情 報 学 基 礎 61－1
自 然 言 語 処 理 142－1
（2001. 3. 5）

Non-negative Matrix Factorization を用いた情報検索

柘植 覚　獅々堀 正幹　北 研二

徳島大学 工学部 知能情報工学科

〒770-8506 徳島市南常三島町 2-1

e-mail: {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単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．本稿では，Non-negative Matrix Factorization (NMF) を用いたベクトル空間モデルの次元圧縮手法を提案する．NMF は非負行列を 2 つの非負行列の積に分解する手法であり，分解された非負の 2 行列は基底行列とその基底のもとでの座標値から成る行列とみなすことができる．基底行列のランクを元の行列のランクより小さくすることにより，次元圧縮が可能となる．NMF は，主成分分析や特異値分解などと異なり，非負制約条件で行列分解を行うため，元の行列を減算を伴わない加算のみの線形結合で表現することができる．これは部分から全体を構成するという我々の直観を反映している．また，NMF は単純な繰り返し演算のみで実行可能であるため，大規模な行列に対して，計算コストや記憶容量の点で他の次元削減手法よりも優れている．MEDLINE コレクションを用いた検索実験を行い，NMF は通常のベクトル空間モデルよりも高い検索性能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

Information Retrieval Using Non-negative Matrix Factorization

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: {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. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are popular techniques for dimensionality reduction based on matrix decomposition, but they contain both positive and negative values in the decomposed matrices. In the work described here, we use non-negative matrix factorization (NMF) for dimensionality reduction of the vector space model. Since decomposed matrices by NMF only contain non-negative values, the original data is represented by only additive, not subtractive, combinations of the basis vectors. This characteristic of parts-based representation is appealing because it reflects the intuitive notion of combining parts to form a whole. Also NMF computation is based on the simple iterative algorithm, it is therefore advantageous for applications involving large matrices. Using MEDLINE collection, we experimentally showed that NMF offers great improvement over the vector space model.

key words　information retrieval, vector space model, non-negative matrix factorization, dimensionality reduction

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への関心が高まっている．これらの研究は，米国における TREC (Text Retrieval Conference)[1] や，日本における IREX (Information Retrieval and Extraction Exercise)[2]，NTCIR (NII-NACSIS Test Collection for IR Systems)[3] のワークショップを中心に広く行われている．

情報検索システムとして，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] が広く使用されている．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，全文書中に含まれる単語を用い，各文書を多次元ベクトル表現するため，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，さらに，類似度計算を行う際の計算コストも増加してしまう．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 (単語文書行列) 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，検索性能が高くなる傾向がある．

単語文書行列の次元圧縮を行う最も代表的な手法として，特異値分解 (SVD; Singular Value Decomposition) を用いた潜在的意味解析 (LSI; Latent Semantic Indexing) が提案されている [5]．この手法は，単語文書行列に対し特異値分解を行い，元の単語文書行列より低いランクの基底行列を求め，その基底に各ベクトルを射影することにより，次元圧縮を行う．しかし，特異値分解は計算コストが高いため，大規模な行列に対して適用することは困難である．

本稿は，Non-negative Matrix Factorization (NMF)[6][7] を用いたベクトル空間モデルの次元圧縮手法を提案する．NMF は，非負行列を 2 つの非負行列の積に分解する手法である．単純な繰り返し演算のみで実行可能であるため，大規模な行列に対しても有効性が高いと考えられる．分解された非負の 2 行列は，基底行列とその基底に対する座標値と考えることができ，基底行列のランクを元の行列のランクより低くすることにより，次元圧縮が可能となる．

以下，2において，Non-negative Matrix Factorization (NMF) の概要を説明し，3では，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，4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．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 [6][7]．

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

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 は元の行列 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 $\boldsymbol{v}$，H の列ベクトルを $\boldsymbol{h}$ とすると，式 (1) は，

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

と書くことができる．この式は，$\boldsymbol{h}$ 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．

