# 技術資料 単語意味属性を使用したベクトル空間法

池原 悟[†]　　村上 仁一[†]　　木本 泰博[†]

従来，ベクトル空間法において，ベクトルの基底数を削減するため，ベクトルの基軸を変換する方法が提案されている．この方法の問題点として，計算量が多く，大規模なデータベースへの適用が困難であることが挙げられる．

これに対して，本論文では，特性ベクトルの基底として，単語の代わりに単語の意味属性（「日本語語彙大系」で規定された約 2,710 種類）を使用する方法を提案する．この方法は，意味属性間の包含関係に基づいた汎化が可能で計算コストもきわめて少なく，容易にベクトルの次元数を圧縮できることが期待される．また，単語の表記上の揺らぎに影響されず，同義語，類義語も考慮されるため，従来の単語を基底とする文書ベクトル空間法に比べて，検索漏れを減少させることが期待される．

BMIR-J2 の新聞記事検索（文書数約 5,000 件）に適用した実験結果によれば，提案した方法は，次元数の削減に強い方法であり，検索精度をあまり落とすことなく，文書ベクトルの基底数を 300〜600 程度まで削減できることが分かった．また，単語を基底とした文書ベクトルの方法と比べて高い再現率が得られることから，キーワード検索におけるKW拡張と同等の効果のあることが分かった．

**キーワード:**　情報検索，ベクトル空間法，意味解析，意味属性，汎化

# Vector Space Model bsased on Semantic Attributes of Words

Satoru Ikehara[†] and Jin'ichi Murakami[†] and Yasuhiro KIMOTO[†]

In order to reduce the dimension of VSM (Vector Space Model) for information retrieval and clustering, this paper proposes a new method, Semantic-VSM, which uses the Semantic Attribute System defined by "A-Japanese-Lexicon" instead of literal words used in conventional VSM.

The attribute system consists of a tree structure with 2,710 attributes, which includes 400 thousand literal words. Using this attribute system, the generalization of vector elements can be performed easily based on upper-lower relationships of semantic attributes, so that the dimension can easily be reduced at very low cost. Synonyms are automatically assessed through semantic attributes to improve the recall performance of retrieval systems.

Experimental results applying it to BMIR-J2 database of 5,079 newspaper articles showed that the dimension can be reduced from 2,710 to 300 or 600 with only a small degradation in performance. High recall performance was also shown compared with conventional VSM.

**KeyWords:**　*Information Retrieval, Vector Space Model, Semantic Analysis, Semantic Attribute, Generalization*

[†] 鳥取大学工学部知能情報工学科，鳥取市, Faculty of Engineering, Tottori University, Tottori-shi, 680-8552, Japan

# 1　はじめに

近年，情報化社会の進展と共に大量の電子化された文書情報の中から，自分が必要とする文書情報を効率良く検索することの必要性が高まり，従来のKW検索に加えて，全文検索，ベクトル空間法による検索，内容検索，意味的類似性検索など，さまざまな文書検索技術の研究が盛んである．その中で，文書中の単語を基底とする特性ベクトルによって文書の意味的類似性を表現するベクトル空間法は，利用者が検索要求を例文で与える方法であり，KW検索方式に比べて検索条件が具体的に表現されるため，検索精度が良い方法として注目されている．

しかし，従来のベクトル空間法は，多数の単語を基底に用いるため，類似度計算にコストがかかることや，検索要求文に含まれる単語数が少ないとベクトルがスパースになり，検索漏れが多発する恐れのあることなどが問題とされている．

これらの問題を解決するため，さまざまな研究が行われてきた．例えば，簡単な方法としては，$tf \cdot idf$ 法 (Salton and McGill 1983) などによって，文書データベース中での各単語の重要度を判定し，重要と判定された語のみをベクトルの基底に使用する方法が提案されている．また，ベクトル空間法では，ベクトルの基底に使用される単語は，互いに意味的に独立であることが仮定されているのに対して，現実の言語では，この仮定は成り立たない．そこで，基底の一次結合によって，新たに独立性の高い基底を作成すると同時に，基底数を減少させる方法として，KL 法 (Borko and Bernick 1963) や LSI 法 (Golub and Vanloan 1996),(Faloutsos and Lin 1995),(Deerwester,Dumais,Furnas,Landauer and Harshman 1990) が提案されている．

KL 法は，単語間の意味的類似性を評価する方法で，クラスタリングの結果得られた各クラスターの代表ベクトルを基底に使用する試みなどが行われている．これに対して，LSI 法は，複数の単語の背後に潜在的に存在する意味を発見しようとする方法で，具体的には，データベース内の記事の特性ベクトル全体からなるマトリックスに対して，特異値分解（SVD) の方法 (Golub and Vanloan 1996) を応用して，互いに独立性の高い基底を求めるものである．この方法は，検索精度をあまり低下させることなく基底数の削減が可能な方法として着目され，数値データベースへの適用 (Jiang,Berry,Donato and Osrtouchov 1999) も試みられている．しかし，ベクトルの基底軸を変換するための計算コストが大きいことが問題で，規模の大きいデータベースでは，あらかじめ，サンプリングによって得られた一定数の記事のみからベクトルの基底を作成する方法 (Deerwester et al. 1990) などが提案されている．このほか，単語の共起情報のスパース性の問題を避ける方法としては，擬似的なフィードバック法（２段階検索法とも呼ばれる）(Burkley,Chris,Singhl,Mitra and Salton 1996),(Kwok and Chan 1998) なども試みられている．また，ベクトルの基底とする単語の意味的関係を学習する方法としては，従来から，Mining Term Association と呼ばれる方法があり，最近，インターネット文書から体系的な知識を抽出するのに応用されている (Lin,Shih,Chen,Ho and Ko 1998)．しかし，現実には，単語間の意味的関係を自動的に精度良く決定することは容易でない．

これに対して，本論文では，ベクトル空間法において，検索精度をあまり低下させることなく，基底数を容易に削減できることを期待して，単語の意味属性をベクトルの基底として使用

