検索質問にあわせた文書ベクトルの次元削減手法

木寺　悠介[*]　陳　幸生[**]　塩原　慶一[***]

[*]神奈川工科大学　工学研究科　情報工学専攻
[**]神奈川工科大学　情報学部　情報工学科
[***]神奈川工科大学　情報学部　情報工学
住所：〒243-0292 神奈川県厚木市下荻野 1030
E-mail: [*]s055808@cce.kanagawa-it.ac.jp, [**]chen@ic.kanagawa-it.ac.jp
[***]s055078@cce.kanagawa-it.ac.jp

あらまし　ベクトル空間モデルによる文章検索では，単語―文書行列により，文書と検索質問両方を表す高次元のベクトル作成し，文書ベクトルと問い合わせベクトルの相関度を計算し，相関度を降順で並べ替えて，検索結果を求める．高次元の文書ベクトルの次元削減は，効率的に文書検索を行うために有効であることが分かっている．本研究では，従来の次元削減手法と異なり，検索質問の内容から，文書ベクトルを作成するとき必要な単語要素を抽出し文書ベクトルを作成する手法を提案する．また，文書ベクトルの次元数を 2 次元までに減らす仕組みを考案し，その仕組みを用い文章検索実験を行い提案手法の有効性を検証する．

A Method for Decreasing the Number of Dimensions of Document Vectors through the Contents of Queries

Yusuke Kidera [*]　Xing Chen [**]　keiichi Shiohara [***]

[*] Dept. of Info. & Comp. Science, Grad. Cause, Kanagawa Institute of Technology
[**] Dept. of Info. & Comp. Sciences, Kanagawa Institute of Technology
[***]Dept. of Info. & Comp. Sciences, Under-Grad. Cause, Kanagawa Institute of Technology
Address: 1030 Simo-Ogino, Atsugi-shi Kanagawa, 243-0292 Japan
E-mail: [*]s055808@cce.kanagawa-it.ac.jp, [**]chen@ic.kanagawa-it.ac.jp
[***]s055078@cce.kanagawa-it.ac.jp

Abstract In the vector space model, the document retrieval is performed by creating a term-document matrix at first. After that, based on the created matrix, documents and queries are represented as high-dimensional vectors. During the query processing, correlations between document vectors and query vectors are calculated and the calculated results are sorted in the down order as the retrieval result. It is known that decreasing the number of the dimensions of the document vectors makes query processing effectively. In this research, we propose a new method based on which the number of the dimensions of the document vectors are greatly decreased into 2 dimensions. This method is different from the other methods proposed previously. In out method, based on the giving contents of queries, terms that are used to great document vectors are extracted for queries. The document vectors are dynamically created during the query processing. Furthermore, we demonstrated that based on our method, only 2 dimensional spaces are needed for representing documents and queries during the query processing. Experimental results are shown for clarifying the effectiveness of the proposed method.

1. はじめに

ベクトル空間モデル[1]では，検索対象となる文章を単語の列に置き換え，元の文書を語彙数と同様な次元数の単語列ベクトルとして扱う．文書ベクトルの各要素は，単語が文書中の出現頻度，または，重みつきの出現頻度である．単語が文書の中に出現しない場合，対応した要素の値は 0（或いは－1 [2]）である．検索処理では，検索質問を文書と同様に語彙数と同様な次元数の単語列ベクトルに置き換えて，検索質問ベクトルと文書ベクトルの相関度を計算し，相関度の高い文書を検索結果として出力する．

ベクトル空間モデルを用いた文書検索は，比較的な長文からなる自然文のような検索質問を与えて，文書と検索質問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分かっている．この方式は，各文章ベクトルが数万から数十万の疎な高次元ベクトルになるため，検索効率の向上を求めるとともに，必要な計算機の記憶容量を考慮し，高次元の文書ベクトルを低次元への縮小が必要である．

Latent Semantic Indexing (LSI) [3]　は，数多く論じられている文書ベクトルの次元縮小手法である[4]．LSI 手法では，文書集合から作成した単語 - 文書行列に特異値分解　（SVD）を行い，高次元の文書ベクトルから低次元ベクトルへのプロジェクションにより文書ベクトルの次元縮小を実現する．LSI 手法では，文書集合に依存するため，文書集合の変化があれば，SVD の

再計算が必要となる．または，SVD 計算は，多くの計算量を必要とするため,近似的な SVD の再計算を行う方法が提案されている[5]が，事前に各分野の特徴を代表できる小規模の文書集合を用意し SVD の計算を行うことが一般的である．

