情 報 学 基 礎　66－12
デジタル・ドキュメント 32－12
（２００２．３．１５）

多変量解析に基づいた情報検索手法の比較検討

日本 IBM(株) 東京基礎研究所 竹内広宜 小林メイ 青野雅樹 寒川光
E-mail: {hironori, mei, aono, samukawa}@jp.ibm.com

ベクトル空間モデルを使った情報検索システムとして Latent Semantic Indexing 法 (LSI 法) が提案されている。LSI 法は文書-キーワード行列の特異値分解によって得られる直交ベクトルを用いて文書ベクトルを低次元空間に射影する方法である。一方、統計学の多変量解析法の一つに多次元データを最も散らばるように低次元空間に射影する方法である主成分分析法があり、LSI 法も主成分分析法であると考えられている。本稿ではキーワード間共分散行列の主成分分析を情報検索に適用する方法について述べ、LSI 法との様々な観点から比較検討を行う。

An Information Retrieval System using a Principal Component Analysis

IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
Hironori Takeuchi Mei Kobayashi Masaki Aono Hikaru Samukawa
E-mail: {hironori, mei, aono, samukawa}@jp.ibm.com

In this paper, we consider a vector space model based information retrieval system using a principal component analysis. Latent Semantic Indexing (LSI) is based on the singular value decomposition to the document by term matrix and is used as the method of dimension reduction. We introduce an information retrieval method using a principal component analysis to the keyword covariance matrix and compare this method with LSI.

1　はじめに

近年、膨大な電子化された文書がデータベース中に蓄えられている。例えば、企業内には様々な形式で文書データが埋もれており、そのようなデータを有効に活用することが必要不可欠になってきている。このような問題を解決する手段としてテキストマイニング技術が近年多く開発され、膨大な文書集合からの知識獲得に応用されている [7]。文書データに対する情報検索技術は、テキストマイニングの技術としては基本的なものであるが、簡単に使用できることからテキストデータからの情報取得方法として最も一般的に使われている技術の一つである。

しかしながら、現在の文書検索技術にも問題がある。例えば、インターネットの世界では Google などのサーチエンジンが活躍しているがキーワードとして入力したクエリーに対して膨大な数のページがヒットしてしまい、なかなか必要な情報がみつからないことが問題となっている。また、Web ページの検索だけでなく企業の文書データベースにおいても、情報の再利用の点からも性能が良い検索システムが必要となっている。

これまで様々な情報検索手法が提案されているが、本稿ではベクトル空間モデルに基づいた類似文書検索技術に焦点をあてる。ベクトル空間モデルは基本的には単語 (キーワード) を属性とし、その重みを要素としたベクトルとして文書を表現するモデルである。したがって、モデル化した時点でかなりの情報を失うという問題があり不十分なモデルであるが、文書をベクトルとして扱えるために処理の大部分が対象の言語に依存しないという利点がある。本稿では多変量解析に基づいた情報検索技術について比較検討を行う。まず、代表的な手法である Latent Semantic Indexing 法 [2] (LSI 法) について述べる。LSI 法は少数の直交する factor を文書-キーワード行列の特異値分解によって求め、文書ベクトルの次元削減を行っている。この次元削減によって類似度計算の効率化や精度の向上が確認されている。LSI 法における次元削減の手法は多変量解析法の一つである主成分分析であるということが言われている。次に、キーワード間の共分散行列に基づいた主成分分析モデルで情報検索システムを表現する。そして、主成分分析法としての LSI 法について述べ、共分散行列の主成分分析法と意味、性質などの観点から比較検討を行う。

