プロジェクションを用いた文書検索に関する研究

大内, 浩仁 / OHUCHI, Hirohito

(発行年 / Year)
2005-03-24
(学位授与年月日 / Date of Granted)
2005-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2004年度　修士論文

プロジェクションを用いた文書検索に関する研究

STUDIES ON DOCUMENT RETRIEVAL
BY USING PROJECTION

指導教員　三浦 孝夫

法政大学大学院工学研究科
電気工学専攻修士課程

03R3208　大内 浩仁（おおうち ひろひと）
Hirohito OHUCHI

目 次

第1章　序論

1.1　問題の背景

情報化社会と呼ばれる今日，電子化された情報が身の回りに溢れている．電子化された情報を有効に活用するためには，必要とされる情報を効率的に見つけ出すための手段が必要である．この要求を満たす手段が，情報検索（Information Retrieval）と呼ばれる技術である．情報検索の研究の歴史は古く，数多くの重要な概念や技術が1940年代後半から1950年代にかけて生み出されている．現在では，計算機システムの高性能化とネットワーク化，これに伴う情報の電子化によって必要な文書を効率的に，かつ高い精度で検索することの重要性は増加している．情報検索は，現在の情報化社会を支えるきわめて重要な基幹技術となっている．

情報は文書，画像，映像，音楽などのメディアに分けることができる．この中では，文書（テキスト）データが最も中心的なメディア情報であると言える．ここで定義される文書データは，自然言語によって記述されたものであり，データベースシステムにおけるスキーマなどの構造を仮定しない．文書検索システムにおいては，検索対象となる文書集合に対して，同じく自然言語で記述された検索質問（query）を処理し，検索結果として，検索質問に適合する文書を出力する．代表的な文書検索システムの例として，サーチエンジンが挙げられる．

文書検索システムの構築においては，適合性の基準をどの様に与えるかが問題となる．その与え方により，情報検索システムは内容型検索と全文検索に分けられる．全文検索では，情報検索システムの目的は検索質問の文字列と一致する部分を持った文書を探し出すことである．対して内容型検索では，検索質問と意味的に類似した文書を見つけ出すことに主眼を置いている．具体的には，文書の内容を特徴づける単語（索引語）を抽出し，索引語の出現頻度などを用いて文書集合および検索質問を表現する．内容型検索のモデルとしては，ベクトル空間モデル，確率モデル，ネットワークモデルなどがある．本研究では，このうち最も代表的なベクトル空間モデルを扱う．ベクトル空間モデルでは，文書集合および検索質問を，索引語を次元とする多次元ベクトルによって表す．このため，適合性の判定をベクトルの類似度計算によって求めることができる．

本研究では，内容型検索に基づく文書検索システムを取り扱う．将来，アクセス可能な文書データ量の増加，および文書データが属する分野の多様化が進むと予想される．電子化された文書が急激に増加する状況においては，字面よりもその意味を重視した検索，即ち内容型検索が，将来的により大きな意義を持つと考えられる．また，文書

検索に特有の問題として，同義語と多義語の問題がある．全文検索では，単語のマッチングで検索を行うため，基本的にこれらの問題を解決することはできない．

しかし現在，汎用的な文書検索システムは，そのほとんどが全文検索に基づいている．その最も大きな理由は，適合性の評価が人間にとって適切であるかどうかを判断するのが難しいことである．全文検索では，検索質問の文字列と文書が一致するかどうかという，計算機システムとユーザの双方に対して明快な基準が存在する．一方，ベクトル空間モデルにおける適合度基準は，検索システムの内部でベクトル表現された環境下におけるもので，ユーザの判定基準と必ずしも一致しない．さらに，文書数の増加に従って索引語が増加するため，数万から数十万次元を持ち，その要素が疎（sparse）であるベクトルが生成されてしまう．これにより，計算機容量の圧迫，検索時間の増大，不要な索引語による検索精度の低下など，さまざまな悪影響が生じてくる．また，索引語の区別自体は字面により行っているため，そのままでは索引語間の多義性，同義性などの意味的な関係を考慮することができない．内容型検索においてユーザの希望を検索結果に反映させるためには，単純な質問文字列以外の情報を与え，付加機能として適切に文書検索システムで機能させる必要がある．

1.2 扱う問題

本研究では，ベクトル空間モデルにおいて表現されたベクトルが張る空間に対して，空間次元の縮小処理を行う．以下では，これらの次元縮小手法を低次元ベクトル空間への射影（projection）という意味を込めてプロジェクション手法と呼ぶ．データの特性を維持したままベクトルの次元を縮小することで，検索の効率化を図ることができる．

1.2.1 多義語と同義語

同じ内容の文書を探す場合でも，ユーザによって入力される検索質問には差異が生じる．ただし，それぞれの単語はある単語の同義語という関係を持つと考えられる．例えば映画やテレビなどの撮影技師について検索する場合，“cameraman” “cinematographer” “camera operator” など．同じ意味を持つ索引語同士を排他的に区別せず，索引語と同じ意味を持つが索引語にはない単語を新たな索引語として取り入れる必要がある．

本研究では，同義語を同一の次元へ縮退させる効果を持つプロジェクション手法である Latent Semantic Indexing（LSI）[5] とシソーラス辞書である WordNet [1]を併用する．一般的に単語は複数の意味を持つ．WordNet は単語の同義語，上位語，下位語，部品語などの意味関係を保持している．それぞれの単語に WordNet を適用し，索引語の持つ全ての意味で同義となる単語を用いて索引語を拡張する．

[1]http://www.cogsci.princeton.edu/~wn/

1.2.2 文書の有効期限

ユーザは一般的に古い文書より新しい文書を優先して検索する．特にニュース記事の検索ではその傾向が顕著になる．文書の時系列データをテキストストリーム，特にニュース記事を対象としたテキストストリームをニュースストリームと呼ぶ．当然ながら，ニュースストリームの検索には即応性が求められる．それと共に，どこまで古い文書なら検索対象として認めるか，あるいは興味を少しづつ失ってゆくことを検索要求として定義できなければならない．

本研究では，古い文書に重み付けを行うことによってこれらの質問要求をモデル化し，データに依存しないプロジェクション手法である Random Projection（RP）[10] を適用する．

1.2.3 複数の応用分野を持つ文書集合

ユーザがある一定の目的をもって一連の検索を行う場合，それが同一の応用分野に偏ることは珍しくない．あらかじめどの応用分野について検索するかを指定する事で，検索効率を向上させることができれば，全体として検索ノイズや検索漏れを低く抑えることができる．

本研究では，2 種類の応用分野を持つ文書集合に対して，従来のプロジェクション手法には無い，“局所的なデータ非依存性” を持つプロジェクション手法を適案する．この手法を，RP 手法と対比させる意味で Skewed Projection（SP）と呼ぶ．

1.3 論文の構成

本研究では，以上の問題について以下の構成で論じる．

第 2 章では，関連研究として現在の代表的なプロジェクション手法である LSI 手法と RP 手法について紹介し，それぞれの特徴と問題点を挙げる．

第 3 章では，プロジェクションと語彙データベースを併用した文書検索について論じる．検索の効率を維持しながら，意味による文書検索の可能性が向上することを示す．

第 4 章では，ニュースストリームの検索に対してプロジェクションを適用する．時間に対するユーザの検索要求をモデル化し，平行してデータに依存しないプロジェクション手法を用いることで，ニュースストリームの持つ特性を損なわずに効率的な検索が可能であることを実験的に検証している．

第 5 章では，単語の頻度分布に基づいたプロジェクション手法を提案している．提案手法が局所的なデータ非依存性を持ち，その非依存性が保たれている範囲では，従来手法に対して高い検索精度を得る事を述べる．

第 6 章では本研究を要約し，本研究では扱うことができなかった問題について言及する．

1.4 発表論文

1. 大内浩仁, 三浦孝夫, 塩谷勇: “多義性を考慮した文書検索”, データ工学ワークショップ（*DEWS*）, 2003.

2. Oh’ Uchi, H. Miura, T. and Shioya, I.: “Document Retrieval by Word Meanings”, *IEEE Pacific Rim Conference on Communications, Computers and Signal processing (PACRIM’ 03)*, pp. 804-807, 2003

3. 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テキストストリームの検索”, データ工学ワークショップ（*DEWS*）, 2004.

4. Oh’ Uchi, H. Miura, T. and Shioya, I.: “Retrieval for Text Stream by Random Projection”, *International conference on Information Systems Technology and its Applications (ISTA)*, pp. 151-164, 2004.

5. 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トリームの検索”, 日本データベース学会 Letters (*DBSJ Letters*) Vol.3, No.3, pp. 1-4, 2004

6. 大内浩仁, 三浦孝夫, 塩谷勇: “頻度分布に基づくプロジェクションを用いた文書検索”, データ工学ワークショップ（*DEWS*）, 2005.

