Simple PCA を用いたベクトル空間情報検索モデルの次元削減

正員 黒岩 眞吾* 正員 柘植 覚*
 非会員 獅々堀 正幹* 非会員 任 福継*
 非会員 北 研二**

Dimensionality Reduction of Vector Space Model for Information Retrieval using
 Simple Principal Component Analysis

Shingo Kuroiwa*, Member, Satoru Tsuge*, Member, Masami Shishibori*, Non-member,
 Ren Fuji*, Non-member, Kenji Kita**, Non-member

In this paper, we propose to use the Simple Principal Component Analysis (SPCA) for dimensionality reduction of the vector space information retrieval model. The SPCA algorithm is a data-oriented fast method which does not require the computation of the variance-covariance matrix. In SPCA, principal components are estimated iteratively so we also propose a criteria to determine the convergence. The optimum number of iterations for each principal component can be determined using the criteria. Experimentally, we show that the SPCA-based method offers improvement over the conventional SVD-based method despite its small amount of computation. This advantage of SPCA can be attributed to its iterative procedure which is similar to clustering methods such as k -means clustering. On the other hand, the proposed method which orthogonalizes the basis vectors also achieved much higher accuracy than the conventional random projection method based on k -means clustering.

キーワード：情報検索，ベクトル空間モデル，次元削減，Simple PCA，クラスタリング

Keywords: Information retrieval, Vector Space Model, Dimensionality Reduction, Simple PCA, clustering

1. まえがき

インターネット技術の急速な発展と普及により，World Wide Web (WWW) を代表とする，オンラインテキストデータ量は増加の一途をたどっている。これに伴い，莫大なテキストデータ中から必要かつ十分な情報のみを効果的に検索できる検索エンジンへの期待が高まっている。その期待に答えるべく多くの研究機関や企業によって，様々な手法を用いた情報検索に関する研究・開発が積極的に進められている⁽¹⁾。これらの研究については，米国における TREC (Text Retrieval Conference)⁽²⁾ や，日本における IREX (Information Retrieval and Extraction Exercise)⁽³⁾，NTCIR

(NII-NACSIS Test Collection for IR Systems)⁽⁴⁾ のワークショップにおいてその成果が報じられているが，今なお発展途上の段階である。

代表的な情報検索モデルの一つとして，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)⁽⁵⁾ がある。VSM を用いた情報検索システムは，質問ベクトルと各文書ベクトル間の距離 (類似度) を計算し，距離の近い文書を検索結果として出力するものであり，検索精度を上げるため様々な距離尺度が提案されている⁽⁶⁾。

しかし，同手法では全文書中に含まれる単語が，各文書中にどの程度出現したかに基づき多次元ベクトルを構成するため，そのベクトルの次元数は巨大なものとなる。また，各文書に出現する単語は限られているため，ベクトルは要素に 0 の多いスパースなベクトルとなる。文書全体をこのようなベクトルで表現すると，莫大な記憶容量が必要になると同時に，類似度計算を行う際の計算コストも増加してしまう。また，文書中に含まれる無意味な単語の有無が距離に影響を与え検索精度を低下させてしまうという現象も生じやすい。このため，これらのスパースなベクトルで表

* 徳島大学 工学部
 〒770-8506 徳島市南常三島町 2-1
 Faculty of Engineering, The University of Tokushima
 2-1 Minami-Josanjima, Tokushima 770-8506

** 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
 Center for Advanced Information Technology, The University of Tokushima
 2-1 Minami-Josanjima, Tokushima 770-8506

現された文書全体 (単語・文書行列) の次元数を削減する手法が数多く提案されている^{(6)~(10)}。

単語・文書行列の次元数を削減する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インデキシング (LSI; Latent Semantic Indexing) がある^{(6)~(8)}。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、次元数を削減する方法である。次元削減された空間では潜在的な意味の近さに文書ベクトル間の距離が反映されると考えられ、実際この空間で検索を行うことで多くの場合、検索精度は高くなる^{(7)~(8)}。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が困難な場合もある。また、情報検索に対象を絞った場合、SVD による基底ベクトルに比べ、より高い検索精度を得ることができる基底が存在する可能性もある^{(1)~(10)~(11)}。