そして簡単な実験を行い、最後にまとめを述べる。

2 Latent Semantic Indexing 法

2.1 ベクトル空間モデル

ベクトル空間モデルでは文書および検索質問 (クエリー) をベクトルとして表現する。ベクトルの各要素には各キーワード (索引語) に対する重みが入る。本稿では文書数 m、キーワード数 n とし、文書集合 D を次のように定義する。

$$D = \begin{pmatrix} \boldsymbol{d}_1^t \\ \vdots \\ \boldsymbol{d}_j^t \\ \vdots \\ \boldsymbol{d}_m^t \end{pmatrix}, \quad \boldsymbol{d}_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{ij} \\ \vdots \\ d_{in} \end{pmatrix} \tag{1}$$

ここで、t は転置を表す。キーワードの重みおよび文書ベクトル間の類似度としてはいくつか提案されているが、本稿では重みとして TF-IDF に基づいた正の値、および類似度として余弦を使用する。

2.2 LSI 法

ベクトル空間モデルを使った情報検索手法の一つのして Deerweste *et al.* [2] が提案した文書-キーワード行列 D の特異値分解に基づく Latent Semantic Indexing 法 (以下 LSI 法) がある。文書-キーワード行列 D は特異値分解 (singular value decomposition: SVD) によって以下のように分解される。

$$D = T\Sigma U^t \tag{2}$$

T は左特異ベクトル (DD^t の固有ベクトル) $\boldsymbol{t}_i$ を列ベクトルとする $m \times m$ の行列、U は右特異ベクトル (D^tD の固有ベクトル) $\boldsymbol{u}_i$ を列ベクトルとする $n \times n$ の行列、Σ は D の特異値 (D^tD の固有値) σ_i を対角要素とする $n \times n$ の対角行列である。LSI 法では特異値の上位 k(経験的に 200 ぐらいが適切と言われている [3]) 番目までをとり、それら k 個の特異値に対応する右特異ベクトルを集めた行列 U_k を用いて文書ベクトルおよび、クエリーベクトルを以下のように n 次元から k 次元に次元削減する。

$$\hat{\boldsymbol{d}}_i = U_k^T \boldsymbol{d}_i, \quad \hat{\boldsymbol{q}} = U_k^T \boldsymbol{q} \tag{3}$$

LSI 法では次元削減されたクエリーベクトル $\hat{\boldsymbol{q}}$ と各文書ベクトル $\hat{\boldsymbol{d}}_i$ との類似度を計算し、類似文書検索を行う。ベクトル空間の次元を減らすことで、類似度計算が効率化されることと、クエリー中のキーワードを含まないが意味的に近い文書が類似文書として検索される可能性があるというのが LSI 法の利点である。

2.3 LSI 法のモデル

Deerwester *et al.* [2] では文書は少数の factor で表現されているという仮定を用いて LSI 法を導入している。以下ではそのモデルを定式化する。今、文書ベクトル $\boldsymbol{d}$ は k 個の factor で表現されていると仮定すると factor を表すベクトル $\boldsymbol{f}_i$ の線形和として以下のように書ける。

$$\boldsymbol{d} = \sum_{i=0}^{k} \alpha_i \boldsymbol{f}_i \tag{4}$$

各文書が k 個の factor の線形和で表現されることから、$\boldsymbol{d}$ および $\boldsymbol{f}_i$ は k 本の基底ベクトル $\boldsymbol{v}_i$ の張る空間上にあり、以下のように表される。

$$\boldsymbol{d} = \sum_{j=0}^{k} \beta_j \boldsymbol{v}_j, \quad \boldsymbol{f}_i = \sum_{j=0}^{k} \gamma_{ij} \boldsymbol{v}_j \tag{5}$$

スペクトル分離のような応用では線形和となった観測ベクトルから factor ベクトルを求める必要があるが、LSI 法では factor ベクトルを求めることを目的とはせず、観測ベクトルの低次元空間表現を求めることを目的としている。そこで、各 factor は直交するという仮定を置き、$\boldsymbol{v}_i$ を factor として求めている。そして、β_j を用いて文書の k 次元空間表現を求めている。その際、k 本の基底ベクトルを行列 D の SVD によって求めている。この基底ベクトル意味については以下の章で述べる。