第2章　関連研究

プロジェクションにおいては，まず文書集合を行列（データ行列）に見立て，射影行列を乗じることで次元縮小を行う．例として，単語数 d，文書数 N の文書集合を大きさ $d \times N$ のデータ行列 X と置く．X の i 行 j 列の要素 x_{ij} は，文書 j における頻度を元に，重み付けや正規化などを行った値である．この場合次元数は d である．この d 次元データ行列を k 次元に縮小する場合は，射影行列として $k \times d$ の行列を決定し，左側から乗じることで，k 次元のデータ行列 X_{DR} を得る．

$$X_{DR}^{k \times N} = S^{k \times d} X^{d \times N}$$

現在知られているプロジェクション手法として，LSI 手法と RP 手法がある．

LSI 手法は，プロジェクション手法として最も広く認知されている手法である．文書集合を行列（データ行列）として見立て，得意値分解（SVD）を行うことで主成分分析と同じアプローチで次元縮小を行う．縮小された行列はフロベニウスノルムの意味で最小に抑えられることが保証されている．しかし，データ行列を用いて射影行列を計算する必要があるため，射影行列は元のデータに依存する．なおかつ SVD 自体が大きな計算量を必要とするため，静的でない文書集合に対する検索に適用するのは困難である．

近年注目されているプロジェクション手法が RP 手法である．RP 手法では，ランダムな要素で構成された行列を射影行列として用いる．行列は正規直交化されている必要があるが，単純な要素の確率分布で近似的に正規直交化の条件を満たすことが可能である [1]．RP 手法における射影行列を R と置くと，R の要素 r_{ij} は次の確率分布を取るように並ぶ．

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

次元縮小された行列は，各ベクトル間の相対的なユークリッド距離を保つ．ベクトル空間モデルにおいてはベクトル間の類似度を適合度の基準とするため，検索質問と文書集合間の適合度も同じく維持される．この射影行列はデータ行列とは関係なく作成することができるため，データに対して非依存であるといえる．しかし，そのランダム性ゆえに数次元から数十次元の低次元では誤差が増幅し，検索精度が低下する．

第3章　多義性を考慮した文書検索

文書検索では通常，単語を索引としている．本論文では単語の多義性を利用した検索方式を提案する．語彙データベースである WordNet を用い，索引を意味の言葉に置き換えることで，意味による文書検索が可能になることを示す．潜在的意味索引付け（Latent Semantic Indexing，LSI）により検索を効率化し，実験によってその有効性を検証する．

3.1　前書き

現在，コンピュータネットワーク上には膨大な量の情報が存在している．膨大な量の情報からいかにして効率的に必要な情報を取り出すかという方式を決めることが重要である．我々は，語の意味関係を文書検索に導入することで，文書検索の機能を拡張する方式を提案する．特に文書検索では，字面による類似性ではなく意味による類義性を意識した検索を行うことができれば，検索の機能を拡張する方法として非常に有効である．

従来の文書検索では，あらかじめ策定された索引語によってのみ検索が可能となっている．このような検索システムでは，類義語，多義語の関係を意識していない．例えば,「学生」と「生徒」は明らかに類義語の関係にあるが，別の索引語として扱われることになる．また，picture という単語が「絵」「写真」「景色」など，複数の意味で用いられている場合に，これを区別することができない．

語の意味を意識した検索方式として，潜在的意味索引付け（Latent Semantic Indexing，LSI）[11, 7] が存在する．LSI では類義語を同一の次元に圧縮することによって，探索空間の次元を縮小するとともに，類義語の発見を可能にしている．しかし，LSI による検索は，文書集合内で定義される局所的な関係を表すに過ぎない．また，意味による検索においては，類義語の発見だけでは不十分である．

本研究では，文書検索機能の拡張を目的とし，単語ではなく単語の意味を用いた検索を行う．共通知識・概念を持つ意味関係を導入することで，単語の意味範囲を拡張する．語彙データベースである WordNet を活用し，索引語の多義語を意味の言葉として置き換えることによって意味検索を実現する．語の置き換えによって増大した次元は，LSI によって縮小する．

提案方式では，単語ではなく意味を直接扱うため，文書の正確な絞込みを行うことができる．また，決まった単語ではなく意味による質問が可能になる．

次元の増加に伴い，LSI における特異値計算のコストが増大する問題がある．これについては，多数の文書からサンプリングを行い，十分信頼性を維持しながら特異値計算を実行する方式 [8] に基づいて，小規模の文書数で実験を行う．

2 節では，LSI とベクトル空間モデルの概要を述べる．3 節では，WordNet の概要について述べ，WordNet を用いた意味関係の導入方法について論じる．4 節で実験を行い，5 節で結びとする．

3.2 LSI とベクトル空間モデル

ここでは，本研究で利用する LSI と，その基礎であるベクトル空間モデルの概要を述べる．

3.2.1 ベクトル空間モデル

文書集合と検索質問をベクトル空間上に表現し，ベクトルの類似度計算によって文書の適合度を判定する検索モデルを，ベクトル空間モデル [11] と呼ぶ．m 個の索引語と n 個の文書から成る文書集合は $m \times n$ 行列によって表現される．この行列をデータ行列と呼ぶ．データ行列の中で文書は m 次元のベクトルとして表現されている．ベクトルの要素は索引語の頻度によって決定される．

3.2.2 LSI

LSI はベクトル空間モデルに基づいた検索手法である．LSI を用いることで，意味関係の導入によって増大した次元の縮小，および意味関係との相互作用が期待できる．ここでは，LSI による質問検索の流れを，次に定義するデータ行列 D を例として述べる．

$$D = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 3 & 1 \end{pmatrix}$$

まずデータ行列を特異値分解する．$m \times n$ 行列の特異値分解は次のように定義されている．

$$D = U\Sigma V^T \tag{3.1}$$

ここで，U は $m \times r$ 直交行列（$U^TU = UU^T = I$ となる行列，I は単位行列），V は $n \times r$ 直交行列（$V^TV = VV^T = I$）である．ここで，$r = \mathrm{rank}(D)$ である．

Σ は $r \times r$ 対角行列である．Σ の対角要素を特異値という．

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \tag{3.2}$$

とした場合，特異値は

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 \tag{3.3}$$

を満たす．

例として，先ほど定義した D の特異値分解を行うと，

$$U = \begin{pmatrix} -0.24 & 0.59 & -0.28 \\ -0.40 & 0.35 & 0.51 \\ -0.42 & 0.38 & -0.44 \\ -0.19 & -0.21 & -0.16 \\ -0.28 & 0.053 & 0.65 \\ -0.70 & -0.59 & -0.15 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 3.82 & 0 & 0 \\ 0 & 2.79 & 0 \\ 0 & 0 & 2.58 \end{pmatrix}$$

$$V^T = \begin{pmatrix} -0.45 & -0.71 & -0.54 \\ 0.82 & -0.57 & -0.073 \\ -0.36 & -0.41 & 0.84 \end{pmatrix}$$

となる．ここで，

$$U = [\mathbf{u_1 u_2} \ldots \mathbf{u_r}] \tag{3.4}$$

$$V = [\mathbf{v_1 v_2} \ldots \mathbf{v_r}] \tag{3.5}$$

と表し，行列 U および V を列ベクトルの集合で表現する．式（3.1）を，

$$D = \sum_{j=1}^{r} \mathbf{u_j} \sigma_j \mathbf{v}_j^T \tag{3.6}$$

と表す．r 個のベクトルによって $m \times n$ のデータ行列を再現できることを示している．特異値が高い項ほど，データ行列 D への影響力は強くなる．

式（3.3）より，1 番目の項が D の復元に最も大きな影響力をもち，r 番目の項が最も影響が少ない．U，V，Σ から最初の k（$< r$）個のベクトル，特異値を選ぶことで，データ行列 D を k 次元で近似する．

k 次元のデータ行列に対し，U，V を構成するベクトルに特異値を重みとして掛け合わせる事で，k 次元の索引語ベクトルおよび文書ベクトルを作成する．

索引語 $t_i (i = 1, 2, \ldots, m)$ を k 次元ベクトル空間に表現する索引語ベクトル $\mathbf{t_{ki}}$ は，

$$\mathbf{t_{ki}} = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, \ldots, \sigma_k u_{ik})^T \tag{3.7}$$

で表される．同様に，文書 $d_j (j = 1, 2, \ldots, n)$ を k 次元ベクトル空間に表現する索引語ベクトル $\mathbf{d_{kj}}$ は，

$$\mathbf{d_{kj}} = (\sigma_1 v_{j1}, \sigma_2 v_{j2}, \ldots, \sigma_k v_{jk})^T \tag{3.8}$$

で表される．先ほどの例で k = 2 として $\mathbf{d_{21}}$ を考えると，

$$\mathbf{d_{21}} = (\sigma_1 v_{11}, \sigma_2 v_{12}) = \begin{pmatrix} 3.82 \times -0.45 \\ 2.79 \times -0.71 \end{pmatrix} = \begin{pmatrix} -1.72 \\ -1.98 \end{pmatrix}$$

となる．

検索を行うためには，検索質問を同じ k 次元空間に表現する必要がある．m 次元の質問ベクトル $\mathbf{q}$ を k 次元空間に表現したベクトルを $\hat{\mathbf{q}}$ とすると，

$$\hat{q}_i = \frac{1}{\sigma_i} \mathbf{q}^T \mathbf{u}_i (i = 1, 2, \ldots, k) \tag{3.9}$$

で $\hat{\mathbf{q}}$ を求められる．例えば

$$\mathbf{q} = (0, 0, 1, 1, 0, 1)^T$$

という質問を与えた場合，k=2 の例では，

$$\hat{\mathbf{q}} = (-0.34, -0.15)^T$$

となる．

質問検索をベクトルの類似度計算によって行う．本研究では質問ベクトルと文書ベクトルの余弦 (cos) の値を用いる．文書集合の中から i 番目の文書を調べる場合，

$$\cos \theta_{ki} = \frac{(\hat{\mathbf{q}}, \mathbf{d}_{ki})}{|\hat{\mathbf{q}}||\mathbf{d}_{ki}|}$$

の値によって，検索質問に対する文書の類似度を調べる．類似度は 1 から-1 の値を取り，大きいほど質問と適合している．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にして表示する．

3.3 語彙データベース

本論文では，語彙データベースとして WordNet を用いる．ここでは，WordNet の特徴および WordNet を使用した意味関係の導入方法について述べる．

3.3.1 WordNet

WordNet はフリーウェアとして提供されている語彙データベースである．シソーラスに近いが，単語ではなく同義語の集合である synset（synonym set）を辞書の構成単位としている．synset によって，多角的かつ階層的な意味関係の表現を可能にしている．

WordNet で検索することができる意味関係は，同義語 (Synonym)，反義語 (Antonym)，上位語 (Hypernym)，下位語 (Hyponym) の他に，部品語 (Meronym) の関係と，部品語

