DEWS2003 1-P-01

多義性を考慮した文書検索

大内　浩仁†　　三浦　孝夫†　　塩谷　　勇††

† 法政大学 工学部 電気電子工学科　〒 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: †{c9943020,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし　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．

キーワード　文書検索，意味質問

Document retrieval in consideration of polysemy

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

† Department of Electrical and Electronic Engineering, 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: †{c9943020,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract　In document retrieval, usually, indexing by term. In this paper, we propose to retrieval method using the polysemy of a word. We show that the Document retrieval by the meaning is attained by transposing an index to the language of a meaning using WordNet which is a lexical database. we estimate the validity is verified by experiment.

Key words　Document Retrieval, Semantic Query

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）[1], [2] 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 [3] に基づいて，小規模の文書数で実験を行う．

2 章では，LSI とベクトル空間モデルの概要を述べる．3 章では，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 章で実験を行い，5 章で結びとする．

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．

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 [1] 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．

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．

$$D = U\Sigma V^T \tag{1}$$

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．

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \tag{2}$$

とした場合，特異値は

$$\sigma_1 \geqq \sigma_2 \geqq \ldots \geqq \sigma_r > 0 \tag{3}$$

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，

$$U = [\mathbf{u_1 u_2 \ldots u_r}] \tag{4}$$

$$V = [\mathbf{v_1 v_2 \ldots v_r}] \tag{5}$$

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（1）を，

$$D = \sum_{j=1}^{r} \mathbf{u_j} \sigma_j \mathbf{v}_j^T \tag{6}$$

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）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，

$$\mathbf{t_{ki}} = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, \ldots, \sigma_k u_{ik})^T \tag{7}$$

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，

$$\mathbf{d_{kj}} = (\sigma_1 v_{j1}, \sigma_2 v_{j2}, \ldots, \sigma_k v_{jk})^T \tag{8}$$

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，

$$\hat{q}_i = \frac{1}{\sigma_i} \mathbf{q}^T \mathbf{u}_i (i = 1, 2, \ldots, k) \tag{9}$$

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. 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．

3.1 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．

3.2 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 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．

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．

4.1 実験環境

元データとして Reuters Corpus [4] 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．

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，

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

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．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．

