社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (10)
2008／9／21

文書ベクトル分布に基づく次元削減による文書検索空間の生成方式に関する評価実験

鷹野 孝典† 陳 幸生†

†神奈川工科大学情報学部情報工学科 〒243-0292 神奈川県厚木市下荻野 1030
E-mail: †{takano, chen}@ic.kanagawa-it.ac.jp

あらまし 本稿では，ベクトル空間の軸上に配置されている文書ベクトルの分布に基づいて，ベクトル空間の次元削減を行うことにより文書検索空間を生成する方式に関する評価実験について述べる．提案方式は，多くの文書ベクトルが均質的に配置された軸を削除することによって，正解文書とノイズ文書についての識別が可能な軸のみで構成される文書検索空間を生成する方式である．提案方式により，適切な次元数にベクトル空間を圧縮させることで，文書ベクトルはそのベクトル空間上で適切に配置され，利用者は問い合わせについて関連性のある文書を獲得することができるようになる．本研究では，評価用の文書コレクション用いた検索実験を行うことにより，提案方式により生成した文書ベクトル空間において，文書間や単語間の意味に応じた文書獲得が実現可能であることを確認した．

キーワード 文書検索，ベクトル空間モデル，意味的検索，ベクトル次元削減

Experiments of a Document Vector Space Creation Method Based on Distributions of Document Vectors

Kosuke Takano† and Xing Chen†

† Dept. of Info. & Comp. Sciences, Kanagawa Inst. of Tech., 1030 Simo-Ogino, Atsugi-shi Kanagawa, 243-0292 Japan
E-mail: †{takano, chen}@ic.kanagawa-it.ac.jp

Abstract In this paper, we describe experiments of a document vector space creation method based on distributions of document vectors. In the semantic search, with the improvement of recall rates of search results, precision rates tend to go down because of noise documents. Based on a basic idea that there is a tendency that basically such noise documents are included in axes of the vector space where many documents are mapped, a proposed method extracts a subspace by deleting axes where a lot of documents are homogeneously mapped so that we can improve precision rates of search results. In this study, we implemented an experimental retrieval system by using a test document collection and clarify the feasibility of our method by showing several experimental results.

Key words document retrieval, vector space, semantic retrieval, reduction of vector dimensions

1. はじめに

ベクトル空間モデルを用いた文書検索方式では，検索文書と問い合わせの内容的な類似性の比較を行う文書検索に対して有効であると確認されており[1][15]，様々な意味的情報検索方式が提案されている[3][9]．

意味や内容に基づいた意味的な検索を実現するための重要な課題の一つは，問い合わせの要求に従って適切なメディアデータを選択することである．適切な情報を獲得するためには，メディアデータの特徴を示すメタデータが的確に付与され，かつそのメタデータに基づいて生成される特徴ベクトルが意味的な検索を可能とするベクトル空間上で適切に配置されている必要がある．

LSI（Latent Semantic Indexing）方式[6]は，特異値分解（Singular Value Decomposition，SVD）を用いたベクトル空間の正規直交化および次元削減方式であり，小さな特異値の軸を削除することによって次元削減を行う．LSI 方式において生成されるベクトル空間上では，ベクトル空間を適切な次元数に削減することにより，文書ベクトルは単語の共起性に基づいて意味的に分類される．これらの方式は，ベクトル空間を構成する基底ベクトルの追加や削除，回転などの操作を行うことにより，文書ベクトル空間上でメタデータに基づいて生成される特徴ベクトルの配置を調整し，単語－単語間，文書－文書間，および，単語－文書間について，元のベクトル空間では得られない意味的な関連性を抽出し，それらの相関量の算出を可能とする方式と捉えることができる．文献[3]で提案されている方式や LSI

方式では，このような文書間や単語間の意味的な関連性を計量可能とするベクトル空間を用いて検索を行うことにより，問い合わせに対してパターンマッチング方式では獲得できない，意味的な文書検索を実現する．しかし，LSI 方式による特異値の大きさに基づいた次元削減が，意味的な検索を実現する文書ベクトル空間の生成において必ずしも有効とは限らない．