する方法を提案する．この方法は，従来の特性ベクトルにおいて基底に使用されている単語を，その意味属性に置き換えるものである．単語意味属性としては，日本語語彙大系 (池原, 宮崎, 白井, 横尾, 小倉, 中岩, 大山, 林 1997) に定義された意味属性体系を使用する．この意味属性体系は，日本語の名詞の意味的用法を約 2,710 種類に分類したもので，属性間の意味的関係（is-a 関係と has-a 関係）が 12 段の木構造によって表現されている．また，日本語の単語 30 万語に対して，どの意味属性（１つ以上）に属す単語であるかが指定されている．従って，本方式では，意味属性相互の意味的上下関係を利用すれば，検索精度をあまり落とさずにベクトルの基底数を削減できる．同時に基底として使用すべき必要最低限の意味属性の組を容易に決定できることが期待される．また，本方式では，検索要求文に使用された単語とデータベース内の記事中の単語の意味的な類似性が，単語意味属性を介して評価されるため，再現率の向上が期待できる．すなわち，従来の単語を基底とした文書ベクトル空間法では，ベクトルの基底として使用された単語間のみでの一致性が評価されるのに対して，本方式では，すべての単語（30 万語）が検索に寄与するため，検索漏れの防止に役立つと期待される．

本論文では，TREC に登録された情報検索テストコレクション BMIR-J2(木谷他 1998) を検索対象とした検索実験によって，従来の単語を用いた文書ベクトル空間法と比較し，本方式の有効性を評価する．

# 2　意味属性体系を基底とした文書ベクトル空間法

## 2.1　単語を基底とした文書ベクトル空間法 (W-VSM)

従来の単語を基底とした文書ベクトル空間法では，文もしくは文書の意味的類似性はその中に出現した単語の組で表現されるものと仮定している．すなわち，文書の意味的類似性を表現するために使用される単語の番号を $i$ $(1 \leq \mathrm{i} \leq n)$ とし，文書中での単語 $i$ の重みを $w_i$ とするとき，文書は，以下のような特性ベクトルで表わされる．

$$V = (w_1, w_2, \cdots, w_i, \cdots, w_n) \tag{1}$$

ベクトルの基底とすべき単語としては，キーワード検索の場合と同様，データベース全体に使用された単語の出現統計から，$tf \cdot idf$ 値などによって重要と判断された単語を通常使用している．また，重み $w_i$ の値としては，文中に単語 $i$ が使用されているときは 1，使用されていないときは 0 とする方法と，文中に使用された単語の出現頻度とする方法がある．また，各文書全体の相対的重みはいずれも等しいとする立場から，ベクトルの絶対値が 1 となるよう正規化する方法も採られている．本論文では以後，式 1 で与えられる特性ベクトルを「単語を基底とした文書ベクトル」と呼び，このベクトルを使用したベクトル空間法を「単語を基底とした文書ベクトル空間法 W-VSM（Word-Vector Space Model)」と呼ぶ．

## 2.2　単語を基底とした文書ベクトル空間法における意味的類似度

単語を基底とした文書ベクトル空間法において．文書の意味類似度を特性ベクトルで表現したとき，異なる文書 $D_i$，$D_j$ 間の意味的類似性 $sim(D_i, D_j)$ は，それぞれの文書に対して求めた特性ベクトルの内積として，式 2 のように表現される．

$$sim(D_i, D_j) = V_i \cdot V_j \tag{2}$$

但し，$V_i \cdot V_j$ は，それぞれ，文書 $D_i$，$D_j$ の特性ベクトルを表す．

従って，単語を基底とした文書ベクトル空間法を用いた情報検索では，利用者の与えた検索要求文について特性ベクトルを求めて，データベースに収録された各文書の特性ベクトルとの間で類似度を計算し，類似度がある一定値以上の文書を抽出している．また，単語を基底とした文書ベクトル空間法では，任意の文書をつなぎ合わせた文書についての特性ベクトルも容易に合成できるから，類似度の高い文書相互間で順にベクトル合成を行えば，文書全体を容易にクラスタリングすることができる．

## 2.3　単語意味属性を基底とした文書ベクトル空間法 (S-VSM)

本論文では，単語の代わりに，その単語の意味属性を使用する方法を提案する．本方式では，すべての単語を $k$ 個の意味属性に分類したのち，分類された意味属性を要素とする特性ベクトルによって文書の意味的類似性を表現する．すなわち，対象とする文書 $D_j$ において $i$ 番目の意味属性を持つ単語全体の重み $S_i$ とするとき，文書 $D_j$ の特性ベクトル $V_j$ は，次式で表現される．

$$V_j = (S_1, S_2, \cdots, S_i, \cdots, S_k) \tag{3}$$

重み $S_i$ の与え方としては，種々の方法が考えられるが，本論文では，単語を基底とした文書ベクトル空間法の場合と同様，$tf \cdot idf$ 法の考えを適用し，以下の方法で得られた値とする．

1. データベースに収録された文書全体に対して，意味属性 $S_i$ に属す単語が出現した頻度の合計を求め，それぞれの $idf$ 値を計算する．
2. 文書 $D_j$ を対象に，意味属性 $S_i$ に属す単語が出現した頻度の合計を求め，その値を文書 $D_j$ の $tf$ 値とする．
3. 上記で得られた $tf$ 値と $idf$ 値から，意味属性 $S_i$ の $tf \cdot idf$ 値を求める．
4. 上記で得られた $tf \cdot idf$ 値を $|V_j| = 1$ となるように正規化する．

なお，式 1 で与えられる特性ベクトルを「単語を基底とした文書ベクトル」と呼んだのに対して，以下では，式 3 で与えられる特性ベクトルを「単語意味属性を基底とした文書ベクトル」と呼び，このベクトルを使用したベクトル空間法を「単語意味属性を基底とした文書ベクトル空間法 S-VSM(Semantic-Vector Space Model)」と呼ぶ．

## 2.4　日本語単語の意味属性体系

単語の代わりに意味属性を基底とする文書ベクトル空間法では，単語の意味属性についての分類体系が必要である．本論文では，意味分類体系として，最近,「日本語語彙大系」(池原他 1997) で提案された日本語名詞の意味属性体系を使用する．図 1 に意味属性体系の一部を示す.