3 主成分分析法に基づく情報検索と LSI 法

3.1 主成分分析法

ここでは、まず一般の主成分分析法 [6] をベクトル空間モデルに基づく情報検索へ適用する手順を示す。

(1) 式の文書-キーワード行列 D が与えられた時、キーワード間の共分散行列は以下のように定義される。

$$C = \frac{1}{m} D^t D - \overline{\boldsymbol{d}}^t \overline{\boldsymbol{d}}, \quad \overline{\boldsymbol{d}} = \frac{1}{m} \sum_{j=1}^{m} d_j \tag{6}$$

$n \times n$ の行列 C の対角成分は各キーワードが出現するばらつき、非対角成分はキーワード間の共起情報が入っていると解釈できる。共分散行列 C は固有値分解によって以下のように分解される。

$$C = V \Lambda V^t \tag{7}$$

(7) で、Λ は C の固有値 $\lambda_i (\lambda_1 \geq \cdots \geq \lambda_n)$ を対角要素に持つ $n \times n$ の対角行列、V は λ_i に対応する固有ベクトル $\boldsymbol{v}_i$ を列ベクトルに持つ $n \times n$ の行列である。LSI 法と同様、得られた固有値のうち正のものから k 番目番目までをとり、それら k 個の固有値に対応する固有ベクトルを集めた行列 V_k を用いて文書ベクトルおよび、クエリーベクトルを以下のように n 次元から k 次元に次元削減する。

$$\hat{\boldsymbol{q}} = V_k^T (\boldsymbol{q} - \overline{\boldsymbol{d}}), \quad \hat{\boldsymbol{d}}_i = V_k^T (\boldsymbol{d}_i - \overline{\boldsymbol{d}}), \tag{8}$$

次元削減を行った後の検索手順は LSI 法と同じである。

(7) での固有値分解では、元の空間 (ここでは n 次元空間) 中の平均を表す点を原点として最もデータが散らばる (分散が大きくなる) ように順番に直交する基底ベクトルを固有ベクトルとして求めている (図 1)。固有ベクトルに相当する新しい基底ベクトル方向でのデータの分散は対応する固有値となる。

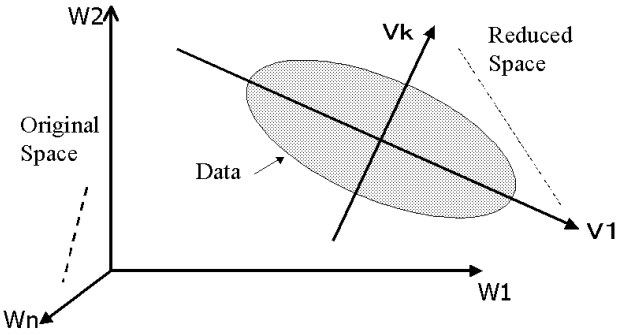

図 1: 主成分分析のイメージ

3.2 LSI 法の主成分分析法としての解釈

LSI 法では文書-キーワード行列 D の SVD によって、基底ベクトルを求めている。SVD に用いた LSI 法が主成分分析法と本質的に等価であることが [5] では述べられている。今、D に対して共分散行列の第 1 項

$$M_D = \frac{1}{m} D^t D \tag{9}$$

をモーメント行列と定義すると、SVD から得られる LSI 法での基底ベクトルは M の固有ベクトルになっている。一方、各文書ベクトルに対して平均ベクトルを引いて平均を 0 にすると共分散行列はモーメント行列と一致する。リモートセンシングにおける多重スペクトル解析や画像解析に分野では、factor ベクトルの推定にこのモーメント行列 M を用いた主成分分析法が主に用いられている。しかしながら、3.1 の主成分分析とは固有値分解を行う行列が異なるため、得られる基底ベクトルの性質は異なるものとなる。それについて以下の節で述べる。