同一の分野に属する文書は似たような単語分布をもつ性質がある．例えば，計算機科学分野に関する論文と化学分野に関する論文では，計算機科学分野に属する論文によく現れる単語は化学分野に関する論文に出現頻度が低い．SVD 計算では，単語の分野ごとの分布特徴を抽出でき，分野の数と同様な次元数のベクトル空間を作成できる．事前に用意した文書集合の単語の分野ごとの分布特徴は，次元圧縮に大きな影響を与える.明確な単語分布特徴を有する文書集合に SVD 計算を行うと，文書ベクトルの次元数を文書集合の分野数と同様な次元数まで圧縮しても，検索精度が落ちないことが分かった．

明確な単語分布特徴を有する文書集合の用意ができれば，重い SVD 計算をしなくても，低次元の文書ベクトルを単語 - 文書行列から直接的に作成する方法（Feature Extracting Model, FEM）が提案されている[6]. この方法では，単語の分布特徴に着目し，ベクトル空間の作成に関連性がある特徴単語を抽出し，文書ベクトルを作成する．文書ベクトルの各要素は，抽出した特徴単語の文書中の出現頻度，または，重みつきの出現頻度である．

本研究では，FEM 手法に基づき，比較的な長文からなる自然文のような検索質問からベクトル空間の作成に関連性がある特徴単語の抽出により，特徴単語－文書行列を作成し，文書ベクトルを作成する．更に，文書ベクトルの次元数を 2 次元までに縮小する手法を考案する．本手法は，文書のベクトル空間が検索質問の変化により動的に作成する特徴と文書ベクトルの次元数が 2 次元までに縮小される特徴がある．提案手法の有効性を検証するため，文書テストデータセット NTCIR 1 [7]を用い検証実験を行う．

2. FEM 手法

本章では、FEM 手法について簡潔に説明する，これからの説明は，明確な単語分布特徴を有する文書集合があることを前提とする．この前提を要求しない文書ベクトルの作成手法について[8]では論じていた．

以下，事前に用意した文書集合をサンプル文書群と呼ぶ．サンプル文書群は，いくつのグループに分けられる.同一グループに属する文書は似たような単語(以下特徴キーワードと呼ぶ）分布を有する．FEM では，それぞれのグループに属する文章から似たような分布をしている単語群（以下特徴キーワード群と呼ぶ）を抽出する．抽出した特徴キーワードは，あるグループに属した文書群に出現した頻度が高いが，他のグループに属した文書群に出現した頻度が低い．

サンプル文章群 $d_1, d_2, d_3 \cdots, d_m$ に対し，文書を q グループに分ける場合，各グループを $C_1, C_2, \ldots, C_q$,で表し，特徴キーワードを t と定義すれば， 表 1のようなテーブルを作成することができる．表 1の第 1 列の各項目は文書グループの特徴キーワード群を表す．例えば，グループ C_1 の特徴キーワード群は $\{t_1, \ldots, t_a\}$ である．

表1 特徴キーワード群と文書グループ

	C_1	C_2	$\cdots$	C_q
$t_1, \cdots, t_a$	$d_1, \cdots, d_a$			
$t_{a+1}, \cdots, t_b$		$d_{i+1} \cdot, d_{i+j}$		
$t_{n-f}, \cdots, t_n$				$d_{m-s}, \cdots, d_m$

FEM では，文書ベクトルが特徴キーワードの各文書中の出現頻度，または，重みつきの出現頻度により作成される.文書ベクトル $\mathbf{d}_j$ の要素を $e_{i,j}$ と表示したら，文書ベクトルは，表 2 のように作成される．表の中の各列は文書ベクトルを表している．各行は，文書グループを表している．

表2 文書ベクトル

	$\mathbf{d}_1$	$\mathbf{d}_2$	$\cdots$	$\mathbf{d}_m$
C_1	$e_{1,1}$	$e_{1,2}$		$e_{1,m}$
C_2	$e_{2,1}$	$e_{2,2}$		$e_{2,m}$
$\cdots$	$\cdots$	$\cdots$		
C_q	$e_{q,1}$	$e_{q,2}$		$e_{q,m}$

