2P-7

NMFによる文書分類精度の改善のための一考察

立川華代 † 小林一郎 †
† お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース

1 はじめに

大量の文書情報の電子化により，テキストデータを処理する機会が増え，昨今，特に文書中の潜在情報を抽出しテキストデータを効率良く扱うテキスト処理手法として，LSI，pLSI，LDA といった様々な手法が提案され，その有用性が示されている．それら手法と同様に文書内の潜在トピックを抽出する手法として，非負値行列因子分解 (NMF:Non-negative Matrix Factorization)[1] がある．NMF は，非負値を要素とするデータを 2 つの非負値の行列に分解し文書のトピック抽出とトピックの下での分類を同時に可能とする．LSI などにおいては正規直交系の座標変換により文書内の主たる成分を抽出する手法であり，そのため分解した行列内に意味を理解しにくい負の数値が出現してしまったが，本来，対象とするデータの性質において負の数値が意味をなさない場合も多く，NMF による非負の行列因子分解手法が注目されている．しかし，NMF は行列因子に分解する際，目的関数の設定および行列の初期値が結果に大きな影響を与えることが知られている．そこで，本研究ではトピックを形成すると考える単語群を文書中から抽出し，それらを制約知識として初期値に与えることで分類精度の向上を検討する．

2 関連研究

NMF による行列因子分解の精度改善の取り組みとして，丸田ら [4] は，因子分解対象となる行列 V における文書ベクトル同士の類似性と NMF により因子分解された行列の内，トピックに対する文書の分類を表す行列 H の類似性を反映するように，目的関数に正規化項として類似行列項を追加し，NMF の行列因子分解の精度向上を示している．また，山口ら [5] は初期値に着目し，行列を特異値分解して得られた直交行列の要素の絶対値を NMF の初期値として利用した．彼らの手法において，手法に対する理論的な説明はなされていないが精度向上が実現されていることを示している．このように NMF の精度を上げる手段はいくつか挙げられるが，本研究では目的関数の改善ではなく初期値の改善に注目し，文書分類の精度向上のための手法を提案する．

3 非負値行列因子分解

NMF は非負値で与えられた行列のデータ V を二つの非負値の行列 W,H に以下に示すように近似的に分解する

$$V \approx WH \tag{1}$$

これにより，次元削減やトピック抽出が可能となる．ここで V は語彙数 n，文書数 m の $n \times m$ の文書の行列，W は語彙数 n，トピック数 r の $n \times r$ の行列，そして H はトピック数 r，文書数 m の $r \times m$ の行列である．W と H の初期値は通常，乱数で与え，式 (2) で与えられるフロベニウスノルムを最小にすることで W, H を V に近似する．

$$J = \|V - WH\|^2 \tag{2}$$

その際，W と H の更新式は以下のようになる．以下において k はトピックを指す．

$$w_{ik} \leftarrow w_{ik} \times \frac{\sum_j v_{ij} h_{kj}}{\sum_j \sum_l (w_{il} h_{lj}) h_{kj}} \tag{3}$$

$$h_{kj} \leftarrow h_{kj} \times \frac{\sum_i v_{ij} h_{ik}}{\sum_i \sum_l (w_{il} h_{lj}) w_{ik}} \tag{4}$$

4 提案手法

本研究では，NMF の因子分解において得られる行列 W の初期値を，分解対象となる文書群の中からトピックを構成すると考えらる単語群から構築される制約知識として与え，因子分解の精度向上に取り組む．ここで，潜在的なトピックは，単語間の共起関係から派生する意味情報であると認識されており，実際に，Newman ら [2] や Stevens ら [3] によって，語の共起関係と潜在トピックごとの語彙に相関関係があることが示されている．このことから，生成される行列 W の初期値が文書内のトピックを反映するように重要語と共起する単語から構成される初期行列を生成する．

4.1 制約知識の構築

制約知識を構築する際に tf-idf と自己相互情報量 (PMI:Pointwise Mutual Information) に着目し，以下の手続きによりトピック行列 W を形成する語彙群を決定する．

