特徴語抽出方式における文書ベクトル空間を対象とした フィードバック機構による検索精度の改善方式の実現

鷹野 孝典† 増田 圭祐‡ 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
‡神奈川工科大学大学院工学研究科情報工学専攻 〒243-0292 神奈川県厚木市下荻野 1030
E-mail: †{takano, chen}@ic.kanagawa-it.ac.jp, ‡s065819 @ cce.kanagawa-it.ac.jp

あらまし 本稿では，特徴語抽出方式（Feature Extraction Model，FEM）による文書ベクトル空間を，自動的なフィードバック機構を用いて検索精度を改善する方式について述べる．FEM 方式では，人間により意味や内容に基づいて分類された異なる文書群により，利用者の検索意図や目的との関連性を有する特徴単語を抽出し，その特徴単語を利用して，文書ベクトル空間を生成する．本研究では，生成した文書ベクトル空間上に射影された文書ベクトルの形成に必要な特徴単語成分を抽出するための基本的アルゴリズムを示し，文書ベクトル空間を構成する特徴単語の追加，削除を行うことにより文書ベクトル空間を再構成する方法を提案する．再構成された文書ベクトル空間上では，文書ベクトル分布が改善されるため，FEM 方式による文書検索システムの検索精度の向上が実現できる．本研究では，学術論文の要旨文書群を対象とした検索実験により，提案方式の有効性および実現可能性を確認した．

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，フィードバック機構

A Method of Improving the Performance of a Document Vector Space Developed on a Feature Extraction Method by A Feedback System

Kosuke TAKANO† Keisuke MASUDA‡ 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: †{takano, chen}@ic.kanagawa-it.ac.jp, ‡s065819 @ ic.kanagawa-it.ac.jp

Abstract In this paper, a document retrieval system is introduced, which is developed based on a method referred to as Feature Extraction Model (FEM). In this method, pre-prepared documents are classified into different clusters by human beings based on the meaning and contents of documents. Terms which closely correlate with the purpose of query are extracted from the documents by using the classification information and documents are represented as vectors on a vector space created by the extracted terms. The created vector space has the characteristics that document vectors on the space cluster together based on the meaning and contents of documents. In this paper, we propose a basic algorithm for an automatic feedback method to reconstruct the vector space by adding or removing some terms which are used to construct the space. Retrieval results are improved by reconstructing the vector space because the distribution of documents on the reconstructed vector space reflect query purpose better. The efficiency of the proposed method are clarified and the possibility for implementing the proposed method are confirmed based on our experiments.

Keyword document retrieval, vector space, semantic retrieval, feature extraction, feedback mechanism

1. はじめに

インターネット上の Web 文書群，PC 上に蓄積されている文書群およびオンライン・データベース群をはじめとして，コンピュータ上で利用可能な様々な形式の情報源が劇的に増加しており，これらの膨大な情報源から，利用者の意図や目的，および状況に応じた情報獲得の実現の重要性が高まっている．

我々は，これまで，人間が文書の意味や内容に基づいて分類した文書群中に出現する単語群の特徴を利用し，小さな計算コストで，低次元のベクトル空間を作成する方式（Feature Extraction Model, FEM）を提案してきた[3, 7, 8, 10].

この方式では，意味や内容に基づいて分類された異なる文書群により，利用者の検索意図や目的との関連を有する特徴単語を抽出し，その特徴単語を利用し，文書ベクトル空間を生成する．生成した文書ベクトル空間は，利用者の検索意図や目的を反映できる特徴がある[7]．文献[10]では，FEM 方式により生成した文書ベクトル空間を対象として，利用者自身がその検索意図に基づいて，文書ベクトルの形成に必要な特徴単語成分を抽出して元の文書ベクトル空間に反映することにより，その文書ベクトル空間を用いた検索システムの検索精度の向上が実現可能であることを確認した．

本研究では，生成した文書ベクトル空間上に射影された文書ベクトルの形成に必要な特徴単語成分を自動的に抽出するための基本的アルゴリズムを示し，文書ベクトル空間を構成する特徴単語の追加，削除を行うことにより文書ベクトル空間を再構成する方法を提案する．

具体的には，文書の意味や内容に基づいて分類された文書群より抽出された特徴単語群において，検索意図に沿った文書群において追加した方が良い特徴単語や，検索意図に合致しない文書群において削除した方が良い特徴単語を，検索処理における各文書の相関量の算出結果に基づいて抽出する基本的アルゴリズムを示す．提案アルゴリムにより抽出された特徴単語群を，元の文書ベクトル空間へ追加・削除するフィードバック処理を行うことにより，文書ベクトル空間の再構成を行い，この文書ベクトル空間を用いた検索システムの検索精度の向上が実現可能なことを検証する．

