自 然 言 語 処 理 154－4
（２００３．３．６）

# Non-negative Matrix Factorization を用いた情報検索モデルの次元圧縮および検索質問拡張

好田 勲† 柘植 覚‡ 獅々堀 正幹‡ 北 研二††

†徳島大学大学院 工学研究科
‡徳島大学 工学部
††徳島大学 高度情報化基盤センター
〒 770-8506 徳島市南常三島町 2-1
e-mail: {kouda, tsuge, bori, kita}@is.tokushima-u.ac.jp

**あらまし** ベクトル空間モデル (Vector Space Model; VSM) は情報検索における代表的な検索モデルであり，検索対象文書および検索質問を多次元ベクトルで表現するという特徴を持っている．しかし，これらのベクトルは一般にスパースかつ高次元であるため，計算機のメモリによる制限や検索時間の増大などの問題が生じる．また，次元が増加するに連れ，文書中に含まれる不必要な索引語がノイズ的な影響を及ぼし検索精度を低下させてしまうという現象も起こってくる．以前，我々はこの問題を解決するため，Non-negative Matrix Factorization (NMF) を用いた VSM の次元圧縮手法を提案した．しかし，メモリの問題がまだ存在する．そこで，本稿では，k-means NMF を用いた VSM の次元圧縮手法を提案する．また，スパースな行列に対し有効な検索手法である検索質問拡張に NMF を用いる手法を提案する．MEDLINE コレクションを用いた検索実験を行った結果，NMF を用いた場合と k-means NMF を用いた場合では，検索精度を劣化することなく計算に必要なメモリを約 1/10 に軽減することができた．また，NMF を用いた検索質問拡張も VSM よりも高い検索精度を示すことができた．

**キーワード** 情報検索，ベクトル空間モデル，Non-negative Matrix Factorization，次元圧縮, k-means, 検索質問拡張

# Dimensionality Reduction and Query Expansion of Information Retrieval using Non-negative Matrix Factorization

Isao Kohda Satoru Tsuge Masami Shishibori Kenji Kita

Department of Information Science & Intelligent Systems
Faculty of Engineering, Tokushima University

2-1, Minami-josanjima, Tokushima, 770-8506

e-mail: {kouda, tsuge, bori, kita}@is.tokushima-u.ac.jp

**Abstract** The Vector Space Model (VSM) is a conventional information retrieval model, which represents a document collection by a term-by-document matrix. Since term-by-document matrices are usually high-dimensional and sparse, they are susceptible to noise and are also difficult to capture the underlying semantic structure. Additionally, the storage and processing of such matrices places great demands on computing resources. Dimensionality reduction is a way to overcome these problems. We proposed non-negative matrix factorization (NMF) for dimensionality reduction of the vector space model. However,this method did not overcome memory problems. Hence, we propose k-means NMF for dimensionality reduction of the vector space model. And, we propose query expansion using NMF in this paper. Using MEDLINE collection, we experimentally showed that k-means NMF offers great improvement over the vector space model.

**key words** information retrieval, vector space model, non-negative matrix factorization, dimensionality reduction, k-means, query expansion

# 1 はじめに

近年のインターネット技術の発展により，World Wide Web (WWW) を代表とする，個人で扱えるオンラインテキストデータの量が増加している．それに伴い，莫大なテキストデータ中から必要な情報を検索する機会も増え，情報検索に関する研究 [1][2][3] への関心が高まっている．

情報検索システムの代表的なモデルとして，検索対象文書と検索質問を多次元ベクトルで表現するベクトル空間モデル (VSM; Vector Space Model)[4] がある．このモデルを用いた情報検索システムは，質問ベクトルと文書ベクトル間の類似度を計算し，類似度の高い文書を検索結果として出力する．

