単語意味属性を用いたベクトル空間法

木本泰博 池原悟 村上仁一
鳥取大学工学部
{kimoto,ikehara,murakami}@ike.tottori-u.ac.jp

1 はじめに

近年、WWWなど電子化された文書情報の氾濫しており、自分が必要とする文書情報を効率良く検索する情報検索システムが必要となっている。従来の検索手法として、キーワード検索方式が一般的であるが、最近では、より検索精度の向上を目指して、ベクトル空間法の研究[1]が盛んである。ベクトル空間法では、検索要求を自然文で与えるため、キーワード検索に比べて、具体的に検索条件を表現することができ、検索精度の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では、多数の単語をベクトルの基底に用いるため、類似度計算のコスト、ベクトルのスパース性により、文書間の類似性が判定できない恐れがあることなどが問題とされており、$tf \cdot idf$法[1]などの頻度統計を利用して、文書データベース中の重要語を基底に選択する方法が一般的である。また、LSI法[2]など、基底となる単語が互いの独立性の高くなるようにベクトル空間の基底軸を変換する方法が提案されているが、変換のための計算コストが高い。これらに対して、本論文はベクトル空間の基底となる単語を日本語語彙大系[3]に定義されている単語意味属性(2,710種)に置き換える方法を提案する。

本方式は、文書間の意味的な類似性を単語の意味で評価するため、従来の基底となる単語のみの評価に比べて、表記の揺れに強く、すべての単語が検索に寄与するため、検索漏れの改善が期待できる。また、意味属性相互の意味的な上下関係を利用すれば、検索精度をあまり落さずにベクトルの基底数を削減することができ、容易に基底とすべき必要最小限の意味属性を決定できることが期待される。

本論文では、情報検索テストコレクション BMIR-J2[4]を検索対象とした検索実験により、従来の単語を基底とした方法と検索精度を比較し、本方式の有効性を評価する。

2 ベクトル空間法

2.1 従来のベクトル空間法

従来のベクトル空間法は、文書の意味を文書内の単語を基底とする特性ベクトルで表現する。特性ベクトル$\vec{V}$の各要素には、文書中での単語W_iの重みを絶対値が1になるように正規化した値w_iを与える(式1)。そして、特性ベクトルの間の距離が近い文書を類似文書として検索する。一般的には、各文書D_i, D_jの特性ベクトルの内積をとり、cosinの値を文書間の類似度$sim(V_i, V_j)$(式2)とする。

$$\vec{V} = (w_1, w_2, ..., w_i, ..., w_m) \quad (1)$$

$$sim(V_i, V_j) = \vec{V_i} \cdot \vec{V_j} \quad (2)$$

$\vec{V_i}, \vec{V_j}$は文書V_i, V_jの特性ベクトル

2.2 単語意味属性を用いたベクトル空間法

従来の単語を基底とする場合では、同義語、類議語が含む文書でも、表記が異なると類性が評価されないため、検索漏れが発生する。この問題を解決するために、本論文では、特性ベクトルの基底を単語から単語の意味に置き換えたベクトル空間法を提案する。

本方式では、文書の特性ベクトル$\vec{V'}$(式3)の基底は日本語名詞の意味的用法を2,710種に分類し、相互の意味的関係を最大12段の木構造で表現している日本語語彙大系[3]の一般名詞意味属性(図1)とし、各要素は各意味属性S_iの重みs_iを与える。

$$\vec{V'} = (s_1, s_2, ..., s_i, ..., s_m) \quad (3)$$

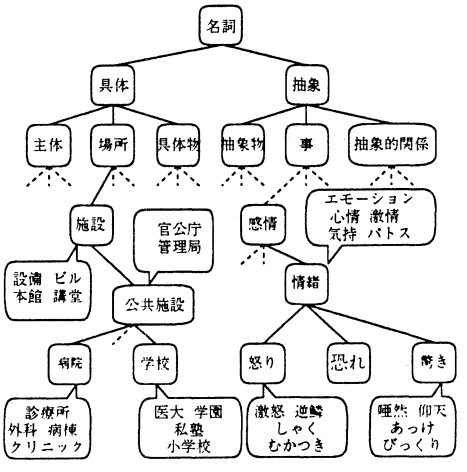

図1: 一般名詞意味体系の一部

また、重み s_i の与え方としては、種々の方法があるが、一般的な $tf{\cdot}idf$ 法を採用する。各意味属性の $tf{\cdot}idf$ 値は以下の方法で求める。

