

特徴語抽出手法による文書検索システムの 精度を向上するための実験

福村 壽晃[†] 木寺 悠介^{††} 鷹野 孝典^{†††} 陳 幸生^{†††}

[†], ^{†††} 神奈川県厚木市下荻野 1030
^{††} 神奈川県厚木市下荻野 1030
E-mail: [†], ^{††}{s035231, s055808}@cce.kanagawa-it.ac.jp, ^{†††}{takan, chen}@ic.kanagawa-it.ac.jp

あらまし 本稿では、FEM (Feature Extraction Model) 方式による文書検索システムについて述べる。本方式の特徴は、人間により互いに意味的に独立であると判断される異なる文書群より文書ベクトル空間を生成する点にある。本方式では、生成した文書ベクトル空間上における文書の分類結果を、人間の意味的な判断による文書の分類結果と一致させることを実現可能とし、FEM 方式による文書検索システムの検索精度を向上することができる。本研究では、提案方式により生成したベクトル空間上に射影した文書群の分布特徴を調べるための評価実験を行い、提案方式の有効性を確認した。

キーワード 文書検索、ベクトル空間モデル、意味的検索、特徴抽出、特異値分解

Experiments for Improving the Performance of a Document Retrieval System Developed on a Feature Extraction Method.

Toshiaki FUKUMURA[†] Yusuke KIDERA^{††} Kosuke TAKANO^{†††} and Xing Chen^{†††}

[†], ^{†††} Dept. of Info. & Comp. Sciences, Kanagawa Institute of Technology 1030 Simo-Ogino, Atsugi-shi Kanagawa,
243-0292 Japan

^{††} Dept. of Info. & Comp. Science, Grad. Cause, Kanagawa Institute of Technology 1030 Simo-Ogino, Atsugi-shi
Kanagawa, 243-0292 Japan

E-mail: [†], ^{††}{s035231, s055808}@cce.kanagawa-it.ac.jp, ^{†††}{takan, chen}@ic.kanagawa-it.ac.jp

Abstract We have developed an document retrieval system based on our proposed method FEM (Feature Extraction Model). In this method, documents are represented as vectors on a vector space created based on FEM. This vectors space has the characteristics that it is created from the documents which are divided into different groups by human. Documents in different groups are considered to be independent from each other by human on semantic meanings. In order to improve the performance of our document retrieval system, we performed experiments to analyze the distribution characteristics of document vectors on the vector space. Based on the retrieval results, we find a method to improve the performance of the system.

Keyword document retrieval, vector space, semantic retrieval, feature extraction, SVD

1. はじめに

文書検索において、文書中出现する単語の分布特性を利用し、検索精度を向上することができると考えられる。同一の分野に属する文書においては、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群は、化学分野に関する論文において出現頻度が低い。

ベクトル空間モデルを用いた文書検索方式では、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確認されている[1]。この方式で

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、重み付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。しかしながら、ベクトル空間モデルを用いた検索方式では、各文章ベクトルが数万から数十万の疎な高

次元ベクトルになり、文書中に出現する単語の分布特性について考慮していない。

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方式である[4]。LSI方式では、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を低次元の文書ベクトルに縮小する。LSI方式では、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。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。

しかし、SVD計算を用いた場合は、文章中出现している単語の分布特徴に基づいて文書の分類を行うため、その分類結果は、必ずしも人間が単語の意味に基づいて文書を分類した結果と一致するとは限らない。

我々は、これまで、人間が単語の意味に基づいて分類した文書群中出现する単語分布の特徴を利用し、計算コストの大きいSVD計算を行わずに、低次元のベクトル空間を作成する方式 (Feature Extracting Model, FEM) を提案してきた[6]。この方式では、単語の分布特徴から、ベクトル空間の生成に必要な特徴単語を抽出する。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、重み付きの出現頻度である。

本研究では、FEM方式による文書検索システムの検索精度を向上することを目的として、FEM方式を用いて意味的に独立した複数の文書群から作成したベクトル空間上に射影した文書群の分布特徴を調べるための評価実験を行った。本実験では、文書ベクトル空間を作成するための文書テストデータセットとしてNTCIR-1[7]を用いた。

2. FEM方式

本章では、FEM方式の概要について述べる。

2.1. FEM方式の概要

以下の説明では、人間が単語の意味に基づいて分類した文書群中出现する単語分布の特徴を有する文書集合があることを前提とする。この前提を必要としない文書ベクトルの作成手法については、文献[8]で述べられている。

