から省略した．

実験結果の概要　Yahoo Answers Topics において，提案手法（Inc.PCA + Leiden）は AMI 0.352 を達成し，K-means (0.299) を上回った．特筆すべきは UMAP + HDBSCAN の結果であり，AMI が 0.119 と著しく低い値にとどまった．これは，HDBSCAN の結果の約 87%のデータがノイズとして棄却され，正解ラベルとの整合性が取れなかったためと考えられる．

大規模データにおける粒度と表現力　約 3,400 万件の Amazon Prodcuts において，提案手法は K-means を大きく上回る性能を示した．特に **Raw + Louvain** 設定は全手法中最高の AMI 0.482 を記録し，494 個のクラスタを検出した．一方，Inc.PCA + Louvain は AMI 0.476 と肉薄しつつ，最高の ARI 0.296 と 1,338 個の詳細なクラスタを検出している．これらの結果は，提案手法がハブ現象の影響を抑制し，商品のサブカテゴリのような詳細な意味構造を捉えられていることを示唆している．また，計算リソースが許すならば Raw データを用いたグラフ構築が情報の損失を防ぎ，最も高い AMI をもたらす可能性があることが確認された．

1）https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

表 2　各データセットにおける性能比較．提案手法は，同一の次元圧縮条件（Inc.PCA）を適用した K-means と比較しても，AMI および ARI において優れた性能を示している．特に大規模データ（Amazon Products）において，Raw + Louvain 設定が最高の AMI を達成している．

Dataset	Method	AMI	ARI	NMI	Clusters	Time(s)
Yahoo Answers Topics	**Proposed (Inc.PCA + Leiden)**	**0.352**	**0.306**	**0.352**	17	77.9
	Proposed (Inc.PCA + Louvain)	0.346	0.286	0.346	14	124.6
	Proposed (Raw + Leiden)	0.327	0.283	0.327	19	64.7
	Proposed (Raw + Louvain)	0.324	0.251	0.324	30	126.2
	MiniBatch K-means (Inc.PCA)	0.327	0.216	0.327	17	54.9
	MiniBatch K-means (Raw)	0.299	0.191	0.299	18	45.4
	UMAP + HDBSCAN	0.119	0.001	0.120	298	18084.4
Arxiv ClusteringP2P	**Proposed (Inc.PCA + Leiden)**	**0.469**	0.220	**0.469**	18	47.0
	Proposed (Inc.PCA + Louvain)	0.456	0.226	0.456	12	60.5
	Proposed (Raw + Leiden)	0.458	**0.235**	0.458	14	45.2
	Proposed (Raw + Louvain)	0.462	0.230	0.463	15	70.4
	MiniBatch K-means (Inc.PCA)	0.448	0.224	0.448	17	19.1
	MiniBatch K-means (Raw)	0.454	0.216	0.454	18	9.8
	UMAP + HDBSCAN	0.136	0.001	0.142	185	5938.7
Amazon Products 2023 (34M)	Proposed (Inc.PCA + Leiden)	0.472	0.237	0.472	819	1484.5
	Proposed (Inc.PCA + Louvain)	0.476	**0.296**	0.476	1338	2850.9
	Proposed (Raw + Leiden)	0.477	0.245	0.477	157	1132.9
	Proposed (Raw + Louvain)	**0.482**	0.271	**0.482**	494	1825.7
	MiniBatch K-means (Inc.PCA)	0.426	0.264	0.426	17	699.5
	MiniBatch K-means (Raw)	0.424	0.282	0.424	18	638.0
	UMAP + HDBSCAN	(Time limit exceeded)				

計算効率とスケーラビリティ　表 2 の Time(s) が示すように，提案手法は計算効率においても優れている．特に既存手法 UMAP + HDBSCAN との比較では，Yahoo Answers Topics において，HDBSCAN は処理に約 5 時間（18,084 秒）を要したが，提案手法（Raw + Leiden）は約 1 分（64.7 秒）で完了しており，約 280 倍の高速化を実現している．

また，Amazon Products（約 2,900 万件）において，提案手法（Raw + Leiden）は約 19 分（1,132 秒）で処理を完了した．興味深いことに，次元圧縮を行わない Raw 設定の方が，Inc.PCA 設定よりも高速な傾向が見られた（例：Amazon Products において Inc.PCA+Leiden は 1,484 秒）．アルゴリズム間の比較では，Leiden 法が Louvain 法よりも一貫して高速（Amazon Products かつ Raw 設定で約 1.6 倍）であり，大規模データにおいて品質と速度のバランスに優れた選択肢であることが確認された．