の逆の関係を表す Holonym がある．部品語の関係とは，例えば日付に対する年，月，日の関係を表す．

同族語 (Coordinate Terms) は，意味の階層関係において同じ階層に位置する語であり，直接の上位語に対する下位語として定義されている．

synset は品詞毎に分けて管理されている．名詞，形容詞，副詞，動詞に対応しており，あわせて約 95,600 語を収録している．1 つの synset には，同じ意味を持つ 1 つ以上の単語が含まれている．例えば「教育機関に所属する学習者」という意味の synset は {student，pupil，educatee} となる．すなわち，同じ synset に属している単語は同義語となる．

複数の意味をもつ単語は複数の synset に表れる．例として「performance」を挙げると，

{performance，public presentation} ,
{performance，execution，carrying out，carrying into action} ,
{operation，functioning，performance} ,
{performance} となる．

WordNet は，品詞ごとの索引ファイルとデータファイル，検索プログラムからなる．

索引ファイルは，それぞれの単語の属している synset，単語の品詞，その単語から検索できる意味関係を示している．データファイルは，synset の識別番号，その synset 持っている意味，synset に含まれる単語数，および synset に含まれる全ての単語のリストを格納している．

3.3.2 WordNet による意味関係の導入

意味関係の導入後における検索では，索引語を含む synset を列挙し，そこに含まれるすべての単語を一つの意味集合と考え，索引語と置き換えている．このため単語数は増大する．意味関係の導入前における検索では，索引語をそのまま質問語として検索を行っている．例として，名詞「performance」の場合をみると，

導入前：performance

導入後：{performance，public presentation，execution，carrying out，carrying into action，operation，functioning}

となる．

品詞によって語彙の取り扱い方は異なる．本研究では，名詞のみを置き換えの対象としている．

3.3.3 WordNet と LSI

意味関係の導入は，データ行列における索引語の増加として表れる．本研究では，意味関係の導入前と後に対応する，2 つのデータ行列を作成する．この 2 つのデータ行列に対して，LSI による質問検索を行う．

導入例として，3 つの索引語と 2 つの文書による

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

というデータ行列を考える．索引語の重み付けには 2 進重みを用いている．1 番目の索引語が，意味の関係として 5 つの単語を含み，他の 2 つの単語は多義性を持たないとする．この場合意味関係導入後のデータ行列 D' は，

$$D' = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

となる．

ベクトル空間モデルにおいては，探索空間の次元がそのまま検索時間に比例する．意味関係の導入によって検索効率が悪化する問題があるが，LSI を併用することによって，質問の機能を拡張しながら，検索効率を維持することを期待できる．

3.4 実験

意味関係の導入による効果を検証するため，2 種類の実験を行う．

3.4.1 実験環境

元データとして Reuters Corpus[15] を使用する．Reuters Corpus はロイター社の新聞記事を XML フォーマットで構成した大規模文書集合で，1 年分，806,791 件のデータを持つ．この中から 507 件を抜き出して文書集合とする．

索引語には，記事のカテゴリを表すトピックス・コードを用いている．トピックスコードを，対応表をもとに自然語に変換し，元の索引語と置き換える．161 語のカテゴリ構成語に対して索引語の置き換えを行う．結果として 795 語の索引語を得る．

索引語の重み付けには，2 進重みを用いる．索引語の頻度は，その語が存在していれば 1，存在していなければ 0 となる．質問ベクトルの頻度も，単語が質問に含まれていれば 1，含まれていなければ 0 とする．

3.4.2 評価方法

実験の評価には，情報検索で広く用いられている再現率と適合率，および 11 点平均適合率を用いる．

再現率は，検索漏れの少なさを示す尺度であり，

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

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，

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

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．再現率・適合率グラフによって情報検索システムの性能を計る．

この実験では，類似度順にランキングされた文書集合に対して，1 位から順に適合文書かどうかを判定し，そのつど再現率と適合率を求めている．再現率が 1，つまり全ての適合文書が検出された時点で評価は終了する．

11 点平均適合率は，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である．この値が，再現率と適合率の関係を総合的に評価する尺度となる．

3.4.3 意味関係の導入前と導入後における検索精度の比較

意味関係の導入前における検索では，索引語をそのまま意味語として用いる．文書集合は 161 × 507 行列で表現される．意味関係の導入後における検索では，索引語を含む synset を列挙し，そこに含まれるすべての単語を一つの意味集合と考え，索引語と置き換えている．このため単語数は増大し，文書集合は 795 × 507 行列で表現される．

単語を検索質問として検索を行う．導入前では performance を質問語とし，質問ベクトルを構成して検索を行う．導入後では，performance，public presentation，execution，carrying out，carrying into action，operation，functioning の 7 語を質問語とし，同様に検索を行う．

意味関係導入前と導入後の再現率－適合率グラフを図 3.1 および 3.2 に示す．

11 点平均適合率による結果を表 3.1 に示す．

意味関係の導入によって，低次元における精度の減少を抑制する効果があるといえる．検索精度は 15 次元の場合を除いて上昇している．特に 7 次元では 35 %の，10 次元では 41 %の精度差が見られる．本実験では，文書中に 100 件近くの同一索引語の不

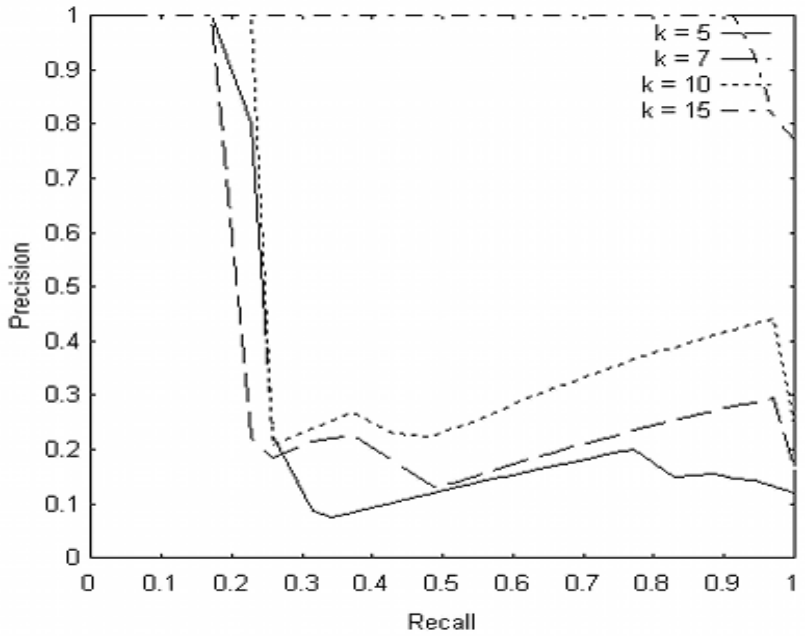

図 3.1: 意味関係導入前

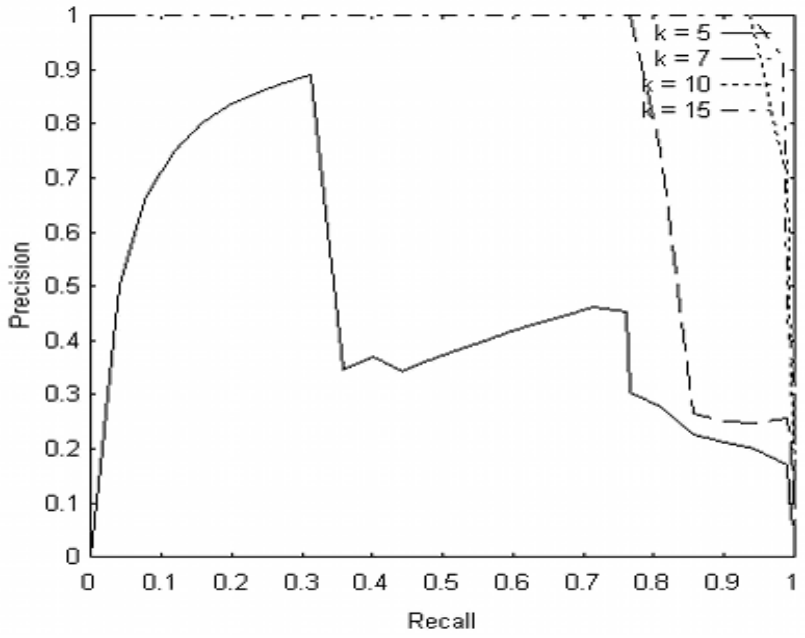

図 3.2: 意味関係導入後

	意味関係導入前	意味関係導入後	
次元	平均適合率	平均適合率	比較結果
5	0.3839	0.5393	+0.1554
7	0.4115	0.8245	+0.4130
10	0.5755	0.9245	+0.3490
15	0.9792	0.9364	-0.0428

表 3.1: 11 点平均適合率

適格文書が存在しているため，不適格文書の固まりが検索結果の上位に入ってしまうと精度が急激に低下する．このため，極端な検索精度の差が発生していると思われる．

ここでは 15 次元を上限としているが，これ以上高次元になると，両方のデータでほぼ 100 %の検索精度となり，比較ができない．また，意味関係導入後の 5 次元の検索結果で，再現率と適合率が共に 0 となる現象が発生している．原因は，類似度で第 1 位と判定された文書が不適合となったためである．その後精度が回復し，11 点平均適合率では意味関係導入前を上回っている．

3.4.4 意味の言葉による質問検索

意味関係導入後の文書集合を用いて，次の 3 種類の検索を行う．

1. 意味関係導入前の索引語を含まない，意味集合内の一部単語による質問
2. 意味関係導入前の索引語 1 語のみによる質問
3. 意味関係導入前の索引語を含む，意味集合の全ての単語による質問

導入前の索引語は performance とする．具体的な質問語は，

1. {carrying out，execution}
2. {performance}
3. {performance，public presentation，execution，carrying out，carrying into action，operation，functioning}

