情報処理学会研究報告
IPSJ SIG Technical Report

文書類似度を考慮したNMFを用いた記事カテゴリ判定

丸田　要[†1]　中村貞吾[†1]

Non-negative Matrix Factorization（NMF) は，索引語文書行列を二つの行列の積に分解することで文書クラスタリングを行う．本論文では NMF の精度を高めるために，単語を要素とする文書ベクトルで類似する文書はクラスタの関係度を要素とする文書ベクトルでも類似するという考えに基づき，NMF の目的関数に文書ベクトル同士の類似度行列を追加した手法を提案する．提案する手法では，NMF の目的関数に，入力データの単語を要素とする文書ベクトルの類似度行列とクラスタリングの途中で求まるクラスタの関係度を要素とする文書ベクトルの差のノルムを追加した目的関数を新たに定義する．定義した目的関数を最小にする式を NMF の更新式に適用し，提案する NMF を実行する．提案する手法の効果を示すため，提案する改良 NMF と元の NMF で実際に記事をクラスタリングし比較する．

Category Determination using NMF with Document Similarity

Kaname Maruta[†1] and Teigo Nakamura[†1]

Non-negative matrix factorization (NMF) is a method for decompose matrix of multivariate date and known to be a useful method for document clustering. In this paper, we propose a method for document clustering based on NMF utilizing similarity between document vectors that tries to preserve the document similarity before and after the clustering. We add an similarity matrix term which means difference between the original document vector and resulting reduced dimensional vector. We present some experimental results that show our method improves accuracy of clustering compared with the original NMF.

†1 九州工業大学情報工学部
Kyushu Institute of Technology School of Computer Science and Systems Engineering

1. はじめに

インターネットが普及して以来，多くの人が気軽にインターネットを利用して情報収集することができるようになった．また，文書群である記事もネット上に膨大な量が存在している．つまり，膨大な量の情報を効率よく処理するために精度の良い記事クラスタリングが必要である．

今回は文書クラスタリング手法の一つである NMF を用いる．NMF は，高次元でスパースな文書行列をクラスタリングするのに適している．しかし，従来の NMF では，目的関数の局所最適解が必ずしもクラスタリングの最適解であるとは言えない．つまり，既存の NMF における目的関数を最小にするようなクラスタリング結果では精度の良いクラスタリングを行うことができないと考えられる．

NMF のクラスタリング精度を上げる方法として様々な方法が提案されている．例えば NMF の局所最適解が初期値に依存するという性質を利用し，他のクラスタリング法によりクラスタリングされた結果を NMF の初期値に与える方法である．このような方法ではクラスタリング精度向上に効果があることが確認されている．

しかし，本論文では他のクラスタリングとの併用ではなく，NMF 単体のクラスタリング精度向上を目的とする．

そこで，NMF における目的関数を改良することで，なるべくクラスタリングの最適解に収束するように制御し NMF 単体のクラスタリング精度向上を目指す．ここで改良する概念として類似度の高い文書ベクトル同士は各クラスタとの関連度も類似し同じクラスタに収束するであろうという考えを用いる．具体的には，単語を要素とする文書ベクトル同士の類似度とクラスタと文書の関連度を要素とする文書ベクトル同士の類似度の差のノルムを既存の NMF における目的関数に追加することで，その目的関数が最小になるにつれてクラスタリングの精度がよくなることを目的とする．

2. NMF

2.1 NMF とは

NMF は Non-negative Matrix Factorization の略で，ベクトル空間モデルで表現されたデータを次元縮約することで，クラスタリングを行う．文書クラスタリングにおいて NMF は，(1) 式のように d 個の文書データと w 個の索引語から作られる $d \times w$ の索引語文書行列 X を $w \times k$ の行列 U と $k \times d$ の行列 V^T の積に分解する．ここで，k はクラスタ数である．

$$X = UV^T \tag{1}$$

この時，U は $w \times k$，V は $d \times k$ であり，各要素は非負値となっている．U は索引語とクラスタの関連度を表し，V^T は文書とクラスタの関連度を表している．