文書ベクトル $\mathbf{d}_j$ の要素 $e_{i,j}$ の値は，グループ C_i の特徴キーワード群中の単語が文書 d_j の中の出現頻度，または，重みつきの出現頻度である．例えば，文書グループ C_i の特徴キーワード群は$\{t_1, t_2, t_3\}$である場合，v_{t1}, v_{t2}, v_{t3}はそれぞれの単語 t_1, t_2, t_3 が文書 d_j 中の出現頻度とすると，$e_{i,j}$ の値は次の通りである．

$$e_{i,j} = v_{t1} + v_{t2} + v_{t3} .$$

或いは，

$$e_{i,j} = w_{t1} \times v_{t1} + w_{t2} \times v_{t2} + w_{t3} \times v_{t3} .$$

式の中の w_{t1}, w_{t2}, w_{t3} は，出現頻度 v_{t1}, v_{t2}, v_{t3} に対し付けた重きである．

文書群が q グループに分かれる場合，文書ベクトルは，式(1)のように q 次元のベクトルとなる．

$$\mathbf{d}_j = [e_{1,j}, e_{2,j}, \cdots, e_{q,j}]^T \qquad (1)$$

検索質問も文書と同様に q 次元のベクトル空間上に射影，ベクトルとして表される．この q 次元のベクト

ル空間は，検索ベクトル空間と呼ばれる．問い合わせ処理では，文書ベクトルと検索質問ベクトルの相関度を求め，相関度の大きい文書を検索結果として出力する．

3. 検索質問から検索ベクトル空間の作成

FEM では，各文書グループに属する特徴キーワードを抽出し，文書ベクトルを作成する．特徴キーワードの抽出処理では，どちらの文書にも出現頻度が高い単語群（以下，共通単語群と呼ぶ）の事前用意が不要である．本研究では，事前に作成した共通単語群がある場合，文書ベクトルの作成について次の手順で考案する．

まず，サンプル文書群の特徴キーワードの分布を次のように理想化する：

グループ C_i の特徴キーワードは，同一グループに属する文書の中のみに出現し，他のグループに属する文書の中には出現しない．

グループ C_i の特徴キーワード群を K_i と定義し，特徴キーワードを t と定義すれば，理想化した特徴キーワードの分布は次の式で表せる．

$t \in C_i$ なら，$t \notin C_j$．そのうち，$i \neq j$，i,j=1,2...,q.

特徴キーワード以外の単語，つまり，どちらの文書にも出現頻度が高い単語群（以下，共通単語群と呼ぶ）を T として定義する．理想化した特徴キーワードの分布により，特徴キーワード t は共通単語群 T の中に出現しない．その理由は，次の通りである．

もし，単語 t が，共通単語群 T の中に出現したら，単語 t は，少なくても 2 つのグループ C_i, C_j（$i \neq j$,）に属した 2 つの異なる文書の中に出現する．文書 d_k と文書 d_l は同様な文書ではなく，また，文書 d_k と文書 d_l がそれぞれ異なるグループ C_i, C_j に属する場合，ある単語 t（$t \in T$）に対し，$t \in d_k$ と $t \in d_l$ が成立する．

$$t \in C_i \text{ と } t \in C_j (i \neq j) \text{ なら，} t \in T.$$

言い換えると、

$$t \in C_i \text{ と } t \notin C_j (i \neq j) \text{ なら，} t \notin T.$$

共通単語群 T を用い，単語 - 文書行列の中から T に属する単語を取り除ければ，残った単語はすべて特徴キーワードになる．つまり，単語 - 文書行列を特徴キーワード - 文書行列に変換することができる．

同一グループに属する特徴キーワードは，他のグループに属する文書の中には出現しない特性により，文書ベクトルの作成を，グループごとに行うことが可能である． 更に，作成した文書ベクトルは，直交特性を持つことが分かる．例えば，グループ C_1 の特徴キーワード群 $\{t_1, \ldots, t_a\}$ は，グループ C_1 に属する文書，d_j の中のみに出現し，他のグループ特徴キーワードは d_j の中に出現しないので，文書 d_j を表すベクトルは，式(2)のように生成した．

$$\mathbf{d}_j = [e_{1,j}, 0, \cdots, 0]^T \qquad (2)$$

式(2)は，文書ベクトル $\mathbf{d}_j$ の値が同じグループ C_1 に属する特徴キーワードの出現頻度 $e_{1,j}$ に決められることを示している．つまり，文書ベクトル $\mathbf{d}_j$ の作成は，文書 d_j と同じグループに属する特徴キーワード群だけを利用すれば可能である．