となる．文書の索引語集合が performance を含んでいれば，その文書は適合文書である．

（3）は完全な意味質問で，前の実験における意味関係の導入後の検索と同一の性質をもつ．(1）は不完全な意味質問である．(2）の場合も，拡張元の単語であっても頻度差は無いので，やはり不完全な意味質問である．(3）に比べて（1）,(2）の精度が落ちなければ，単語の字面ではない，意味による検索の可能性が実証できる．

再現率－適合率のグラフを図 3.3，3.4 および 3.5 に示す．

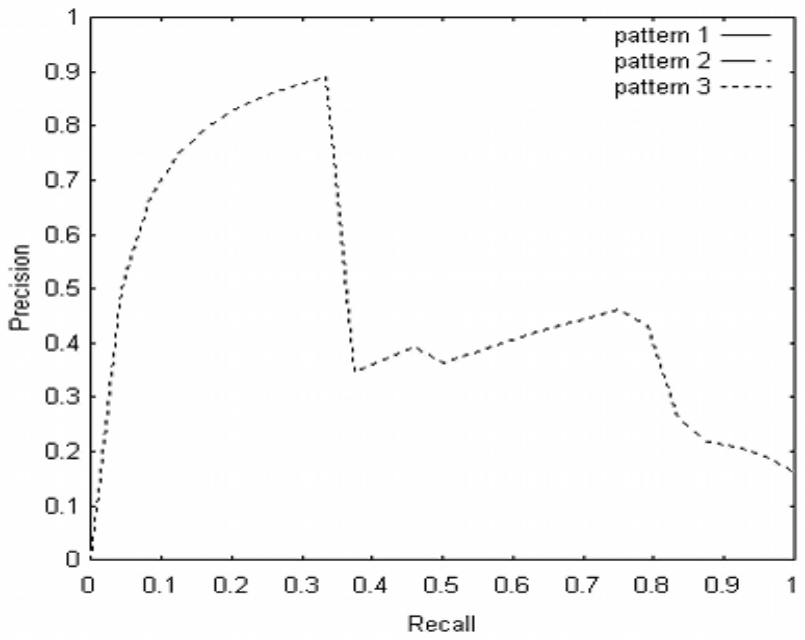

図 3.3: 5 次元における意味検索

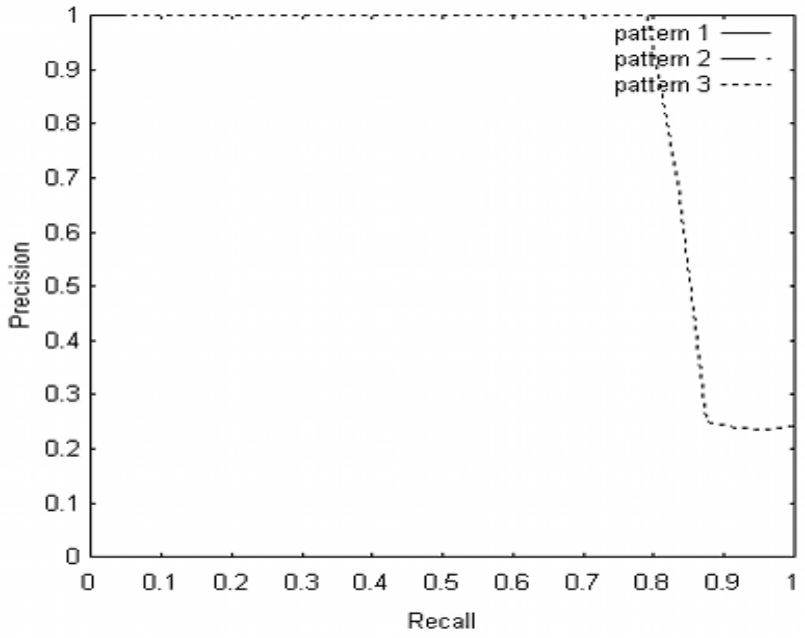

図 3.4: 7 次元における意味検索

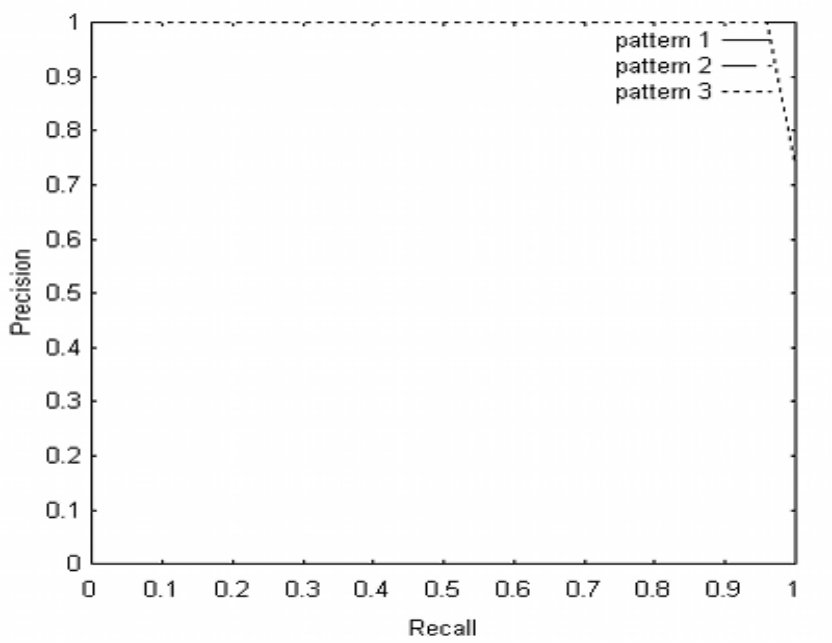

図 3.5: 10 次元における意味検索

次元	質問 1	質問 2	質問 3
5	0.5445	0.5445	0.5445
7	0.8337	0.8337	0.8337
10	0.9752	0.9752	0.9752

表 3.2: 11 点平均適合率

11 点平均適合率による結果を表 3.2 に示す.

意味検索の有効性が実証されている．グラフは完全に重なり，11 点平均適合率も同じ値である．3 種類の質問に対して，まったく同じ結果を出力している．意味集合に属する単語が全て同じ次元に圧縮されていることが推測できる．意味関係の導入前には存在しなかった単語を新たな索引語として検索することが可能となる．

3.5 結論

意味関係の導入によって，多義語の概念を文書検索に取り入れることができるようになった．また，LSI を併用することで，意味関係の導入後も検索効率を維持できるだけでなく，検索精度の上昇も期待できる．

利用できる索引語の幅が広がることで，意味による検索質問が可能となっている．

今後は，出現頻度の重み，WordNet の適用時における重み，などによる調整を進めて

いくことが課題となる．今回の実験では，最も単純な2進重みを使用している．TF*IDF法などの大域的重みも含んだ重み付けを考慮する必要がある．また，多義性を持ち，意味関係の導入によって展開された単語と，多義性を持たない単語との頻度に格差が存在していないことも問題である．

第4章　ランダムプロジェクションを用いたニュースストリームの検索

本論文は効率的なニュースストリームの検索方法を提案する．ニュースストリームの検索においては，更新情報をどのように取り扱うかが検索処理の効率向上において問題となる．ランダムプロジェクションを用いた動的な次元縮小が検索性能および検索時間の双方で充分実用的な処理を可能とすることを述べ，実験によりその有効性を示す．

4.1　前書き

近年，情報検索の分野において，TDT（Topic Detection and Tracking）プロジェクトに代表される時系列データへの要求が高まっている．特に，文書の時系列データとして配信されるニュース記事，いわゆるニュースストリームから記事を検索するためにはいくつかの問題がある．

一般にニュースストリームには大きく異なる2つの特性が含まれている．新聞のように，新たなトピックが発生すると爆発的な量のニュースが生成され，時間と内容の関連性を追跡することが重視される．これをトピック特性という．これに対し，TV ニュース放送では，一定時間に一定量のニュースが定常的に生成される．このことを実時間特性という．ニュースストリームに対する検索では，両特性に関して検索基準や結果の評価が異なる．このため，検索手法の有効性を検証するためには，トピック性および実時間性に対する検討が必要である．また，ニュースストリームに対する検索は即応性を持たなければならない．

一般に，テキスト集合に対する検索は主な単語を取り出し，これらを用いたベクトル空間モデルに基づいて処理される [11]．テキストデータは語いの数だけ次元が存在し，一般的に数万から数十万次元の高次元データとなる．高次元データをそのまま扱うと，計算機容量の確保および即応性への対応が困難になる．計算機容量の効率化と，即応性を実現するためには，テキストデータの次元を縮小して格納する必要がある．

テキストデータにおける次元縮小の方式として Latent Semantic Indexing（LSI）[5, 12, 14, 2] が知られている．LSI 手法では，特異値分解（SVD）を用いて検索空間の次元を縮小する．これにより検索精度を維持したまま次元を大きく縮小することができるため，検索効率と検索精度を両立することができる．しかし，LSI 手法にはデータ

の更新に対してSVDのための再計算が必要となる．LSI手法をニュースストリームの差分情報に対応させるのは容易ではない．

本研究では，ランダムプロジェクション（RP）[14]を用いることによって，計算機容量および検索効率において効率的なニュースストリームの検索方式を提案する．RP手法による次元縮小は計算処理が少なく，データの更新時に再計算を行う必要がないため，差分情報に対して動的に次元縮小を行い，検索質問に対する即応性を保持することができる．

2節ではテキストデータの次元縮小方式としてRP技術とLSI技術について述べ，両者の比較を行う．3節でRP技術を用いたニュースストリーム検索の基準について述べる．4節に実験結果を示し，5節で結びとする．