1. 文書 DB に収録された文書全体に対して、意味属性 S_i に属する単語が出現する文書数を求め、idf 値を計算する。

$$idf = \log \frac{\text{文書 } DB \text{ の総文書数}}{DF_i} \quad (4)$$

DF_i : 意味属性 S_i に属する単語の出現する文書数

2. 各文書を対象に、意味属性 S_i に属する単語の出現頻度を求め、tf 値を計算する。

$$tf = \text{意味属性 } S_i \text{ に属する単語の出現頻度} \quad (5)$$

3. (4 式)、(5 式) で得られた tf 値と idf 値から $tf{\cdot}idf$ 値を求める。

$$tf \cdot idf = tf \times idf \quad (6)$$

なお、以下では、単語を基底とするベクトル空間法を単語ベクトル空間法、意味属性を基底とするベクトル空間法を意味ベクトル空間法と呼ぶ。

3 必要最小限の意味属性の決定法

ベクトル空間法では、類似度の計算コスト削減の観点から、ベクトルの基底数削減が望まれる。しかし、一般に、基底数が減少すると検索精度は低下する。本節では、意味属性相互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を発見する方法について述べる。

3.1 意味属性の汎化方法

意味属性の汎化とは、下位の意味属性をその上位の意味属性へ縮退させ、上位属性で代表することである。

汎化の対象となる意味属性の選択方法としては、意味属性の粒度に注目して、意味の粒度が細かい、最下位の意味属性から順次、選択する方法、また、意味属性の重みに注目し、検索にあまり寄与しないと思われる重みの少ない意味属性を選択する方法が考えられる。図 2 に具体的な汎化の例を示す。

文書 DB 内に収録された文書が検索対象となる確率はすべて等しいとし、文書 DB 内、すべての文書を対象に求めた特性ベクトルの和を V_t(式 5) とする。V_t の要素 n_i の値が小さい意味属性は検索精度に与える影響が少なくなるから、少ない基底数で高い検索精度を得

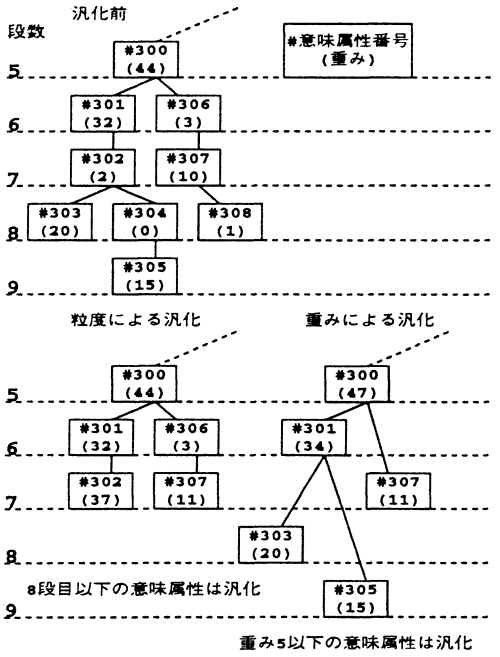

図 2: 汎化の方法

るには、各 n_i の値が均等していることが必要である。つまり、各 n_i の値が均等になるように意味属性を選択し、汎化すれば、検索に寄与しない意味属性を順次削減することが期待できる。

$$\vec{V_t} = (n_1, n_2, ..., n_i, ..., n_m) \quad (7)$$

n_i は意味属性 i の文書 DB 内の出現頻度

ここで、各 n_i の均等さを n_i の平均値 $\bar{n}$(式 9) との変動で評価し、評価関数 H(式 8) を定義する。

$$\begin{aligned} H &= (\bar{n} - n_1)^2 + ... + (\bar{n} - n_i)^2 + \\ &\quad (\bar{n} - n_j)^2 + ... + (\bar{n} - n_m)^2 \\ &= \sum_{i=1}^{m} (\bar{n} - n_i)^2 \end{aligned} \quad (8)$$

$$\bar{n} = \sum_{i=1}^{m} n_i / m \quad (9)$$

H の値が減少するように意味属性 S_i を直属の上位属性 S_j に汎化した場合の H を H' とすると、以下の式が成立する。

$$\begin{aligned} H - H' &= (\bar{n} - n_i)^2 + (\bar{n} - n_j)^2 - \\ &\quad (\bar{n} - n_i - n_j)^2 > 0 \end{aligned} \quad (10)$$

