F-022

入力ベクトル疎密がデンドログラム特徴量へ与える影響分析

方　越洋 †　　斉藤　和巳 †
† 神奈川大学大学院 理学研究科

1　はじめに

大規模文書データの階層的クラスタリングは情報抽出の基盤技術として重要である．特に，文書埋め込みモデルの増加に沿って，異なる文書ベクトルの処理手法を選ぶことは課題になる．著者らは従来のけんきゅ [1] において，分割法と凝集法を統合したハイブリッド手法（avgd/ward/cosd）を提案した．しかし，入力ベクトルの疎密特性がデンドログラム構造に与える影響は未解明であった．本研究ではこの課題を解決するため，以下の新規分析法を提案する：

第一に, TF-IDF 疎ベクトルの初期分割型クラスタリングを実施し, 得られたクラスタリング結果と TF-IDF ベクトルを凝集型階層的クラスタリングアルゴリズムに入力することで, デンドログラムを構築する. 第二で, GloVe 埋め込み [2] により高次元疎ベクトルを 300 次元密ベクトルに写像し, 同一初期クラスタリング結果で, 写像したベクトルを同一の階層的クラスタリングプロセスに統合することで, デンドログラムを構築する. 最終的に, 正規化相互情報量（NMI） [3] という評価指標を導入し, 両クラスタリング結果の類似性を評価する定量的比較分析を実施する. 同時に，ベースラインとしてランダム併合法（rand）を用いる．

2　提案手法

文書数 N，初期クラスタ数 k で初期クラスタ集合 $\{C_1, ..., C_k\}$ を生成する．クラスタ j のセントロイドは $\mathbf{y}_j$ で表す．各表現 $r \in \{疎, 密\}$ と手法 $m \in \{avgd, ward, cosd, rand\}$ の組合せに対し，クラスタ数が j になるまで併合したクラスタ集合を $C_j^{(r,m)}$ 求める．

比較対象のクラスタ集合 $\mathcal{P}_j^{(r_1,m_1)} = C_1^{(r_1,m_1)}, ..., C_j^{(r_j,m_j)}$ に対し，クラスタ間の相互情報量 $I(\mathcal{P}_j^{(r_1,m_1)}; \mathcal{P}_j^{(r_2,m_2)})$

$$I(\mathcal{P}_j^{(r_1,m_1)}; \mathcal{P}_j^{(r_2,m_2)}) = \sum_{x \in \mathcal{P}_j^{(r_1,m_1)}} \sum_{y \in \mathcal{P}_j^{(r_2,m_2)}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

とエントロピー $H(\mathcal{P}_j^{(r_1,m_1)})$

$$H(\mathcal{P}_j^{(r_1,m_1)}) = -\sum_{x \in \mathcal{P}_j^{(r_1,m_1)}} p(x) \log p(x) \quad (2)$$

を定義すると, 同一手法で異なる表現や、同一表現で異なる手法での併合結果での一致度を NMI で評価できる．

$$\mathrm{NMI}\left(\mathcal{P}_j^{(r_1,m_1)}, \mathcal{P}_j^{(r_2,m_2)}\right) = \frac{2I(\mathcal{P}_j^{(r_1,m_1)}; \mathcal{P}_j^{(r_2,m_2)})}{H(\mathcal{P}_j^{(r_1,m_1)}) + H(\mathcal{P}_j^{(r_2,m_2)})} \quad (3)$$

ここで $p(x) = \frac{|x|}{N}$, $p(x) = \frac{|x \cap y|}{N}$ はクラスタサイズ比率である．

一方，初期クラスタからクラスタ数が j になるまで併合する際に，クラスタ間距離を $d_*(i,j)$ を定義すると，各手法の距離計算は

- 群平均法（$avgd$）：$d_a(i,j) = 2(1 - \mathbf{y}_i^T \mathbf{y}_j)$
- Ward 法 ($ward$)：$d_w(i,j) = |C_i|\|\mathbf{y}_i\|^2 + |C_j|\|\mathbf{y}_j\|^2 - (|C_i| + |C_j|)\|\mathbf{y}_{i\cup j}\|^2$
- コサイン法 ($cosd$)：$d_c(i,j) = |C_i|\|\mathbf{y}_i\| + |C_j|\|\mathbf{y}_j\| - (|C_i| + |C_j|)\|\mathbf{y}_{i\cup j}\|$

になる．さらに，Rand 法はランダムでクラスタペアを選択する．

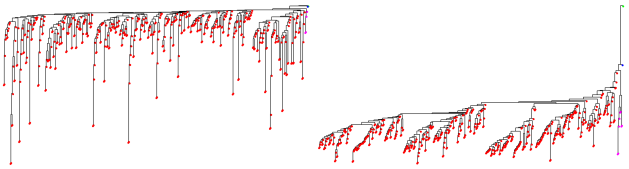
3　実験による評価

本実験では, 文書数 $N = 299,752$ で, 語彙数 $M = 102,660$ の New York Times news articles(NYT) データと文書数 $N = 8,200,000$ で, 語彙数 $M = 141,043$ の PubMed データを利用した．

図 1, 2, 3 に, NYT 結果の疎ベクトルと写像した密ベクトルの 3 つ手法でのデンドログラム可視化結果の例である. ノード配色については，4 つのクラスタとなるようにカットオフを設定し，それぞれのリーフに対し，赤，青，緑，紫の 4 色を割り当てた．