4.2 テキストデータの次元縮小

テキストデータの次元を縮小する方法として，RP手法について述べる[1, 3, 10]．以下ではベクトル空間モデルに基づき，単語数 d，文書数 N の単語・文書行列 X を考える．行列の大きさは d 行 N 列であり，それぞれの列ベクトルが1件の文書を表している．行列 X の i 行 j 列の要素 X_{ij} は，文書 j における単語 i の頻度である．

ランダムプロジェクション（RP）は要素をランダムに決定した行列である．これにより高次元データを低次元の部分空間に射影することができる．以下では，大きさ $d \times N$ の単語・文書行列 X を大きさ $k \times N (k << d)$ の単語・文書行列 X^{RP} に射影する．このため大きさ $k \times d$ のRP行列 R を決定する．単語・文書行列 X のRP技術による次元縮小は，次の計算で行う．

$$X^{RP}_{k \times N} = R_{k \times d} X_{d \times N} \tag{4.1}$$

RP行列 R の要素を構成する際に，非常に単純な要素の分布[1]が提案されている．RP行列 R の要素 r_{ij} は，次のような独立した分布をとるように並ぶ．

$$r_{ij} = \sqrt{3} \cdot \begin{cases} +1 & \text{確率 } 1/6 \\ 0 & \text{確率 } 2/3 \\ -1 & \text{確率 } 1/6 \end{cases} \tag{4.2}$$

この分布に基づいて構成される R では，列ベクトルの長さの期待値が全て1になる．

質問検索は，質問をベクトル表現した $\mathbf{q}_{d \times 1}$ をk次元空間に射影して行う．

$$\mathbf{q}^{RP}_{k \times 1} = R \mathbf{q}_{d \times 1} \tag{4.3}$$

検索結果として質問ベクトルと文書ベクトルとの類似度をコサイン尺度で計算し，ランキングを表示する．

RP手法の次元縮小にともなう誤差は，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 $\mathbf{x}_2$ とお

く．$\mathbf{x}_1$ と $\mathbf{x}_2$ の d 次元におけるユークリッド距離は，$|\mathbf{x}_1 - \mathbf{x}_2|$ で定義される．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再現することができる．[3]

$$\sqrt{d/k}|R\mathbf{x}_1 - R\mathbf{x}_2| \tag{4.4}$$

式（5.7）が成り立つためには，R が直交行列（$R^T R = I$）である必要がある．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．

$$\epsilon = R^T R - I \tag{4.5}$$

このとき ϵ の要素は，平均 0，分散 $1/k$ の正規分布をとる [10]．よって縮小次元数 k を大きくするほど，誤差は減少する．

LSI 手法では，特異値分解（SVD）によって射影行列を求める．単語・文書行列 X の SVD は，$X = U \times S \times V^T$ で表される．U は $d \times r$ のユニタリ行列である．S はランク r の対角行列でこの対角要素を特異値と呼ぶ．V は $r \times N$ のユニタリ行列である．次元縮小には U を射影行列として用いる．SVD は大きい計算量を必要とし，元の行列に依存する処理である．このため更新に対して再計算が必要となる．対して RP 行列は少ない計算量で作成可能である．また，データに依存しないため，更新に伴う再計算は必要ない．RP 手法は本質的にニュースストリーム検索に対して有効といえる．

4.3 ニュースストリーム検索の基準

ニュースストリームでは，時系列的に配置されたテキストデータが順次入力されていくことになる．そのため，更新されたばかりの新しい文書と過去の文書が混在する．一般的に，質問者にとっては過去の情報より新しい情報に価値があることが多い．その場合，過去のデータと新しいデータを同列に扱うことはできない．

トピック性ニュースストリームに対する検索は，一般的に最新のトピックに対して行われる．検索要求を満たすためには，新しい文書に対して優先順位を与える仕組みが必要となる．一方，実時間性ニュースストリームでは，過去 1 週間のニュースなど，検索期間のみ指定する質問が多い．この場合，期間内の文書だけを検索対象とする代わりに，時間差による優先度の格差は存在しない．

本研究では，数学的関数を用いた重み係数を計算することで，ニュースストリームへの検索要求を表現する．トピック性ニュースストリームに対しては，指数関数的に減少する重み係数を使用する．t を文書の経過時間とおくと，重み係数 $w_a(t)$ は，以下の式によって表される．

$$w_a(t) = \exp(-t/a) \tag{4.6}$$

a は重み係数の減少度を調節するパラメータである．a が大きいほど重み係数の減少は緩やかになり，$a = \infty$ で減少が完全に止まる．

実時間性ニュースストリームの検索では，窓関数による重み係数を使用して検索期間を限定する．窓関数による重み付けでは，検索期間を表すパラメータを p と置いた場合，重み係数 $w_a(t)$ を次のような式で表す．

$$w_a(t) = \begin{cases} 1 & (t \leq p) \\ 0 & (t > p) \end{cases} \tag{4.7}$$

いずれの場合も，質問者が関数のパラメータを与えることで，様々な優先順位や検索期間の与え方を制御できる．

4.4 実験

本章では，RP 手法を用いたトピック性および実時間性ニュースストリームの検索実験を行う．まず LSI 手法と RP 手法を比較する．次にトピック性および実時間性ニュースストリームの検索を行い，検索結果について考察する．

4.4.1 実験環境

以下の実験は，FreeBSD4.6.2，Pentium4 2.8GHz，メモリ 1GB の計算機上で行う．ニュースストリームのデータとして，Reuter-21578 コーパス[1] および TDT2 コーパス[2]を使用する．これらのコーパスの詳細を表 4.1 に示す．

	Reuter コーパス	TDT2 コーパス
特性	記事データ	放送データ (CNN)
有効文書数	19042 件	15785 件
文書期間幅	1 年間	6ヶ月間
索引語数	2662 語	2859 語
検索間隔	6 時間ごと	1 件ごと
検索回数	199 回	15785 回

表 4.1: Reuter コーパスおよび TDT2 コーパスの実験環境

Reuter コーパスでは，6 時間ごとに出現する文書数に 0 件から 422 件までの幅がある．この特徴から Reuter コーパスをトピック性ニュースストリームとして扱う．

TDT2 コーパスは実時間性ニュースストリームに対応する．CNN の放送時間が決まっているため，検索期間を区切った場合に検索文書数がほぼ一定になる．

[1] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[2] http://www.ldc.upenn.edu/Projects/TDT2/

4.4.2 評価方法

検索結果の評価として，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，

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

で表す．適合率は，検索ノイズの少なさを示す尺度であり，

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

で表す．再現率と適合率はトレード・オフの関係にあるため，11 点平均適合率が検索精度を示す指標となり得る．

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，その結果類似度が 0.5 以上となった文書を選ぶ．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．

ニュースストリームの検索では，検索質問による 11 点平均適合率の推移を図示し，11 点平均適合率の平均値を求めて総合的な指標とする．

4.4.3 LSI 技術と RP 技術の比較

Reuter コーパスを用いて，RP 手法と LSI 手法を使用した検索を行う．次元縮小の処理に必要な時間および次元縮小時の検索精度を実験により比較し，考察する．次節以降では，ニュースストリームについての考察のみを行う．

処理時間

LSI 手法による検索は，計算機容量の問題から，先頭の 10000 件のみを処理する．RP 手法による検索では，19042 件の文書を 1 度に処理する．LSI 手法および RP 手法が次元縮小に要した時間を表 4.2 に示す．

処理時間の比較では，RP 手法が圧倒的に勝っている．LSI では，次元数を減らしても処理時間は減少しない．他方，RP 行列は少ない計算量で作成が可能であるとともに，縮小次元数に応じて行列の大きさを減らすことができる．

検索精度

縮小後の次元数を 5 次元から 250 次元までの 10 段階に設定し，それぞれの次元数で LSI 手法と RP 手法による質問検索の検索精度を求める．

RP 手法を用いた検索では，それぞれの次元で 3 回ずつ実験を行い，11 点平均適合率の平均値を最終的な評価とする．同時に RP 手法による次元縮小で生じる分散を計る．

	次元数	処理時間（秒）
（RP）	100	74
	200	150
	300	231
	400	308
	500	387
（LSI）	2662	21469

表 4.2: RP 手法および LSI 手法の処理時間

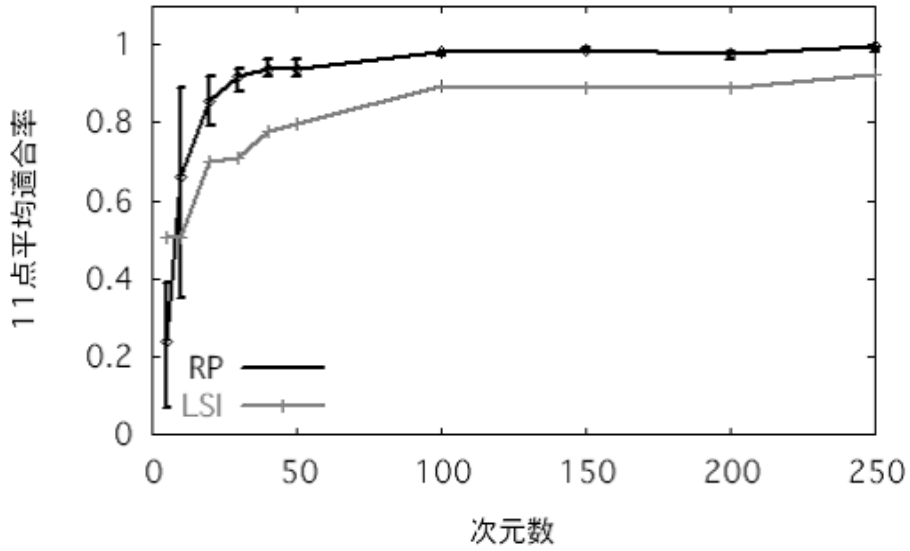

図 4.1: RP 手法および LSI 手法の検索精度

検索の結果を図 4.1 に示す.

