頻度分布に基づくプロジェクションを用いた文書検索

大内　浩仁† 　三浦　孝夫† 　塩谷　勇††

† 法政大学 工学研究科 電気工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
†† 産能大学 経営情報学部　〒 259–1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{i03r3208,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし　ランダムプロジェクション（RP）による文書検索では，ランダムな射影行列を作成することで高速かつデータに依存しない検索を行うことができる．しかし，そのランダム性により，特に低次元で検索の安定性が低下する．本研究では，単語の頻度分布に基づいて射影行列を構成するプロジェクション手法として Skewed Projection（SP）を提案する．この手法を用いることにより，誤差を保存しつつ，分布に特有な応用分野に属する文書集合に対して，局所的に非依存かつ効率的な文書検索が行えることを示す．

キーワード　情報検索，ベクトル空間モデル，頻度分布に基づくプロジェクション，ランダムプロジェクション，次元縮小

Document Retrieval by Projection Based Frequency Distribution

Hirohito OHUCHI†, Takao MIURA†, and Isamu SHIOYA††

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University　1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{i03r3208,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract　In document retrieval task, random projection (RP) is a useful technique of dimension reduction. It can be obtained very quickly yet the recalculation is not necessary to any changes. However, in lower dimension, random projection has instability by randomness in itself. In this investigation, we propose a new technique, called *skewed projection* (SP) for dimension reduction technique based term frequency distribution. We show that this technique can take advantages of local independency thus we can obtain efficient retrieval for documents which belong to specific application area.

Key words　Information Retrieval, Vector Space Model, Skewed Projection, Random Projection, Dimension Reduction

1. 前　書　き

近年，計算機環境における文書データの種類と量はますます多様化している．それに伴い，適切な文書を効率よく検索する技術の重要性は増大している．

一般的に，文書データのモデル化はベクトル空間モデル [9] を用いて各文書をベクトルに置き換えることで行う．語彙の数がベクトルの次元数となるため，各文書ベクトルは数万から数十万の高次元で疎（sparse）なベクトルになる．高次元データをそのまま扱うと，検索の効率を損なうと同時に計算機容量を圧迫する．このため，プロジェクション手法を用いて文書ベクトルの要素を保持したまま低次元のベクトル空間へデータを射影することが必要である．

代表的なプロジェクション手法として，Latent Semantic Indexing（LSI）が存在し，数多く論じられている [5], [11]．LSI 手法は得意値分解（SVD）に基づいて構成されており，検索精度を維持したまま大幅に次元を縮小することが可能である．このため，検索効率と検索精度を両立した検索質問を行うことができる．しかし，SVD は文書集合に依存するため，データの更新に対して SVD の再計算が必要となる．SVD の計算には多くの計算量を必要とするため，動的に新しい文書データが出現するような環境に対応する事が困難である．近似的に SVD の再計算を行う方法が提案されている [2] が，更新に従って検索精度が低下するため，動的な環境下では LSI 手法の適用は難しい．

これに対し，近年注目されているプロジェクション手法にランダムプロジェクション（RP）がある [8]．RP 手法ではラン

ダムな要素で射影行列を構成する．LSI 手法に対してはるかに行列の作成が容易であり，かつプロジェクションがデータに対して独立である．データに独立なプロジェクションは射影行列の再計算が不要であることを意味し，検索効率と検索精度の両立が可能となる．しかし，そのランダム性ゆえに，特に低次元でプロジェクションの安定性が低下する問題がある [3]．

本研究では，文書データの単語分布を元に射影行列を構成するプロジェクション手法（Skewed Projection:SP）を提案する．文書データの傾向が変わればその単語分布は変化する．例えば，計算機科学に関する論文とスポーツのニュース記事では，よく現れる単語は自ずと異なってくる．逆に言えば，同一の応用分野に属する文書は似たような単語分布を持つ．そのため SP 手法は “局所的な非依存性” を保持する．この非依存性が維持されている限りは，汎用的なプロジェクションである RP 手法に対して，低次元における安定性と検索効率において RP 手法を上回ることが期待できる．また，対象とする応用分野が同じである限り，射影行列の再計算が不要になる．

単語の頻度を考慮した研究では，各文書から低頻度の単語を無視して文書検索を行っている [12]．次元縮小にかかる時間と検索精度を総合的に判断した結果，優れた検索効率を実現することが述べられている．しかし，各文書の単語頻度をその文書のみに適用するだけで，局所的な非依存性という概念は無い．