理想サンプル文書群の特徴キーワードの分布により，文書 d_j から共通単語を取り除いたら，残った単語はすべて特徴キーワードである．もし，文書 d_j がすべての特徴キーワードを有する場合，共通単語の取り除く処理を行うだけで，すべての特徴キーワードが抽出されることができる．

次は，理想化しない普通のサンプル文書群を対象として，文書ベクトルの生成方法について述べる．単語が文書 d_j の中に理想的に分布していない場合，文書 d_j から共通単語を取り除いたら，残った単語について次のような特性がある．

1．残った単語は特徴キーワードであるが，同一グループに属するすべての特徴キーワードではない．つまり，残った単語は同一グループに属する特徴キーワード群の部分集合である．この状況について，我々は，作成した文書ベクトルには“要素不足”と定義する．この場合，作成した文書ベクトルは，理想的な特徴キーワード分布を有するとき作成した文書ベクトルと比べると，ベクトルの長さが短い．

2．残った単語の中に特徴キーワードではない単語がある．この状況について，我々は，作成した文書ベクトルに“ノイズ”があると定義する．この場合，作成した文書ベクトルは，理想的な特徴キーワード分布を有する場合と比べ，ベクトルの長さが長い．

比較的な長文からなる自然文のような検索質問は，文書として扱うことができる．検索質問の中から共通単語を取り除いたら，残った単語は検索質問と同じグループ(分野)に属する特徴キーワードと考えられる．検索質問から抽出した特徴キーワードを用い FEM 方式で検索対象となる文書をベクトルとして表すことができる．

検索質問から抽出した特徴キーワードにより作成した文書ベクトルは，2 次元の空間上に分布する．空間の一つの軸は特徴キーワードにより構成されている．

もう一つの軸は,共通単語群 T により構成されている.

共通単語群 T により構成された軸上の値は，検索質問と直交するので,その値を検索処理中に無視できる.そのため，検索処理では，文書ベクトルの特徴キーワードにより構成された軸上の値だけを処理する．この値が大きくなればなるほど，検索質問から抽出した特徴キーワードが文書の中に出現頻度が高いことを示すので，値の大きい順で関連文書を検索結果として出力する.

本研究で考案した手法は，従来のストップリストを用い検索空間を構築手法[9]と異なる．従来の手法では,語彙からストップリストに属した単語を取り除き，残った単語を用い検索空間を構築する．構築した空間の次元数が縮小したが，残った単語数と同様数の高次元の空間である．本手法は，文書ベクトルの次元数が 2 次元までに縮小される特徴がある．また，本手法は，検索質問の単語と検索対象にある単語の論理演算により検索結果を求める方法と異なる．本手法では，文書のベクトル空間が検索質問の変化により動的に作成する．つまり，検索対象は検索質問により動的に分類される．検索結果はその分類の結果である.

4. 実験

実験では，NTCIR-1[7]を使用する．NTICIR-1 では，「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 万件が集められた．NTCIR-1 は，日本語のみの J コレクション，英語のみの E コレクション，日本語と英語の両方を含む JE コレクションがある.実験には，日本語のみの J コレクションを文書データとして使用した.

J コレクションは(1)文書，(2)検索課題，(3)正解文書リストなどから構成した.

(1)文書

日本国内 65 学協会が主催する全国大会，研究会などで発表された論文の著者抄録，約 33 万件がある.

(2)検索課題

利用者の検索要求を，自然言語を用いて，一定の書式で記述したものである.日本語で記述された 83 個の検索課題がある.

(3)正解判定

各検索課題に適合する文書のリストである．リストには，正解判定のファイル番号不正解判定のファイル番号が載っている．正解判定のファイル番号には，英文字（A）が付いているので，“A 判定”とも呼ぶ．不正解判定のファイル番号には,英文字(C)が付けられ，以下では，“C 判定”とも呼ぶ．リストには，“B 判定”もある．それは，正解ではないが検索質問にある程度関連性があるファイルへの判定である.

NTCIR-1 の検索課題は，(1)検索課題<TOPIC>，(2)タイトル<TITLE>，(3) 検索要求<DESCRIPTION>，(4) 検索要求説明<NARRATIVE>，(5) 概念<CONCEPT>と (6) 分野<FIELD>から構成される.

NTCIR-1 の検索課題は，システムに投入する文字列である「問い合わせ（または検索式)」ではなく，利用者の検索要求の自然言語によってしたものである．また，検索要求文<DESCRIPTION>には，正解判定に必要な概念がすべて含まれているが，検索対象文書中の語句との単純照合だけではすべての正解文書を検索することができない．更に，<DESCRIPTION>中の語句を含む文献が必ずしもすべて正解とはならない.