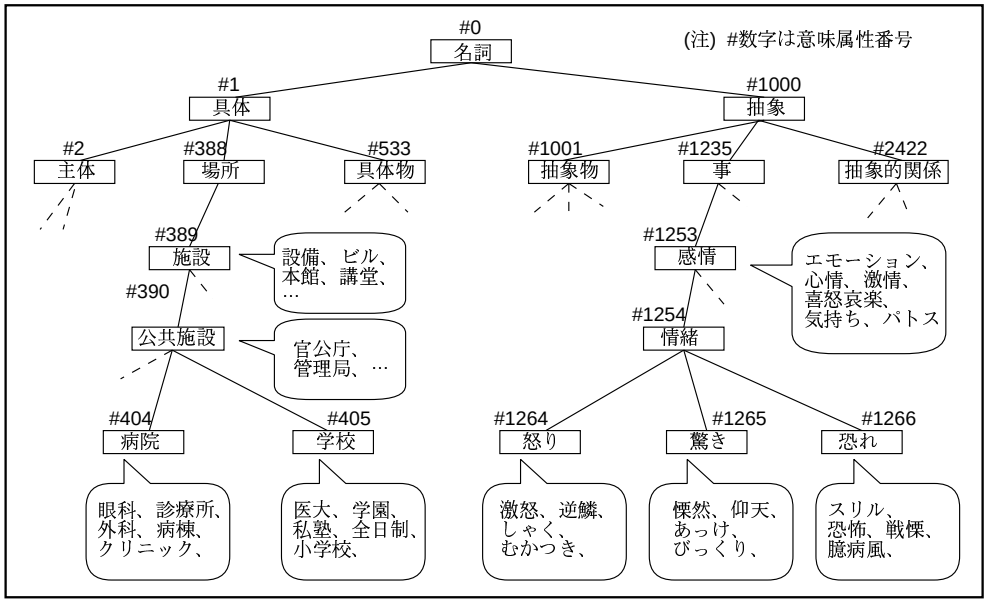


図 1: 一般名詞意味属性体系の一部

Fig. 1: Portion of General Noun Semantic Attributes System

この意味属性体系は，日本語名詞の意味的な用法を 2,710 種類の意味属性に分類したもので，意味属性間の意味的関係（is-a 関係，has-a 関係）が，12 段の木構造で表現されている．また，単語意味辞書では，日本語名詞 30 万語のそれぞれが，どのような意味属性を持つか（一つ以上）が規定されている．従って，文書中に使用された名詞の出現頻度が分かれば，3 式のベクトルの要素 $S_i$ は，$i$ 番目の意味属性を持つ名詞の出現頻度から 2.3 章で述べた方法で容易に求めることができる.

## 2.5　単語意味属性を基底とした文書ベクトルの効果

情報検索において，従来の単語を基底とした文書ベクトル空間法 W-VSM に比べて，単語意味属性を基底とする文書ベクトル空間法 S-VSM が，どのような効果を持つかについて考察する.

1. ベクトルの基底数削減の可能性
   従来の単語を基底とした文書ベクトル空間法では，ベクトルの基底として使用される名詞の意味は，互いに独立であることが仮定されているが，現実にはこの仮定は成り立たない．そのため，ベクトルの基底数を減少させるため，従来，基底をクラスタリングで得られたクラスターのベクトルとしたり，特異値分解 (SVD: Singular Value Decomposition) によって得られたベクトルに変換する方法の研究 (Deerwester et al. 1990) が行われてきた．しかし，これらの方法は，ベクトルの変換に多くのコストを要する点が問題であった．
   これに対して，本論文で基底として使用する単語意味属性は，木構造によって意味的上下関係（is-a 関係と has-a 関係）が規定されている（2.4 節参照）．この関係を利用して基底数を削減するため，計算コストはきわめて小さい．また，あまり効果のない意味属性を上位の意味属性で代用できるので，削減された意味属性も検索精度に寄与できるため，従来の方法と同様，検索精度をあまり落とすことなく，基底数が削減できると期待される．
2. 検索漏れの減少の可能性
   従来の単語を基底とした文書ベクトル空間法では，文書中に出現した単語のうち，ベクトルの基底として選択された単語のみがその文書の意味に反映する．そのため，意味が同じであっても，表記が異なる語は別の語として判定される．また，同義語や類義語を含む文書であっても，それが基底として採用されない限り検索の対象とならない．
   これに対して，単語意味属性を基底とした文書ベクトル空間法では，2.3，2.4 節で述べたように，30 万語の名詞が 2,710 の意味属性にマッピングされ，検索要求文に使用された単語とデータベース内の記事中の単語の意味的な類似性が，単語意味属性を介して評価される．すなわち，文書中に使用される語は，それが異表記語，同意語，同義語のいずれであっても，その意味が特性ベクトルに反映するため，情報検索において，検索漏れの削減の効果が期待できる．
3. 適合率の低下
   単語意味属性を基底とした文書ベクトル空間法では，1 つの単語に対して意味属性による検索をおこなうため，複数の単語を検索するのと等価になる．そのため適合率の低下が予想される．

# 3　必要最小限の意味属性の決定

本論文では，2 章で述べた単語意味属性を基底とした文書ベクトルの効果を評価するため，日本語語彙大系で定義された意味属性 2,710 種類のすべてを使用する場合と，その中から必要最小限と見られる意味属性を選択して使用する場合について検索精度を調べる．本章では，意味属性の上下関係に着目した汎化により，ベクトルの基底として使用すべき必要最小限の意味属性の組を発見する方法について述べる．

## 3.1　汎化の方法

汎化とは，モデル学習において，事例から規則を発見するための帰納的推論の一種である．ここでは，特性ベクトルの基底数を減少させるため，情報検索に効果が少ないと推定される意味属性を直属上位の意味属性に縮退させることを汎化と呼ぶ．本論文では，汎化によって基底から削除された意味属性の $tf \cdot idf$ 値 は，その上位の意味属性の $tf \cdot idf$ 値に加えることとする．汎化の対象とする意味属性の選び方については，様々な方法が考えられるが，ここでは，意味属性の粒度と意味属性の $tf \cdot idf$ 値に着目する方法を考える．

### 3.1.1　粒度による汎化　S-VSM(g)

ベクトルの基底に使用される意味属性は，12 段の木構造からなり，下位になるほど意味の粒度が相対的に小さくなる．そこで，各意味属性の位置する段数を粒度と考え，ある一定の粒度より小さい意味属性を汎化する．図 2 に，8 段以下の意味属性を 7 段目の意味属性に汎化する場合の例を示す．