$$n_i \cdot n_j < \frac{\bar{n}^2}{2} \quad (11)$$

以上から、汎化すべき意味属性は (式 11) を満たす意味属性を選択すれば良いことがわかる。以下に、具体的な汎化手順を示す。

1. 上下関係にある意味属性の $n_i \cdot n_j$ の値が最小の意味属性を汎化する。
2. 検索実験を行い、検索精度の低下が、ある値以上であれば、汎化を停止する。
3. 1へ戻る

4 従来法との比較実験

本節では、検索精度と必要最小限の基底数に関する実験を行い、提案した方法と従来の方法の比較を行う。

4.1 実験条件

(1) 検索対象

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」[4](以下 BMIR-J2) を対象とする。BMIR-J2 は、1994 年の毎日新聞より国際十進分類 (UDC) で経済、工学、工学技術一般に分類される記事 5,080 件を対象とし、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」と形式で統一され、「～」の部分の相当する名詞句が列挙されている。検索要求に対する正解記事は、ランクA (～を主題としている記事)、ランクB (～の内容を少しでも含む記事) の 2 種類に分類している。

(2) 評価パラメータ

実験結果は以下のパラメータで評価する。

$$\text{類似度} : sim(D_i, D_j) = \vec{V_i} \cdot \vec{V_j} \qquad (12)$$

$\vec{V_i}, \vec{V_j}$ は文書 D_i, D_j の特性ベクトル

$$\text{再現率} : R = \frac{\text{抽出された正解文書数}}{\text{総正解文書数}} \qquad (13)$$

$$\text{適合率} : P = \frac{\text{抽出された正解文書数}}{\text{抽出された文書数}} \qquad (14)$$

$$F\text{ 値} : F = \frac{(b^2+1) \cdot P \cdot R}{b^2 \cdot P + R} \qquad (15)$$

F 値は適合率と再現率を総合的に評価する指標で、パラメータ b は適合率に対する再現率の相対的な重みを示す。本論文では、両者を対等の重みとし、$b = 1$ の場合で評価する。

(3) 実験方法

検索要求として、新聞記事が与えれ、類似記事を検索することを考え、「主題の一致する記事」を正解記事とする。具体的には、BMIR-J2 のランクAの記事の集合から 1 記事を検索要求用として抽出し、残りのランクAの記事を正解候補とする。

検索精度は 90 件の検索要求の平均検索精度で評価する。

4.2 検索精度に関する実験

意味属性 2,710 種すべてを使用した意味ベクトル空間法について検索実験を行い、検索精度を単語ベクトル空間法 (基底数 2,710) と比較した。文書検索では、類似度がある一定値以上を持つ記事を検索の対象とするが、その選び方によって、適合率、再現率は変化する。そこで、検索精度の評価では、総合的な F 値が最大となるように類似度を設定する。

類似度がある一定値以上を抽出した場合の類似度と適合率、再現率の関係を図 3 に示す。なお、類似度 0.7 以上の場合は、抽出される文書が高々1 件程度となるため、グラフから削除した。

F 値が最大となるように類似度を設定した場合の検索精度を表 1 に示す。

表 1: 基底数 2,710 の時の検索精度

基底数	再現率	適合率	F 値
本手法	0.55	0.50	0.45
従来法	0.52	0.56	0.45

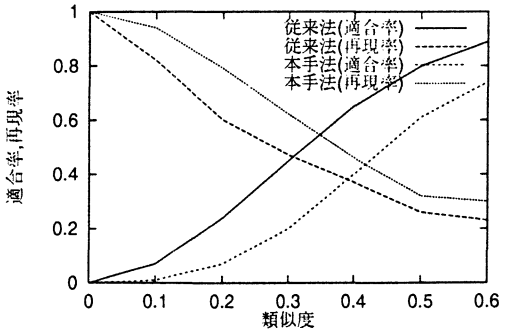

図 3: 類似度と適合率・再現率の関係

以上の図 3 の結果から、本方式は表記の揺らぎに強く再現率は高いが、その反面、検索ゴミを拾いやすく適合率は低いことが分かる。また、表 1 の結果から総合的な検索精度 (F 値) はほとんど変わらないことがわかる。