上述の通り，NMF は主成分分析 (PCA; Principal Component 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の行列を表現できる．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，元の行列に近似する手法について述べる [6]．近似行列 W, H は，

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

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

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，$\overline{H} \rightarrow H$, $\overline{W} \rightarrow W$ と変換し，再度，式 (4)，式 (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，元の行列 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，近似行列と元の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 [7]．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．

$$\overline{H}_{ij} = H_{ij} \sum_k W_{ki} \frac{V_{kj}}{(WH)_{kj}} \tag{7}$$

$$\hat{W}_{ij} = W_{ij} \sum_k \frac{V_{uk}}{(WH)_{ik}} H_{jk}$$

$$\overline{W}_{ij} = \frac{\hat{W}_{ij}}{\sum_k \hat{W}_{kj}} \tag{8}$$

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，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，元の行列を近似した 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．

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

3 NMF を用いた次元圧縮手法

NMF を情報検索に用いるため，検索対象文書を VSM により多次元ベクトルで表現した単語文書行列の作成を行う．この行列を NMF の元の行列 V として用い，更新規則の繰り返しにより近似分解行列 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．そのため，W のランクを低くし，V をこれらの基底に射影することにより，V の次元圧縮を行うことができる．この時，W のランク r が，射影された行列の次元数となる．検索質問に対しても，同様に，ベクトル表現を行い，W と線形結合させ，次元圧縮を行う．これらの次元圧縮されたベクトルの類似度を比較することにより，情報検索を行う．

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．

4 検索実験

情報検索における 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．

4.1 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 1033 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 1033 文書全体から，前処理として，“a” や “about” などの一般的な 439 個は，文書の内容とほとんど無い関連の単語 (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単語に対し，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 5526 から 4328 単語に削減をし，それらの処理を施したこの 4328 単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，ベクトル空間モデル (VSM; Vector Space Model) に基づいた情報検索システムを構築した．VSM で作成を行った単語文書行列の各要素 d_{ij} は，文書 j に対する索引語 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，

$$d_{ij} = L_{ij} \cdot G_i \tag{10}$$

である．ここで，L_{ij} は文書番号 j の索引語 i 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 [8] を用いた．この重みは，

ローカル重み:

$$L_{ij} = log(1 + f_{ij}) \tag{11}$$

グローバル重み:

$$G_i = 1 + log\left(\sum_j \frac{p_{ij} \cdot log(p_{ij})}{log(m)}\right) \qquad (12)$$

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 では，元の行列が $n \times m$ 行列ならば，圧縮後の次元数，すなわち近似行列 WH のランク数 r を $(n+m)*r < n*m$ とすることにより，元の行列を圧縮した近似行列が得られる．本稿で用いたテストコレクションの場合，元の行列の行数 (検索索引語数) $n = 4328$，列数 (文書数) $m = 1033$ であるため，r は，833 未満の値を取ることにより元の行列を圧縮する W, H を得ることができる．

検索システムの精度の評価には，一般的に用いられている適合率 (Precision) と再現率 (Recall) を用いた [9][10]．再現率と適合率は，それぞれ個別に用いて，システム評価を行うことができるが，本稿では，一般にランクづけ検索システムの評価に用いられる再現率-適合率曲線を用い，システムの評価を行った．この曲線は，各質問に対しひとつの曲線が作成されるが，本稿では，全質問の平均再現率-適合率曲線を用いた．また，これらの平均から得られる平均適合率においても評価を行った．適合率，再現率の計算は，

$$\text{適合率} = \frac{\text{関連のある文書数のうち検索できた文書数}}{\text{検索文書数}}$$

$$\text{再現率} = \frac{\text{関連のある文書数のうち検索できた文書数}}{\text{関連のある文書数}}$$

で行った．

4.2 検索結果

4.2.1 NMF を用いた次元圧縮の有効性

NMF の有効性を検証するため，繰り返し回数を 20 回の場合の次元数に対する検索性能の比較を行った．比較を行った圧縮後の次元は，$r < 833$ で元の行列を圧縮することが可能であるため，圧縮後の次元数 (WH のランク) は，$r = 20, 40, 100, 400, 800$ の 5 種類を用いた．