しかし，検索対象文書が膨大となると，各文書のベクトルは要素に 0 が多い非常にスパースなベクトルとなる．文書全体をこのようなスパースなベクトルで表現すると記憶容量が莫大となり，類似度計算を行う際の計算コストも増加する．また，不必要な索引語がノイズとなり，検索精度の劣化につながる．このため，これらのスパースなベクトルで表現された文書全体 (索引語文書行列) を圧縮する手法が現在までに数多く提案されている．一般に，情報検索に次元圧縮を行った行列を用いると圧縮を行わない行列より，検索精度が高くなる傾向がある．

我々は，索引語文書行列の次元圧縮を行う手法として Non-negative Matrix Factorization[6][7] (NMF) を用いた次元圧縮手法を提案した [8]．しかし，NMF を用いて次元圧縮を行った場合，基底行列を求める際メモリの問題が生じる．この問題を解決するため索引語文書行列を分野毎に分類し，NMF により基底行列を求め次元圧縮する k-means NMF を用いた VSM の次元圧縮手法を提案する．NMF で求められる各基底ベクトルには，その基底を代表する要素に対して強い重みがかかっている．以下，**2**において，NMF の概要を説明し，**3**では，NMF，k-means NMF を用いた情報検索のための次元圧縮手法について述べる．提案手法の有効性を検証するため，**3.3**において，英文情報検索テストコレクション MEDLINE を用いた情報検索実験を行い，それらの結果に対する考察を行う．そして，**4**において NMF を用いた情報検索のための検索質問拡張手法について述べ，検索質問拡張手法の有効性を検証するため，**4.2**において MEDLINE を用いた情報検索実験を行い，結果の考察を行う．最後に，**5**において，本稿のまとめと今後の課題について述べる．

# 2 Non-negative Matrix Factorization

NMF は，非負の $n \times m$ 行列 $V$ を非負の $n \times r$ 行列 $W$ および非負の $r \times m$ 行列 $H$ に分解する手法である．

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

一般に近似行列 $WH$ のランク $r$ を

$$(n+m) * r < n * m \tag{2}$$

の範囲で選択することにより，$WH$ は元行列 $V$ を圧縮した行列とみなすことができる．

$V$ の各列ベクトルを$\boldsymbol{v}_i$，$H$ の列ベクトルを$\boldsymbol{h}_i$とすると，式 (1) は，

$$\boldsymbol{v}_i \approx W\boldsymbol{h}_i \tag{3}$$

を書くことができる．この式より，$\boldsymbol{v}_i$は$\boldsymbol{h}_i$ の要素で重み付けされた $W$ の線形結合であるとみなすことができる．これより，$W$ は $V$ 内のデータを線形近似するための基底行列であると考えることができる．

NMF は主成分分析 (PCA;Principal Compnent Analysis) や SVD などと異なり，非負制約条件で行列分解を行う．そのため，得られた分解行列は減算を伴わない加算のみの線形結合で元行列を表現できる．これは，特定要素のみで全体の行列を表現可能であることを示し，我々の直観を反映している．

## 2.1 分解行列 $W, H$ の更新規則

NMF では，行列 $V$ を 2 つの行列の積で近似するが，この際の近似の良さの尺度として，2 行列間の距離と 2 行列間の相違が用いられる [6] [7].

はじめに，2 行列間の距離を最小にするように $W, H$ を更新し，元行列に近似する手法について述べる．

近似行列 $W, H$ は，

$$\overline{H}_{ij} = H_{ij}\frac{(W^TV)_{ij}}{(W^TWH)_{ij}} \tag{4}$$

$$\overline{W}_{ij} = W_{ij}\frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{5}$$

の規則で更新される．ここで，$\overline{H}, \overline{W}$ は更新された分解行列であり，繰り返し演算を行う場合には，$\overline{H} \to H$,$\overline{W} \to W$ と変換し，再度，式 (4)，(5) を適用する．この更新式を更新規則 1 とする．この更新規則は，式 (6) に示す 2 行列間のユークリッド距離を用いた目的関数が収束するまで繰り返しを行い，元行列 $V$ を近似する行列 $W, H$ を得る．

$$F = \sum_i \sum_j (V_{ij} - (WH)_{ij})^2 \tag{6}$$

また，近似行列と元行列間の相違を最小にするように，$W, H$ を更新し，元の行列を近似する手法について述べる．この手法は，次に示す更新規則を用い近似行列 $W, H$ の更新を行う．