2 章では RP 手法と SP 手法について述べる．3 章で両手法の理論的考察と，文書検索における両手法の比較を行う．4 章で実験結果を示し，5 章で結びとする．

2. 文書検索における次元縮小

ベクトル空間モデルにおけるプロジェクション手法としての RP 手法および SP 手法について述べる．

2.1 ベクトル空間モデル

ベクトル空間モデルでは，文書集合をデータ行列で表現する．各文書はデータ行列の列ベクトルとして構成する．単語数 d ，文書数 N の文書集合は大きさ $d \times N$ のデータ行列となる．データ行列 X の i 行 j 列の要素 x_{ij} は，文書 j における単語 i の重みである．重みの与え方としては，一般的に単語頻度（TF），もしくは単語頻度に文書頻度の逆数をかけた値（TF*IDF）が用いられる．

検索を行うための検索質問は，ベクトル $\mathbf{q}^{d\times 1}$ で表現される．質問ベクトルと文書ベクトルの類似度を測定し，文書の類似度を降順にソートすることで，検索結果をランキングとして表示する．類似度は，質問ベクトルと文書ベクトルの余弦（cos）で定義する．文書集合の中から i 番目の文書を調べる場合，

$$\cos\theta_i = \frac{(\mathbf{q}, \mathbf{x}_i)}{|\mathbf{q}||\mathbf{x}_i|}$$

の値によって，検索質問に対する文書の類似度を求める．$\mathbf{x}_i$ は，X の i 番目の列ベクトルを意味する．類似度は 1 から-1 の値を取り，1 に近いほど質問と適合している．

2.2 ランダムプロジェクション手法

RP 手法による文書データの次元縮小について述べる．以下では，大きさ $d\times N$ のデータ行列 X を大きさ $k\times N(k << d)$ のデータ行列 X_RP に射影する．このための射影行列として，要素をランダムに並べた大きさ $k\times d$ の RP 行列 R を決定する．データ行列 X の RP 手法による次元縮小は，次の計算で行う．

$$X_{RP}^{k\times N} = R^{k\times d} X^{d\times N} \tag{1}$$

この処理の計算量は $O(dkN)$ [11] である．すなわち，次元数を縮小するほど計算時間は短縮される．

この RP 行列の要素は，すべて文書集合に独立して決定される．このため，文書集合が更新されても射影行列を変更する必要はない．ただし，RP 行列 R の要素を構成する際は，以下の条件を満たす必要がある．[3], [8]

- 各要素が平均 0，分散 1 の独立正規分布に従う
- 各列ベクトルの長さが 1（単位ベクトルと等しい）
- R が直交行列

このうち特に，行列の直交化は大きな計算量を必要とする．そこで，これらの条件を近似的に満たすような要素の生成方式が提案されている [1]．RP 行列 R の i 行 j 列における 要素 r_{ij} を次の確率分布に従うように選ぶ．

$$r_{ij} = \sqrt{3}\cdot \begin{cases} +1 & \text{確率 } 1/6 \\ 0 & \text{確率 } 2/3 \\ -1 & \text{確率 } 1/6 \end{cases} \tag{2}$$

この分布に従う行列を作成するための計算量は $O(kd)$ であり，更に $k << d$ であることから，実際の処理時間は非常に少ない．

質問検索を行う際は，文書と同様に，質問ベクトルを同一の R で射影する．

$$\mathbf{q}_{RP}^{k\times 1} = R\mathbf{q}^{d\times 1} \tag{3}$$

文書集合 X_{RP} の各列ベクトルとの類似度を計算し，検索結果としてランキングを表示する．

2.3 頻度分布に基づくプロジェクション手法

本稿では，単語の頻度分布（以下，頻度分布）に基づくプロジェクション手法を提案する．この手法では単語の分布にもとづいて確率密度関数を作成し，これを用いて射影行列の要素を構成する．以下では，この手法を RP 手法に対比して Skewed Projection（SP）手法と呼ぶ．