以下、ベクトル空間を作成するため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、文書の意味に基づいて、人間によりいくつかのグループに分類される。同一グループに属する文書は似たような単語 (以下、特徴単語と呼ぶ) 分布を有する。FEM

方式では、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 (以下特徴単語群と呼ぶ) を抽出する。抽出した特徴単語は、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という性質を持つ。

サンプル文章群 $d_1, d_2, d_3, \dots, d_m$ 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, \dots, C_q で表す。特徴単語を 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単語群を表す。例えば、グループ C_1 の特徴単語群は $\{t_1, \dots, t_a\}$ である。

表1 特徴単語群と文書グループ

	C_1	C_2	...	C_q
T_1, \dots, t_a	d_1, \dots, d_a			
t_{a+1}, \dots, t_b		d_{i+j}, \dots, d_{i+j}		
t_{n-f}, \dots, t_n				d_{m-s}, \dots, d_m

FEM方式では、各文書ベクトルは、各文書中における特徴単語の出現頻度、または、重み付きの出現頻度により作成される。文書ベクトル d_j の要素を $e_{i,j}$ とすると、文書ベクトルは、表2のように表される。表2中の各列は、それぞれ文書ベクトルを表しており、各行は、それぞれ分類した文書のグループを表している。

表2 文書ベクトル

	d_1	d_2	...	d_m
C_1	$e_{1,1}$	$e_{1,2}$		$e_{1,m}$
C_2	$e_{2,1}$	$e_{2,2}$		$e_{2,m}$
...		
C_q	$e_{q,1}$	$e_{q,2}$		$e_{q,m}$

文書ベクトル d_j の要素 $e_{i,j}$ の値は、グループ C_i における特徴単語が文書 d_j 中に出現する頻度、または、重み付きの出現頻度である。例えば、文書グループ C_1 の特徴単語群を $\{t_1, t_2, t_3\}$ 、および、 v_{11}, v_{12}, v_{13} をそれぞれ、単語 t_1, t_2, t_3 が文書 d_j 中に出現する頻度とすると、 $e_{i,j}$ の値は次の通りである。

$$e_{i,j} = v_{i1} + v_{i2} + v_{i3}.$$

あるいは、

$$e_{i,j} = w_{i1} \times v_{i1} + w_{i2} \times v_{i2} + w_{i3} \times v_{i3}.$$

式の中の w_{i1}, w_{i2}, w_{i3} は、出現頻度 v_{i1}, v_{i2}, v_{i3} に対する重み係数である。

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。

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

検索質問も文書と同様に q 次元のベクトル空間上に

射影され、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。

2.2. 特徴単語抽出方式

FEM方式では、各文書グループに属する特徴単語を抽出し、文書ベクトルを作成する。文書ベクトルの作成について次の手順で実現する。

まず、サンプル文書群の特徴単語を次のように抽出する：

グループ C_i の特徴単語は、同一グループに属する文書の中のみ出現し、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 の特徴単語群を K_i と定義し、特徴単語を t と定義すれば、特徴単語の分布は次の式で表せる。

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

特徴単語以外の単語、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。特徴単語の分布により、特徴単語 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、少なくとも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 を使い、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単語になる。つまり、単語 - 文書行列を特徴単語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単語は、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_i の特徴単語群 $\{t_1, \dots, t_a\}$ は、グループ C_j に属する文書、 d_j の中のみ出現し、他のグループ特徴単語は 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。

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

式(2)は、文書ベクトル \mathbf{d}_j の値が同じグループ C_j に属する特徴単語の出現頻度 $e_{1,j}$ 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単語群だけを利用すれば可能である。

サンプル文書群の特徴単語の分布により、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単語である。もし、文書 d_j がすべての特徴単語を有する場合、共通単語の取り除く処理を行うだけで、すべての特徴単語が抽出されることができる。

空間上それぞれの軸は互いに直交するため、検索処理では、文書ベクトルの特徴単語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単語が文書の中に出現頻度が高いことを示すので、値の大きい順に関連文書を検索結果として出力する。

3. 実験

3.1. 実験環境

実験では、NTCIR-1[7]を使用する。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。実験では、日本語のみを対象とした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されている。

(1)文書

日本国内65学協会が主催する全国大会、研究会等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。正解判定のファイル番号には、英文字(A)が付いており、“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。また、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について“B判定”がある。