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 1 および 2 に示す．

11 点平均適合率による結果を表 1 に示す．

	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 1　11 点平均適合率

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

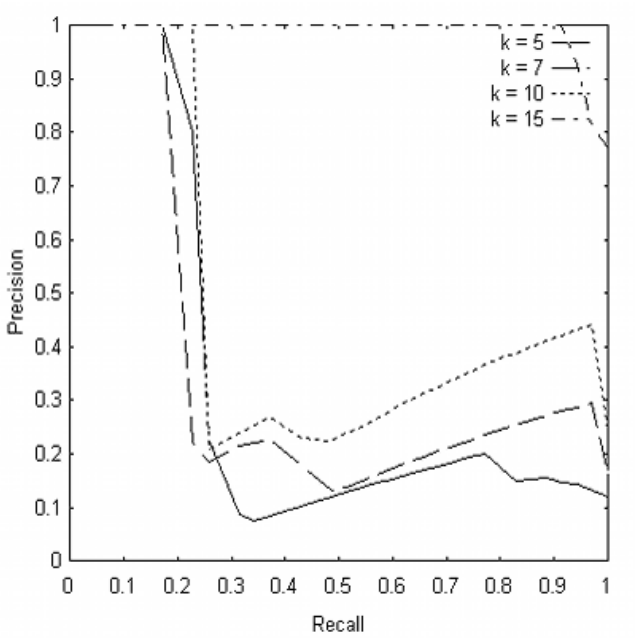

図 1　意味関係導入前

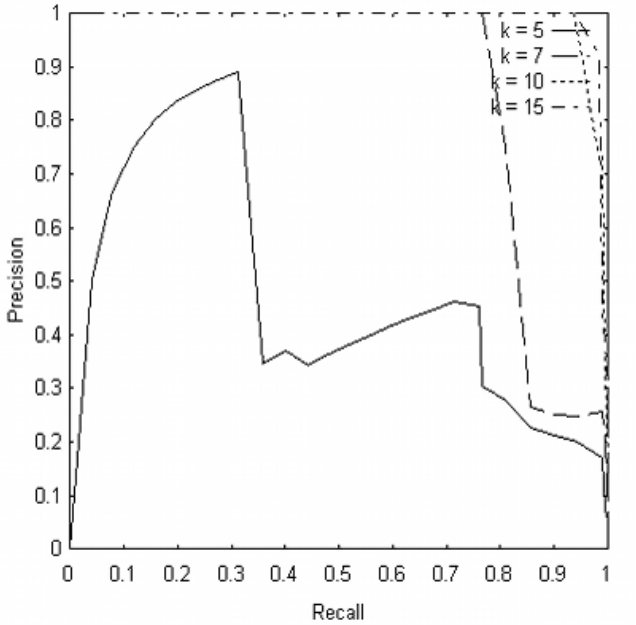

図 2　意味関係導入後

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．

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 3，4 および 5 に示す．

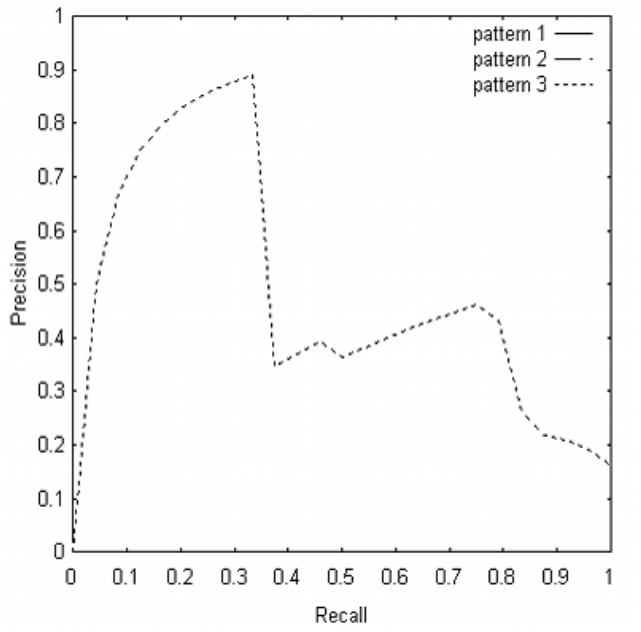

図 3　5 次元における意味検索

11 点平均適合率による結果を表 2 に示す．

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 2　11 点平均適合率

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，

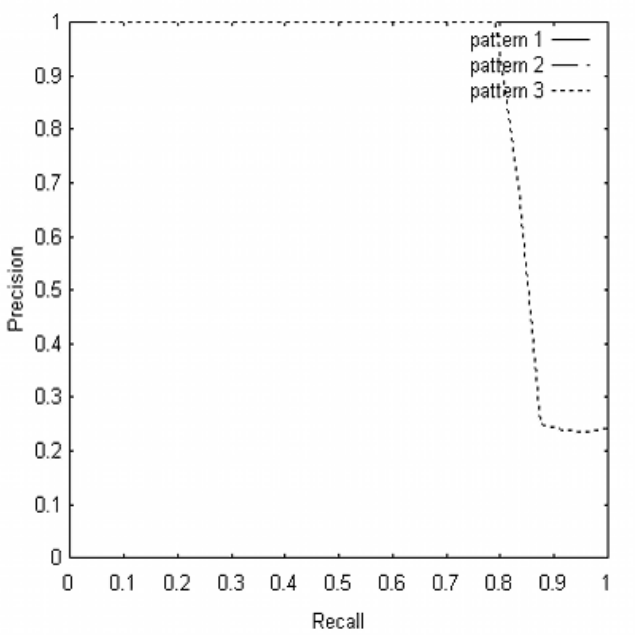

図 4　7 次元における意味検索

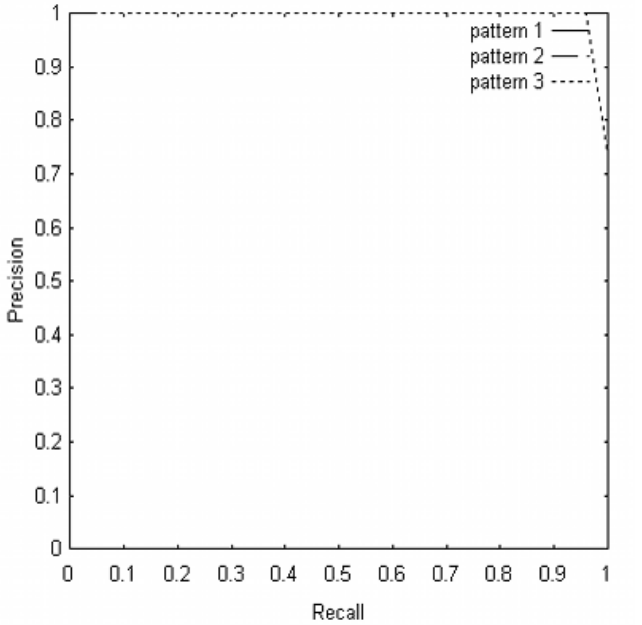

図 5　10 次元における意味検索

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．

5.　結　　び

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．

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 14580392) の支援による．

文　　献

[1] 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

[2] 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”, *DEWS*, 2002

[3] 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

[4] 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