SP 手法における射影行列を構成するために，対象となる文書集合と同一分野のサンプル文書集合を用意し，頻度分布を得る．以下では単語数 d 文書数 M のサンプル文書集合を仮定する．まずそれぞれの単語が出現している文書数を調べる．単語 i の出現文書数を f_i とおくと，$0 \leqq f_i \leqq M$ となる．この f_i より，サンプル文書集合に基づく単語 j の確率分布を

$$Pr(f_j) = \frac{f_j}{\sum_{i=1}^{d} f_i} \tag{4}$$

で決定する．

この確率分布に従って SP 行列を構成する．SP 行列 S の大きさは RP 行列と同じく $k\times d$ である．まず SP 行列の第 1 行目に関して，式（4）に基づく確率分布に従って単語を 1 つ選ぶ．k 番目の単語が選ばれた場合，1 行 k 列の要素 s_{1k} に 1 を

加える．これを d 回試行し，SP 行列の全ての行について同じ処理を行う．この結果，S の各行ベクトルは，サンプル文書集合の単語頻度分布から生成される疑似的な文書ベクトルとして表される．最後に S を正規直交化し，SP 行列の要素を決定する．

実際の次元縮小は RP 手法と同じく，データ行列および質問ベクトルに対して SP 行列 S を左側から乗じることで行う．

$$X_{SP}^{k\times N} = S^{k\times d}X^{d\times N} \tag{5}$$

$$\mathbf{q}_{SP}^{k\times 1} = S\mathbf{q}^{d\times 1} \tag{6}$$

文書ベクトルとの類似度を計算し，検索結果としてランキングを決定する．

3. RP 手法および SP 手法の誤差保証

RP 手法と SP 手法における次元縮小に伴う誤差の保証と，文書検索における両プロジェクション手法の比較について述べる．

3.1 誤差保証と正規直交系

RP 手法の元となる考え方は，Johnson と Lindenstrauss の補題 [7] に端を発し，今日では次の定義で表される．

> d 次元ユークリッド空間上の M 個の点集合は，$k \leqq O(logM/\epsilon^2)$ で得られる k 次元ユークリッド空間上に写像することができる．その際，点集合における任意の 2 点間の距離は誤差 $(1 \pm \epsilon)$ で保存される [4].

RP 手法によるプロジェクションがデータに依存しないのは，ベクトル間の相対的な距離関係を保証しているためである．

文書検索における RP 手法の誤差は，ベクトル間のユークリッド距離に対して定義される．$d \times N$ 行列 X から任意の 2 つの列ベクトルを取り出し，$\mathbf{x}_1$ および $\mathbf{x}_2$ と置く．$\mathbf{x}_1$ と $\mathbf{x}_2$ の d 次元におけるユークリッド距離を $|\mathbf{x}_1 - \mathbf{x}_2|$ で表す．RP 行列 R により k 次元に縮小された空間における $\mathbf{x}_1$ と $\mathbf{x}_2$ のユークリッド距離は以下の式で近似することができる [3].

$$\sqrt{d/k}|R\mathbf{x}_1 - R\mathbf{x}_2| \tag{7}$$

式（7）が成り立つためには，R が直交行列である必要がある．ただし，十分な高次元空間でランダムな方向を有するベクトル集合は，近似的に直交性を満たすことが知られている [6].

R の直交性に対する誤差を表す $d \times d$ 行列 ϵ を次の式で定義する．R^TR が 単位行列に近似するほど，R は直交行列に近くなる．

$$\epsilon = R^TR - I \tag{8}$$

このとき，ϵ の要素は，平均 0，分散 $1/k$ の正規分布をとる [8]. 従って縮小次元数 k を大きくするほど，ユークリッド距離における誤差は減少する．

このため，射影行列を用いた次元縮小には，射影行列の各ベクトルが正規直交系であることが求められる．近似的に直交性を満たすためには偏りのないベクトル集合であることが求められるため，SP 行列はこれを満たさない．そのため，SP 行列に対して正規直交化を行う必要がある．その結果得られた射影行列は正規直交系であり，RP 手法と同様の誤差保証を得る．しかし要素の分布に偏りが生じているため，RP 手法と比較して式（7）による近似がされにくくなる．

3.2 文書検索における RP 手法と SP 手法

文書検索のタスクで RP 手法と SP 手法がどのように次元縮小を行うかについて述べる [10]. 射影行列を用いた d 次元から k 次元への次元縮小は，d 個の単語から k 個の単語を再構成する作業である．式（2）の分布に基づいて RP 行列を構成した場合，RP 手法の次元縮小は次の手順により単語の再構成を行う [1].