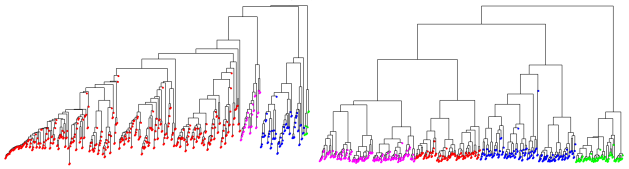
NYT の 3 つ図を横断的に比較すると, 同一の文書データに対して同じ手法を用いた場合でも, 異なるベクトル表現によって構築されたデンドログラムに差異が認められる. コサイン法では両方ベクトル表現間に顕著な相違は見られなかった. ウォード法で密ベクトルはより優れたデンドログラム特性を示す. 一方，群平均法は文書ベクトルの疎密表現にかかわらず，デンドログラムが偏りが高い傾向が見て取れる．これに対して, 縦断比較では, いずれの手法においても，コサイン法は比較的平衡の取れたデンドログラムを生成することが確認された．

†Yueyan FANG　†Kazumi SAITO
†Kanagawa University



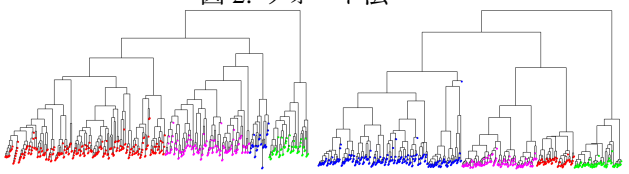
疎ベクトル　　密ベクトル

図 1: 群平均法



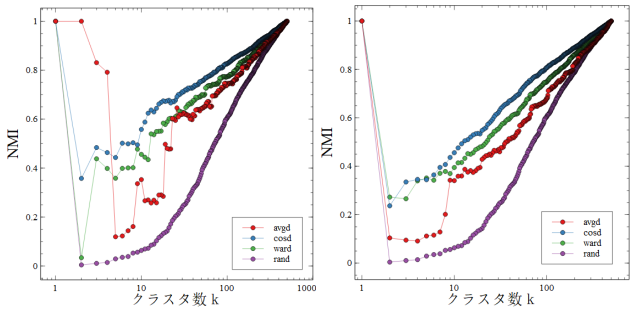
疎ベクトル　　密ベクトル

図 2: ウォード法



疎ベクトル　　密ベクトル

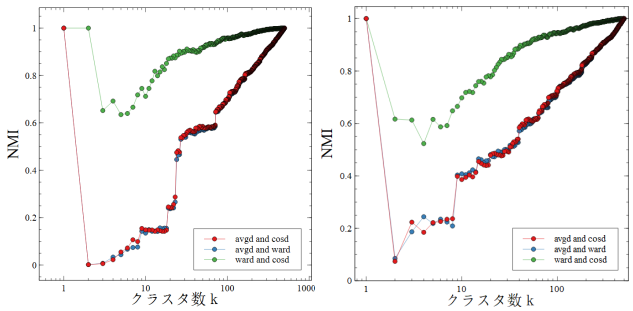
図 3: コサイン法



NYT　　PubMed

図 4: 疎ベクトルと密ベクトルの NMI

図 4 は，疎密ベクトルでクラスタを 500 から 1 まで変化させた際の毎回併合後を NMI を計算し，併合過程における NMI の値を示している．$k = 500$ の場合，初期クラスタは同一の球面 k -means クラスタリングに由来するため，完全に一致する．$k = 1$ の場合，全てのサンプルは単一のクラスタに集約される．クラスタ数が中程度の範囲においては，NMI が低下するが，ランダムに併合する rand によるデンドログラムと比較して，群平均法, Ward 法, コサイン法の NMI は十分に高く，特に，コサイン法が安定して最も良い性能を示した．よって，本実験で，文書ベクトルの疎密表現にかかわらず，コサイン法を用いれば，クラスタ中身の一致傾向が最も高いことを確認した．さらに GloVe は Wikipedia データで単語埋め込んでいることから，PubMed より NYT データとの親和性が高く，結果として意味的関連性の精度が向上する可能性がある．



NYT　　PubMed

図 5: 密ベクトル内部の NMI

図 5 は密ベクトル内部三つ手法の NMI の値を示している．この結果より，いずれの文書データにおいても，コサイン法と Ward 法の NMI は，群平均法との比較結果を大幅に上回った．またデンドログラムからも明確に観察されるように，両手法は類似したより均衡性の高いクラスタ構造を形成する傾向が確認された．

4　おわりに

本研究では，同一データを基に球面 k-means 法で同一クラスタラベルを構築し，疎密ベクトルを用いた凝集型クラスタリングでデンドログラムを構築した．結果から見ると，階層的クラスタリングの併合過程において，入力ベクトル空間の相違により距離計算と併合順序に差異が生じるが，本実験で，コサイン法は 2 つベクトル表現に対して比較的安定し，2 つ表現手法の差異を超えて一貫したクラスタ構造を抽出可能にすることを示す．

今後の研究課題として，より多様なデータ表現とクラスタリング手法を導入し，異なるデータ形式下におけるハイブリッドクラスタリングの汎用性と実用性を解明することが挙げられる．

参考文献

[1] 方 越洋. 斉藤和巳. コサイン類似度に基づく分割型と凝集型ハイブリッド文書クラスタリング法の高速化と空間拡散性の評価. 論文誌数理モデル化と応用（TOM71）, 2025. 掲載予定

[2] Jeffrey Pennington , Richard Socher , Christopher D. Manning. GloVe: Global Vectors for Word Representation Empirical Methods in Natural Language Processing (EMNLP) .pp.1532–1543.2014

[3] Witten, Ian H. Frank, Eibe . Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Amsterdam. ISBN 978-0-12-374856-0.2005