NTCIR-1 の検索課題から 3 種類の検索質問の構成が可能である：(1) 最短の問い合わせ，(2) 短い問い合わせ，(3) 長い問い合わせ．また，概念<CONCEPT>を利用すると検索性能が向上することが知られている.

本実験の検索質問は，NTCIR-1 が用意した 83 個の検索課題から，“タイトル<TITLE>”，“検索要求<DESCRIPTION>”と“検索要求説明<NARRATIVE>”の 3 部分を用い検索質問を構成する．検索性能を向上できた“概念<CONCEPT>”部分を利用しない.

構成した 83 個の検索質問は，理想的な単語分布を持っていないため，本研究の提案手法を用い構成した文書ベクトルには，ベクトルの構成“要素不足”とベクトルの構成要素に“ノイズ”が含まれるとの現象が発生した．実験では，その現象が検索精度にどのような影響を与えるかを検証する.

実験では，J コレクションの“(1) 文書データ”から論文のタイトルと抄録を，実験用の文書データとして，抽出した．また，83 個の検索課題に対し，正解判定リストから，正解文書 10 個（10 個未満の場合，すべての正解文書）と不正解文書 20 個と 90 個をランダムで抽出し，二つの文書グループ、第 1 文書群と第 2 文書群を実験用の文章集合として構成した.

提案手法は，検索質問と文書中の語句への自然言語理解ができない．不正解文書として検索結果に出現したものは，必ずしも“要素不足”と “ノイズ”の影響とはいえない．語句への自然言語理解が必要になる場合，本提案手法を用い望ましい検索結果が求めなれない.

実験では，文書の中から特徴キーワード群の抽出処理が必要である．英語文書の中に単語と単語の間に空白（space）があり，単語の抽出処理では，単語間の空白により，単語の抜き出しができる．一方，日本語文書には，単語と単語の間に空白がないので，単語の抽出処理では，別の単語抜き出しツールが必要である．本実験では，奈良先端科学技術大学院大学が開発した日本語形態素解析システム茶筌[10]を使用した.