本研究では，LSI 方式のように特異値の大きさに基づくのではなく，ベクトル空間の軸上に配置されている文書ベクトルの分布に基づいて，ベクトル空間の次元を削減し，意味的な検索を実現する方式を示す．意味的な検索では，検索結果の再現率の改善に伴い，問い合わせに関連のない文書（以下，ノイズ文書）が増加するため，適合率が低下する傾向にある．ベクトル空間において多くの文書ベクトルが均質的に配置された軸上では，問い合わせに関連のある文書（以下，正解文書）とノイズ文書を識別することは難しい．このような軸を削除することで，正解文書と識別困難なノイズ文書を検索結果から減らすことにより，検索結果の適合率を改善することができる．

本研究では，NTCIR-1[11]により提供される検索評価用テストコレクションを検索対象データとした検索実験において，LSI 方式を用いた場合の検索結果との比較を行い，提案方式が実現可能であることを確認した．

2. 特異値分解により生成されるベクトル空間についての考察

本章では，LSI 方式に基づいて特異値分解により生成されるベクトル空間を用いた検索における考察を行う．ここでは，意味的な検索を可能とするベクトル空間の生成過程における，ベクトル空間の軸削除による次元圧縮の必要性を明らかにし，LSI 方式による特異値の大きさに基づいた軸削除方式が，必ずしも有効とは限らないことについて述べる．

例　ベクトル空間 **F**，および，**F** 上の 5 つの検索文書ベクトルからなる行列 $\mathbf{D}^{(f)}$について考察する．各文書は，二つの索引語 $\mathbf{f}_1$ と $\mathbf{f}_2$ で表されている．

$$\mathbf{D}^{(f)} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{d}_3 \\ \mathbf{d}_4 \\ \mathbf{d}_5 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 5 \\ 1 & 2 \\ 3 & 2 \\ 2 & 3 \end{bmatrix} \tag{1}$$

行列 $\mathbf{D}^{(f)}$の特異値分解により，次のような $\mathbf{r}_1$ と $\mathbf{r}_2$ を基底とする行列 **R** を得る．($\mathbf{D}^{(f)} = \mathbf{LSR}$)

$$\mathbf{R} = \begin{bmatrix} -0.46824 & 0.8836 \\ -0.8836 & -0.46824 \end{bmatrix} \tag{2}$$

行列 $\mathbf{D}^{(r)}$は，ベクトル空間 **R** に射影された文書ベクトルを表している．

$$\mathbf{D}^{(r)} = \mathbf{D}^{(f)} \times \mathbf{R} = \begin{bmatrix} -1.4047 & 2.6508 \\ -4.418 & -2.3412 \\ -2.2354 & -0.05288 \\ -3.1719 & 1.7143 \\ -3.5873 & 0.3628 \end{bmatrix} \tag{3}$$

図 1 は，二つのベクトル空間における文書ベクトルの分布を表している．一方のベクトル空間 **F** は $\mathbf{f}_1$ と $\mathbf{f}_2$ を基底とし，もう片方のベクトル空間 **R** は $\mathbf{r}_1$ と $\mathbf{r}_2$ を基底としている．

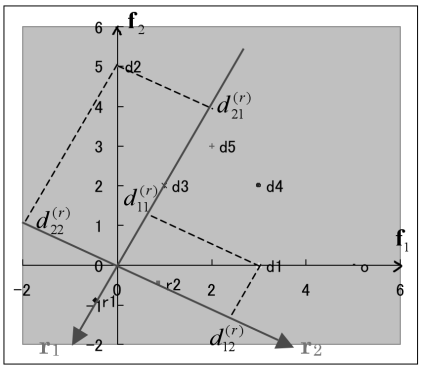

図 1　2 つのベクトル空間における文書ベクトルの分布

ここで，問い合わせベクトル **q** と文書ベクトル $\mathbf{d}_i$ がベクトル空間で直交している場合，それらの内積値は 0 である．これらのベクトルが新しい直交空間 **R**（$\mathbf{R} \times \mathbf{R}' = \mathbf{I}$）に射影された場合もまた，それらの内積値は 0 である．すなわち，新しいベクトル空間 **R** では，問い合わせベクトル **q** を用いて，文書ベクトル $\mathbf{d}_i$ を内積値やコサイン値に基づいた順位付けにより獲得することができない．図 1 で示したように，文書ベクトル $\mathbf{d}_1$ と $\mathbf{d}_2$ は，$\mathbf{f}_1$ と $\mathbf{f}_2$ を基底とするベクトル空間 **F** で直交している．この直交した性質は，$\mathbf{r}_1$ と $\mathbf{r}_2$ を基底とする新しいベクトル空間 **R** でも変わらない．