本研究では，学術論文の要旨に関する文書群（NTCIR-1[11]）を用いた検索実験を行い，提案方式の実現可能性および有効性を検証した．

2. 関連研究

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, 16]．

Latent Semantic Indexing (LSI) [6]は，特異値分解（SVD）を用いた文書ベクトルの次元縮小方式である[13]．しかし，SVD 計算を用いた場合は，文章中に出現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．文献[7]において，FEM 方式は，ベクトル空間上の各軸上に射影した文書群の分布が，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確認した．

利用者のフィードバックによる検索精度の向上を実現する方式として，適合フィードバック[1]や文献[4]の方式等が提案されている．適合性フィードバックは，検索結果から正解文書や不正解文書を指定し，それに基づいて検索質問を修正する方式である．また，文献[4]は，利用者の意図に合った情報を効率よく抽出するための，構造化されたレコード抽出方式を提案している．検索精度を向上するために，利用者のフィードバックを利用している．

これらに対して，FEM 方式は，異なる複数のサンプル文書群より，それぞれに共通な特徴単語を排除した軸より構成される直交ベクトル空間を構築する方式である．本研究では，FEM 方式による文書検索システムの検索精度の向上を目的としたフィードバック機構を実現するために，FEM 方式により生成された文書ベクトル空間上に射影された文書ベクトルに着目し，その文書ベクトルの形成に必要な特徴単語成分を自動的に抽出する基本的アルゴリズムを示し，元の文書ベクトル空間を再構成する方法を提示する．

3. FEM 方式の概要

本章では，FEM 方式の概要について述べる．詳細は，文献[3]に述べられている．FEM 方式は，(1)文書ベクトル空間の生成プロセス，および，(2)文書検索プロセス，により実現される．これらのプロセスについて，下記に説明を行う．

3.1. 文書ベクトル空間の生成プロセス

文書ベクトル空間の生成プロセスでは，まず，文書ベクトル空間を作成するためのサンプル文書群を用意する．サンプル文書群は，文書の意味や内容に基づいて，人間によりいくつかのクラスタに分類される．同一クラスタ中のサンプル文書群では，それぞれ意味や内容が類似している文書が選択されているので，同一クラスタに属する各サンプル文書中には，似たような単語（以下，特徴単語と呼ぶ）が出現すると仮定される．この仮定のもとで，これらの特徴単語は，あるクラスタに属するサンプル文書群における出現頻度が高いが，他のクラスタに属したサンプル文書群における出現頻度が低いという性質を持つ．サンプル文書群により生成された文書ベクトル空間上に，検索対象となる文書群を射影することで，文書ベクトル空間の生成が実現される．

表1 特徴単語群と文書クラスタ

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

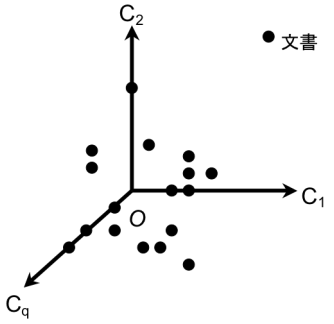

図 1 文書ベクトル空間

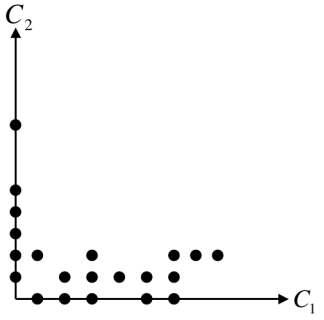

図 2 文書ベクトル空間における文書分布の例

3.2. 検索空間の生成

サンプル文章群 $d_1, d_2, d_3 \cdots, d_m$ に対し，文書を q 個のクラスタに分ける．各クラスタを $C_1, C_2, \ldots, C_q$ で表す．クラスタ中の特徴単語を t とし，クラスタ C_i の特徴単語群を K_i とすると，

$$K_i = \left\{ t \mid t \in C_i \wedge t \in C_j \wedge i \neq j \right\} (i, j = 1,2,\cdots,q)$$

文書クラスタとサンプル文書，および特徴単語群の関係を表 1 に示す．第 1 列の各項目は，文書クラスタ C_i の特徴単語群 K_i を表している．例えば，クラスタ C_1 の特徴単語群 K_1 は $\{t_1, \ldots, t_a\}$ である．

クラスタ C_i を表現するベクトルを $\mathbf{C}_i$ とし，$\mathbf{C}_i$ の単位ベクトルを $\mathbf{c}_i$ とする．サンプル文書が q 個のクラスタに分類されるとすると，q 個のクラスタベクトル $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q$ が生成される．この q 個の単位クラスタベクトルで構成される空間を，検索空間(図 1)と呼ぶ．2 つの異なるクラスタベクトルの内積は $\mathbf{c}_i \cdot \mathbf{c}_j = 0$ であるので，検索空間は q 次元の直交空間である．