このような問題に対し佐々木らは球面 k 平均アルゴリズムとランダム・プロジェクションを情報検索に適用することを試みている (以下、本稿では、球面 k 平均アルゴリズムとランダム・プロジェクションを組み合わせたこの手法をランダム・プロジェクションと表記する)⁽¹¹⁾。これは、球面 k 平均アルゴリズムによりデータをクラスタリングし、各クラスタのセントロイドを概念ベクトルと考え、それらのベクトルを基底ベクトルとすることで、高速化および検索精度向上の 2 つを同時に狙った手法である^{(1)~(11)}。文献⁽¹¹⁾では、SVD の高速アルゴリズムである Lanczos 法に比べ、約 2 分の 1 の時間でプロジェクションのための基底ベクトルを計算できたという実験結果が示されている。しかし、同手法には SVD による手法と同程度の検索精度を達成するために、SVD に比べ遥かに大きい次元数を必要とするという問題がある。また、その最高精度も SVD による手法には達していない。これは、ランダムプロジェクションによる基底が互いに直交していないことに原因があると予想される。

これらの問題に対し、本稿では Simple Principal Component Analysis (SPCA)⁽¹²⁾ を用いたベクトル空間モデルの次元削減を試みる。SPCA は、Patridge らによって提案された手法であり、手書き文字の情報圧縮においてその効果が確認されている⁽¹²⁾。SPCA は、主成分分析を行うことを第一義の目的としており、デフレーション (deflation) と呼ばれる操作により得られる各基底ベクトルは互いに直交する。また、SPCA では、推定された主成分ベクトルと直交する超平面でデータを 2 分割し、各々の空間に含まれる文書ベクトルの平均に基づき再度主成分ベクトルを推定するという演算を、データ分割面 (推定される主成分ベクトルの方向) が変化しなくなるまで繰り返し行うことになる。この繰り返し演算により、一般的には推定される主成分ベクトルの方向は分散最大の方に収束するものと期待できる⁽¹²⁾。しかし、文書ベクトルのようにスパースなベクトルデータに対しては、必ずしも分散最大の方に収束す

るのではなく、データが 2 つのクラスタにきれいに分割された時点で推定する主成分の方向が収束することも考えられる。この場合、SPCA で得られる基底ベクトルはランダム・プロジェクションで期待していたような概念ベクトルに近い基底ベクトルとなる可能性もある。そこで、本稿では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通じ、SPCA を情報検索に適用した場合のアルゴリズムの収束速度および検索精度についての検討を行う。

以下、2 章において、SPCA のアルゴリズムおよびテキスト情報検索での利用方法を説明する。3 章では、SPCA を用いた次元削減法の有効性を検証するため、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うとともに、それらの結果に対する考察を行う。最後に、4 章において、本稿のまとめと今後の課題について述べる。

2. Simple PCA を用いた次元削減法

Simple Principal Component Analysis (SPCA) は、主成分分析の高速化を目的に Patridge らによって提案された手法であり、手書き文字の情報圧縮においてその効果が確認されている⁽¹²⁾。同手法は、単純な繰り返し演算により主成分を近似的に抽出する手法で、SVD 等の行列演算に基づく手法 (Matrix Method) とは異なり、ニューラルネットワーク法⁽¹³⁾等のデータに基づく手法 (Data Method) に分類される。

2.1 Simple PCA 本節では SPCA のアルゴリズムを説明する。

ベクトル集合 $V = \{v_1, v_2, \dots, v_m\}$ の主成分を抽出する問題を考える。まず、通常的主成分分析と同様に、集合の重心を原点とするために各ベクトルから全ベクトルの平均ベクトルを減じ新たなベクトル集合 $X = \{x_1, x_2, \dots, x_m\}$ (以下、入力データと呼ぶ) を求める。

$$x_i = v_i - \frac{1}{m} \sum_{j=1}^m v_j$$

これらの入力データの第一主成分を表す固有ベクトルを α_1 としたとき、以下の出力関数 (output function) y を考える。

$$y_1 = \alpha_1^T x_j \dots \dots \dots (1)$$

ここで、 T は転置を表す。この関数は、 α_1 と直交する原点を通る超平面に対し、入力ベクトルが α_1 と同じ側に存在すれば正の値を、反対側に存在すれば負の値を返す関数である。この関数を用いた閾値関数 (threshold function) を導入する。

$$\Phi_1(y_1, x_j) = \begin{cases} x_j & \text{if } y_1 \geq 0 \\ 0 & \text{otherwise} \end{cases} \dots \dots \dots (2)$$