LSI 方式は，特異値分解を用いて直交空間を生成する方式であり，直交空間を生成した後，n 次元から k 次元（$k < n$）に空間を圧縮するための操作が必要である．LSI 方式では，行列 $\mathbf{D}^{(f)}$の特異値分解により，ベクトル空間 **R** を生成する．本稿で示す方式は，文献[6]で述べられているのとは異なり，ベクトル空間 **R** は，ベクトル空間 **F** の基底ベクトルを回転させることによって生成される空間であるとみなす．

空間の次元削減は，ベクトル空間 **R** から基底ベクトル（以下，軸）を取り除く操作である．LSI 方式では，小さな特異値の軸から順に取り除く．すなわち，**R** が次の n 基底ベクトルで構成される場合，

$$\mathbf{R} = \left[\mathbf{r}_1 \quad \mathbf{r}_2 \quad \cdots \quad \mathbf{r}_n\right] \tag{4}$$

軸 $\mathbf{r}_n$ が最初に取り除かれ，次いで，$\mathbf{r}_{n-1}$，$\mathbf{r}_{n-2}$，…と削除される．しかし，取り除かれた軸が，メディアデータに基づいて生成される特徴ベクトルを意味的な要素に従い適切に配置するために不要であり，取り除かれていない軸が，本当にメディアデータに基づいて生成される特徴ベクトルを意味的な要素に従い適切に配置するために必要であるかは明確ではない．

図 1 中の $\mathbf{r}_1$ と $\mathbf{r}_2$ を基底とする新しいベクトル空間 **R** では，LSI 方式による次元圧縮により，小さな特異値の軸 $\mathbf{r}_2$ が削除される．しかし，この場合，軸 $\mathbf{r}_1$ を取り除くほうがよいと考えられる．このことについて，ベクトル空間 **R** を用いて，3 通りの問い合わせに対して検索を行い，軸 $\mathbf{r}_2$ を削除した場合と軸 $\mathbf{r}_1$ を削除した場合で，検索結果を比較する．

式(1)の行列 $\mathbf{D}^{(l)}$ の二つの索引語 $\mathbf{f}_1$ と $\mathbf{f}_2$ を用いて，$\mathbf{q}_1$=[1 0]，$\mathbf{q}_2$=[0 1]，$\mathbf{q}_3$=[1 1]三つの問い合わせベクトルについて考察する．各問い合わせに対して，軸 $\mathbf{r}_2$ を削除した場合と軸 $\mathbf{r}_1$ を削除した場合で，それぞれ検索した結果を表 1 から表 3 にまとめる．なお，類似度の計量にはコサイン尺度を用いている．

表 1 問い合わせ $\mathbf{q}_1$ に対する検索結果の比較

軸 $\mathbf{r}_2$ を削除（LSI 方式）			軸 $\mathbf{r}_1$ を削除		
順位	文書	相関量	順位	文書	相関量
1	d_1	1	1	d_1	1
1	d_2	1	1	d_4	1
1	d_3	1	1	d_5	1
1	d_4	1	4	d_2	-1
1	d_5	1	4	d_3	-1

表 2 問い合わせ $\mathbf{q}_2$ に対する検索結果の比較

軸 $\mathbf{r}_2$ を削除（LSI 方式）			軸 $\mathbf{r}_1$ を削除		
順位	文書	相関量	順位	文書	相関量
1	d_1	1	1	d_2	1
1	d_2	1	1	d_3	1
1	d_3	1	3	d_1	-1
1	d_4	1	3	d_4	-1
1	d_5	1	3	d_5	-1

表 3 問い合わせ $\mathbf{q}_3$ に対する検索結果の比較

軸 $\mathbf{r}_2$ を削除（LSI 方式）			軸 $\mathbf{r}_1$ を削除		
順位	文書	相関量	順位	文書	相関量
1	d_1	1	1	d_1	1
1	d_2	1	1	d_4	1
1	d_3	1	1	d_5	1
1	d_4	1	4	d_2	-1
1	d_5	1	4	d_3	-1

表 1 から表 3 に示すように，軸 $\mathbf{r}_2$ を削除した場合は，全ての検索文書が同じ類似度を持つことになり，順位付けすることが難しい．しかし，軸 $\mathbf{r}_1$ を削除した場合では，二つの分類（d_1，d_4，d_5）および（d_2，d_3）を作り，順位付けすることができる．