3.3. 文書ベクトルの射影

検索空間上へ射影された各文書ベクトルは，クラスタ・ベクトル $\mathbf{c}_i$ を用いて，下記の式で表される．

$$\mathbf{d}_j = \sum_{i=1}^{q} e_{i,j} \mathbf{c}_i \qquad (1)$$

文書群が q 個のクラスタに分かれる場合，文書ベクトルは，式(1)のように q 次元のベクトルとなる．文書ベクトル $\mathbf{d}_j$ の要素 $e_{i,j}$ の値は，クラスタ C_i における特徴単語が文書 d_j の中に出現する頻度，または，重みつきの出現頻度である．クラスタ C_i の各特徴単語 $t_1, t_2, \ldots, t_a$ が文書 d_j に出現する頻度を $v_{t1}, v_{t2}, \ldots, v_{ta}$，および，それぞれの出現頻度に対する重み計数を $w_{t1}, w_{t2}, \ldots, w_{ta}$ とすると，

$$e_{i,j} = v_{t1} + v_{t2} + v_{t3} \qquad (2)$$

あるいは，

$$e_{i,j} = w_{t1} \times v_{t1} + w_{t2} \times v_{t2} + w_{t3} \times v_{t3} \qquad (3)$$

図 2 は，クラスタ C_1 と C_2 から構成される 2 次元文書ベクトル空間上に，各文書を射影した場合の文書分布の例を示している．

3.4. 文書検索プロセス

文書検索プロセスは，(1)指定された問い合わせに基づいた部分空間の選択処理，および，(2)部分空間上で文書のランキング処理，という 2 つの処理により実現される．部分空間は，問い合わせに基づき，文書ベクトル空間上で選択される．q 次元の文書ベクトル空間より選択される部分空間は，v 次元のベクトル空間である．v 次元部分空間は，v 個のクラスタに相関を持つ．部分空間を選択するステップを下記に示す．

ステップ(1) 問い合わせ Q が与えられた場合，パターンマッチングにより，問い合わせ中の単語を含む文書を検索する．

ステップ(2) 検索された全ての文書ベクトルについて，$|e_{i,j}\mathbf{c}_i|$ が最大となるクラスタ・ベクトル $\mathbf{c}_i$ を抽出する．抽出されたクラスタ群を表すベクトルを $\mathbf{q}$ 初期値 $\mathbf{0}$ とすると，抽出されたクラスタ・ベクトル $\mathbf{c}_i$ は，$\mathbf{q}$ に加えられる．

$$\mathbf{q} = \mathbf{q} + \mathbf{c}_i \qquad (4)$$

例えば，ステップ(1)で得られた全文書が，$\mathbf{d}_1 = 5\mathbf{c}_1 + 2\mathbf{c}_2$, $\mathbf{d}_2 = 3\mathbf{c}_2$ のとき，$\mathbf{d}_1$, $\mathbf{d}_2$ において，$|e_{i,j}\mathbf{c}_i|$ を最大とするクラスタ・ベクトル ${}_j\mathbf{c}_i$ は，それぞれ $\mathbf{c}_1$，$\mathbf{c}_2$ であるので，$\mathbf{q} = \mathbf{c}_1 + \mathbf{c}_2$ となる．

ステップ(3)問い合わせに相当する部分空間 S の選択は，$\mathbf{q}$ と $\mathbf{c}_i$ の内積計算により行う．部分空間選択のための閾値を ε とすると，$\mathbf{q} \cdot \mathbf{c}_i < \varepsilon$ を満たす $\mathbf{c}_i$ が部分空間 S に加えられる．ステップ(2)で $\mathbf{q} = \mathbf{c}_1 + \mathbf{c}_2$ が得られたとすると，閾値が 1 である場合，$\mathbf{q}$ と $\mathbf{c}_1$ および $\mathbf{c}_2$ の内積は，それぞれともに 1 であるので，$\mathbf{c}_1$ および $\mathbf{c}_2$ が部分空間 S に加えられ，$S = \{\mathbf{c}_1, \mathbf{c}_2\}$ となる．

部分空間上に射影された文書ベクトル $\mathbf{d}_j$ は，クラスタ・ベクトル $\mathbf{c}_i$ を用いて次式のように表される．

$$\mathbf{d}_j = \sum_{i=1}^{v} e_{i,j} \mathbf{c}_i (\mathbf{c}_i \in S) \qquad (5)$$

次に，各文書は，部分空間上における各文書ベクト

ルのノルムを算出し，ランキングされる．

$$\left|\mathbf{d}_j\right| = \left|\sum_{i=1}^{v} e_{i,j}\mathbf{c}_i\right| \left(\mathbf{c}_i \in S\right) \qquad (6)$$

4. 文書ベクトル空間の検索精度の改善のためのフィードバック機構

本章では，提案方式である文書ベクトル空間の検索精度の改善のためのフィードバック機構について述べる．提案方式は，文書ベクトル空間を形成する特徴単語群について，追加，削除を行うフィードバック操作(図 4)により，文書ベクトル空間を再構成する方式である．提案方式により，FEM 方式により実現した文書ベクトル空間の検索精度の向上が実現可能となる．