3.3 幾何学的解釈による比較

3.1 で述べたように、主成分分析は高次元データを散らばりが最も保持される形で低次元空間に射影する方法である。従って、類似したデータはより類似する方向に、類似していないデータはより類似しない方向に射影される。この時、射影方向のデータの分散および射影方向を表す方向ベクトルは共分散行列の固有値、およびそれに対応する固有ベクトルをして求められる。このような幾何学的解釈に基づいて、SVD から求められる LSI 法での基底ベクトルを考察する。

元の空間において文書ベクトル (重みは全て正と仮定) は、全ての座標が正である第 1 象限に存在する。今、全ての文書ベクトルを原点に対して対称変換した文書ベクトルの集合を考える。この文書ベクトルの集合は $-D$ で表される。ここで、元の文書集合と対称変換した文書集合を合わせた $2m \times n$ の文書-キーワード行列

$$\tilde{D} = \begin{pmatrix} D \\ -D \end{pmatrix} \tag{10}$$

を考える。$\tilde{D}$ 中の文書ベクトルの平均は 0 である。したがって、$\tilde{D}$ の共分散行列は

$$\begin{aligned} C_{\tilde{D}} &= \frac{1}{2m}\tilde{D}^t\tilde{D} \\ &= \frac{1}{2m}\begin{pmatrix} D^t & -D^t \end{pmatrix}\begin{pmatrix} D \\ -D \end{pmatrix} \\ &= M_D \end{aligned} \tag{11}$$

となり、D のモーメント行列と一致する。したがって、LSI 法で得られる基底ベクトルは $\tilde{D}$ の共分散行列の固有値分解によって得られる固有ベクトルと一致する。

共分散行列による主成分分析の場合、得られた固有ベクトルは対応する固有値の大きい順にデータの分散情報を保持している。これを幾何学的に示すと、D および $\tilde{D}$ の共分散行列による主成分分析は図 2 および図 3 のようになる。 図で、実線は元の空間の基

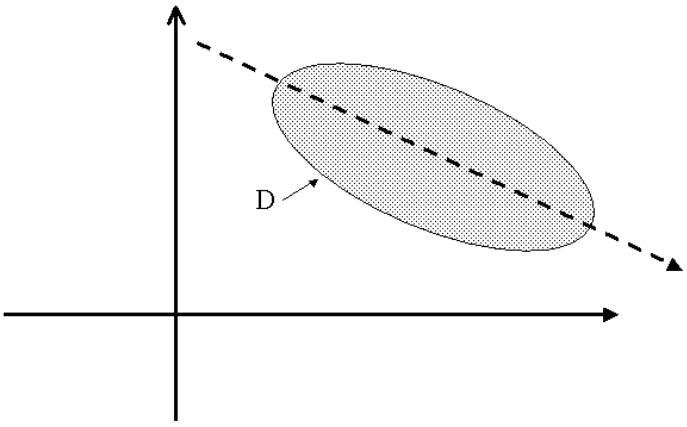

図 2: D の共分散行列による主成分分析

底ベクトル、破線は固有値分解で得られるデータを射影する方向ベクトルを示している。D の共分散行列による主成分分析法では文書集合中のデータが最も散らばる方向に射影空間を構築しているが、LSI 法では $\tilde{D}$ 中のデータが最も散らばる方向に射影空間を構築するため、元の文書集合中のデータは必ずしも散らばる方向に射影されるとは限らないことがわかる。また、低次元空間へのデータの射影では D の共分散行列による主成分分析法では文書集合の平均ベクトル位置に原点が移動するが、LSI 法では原点の位