3. 提案方式

本章では，提案方式である，ベクトル空間の軸上に配置されている文書ベクトルの分布に基づいて，ベクトル空間の次元を削減することにより，元のベクトル空間では得られない文書間や単語間の意味的な関連性の計量を可能とする文書ベクトル空間を生成する方式について述べる．

提案方式は，ベクトル空間において，多くの文書ベクトルが均質的に配置された軸を削除することによって，正解文書とノイズ文書についての識別が可能な軸のみで構成される文書ベクトル空間を生成する方式である．本研究では，特異値分解により生成したベクトル空間を対象として提案方式を適用する．特異値分解は，検索文書と問い合わせを射影する直交ベクトル空間を生成するために実行される．提案方式により，適切な次元数にベクトル空間を圧縮させることで，文書ベクトルはそのベクトル空間上で適切に配置され，利用者は問い合わせについて意味的に関連性のある文書を獲得することができる．

3.1. 基本的な考え方

意味的な検索では，検索結果の再現率の改善に伴い，ノイズ文書が増加するため，適合率が低下する傾向にある．すなわち，再現率と適合率の間にはトレードオフの関係が見られる．ベクトル空間において多くの文書ベクトルが均質的に配置された軸上では，正解文書とノイズ文書を識別することは難しい．このような軸を削除することで，正解文書と識別困難なノイズ文書を検索結果から減らすことにより，検索結果の適合率を改善することができる．

提案方式により，ベクトル空間の空間軸を削除するための基本的な考え方について説明する．図 2 は，軸 C_1 および軸 C_2 の二つの軸より構成される文書ベクトル空間 **S**，および，その文書ベクトル空間上に配置されている文書ベクトルを表している．図中の白点は，ある問い合わせについて，意味的な関連性を持った正解文書を表す文書ベクトル(以下,正解文書ベクトル）を示しており，黒点は，意味的な関連性を持たないノイズ文書を表す文書ベクトル（以下，ノイズ文書ベクトル）を示している．この図において，軸 C_2 上では，正解文書ベクトルとノイズ文書ベクトルが均質的に，混在して配置されているため，軸 C_2 上で，これらの正解文書とノイズ文書を区別することが困難であることがわかる．しかしながら，軸 C_1 上では，正解文書ベクトルとノイズ文書ベクトルが分散して配置されているため，軸 C_1 上で，正解文書とノイズ文書を区別することができる．したがって，文書ベクトル空間 **S** を用いた検索において，軸 C_2 を削除し，軸 C_1 を使用することにより，検索結果からノイズ文書群を除外し，この問い合わせに対する適合率を改善することが期待される．

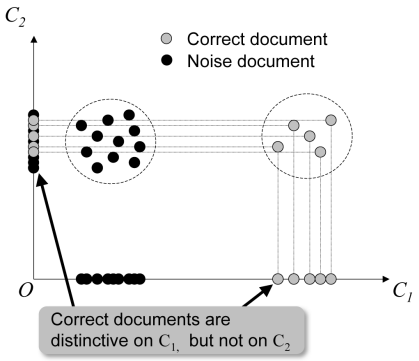

図 2 文書ベクトル空間 S 上に配置されている文書ベクトル

3.2. 次元圧縮による文書空間の生成方式

前節に示した基本的な考え方に基づき，提案方式では，文書ベクトル空間から，多くの文書ベクトルが均質的に配置された軸を削除することにより，元のベクトル空間では得られない文書間や単語間の意味的な関連性の計量を可能とする文書ベクトル空間を生成する．提案方式による文書ベクトル空間の生成ステップを下記に示す．

Step-1 検索対象となる文書集合 D を用意し，この文書集合 D より，文書集合と各文書に出現する索引語の関係を表す文書・索引語行列 **D** を作成する．

$$\mathbf{D}=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \tag{5}$$

ここで，文書・索引語行列 **D** の各要素 d_{ij} は，文書集合 D 中の文書 d_i における単語 t_j の出現回数を示している．単語 t_j は，文書集合 D 中に出現する単語集合 T の要素である．