検索空間における検索精度を向上させるためには，検索空間を構成する各クラスタ･ベクトル $\mathbf{c}_i$ において，$\mathbf{c}_i$ に射影する文書分布が適切になっている必要がある．適切な文書分布の下で，3.4 節において述べた文書検索処理において問い合わせと関連のある部分空間が選択され，検索の意図や目的に応じた，文書のランキングが実現される．

特徴単語の追加・削除によるフィードバックは，図 4 に示すに 3 つのプロセスが考えられる．

Feedback1 ストップワードへのフィードバック

Feedback2 検索空間を再構成する際のフィードバック (提案方式)

Feedback3 文書を射影する際のフィードバック

Feedback1 では，特徴単語の追加が行えず，Feedback3 では，特徴単語を追加した場合，検索空間の直交性が保持されない．提案方式では，Feedback2 により，特徴単語の追加・削除を行うとともに，直交性を保持した検索空間を再構成を行う．Feedback2 をクラスタ C_i に対して実行する基本的アルゴリズムを下記に示す．

ここで，式(2)において，式の各項 v_{tp}(または，$v_{tp}\times w_{tp}$)を，特徴単語 t_p の寄与値と定義する．寄与値は，文書 d_j 中の特徴単語 t_p がクラスタ C_i において，どの程度の相関量の算出に寄与しているかを示す値である(図 3)．クラスタ C_i における文書ベクトル $\mathbf{d}_j$ において，特徴単語の寄与値を分析することにより，クラスタ C_i におけるその文書の意味や内容を判別できる．

フィードバック処理の基本的アルゴリズム

Step-1 3 章で示した方式により，サンプル文書より，カテゴリ毎に特徴単語 T_i を(i=1,2,...q)を抽出する．このとき除外された共通単語を l_+する．l_+ は，フィードバック処理により，追加される候補となる特徴単語群である．

Step-2 特徴単語群 T_i より，文書ベクトル空間を生成し，各文書群を射影する．文書ベクトルは，無限大ノルムで正規化を行う．

Step-3 正解文書と上位 n 件の不正解文書について座標値の総和平均の比率 R_n を算出する．正解文書集合を K，上位 n 件の不正解文書を I_n，軸 i 文書 d の座標値を $V_i(d)$，および文章集合 S の要素数を $N(S)$で表すと，

$$R_n = \left(\frac{1}{N(K)}\sum_{d\in K} V_i(d)\right) \Big/ \left(\frac{1}{N(I_n)}\sum_{d\in I_n} V_i(d)\right) \qquad (7)$$

Step-4 最大座標を持つ不正解文書 $d_{\max}$ に含まれる特徴単語群を削除候補 l_-とする．l_-の各特徴単語のうち，正解文書の寄与値の総和が閾値 ε_1 以下，かつ不正解文書 $d_{\max}$ の寄与値が閾値 ε_2 以上の特徴単語 t を T_i から削除する．文書 d における特徴単語 t の寄与値を $H_d(t)$とすると，

$$\sum_{d\in K} H_d(t) \le \varepsilon_1 \wedge H_{d_{\max}}(t) \ge \varepsilon \qquad (8)$$

Step-5 l_+から，特徴単語 t を追加し，文書ベクトル空間を再構成した場合に，正解文書の寄与値の総和が最大，かつ不正解文書の寄与値の総和の差分が閾値 ε_3 以下である特徴単語 t を T_i に追加する．

$$\sum_{d\in K} H_d(t) = \max \wedge \sum_{d\in I_n^{new}} H_d(t) - \sum_{d\in I_n^{old}} H_d(t) \le \varepsilon_3 \qquad (9)$$