以上の関数を用い、初期値として適当に与えた任意のベクトル a_1 を繰り返し演算により、 α_1 と同じ方向に近づける

ことを考える。この繰り返し演算は以下の式で表現される。

$$a_1^{k+1} = \frac{\sum_j \Phi_1(y_1, x_j)}{\left\| \sum_j \Phi_1(y_1, x_j) \right\|}; \quad y_1 = (a_1^k)^T x_j \quad \dots\dots\dots (3)$$

ここで、 k は繰り返し回数であり、出力関数 y_1 は 1 つ前の繰り返しで推定した a_1^k により計算される。この演算は a_1 を分散のより大きな方向に修正する性質があるので、同演算を繰り返すことにより、 a_1^k は a_1 と同じ方向に収束すると期待できる。

以上の演算により第一主成分が計算できるが、引き続き第二主成分を求めるために入力データより第一主成分を取り除く必要がある。そこで、下記の式により各データ x_i から第一主成分を除去する（この操作をデフレーションと呼ぶ）。

$$x'_i = x_i - ((a_1^M)^T x_i) a_1^M \quad \dots\dots\dots (4)$$

ここで、 M は収束時の繰り返し回数である。この x'_i に対し式 (3) により a_2 を求めることで第二主成分が得られる（ただし、 a_1 は a_2 に、 x_j は x'_j に置き換える）。以下、同様の操作を繰り返すことで寄与率の大きい順に主成分を得ることができる。

なお文献⁽¹²⁾においては、閾値関数として式 (2) に加え以下の 2 つの関数を閾値関数として提案している。

$$\Phi_2(y, x_j) = \begin{cases} +x_j & \text{if } y \geq 0 \\ -x_j & \text{otherwise} \end{cases} \quad \dots\dots\dots (5)$$

$$\Phi_3(y, x_j) = y x_j \quad \dots\dots\dots (6)$$

3 章の検索実験では、式 (2) に加え、この 2 つの閾値関数についても性能比較を行う。

なお、SPCA の計算量は $O(dmn)$ (ただし、 d は求める主成分数、 m はデータ数、 n はベクトルの次元数、また収束回数が定数となる) であり、佐々木らの球面 k 平均アルゴリズムと等しい。これは、SVD 等の行列演算に基づく方法に比した場合格段に少ない演算量で主成分を近似的に求めることが可能であることを意味する。また、SVD の高速アルゴリズムである Lznczos 法と比した場合でも半分程度の時間で基底ベクトルを計算できることが実験的に示されている。

2.2 収束条件 SPCA は式 (3) の繰り返し演算により各々の基底ベクトルを計算する。本稿では、以下の 2 つの方法で収束を判定し、その性能を後述の実験で比較する。

- (1) 繰り返し数をあらかじめ指定し、その繰り返し数に達したら終了する。
- (2) 更新前後の基底ベクトル a_1^k, a_1^{k+1} が下記の収束判定条件を満足したら終了する。

$$1.0 - \frac{(a_1^{k+1})^T a_1^k}{\|a_1^{k+1}\| \|a_1^k\|} \leq Th \quad \dots\dots\dots (7)$$

ここで、 Th はあらかじめ指定するしきい値である。

この式 (7) の左辺、 a_1^{k+1} と a_1^k のなす角に相当するものであり、更新によるベクトルの方向に変化がなくなれば 0 となる。以下、式 (7) の左辺の値を推定係数 (estimation factor) と呼ぶ。

なお、本収束条件は個々の基底ベクトルを求める際の収束条件であり、とりあげる主成分 (基底ベクトル) の数については、従来研究⁽¹⁾⁽⁶⁾⁻⁽¹⁰⁾と同様に実験において検討する。

2.3 検索方法 本提案手法は、「潜在的意味インデキシング (LSI) による検索」における次元削減部に Simple PCA を適用したものである。そのため、検索法は LSI による検索と同一である。