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。本実験では、文書の中から特徴単語群の抽出処理に、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釜[10]を使用した。

3.2. 実験 1

3.2.1. 目的

FEM 方式による文書検索システムの検索精度を向上することを目的として、FEM 方式を用いて意味的に独立した複数の文書群から作成したベクトル空間上に射影した文書群の分布特徴に関する評価を行う。

3.2.2. 実験方法

NTCIR-1 より、3 つの検索課題を抽出し、それぞれの検索課題より文書群を 15 件ずつ、合計 45 件を用意する。それぞれの文書群には、文書番号とは別に、 d_1 から d_{45} までの文書 ID を付与する。また、15 件のうち正解判定の文書を 5 件ずつ含める ($d_1 \sim d_5$, $d_{16} \sim d_{20}$, および $d_{31} \sim d_{35}$)。本実験では、以下の 2 つの文書ベクトル空間を生成する。

1. 提案方式により作成したベクトル空間：正解判定の文書群 $d_1 \sim d_5$, $d_{16} \sim d_{20}$, および $d_{31} \sim d_{35}$ より、それぞれ 1 つの文書を選択して得られる、3 つの文書より抽出された特徴単語群を用いて、提案方式により 3 次元の文書ベクトル空間を作成する。各文書群より抽出された特徴単語群より構成される軸を、それぞれ、 C_1 , C_2 , C_3 とする (図 1)。

2. 特異値分解により作成したベクトル空間：全 45 件の文書群を対象として特異値分解を行い、その結果に基づき文書ベクトル空間を作成する。

本実験では、それぞれの文書ベクトル空間上に文書 45 件を射影し、文書群の分布の特徴を分析する。

3.2.3. 実験結果

3 つの文書 d_5 , d_{16} , および、 d_{34} より抽出された特徴単語群を用いて、文書ベクトル空間を作成した場合の実験結果を表 3~5、および図 2, 3 に示す。表 3~5 において文書番号に下線の引いてある文書は、各軸への文書群の射影結果における正解文書であることを示している。ここで、各文書の座標は、文書と各軸を構成する特徴単語群間の相関値を示している。

提案方式による検索空間では、 C_1 軸 (d_5) および C_2 軸 (d_{16}) へ射影した文書群において、上位 5 件中に正解文書が 4 件または 5 件があり、意味に基づいた適切な分布となっていることがわかる。ただし、 C_3 軸 (d_{34}) へ射影した文書群において、上位 5 件中に正解文書 3 件があるが、他の 2 軸へ射影した場合に比べて、やや分布が適切に行われていないことが確認できる。これは、文書 d_{34} においては、人間の判断に基づいた文書分類と単語に基づいた分類が一致していないことが要因であると考えられる。

一方、特異値分解に基づいた検索空間を 3 次元に圧縮した場合の文書の分布図を図 2 に示す。図 2 において、軸 1 を構成する文書群の正解文書 ($d_1 \sim d_5$) は、高い相関量を示していない。これは軸 1 のグループの

文書間において、文書の意味に基づいた相関量が出てこないことを表している。軸 2 を構成する文書群の正解文書 ($d_{16} \sim d_{20}$)、および軸 3 を構成する文書群の正解文書 ($d_{31} \sim d_{35}$) に着目した場合も、同様である。これらの結果は、特異値分解に基づいた文書ベクトル空間では、その分類結果が、必ずしも人間が単語の意味に基づいて文書を分類した結果と一致するとは限らないことを示している。

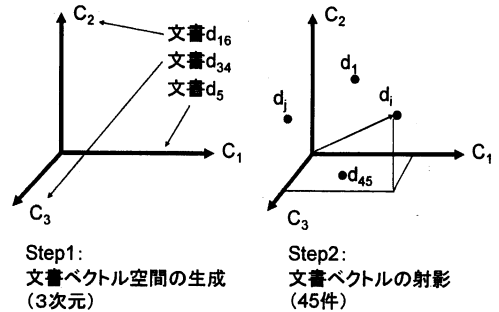


図 1 提案方式による文書ベクトル空間生成(実験 1)

表 3 C_1 軸 (d_5) に文書群を射影した結果 (実験 1)