Step-6 下記の場合に，フィードバック処理を終了する．それ以外の場合は，Step-2 を繰り返す．

(a) 追加単語のリスト l_+が無くなる．

(b) 式(7)の R_n が閾値 ε を超える．

(c) 条件式(8), (9)を満たす特徴単語 t が無くなる．

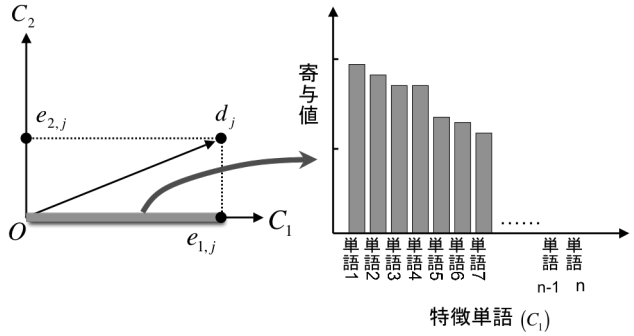

図 3 特徴単語の寄与値

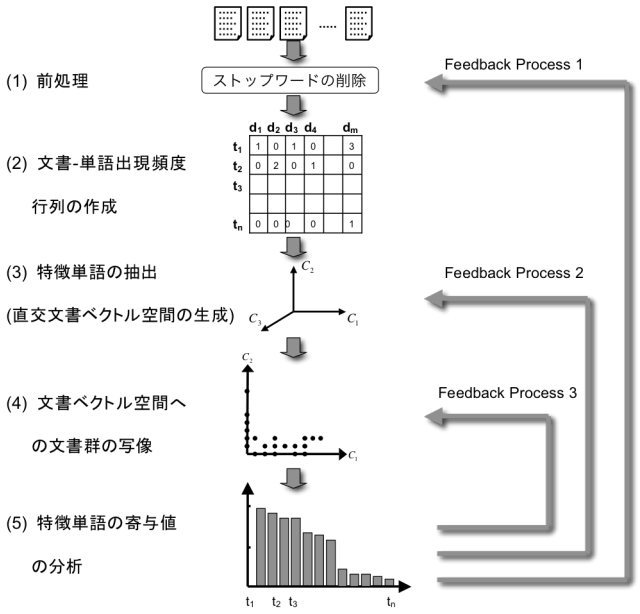

図 4 特徴単語のフィードバック

5. 実験

5.1. 実験環境

実験では,NTCIR-1[11]を使用して検索システムを実現する．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，英語のみを対象としたEコレクションを文書データとして使用した．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”が設定されている．実験では，E コレクションの文書群から論文のタイトルと抄録 443 件を，文書データとして抽出した．検索課題については，A 判定が 10 文書以上設定されているものを任意に選んでいる．また，各検索課題について，A，B，および C 判定である文書は，それぞれ，10 件，20 件，70 件を上限の文書件数にして，検索課題番号順に抽出を行った．実験に用いた文書データの詳細を表 2 に示す．

本実験では，表 3 に示すサンプル文書 gakkai-0000056563, gakkai-0000143277, および gakkai-0000179708 を用いて，3 次元の文書ベクトル空間の生成を行った．各サンプル文書より形成される座標軸 ID を，それぞれ C_1，C_2，および C_3とする． なお，前処理として特徴単語に対してステミング処理を行っている．

表 2 実験用文書データの詳細

課題番号	トピック	判定毎の文書件数			合計
		A	B	C	
4.	文書画像理解	10	0	70	80
10.	キーワード自動抽出	10	2	65	77
12.	マイニング手法	10	19	59	88
57.	創造的思考のモデル化及び支援	10	0	60	70
61.	階層関係の自動抽出	10	1	48	59
83.	骨形成の分子メカニズム	10	1	58	69
合計		60	23	360	443

表 3 文書ベクトル空間生成用のサンプル文書

軸 ID	サンプル文書	
	文書番号	課題番号 (判定)
C_1	gakkai-0000056563	4. (A 判定)
C_2	gakkai-0000143277	12. (A 判定)
C_3	gakkai-0000179708	83. (A 判定)

5.2. 実験 1

5.2.1. 目的

提案アルゴリズムによるフィードバック処理の性能を確認する．

5.2.2. 実験方法

軸 C_1 を形成するサンプル文書について，検索課題 4.の A 判定文書 5 件を用いる．それぞれの文書を用いて構築した文書ベクトル空間に対して，フィードバック処理を 10 回行う.各回のフィードバック処理において，正解文書と不正解文書の座標値の総和平均の比率を算出し，考察を行う．

正解文書群には，検索課題 4 の A 判定文書 10 件を用いる．それ以外の文書群を不正解文書群とする．

5.2.3. 実験結果

実験結果を図 5ー図 7 に示す.

図 5 は，検索課題 4.の A 判定文書をサンプル文書としてそれぞれ用いて，文書ベクトル空間を生成した場合の,正解文書 10 件の座標値の総和平均と不正解文書の上位 10 件の座標値の総和平均の比率（以下，R_{10}とする）を，フィードバック回数毎に示したグラフである．同様に図 6 は，正解文書 10 件の座標値の総和平均と不正解文書の上位 300 件の座標値の総和平均の比率（以下，R_{300} とする）を，フィードバック回数毎に示したグラフである．

R_{300} では，わずかながら増加傾向にあることが確認できる．これは，提案アルゴリズムによるフィードバック処理によって，文書全体的に正解文書の座標値のが比率が増加する傾向にあることを示している.また，R_{10} も同様に，緩やかに増加する傾向にあるが，ある

フィードバック回をピークに減少傾向にある.例えば，文書 043195 では，3 回目のフィードバックがピークであり，文書 056563 では，5 回目のフィードバックがピークとなっている．

これについて，図 7 文書 056563 における正解文書と不正解文書における座標値の総和平均の比較から確認すると，フィードバック処理の回数を重ねる毎に，正解文書と不正解文書における座標値の総和平均の差が小さくなる傾向にあることがわかる．これは，提案アルゴリズムによるフィードバック処理においては，フィードバック処理の回数を重ねる毎に，文書群全体の上位から下位の文書について削除可能な特徴単語を抽出するため，フィードバック処理の回数が増えるにつれて,上位の不正解文書に対して，正解文書と同様のフィードバック効果が反映されてしまうことが原因である．