図 1に，更新規則 1 を用いた場合の各次元に対する再現率-適合率曲線を示す．また，図 2に，更新規則 2 を用いた場合の各次元に対する再現率-適合率曲線を示す．比較のため，次元圧縮を行わない VSM モデルで作成した単語文書行列を用いた情報検索結果 (再現率-適合率曲線) を同図に示す．

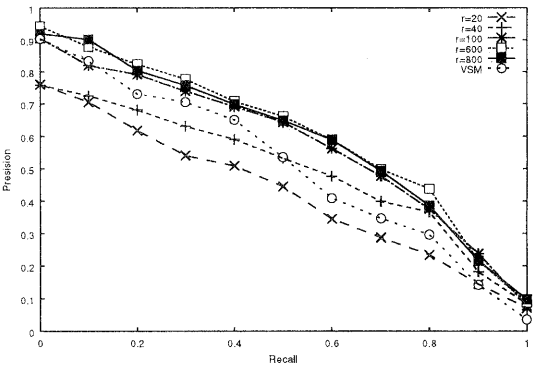

図 1: 更新規則 1 を用いた場合の次元数に対する検索性能

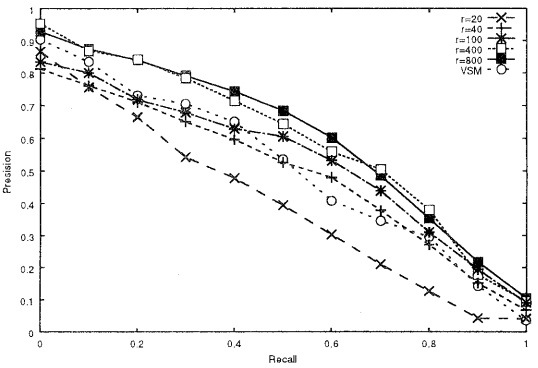

図 2: 更新規則 2 を用いた場合の次元数に対する検索性能

図 1，図 2より，いずれの更新規則においても，次元数がある一定を越えると次元圧縮を行っていない VSM モデルより検索性能が向上することがわかる．

また，いずれの更新規則ともに，圧縮を行った次元数がある一定を越えると検索精度は，ほぼ同程度となることがわかる．これは，NMF により分解を行った，基底行列 W の各列が単語文書行列 V を十分に表す基底をある程度の基底数で表現できているのではないかと推測できる．

また，各次元における繰り返し 20 回が終了した場合の目的関数が与えたコスト (これは，V と WH の類似度と取ることができる) を図 3，図 4に示す．この図より，各次元，繰り返し回数 20 回と固定した場合には，次元数が増加することにより，もとの行列をより近似する分解行列 W, H が得られていることがわかる．しかし，先