具体的には、2.1 で示したベクトル集合 $V = \{v_1, v_2, \dots, v_m\}$ の各要素 v_i が検索対象となる個々の文書ベクトルとなる。各文書ベクトルを V から求めた d 個の要素からなる基底ベクトル集合 $A = \{a_1, a_2, \dots, a_d\}$ により射影し、次元削減後の検索対象ベクトル集合 $V' = \{v'_1, v'_2, \dots, v'_m\}$ を得る。同様に、検索質問ベクトル q も A により射影し、 q' とする。検索時には、 q' と各検索対象ベクトル v'_i との類似度を計算し、類似度が高い順に対応する文書を検索結果として出力する。なお、実験で用いる具体的な文書ベクトル構成法および類似度に関しては、次章で述べる。

3. 検索実験

情報検索における SPCA を用いた次元削減手法の有効性を検証するため、テキスト情報検索分野における標準的な評価用テストコレクションである MEDLINE⁽⁷⁾ を用い以下の 3 つの手法による情報検索実験を行った。

- (1) ベクトル空間モデル
- (2) SVD を用いた LSI
- (3) SPCA を用いた LSI (提案手法)

検討した項目は、文書ベクトル集合に対する SPCA の収束性および収束の判定条件、次元削減時の最適次元数、平均適合率、再現率・適合率曲線からの提案手法の有効性である。

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 個の検索質問文と各検索質問に対する正解 (適合) 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。

前処理として、このテストコレクションに含まれる 1,033 文書全体から、“a” や “about” などの一般的な 439 語を、文書の内容と関連の無い単語 (不要語) として削除した。また、全文書中に 1 回しか出現しなかった単語も削除した。この処理により削除されなかった単語に対し、Porter 法による接辞処理⁽¹⁴⁾を施し、語幹への変換を行った。以上の処理の結果、文書全体に存在した異なり単語数は 5,526 から 4,329 単語に削減された。実験ではこれら 4,329 単語を

検索に用いる索引語とした。

3.1.1 VSM 前処理によって得られた索引語を用い、4,329 次元の文書ベクトルを構成しベクトル空間モデル (VSM) に基づく情報検索システムを構築した。文書 j に対応する文書ベクトル v_j の要素 v_{ij} は、文書 j の索引語 i のローカル重み L_{ij} と索引語 i のグローバル重み G_i の積で表現される。すなわち、

$$v_{ij} = L_{ij} \cdot G_i \dots\dots\dots (8)$$

となる。

各々の重みとしては、対数エントロピーに基づく重み⁽¹⁵⁾を用いた。この重みは、

$$L_{ij} = \begin{cases} 1 + \log f_{ij} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases} \dots\dots\dots (9)$$

$$G_i = 1 + \sum_{j=1}^m \frac{f_{ij} \log \frac{f_{ij}}{F_i}}{\log m} \dots\dots\dots (10)$$

で与えられる。 m はテストコレクション中の文書数を、 f_{ij} は文書 j における索引語 i の出現頻度を、 F_i は全文章中における索引語 i の出現頻度を各々表す。

類似度は、次式に示される、質問ベクトル q と各文書ベクトル v_j 間のコサイン尺度を用いた。

$$\cos(q, v_j) = \frac{q^T v_j}{\|q\| \|v_j\|} \dots\dots\dots (11)$$

3.1.2 SPCA まず VSM の文書ベクトル集合 $\{v_j\}$ を用い 2 章に示したアルゴリズムにより寄与率の高い順に N 個の主成分を求めた。次にこれらの主成分を基底ベクトルとして次元数の削減を行い、次元削減されたベクトル間のコサイン尺度により検索を行った。閾値関数としては式 (2)、式 (5)、式 (6) の 3 種類を用い収束性および次元数に関し平均適合率を尺度に比較を行った。なお、初期ベクトル a^0 は単位ベクトルとした。