このため，フィードバック処理を適切な文書分布の状態で停止するには，フィードバック終了のための適切な閾値を設定するか，数回のフィードバック処理のうち座標値の総和平均の比率が最大となった場合の文書分布を実現する文書ベクトル空間を選択するのが有効な方式として考えられる．

5.3. 実験 2

5.3.1. 目的

提案アルゴリズムによるフィードバック処理により，文書ベクトル空間の文書分布を改善できることを確認する．

5.3.2. 実験方法

軸 C_1 上に射影された文書群においてフィードバック処理を 5 回行い，検索課題 4.の A 判定文書 10 件を対象として，処理の前後における座標値を比較する．フィードバック後の座標値が，フィードバック前よりも大きくなる傾向にあることを確認する．

5.3.3. 実験結果

実験結果を表 4，表 5 および図 8，図 9 に示す．表 4 は,フィードバック処理前後における軸 C_1 上の文書分布の比較を示している．文書番号の下線は，検索課題 4.の A 判定文書であることを示しており，網掛けの文書は軸 C_1 を形成しているサンプル文書であることを示している．表 4 より，フィードバック処理前では，座標値の上位 10 件中に 5 つの A 判定文書が分布しているが，フィードバック処理後では上位 10 件中に 6 つの A 判定文書が分布しており，座標値も大きくなっていることが確認できる．

表 5 は，各回のフィードバック処理によりの追加・削除された単語群を示している．検索課題 4.は，文書画像理解をトピックとしており，domain や structur など，この分野に関連する単語が追加されていることが分かる．また，削除単語についてもこの分野に関連が少ない単語が主に削除されているが，adapt や readabl など関連性があると判断される単語も含まれている．

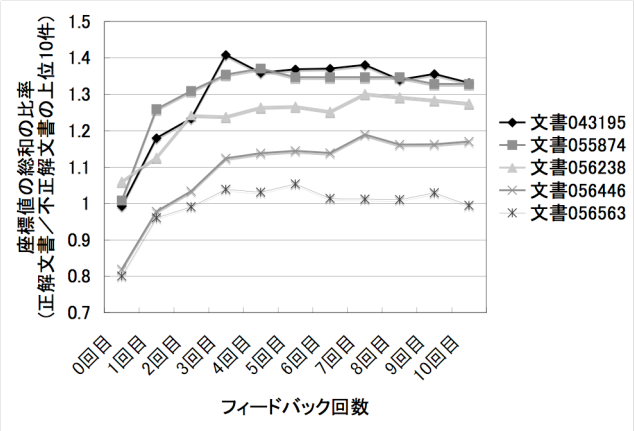

図 5 座標値の総和の比率 (正解文書／不正解文書の上位 10 件)

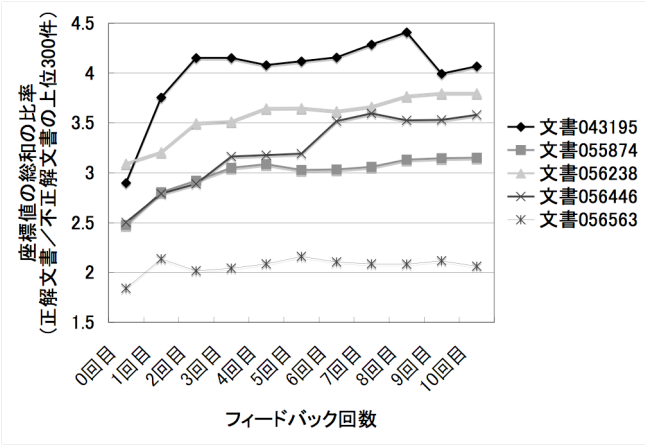

図 6 座標値の総和の比率 (正解文書／不正解文書の上位 300 件)

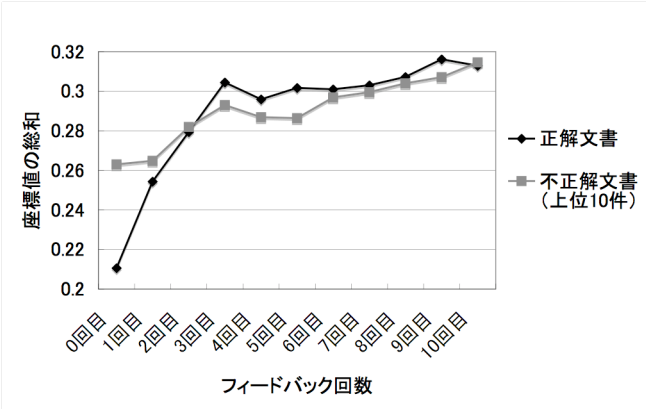

図 7 文書 056563 における正解文書と不正解文書における座標値の総和の比較

図 8 および図 9 は，フィードバック処理前後の，軸 C_1 上の文書 054566 の特徴単語の寄与値を示している．フィードバック処理後は，特徴単語 structur の寄与値が大きく，フィードバック処理により，文書の内容に適切な特徴単語が座標値の算出に寄与していることが分かる．