（1） 元の単語から，ランダムに 2/3 を破棄する．

（2） 残りの単語を 2 つのグループにに分割する．各グループ内の単語数は（確率的に）等しい．

（3） それぞれのグループの単語頻度に重み（$+\sqrt{3}, -\sqrt{3}$）を掛けた値の和を新しい単語の頻度として決定する．

RP 手法では，単語の重みをランダムに決定している．即ち，確率的に出現頻度の低い単語に重みを与えてしまうことがある．このため，単語の偏りを考慮せず，汎用的なプロジェクション手法としてデータに依存しない次元縮小が可能である．

SP 手法では，射影行列の要素が偏った分布により構成されている．サンプル文書集合で出現確率 $Pr(f_j)$ が高い単語ほど，より大きな重みを持つ確率が高い．また，出現頻度の低い単語にはほとんど重みを与えることがない．このため，検索する文書集合がサンプル文書集合と同じ頻度分布を持つ（同一の応用分野に属する）場合，局所的に RP 手法よりも良く文書ベクトル間の距離を保つといえる．SP 手法の重要な特性である，同一の頻度分布内でのデータに依存しない次元縮小を可能としている．逆に言えば，サンプル文書集合と異なる単語分布をもつ文書を検索する場合は，RP 手法よりも検索精度が低下することになる．

4. 実　　験

まず実験環境として検索を行う文書データの詳細と，検索質問に対する答えの評価方法について述べる．次に RP 手法，SP 手法，異なる頻度分布に基づく SP 手法の 3 種類のプロジェクション手法について実験を行い，それらの実験結果について考察する．

4.1 実 験 環 境

文書データとして，NTCIR-1(注1) を使用する．NTCIR-1 は「学会発表データベース」から抽出した，学会発表論文の要旨を集めたテスト・コレクションである．日本国内の 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妙録を収録している．日本語と英語の両方を含む JE コレクション，日本語のみの J コレクション，英語のみの E コレクションの 3 つのコレクションがある．

本実験では，E コレクションの中から土木学会に属する文

(注1)：http://research.nii.ac.jp/ntcir/

書集合（以下，土木文書）と計測自動制御学会に属する文書集合（以下，計測文書）を抜粋して使用する．E コレクション 187,080 件中，土木文書は 12972 件，計測文書は 10781 件存在する．それぞれの文書について，妙録部分を索引語として，不要語（stop word）の削除および単語のステミング [9] を行う．結果として，土木文書では 27155 語，計測文書では 22491 語の索引語を得る．本実験では，Zipf の法則 [13] に基づき，土木文書から 1004 語，計測文書からは 1030 語の索引語を抽出している [10] (注2)．なお，土木文書と計測文書の双方に共通する索引語数は 623 語である．

RP 行列および SP 行列の生成には，高品質の乱数を高速に生成する Mersenne Twister (注3)を用いる．

4.2 異なる文書集合に対する単語分布の相違

SP 手法を適用する際は，それぞれの文書集合を 2 つに分ける．偶数番目の文書をサンプル文書集合，奇数番目の文書を検索文書集合とする．以下の表 1 の通りに頻度分布を作成する．

文書集合	索引語	同一の頻度分布	異なる頻度分布
土木検索文書	土木文書	土木サンプル文書	計測サンプル文書
計測検索文書	計測文書	計測サンプル文書	土木サンプル文書

表 1　文書集合および頻度分布の構成

頻度分布の違いを測るため，文書集合に対して χ^2 検定を行い，2 種類の文書集合が独立であるといえるか，そうでないかを判定する．まず土木サンプル文書集合と土木検索文書集合の頻度分布を下の図 1 に示す．