3.1.3 SVD 比較のために、SVD により SPCA と同次元に次元数を削減した場合の実験も行った。SVD にはランチョス法 (SVDPACK)⁽¹⁶⁾ を用いた。

3.1.4 評価方法 検索システムの精度の評価には、“trec_eval” プログラム⁽²⁾ を用いて平均適合率 (non-interpolated average precision) および再現率・適合率曲線 (recall-precision curve) を求めた。ただし、平均適合率を求めるにあたっては、実用性を考慮し上位 50 位までの検索結果を用いた。適合率および再現率の定義は以下のとおりである。

$$\text{適合率} = \frac{\text{検索できた関連文書数}}{\text{検索文書数}}$$

$$\text{再現率} = \frac{\text{検索できた関連文書数}}{\text{関連のある文書数}}$$

再現率・適合率曲線を求めるにあたっては、各質問ごとの適合率を再現率 0.0, 0.1, ..., 0.9, 1.0 の 11 点で各々平均を

図 1 繰り返し回数と平均適合率

Fig. 1. Average precision as a function of number of the iterations.

とることにより求めた⁽¹⁾。一方、平均適合率は再現率レベル 0.0, 0.1, ..., 0.9, 1.0 の適合率の平均により求める。ただし、検索結果を上位 50 位までに絞った場合、再現率があるレベルで頭打ちになるため、それ以上のレベルの再現率に対応する適合率は 0.0 とし計算する⁽²⁾。

3.2 繰り返し回数と収束性 SPCA の繰り返し演算による収束性を調べるために、各閾値関数を使った場合の繰り返し回数と平均適合率の関係を調べた。次元削減後の次元数を 50 とした場合 (この次元数において SVD では、最高の平均適合率が得られた) の結果を図 1 に示す。図中の Eq.(2)、Eq.(5)、Eq.(6) は、式 (2)、式 (5)、式 (6) の閾値関数を用いた場合の結果を各々示している。比較のために SVD の結果も示した。どの閾値関数を用いた場合も同様に、SPCA は繰り返し回数 5 で一旦 SVD の平均適合率を上まわっている。その後、繰り返し回数 20 回 ~ 30 回で平均適合率は収束し、すべての閾値関数で SVD より高い性能を示した。十分収束するまで繰り返すという条件を満足できれば、式 (5) の閾値関数が高性能を示すと考えられる。なお、次元数を 20 ~ 200 程度に変化させた場合でも、収束性に関しては同様な結果が得られている。

3.3 しきい値による収束判定 前節の実験で、収束が最も遅くかつ最高性能であった式 (5) の SPCA について、2.2 節で示した式 (7) を収束の判定に用いる実験を行った。結果を図 2 に示す。しきい値 Th が小さいほど収束条件は厳しくなる。平均適合率に多少の変動はあるものの、概ねしきい値を厳しくすることで性能の向上が見られる。また、しきい値 0.0001 と 0.00005 の場合の平均適合率は同一で最大値に到達していることから、しきい値 0.0001 で収束していると考えられる。

しきい値 0.0001 とした場合に各次元毎に収束に要した繰り返し回数を表 1 に示した。1 ~ 40 次元までの平均繰り返し回数は 16.1 である。次元によって収束までの回数にばらつきがあり、しきい値を用いることにより、収束までの

図 2 推定係数のしきい値と平均適合率

Fig. 2. Average precision as a function of the threshold of estimation factor.

表 1 各次元の収束までに要した繰り返し回数

Table 1. A number of iterations required until convergence of each principal component.

Component	Iterations	Component	Iterations
1	18	21	16
2	9	22	21
3	13	23	16
4	27	24	13
5	18	25	22
6	10	26	23
7	17	27	17
8	13	28	16
9	15	29	19
10	23	30	9
11	17	31	17
12	18	32	16
13	14	33	21
14	9	34	22
15	20	35	19
16	14	36	17
17	11	37	13
18	13	38	12
19	11	39	12
20	16	40	16

延べ繰り返し回数をかなり削減できたことがわかる。

3・4 最適次元数 次元削減後の次元数を変化させ平均適合率を調べた。結果を図 3 に示す。比較のために SVD を用いた場合の結果もプロットした。SPCA を用いることで、次元数 80 以下の場合、すべての閾値関数で SVD より高い平均適合率が得られた。SPCA を用いた場合の最高の平均適合率は式 (5) の閾値関数を用いた次元数 40 のときで 0.694 である。これに対し SVD では次元数 50 のときの 0.663 が最高の平均適合率であり、SPCA を用いることで誤り率を 1 割削減できたことになる。なお、VSM を用いた場合の平均適合率は 0.494 で図 2 の範囲外となっている。また、直交化を行わないランダム・プロジェクションは、次

図 3 次元数と平均適合率

Fig. 3. Average precision as a function of number of dimensions.