### 3.1.2　$tf \cdot idf$ 値による汎化　S-VSM(w)

検索対象となるデータベースの文書全体での $tf \cdot idf$ 値の小さい意味属性は，検索に寄与する程度が小さいと考えられるため， $tf \cdot idf$ 値の小さい意味属性を汎化の対象とする．汎化によって削除された意味属性の $tf \cdot idf$ 値は，上位直属の意味属性の $tf \cdot idf$ 値に加算する．直属の意味属性が削除されているときは，さらに上位の意味属性の $tf \cdot idf$ 値に加算する．図 2 に，$tf \cdot idf$ 値が 5 以下の意味属性を汎化する場合の例を示す．

## 3.2　必要最小限の意味属性の決定

ベクトル空間法では，計算量を削減する観点から，ベクトルの基底数を減少させることが望まれる．しかし，多くの場合，検索精度は低下させずに基底数を削減することは困難である．そこで，前節で述べた汎化の方法を使用し，検索精度をある一定値以上低下させない範囲で，必要最小限の意味属性の組を求める方法を考える．

### 3.2.1　粒度による汎化　S-VSM(g)

元来，特性ベクトルで表現される文書の意味の粒度は，ベクトルの基底に単語そのものを使用する場合が最も細かい．意味属性を使用する方法では，すでに意味的な汎化が行われており，意味の粒度は荒くなっている．粒度に着目した汎化がさらに進めば検索精度は次第に低下すると考えられるため，必要最小限の意味属性の組を発見するには，順次，汎化を進めながら，検

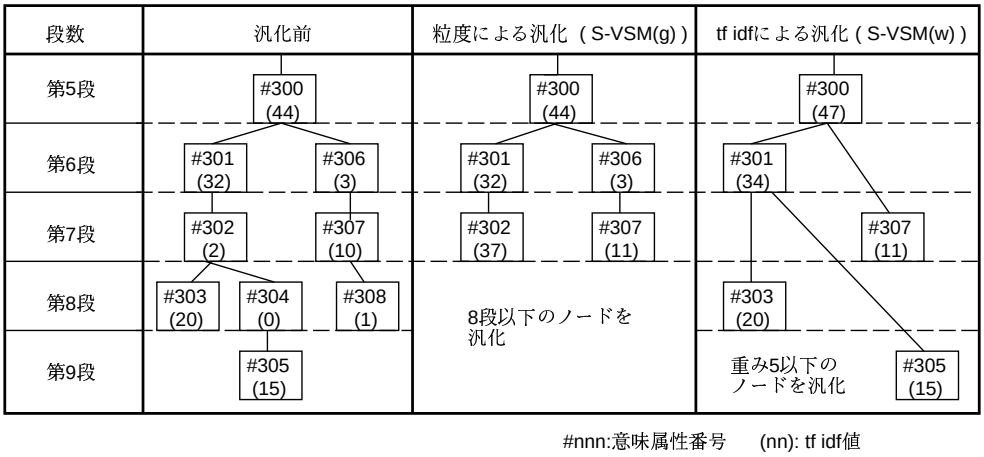


図 2: 汎化の方法

Fig. 2: Generalization Methods

索精度の変化を追跡する必要がある．その結果，検索精度が低下する直前に使用した意味属性の組を必要最小限の組とする．

### 3.2.2 $tf \cdot idf$ 値による汎化　S-VSM(w)

1. 基本的な考え方
   データベース中で $tf \cdot idf$ 値の小さい意味属性が汎化の対象となる．しかし，必ずしも，$tf \cdot idf$ 値の小さい意味属性のすべてを汎化すればよいとは限らない．いま，データベース内に収録された文書が検索対象となる確率はすべて均等だとし，すべての文書を対象に求めた特性ベクトルの和を $V_t$ とする．$V_t$ 要素 $n_i$ の値の小さい意味属性 $\#i$ は，検索精度に与える影響が少ないから，情報検索において少ないベクトルの基底数で高い検索精度を得るには，各 $n_i$ の値がバランスしていることが必要である．すなわち，$tf \cdot idf$ 値の低い意味属性でも，基底間でアンバランスが増大するような汎化は，検索精度低下の原因となるから，高い検索精度を得るためには，データベース内の文書全体で出現する $tf \cdot idf$ 値がバランスするような意味属性を特性ベクトルの基底に選定する必要がある．
2. 汎化すべき意味属性の選択基準
   汎化すべき意味属性の選択基準について考える．データベース内に収録された文書全体の特性ベクトルを式 4 とする．

$$V_t = (n_1, n_2, \cdots, n_i, \cdots, n_m) \tag{4}$$

   ただし，$n_i$ は，意味属性 $\#i$ に属す単語のデータベース全体での $tf \cdot idf$ 値の和を，また，

$m$ は，基底に使用される意味属性の数を示す．ここで，各 $n_i$ の値の均等さを変動によって評価するとし，評価関数 $H$ を以下のように定義する．

$$H = (n_1 - n)^2 + (n_2 - n)^2 + \cdots + (n_i - n)^2 + \cdots + (n_m - n)^2 \tag{5}$$

但し $n$ は $n_i$ の平均値とする．

$$n = \sum_{i=1}^{m} \frac{n_i}{m} \tag{6}$$

基底のバランスを向上させるには，$H$ の値が，減少するような基底（意味属性 $\#i$ ）を選んで汎化を行う．そこで，意味属性 $\#i$ を汎化することを考える．$\#i$ の直属上位の意味属性の番号を $\#j$ とすると，汎化では，$n_i$ の値が $n_j$ に加算され，基底数 $m$ が 1 だけ減少する．従って，このようにして得られた $H$ の値を $H_1$ とすると，$H$ と $H_1$ の差は，近似的に[1]式 7 が得られる．

$$H - H_1 \simeq (n_i - n)^2 + (n_j - n)^2 - (n_i + n_j - n)^2 \tag{7}$$

ここで，条件から，$H - H_1 > 0$ とおくと，式 8 が得られる．

$$n_i \cdot n_j < n^2/2 \tag{8}$$

以上から，汎化すべき基底は，その重 $tf \cdot idf$ 値と直属上位の基底の $tf \cdot idf$ 値との積が，基底の平均値の二乗値の半分以下になるものを選択する．