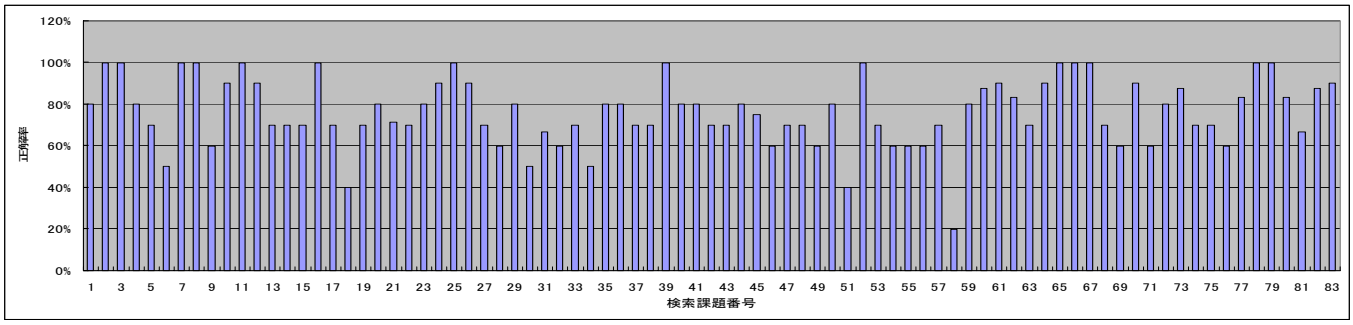

図 1 第 1 文書群の 83 個検索質問に対した検索結果

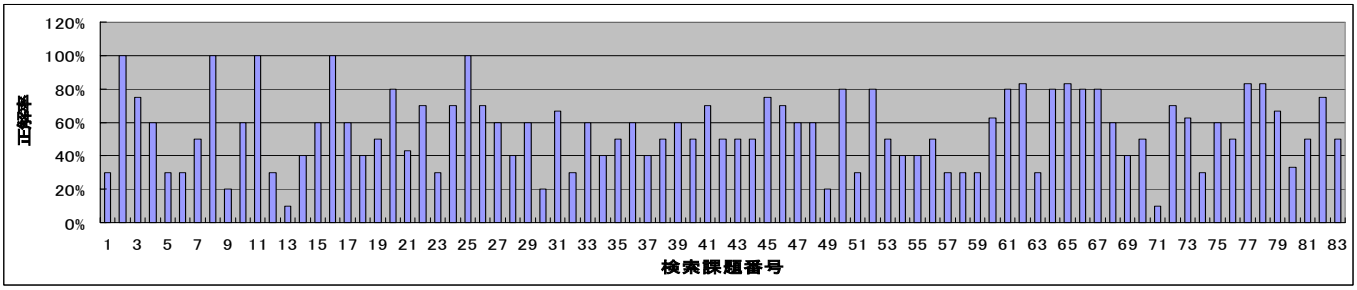

図 2 第 2 文書群の 83 個検索質問に対した検索結果

図 1 では，第 1 文書群の 83 個の検索質問に対して行った実験の結果を示した．図の中では，横軸は NTCIR-1 の検索課題番号であり，縦軸は各検索質問の正解率である．正解率は，検索結果の上位 10 件の中にある正解ファイルの数と該当検索課題を対象とした検索対象中の正解ファイル数の割合である．例えば，検索課題 1 番に対し，正解判定リストからランダムで 10 個の正解ファイルが選ばれた．検索結果の中に，10 個の正解ファイルのうちに 8 個が上位 10 位にあるため，正解率が 80%となる．

表 3 では，各検索課題の正解率を降順で並べ替えて示した．83 個の検索課題に対し行った検索実験の結果の中に，正解率が 60%以上達した課題は 77 個があり，全体課題数の約 93%を占める．そのうち，正解率が 100%になった課題は 14 個があり，全体課題数の約 17%を占める．正解率が 80%以上の課題は 41 個がり，全体課題数の約 49%である．正解率が 60%以下になった課題は 6 個があり，全体課題数の約 7%を占める．そのうち，正解率が 50%になった課題は 3 個，正解率が 40%になった課題は 2 個，正解率が 20%になった課題は 1 個である．実験結果により，平均正解率は 75%であることが分かった．

図 2 では，第 2 文書群の 83 個の検索質問に対し行った実験の結果を示した．第 2 文書群の中には，不正解ファイルの数は，第 1 文書群と比べ4.5倍に増加した．不正解ファイルの数の増加に伴い，正解率が低下したことがわかる．正解率の低下は，文書ベクトルの作成過程で，‘要素不足’と‘ノイズ’があると考えられる．次に，NTCIR の検索課題 22 番を例として，‘要素不足’と‘ノイズ’が文書ベクトルへの影響を分析する．

課題番号	正解判定	文章番号
22	正解	gakkai-0000308112
22	不正解	gakkai-0000156927
22	不正解	gakkai-0000284048
22	不正解	gakkai-0000183747
22	正解	gakkai-0000129828

図3 課題 22 番に対し行った実験結果

図 3では，検索課題 22 番に対し行った実験結果を示した．三つの不正解文書が上位 5 に入ったことが分かった．

図 3の中の文書番号は，NTCIR-1 の文書番号と一致している．検索課題 22 番によって作成した検索質問の中から特徴キーワードが抽出された．図 4では，抽出した特徴キーワード群が示した．特徴キーワード群の中には，文書ベクトルに“ノイズ”を与える単語，“たい”，“欲しい”が含まれたが，抽出された単語の全体数が少ないので，検索精度の低下は‘要素不足’と考えられる．