以上の実験結果は，提案アルゴリズムによるフィードバック処理により，文書ベクトル空間の文書分布を改善できることを示している．

表 4 軸 C_1 上の文書分布の比較

	フィードバック前		フィードバック後	
順位	文書番号 (下 6 桁)	座標 (相関値)	文書番号 (下 6 桁)	座標 (相関値)
1	056563	0.8815	056563	0.9767
2	056445	0.3526	055955	0.4635
3	266769	0.3379	056445	0.4635
4	010315	0.3085	054566	0.3476
5	055955	0.2938	056459	0.3476
6	043313	0.2938	055874	0.3145
7	020258	0.2791	010612	0.3145
8	055874	0.2644	009028	0.2979
9	056459	0.2497	266769	0.2979
10	009028	0.2497	052462	0.2979

表 5 フィードバック処理により追加･削除された特徴単語群

フィードバックの回数	削除された単語	追加された単語
1 回目	research, rule	domain
2 回目	depend, part	propos
3 回目	adapt, high, necessari	structur
4 回目	propos	
5 回目	describ, readabl	

5.4. 実験 3

5.4.1. 目的

実験 2 でフィードバック処理により再構成を行った文書ベクトル空間を用いた検索システムにおいて検索精度が向上していることを確認する．

5.4.2. 実験方法

問い合わせとして，rectangle, understand の単語列を用いて検索を行う．実験 2 で行ったフィードバック処理において，フィードバック前とフィードバック後で検索精度を比較する．検索課題 4.の A 判定文書 10 件を正解文書として，フィードバック処理後では，フィードバック処理前よりも検索精度が向上していることを確認する．

5.4.3. 実験結果

実験結果を表 6 および図 10 に示す．

表 6 は，フィードバック処理前後における検索結果の比較を示している．文書番号の下線は，問い合わせの正解文書であることを示し，網掛けの文書は軸 C_1 を形成するサンプル文書であることを示している．表 6 より，フィードバック処理前の検索結果では上位 10 件中に 2 件の正解文書のみ獲得されているが，フィードバック処理後では，上位 10 件中に 5 件の正解文書が獲得されていることが確認できる．また，図 10 は，フィードバック処理前後における検索精度の比較を示しており，フィードバック処理後では，フィードバック処理前より検索精度が向上していることが分かる．

以上の実験結果により，フィードバック処理により再構成を行った文書ベクトル空間を用いた検索システムにおいて検索精度が向上していることが確認できる．

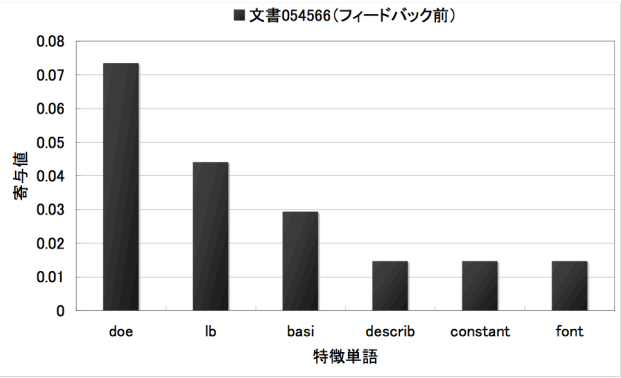

図 8 軸 C_1 上の文書 054566 の特徴単語の寄与値（フィードバック処理前）

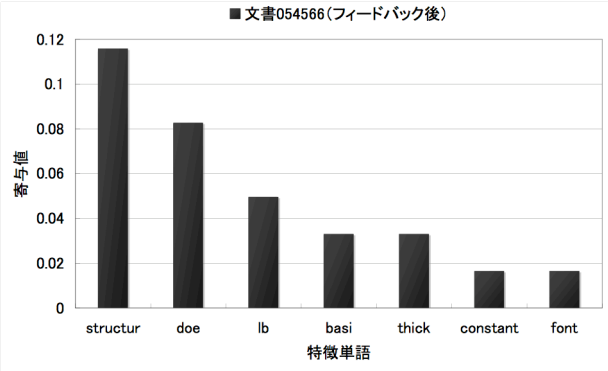

図 9 軸 C_1 上の文書 054566 の特徴単語の寄与値（フィードバック処理後）

6. まとめ

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，生成した文書ベクトル空間上に射影された文書ベクトルの形成に必要な特徴単語成分を自動的に抽出するための基本的アルゴリズムを示し，文書ベクトル空間を構成する特徴単語の追加，削除を行うことにより文書ベクトル空間を再構成する方法を提案した．FEM 方式による文書検索システムでは，人間により意味や内容に基づいて分類された