Step-2 文書・索引語行列 **D** を特異値分解することにより，正規直交空間 **L** を得る．式(4.2)で，**S** は特異値を含む対角行列であり，**L** は行列 **S** の左行列，**R** は行列 **S** の右行列である．直交空間 **LS** は，文書間の意味的な関連性を計量するための文書ベクトル空間である．

$$\mathbf{D}=\mathbf{L}\times\mathbf{S}\times\mathbf{R}^T \tag{6}$$

Step-3 文書ベクトル空間 **LS** において，軸上に配置されている文書ベクトルの割合が，閾値 ε_1 より大きい軸を選択し削除する．これは，文書ベクトル空間 **LS** において，多くの文書ベクトルが均質的に配置された軸を削除する操作を意味している．

$$\frac{n_i-n_d}{N}\geq\varepsilon_1 \tag{7}$$

ここで，N は検索対象となる文書ベクトルの総数，n_i は軸 C_i 上に配置されている文書ベクトル数を表している．n_d は軸 C_i 上において座標値の絶対値が閾値 ε_2 より小さい文書ベクトル数を表しており，式(8)を満たす文書ベクトル数である．$v_i(\mathbf{d}_k)$は軸 C_i 上の文書ベクトル $\mathbf{d}_k$ の座標値の絶対値である．

$$v_i(\mathbf{d}_k)\leq\varepsilon_2 \tag{8}$$

文書ベクトル空間 **LS** において，**S** は特異値を含む対角行列であるので，文書ベクトル空間 **LS** は，文書ベクトル空間 **L** の各軸上にある文書ベクトルの座標値を，各軸に対応する特異値の分だけ拡大縮小した座標値を持つ文書ベクトルの分布を表している．文書ベクトル空間 **LS** における各軸上の文書ベクトルの分布は，各軸に対応する特異値に依存するため，式(8)を満たす文書ベクトル数 n_d を求める際に，閾値 ε_2 を設定することが困難である．このため文書ベクトル空間 **LS** において，式(7)および式(8)により，削除する空間軸を判定するために，文書ベクトルの座標値が特異値に依存しない直交空間 **L** の各行を 2 ノルム正規化した行列を用いる．

さらに，軸 C_i において，正負別の判定を行う場合は，式(7)の代わりに次式を用いる．

$$\frac{n^+{}_i-n^+{}_d}{N}\geq\varepsilon_1 \ \wedge\ \frac{n^-{}_i-n^-{}_d}{N}\geq\varepsilon_1 \tag{9}$$

ここで，n_i^+ と n_i^-は軸 C_i 上に正軸，および負軸に配置されている文書ベクトル数を表している．また，n_d^+と n_d^-は軸 C_i 上の正軸，および負軸において座標値の絶対値が閾値 ε_2 より小さい文書ベクトル数を表している．

Step-4 Step-3 により，文書ベクトル空間 **LS** から，多くの文書ベクトルが均質的に配置された軸を削除することにより，元のベクトル空間 **LS** では得られない文書間や単語間の意味的な関連性の計量を可能とする文書ベクトル空間 $\mathbf{S}_u$ を生成する．

3.3. 計量方式

Step-3 で述べたように，特異値分解により生成された文書ベクトル空間 **LS** では，各文書ベクトルは，各軸に対応する特異値に応じて拡大縮小する．そのため，文書ベクトルの配置の度合いは，ベクトル空間の各軸上で異なっている．したがって，ユークリッド距離や内積は，問い合わせベクトルと文書ベクトルの距離を

測定するために有効とはいえない．そこで，Step-4 により生成された文書ベクトル空間 $\mathbf{S}_u$ を用いた検索において，問い合わせと検索文書の関連性を計量するために，問い合わせベクトル $\mathbf{q}$ と文書ベクトル $\mathbf{d}_i$ の角度を示すコサイン尺度を用いる．

$$\cos(\mathbf{d}_i,\mathbf{q}) = \frac{\mathbf{d}_i' \times \mathbf{q}}{\sqrt{(\mathbf{d}_i' \times \mathbf{d}_i)}\sqrt{(\mathbf{q}' \times \mathbf{q})}} \quad (10)$$

4. 実験

本章では，提案方式により生成した文書ベクトル空間を用いた検索実験について述べる．ここでは，生成した文書ベクトル空間を用いた検索の実現可能性を確認する．本実験では，疑似文書データおよびNTCIR-1により提供される検索評価用テストコレクションを検索データとして検索実験を行う．提案方式とLSI方式により生成した文書ベクトル空間を用いた検索を行い，両者の検索結果を比較することで，提案方式が有効であることを確認する．