NMF では，縮約後の各列が文書のクラスタを表現していると考える．つまり，V^T が索引語文章行列 X の各文書を次元縮約した結果であり，クラスタリングした結果である．具体的には，V^T の h 行目の要素の値が，各文書と h 番目のクラスタとの関連度の大きさを表している．そのため，i 番目での文書データのクラスタは (2) 式で得られる．

$$\arg\max_h \ v_{ih} \tag{2}$$

ここで，v_{ih} は行列 V の i 行 h 列の要素を表す．

2.2 アルゴリズム

まず，索引語文書行列 X を与え，U と V を求める．この時，U と V は (3) 式の J を最小にするような更新式の繰り返しにより求める．

$$J = ||X - UV^T||_F \tag{3}$$

U と V を求める更新式は，(4) 式と (5) 式である．

$$v_{ij} \leftarrow v_{ij} \frac{(X^T U)_{ij}}{(VU^T U)_{ij}} \tag{4}$$

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^T V)_{ij}} \tag{5}$$

ここで u_{ij} と v_{ij} はそれぞれ U と V の i 行 j 列の要素を表し．$(X)_{ij}$ により行列 X の i 行 j 列の要素を表す．求めた V から (2) 式を用いて各文書の所属するクラスタを定める．

また各繰り返しの後に U と V の値が発散するのを防ぐため，U を (6) 式のように正規化する．

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{6}$$

NMF の更新式である (4) 式と (5) 式には初期値 $U^{(0)}$ と $V^{(0)}$ が必要である．本論文では，初期値にランダムな正の値を与える．

2.3 NMF の問題点

- J の局所最適解が必ずしもクラスタリングの最適解であるとは言えない

この問題は，J を最小とする U と V の局所解が必ずしも精度の良いクラスタリング結果であるとは言えないということである．なぜなら，クラスタリング結果が悪い V の局所解でも J を最小にすることができるためである．この問題の解決法の一つとして，初期値の設定をランダムではなく，他のクラスタリング法により整理された初期値を与えることである．これは，U と V の局所解の収束が初期値に大きく依存するためである．しかし，本論文では，NMF 単体の精度向上が目的であるため別の解決法を提案する必要がある．

3. 関連研究 (類似研究との比較)

NMF の精度を向上する手法は多々他の論文で提案されている．

- ピンポン型文書クラスタリングや Hybrid Method

ピンポン型文書クラスタリングも Hybrid Method も NMF の初期値を整理する方法を行っている．その方法は，k-means や pLSI などの他のクラスタリング法で求めたクラスタリング結果を NMF の初期値として与えることでクラスタリングの精度向上を図っている．この方法では，整理された初期値を使うことで，クラスタリング精度の悪い V の局所解に収束することを防いでいる．

4. 提 案 手 法

本論文では，2.3 節で挙げた「J の局所最適解が必ずしもクラスタリングの最適解であるとは言えない」という問題の解決のために，類似する文書はクラスタとの関係も類似するという考えに基づいた解決法を提案する．つまり，以下に示す二つの項目が類似するという考えである．

- 索引語文書行列 X の列ベクトルで表現された単語が要素である文書ベクトル同士の類似度
- NMF のクラスタリング結果である V^T の列ベクトルで表現されるクラスタと文書の関連度が要素である文書ベクトル同士の類似度

具体的には，(3) 式の目的関数に (7) 式のように類似性を反映した類似行列項を代入する．

$$J_1 = ||X - UV^T||_F + \mu||sim(X) - sim(V^T)||_F \tag{7}$$

ここで，$sim(A)$ は A 行列の列ベクトル同士のコサイン類似度行列であり (8) 式で求める．

$$sim(A)_{ij} = \frac{A_i \cdot A_j}{||A_i|| \times ||A_j||} \tag{8}$$

ここで，A_g は行列 A の g 列目の列ベクトルである．また，μ は類似行列項の重みである．μ は実験を行い設定する．

U と V の新たな更新式は (7) 式をそれぞれ U と V に対して微分することで求めることができる. 微分して求めた更新式が (9) 式と (10) 式である．

$$v_{ij} \leftarrow v_{ij} \frac{(X^T U)_{ij}}{(VU^T U)_{ij}} - \mu \frac{\sum_{m \neq j}^{D} (M_{mj} + M_{jm})}{||sim(X) - sim(V^T)||_F} \tag{9}$$