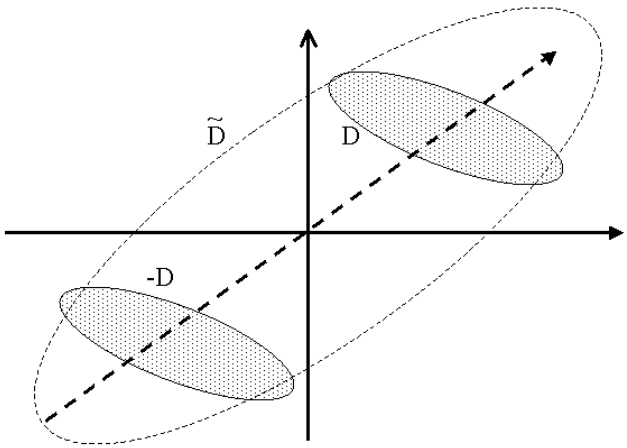

図 3: $\tilde{D}$ の共分散行列による主成分分析

置は移動しないことになる。そのため、LSI 法では余弦で類似度を測った場合に異なる位置にある 2 点を高い類似度で検出する可能性がある。

情報検索では、クエリーに対して類似性のあるものだけを抽出したい。この考えに基づくと、低次元空間への射影には、最も分散情報を保持した形で射影できる D の共分散行列の主成分分析を用いる方が適切だと考えられる。また、D の共分散行列の主成分分析で得られる、固有値の和は共分散行列の対角要素 (各キーワードの分散) の和と等しいため、降順に固有値を調べることにより新しい基底ベクトル空間でデータの変動の何 % を説明できるているか (寄与率、累積寄与率) を求めることができ、この累積寄与率を指標としてデータに依存して次元削減後の次元数を決定することができる。一方で LSI 法で得られる特異値の 2 乗和は $\tilde{D}$ の共分散行列の対角要素の和と等しいため、元のデータの分散情報を保持していないため、累積寄与率によって次元数を決定することは難しい。

3.4 次元削減計算における比較

LSI 法では特異値分解 (SVD)、3.1 の主成分分析ではキーワード間共分散行列の固有値分解における行列計算が重要な位置を占める。一般に対称行列の固有値分解は行列の 3 重対角化に多くの計算時間がか

かる。ここでは、両手法に用いられる固有値分解における行列計算 (3 重対角化) とその特徴についてまとめる。

大型対称行列の固有値分解は、いったん 3 重対角行列に変換する方法が一般である。対称行列 A の 3 重対角化とは $P^tAP = \Phi$

$$\Phi = \begin{pmatrix} \alpha_1 & \beta_1 & & & O \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1} \\ O & & & \beta_{n-1} & \alpha_n \end{pmatrix} \tag{12}$$

の形に行列を変換することである (P は直交行列)。3 重対角行列の固有値分解は比較的用意で、大型行列の場合 3 重対角化に要する手間に比べて無視できるほど少ない。Φ の固有値、固有ベクトルの組を $(\lambda_{\Phi i}, \boldsymbol{x}_i)$ として求まると、A の固有値、固有ベクトルは $(\lambda_{\Phi i}, P\boldsymbol{x}_i)$ として求まる。

LSI 法では文書-キーワード 行列 D に対して SVD を行う。D は高次元な行列であるが、非常に疎な行列である。対称疎行列の固有値および固有ベクトルを求める方法として Lanczos 法がある [4]。この方法は固有値の大きい順に必要な数だけ固有値およびそれに対応する固有ベクトルを求める方法であるため、次元削減という観点で考えると非常に望ましい。LSI では D の特異値分解は次のような対称行列 B

$$G = \begin{pmatrix} O & D \\ D^t & O \end{pmatrix} \tag{13}$$

の固有値分解という形で行う。この対称疎行列 G の固有値分解は

$$G = Q \begin{pmatrix} \Sigma & & O \\ & -\Sigma & \\ O & & O \end{pmatrix} Q^t \tag{14}$$