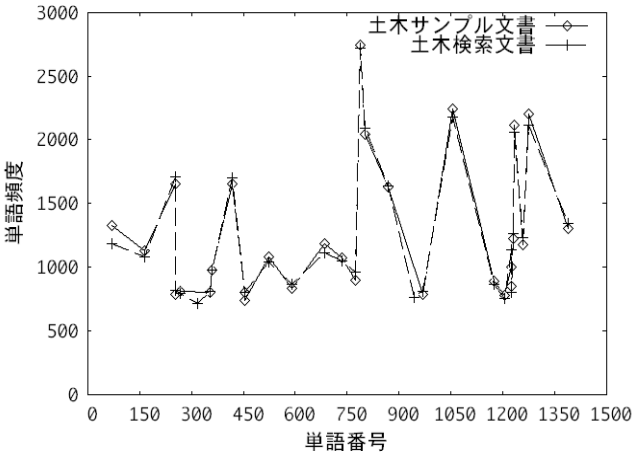

図 1　土木文書の頻度分布

これらの文書集合に対して自由度 1410 での χ^2 検定を行った結果，X^2 の値は 1512.614 となり，有意水準 2%で独立ではない．

次に，計測サンプル文書集合と計測検索文書集合について同様の検定を行う．頻度分布を図 2 に示す．

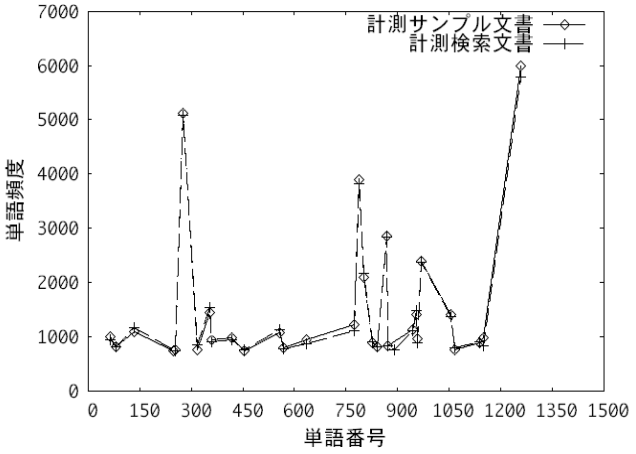

図 2　計測文書の頻度分布

χ^2 検定の結果，X^2 の値は 1663.129 となる．自由度 1410，有意水準 2%未満で独立ではない．

最後に，異なる分布の場合として，土木サンプル文書と計測検索文書について検定を行う．頻度分布を図 3 に示す．

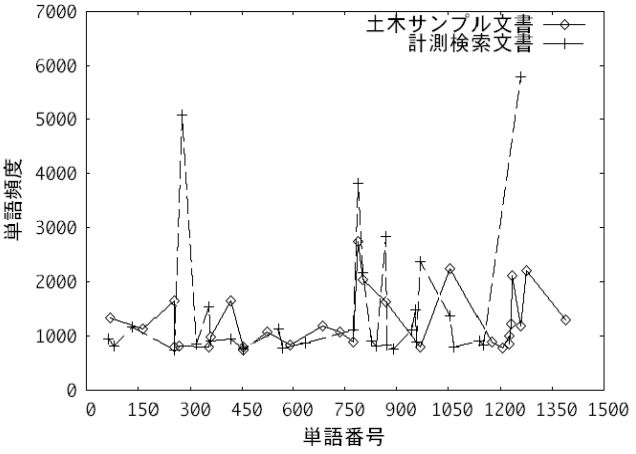

図 3　土木サンプル文書と計測検索文書の頻度分布

(注2)：Zipf の法則には，高頻度の単語で成り立つ Zipf の第 1 法則と，低頻度の単語で成り立つ Zipf の第 2 法則がある．低頻度の単語をどの程度削除するかの基準として，まず「中程度の頻度」を決める必要がある．頻度 1 の単語数を F_1 とすると，2 つの法則を同時に満たす中程度の単語頻度 f_k は，以下の式で求められる．