3. 汎化の手順

具体的には，以下の手順で汎化を行う．

(a) 汎化

上下関係にある意味属性 $n_i$ , $n_j$ のすべての組のうち，積が最も小さい組を汎化する．

(b) 検索

情報検索実験を行い，検索精度を求める．

(c) 停止

検索精度の低下がある閾値以下の値のときは (a) に戻り，それ以上の時は，汎化を停止する．

## 3.3 ベクトル変換のための計算コスト

2 節で述べた汎化は，基本となるベクトルの軸を変換する点では，従来の KL 法や LSI 法と同様である．そこで，そのために必要な計算コストを比較する．まず，ベクトルの基底数を削減するのに要するコストについて考える．

[1] $H_1$ の平均 $n'$ は $n' = \frac{m}{m-1}n$ となる．ここで $m >> 1$ とすると $\frac{m}{m-1} \simeq 1$ から $n' \simeq n$ となる．

データベースに収録された文書の総数と削減前のベクトルの基底数の和を $N$，削減後のベクトル基底数を $k$ とすると，単語を基底とした文書ベクトル空間法の場合，通常，計算量は $N^4$ もしくは $N^5$ に比例すると言われている．LSI 方式でも，特異値分解に必要な計算量は，$N^2 \cdot k^3$ に比例する．このため，データベースの規模が増大すると急激に計算量が増大することが大きな問題であった．

[2]

これに対して，使用される意味属性の総数を $M$，段数を $d$（日本語語彙大系の場合 $M = 2,710, d = 10$ ）とすると，単語意味属性を基底とした文書ベクトルにおいて粒度による汎化を行うときは，必要最小限の意味属性の数を求めるための計算コストは，ほぼ，$M \cdot d$ に比例する．また $tf \cdot idf$ 値による汎化の場合は，ほぼ，$M^2 - k^2$ に比例する．また，必要最小限の意味属性の組が決定した後，文書毎の特性ベクトルを変換することは容易で，その計算コストは，文書量に比例する．

# 4　実験

本章では，情報検索の精度と必要最小限の意味属性の組に関する実験を行い，提案した方式の特徴を評価する．

## 4.1　使用する文書

実験には，TREC に登録された「情報検索評価用テストコレクション BMIR-J2」(木谷他 1998) (以下 BMIR-J2) を利用する．BMIR-J2 は，1994 年の毎日新聞より国際十進分類（UDC）で経済，工学，工業技術一般に分類される記事 5,080 件を対象とするもので，文書集合，検索要求，正解判定結果から構成される．検索要求は「～ に関する記事が欲しい」という形式で統一され,「～」の部分にあたる名詞句が列挙されている．また，検索要求に対する正解として，下記の通り，2 種類の記事が示されている．

1.　ランクA
　検索要求を主題としている記事
2.　ランクB
　検索要求の内容を少しでも含む記事

[2] なお，大規模疎行列の固有値計算アルゴリズムは Krylov 部分空間法の一種である Lanczos 法を用いて高速に解くことができる．この方法は，一定次元の部分空間における近似固有ベクトルをもとに新たに初期ベクトルを計算し, 反復法として用いることによって記憶容量を低減させる．反復 Lanczos 法は，特に疎行列を扱う場合に実際的な解法であるといえるが，固有値が近接している場合，正確な計算が難しいことが知られている．

## 4.2 評価のパラメータ

実験結果は，以下の 4 つのパラメータを用いて評価する．

1. $sim$ :文書類似度

$$sim(D_i, D_j) = V_i \cdot V_j \tag{9}$$

（但し，$V_i \cdot V_j$ は，それぞれ，文書 $D_i, D_j$ の特性ベクトル）

2. $R$ : 再現率 (recall factor)

$$R = \frac{\text{抽出された正解文書数}}{\text{データベース中の正解文書数}} \tag{10}$$

3. $P$ : 適合率 (precision factor)

$$P = \frac{\text{抽出された正解文書数}}{\text{抽出された文書数}} \tag{11}$$

4. $F$ : 検索精度 (f-parameter)

$$F = \frac{(b^2+1) \cdot P \cdot R}{b^2 \cdot P + R} \tag{12}$$

但し，式 12 のパラメータ $b$ は，$P$ に対する $R$ の相対的な重みを示す．実験では，両者を対等と考え，$b = 1$ とする．

## 4.3 実験の方法

検索要求として新聞記事が与えられたとき，類似した新聞記事を検索することを考え,「主題が一致している新聞記事」を正解とする．具体的には，主題が一致している記事（ランク A）のうちの 1 つを検索要求用の記事に使用し，データベース内に収録された 5,079 件の記事の中から残りのランク A の記事を検索する．検索要求用の記事を替えながら，この手順を 90 回繰り返し，平均の検索精度で評価する．従来の単語を基底とした文書ベクトル空間法による実験では，データベース記事全体を対象に使用されている名詞の $tf \cdot idf$ 値を求め，その値の大きい順に基底とする名詞を決定する．また，基底毎の重要度を考慮し，各単語ベクトルの要素の値には，単語の文書中での出現頻度に $idf$ 値を掛けた値を使用する．なお，情報検索では，ある一定値以上の類似度を持つ文書を抽出の対象とするが，その値の選び方によって，再現率，適合率の値は変化する．そこで，検索の精度評価では，いずれの場合も，$F$ 値が最大となるよう類似度を設定する．

# 5　実験結果

## 5.1　単語意味属性を基底とした文書ベクトル (S ー VSM) と単語を基底とした文書ベクトル空間法（W-VSM）の比較

2,710 種類の意味属性のすべてを使用する場合について情報検索実験を行い，従来の単語を基底とした文書ベクトル空間法（W-VSM）と検索精度を比較する．

本論文の方法による検索精度を従来の単語を基底とした文書ベクトル空間法と比べた結果を図 3 に示す．図 3 では，情報検索において類似度 $\alpha$ 以上の文書を抽出した場合について，$\alpha$ と再現率 $R$，適合率 $P$ の関係を示している．なお，類似度 0.7 以上とする場合は，検索される文書が 1 件程度となってしまい，信頼できないので，グラフから削除した．