$$\begin{aligned} \overline{H}_{ij} &= H_{ij} \sum_k W_{ki} \frac{V_{kj}}{(WH)_{kj}} \qquad (7) \\ \widehat{W}_{ij} &= W_{ij} \sum_k \frac{V_{ik}}{(WH)_{ik}} H_{jk} \\ \overline{W}_{ij} &= \frac{\widehat{W}_{ij}}{\sum_k \widehat{W}_{kj}} \qquad (8) \end{aligned}$$

ここで，$\overline{H}, \overline{W}$ は，それぞれ更新された $H, W$ を示す．更新規則 1 と同様に，繰り返し適用を行う場合には，更新された行列をそれぞれ $H, W$ とし，再度この更新規則を適用する．この更新規則を以下では，更新規則 2 とする．

この更新規則は，式 (9) に示す目的関数が局所的に最大となるように繰り返し適用することにより，元行列を近似した $W, H$ を得る．この目的関数は，近似尺度として，Kullback-Leibler divergence を用いている．

$$F = \sum_i \sum_j (V_{ij} \log((WH)_{ij}) - (WH)_{ij}) \tag{9}$$

## 2.2 基底ベクトルに関する考察

本節では，NMF により求められた基底行列に関する考察を行う．NMF は非負行列分解手法であるため，求められた基底行列は負要素を含まない．そのため，各基底の重み付け和のみで元行列を表現できる．これは部分和により全体を把握するという我々の直観に近い．また，各基底は元行列の意味的空間を張る軸と考えることもできる．そこで本節では，NMF により求めた基底がどのような意味空間を張る軸を構成しているかを調べる．

表 1: NMF により求められた基底

| 基底 1 | | 基底 2 | |
|---|---|---|---|
| 索引語 | 重み | 索引語 | 重み |
| 投手 | 0.108 | 尾崎 | 0.230 |
| オリックス | 0.061 | 合田 | 0.149 |
| 巨人 | 0.042 | ツアー | 0.027 |
| ダイエー | 0.038 | バーディー | 0.027 |
| 日本ハム | 0.038 | 将司 | 0.023 |
| ヤクルト | 0.036 | パー | 0.023 |
| キャンプ | 0.028 | 将 | 0.020 |
| 横浜 | 0.026 | 最終 | 0.020 |
| 郭 | 0.021 | ミノザ | 0.019 |
| 阪神 | 0.021 | プロ | 0.018 |
| **基底 3** | | **基底 4** | |
| 索引語 | 重み | 索引語 | 重み |
| 大会 | 0.124 | ボクシング | 0.113 |
| ＰＬ学園 | 0.087 | 世界 | 0.061 |
| 拓大一 | 0.054 | フィールド | 0.040 |
| 投手 | 0.041 | 米国 | 0.038 |
| 安打 | 0.038 | 辰吉 | 0.035 |
| 小達 | 0.033 | プロ | 0.032 |
| 中村 | 0.026 | リ | 0.030 |
| 大阪 | 0.024 | バンタム | 0.030 |
| 宇高 | 0.017 | タイトルマッチ | 0.030 |
| 宇島東 | 0.015 | 日本 | 0.027 |

考察には，毎日新聞 94 年新聞記事コーパスからスポーツに関連する 100 記事を用いた．これら 100 記事はラグビー，野球，サッカー，陸上，相撲，柔道，ボクシング，テニス，競馬，ゴルフの 10 分野からそれぞれ 10 記事ずつ選択した．選択した 100 記事に対し，茶筌 [9] を用い形態素解析を行い，名詞と判断された索引語を索引語として用いた．この結果，索引語数は 1,356 となった．これらの索引語に対し，ベクトル空間モデルを用い索引語文書行列を作成した．索引語文書行列の各要素に対する重み付けには tf-idf 法 [4] を用いた．