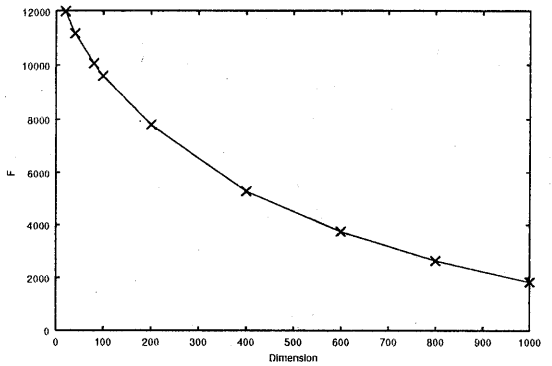

図 3: 更新規則 1 を用いた場合の繰り返し回数 20 回の目的関数から得られる F の値

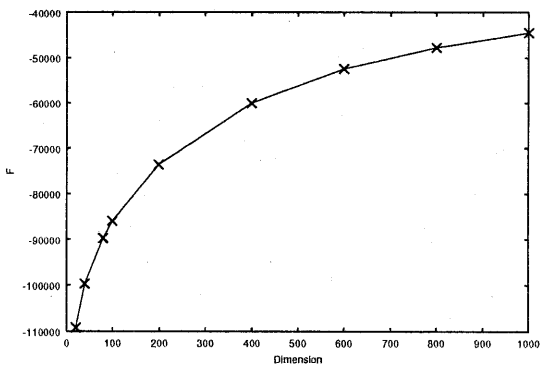

図 4: 更新規則 2 を用いた場合の繰り返し回数 20 回の目的関数から得られる F の値

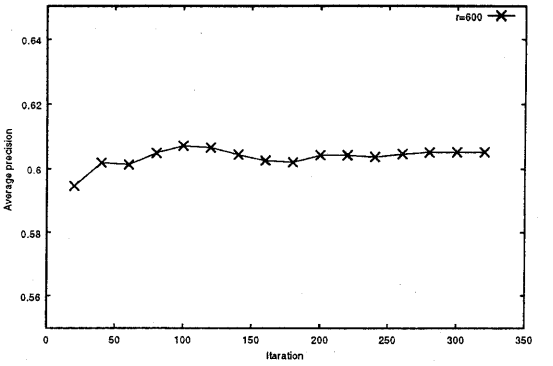

図 5: 更新規則 1 を用いた場合の繰り返しによる検索性能比較

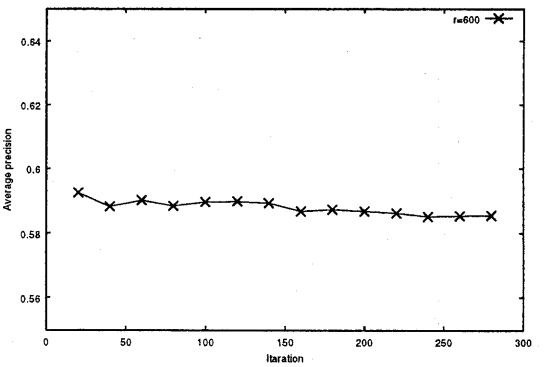

図 6: 更新規則 2 を用いた場合の繰り返しによる検索性能比較

に述べた通り，WH のランクがある一定以上になると情報検索性能は，ほぼ同等であるため，近似する分解行列が得られたとしても，必ずしも検索性能が高くなるわけではないことが示せた．

4.2.2 繰り返し回数に対する検索性能比較

NMF は，繰り返し更新規則を適用することにより，近似行列 WH を元の行列に近似させて行く手法である．そのため，繰り返し回数により，近似度合が変化し，情報検索性能に大きな影響を与えると考えられる．そこで，繰り返し回数に対する検索性能の比較を行った．

繰り返し回数と検索性能の比較には，**4.2.1**の実験において，更新規則 1 で高い検索性能を示した圧縮後の次元数，$r = 600$，更新規則 2 において，高い検索性能を示した圧縮後の次元数，$r = 400$ を用いた．これらの r に対し，繰り返し更新規則を適用し，繰り返し回数による検索性能の変化を調べた．

図 5，図 6にそれぞれ，更新規則 1 を用いた場合の繰り返し回数に対する平均適合率，更新規則 2 を用いた場合の繰り返し回数に対する平均適合率を示す．更新規則 1, 2 ともに，多少のバラツキは存在するが，更新規則 1 の場合，繰り返し 100 回程度，更新規則 2 の場合，50 回程度で平均適合率がほぼ収束していることがわかる．これは，VSM モデルでベクトル化を行った単語文書行列は，非常に 0 の多いスパースな行列であるため，少ない繰り返し回数でもとの行列 V に近似した分解行列が得られ，少ない繰り返し回数で検索性能が収束したと考えられる．