図 4 再現率・適合率曲線

Fig. 4. Recall-precision curve.

元数が 900 の時、最高の平均適合率 0.604 が報告されている⁽¹¹⁾。また、同手法では次元数が下がるにつれ平均適合率は低下し、次元数 100 では 0.398 となる。

3・5 再現率・適合率曲線 SPCA, SVD, VSM の性能の違いをより詳細に調査するために、各手法において平均適合率が最高となった条件下で再現率・適合率曲線を作成した。結果を図 4 に示す。図中の () 内の数字は各手法での次元数である。VSM に比べ、次元数を削減した SPCA, SVD では顕著な性能の向上が見られる。SPCA と SVD を比較した場合、SPCA が平均的に高い適合率を示している。また、その差は再現率 0.5 で最大で 0.065 となる (このときの SPCA の適合率は 0.795)。一方で、再現率 0.8 以上、0.2 未満では両者の差はあまりない。

3・6 考 察 SPCA を用いることで、SVD に比べ平均適合率で約 1 割の誤りを削減することができた。また、そのときの次元数は 40 で、SVD の最適次元数と同等

以下になっている。さらに、図4の再現率・適合率曲線に着目すると、再現率0.2～0.8でSPCAがSVDに比べ高い性能が得られていることが分かる。例えば、再現率0.5において、SVDでは検索文中の73%が関連文書であるのに対し、SPCAでは80%が関連文書となる。

一方で、SPCAがSVDに比べ平均適合率が高かったことは、SPCAにより得られた基底ベクトルがSVDと異なっていたことを示している。SVDによる次元削減は主成分分析による次元削減と等価であり⁽⁴⁾、SPCAで平均適合率が高くなったということは、SPCAで主成分を正しく推定できていなかったことを意味する。図3に着目すると、SPCAが特に次元数（とりあげる主成分数）が小さい場合にSVDと異なる射影を行っていることがわかる。また、次元数が小さい場合はSPCAで用いた各閾値関数によっても平均適合率に違いが見られる。一方で、次元数が大きくなるとSPCAとSVD、および各閾値関数による平均適合率の違いはほとんどなくなっている。これは、情報検索の対象となる文書ベクトル空間がスパースであることが原因の一つであると考えられる。SPCAは、データの重心を通る超平面でデータを2分し、その結果に基づき主成分を推定するアルゴリズムである。SPCAにおける各閾値関数は、分割されたデータによる主成分の推定法に対応している。式(2)では、分割面的一方に属するデータの平均ベクトルを主成分としている。式(5)では、分割された各領域のデータの平均を結ぶベクトルを主成分としており、式(6)では単なる平均ではなく、全データの重心からの距離に応じた重み付き平均を利用している。特に、式(5)による分割は、分割面が全データの重心を通ることを条件とした、 k 平均アルゴリズム（ただし、 $k=2$ ）になる。これらのアルゴリズムは、推定した主成分と直交する超平面によりデータがきれいに分割され、各領域のデータの平均が変化しなくなった時点で収束に至る。そのため、スパースなデータに対しては必ずしも分散最大の方ではなく、データを2つのクラスに準最適に分割するためのベクトルに収束することとなる。実際、本実験でSPCAにより平均適合率が改善したことは、(a) SPCAで得られた基底ベクトルが分散最大の方以外に収束していること、(b) その基底ベクトルが情報検索においては有効であるということ、を示している。このうち、(b)は、情報検索のための次元削減ではクラスタリングにより基底ベクトルを選択することが有効であることを示唆している。なお、次元数40以上では次元数が大きくなるに従いSVDやSPCAの各閾値関数による平均適合率は類似した値となっている。これは、デフレーションによりデータのスパースネスが失われていくことにより生じていると考えられる。以上の考察は現段階では一仮説にすぎないが、今後、いくつかの教師なしクラスタリングによる次元削減法の検討を進め、さらなる検証を行う予定である。