作成した索引語文書行列に対し NMF を適用し，基底行列を求めた．NMF の条件として，繰り返し回数 20 回，$W, H$ の初期値：0.0〜1.0 をランダムに発生，$r = 20$ とした．これらの条件により 20 個の基底ベクトルを求めた．

表 1に NMF により求められた基底行列の 1 部を示す．これらの索引語は各基底において重みが高い上位 10 索引語である．表 1より，NMF により求められた基底は適切に各分野を示す意味的な索引語に高い重みがつけられていることがわかる．特に，基底 1,3 は主に「野球」の内容を示している軸であると考えられる．その中でも，基底 1 は特に「プロ野球」に重みをおいた軸，基底 3 は「高校野球」に重みをおいた軸であると分類することができる．同様に基底 2 は「ゴルフ」，基底 4 は「ボクシング」と分類することができる．また，基底 2 と基底 4 の「ゴルフ」,「ボクシング」と分類できる基底には同じ索引語「プロ」が含まれている．同じ索引語においても重み付けされの違いにより新たな軸表現されるため，検索精度の向上が可能であるとが期待できる．

# 3 NMF を用いた VSM の次元圧縮

本節では，NMF を用いた VSM の次元圧縮手法について述べる．

## 3.1 次元圧縮手法

NMF は行列分解手法であるため，テキスト形式で書かれた文書集合に直接適用することはできない．そこで，VSM[4] により文書を索引語の多次元ベクトル (文書ベクトル) として表現し，テキスト形式である文書集合を NMF が適用可能な行列 (索引語文書行列) に変換する．NMF における元行列 $V$ として索引語文書行列を用い，分解行列 $W, H$ を得る．

**2**で述べたように，$W$ は $V$ を表現する基底ベクトルで構成された行列であると考えることができる．そこで，基底行列 $W$ のランクを元行列の次元数より低くし，

$$\bar{V} = W^T V \tag{10}$$

と索引語文書行列 $V$ を基底行列 $W$ に射影することにより索引語文書行列の次元を削減することができる．ここで，$W^T$ は $W$ の転置を示し，$\bar{V}$ は射影後の索引後文書行列である．この際，行列 $W$ のランクが $\bar{V}$ の各ベクトル，すなわち文書ベクトルの次元数となる．検索質問に対しても同様に VSM によりベクトル表現し，基底行列 $W$ に射影し，次元削減を行う．次元削減された文書ベクトルと次元削減された検索質問ベクトル間の類似度を計算することにより，類似文書検索を行う．

文献 [7] では，基底行列 $W$ の各列ベクトルは，その基底を代表する要素に強い重みがかかっていると報告されている．これは，$V$ 内に含まれる潜在的な意味ととらえることができ，これらの基底に射影することにより，LSI 同様に高い検索精度が期待できる．

## 3.2 k-means NMF

**2**により基底行列を求める場合，検索対象文書が大規模になると索引語文書行列が大きくなり，NMF の計算に莫大なメモリが必要となる．

そこで，検索対象文書をクラスタリングし，各クラスタに対し NMF を適用し基底ベクトルを求める手法を提案する．これより，NMF の計算に必要なメモリ量を軽減することが可能である．以下にその手順を示す．

1. 索引語文書行列をクラスタリングにより，分類.
2. 分類された各行列に対し NMF を適用し，基底行列を求める．
3. 求めた基底行列に索引語文書行列を射影.

また，クラスタリングにより分野が似通るため，分野に特化したよりよい基底が計算可能と推測でき，さらなる検索精度を得ることができると期待できる．本稿ではクラスタリング手法として k-means を用いた．

## 3.3 検索実験

情報検索における k-means NMF を用いた次元圧縮手法の有効性を検証するため，情報検索評価用テストコレクション MEDLINE を用いた情報検索実験を行った．以下で，この実験について説明する．

### 3.3.1 実験条件

MEDLINE は，医学・生物学分野における英文の文献情報データベースである．このテストコレクションは，検索対象文書 1,033 文書で構成される，約 1Mbyte の容量を持つテキストデータである．情報検索評価用データとして，30 個の検索質