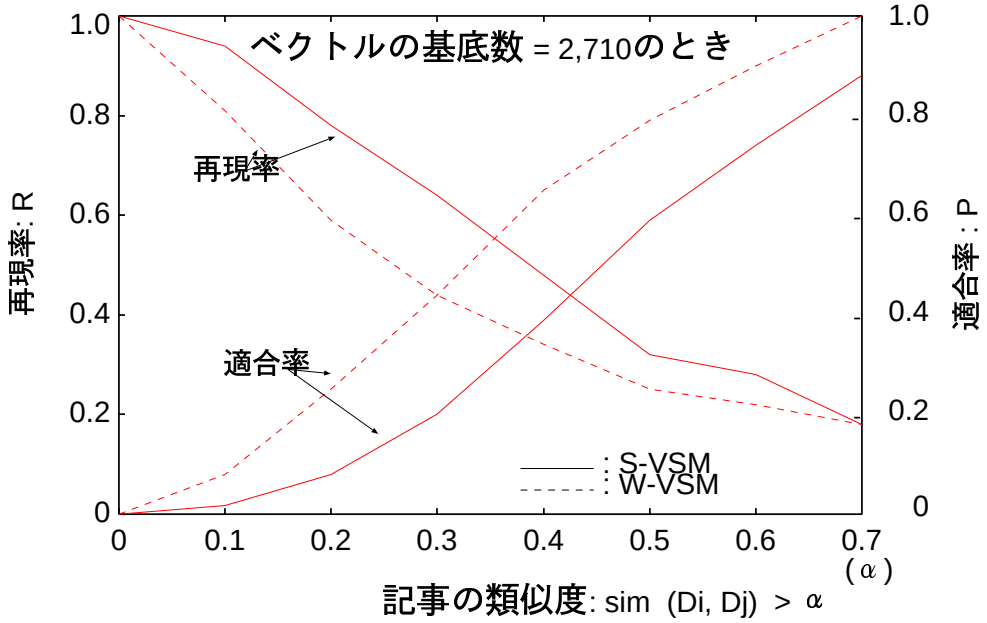


図 3: 記事の類似度と検索の精度の関係

Fig. 3: Similarity and Performance of Information Retrieval

また，この結果から得られた類似度 $sim$ と検索精度 $F$ 値の関係を図 4 に示す．

これらの図から，以下のことが分かる．

1. 単語意味属性を基底とした文書ベクトルは，単語を基底とした文書ベクトル空間法に比べて，すべての類似度領域で，再現率が高く，適合率が低い．
2. 検索精度（$F$ 値の最大値）は，両者は殆ど変わらない．

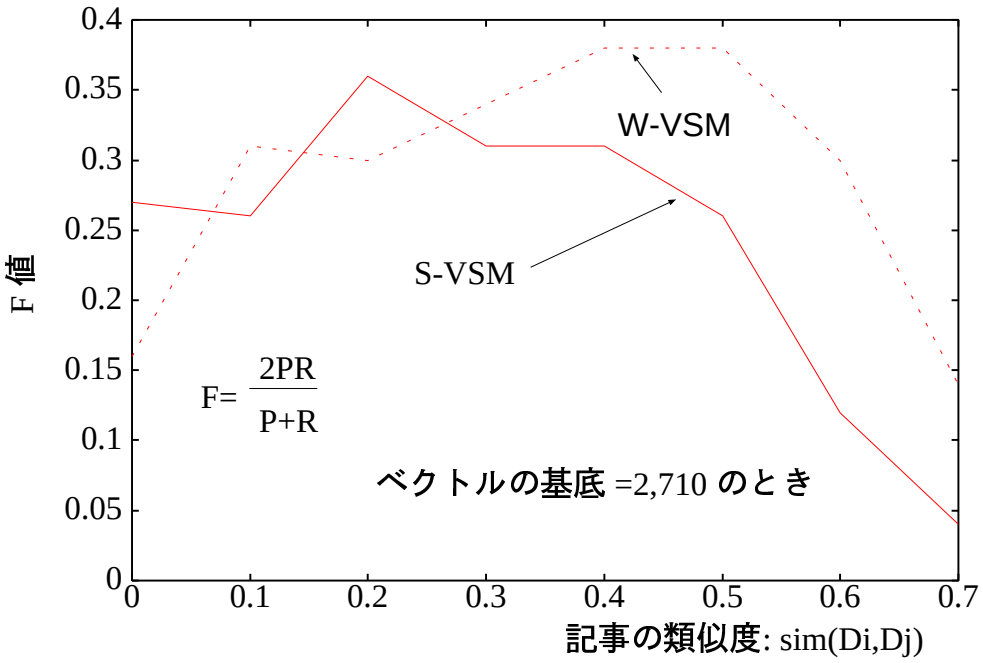


図 4: 記事の類似度と検索の精度の関係

Fig. 4: Similarity and F value

## 5.2 粒度による汎化（S-VSM(g)）と $tf \cdot idf$ 値による汎化（S-VSM(w)）の比較

3.2 節で述べたような，意味属性の粒度に着目する汎化（S-VSM(g)）と意味属性の $tf \cdot idf$ 値に着目する汎化（S-VSM(w)）の 2 つの汎化の方法を用いて，ベクトルの基底として使用する意味属性の数と検索精度の関係を求めた．その結果を図 5 に示す．また，このうち，意味属性の $tf \cdot idf$ 値による汎化の場合について，汎化に伴う評価関数 $H$ の値の変化を同図に示す．なお，ここでは，$b = 1$ とした．

図 5 の結果から，検索精度をあまり低下させない範囲（ピーク値の 10 ～ 20% 以内の低下）で必要最小限のベクトルの基底数を求めると表 1 の結果を得る．

これらの図表から，以下のことが示される．

1. 今回の実験では，単語意味属性を基底とする文書ベクトル空間法は，従来の単語を基底とする文書ベクトル空間法に比べて，基底数が小さくても検索精度が高いことが示された．
2. 汎化の方法としては，粒度による汎化（S-VSM(g)）より $tf \cdot idf$ 値による汎化（S-VSM(w)）の方が基底数削減に強い．