順位	文書番号	文書 ID	座標 (相関値)
1	<u>gakkai-0000012880</u>	d5	8.2825
2	<u>gakkai-0000010423</u>	d2	3.7187
3	<u>gakkai-0000010501</u>	d3	2.7045
4	<u>gakkai-0000005053</u>	d1	1.8593
5	<u>gakkai-0000000033</u>	d21	1.6903
6	<u>gakkai-0000000899</u>	d37	1.3522
7	<u>gakkai-0000000955</u>	d39	1.1832
8	<u>gakkai-0000000657</u>	d24	1.1832
9	<u>gakkai-0000001356</u>	d9	1.1832
10	<u>gakkai-0000001545</u>	d11	1.0142

表 4 C_2 軸 (d_{16}) に文書群を射影した結果 (実験 1)

順位	文書番号	文書 ID	座標 (相関値)
1	<u>gakkai-0000017019</u>	d16	11.9097
2	<u>gakkai-0000019367</u>	d17	5.7287
3	<u>gakkai-0000133732</u>	d18	5.1257
4	<u>gakkai-0000180412</u>	d19	3.7689
5	<u>gakkai-0000183936</u>	d20	3.6181
6	<u>gakkai-0000000657</u>	d24	1.9598
7	<u>gakkai-0000000575</u>	d23	1.9598
8	<u>gakkai-0000001119</u>	d27	1.8091
9	<u>gakkai-0000001109</u>	d24	1.8091
10	<u>gakkai-0000010423</u>	d2	1.6583

表 5 C_3 軸 (d_{34}) に文書群を射影した結果 (実験 1)

順位	文書番号	文書 ID	座標 (相関値)
1	<u>gakkai-0000017019</u>	d34	11.1962
2	<u>gakkai-0000019367</u>	d35	4.166
3	<u>gakkai-0000133732</u>	d37	2.2132
4	<u>gakkai-0000180412</u>	d32	2.083
5	<u>gakkai-0000183936</u>	d43	1.9528
6	<u>gakkai-000000657</u>	d31	1.8226
7	<u>gakkai-000000575</u>	d45	1.6925
8	<u>gakkai-0000001119</u>	d44	1.6925
9	<u>gakkai-0000001109</u>	d17	1.6925
10	<u>gakkai-0000010423</u>	d15	1.6925

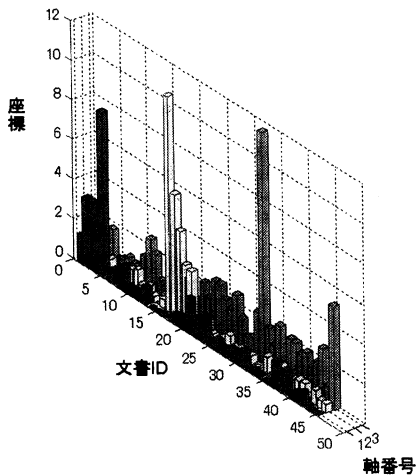


図 2 提案方式による検索空間上 (d_5 , d_{16} , d_{34}) における文書の分布 (実験 1)

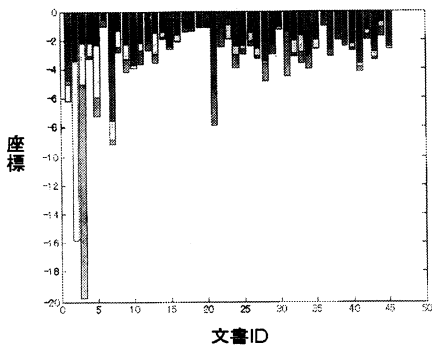


図 3 特異値分解空間上における文書の分布 (実験 1)

3.3. 実験 2

3.3.1. 目的

実験 1 において、 C_3 軸 (d_{34}) では、他の 2 軸 (C_1 および C_2) の分布よりも意味に基づく分類が適切に行われていなかった。本実験では、 C_3 軸を構成する文書群に対して、新たに文書を統合した場合において、その統合により生成された軸への文書群の射影結果が、人間の判断による文書の分類結果と一致させることが実現可能であることを検証する。

3.3.2. 実験方法

d_{34} に加え、 d_{33} より抽出した特徴単語より構成される軸を C'_3 とする。 C'_3 軸への 45 件の文書群を射影し、 C_3 軸 (d_{34}) へ射影した場合よりも、正解文書が上位の座標 (相関値) に射影されていることを確認する。