となり、Q の中に T, U の情報が含まれている。ここで、上位 k 個の固有値、固有ベクトルをが必要な場合、k 次まで、G の 3 重対角化を行えばよい。Lanczos 法は k 次までの 3 重対角化を反復計算で行う方法である。G が 3 重対角化によって (12) 式で示される Φ に変換されるとする。まず、適当なノルム 1 のベクトル $\boldsymbol{p}_1$ を作り、$\alpha_1, \beta_1, \boldsymbol{p}_2$ を以下のように求める。

$$\begin{cases} \alpha_1 = \boldsymbol{p}_1^t G \boldsymbol{p}_1 \\ \beta_1 = \|G\boldsymbol{p}_1 - \alpha_1 \boldsymbol{p}_1\| \\ \boldsymbol{p}_2 = (G\boldsymbol{p}_1 \alpha_1 \boldsymbol{p}_1)/\beta_1 \end{cases} \tag{15}$$

を求める。(15) 式から反復的に以下のように k 次までの 3 重対角化を行う。

$$\begin{cases} \alpha_i = \boldsymbol{p}_i^t G \boldsymbol{p}_i \\ \beta_i = \|G\boldsymbol{p}_i - \beta_{i-1}\boldsymbol{p}_{i-1} - \alpha_i \boldsymbol{p}_i\| \\ \boldsymbol{p}_{i+1} = (G\boldsymbol{p}_i \beta_{i-1}\boldsymbol{p}_{i-1} - \alpha_i \boldsymbol{p}_i)/\beta_i \end{cases} \tag{16}$$

行列 G が疎行列である場合、(16) 式の反復計算で最もコストがかかる $G\boldsymbol{p}_i$ の計算を短縮できるため、効率よく計算できる。文書-キーワード 行列の非零の割合を r(通常の行列計算では $r \approx 0.01$) とすると計算量は $O(rmn)$ となる。文書-キーワード 行列の特異値分解はこのように Lanczos 法を用いることで効率よく行うことができる。しかしながら、文書-キーワード 行列の特異値分解の計算量、および必要とするメモリの大きさは文書数 m に比例することになる。そのため文書数が非常に大きくなってくると計算が困難になってしまうという問題が残ってしまう。

次にキーワード 間共分散行列の固有値分解を 2 種類考える。式 (6) の共分散行列 C は密行列である。したがって、C を陽に作ってしまった場合、Lanczos 法は用いることができない。このような密対称行列に対しては Householder 変換を用いて次のように 3 重対角化を行う。

$$H_n \cdots H_1 C H_1 \cdots H_n = \Phi \tag{17}$$

Householder 変換は密行列の 3 重対角化で最もすぐれた方法として知られているが $O(n^3)$ の計算量と $O(n^2)$ のメモリを必要とする。Lanczos 法のような効率化はできないが構成度である。また、文書数が非常に大きい集合に対してもキーワード 数を固定することで固有値分解の計算が可能となる。一方で、C を陽に作らない方法も考えられる。密行列 C を Lanczos 法に適用することができないのは $C\boldsymbol{p}_i$ の計算が効率良く行うことができないからだが、$C\boldsymbol{p}_i$ の計算を C を陽に作らずに

$$C\boldsymbol{p}_i = \frac{1}{m} D^t D \boldsymbol{p}_i + \overline{\boldsymbol{d}\boldsymbol{d}}^t \boldsymbol{p}_i \tag{18}$$

と分け、それぞれの項で右側からの演算 ($\boldsymbol{p}_i$ から) を行うことを考える。すると第 2 項の計算は $O(n)$ で行え、第 1 項は $O(rmn)$ の計算で行えることがわかる。このように共分散行列を陽に作らないことで Lanczos 法を適用して効率化することができる。ただし、この場合は LSI 法と同様、文書数が非常に大きくなった場合は計算が困難になってしまう。

SVD に基づいた LSI 法は疎行列の特性を使って効率よく次元削減を行うことができるが、文書数が膨大になった場合は計算が困難という問題があるが、共分散行列の主成分分析の場合は文書数が膨大になった場合になっても陽に共分散行列を作ることで対応することができることがいえる。