必要最小限の基底数について見ると，十分な基底数を持つ場合に比べて，検索精度を 10 ～

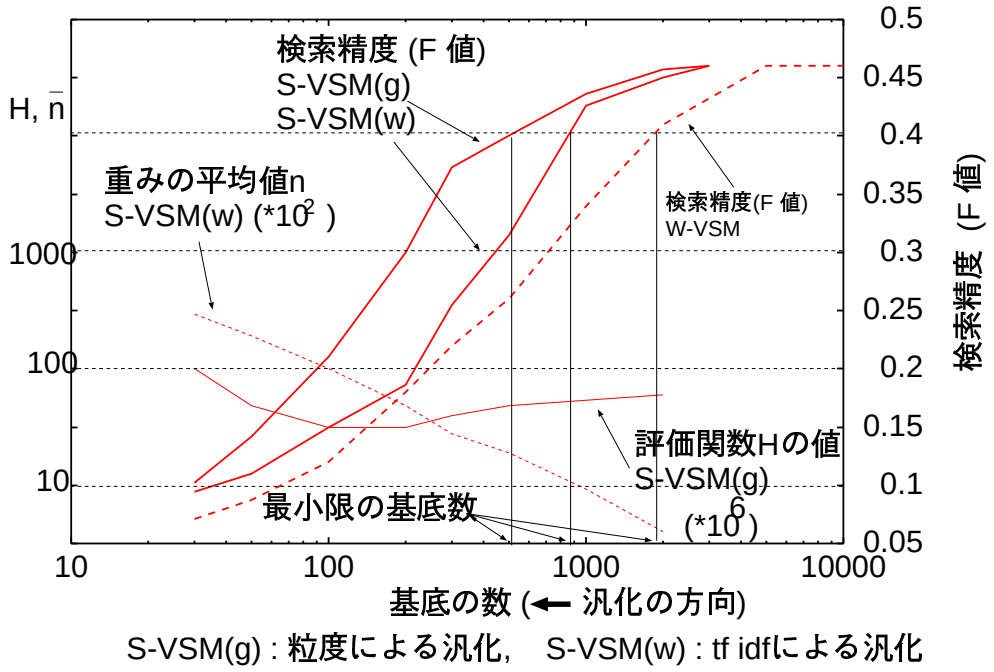


図 5: 必要最小限の基底数の決定

Fig. 5: Determination of Minimum Number of Vector Bases

表 1: 必要最小限の基底数

Table 1: Minimum Number of Vector Bases

| 方式種別 | 基底数削減の方法 | 検索精度 ($F$ 値) 低下の許容度 | |
|---|---|---|---|
| | | ピーク値の 10% | ピーク値の 20% |
| 本論文の方法 (単語意味属性を基底) | 粒度による汎化 ( W-VSM(g) ) | 900 属性 | 700 属性 |
| | $tf \cdot idf$ 値による汎化 ( W-VSM(w) ) | 600 属性 | 300 属性 |
| 従来の方法 (単語を基底) | $tf \cdot idf$ による方法 | 2,200 属性 | 1,500 属性 |

(注) 意味属性を上位 8 段まで使用

20% 以上低下させないためには，単語を基底とする文書ベクトル法では，最低 2,000 程度の基底数が必要とされるのに対して，単語意味属性ベクトルを用いて，$tf \cdot idf$ 値による汎化では，基底数を約 300 ～ 600 程度まで削減できる．

# 6 考察

## 6.1 単語意味属性を基底とする文書ベクトル空間法と単語を基底とする文書ベクトル空間法の比較

実験によれば，単語意味属性を基底とする文書ベクトルは，単語を基底とする文書ベクトル空間法に比べて，再現率が高いことが分かった．本研究では，簡単のため，文書中に使用された単語の頻度から直接，意味属性の $tf \cdot idf$ 値を求めることとし，複数の意味を持つ単語は，その $tf \cdot idf$ 値 を，該当する複数の意味属性に均等に加える方法を採った．これは，単語を基底とする文書ベクトルの場合と同じ扱いであるが，適合率を減少させる原因の一つと考えられる．これに対して，文書中で使用された単語の多義解消を行うことができれば，適合率の向上は可能であると期待される．

ただし，今回の実験は，BMIR-J2 における新聞記事検索のタスクであり，文書数も約 5,000 件と少ない．今後検索する分野が変化したときや，文章数が増加した場合，これらの結論が変わってくる可能性がある．今後，これらの課題を追求する必要がある．

## 6.2 意味属性体系

本研究に使用した意味属性体系は，元来，単語多義の解消を狙って開発されたものであり，複数の語義を持つ単語は，通常，複数の意味属性を持つ構造となっている．日本語語彙大系には，さらに，動詞と名詞の共起関係から，両者の文中での意味を特定するための仕組みが定義されている．そこで，これらの情報を使用した意味解析によって文書中で使用された単語の意味的用法を決定し，その後，該当する意味の重みを求めることにすれば，質問文と同じ単語が使用された文書でも意味の異なる用法の文書は検索対象外とすることができるため，適合率は向上すると期待される．

## 6.3 基底数の削減のためのテストデータ

実験では，提案した単語意味属性を基底とした文書ベクトル空間法と従来の単語を基底とした文書ベクトル空間法が基底数削減にどれだけ強いかを比較評価するため，情報検索方式の評価実験用として広く提供されている BMIR のデータセット（検索条件と正解付き）を使用した．実験はいずれもオープンテストである．これは，以下に述べるように，この種の研究では大量のデータを対象としたオープンテストは困難なためである．

すなわち，本手法では，検索対象とするデータベースに対して必要最小限の意味属性の組を発見することが必要であるが，そのためには，汎化を進める過程で検索精度が低下するかどうかの評価が必要で，検索結果についてあらかじめ正解を知っておく必要がある．しかし，大規模なデータベースの場合，様々な検索条件について，あらかじめ正しい検索結果を知ることは

通常難しい（この事情は他の検索方式の場合も同様である）.

ところで，本方式を現実のシステムに応用するには，部分的な標本（例えば，数千件程度の記事）に対して今回と同様の実験により必要最小限の意味属性の組決める必要があるが，必要な意味属性の数（これを $n$ 個とする）が分かれば，$n$ 個を構成する意味属性の種類は，データベースの規模に応じてさらに最適化することができる．すなわち，大規模なデータベースでも単語の出現頻度統計を取るのは比較的容易であるから，単語統計から作成された意味属性を初期値とし，意味属性数が $n$ となるまで汎化すれば，残った $n$ 個の意味属性は，データベース全体から見て最適な組み合わせとなり，運用段階においてもクローズドテストに近い検索精度が得られるものと期待できる.

### 6.4　必要最小限の意味属性

粒度による汎化（S-VSM(g) において文書ベクトル数を 700 に汎化したときに残った単語意味属性を調査した．この結果．汎化で残った単語意味属性の多くは，汎化をする前に $tf \cdot idf$ 値が大きく，かつ頻度も多い単語意味属性であった．例として「抽象」,「名詞」,「事」など意味意味属性であった.