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^T V)_{ij}} \tag{10}$$

ここで，$(V^T)_{ij}$ は $sim(V^T)$ の i 行と j 列目に含まれているため M_{ab} において場合分けして求めている．M_{mj} は $sim(V^T)$ の j 行目の行ベクトルに，M_{jm} は $sim(V^T)$ の j 列目の列ベクトルに対する部分である．M_{ab} は (11) 式に示す．

$$M_{ab} = (sim(V^T)_{ab} - sim(X)_{ab}) \times (\frac{d}{dv_{ij}} sim(V^T)_{ab}) \tag{11}$$

ここで，$sim(V^T)_{ab}$ に対する微分は (12) 式である．

$$\frac{d}{dv_{ij}} sim(V^T)_{ab} = \frac{(V^T)_{ia}}{||V_a^T|| \times ||V_b^T||} - (V^T)_{ij} sim(V^T)_{ab} ||V_a^T||^{-\frac{1}{2}} \tag{12}$$

ここで，$||V_g^T||$ は V^T の g 列目の列ベクトルのノルムである．

5. 実　　験

5.1 実験データ

Yahoo!ニュースで公開されている 2011 年の記事から取得した日時の範囲が異なる 4 つの記事集合を用いた．各記事集合の記事数は 250 から 300 である．そのような記事集合をそれぞれ 6 クラスにクラスタリングする．各記事集合に関するデータの詳細は以下の表 1 に示す．

表 1　データセット

Data	Document	Words	Clus
yahoo-1	204	6183	6
yahoo-2	300	9590	6
yahoo-3	300	9870	6
yahoo-4	300	9229	6

表 1 において，Data は各文書集合であり，Document と Words と Clus はそれぞれ各文書集合における文書数と索引単語数とクラスタ数を表している．

実験に用いる記事の文書データは形態素解析器 MeCab を用いて名詞を抽出し，各文書データにおいて TF-IDF 値を求め，文書ベクトルの大きさが 1 になるように正規化した値である．

5.2 実験の概要と評価方法

実験では，既存の NMF と改良した NMF とのクラスタリング結果を比較する．重み μ を 0.001 から 0.00025 に変化させて表 1 の記事データを国内・海外・経済・スポーツ・エンタメ・科学の 6 クラスタに分ける．

各記事集合を 20 回クラスタリングし各評価方法での平均をとる．この時，U と V の更新式での繰り返しは 30 回とする．

クラスタリング結果の評価方法は Entropy と Purity と Accuracy を用いる．

Entropy は式 (13) より求める．

$$Ent = \sum_{i=1}^{k} \frac{|C_i|}{N} \times (-\sum_{h=1}^{k} P(A_h|C_i) \log P(A_h|C_i)) \tag{13}$$

Purity はクラスタリング結果であるクラスタに一番多く含まれている正解クラスタを用いて，どの程度クラスタリング結果が良いかを示す指標である．クラスタリング結果の Purity は，各クラスタのデータ数による重み付き平均をとるように定義し，式 (14) に示す．

$$Pur = \frac{1}{N} \sum_{i=1}^{k} \max_{h} |C_i \cap A_h| \tag{14}$$

Accuracy は，正解クラスタとクラスタリング結果のクラスタの全ての組み合わせにおいて一番精度のよい組み合わせをその結果の精度とする．Accuracy は式 (15) で求める．

$$Acc = \sum_{i=1}^{k}(\frac{|C_i|}{N} \times \frac{|A_h \cap C_i|}{A_h}) \quad (15)$$

式 (13)，式 (14)，式 (15) において C_i はクラスタリング結果に対する i 番目の文書のクラスタであり，A_h は正解データに対する h 番目の文書のクラスタである．$A_h \cap C_i$ は正解データであるクラスタ A_h とクラスタリング結果のクラスタ C_i が共通している記事数である．またここで N は文書数を示す．