文書ベクトルの作成要素を増加するために，検索課題 22 番と正解文書 gakkai-0000185241 を併用し特徴キーワードを抽出することにした．

表3　各検索課題の正解率

検索課題番号	検索キーワード	上位10位にある正解ファイル数	検索対象にある正解ファイル総数数	正解率
79	β アミロイドタンパク	6	6	100%
78	テロメア	6	6	100%
67	自然 災害 時 学校 施	10	10	100%
66	大学 図書館 建築 空間	5	5	100%
65	遺伝 アルゴリズム 用い 画像 検索	6	6	100%
52	ソーラーカー	5	5	100%
25	LFG	10	10	100%
16	最大 共通 部分 グラフ	4	4	100%
11	連結 全域 グラフ	5	5	100%
8	associative rule	7	7	100%
7	認知 側面	2	2	100%
3	サンプル 複雑 性	4	4	100%
2	複合 名詞 構造 解析	5	5	100%
39	WWWトラフィック	10	10	100%
83	骨 形成 分子 メカニズ	9	10	90%
70	高齢 者 行動 特性	9	10	90%
64	生涯 学習 施設 学校	9	10	90%
61	階層 関係 自動 抽出	9	10	90%
26	語彙 機能 文法	9	10	90%
24	機械 翻訳 システム	9	10	90%
12	マイニング 手法	9	10	90%
10	キーワード 自動 抽出	9	10	90%
82	抗 マラリア 薬剤	7	8	88%
73	マウス 精子 形成	7	8	88%
60	占領 期 教育 事情	7	8	88%
80	神経 再生	5	6	83%
77	点字 翻訳	5	6	83%
62	生涯 学習 ボランティア	5	6	83%
72	博物館 資料	8	10	80%
59	シソーラス 自動 構築	8	10	80%
50	人工 知能 将棋	8	10	80%
44	画像 電子 透かし	8	10	80%
41	超 高 精細 画像 医療 技術 応用	8	10	80%
40	高 精細 画像 生成 手	8	10	80%
36	モバイル 環境 グループ ウェア 問題 点	8	10	80%
35	電子 図書館	8	10	80%
29	位置 計測	8	10	80%
23	新聞 記事	8	10	80%
20	カタカナ 外来 語	8	10	80%
4	文書 画像 理解	8	10	80%
1	ロボット	8	10	80%
45	リング 型 ネットワーク アクセス 制御 方式	3	4	75%
21	機械 翻訳 評価	5	7	71%
75	運動 シミュレーション	7	10	70%
74	細胞 周期 制御	7	10	70%
68	デジタル 著作 物	7	10	70%
63	多国 語 文字	7	10	70%
57	創造 思考 モデル 化	7	10	70%
53	電波 人体 影響	7	10	70%
48	ホーム エリア ネット	7	10	70%
47	マルチ キャスト 実装	7	10	70%
43	動画 像 圧縮 センサ	7	10	70%
42	次元 動画 像 通信	7	10	70%
38	TCP/IP 通信 スループット 特性	7	10	70%
37	バッファ 制御	7	10	70%
33	メディア 同期	7	10	70%
27	シソーラス 用い 検索	7	10	70%
22	知識 獲得	7	10	70%
19	係り 受け 解析	7	10	70%
17	排他 制御	7	10	70%
15	コロケーション	7	10	70%
14	故障 診断	7	10	70%
13	ループ 領域 解析	7	10	70%
5	特徴 次元 リダクション	7	10	70%
81	脳 性差	4	6	67%
31	データ 品質 制御	4	6	67%
76	多面体 干渉 箇所 検出	6	10	60%
71	著作 権 使用 料 課金 体系	6	10	60%
69	コンピュータ 利用 授業	6	10	60%
56	人工 物 ライフサイクル 情報 情報 知識 共有	6	10	60%
55	構造 化 文書 検索 用 全文 データベース 索引 技術	6	10	60%
54	光 ファイバ 通信 速度	6	10	60%
49	ネットワークトポロジー	6	10	60%
46	周期 リセット アルゴリ	6	10	60%
32	ネットワーク コラボレーション	6	10	60%
28	ニューラルネットワーク	6	10	60%
9	インターネット トラヒック 統計	3	5	60%
34	無線 通信 制御	5	10	50%
30	データ 駆動 画像 処理	5	10	50%
6	知的 エージェント	5	10	50%
51	次世代 インターネット	4	10	40%
18	通信 品質 保証	4	10	40%
58	Zipf 法則 応用	2	10	20%

たい	手法
システム	除外
テキスト	対話
獲得	知り
見出し	知識
言語	法
限定	欲しい
自然	理論
辞書	論文

図4　検索課題 22 番に対応した検索質問の中から抽出した特徴キーワード

図 5では，抽出した特徴キーワード群が示された．特徴キーワード群の中には，文書ベクトルに“ノイズ”を与える単語，‘しばしば’,‘たい’，‘たら’などが多数含まれた．‘ノイズ’の影響で，正解率が向上できないと考えられる．

gakkai	現れる	除外	文
しばしば	言語	照応	方法
たい	限定	条件	法
たら	後	性質	本稿
という	後件	前件	問題
において	語	操作	用
により	行なう	対話	欲しい
に関して	際	代名詞	利用
システム	自然	知り	理論
ゼロ	辞書	知識	例文
テキスト	実際	調べ	論
マニュアル	主	提案	論文
一つ	手がかり	的	
解決	手順	動作	
獲得	手法	必要	
結果	種類	表現	
見出し	述語	分類	