4.3 基底数削減に関する実験

単語ベクトル空間法 (最大基底数 5,000) と前節で説明した粒度の汎化と重みの汎化による基底削減法を適用した意味ベクトル空間法の基底数と検索精度の関係を求め、基底数削減の可能性と必要最小限の基底数について比較する。

図 4 に基底数と検索精度 (F 値) の関係を示す。

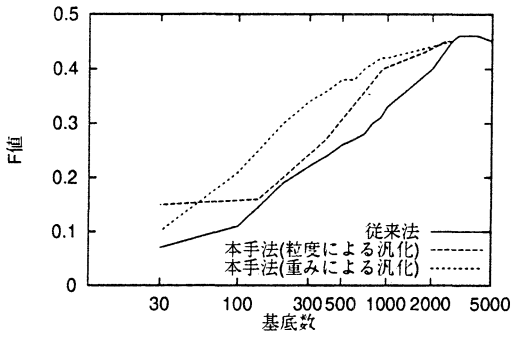

図 4: 基底数と検索精度 (F 値) の関係

また、検索精度の低下の許容範囲を 10 ~ 20%程度とした場合の必要最小限の基底数を表 2 に示す。

表 2: 必要最小限の基底数

方式種別	基底数削法	検索精度低下の許容度	
		ピーク値の 10%	ピーク値の 20%
本手法	粒度による汎化	700 属性	500 属性
	重みによる汎化	300 属性	200 属性
従来法	$tf \cdot idf$ 法	2,500 属性	1,500 属性

以上の図 4、表 2 の結果から、以下のことが分かる。

(1) 意味ベクトル空間法は、従来の単語ベクトル空間法に比べて、基底数削減に強い。

(2) 汎化の方法は、粒度による汎化より重みによる汎化の方がより基底数削減に強い。

(3) 基底数が約 2,500 種以下では、検索精度は意味ベクトル空間法の方が優れている。

必要最小限の基底数については、検索精度の低下の許容範囲を 10 ~ 20%程度とすると必要最小限の基底数は 1500 ~ 2000 種である。これに対し、本方式は約 200 ~ 300 種程度まで基底数を削減できることがわかる。

5 おわりに

従来、ベクトル空間法では、文書の意味的な特徴を表す特性ベクトルの基底に、文書中に現れる単語を使用するのが一般的であったが、本論文では、単語の代わりに単語意味属性を使用する方法を提案した。また、意味属性間の意味的な上下関係を利用し、検索精度をあまり低下させないで基底を削減する方法を示した。BMIR-J2 の 5,080 記事を検索対象とした検索実験の結果によれば、提案した方法は、単語の揺らぎに影響されず、同義語、類義語の存在も検索の対象となるため、キーワード検索における、シソーラスを使用したキーワード拡張と同等の効果があり、従来の単語ベクトル空間法に比べて、高い再現率を得られるが、反面、検索ゴミを拾いやすく、適合率が低下することが分かった。

本論文では、単語の多義性については考慮しなかったが、意味属性体系の持つ能力を用いて単語の多義を解消すること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属性から順に検索ゴミを発生しやすいと意味属性を削除する方法が考えられ、今後は基底数削減について LSI 法との比較も検討していく予定である。

謝辞

本研究では、(社) 情報処理学会・データベースシステム研究会が、新情報処理開発機構との共同開発により、毎日新聞 CD-ROM'94 データ版を基に構築した情報処理検索システム評価用テストコレクション BMIR-J2 を利用した。毎日新聞社ならびに BMIR-J2 の開発に携わられた方々に感謝します。

参考文献

[1] Salton, G. and McGill, M.J.: "Introduction to Modern Information Retrieval",McGraw-Hill,(1983)

[2] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A.:"Indexing by latent semantic analysis." Journal of the American Society for Information Science,pp.391-407,(1990)

[3] 池原他：日本語語彙大系, 岩波書店,(1997)

[4] 木谷他: 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,(1998)

[5] 木本、池原、白井：日本文意味的検索に必要な最小単語意味属性の組の決定, 情報処理学会第 57 回全国大会,pp.241-242,(1998)

[6] 木本、池原、村上：意味属性を用いたベクトル空間法の検索精度, 電気情報通信学会研究報告 NLC99-24,pp.37-44,(1999)