表 2: 検索実験結果

| 手法 | 平均適合率 | メモリ (Mbyte) |
|---|---|---|
| VSM | 0.4954 | – |
| NMF(1) | 0.5755 | 39.3 |
| k-means NMF(1) | 0.5803 | 3.6 |
| NMF(2) | 0.5964 | 28.6 |
| k-means NMF(2) | 0.6117 | 2.7 |

問文書と各検索質問に対する正解 (関連) 文書が用意されている．各検索質問に対する平均関連文書数は 23.2 文書である．

このテストコレクションに含まれる 1,033 文書全体から，前処理として，“a” や “about” などの一般的な 439 個を，文書の内容とほとんど関連の無い索引語 (不要語) として削除した．この処理により削除されなかった索引語に対し，接辞処理を施し，語幹の変換を行った．この前処理の結果，文書全体に存在した単語数 5,526 から 4,328 に削減をし，それらの処理を施したこの 4,328 索引語を検索に用いる索引語として抽出し，実験データとして用いた．

前処理によって得られた索引語を用い，VSM に基づいた情報検索システムを構築した．VSM で作成を行った索引語文書行列の各要素 $d_{ij}$ は，文書番号 $j$ の索引語 $i$ に対する重みを表し，各索引語の頻度に重みを加えた数値である．これは，

$$d_{ij} = L_{ij} \cdot G_i \tag{11}$$

である．ここで，$L_{ij}$ は文書番号 $j$ の索引語 $i$ に対するのローカル重みをし，$G_i$ は索引語 $i$ のグローバル重みを示す．

これらの索引語の重み付けとして，本稿では，対数エントロピー手法 [4] を用いた．この重みは，

ローカル重み:

$$L_{ij} = log(1 + f_{ij}) \tag{12}$$

グローバル重み:

$$G_i = 1 + log\left(\sum_j \frac{p_{ij} \cdot log(p_{ij})}{log(m)}\right) \tag{13}$$

であたえられる．$m$ はテストコレクション中の文書数，$f_{ij}$ は文書番号 $j$ における索引語 $i$ の出現頻度を表す．また，$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}$ を示す．

NMF における分解行列 $W, H$ の初期値として，0.0〜 1.0 間の数字をランダムに発生させたものを用いた．NMF における繰り返し回数は 20 回とし，圧縮後の次元数は 500 次元とした．また，k-means NMF における k-means のクラスタ数は 10 とする．

検索システムの精度の評価には，一般的によく用いられる平均適合率を用いた．

#### 3.3.2 検索実験結果

情報検索実験結果を表 2に示す．なお，表中の ( ) 内の数字は，更新規則番号を示す．

表 2の結果より，k-means NMF により 500 次元に圧縮することにより，VSM の平均適合率 0.4954 を更新規則 1 では 0.5803，更新規則 2 では 0.6117 と，約 16.8%，約 23% 改善ができた．また，NMF と k-means NMF では検索精度は更新規則 1，2 共にほぼ同等である．しかし，NMF を計算する際に必要なメモリが，NMF に比べ約 1/10 に軽減することができた．これらのことから，本手法は有効であるといえる．

## 4 NMF を用いた検索質問拡張

本節では，NMF を用いた VSM の検索質問拡張法について述べる．一般にユーザが検索したい文書に対する検索質問文書を提示した場合には，通常検索質問文書に含まれる索引語に対してのみ検索が行われる．しかし，検索対象文書が，検索質問文書と異なった索引語 (類義語) で記述されていた場合，検索が行われない．そこで，ユーザが提示した検索質問文書に対し，関連のある索引語を新たに付け加えそのような類義語に対して検索が行われるにする手法が検索質問拡張である．そこで，NMF を用いた検索質問拡張法を提案する．NMF で求められた基底ベクトルは，**2.2**で述べたように，各分野を適切に示す意味的な索引語に高い重みがつけられている．よって，NMF で求めた基底を検索質問拡張に利用することは，検索精度の向上に有効だと考えられる．

### 4.1 検索質問拡張法

NMF を検索質問拡張に用いるため以下にその手順を示す．

1. 索引語文書行列に対し NMF により，基底行列 $W$ を求める．

表 3: 検索実験結果

| 手法 | VSM | NMF | | |
|---|---|---|---|---|
| N | | 10 | 20 | 30 |
| 平均適合率 | 0.4954 | 0.5375 | 0.5353 | 0.5382 |

2. 検索質問ベクトル$\boldsymbol{q}$に含まれる索引語 $q_j$ の重みが最も高い基底ベクトル $W_i$ を選択.