検索精度の比較では，平均値のみを考えれば 5 次元以外の全ての次元で RP 手法が上回っている．ベクトル間の距離を保存するという RP 手法の特性が，検索精度の維持に寄与していると考えられる．分散は次元数が増加するほど少なく，100 次元以上では最低値と最大値の差が 1〜2%に収束している．よって本実験においては，100 次元以上で RP 手法が安定して LSI 手法と同等以上の検索性能を発揮するといえる．

4.4.4 RP 手法によるトピック性ニュースストリームの検索

Reuter コーパスに対して指数関数（式（4.6））に基づく重み付けを行う．t の単位は日数とする．6 時間なら 0.25 である．

過去のデータに対する重み付けのパラメータ a は，急激な重み付け $(a = 10)$ ，緩や

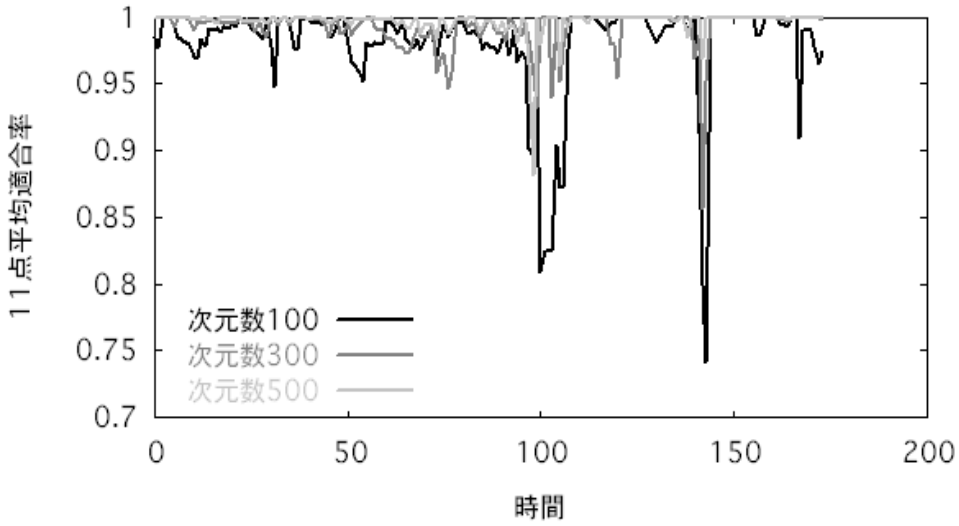

図 4.2: 急激な重み付け（$a = 10$）の検索精度

かな重み付け ($a = 45$)，重み付けなし ($a = \infty$) の 3 種類とする．急激な重み付けでは，重み係数は約 7 日間で 0.5 に減少する．緩やかな重み付けでは，約 30 日間で 0.5 に減少する．重み付けなしの場合は，重み係数は常に 1 である．

それぞれの重み付けで 100 次元，300 次元，500 次元の 3 つの次元数における検索質問を行い，11 点平均適合率の推移を求める．急激な重み付けの検索結果を図 4.2 に，緩やかな重み付けの検索結果を図 4.3 に，重み付けなしの場合の検索結果を図 4.4 に示す．

トピック性ニュースストリームにおける 11 点平均適合率の平均値を表 4.3 に示す．

	100 次元	300 次元	500 次元
急激な重み付け	0.968	0.980	0.992
緩やかな重み付け	0.979	0.992	0.997
重み付けなし	0.982	0.998	0.995

表 4.3: 11 点平均適合率の平均値—Reuter コーパス

4.4.5 RP 手法による実時間性ニュースストリームの検索

実時間性ニュースストリームとして TDT2 コーパスを用いる．重み付けには窓関数（式（4.7））を用いる．検索期間のパラメータ p を日数として，1 日間 ($p = 1$)，7 日間 ($p = 7$)，30 日間 ($p = 30$) の 3 種類を与える．前節と同様，100 次元，300 次元，500 次元において検索質問を行い，11 点平均適合率の推移を求める．検索期間 1 日の検索