本論文では Yahoo!ニュースでのカテゴリにおいて，同一のカテゴリに属する記事同士は文書ベクトルが類似しているという仮定しているため Accuracy の値が高いと類似する文書同士が同一のクラスタにクラスタリングされたということを示す．

5.3 実験結果

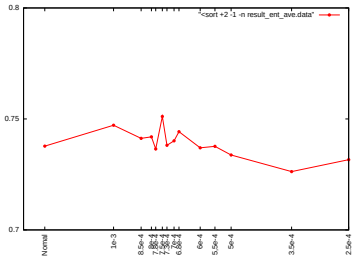

図 1 Entropy の μ による推移

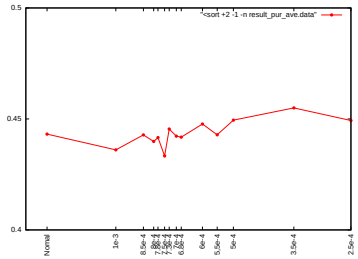

図 2 Purity の μ による推移

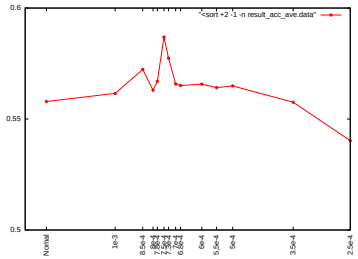

図 3 Accuracy の μ による推移

4 つの記事集合を既存の NMF と改良した NMF($\mu = 0.001 \sim 0.00025$) で各 20 回クラスタリングを行った結果を図 1，2，3 に示す．図 1，2，3 での Entropy と Purity と Accuracy の値は，4 つの記事集合をクラスタリングして得られた各記事における 20 回の平均を，さらに 4 つの記事集合で平均した値である．

5.4 実験考察

図 1 を見ると $\mu = 0.0005$ 以降で Entropy は減少しているためそれ以降で Entropy は良くなっている．また，図 2 を見ると $\mu = 0.0006$ 以降で Purity は増加しているためそれ以降で Purity は良くなっている．そして，図 3 を見ると $\mu = 0.00075$ 周辺において増加しているためその周辺では Accuracy は良くなっていると言える．しかし，図 1，2，3 において Accuracy が最も良い値を示している $\mu = 0.00075$ では，Entropy と Purity はあまり良い値を示していない．これは，外れ値が影響していると考えられる．

結果から $\mu = 0.00073$ が最良であると判断しその結果を表 2 に示す．

表 2 $\mu = 0.00073$ での結果

NMF	Entropy	Purity	Accuracy
既存	0.738	0.443	0.557
改良	0.738	0.445	0.577

表 2 では，Entropy と Purity はあまり変化していないが Accuracy の値は増加している．そのため，類似する文書データは同一のクラスタにクラスタリングされやすくなったと考えられる．

6. 終わりに

結果から，既存の NMF と比較して $\mu = 0.00073$ に設定した改良 NMF での記事クラスタリングは Entropy と Purity を保ちながら Accuracy の向上する効果があることがわかった．つまり，既存の NMF における目的関数に類似行列項を導入することは類似する記事同士を同一のクラスタにクラスタリングすることに効果があると言える．

Accuracy の更なる向上や NMF によるクラスタリングの分散度合いの改善が今後の課題である．

参 考 文 献

1) D.D.Lee , H.S.Seung : “AlgorithmsforNon-negative Matrix Factorization”, NIPS , pp.556-562 , (2000).
2) C.Ding,T.Li,W.Peng : “On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing”, Computational Statistics and Data Analysis 52 , 3913 - 3927 (2008).
3) 萩野広樹, 吉田哲也: “トピックグラフに基づく NMF を用いた転移学習” , IPSJ SIG Technical Report , Vol.2011-MPS-82 No17.
4) 新納浩幸，佐々木稔 : “Mcut + NMF による文書クラスタリング” , C3-7,pp,558-561,(2007)
5) 新納浩幸，佐々木稔: “NMF とリンクベースの修正法によるピンポン型文書クラスタリング” , 情報処理学会，自然言語処理研究会報告，Vol.2007,no.47,p.7-12.