$$f_k = \frac{\sqrt{8F_1 + 1} - 1}{2} \tag{9}$$

ここで得られた出現頻度 f_k が索引語の頻度順位において中間地点であることを仮定すれば，以下の手順で索引語数を決定できる．

（1） 出現頻度 f_k を持つすべての語を索引語とする

（2） 第 1 順位から $f_k - 1$ 個の頻度を持つ語までのすべてを索引語とする．全部で K 個の語があるとする

（3） $f_k + 1$ 以下の出現頻度の語のうち，上位 K 個を索引語とする

本実験では，土木文書に対して $F_1 = 16060, f_k = 178$ ，計測文書に対して $F_1 = 12467, f_k = 157$ を得る．

(注3)：http://www.math.keio.ac.jp/ matumoto/mt.html

図 3 から頻度分布の違いを見て取ることが出来る．χ^2 検定の結果も X^2 の値が 100620.8 であり，自由度 1410，有意水準 2%で独立である．

以上より，土木文書と計測文書がお互いに異なる頻度分布を持っていることが言える．

4.3 評価方法

検索精度の評価には，11 点平均適合率を用いる．11 点平均適合率とは，0.0 から 0.1 刻みで 1.0 までの再現率における適合率の平均値である．

再現率は，検索漏れの少なさを示す尺度であり，

$$\text{再現率} = \frac{\text{検索された文書中の適合文書の数}}{\text{全文書中の適合文書の数}}$$

で表される．適合率は，検索ノイズの少なさを示す尺度であり，

$$\text{適合率} = \frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}}$$

で表される．再現率と適合率はトレード・オフの関係にある．理想的な情報検索システムでは再現率と適合率が共に 1 となる．しかし，実際には検索漏れを無くそうとすれば不適合文書が混じり，適合文書だけを取り出そうとすれば検索漏れが発生する．

適合文書として，次元縮小を行わない状態で質問検索を行い，その結果類似度 0.4 以上となった文書を選ぶ．これにより，次元縮小による検索精度への影響を調べることができる．縮小次元数の推移による 11 点平均適合率の変化を評価の指標とする．

4.4 RP 手法と SP 手法の比較

土木文書と計測文書，2 つの文書集合を用いて RP 手法と SP 手法の検索精度，検索の安定性および検索効率について実験を行う．本実験では，同一の頻度分布（同分布）に基づく SP 手法による検索，RP 手法による検索，および異なる頻度分布（異分布）に基づく SP 手法による検索の 3 種類を適用する．縮小次元は 10 次元から 10 刻みで 300 次元までとし，それぞれの次元軸について各 5 回ずつ次元縮小および検索質問を行う．1 回ごとに射影行列を作り替えて次元縮小を行い，11 点平均適合率の平均値，誤差，および同一精度で削減可能な次元数を計測する．全体の検索回数は $2 \times 3 \times 30 \times 5 = 900$ 回となる．

まず，各次元における 11 点平均適合率の平均値を求める．次元軸ごとに検索精度がどの様に変化するかを見ることができる．土木文書に対する検索結果を図 4 に，計測文書に対する検索結果を図 5 に示す．

各次元軸で 5 回ずつ検索質問を行った際に，最高の精度と最低の精度の差がどの程度広がるかを計測する．これによりプロジェクションの安定性を測ることができる．土木文書に対する検索結果を図 6 に，計測文書に対する検索結果を図 7 に示す．