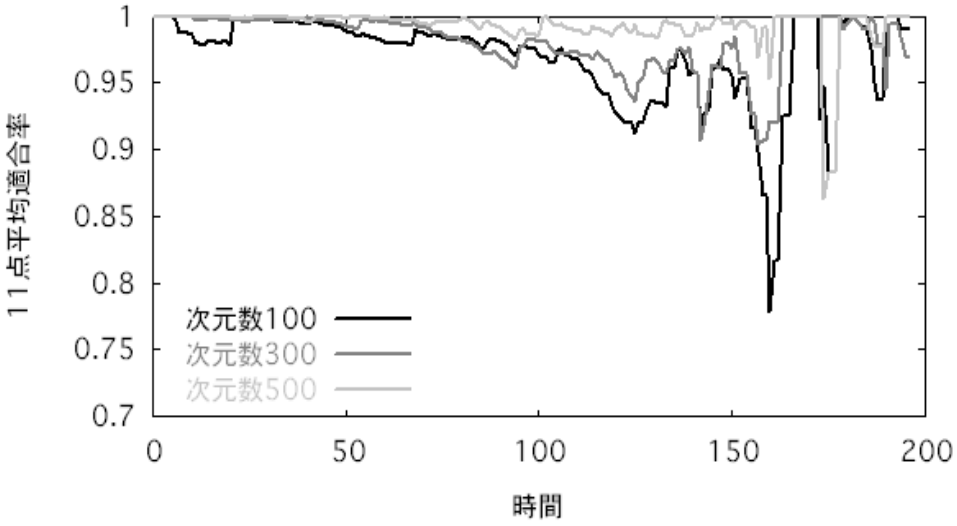

図 4.3: 緩やかな重み付け（$a = 45$）の検索精度

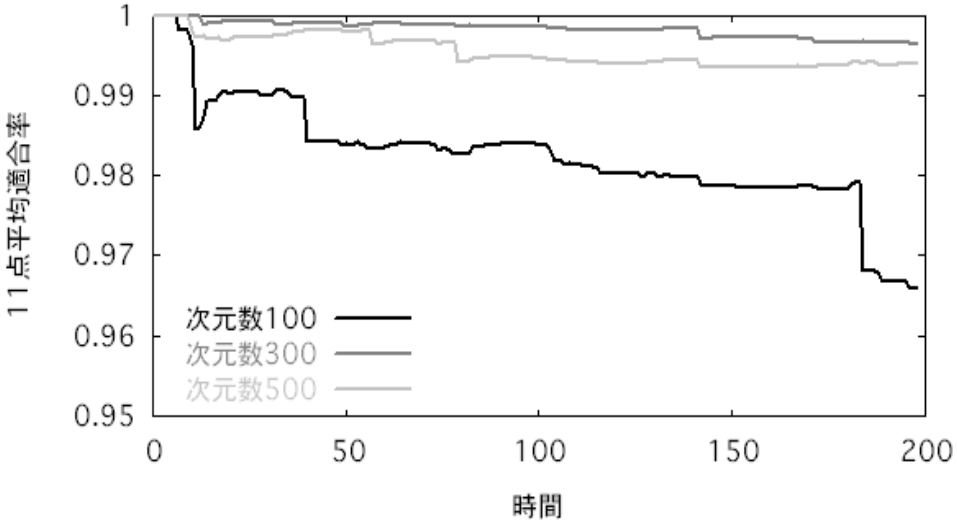

図 4.4: 重み付けなし（$a = \infty$）の検索精度

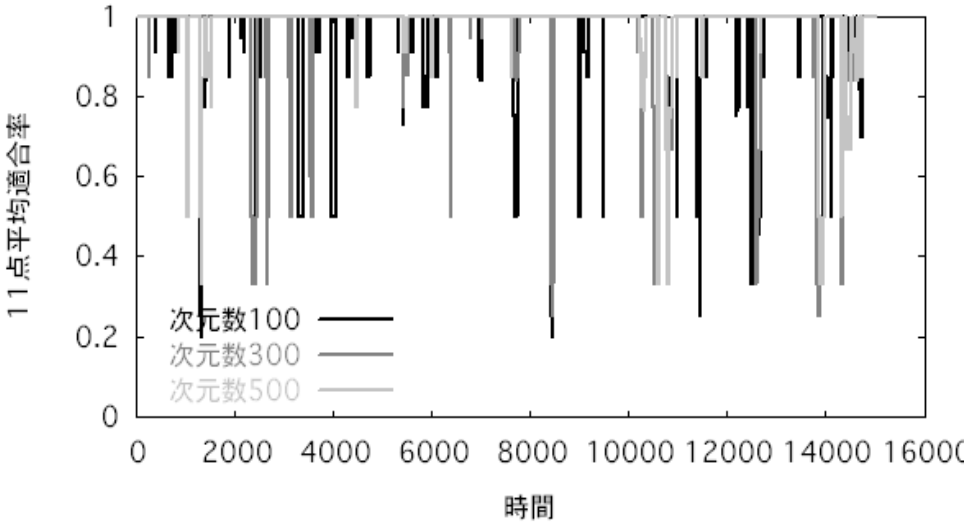

図 4.5: 検索期間 1 日 $(p = 1)$ の検索精度

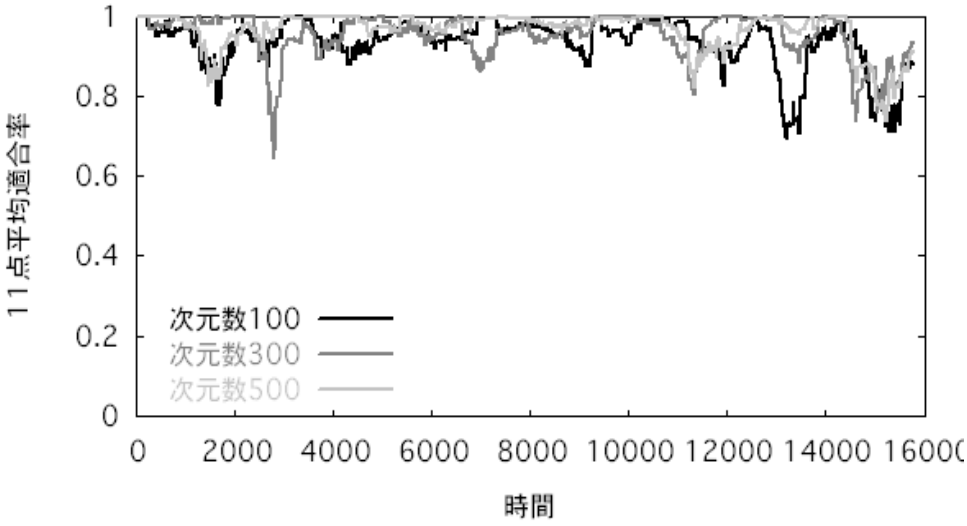

図 4.6: 検索期間 1 日 $(p = 7)$ の検索精度

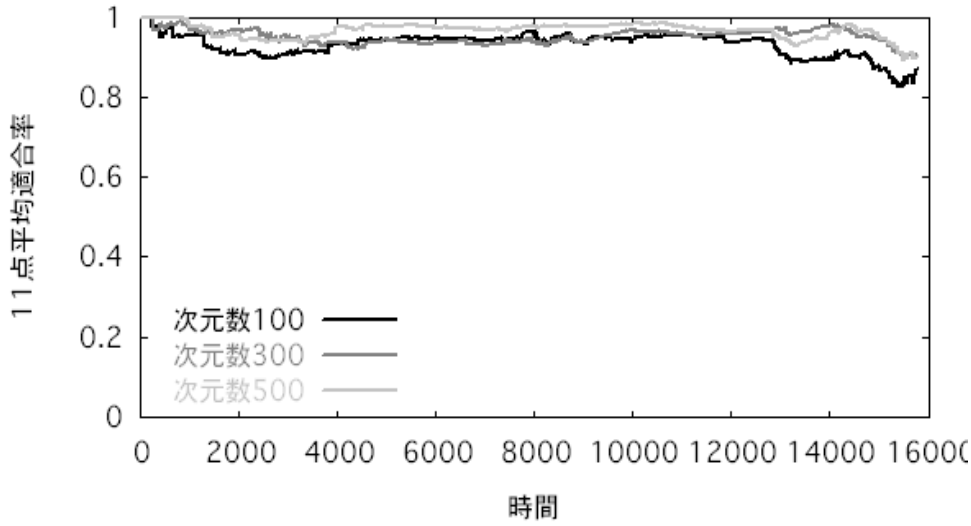

図 4.7: 検索期間 1 日 ($p = 30$) の検索精度

結果を図 4.5 に，検索期間 7 日の検索結果を図 4.6 に，検索期間 30 日の場合の検索結果を図 4.7 に示す.

実時間性ニュースストリームの検索に対する 11 点平均適合率の平均値を表 4.4 に示す.

	100 次元	300 次元	500 次元
急激な重み付け	0.957	0.965	0.981
緩やかな重み付け	0.933	0.952	0.965
重み付けなし	0.931	0.951	0.961

表 4.4: 11 点平均適合率の平均値—TDT2 コーパス

4.5 考察

トピック性ニュースストリームでは，11 点平均適合率の平均値は全ての場合で 90%を超えており，ニュースの更新に対して安定した検索を行っている．また，次元数が高くなるに従って平均値も上昇する傾向にあり，RP 行列の誤差保証が正しいことを示している.

11 点平均適合率の推移を見ると，重み付けなしの場合では適合率が継続して低下しているのに対し，重み付けがある場合には適合率の低下は局所的になっている．重み付けなしの場合は，検索もれ文書が更新によって除外されることが無いため，適合率が低下し続けると考えられる．検索漏れ文書が残り続ける重み付けなしの場合で最も

高い平均値を得たという結果は，全体を通して次元縮小の誤差が小さく抑えられていることを示している．

実時間性ニュースストリームでは，11 点平均適合率の平均値がトピック性と同じく 90%以上の値を保持している．次元が高いほど平均値が高くなる傾向も同様に見られる．検索期間ごとの比較では，短いほど平均値が高くなっており，トピック性とは逆である．

11 点平均適合率の推移を見ると，検索期間 1 日 (図 5) の場合においてごく局所的にかなり適合率の低い検索質問が存在する．この原因として，検索質問ごとの適合文書の数が極端に少ないことが挙げられる．適合文書が 2，3 件しかない場合には，1 件の検索もれが大幅な適合率の減少を引き起こすことになる．

結果的にトピック性，実時間性ニュースストリームの双方で高い検索精度を得ていることから，RP 手法がニュースストリームの検索にきわめて有効であるといえる．

4.6 結論

本研究では，ニュースストリームの検索において RP 手法を適用した．RP 手法における誤差の保証から動的な検索処理が行えることを述べ，これにより検索効率と検索時間を充分実用的な範囲で両立したニュースストリームの検索が可能になることを示した．今後は使用される記憶域についての議論を加え，モバイル環境下での利用を考えていく予定である．

第5章　頻度分布に基づくプロジェクションを用いた文書検索

ランダムプロジェクション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．このプロジェクション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．

5.1　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 [11] 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．

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 [10]．RP 手法ではランダムな要素で射影行列を構成する．そのため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．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．

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 [16]．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．

5.2 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．

5.2.1 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 の頻度である．

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．

5.2.2 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 $X_R P$ 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．

$$X_{RP}^{k\times N} = R^{k\times d} X^{d\times N} \tag{5.1}$$

この処理の計算量は $O(dkN)$ [14] 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．

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, 10]

- 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．

$$r_{ij} = \sqrt{3} \cdot \begin{cases} +1 & \text{確率 } 1/6 \\ 0 & \text{確率 } 2/3 \\ -1 & \text{確率 } 1/6 \end{cases} \tag{5.2}$$

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．

$$\mathbf{q}_{RP}^{k\times 1} = R\mathbf{q}^{d\times 1} \tag{5.3}$$

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．

5.2.3 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．この手法では単語頻度の偏りに基づ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，$0 \leq f_i \leq M$ となる．この f_i より，サンプル文書集合に基づく単語 j の出現確率を

$$Pr(f_j) = \frac{f_j}{\sum_{i=1}^{d} f_i} \tag{5.4}$$

で求める．

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（5.4）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．

$$X_{SP}^{k\times N} = S^{k\times d}X^{d\times N} \tag{5.5}$$

$$\mathbf{q}_{SP}^{k\times 1} = S\mathbf{q}^{d\times 1} \tag{5.6}$$

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．

5.3 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．

5.3.1 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 [9] 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，$k \leq O(logM/\epsilon^2)$ 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]．

$$\sqrt{d/k}|R\mathbf{x}_1 - R\mathbf{x}_2| \tag{5.7}$$

式（5.7）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．

$$\epsilon = R^TR - I \tag{5.8}$$

このとき，ϵ の要素は，平均0，分散 $1/k$ の正規分布をとる [10]．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 行列に対しては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（5.7）による近似がされにくくなる．

5.3.2 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 [13]．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（5.2）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]．

まず元の単語から，ランダムに 2/3 を破棄する．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．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．

単語を属性に置き換えれば，ユークリッド空間上のあらゆる種類のデータに対して式（5.7）に準じる一定の誤差保証を得ることが期待できる．

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．同一の頻度分布内であれば，データに依存しない次元縮小が可能である．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．

5.4 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．

5.4.1 実験環境

文書データとして，NTCIR-1 [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

[1] http://research.nii.ac.jp/ntcir/

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 [11] 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 [17] 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 [13] [2]．

SP 手法を適用する際は，それぞれの文書集合について，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 5.1 の通りに頻度分布を作成する．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 [3]を用いる．

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 5.1: 文書集合および頻度分布の構成

異なる頻度分布では，異なる文書集合の索引語によって頻度分布を決定する．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．

[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．

$$f_k = \frac{\sqrt{8F_1 + 1} - 1}{2} \tag{5.9}$$

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．

[3]http://www.math.keio.ac.jp/˜matumoto/mt.html

5.4.2 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．

5.4.3 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 5.1 に，計測文書に対する検索結果を図 5.2 に示す．

各次元で 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 5.3 に，計測文書に対する検索結果を図 5.4 に示す．

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 5.2 に，計測文書に対する検索結果を表 5.3 に示す．