### 6.5　$tf \cdot idf$ 値による汎化 と 頻度による汎化

2 節において，必要最小限の意味属性の決定するために，粒度による汎化 (S-VSM(g)) と $tf \cdot idf$ 値による汎化 (S-VSM(w)) を示した．本論文でしめした両方法は，どちらも $tf \cdot idf$ 値を利用しているが，単語の出現頻度を利用する方法も考えられる．そこで 4 節の実験の前に，予備実験として，出現頻度を利用する場合と $tf \cdot idf$ 値を利用する場合で $F$ 値がどちらが高くなるか調査した．この結果，$tf \cdot idf$ 値を利用する場合のほうが良い値を示した．そのため，以後の実験においては $tf \cdot idf$ 値を利用した.

なお，頻度が大きいが $tf \cdot idf$ 値が小さくなる単語意味属性は,,「自尊，卑下」,「敬称 (女)」,「自信，自棄」,「生産行程」,「自信」などであった．また，比較的頻度が小さいが $tf \cdot idf$ 値が大きくなる単語意味属性は,「乗り物」,「親、祖父母，先祖」,「親」,「報償」,「庭園」,「休養」,「余暇」などであった.

## 7　結論

従来，ベクトル空間法では，文書の意味を表す特性ベクトルの基底に，文中に現れる単語を使用していた．本論文では，単語の代わりに単語の意味属性（「日本語語彙大系」で規定された約 2,710 件）を使用する方法を提案した．また，意味属性間の意味的上下関係に着目したベクトルの基底の汎化の方法を提案し，情報検索の精度を低下させない範囲で，基底数を削減する

方法を示した．この方法は，基底数を削減するための計算量が，データベース内の文書数に依存しないため，大規模なデータベースへの適用が容易である．

BMIR-J2 の新聞記事（5,080 記事）の検索に適用した実験結果によれば，提案した方法は，単語の表記上の揺らぎに影響されず，同義語や類義語の存在も検索の対象となることから，従来の方法と比べて高い再現率が得られた．その反面，単語を基底とする文書ベクトルの場合に比べて，不適切な記事を拾いやすく，適合率が低下することが分かった．この効果は，キーワード検索においてシソーラスを使用したKW拡張の効果に相当する．また，本方式は，次元数の削減に強い方法であり，従来の方法に比べて，検索精度を落とすことなく，ベクトルの基底数を大幅に削減できることが分かった．

今回は，単語の多義性の問題は考慮しなかったが，単語意味属性を基底とする文書ベクトルでは，意味属性体系の持つ能力を用いて単語の多義を解消した後，基底とする意味属性の重みを計算する方法が可能と考えられるので，今後は，この方法についても検討していきたい．また，基底数をさらに削減する方法として，意味属性体系の上位のノードから順に，不適切な記事を拾いやすいノードを選択してベクトルの基底から削除する方法についても検討していく予定である．

# 参考文献

Borko, H. and Bernick, M. D. (1963). “Automatic Document Classification.” In *Journal of the ACM*, Vol. 10, pp. 151–162.

Burkley,Chris,Singhl, A.,Mitra, M.,and Salton, G. (1996). “New Retrieval Approaches using SMART.” In *TREC4. In D. K. Harman (ed.) The second Text Retrieval Conference (TRC2)*, pp. 25–48.

Deerwester, S.,Dumais, S. T.,Furnas, G. W.,Landauer, T. K.,and Harshman, R. (1990). “Indexing by Latent Semantic Anlysis.” In *Journal of the Society for Information Science*, Vol. 41, pp. 391–407.

Faloutsos, C. and Lin, K. I. (1995). “A fast algorithm for indexing data-mining and visualization of traditional and multimedia datasets.” In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pp. 163–174.

Golub, G. H. and Vanloan, C. F. (1996). *Matrix Computations*.

Jiang, M. W.,Berry, J. M.,Donato, J. M.,and Osrtouchov, G. (1999). “Mining Consumer Product Data Via Latent Semantic Indexing.” In *Intelligent Data Analysis*, Vol. 3, pp. 377–398.

Kwok, K. L. and Chan, M. (1998). “Improving Two Stage ad-hoc retrieval for short queries.” In *In SIGIR’98*, pp. 250–256.

Lin, S. H.,Shih, C. S.,Chen, M. C.,Ho, J. M.,and Ko, M. T. (1998). “Extracting Classification

Knowledge of Internet Documents with Mining Term Association." In *A Semantic Approach Proc. of the 21st Annual Inernational ACM SIGIR Conference on Reaserach and Development in Information Retrieval*.

Salton, G. and McGill, J. M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill New York.

池原, 宮崎, 白井, 横尾, 小倉, 中岩, 大山, 林 (1997). 日本語語彙大系. 岩波書店.

木谷他 (1998). "日本語情報検索システム評価テストコレクション BMIR-J2." 情報処理学会研究報告 98-DBS-114-3.

## 略歴

**池原悟:**　1967 阪大・基礎工・電気卒．1969 同大大学院修士課程修了．同年日本電信電話公社入社，電気通信研究所配属．1996 鳥取大学工学部教授に着任，現在に至る．工博．この間，数式処理，トラフィック理論，自然言語処理の研究に従事．1996 スタンフォード大学客員教授．1982 情報処理学会論文賞，1993 同研究賞，1995 日本科学技術情報センタ賞（学術賞），同年人工知能学会論文賞受賞．2002 電気通信普及財団テレコムシステム技術賞，電子情報通信学会，人工知能学会，言語処理学会，各会員．

**村上仁一:**　1984 筑波大・第 3 学群基礎工学類卒．1986 同大大学院終了．同年 NTT 情報処理研究所に入社．1991 より 1995 ATR 自動翻訳電話研究所に出向．1998 鳥取大学助教授就任，現在に至る．この間，自然言語処理，音声認識の研究に従事．日本音響学会，電子情報通信学会，各会員．

**木本泰博:**　1998 鳥取大・工学部・知能情報工学科卒．2000 同大大学院修士課程修了．同年積水ハウス入社．現在に至る．

(2002 年 5 月 1 日 受付)
(2002 年 5 月 1 日 再受付)
(2002 年 5 月 1 日 採録)