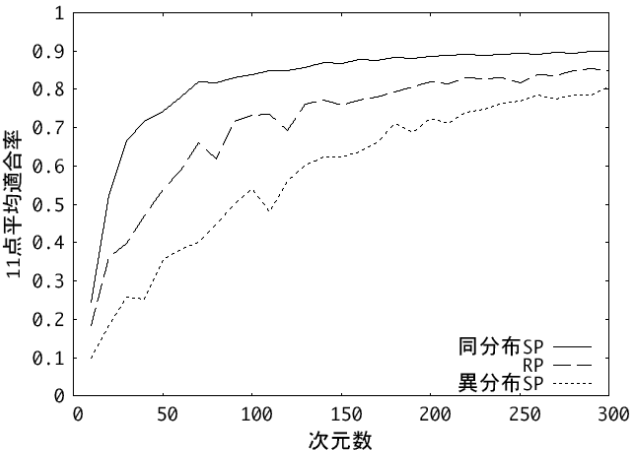

図 4　土木学会：検索精度（平均）

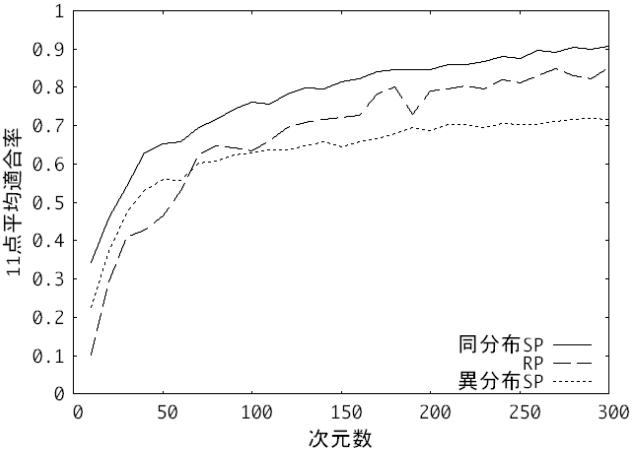

図 5　計測自動制御学会：検索精度（平均）

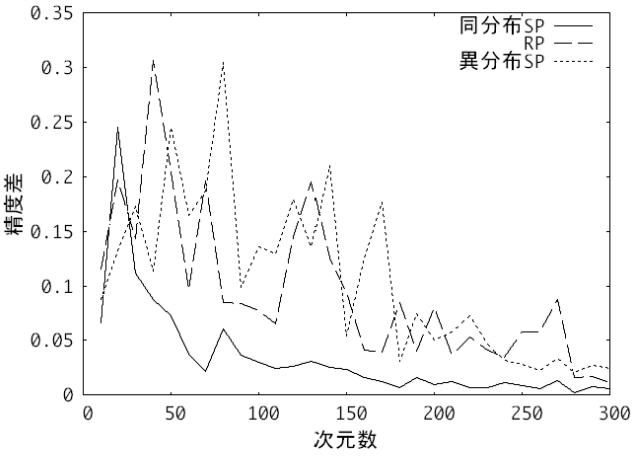

図 6　土木学会：検索精度（最高値 – 最低値）

11 点平均適合率を一定のラインで固定した場合，どの次元軸でそのラインを超えるかを計測する．より低い次元数で同じ精度に達することが出来れば，検索効率が向上していると言える．5 回の平均値を基準とする．土木文書に対する検索結果を表 2 に，計測文書に対する検索結果を表 3 に示す．

平均適合率	同分布 SP	RP	異分布 SP
0.5	20	50	100
0.6	30	70	130
0.7	40	90	180
0.8	70	190	300

表 2　土木学会：同一精度内での最小次元数

平均適合率	同分布 SP	RP	異分布 SP
0.5	30	60	40
0.6	40	70	70
0.7	80	130	210
0.8	150	180	——

表 3　計測自動制御学会：同一精度内での最小次元数

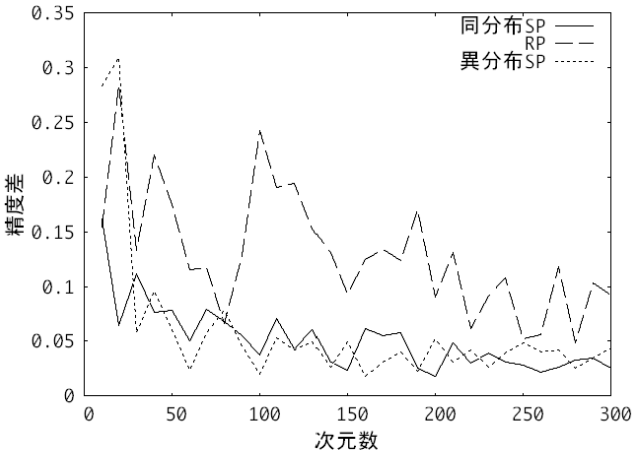

図 7　計測自動制御学会：検索精度（最高値 – 最低値）

4.5 考　　察

平均精度では，全ての次元で同分布 SP が最も良い．また，計測文書に対する検索で低次元の場合のみ 異分布 SP 手法 $>$ RP 手法 で，その他の場合では全て 同分布 SP 手法 $>$ RP 手法 $>$ 異分布 SP 手法 が成り立っている．これらの結果から，頻度分布を適切に考慮することが文書検索の精度向上について大きな役割を果たすと言える．また，的はずれな頻度分布を選んだ場合，ランダムな単語選択にも劣る結果を得ることになる．計測文書の検索では低次元において RP 手法が異分布 SP 手法を下回るのは，RP 手法の直交性に関する誤差が影響していると思われる．全体の傾向として次元が上がるほど精度も上昇しており，いずれのプロジェクション手法も次元縮小に伴う誤差保証が正しく行われていることを示している．