3.3.3. 実験結果

実験結果を表 6 および図 4 に示す。表 6 は、 C'_3 軸 ($d_{33}+d_{34}$) へ文書群を射影した結果を示している。表 6 では、上位 5 件中に正解文書の全 5 件があり、意味的に適切な文書分布となっている。この結果により、 C'_3 軸 ($d_{33}+d_{34}$) へ文書群を射影した場合は、 C_3 軸 (d_{34}) へ射影した場合よりも、人間の判断による文書の分類結果と一致していることが確認できる。また、図 4 のグラフより、他の軸への射影した文書群の分布も実験 1 の結果を保ち、人間の意味的な判断に基づいた適切な分布となっていることがわかる。

この実験結果は、人間が単語の意味に基づいて文書を分類した結果と一致しない場合において、適切な特徴単語を抽出する文書群を統合することにより、人間の判断による文書の分類結果と一致させることが実現可能なことを示している。このような人間の判断による文書の分類結果と一致させることにより、提案方式による検索システムにおいて、検索精度を向上させることが可能となる。

表 6 C'_3 軸 ($d_{33}+d_{34}$) に文書群を射影した結果 (実験 2)

順位	文書番号	文書 ID	座標 (相関値)
1	<u>gakkai-0000185065</u>	d34	7.9169
2	<u>gakkai-0000065098</u>	d33	7.6813
3	<u>gakkai-0000234682</u>	d35	3.0576
4	<u>gakkai-0000007590</u>	d31	2.8541
5	<u>gakkai-0000008528</u>	d32	2.2555
7	<u>gakkai-0000003281</u>	d15	2.0912
7	<u>gakkai-0000003281</u>	d45	2.0912
8	<u>gakkai-0000002753</u>	d43	2.0517
9	<u>gakkai-0000010501</u>	d37	2.0189
10	<u>gakkai-0000000899</u>	d17	1.7886

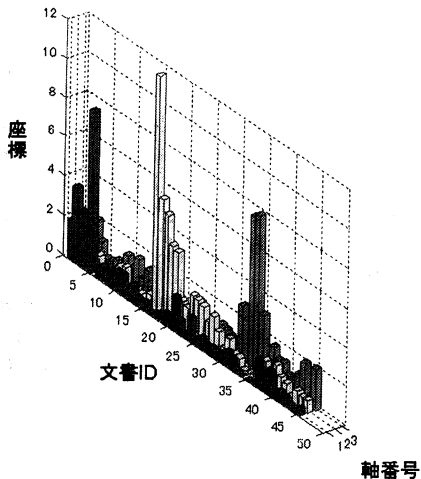


図 4 提案方式による検索空間上(d_5 , d_{16} , d_{33+34})における文書の分布 (実験 2)

4. 結論

本稿では、FEM(Feature Extraction Model)方式による文書検索システムについて提案した。提案方式により生成されるベクトル空間の特徴は、人間により互いに意味的に独立であると判断されるグループからなる文書群より生成されるという点にある。

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用いて、FEM 方式による文書検索システムの検索精度を向上することを目的として、提案方式により生成したベクトル空間上に射影した文書群の分布特徴を調べるための評価実験を行い、提案方式の有効性を確認した。

実験において、人間の判断による意味的に独立な 3 文書を用いて、3 次元の文書ベクトル空間の生成を行い、それぞれの軸への文書群の射影結果が、意味的に適切な分布となっていることを確認した (実験 1)。また、文章中出现している単語の分布特徴に基づく文書の分類結果が、人間が単語の意味に基づいて文書を分類した結果と一致しない場合において、適切な特徴単語を抽出する文書群を統合することにより、人間の判断による文書の分類結果と一致させることが実現可能なることを確認した (実験 2)。これにより、提案方式による検索システムにおいて、検索精度を向上させることが可能となる。

今後の課題として、実験 2 で示したケースにおいて、文書ベクトル空間上の文書群を対象とした、人間の意味的な判断に近い文書分布を実現する軸を構成するた

めの、特徴単語を抽出する文書群を選択するアルゴリズムの開発を行う予定である。

5. 謝辞

本研究に関して、実験のサポートをして頂いた神奈川県立川崎大学大学院工学研究科情報工学専攻増田圭祐氏、同大学情報学部情報工学科塩原慶一氏にこの場を借りて感謝申し上げます。また、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「茶釜」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/>
- [11] 木寺悠介, 陳幸生, 塩原慶一, "検索質問にあわせて文章ベクトルの次元削減手法," 電子情報通信学会第 17 回データ工学ワークショップ (DEWS2006) 論文集, (8 pages), March, 2006.