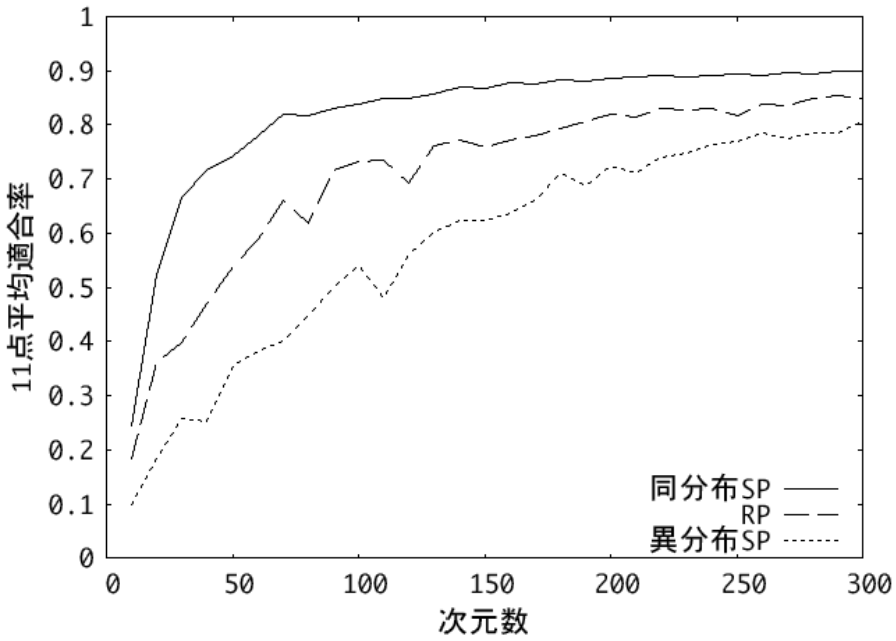

図 5.1: 土木学会：検索精度（平均）

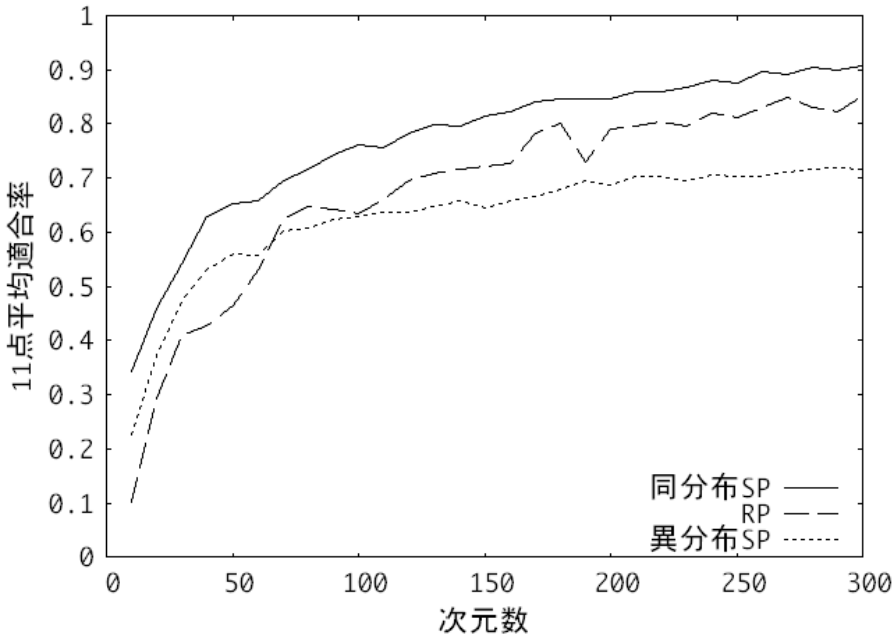

図 5.2: 計測自動制御学会：検索精度（平均）

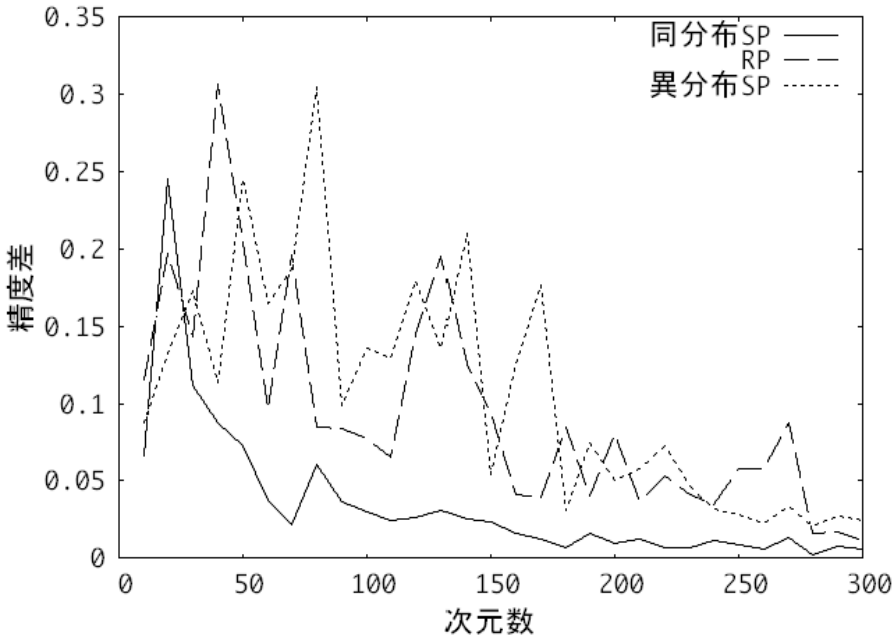

図 5.3: 土木学会：検索精度（最高値 – 最低値）

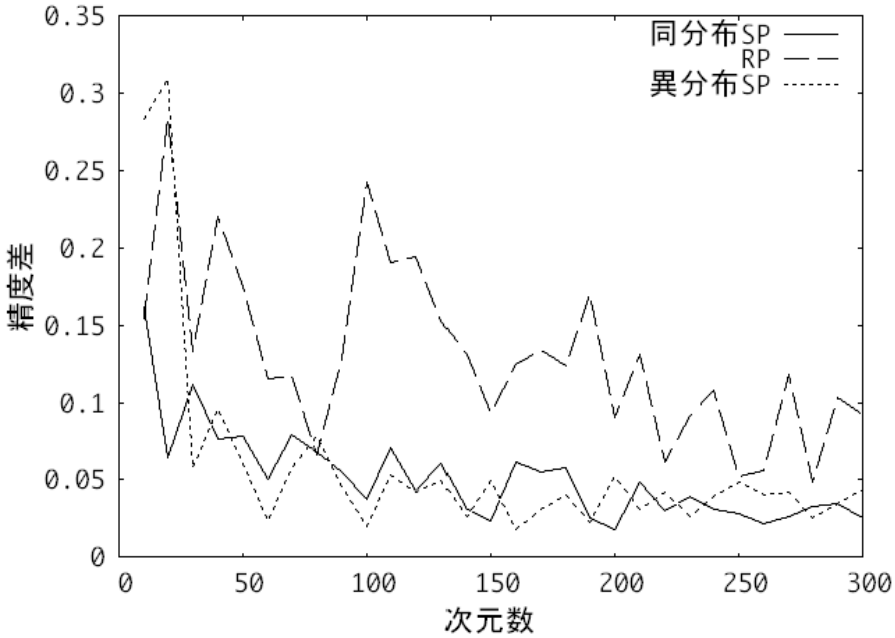

図 5.4: 計測自動制御学会：検索精度（最高値 – 最低値）

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 5.2: 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 5.3: 計測自動制御学会：同一精度内での最小次元数

5.4.4 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 手法の検索効率が際だっている．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，類似した頻度分布を持つ文書集合に対して，きわめて有効なプロジェクション手法になりうる．

5.5 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，頻度分布の違いを定量的にとらえ，同分布 SP 手法と異分布 SP 手法の境界線を明確に判断するための方法を考えていく予定である．

第6章　要約

本研究では，内容型検索のモデルであるベクトル空間モデルにおいて，ユーザの希望を文書検索に反映させるための手法について論じた．プロジェクション手法を用いることで，より高度な検索機能を同等，あるいはより優れた検索効率で実現できることを示した．

まずシソーラス辞書とプロジェクション手法を併用する文書検索について論じた．シソーラス辞書 WordNet を用いた意味関係の導入により，導入前には存在しなかった単語を新たな索引語として検索することが可能となった．LSI 手法により拡張前と同等以下に次元を縮小することで，検索効率を損なうことなく，検索処理をユーザの意図に近づける事ができた．

次に，時系列の文書データ，特にニュースストリームについて検索期間の指定をモデル化し，データ非依存の RP 手法と組み合わせる文書検索方法を論じた．ユーザによって異なる，時間を考慮した検索要求に対して，ニュースストリームの検索に対して求められるレベルの即応性を持つ，検索性能，検索時間の双方で十分に実用的な検索を行うことが可能となった．

また，局所的な非依存性を保持するプロジェクション手法である SP 手法を提案した．同一の応用分野に属する文書を検索する限りは再計算の必要がなく，RP 手法に対して高い検索効率を持つ検索が可能になった．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．対象とする応用分野を切り替えることで，ユーザの指定する検索対象を常に効率よく検索する事ができた．

表 6.1 に各プロジェクション手法の特性を示す．

	インスタンス依存	インスタンス独立
カテゴリ依存	LSI	SP
カテゴリ独立	LSI	RP

表 6.1: 各プロジェクション手法の特性

今後の課題として，文書検索時に必要となる記憶域についての議論が必要である．プロジェクション手法によって，文書ベクトルの次元は縮小し，基本的に記憶域の消費は抑えられると言える．しかし，LSI 手法においては得意値分解の処理時に大量の記憶域を消費してしまう．これら次元縮小時の議論も含め，文書検索の全体的な流れに

おいて記憶域の最大消費，平均，実用時のデータなどを取得し，定量的に考察を行う必要があると考える．定量的な考察は，特にモバイル環境など記憶域が制限されている分野での利用を論じる場合に必要不可欠である．

謝辞

本研究を遂行するにあたり，日頃より数々のご指導をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます．

また，産能大学経営情報学科 塩谷勇教授にも多くのご指導をいただきました．深く感謝いたします．

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適な研究環境の整備をして頂きました．御礼申し上げます．

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます．

最後に，今までの学生生活を支えてくださった私の両親に感謝したいと思います．

参考文献

[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] 伊藤 拓, 中西 崇文, 北川 高嗣, 清木 康: "潜在的意味抽出方式と意味の数学モデルによる意味的連想検索方式の比較", データ工学ワークショップ（*DEWS*）, 2002.

[8] F. Jiang, R. Kannan, M. L. Littman and S. Vempala: "Efficient Singular Value Decomposition via Improved Document Sampling", Technical Report CS-99-5, Department of Computer Science, Duke University, 1999

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

[10] 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.

[11] 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.

[12] Oh ’ Uchi, H. Miura, T. and Shioya, I.: “Retrieval for Text Stream by Random Projection”, *International conference on Information Systems Technology and its Applications (ISTA)*, pp. 151-164, 2004.

[13] 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

[14] 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.

[15] T.G. Rose, M. Stevenson and M. Whitehead: “The Reuters Corpus Volume 1 - from Yesterday’s News to Tomorrow’s Language Resources”, In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, 29-31 May 2002

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

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