このことを確かめるために，図 7，図 8に，それぞれ更新規則 1，更新規則 2 を用いた場合の目的関数から

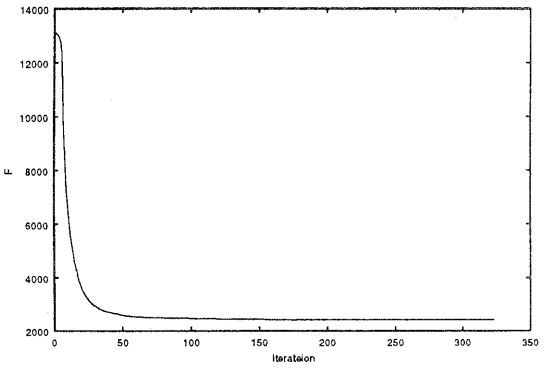

図 7: 更新規則 1 を用いた場合の繰り返しによるコストの変化

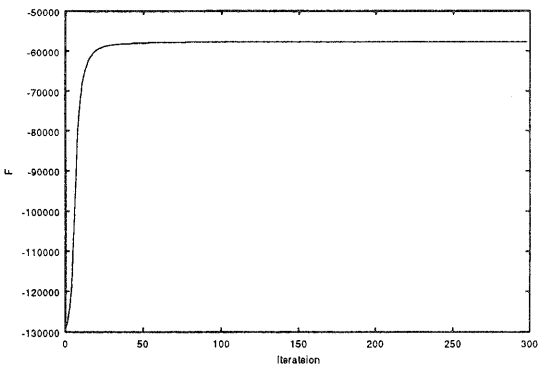

図 8: 更新規則 2 を用いた場合の繰り返しによるコストの変化

得られるコストを示す．これらの図から，検索結果と同様に少ない繰り返し回数で収束していることがわかる．これより，大まかであるが検索性能と繰り返しに対する目的関数から得られるコストとの相関関係があることがわかる．

本稿では，繰り返し回数の終了条件に目的関数からのコストを用いずに，繰り返し回数で制限を行っているが，検索精度と目的関数からのコストに相関関係が見られるため，コストの増加割合による繰り返し回数の制限を行うことが可能であることがわかった．

5 まとめ

本稿では，情報検索の代表モデルであるベクトル空間モデル (VSM; Vector Space Model) の次元圧縮手法として，Non-negative matrix factorization (NMF) を用いることを提案した．NMF は繰り返し演算により，非負行列 V を二つの非負行列 W, H に分解する手法であり，W は，V を表現する基底であると考えられる．そのため，基底に適したランク数の少ない W を選択することにより，V の表現能力を減少させることなく次元の圧縮が行うことが可能である．

MEDLINE を用いた検索実験を行い，情報検索における NMF を用いた次元圧縮手法の有効性を調べた．繰り返し回数 20 回の場合，次元数 100 次元に圧縮を行っても，従来の VSM の検索性能を上回った．また，繰り返し回数に対する情報検索性能を調べ，繰り返しをさほど多く行わなくても，従来の VSM より高い検索性能を得ることできるがわかった．

本稿では，繰り返し回数の終了条件は人手により決定を行ったが，目的関数から得られるコストと情報検索性能間に相関関係が見られるため，目的関数のコストを用いた繰り返し終了条件の検討を行う予定である．また，本稿では，基底行列 W の解析を行っていないが，実際にどのような値をとり，どこの検索索引語に強い重みがかかっていることを調べてることにより，それらの基底で表現される特に重要な検索索引語がわかると考えられる．そのため．今後，得られた基底行列の分析を行い，重要と考えられる検索索引語の解析を進めて行く予定である．

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] G. Salton and J. McGill. *Introduction to Modern Information Retrieval*. 1983.

[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] E. Chisholm an T. Kolda. New term weighting formulas for the vector space method in information retrieval. *Technical Memorandum ORNL-13756*, 1999.

[9] D. Lewis. Evaluating text categorization. *Proc. of Speech and Natural Language Workshop*, pp. 312–318, 1991.

[10] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York, 1994.