また、ランダム・プロジェクションとSPCAを比較した場合、SPCAが40次元で最大平均適合率0.694を得てい

るのに対し、ランダム・プロジェクションでは、次元数900で最大平均適合率0.604となっている⁽¹¹⁾。このことより、SPCAが最大平均適合率を遥かに低い次元数で達成していることがわかる。また、その最大平均適合率もランダム・プロジェクションを大きく上まわっている。ランダムプロジェクションもクラスタリングに基づき基底ベクトルを選択しているが各基底ベクトルは互いに直交していない。これに対し、SPCAではデフレーションにより基底ベクトルを直交化しており、このことが、情報検索において有効であることを示唆している。

4. 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)の次元削減手法として、Simple Principal Component Analysis (SPCA)⁽¹²⁾を用いることを提案した。SPCAは簡単な繰り返し演算により、主成分を寄与率の大きい順に求めていく近似的な主成分分析手法であり、SVD等の行列演算を基本とする手法に比べ計算量が格段に少ないという特長がある。

MEDLINEを用いた検索実験を行った結果、削減後の次元数によらず、SVDに比べて同等以上の平均適合率を得た。また、この実験における最高の平均適合率はSPCAで0.694(次元数40)、SVDで0.663(次元数50)であり、SPCAを用いることで誤り率は約1割削減されている。これは、文書ベクトルデータがスパースなことやSPCAが k 平均クラスタリングと類似したアルゴリズムを採用していることが要因の一つではないかと考えられる。一方で、球面 k 平均クラスタリングによるランダム・プロジェクションと提案手法を比べた場合も、次元数、最高性能共に提案手法の有効性が認められた。これは、クラスタリングに加え、基底ベクトルの直交化を行うことが情報検索においても重要であることを意味している。

残された課題として、より大規模な文書データベースおよび日本語文書データベースでの提案手法の有効性の検証がある。また、平均適合率が最大となるような削減後の次元数においてSPCAがSVDに比べ高い平均適合率を示す基底を得ることができた原因について、詳細な検証を行う予定である。このとき、寄与率と平均適合率の関係についても検討し、情報検索に最適となる削減後の次元数を基底ベクトル作成時に推定する方法についても考えていく予定である。さらに、クラスタリングに基づく次元削減法として、現在、教師なしクラスタリング手法の1つであるベクトル量子化を用いた次元削減法の実験を進めている。

謝辞

本研究の一部は文部科学省科学研究費基盤研究(B)14380166, 17300036, 17300065, 国際コミュニケーション基金, 電気通信普及財団および放送文化基金の補助を受け行った。

(平成16年8月24日受付, 平成17年6月10日再受付)

文 献

- (1) K. Kita, M. Shishibori, and K. Tsuda: "Information Retrieval Algorithms", KYORITSU SHUPPAN (2002) (in Japanese)
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)
- (2) TREC homepage: <http://trec.nist.gov/>
- (3) IREX homepage: <http://cs.nyu.edu/cs/projects/proteus/irex>
- (4) NTCIR homepage: <http://www.rd.nacsis.ac.jp/~ntcadm/>
- (5) G. Salton and J. McGill: Introduction to Modern Information Retrieval. McGraw-Hill Book Company (1983)
- (6) M. W. Berry, Z. Drmac, and E. R. Jessup: "Matrices, vector spaces, and information retrieval", *SIAM Review*, Vol.41, No.6, pp.335-362 (1999-6)
- (7) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. 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)
- (8) T. A. Letsche and M. W. Berry: "Large-scale information retrieval with latent semantic indexing", *Information Sciences - Applications*, Vol.100, Issue 1-4 pp.105-137 (1997-8)
- (9) X. Tai, M. Fukumi, and K. Kenji: "A new approach for improvement in vector space information retrieval model based on supervised learning", *IEEEJ Trans.EIS*, Vol.121, No.10, pp.1647-1653 (2001-10) (in Japanese)
Tai Xiaoying・福見 稔・北 研二:「教師あり学習によるベクトル空間情報検索モデルの精度改善」, 電学論 C, 121, 10, pp.1647-1653 (2001-10)
- (10) S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kenji: "Information retrieval using non-negative matrix factorization", *IEEEJ Trans.EIS*, Vol.124, No.7, pp.1500-1506 (2004-7) (in Japanese)
柘植 覚・獅々堀正幹・黒岩眞吾・北 研二:「Non-negative Matrix Factorization を用いたベクトル空間情報検索モデルの次元削減手法」, 電学論 C, 124, 7, pp.1500-1506 (2001-7)
- (11) M. Sasaki and K. Kita: "Deimensionality Reduction of Vector Space Information Retrieval Model based on Random Projection", *Journal of Natural Language Processing*, Vol.8, No.1, pp.5-19 (2001-1) (in Japanese)
佐々木稔・北 研二:「ランダム・プロジェクションによるベクトル空間情報検索モデルの次元削減」, 自然言語処理, 8, 1, pp.5-19 (2001-1)
- (12) M. Partridge and R. Calvo: "Fast dimensionality reduction and simple PCA", *IDA*, Vol.2, No.3, pp.292-298 (1997-2)
- (13) K. I. Diamantaras and S. Y. Kung: Principal component neural networks: theory and applications. Wiley, New York (1996)
- (14) W. B. Frakes and R. Baeza-Yates: Information Retrieval: Data Structures and Algorithms. Prentice Hall (1992)
- (15) E. Chisholm and T. Kolda: "New term weighting formulas for the vector space method in information retrieval", *Technical Memorandum ORNL-13756* (1999)
- (16) M. W. Berry, T. Do, G. O'Brien, and V. D. Pietra: SVD-PACKC (Version 1.0) User's Guide. Department of Computer Science, University of Tennessee (1993)