3. 選択された基底ベクトル $W_i$ に対し，閾値以上の索引語を拡張し，次式で行う.

$$q_j = \sum_{j=0}^{N}(W_{ij} + q_j) \tag{14}$$

4. 検索質問ベクトルに含まれる索引語に対して 2，3 を行う.

検索質問拡張した検索質問ベクトルと検索対象文書ベクトルとの類似度を計算を行うことにより，文書検索を行う.

### 4.2 検索実験

情報検索における NMF を用いた検索質問拡張の有効性を検証するため，情報検索評価用テストコレクション MEDLINE を用いた情報検索実験を行った．以下で，この実験について説明する.

#### 4.2.1 実験条件

用いるテストコレクション MEDLINE の実験条件，NMF の初期値は繰り返し回数は **3.3.1**と同様である．NMF における求める基底数は 400 とした.

拡張する検索質問の閾値は，各質問ごとの各索引語の上位 $10, 20, 30$ 個を拡張する値とした．検索システムの精度の評価には，一般的によく用いられる平均適合率を用いた.

### 4.3 検索実験結果

MEDLINE による情報検索実験を行った．表 3 に結果示す．この結果より，全ての検索質問拡張法において VSM より検索精度が向上していることがわかる．このことから，ユーザが提示した検索質問文で検索を行った際に上位に位置付けられなかった文書が，NMF を用いて検索質問拡張を行うことにより，上位に位置付けられたからだと考えられ，本手法が有効であることがわかる.

## 5 まとめ

本稿では，k-means NMF を用いたベクトル空間モデル VSM の次元圧縮手法提案した．本手法は，あらかじめ索引語文書行列をクラスタリングし，各クラスタに対し NMF により基底ベクトルを求める手法である．MEDLINE を用いた情報検索実験を行い，情報検索における NMF を用いた次元圧縮手法と k-means NMF を用いた次元圧縮手法との比較実験を行った．その結果，k-means NMF を用いた次元圧縮手法は，NMF の計算時に必要とするメモリ量を約 1/10 に軽減することができ，検索精度も向上した.

また，NMF を用いた検索質問拡張手法においても検索精度が VSM よりも向上した．これは NMF を用いることにより適切な検索質問拡張がされ，検索漏れのあった検索対象文書が上位に位置づけされたからだと考えられる.

本稿では，情報検索実験として，テストコレクション MEDLINE を用いた．今後さらに，大規模なテストコレクションでの情報検索実験を行う予定である.

## 参考文献


[1] TREC homepage. http://trec.nist.gov/.

[2] IREX homepage. http://cs.nyu.edu/cs/projects/proteus/irex.

[3] NTCIR homepage. http://www.rd.nacsis.ac.jp/˜ntcadm/.

[4] 北 研二，津田 和彦，獅々堀 正幹. 情報検索アルゴリズム，共立出版 2002.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[6] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *NIPS 2000*, 2000.

[7] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, Vol. 401, pp. 788–791, 1999.

[8] 柘植 覚，獅々堀 正幹，北 研二. Non-negative Matrix Factorization を用いた情報検索 自然言語処理研究会, 2001, NL142–1, pp.1–6

[9] 松本裕治，北内 啓，山下 達雄，平野 善隆，松田 寛，浅原 正幸. 日本語形態素解析システム『茶筌』 version 2.0 使用説明書 第二版. 奈良先端科学大学院大学, Technical Report NAIST-IS-TR99012, 1999.