文書群より特徴単語を抽出し，利用者の検索意図や目

表 6 検索結果の比較(フィードバック処理前後)

	フィードバック前		フィードバック後	
順位	文書番号 (下 6 桁)	座標 (相関値)	文書番号 (下 6 桁)	座標 (相関値)
1	143277	1.0000	143277	1.0000
2	179708	0.9473	056563	0.9767
3	056563	0.8815	055955	0.5101
4	266769	0.4845	056445	0.4652
5	218912	0.3879	266769	0.4568
6	143957	0.3812	009028	0.3996
7	179466	0.3748	218912	0.3893
8	009028	0.3655	054566	0.3636
9	055955	0.3637	234836	0.3599
10	234836	0.3632	056459	0.3539

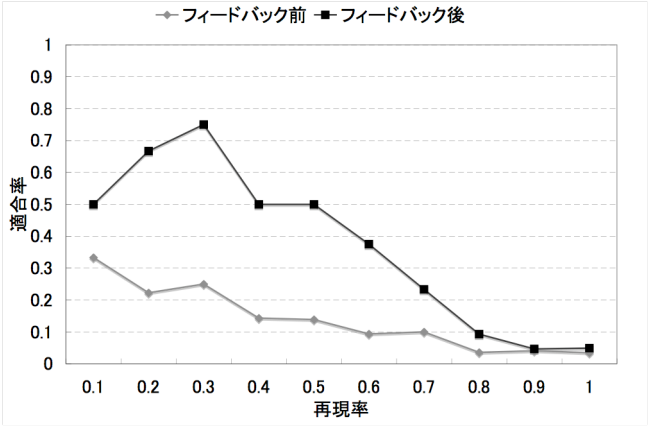

図 10 検索精度の比較(フィードバック処理前後)

的を反映した文書ベクトル空間を生成する．

提案方式では，文書の意味や内容に基づいて分類された文書群より抽出された特徴単語群において，検索意図に沿った文書群において追加した方が良い特徴単語や，検索意図に合致しない文書群において削除した方が良い特徴単語を，検索処理における各文書の相関量の算出結果に基づいて抽出する基本的アルゴリズムを示した．提案アルゴリズムにより抽出された特徴単語群を追加・削除するフィードバック操作を行うことにより，文書ベクトル空間の再構成を行い，FEM 方式により生成された文書ベクトル空間を用いた検索システムの検索精度の向上が実現可能となる．

本研究では，NTCIR より提供されている学術論文の要旨に関する文書群を対象とした実験により，下記の項目を確認した．

（１）提案アルゴリズムによるフィードバック処理の性能を確認を行った．（実験 1）

（２）提案アルゴリズムによるフィードバック処理により，文書ベクトル空間の文書分布を改善できることを確認した．（実験 2）

（３）フィードバック処理により再構成を行った文書ベクトル空間を用いた検索システムにおいて検索精度が向上していることを確認した．（実験 3）

今後の課題として，より多くのサンプル文書を用いて生成した検索空間を用いて，多くの検索文書を対象とした大規模な検索実験を行い，提案方式の定量的評価およびスケーラビリティを検証していく予定である．

文　献

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

[2] 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.

[3] 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, 2005.

[4] 張 建偉，黒川沙弓，石川佳治，北川博之，``フィードバックを利用した情報源の選択に基づくレコード抽出手法," 情報処理学会研究報告 DBS-140(II), pp.291-298, 2006.

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

[6] 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.

[7] 福村壽晃，木寺悠介，鷹野孝典，陳 幸生，``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," 情報処理学会研究報告 Vol. 2006, No. 77, 2006-DBS-140(II). p.299-304, 2006.

[8] 木寺悠介，陳 幸生，塩原慶一，``検索質問にあわせた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), 2006.

[9] Lewis, D.D., Schapire, R.E., Callan, J.P., and Papka, R., ``Training algorithms for linear text classifiers," SIGIR, pp.298-315, 1996.

[10] 増田圭祐，鷹野孝典，陳 幸生，``利用者の検索意図や目的を反映した文書ベクトル空間の動的形成による検索精度向上についての評価，" データベースと Web 情報システムに関するシンポジウム（DBWeb2006）論文集，Vol.2006, No.16，pp.143-152，2006

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

[12] 大橋英博，清木 康，``意味的連想検索方式における意味表現ベクトルを対象とした学習機構の実現," 情報処理学会研究報告, 2004-DBS-134(II)，pp. 499-504, 2004.

[13] 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.

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

[15] 鷹野孝典，図子泰三，清木 康，``事象間の因果関係を扱う動的な文脈解釈機能を有する意味的連想検索方式の実現," 情報処理学会論文誌：データベース，Vol.46, No. SIG 5(TOD25), pp.40-55, 2005.

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