NTCIR-1 は，「学会発表データベース」から抽出した学会発表論文要旨約 33 万件が集められている．NTCIR-1 には，日本語版を対象としたJコレクションと，英語版を対象としたEコレクションがある．実験2では，Eコレクションを文書データとして使用する．E コレクションは，(1) 文書，(2) 83 個の検索課題，(3)正解文書リストなどから構成されている．正解文書リストは，各検索課題に適合する文書の正解判定のリストである．各検索課題の正解文書は，“A 判定”が設定され，不正解文書は，“C 判定”が設定されている．また，正解ではないが，検索課題にある程度の関連がある文書については，“B 判定”が設定されている．

本実験では，E コレクションの文書集合から論文のタイトルと抄録 95 件を文書データとして抽出した．検索課題については，A 判定である文書が 10 件以上設定されており，分野が異なるものを 5 つ選んでいる．また，各検索課題について C 判定である文書は，それぞれ 15 件を上限にして抽出を行った．表 4 に実験に用いた文書データの詳細を示す．なお，前処理として単語に対してステミング処理を行っている．

表 4 文書データの詳細

課題番号	トピック	A	B	C	合計
108	XML	5	-	15	20
116	Barrier-free design of public facilities	5	-	15	20
119	Values (in Japan)	5	-	15	20
123	Biofilms	5	-	12	17
135	Phase transition of vacuum	5	-	13	18
合計	-	25	-	70	95

表 4 に基づいて文書・単語行列 $\mathbf{D}_1$ を作成する．3.2. 節の Step-2 に示したように，文書・単語行列 $\mathbf{D}_1$ を特異値分解することにより得られる正規直交空間 $\mathbf{L}$ を対象として提案方式を適用する．直交空間 $\mathbf{L}$ 上の軸には，大きな特異値の順に連続して 1，2，3，…と番号を割り当てる．3.2.節の Step-3 の式(7)，(8)に従い，多くの文書ベクトルが均質的に配置された軸を削除することにより，文書ベクトル空間 $\mathbf{S}_1$ を生成する．また，文書ベクトル分布を正負の軸ごとに区別するために，式(8)，(9)に従って，文書ベクトル空間 $\mathbf{S}_2$ を生成する．

問い合わせとして“japanese lifestyle”を使用し，文書ベクトル空間 $\mathbf{S}_1$，$\mathbf{S}_2$ と LSI 方式より生成した文書ベクトル空間を用いて検索実験を行い，検索結果を比較する．この問い合わせに対して，文書 d_{41}～d_{45} を正解文書として設定している．

文書ベクトル空間 $\mathbf{S}_1$，$\mathbf{S}_2$ と LSI 方式により生成した文書ベクトル空間を用いて検索を行った結果として，再現率・適合率グラフを図 3 に示す．表 5，表 6 は，文書ベクトル空間 $\mathbf{S}_1$，$\mathbf{S}_2$ を次元削減したときに構成する軸番号を示している．なお，$\mathbf{S}_1$，$\mathbf{S}_2$ の最初の次元数は，95 次元である（95 文書・1908 単語）．

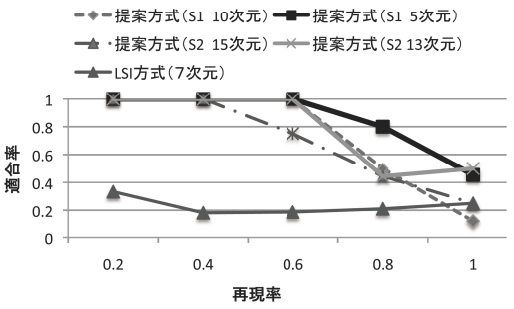

図 3 再現率・適合率グラフ

表 5 文書ベクトル空間 S_1 を構成する軸番号

次元数	軸番号
10	4，16，19，73，81，90，91，93，94，95
5	4，16，93，94，95

表 6 文書ベクトル空間 S_2 を構成する軸番号

次元数	軸番号
15	4, 14, 15, 16, 73, 78, 81, 87, 88, 90, 91, 92, 93, 94, 95
13	14, 15, 16, 73, 81, 87, 88, 90, 91, 92, 93, 94, 95

提案方式では，$\mathbf{S}_1$ では 5 次元，$\mathbf{S}_2$ では 13 次元，で最も適合率が改善されていることが確認できた．更に，提案方式による $\mathbf{S}_1$ (5 次元) および $\mathbf{S}_2$ (13 次元)の検索結果では，問い合わせである“japanese lifestyle”を含