4 実験

ここではキーワード間共分散行列による主成分分析に基づいた次元削減手法についての簡単な検索実験結果を述べる。データとしては TREC のドキュメントセットのうち LA Times の 127741 文書を用いた。このデータに対してキーワードを切り出し、127741×9770 の文書-キーワード行列を作成した。この行列からキーワード間共分散行列を作成し、固有値分解を行い 200 次元に次元を削減した。

これに対して TREC6 の検索課題 (301-310) を使って検索の評価を行った。検索結果の評価方法にはいろいろな手法があり、それぞれいろいろな特徴、問題点がある [8]。Voorhees *et al.* [9] では、いくつかの評価尺度についてその相関関係を述べている。ここでは、P(10) と呼ばれる類似度ランキング結果の上位 10 位での精度 (precision) [9] を用いた。共分散行列を用いた主成分分析法での P(10) の平均は 13.8% であった。精度が高いと評価されているシステム [1] で P(10) の平均が約 26% であるから、本手法は検索手法としては未だに不十分なものだといえる。

主成分分析を用いた次元削減では、類義語を抽出することができる。実用的な検索システムの構築を考える場合には、この情報を検索結果に加えることで、キーワードによる絞り込み検索のナビゲーションを行うことができる。例えば、まず最初に思い付くままにクエリーを入力し、類似文書検索を行い、その後、得られた結果および類義語などの付加情報を使って絞り込み検索を行うということが考えられる。

5 まとめ

本稿では、ベクトル空間モデルに基づいた検索手法について検討した。LSI 法は代表的な手法であるが、主成分分析法の 1 つとして考えられているが、統計学で通常用いられる共分散行列に基づく主成分分析とは次元を削減する手法が異なることを示した。共分散行列の主成分分析を用いることで、似ているものを近くに配置し、似ていないものを遠くに配置することができるので情報検索には適した次元削減だと考えられる。計算量の観点で見ると、共分散行列の主成分分析はたとえ、共分散行列という密対称行列であっても陽に行列を計算しないことで効率良く固有値分解が可能であるし、文書数が膨大になっても行列を陽に作ることで次元削減計算が可能であることがわかった。

ベクトル空間モデルはモデルが簡単であるため容易に扱えるが、モデル化した段階で非常に情報を失っており、性能に影響を与えていると考えられる。そのため実用的なシステムを考えた場合には類義語の情報などを検索結果に付与し、利用者に絞り込み検索などが簡単に行えるようにする必要がある。一方で、モデル化の段階でベクトルの属性に用いるキーワードの選択が重要である。多くの場合、頻度に基づいているが、文書分類などに用いるキーワード選択 [10] のようにベクトルの属性を選択することで性能がある程度向上することが考えられる。

参考文献

[1] D.Carmel, D.Cohen, R.Fagin, E.Farchi, M.Herscovici, Y.S.Maarek and A.Soffer. Static Index Pruning for Information Retrieval Systems, In *Proc. SIGIR'01*, 43-50, 2001.

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

[3] S.T.Dumais. LSI meets TREC:A status report. In D.Harman(Ed.) *The First Text REtrieval Conference(TREC-1)*. NIST special publication 500-207, 137-152, 1993.

[4] A.Jennings and J.J.McKeown, Marix Computation, Wiley, 1992.

[5] 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, 2001.

[6] K.V.Mardia, J.T.Kent and J.M.Bibby. Multivaroate Analysis, Academic Press, 1980.

[7] 那須川 哲哉, 河野 浩之, 有村 博紀. テキストマイニング基盤技術, 人工知能学会誌, 16(2), 2001.

[8] 徳永 健伸. 情報検索と言語処理, 東京大学出版会 1999.

[9] E.M.Voorhees and D.K.Harman. Overview of the Seventh Text REtrieval Conference(TREC-7). NIST, http://trec.nist.gov/pubs.html, 1998.

[10] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization, In *Proc. IMCL'97*, 412-420, 1997.