6　おわりに

本研究では，大規模かつ高次元な文埋め込みデータセットに対し，意味的な連続性と密度の不均一性を考慮したロバストなトピック抽出を行うためのグラフベースクラスタリング手法を提案した．提案手法は，近似近傍探索によるグラフ構築を基盤とし，各データ点の局所的な密度に応じた適応的スケーリングと，構造的類似度である共有近傍（SNN）を統合することで，高次元空間におけるハブ現象やノイズの影響を効果的に低減した．

MTEB ベンチマークおよび 2,900 万件の実データを用いた実験の結果，提案手法は以下の優位性を示した．第一に，トピックが連続的に分布するデータにおいて，従来の密度ベース手法のようにデータを過度に棄却することなく構造を抽出できた点である．第二に，K-means と比較して，ロングテールを含む実データの多様なトピック分布を，より詳細な粒度で捉えられた点である．第三に，GPU による並列処理を活用することで，数千万件規模のデータセットに対しても実用的な時間内で解析可能なスケーラビリティを実証した点である．

今後の課題として，ハイパーパラメータ（特に構造的類似度と密度情報のバランス係数 α や 最近傍と距離ギャップの分位点 q）の自動調整手法の確立が挙げられる．また，Leiden 法の特性を活かし，トピックの階層構造を抽出・可視化するフレームワークへの拡張も，探索的データ分析の観点から有用であると考えられる．他にもトピックモデルとしての評価（例えば，トピックごとに算出した top-k 単語に対する NPMI など）も必要だろう．本手法が，大規模テキストデータに埋め込まれた複雑な意味構造を解き明かすための一助となることを期待する．

参考文献

[1] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3980–3990, 2019.

[2] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[3] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. **Journal of Open Source Software**, Vol. 3, No. 29, p. 861, 2018.

[4] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. **Density-Based Clustering Based on Hierarchical Density Estimates**, pp. 160–172. Springer Berlin Heidelberg, 2013.

[5] Miloš Radovanovi, Alexandros Nanopoulos, and Mirjana Ivanovi263;. Hubs in space: Popular nearest neighbors in high-dimensional data. **Journal of Machine Learning Research**, Vol. 11, No. 86, pp. 2487–2531, 2010.

[6] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. **Nature Biotechnology**, Vol. 33, No. 5, pp. 495–502, 2015.

[7] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. A network approach to topic models. **Science Advances**, Vol. 4, No. 7, 2018.

[8] Mahfuzur Rahman Chowdhury, Intesur Ahmed, Farig Sadeque, and Muhammad Yanhaona. Topic modeling using community detection on a word association graph. In **Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing**, pp. 908–917, 2023.

[9] Nicolas Stefanovitch, Guillaume Jacquet, and Bertrand De Longueville. Graph and embedding based approach for text clustering: Topic detection in a large multilingual public consultation. In **Companion Proceedings of the ACM Web Conference 2023**, pp. 694–700, 2023.

[10] Venkatesh Bollineni, Igor Crk, and Eren Gultepe. Mapping hymns and organizing concepts in the rigveda: Quantitatively connecting the vedic suktas. In **Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities**, pp. 514–523, 2025.

[11] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In **Proceedings of the 20th international conference on World wide web**, pp. 577–586, 2011.

[12] David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. **International Journal of Computer Vision**, Vol. 77, No. 1–3, pp. 125–141, 2007.

[13] Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In **Advances in Neural Information Processing Systems**, 2004.

[14] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. **Scientific Reports**, Vol. 9, No. 1, 2019.

[15] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, Vol. 2008, No. 10, p. P10008, 2008.

[16] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, 2023.

[17] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation, 2024.

[18] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. **J. Mach. Learn. Res.**, Vol. 11, pp. 2837–2854, 2010.

[19] Lawrence Hubert and Phipps Arabie. Comparing partitions. **Journal of Classification**, Vol. 2, No. 1, pp. 193–218, 1985.

[20] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. **J. Mach. Learn. Res.**, Vol. 3, No. null, pp. 583–617, 2003.

[21] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining**, pp. 2623–2631, 2019.