精度差については，土木文書，計測文書どちらの場合でも同分布 SP 手法が RP 手法より最小で推移している．この理由として，射影行列を正規直交化していることが大きい．RP 手法に対しては，安定性の面で確実に優位に立っていると言える．異分布 SP 手法の場合は，土木文書の場合は RP 手法とほぼ同じ，しかし計測文書では同分布 SP 手法とほぼ同じ誤差に抑えられている．このために低次元では RP 手法より良い精度を得られたと思われる．異分布 SP 手法においても正規直交化を行っているため，本来は常に RP 手法よりも精度差が低く抑えられるべきである．しかし，異分布 SP 手法では，頻度分布がどの様に異なるかによって，安定性が低下することもあれば，低下しないこともあると考えられる．土木文書に対する計測文書の頻度分布が前者で，計測文書に対する土木文書の頻度分布が後者である．この意味で，異分布 SP 手法による検索は不安定といえる．

同一精度内の最低次元数は，平均精度の場合と同じく全ての場合で同分布 SP 手法が最も良い結果を得ている．特に平均適合率 0.5 および 0.6 の到達次元数は，RP 手法の半分近くに抑

えられている．低次元における同分布 SP 手法の検索精度が際だっている．また，異分布 SP 手法では平均適合率 0.7 および 0.8 で到達次元数が大きく上昇している．これは，本来重要な単語に対して大きな重みを与えられないために，検索精度が他の手法より低い段階で頭打ちになってしまうためと考えられる．

全体として，文書検索においては同分布 SP 手法が検索精度および安定性の両面で RP 手法を上回る結果となる．より低い次元数で同じ精度を達成できるため，検索効率についても同様である．しかし，サンプル文書集合と異なる分野について検索を行うと精度，安定性，検索効率の全てで RP 手法を下回ることがありえる．実際には，SP 手法では射影行列の直交化が必要な分，検索効率は低下する．それでも総合的には同分布 SP 手法は RP 手法を上回ると言える．SP 手法は，文書の応用分野に関する背景知識を生かした文書検索を可能とする，きわめて有効なプロジェクション手法になりうる．

5. 結　　論

本研究では，特定の応用分野に属する文書集合に対して，単語の頻度分布を考慮したプロジェクション手法，SP 手法を提案した．汎用的なプロジェクション手法である RP 手法と比較を行い，サンプル文書と同じ頻度分布を持つ文書集合に対しては RP 手法より優れた検索が可能になることを示した．SP 手法は局所的な非依存性を維持するため，文書集合について局所的な非依存性が存在することを実証した．

今後は，同じ頻度分布を，“似て非なる” 応用分野に対してどの様に適用するかを論じる必要がある．

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 16500070) の支援をいただいた．

本実験に対しては国立情報学研究所より，NTCIR-1 テストコレクション 1，NTCIR-1 テストコレクション 1 付属言語タグつきデータコレクション，および NTCIR-2 テストコレクション 1 の提供をいただきました．関係各位に深く感謝します．

文　　献

[1] Achlioptas, D.: ”Database-friendly random projections”, In *Proc. ACM Symp. on the Principles of Database Systems*, pp. 274-281, 2001.

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

[3] Bingham, E. and Mannila, H.: “Random projection in dimentionality reduction: Applications to image and text data”, In *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pp. 245-250, 2001.

[4] Dasgupta, S. and Gupta, A.: “An elementary proof of the Johnson-Lindenstrauss Lemma”, Technical Report TR-99-006, International Computer Science Institute, 1999.

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

[6] Hecht-Nielsen, R.: “Context vectors: general purpose approximate meaning representations self-organized from raw data” In *Computational Intelligence: Imitating Life(Zurada et al. eds.)*, pp. 43-56, IEEE Press, 1994

[7] Johnson, H. and Lindenstrauss, J.: “Extensions of lipschitz mapping into a hilbert space”, In *Conference on Modern Analysis and Probability*, pp. 189-206, 1984.

[8] Kaski, S.: “Dimensionality reduction by random mapping: Fast Similarity Computation for Clustering”, In *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, Vol 1, pp. 413-418, 1998.

[9] 北 研二, 津田 和彦, 獅子堀 正幹: “情報検索アルゴリズム”, 共立出版, 2002.

[10] 大内 浩仁, 三浦 孝夫, 塩谷 勇: “ランダムプロジェクションを用いたニューススストリームの検索”, 日本データベース学会 Letters (DBSJ Letters) Vol.3, No.3, 2004

[11] 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.

[12] Schutze, H. and Silverstein, H.: “Projections for Efficient Document Clustering”, In *Proc. Special Interest Group on Information Retrieval(SIGIR)*, pp. 74-81, 1997.

[13] Zipf, G, K.: “The human behavior and the principle of least effort”, Addison Wesley, 1949.