黒 岩 眞 吾 (正員) 1986 年, 電通大・電気通信・通信卒。1988

年, 同大学院修士課程了。博士 (工学)。同年国際電信電話 (株) 入社。1988 年~2001 年, 同社研究所において電話音声認識システムの研究・開発に従事。2001 年, 徳島大学工学部助教授。音声認識, 話者照合, 情報検索の研究に従事。電子情報通信学会平成 8 年度学術奨励賞, 日本音響学会第 3 回および第 5 回技術開発賞各受賞。日本音響学会, 電子情報通信学会, 人工知能学会各会員。

柘 植 覚 (正員) 1996 年, 徳島大学工学部知能情報工学科卒業。1998 年, 同大学院工学研究科博士前期課程知能情報工学専攻修了。2001 年, 同大学院工学研究科博士後期課程システム工学専攻修了。2000 年, 徳島大学工学部助手, 博士 (工学), 音声認識, 情報検索等の研究に従事。

獅々堀 正 幹 (非会員) 1991 年, 徳島大学工学部情報工学科卒業。1993 年同大学院博士前期課程修了。1995 年同大学院博士後期課程退学。同年同大学工学部知能情報工学科助手。1997 年同大学工学部知能情報工学科講師。2001 年同大学工学部知能情報工学科助教授。現在同大学工学部知能情報工学科助教授。博士 (工学)。マルチメディア情報検索, 自然言語処理の研究に従事。著書「情報検索アルゴリズム」共立出版, 情報処理学会第 45 回全国大会奨励賞受賞。電子情報通信学会, 言語処理学会各会員。

任 福 継 (非会員) 1982 年, 北京郵電大学電信工程卒。1985 年同大学院修士課程了。1991 年北海道大学大学院工学研究科博士後期課程了。博士 (工学)。CSK 研究員, 広島市立大学情報科学部助教授を経て, 2001 年徳島大学工学部教授。1996 年~1997 年アメリカニューメキシコ大学訪問教授。自然言語処理, 感性情報処理, 学習システム, 人工知能, 多言語多機能多メディア知的システムに関する研究に従事。IEEE, ACL, 情報処理学会, 言語処理学会, 教育システム情報学会各会員。

北 研 二 (非会員) 1981 年, 早稲田大学理工学部数学科卒業。1983 年, 沖電気工業 (株) 入社。1987 年, ATR 自動翻訳電話研究所出向。1992 年, 徳島大学工学部講師。1993 年, 同助教授。2000 年, 同教授。2002 年, 同大学高度情報化基盤センター教授。工学博士。自然言語処理, 情報検索等の研究に従事。平成 6 年日本音響学会技術開発賞受賞。著書『確率的言語モデル』(東京大学出版会), 『情報検索アルゴリズム』(共著, 共立出版) など。