図 5　検索課題 22 番と正解ファイル gakkai-0000185241 を併用し抽出した特徴キーワード

課題番号	正解判定	文章番号
22	正解	gakkai-0000185241
22	不正解	gakkai-0000234764
22	不正解	gakkai-0000044911
22	不正解	gakkai-0000261284
22	正解	gakkai-0000234722

図 6　検索課題 22 番と正解ファイル gakkai-0000185241 を併用し抽出した特徴キーワードにより得られた検索結果

課題番号	正解判定	文章番号
22	正解	gakkai-0000185241
22	正解	gakkai-0000234722
22	不正解	gakkai-0000044911
22	正解	gakkai-0000129828
22	正解	gakkai-0000308112

図 7　検索課題 22 番と正解ファイル gakkai-0000185241 を併用し，‘ノイズ’単語を取り除いた後の検索結果

図 6では，‘ノイズ’単語が削除される前の検索結果を，**図 7**では，‘ノイズ’単語が削除された後の検索結果を示している．**図 6**と**図 7**を比べると，‘ノイズ’単語が削除された後，検索精度が向上したことがわかった．

しかし，**図 7**で示したように，特徴キーワードの増加と‘ノイズ’単語の削除では，完全に不正解文書を上位から排除することできない．**図 7**では，一個の不正解文書 gakkai-0000044911 が上位 5 件に含まれることを示した．この原因は，検索質問で，「辞書の見出しなどからの知識獲得は除外する」と要求しているが，本提案方式では，単語から文書ベクトルを作成し検索を行い，自然言語の理解をサポートしていないので，検索要求に満足することができない．

文書 gakkai-0000044911 の中を見ると，図 8に示したように，特徴キーワードが多数含まれることがわかった．単に検索質問から抽出した特徴キーワード群の視点から見ると，文書 gakkai-0000105058 は，正解文書として考えられる．

単語	個数
言語	3
手法	1
述語	1
知識	9
表現	6

図8　ファイル gakkai-0000044911 の中に含まれた特徴キーワード

5. 結論

本研究では、従来の次元削減手法と異なる、高次元の文書ベクトルの次元削減手法を提案した．提案手法では，比較的に長文になる検索質問から文書ベクトルを作成するための特徴キーワードを抽出し，それに基づき低次元の文書ベクトルを作成する．本手法は，文書ベクトルの次元数が 2 次元まで縮小する特徴がある．更に，本手法では，文書のベクトル空間が検索質問の変化により動的に作成する特徴がある．この特徴は，静止的に検索空間の作成手法により，検索質問に対しよい検索結果を求められると考えられる．

本論文では，国立情報学研究所が作成した NTCIR-1 文書コレクションの中にある日本語文書コレクションを実験用データとして実験を行った実験結果より，本手法を用い行った文書検索は，自然言語理解が必要な場合を除き,高い検索精度を実現したことを確認した．

本手法を用いて他の文章検索手法，例えば，LSI を使用した文章検索との対比実験を行うことを今後の研究テーマとしている．

謝辞

本実験に対し，国立情報学研究所より，NTCIR-1 テストコレクション 1 の提供をいただき，深く感謝いたします．また，日本語形態素解析システム“茶筌” を無償で配布していただいた奈良先端科学技術大学院大学に深く感謝いたします．

文　献

[1] S. K. Michael Wong, Wojciech Ziarko, P. C. N. Wong, “Generalized Vector Space Model in Information Retrieval,” SIGIR, pp. 18-25, 1985

[2] Cooper, W.S., “On deriving design equations for information retrieval systems,” JASIS, Nov. pp. 385-395, 1970.

[3] Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K. and Harshman, R., “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, Vol. 41, No. 6, pp.391-407, 1990.

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

[5] Berry, M. W., Dumais, S. T. and O’Brien, G. W., “Using linear algebra for intelligent information retrieval,” SIAM Review, Vol. 37, No.4, pp. 573-595, 1995.

[6] Chen, X. and Kiyoki, Y.: “A Query-Meaning Recognition Method with a Learning Mechanism for Document Information Retrieval,” Information Modelling and Knowledge Bases (IOS Press), Vol. XV, pp.37-54, (June 2003).

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

[8] Chen, X. and Kiyoki, Y.: “A Dynamic Retrieval Space Creation Method for Semantic Information Retrieval,” Information Modelling and Knowledge Bases, Vol.XVI, IOS Press, pp.46-63, (May 2005).

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

[10] http://chasen.naist.jp/hiki/ChaSen/