A Study on Improving the Accuracy of Document Clustering using NMF

† Kayo TATSUKAWA(tatsukawa.kayo@is.ocha.ac.jp)
†† Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Advanced Sciences, Graduated School of Humanities and Sciences, Ochanomizu University (†)
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Step.1 文書中の重要語の決定

トピックを代表するような語を選択するため，文書中からtf-idf値の高く，かつ，その語の属する文書の語彙数が多いものを抽出し，これらを重要語とする．文書の語彙数が多いものを選ぶ理由は，Step.3において共起する語を追加する際に，共起語が少なくなりトピックを表現する語彙の数が少数になることを防ぐためである．

Step.2 トピックを代表する重要語の決定

Step.1で選んだ重要語のうち互いに同じトピックに入らないと考えられるものをトピックの数だけ選択する必要がある．そこでそれぞれの重要語に対して全語彙とのPMIをはかりPMIベクトルを作成する．このベクトルのサイズは重要語の数 × 全語彙数である．このPMIベクトルの行同士のコサイン類似度を測り，その値が小さいもの，つまり互いに同じ語と共起しないような語を抽出する．ただし，ここで抽出する語の数はトピック数と同じ数である．これにより，各トピック固有の重要単語を抽出する．

Step.3 重要単語に共起する語彙群の決定

Step.2で抽出した語に対して共起する語を追加し，制約知識とする．

4.2 制約知識による文書分類

上記の手続きによって構築した制約知識をNMFの初期値に挿入する．NMFによる因子分解後の行列 W は文書のトピックがどのような語で構成されているのかを表現している．この W は通常0から1の乱数を初期値に設定するが，制約知識となる語を強調するため，該当部分に10をその値として代入する．これによりトピックを表す語の特徴を表すとする．

また，文書のクラスタリングについては行列 H の結果をそのまま利用する．つまり，文書 d_j が割り当てられるクラス c_j は，以下の式で決定される．

$$c_j = \arg\max_k h_{kj} \tag{5}$$

5 実験

5.1 実験仕様

20 Newsgroupsのデータの4トピックから文書を選択し実験を行った．文書の量を変化させ，各トピックから3文書ずつ(計12文書，語彙数:1053)，21文書ずつ(計84文書，語彙数:3794)の2セットを利用した．結果はPurityと相互情報量により評価し比較する．

$$\text{相互情報量} = \Sigma_{i=1}^{K} \frac{C_i}{N} \times (-\Sigma_{n=1}^{K} P(A_h|C_i) \log P(A_h|C_i)) \tag{6}$$

相互情報量は式(6)により求める．ここで $P(A_h|C_i)$ は正解集合 A_h にクラスタ C_i の文書が属する確率である．

5.2 実験結果

実験結果を表1に示す．

表 1: 分類の精度

3文書ずつ	Purity	相互情報量
制約なし	0.633	0.702
制約あり	**0.658**	**0.648**
21文書ずつ	Purity	相互情報量
制約なし	0.465	1.136
制約あり	**0.487**	1.136

5.3 考察

表1より，制約知識を反映させた初期値を利用した方が僅かであるが精度が上がっていることがわかる．

6 おわりに

本研究では，NMFで分解される行列の初期値について，通常乱数であるものに，tf-idfにより重要語を抽出し，さらに抽出された語それぞれに対しPMIが高いものを加えることで構築した制約知識を適用した．その結果，分類精度に大きな向上は見られなかったが，僅かながら向上することを確認した．今後は，制約となる単語数を増やすなどして，より制約知識適用の効果を検討したいと考えている．

参考文献

[1] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pp. 556–562. MIT Press, 2000.

[2] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *HLT: The 2010 North American Chapter of the ACL*, pp. 100–108, 2010.

[3] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961, 2012.

[4] 丸田要, 中村貞吾. 文書類似度を考慮したnmfを用いた記事カテゴリ判定. 情報処理学会研究報告, 2012.

[5] 山口桂吾, 堀田政二, 宮原末治. 非負行列因子分解の初期値設定法とその応用. 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理, Vol. 87, pp. 923–928, mar 2004. 3.