んでいない文書 d_{42}, d_{44} および d_{45} を上位に確認することができた．これは，提案方式により生成した文書ベクトル空間では，パターンマッチングではなく文書の意味的な検索を可能にしていることを示している．

一方，LSI 方式で最も良い検索結果を得られたのは 7 次元であるが，その適合率は提案方式による検索結果より低い．これは，問い合わせについて，意味的に関連性のある適切な文書を獲得するためには，大きな特異値の軸が，常に有効であるとは限らないことを意味している．

以上の実験結果により，提案方式により生成した文書ベクトル空間を用いた検索において，意味的な文書獲得が実現可能であり，提案方式が有効であることを確認できた．

5. まとめと今後の課題

本稿では，ベクトル空間の軸上に配置されている文書ベクトルの分布に基づいた，ベクトル空間の次元削減による文書ベクトル空間を生成する方式を示した．さらに，特異値分解により得られたベクトル空間を対象として，提案方式により生成された文書ベクトル空間を用いた場合の検索と，LSI（Latent Semantic Indexing）方式により，特異値の大きさに基づいた次元圧縮により生成されたベクトル空間を用いた場合の検索との比較実験を行うことで，提案方式の実現可能性を確認した．

LSI 方式において生成される文書ベクトル空間上では，ベクトル空間を適切な次元数に削減することにより，文書ベクトルは単語の共起性に基づいて意味的に分類され，意味的な検索が可能となる．しかし，単語の共起性に基づいた分類が困難である場合，LSI 方式では検索精度が低下する可能性がある．本研究では，このような場合においても，提案方式により，多くの文書ベクトルが均質的に配置された軸を削除することにより，正解文書とノイズ文書についての識別が可能な軸のみで構成される文書ベクトル空間を生成し，意味的な文書検索を実現することができる可能性を示した．

本研究の今後の課題として，様々な分野の大規模文書データ群を対象とした検索実験により，提案方式のスケーラビリティに関する有効性を検証するとともに，利用者の問い合わせの意味や内容に応じたベクトル空間軸の選択機能による，利用者の意図や関心に応じたダイナミックな検索方式の実現が期待される．

文　献

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[2] M. W. Berry, S. T. Dumais and G. W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Review, 37(4), 1995, 573-595.

[3] X. Chen and Y. Kiyoki, *A query-meaning Recognition Method with a Learning Mechanism for Document Information Retrieval*, Information Modelling and Knowledge Bases XV (IOS Press), Vol. 105, 2004, 37-54.

[4] X. Chen, Y. Kiyoki, K. Takano and K. Masuda, *A Semantic Space Creation Method with an Adaptive Axis Adjustment Mechanism for Media Data Retrieval*, The 17th European Japanese Conference on Information Modelling and Knowledge Bases, 2007.

[5] W.S. Cooper, *On deriving design equations for information retrieval systems*, JASIS, 1970, 385-395.

[6] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, 41(6), 1990, 391-407.

[7] H. Ohuchi, T. Miura and I. Shioya, *Document Retrieval by Projection Based Frequency Distribution*, Workshop on the Institute of Electronics, Information and Communication Engineers, 2005.

[8] S. Kaski, *Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering*, Proceedings of IJCNN'98, International Joint Conference on Neural Networks, 1998, 298-315.

[9] T. Kitagawa and Y. Kiyoki, *A mathematical model of meaning and its application to multidatabase systems*, Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, April 1993, 130-135.

[10] D. D. Lewis, R. E. Schapire, J. P. Callan and R. Papka, *Training algorithms for linear text classifiers*, SIGIR, 1996, 413-418.

[11] NTCIR: http://research.nii.ac.jp/ntcir/

[12] C. H. Papadimitriou, P. Raghavan, H. Tamaki and S. Vempala, *Latent semantic indexing: A probabilistic analysis*, In Proc. 17th ACM Symp. On the Principles of Database Systems, 1998, 159-168.

[13] G. Salton, *The SMART retrieval system - Experiments in automatic document processing*, Prentice-Hall Inc, Englewood Cliffs, New Jersey, 1971.

[14] H. Schutze and C. Silverstein, *Projections for efficient document clustering*, Proceedings of SIGIR'97, 1997, 74-81.

[15] S. K. M. Wong, W. Ziarko, P. C. N. Wong, *Generalized Vector Space Model in Information Retrieval*, SIGIR, 1985, 18-25.