情 報 学 基 礎 60－3
（2000．9．27）

意味情報に基づく検索と全文検索の統合

北内 啓　高木 徹　岩城 修
株式会社 NTT データ
{kitauchi,takaki,iwaki}@rd.nttdata.co.jp

単語の意味情報に基づく検索手法と全文検索手法の統合手法を提案する．まず，意味分類辞書および検索対象文書中の単語共起を用いることにより，単語に対して意味ベクトルを付与し，これを参照することで意味情報に基づく検索を実現する．また，意味の広い単語の検索には意味情報に基づく検索が適していることを利用し，意味分類辞書において検索語が上位の階層の意味分類に属するときは，意味情報に基づく検索のスコアの比重を高くする．BMIR-J2 を用いた評価実験を行った結果，全文検索と比較して平均適合率が 7.3%向上することを確認した．

[キーワード] 情報検索, 意味情報, シソーラス, BMIR-J2

Integration of Semantics-based Search and Full-text Search

Akira KITAUCHI　Toru TAKAKI　Osamu IWAKI
NTT Data Corporation

This paper proposes a method for integrating a semantics-based search with a full-text search. Semantic vectors for all words are made using a thesaurus and word co-occurrences within the documents, and referred for the semantics-based search. Since the semantics-based search works better than traditional full-text search for query words which have wide meaning, a weight of ranking score for the semantics-based search is increased when the query words belong to upper categories in the thesaurus. Experimental results using a test collection BMIR-J2 show that the proposed method improves average precision about 7.3% compared to the traditional full-text search.

[keyword] information retrieval, semantics, thesaurus, BMIR-J2

1　はじめに

近年，インターネットをはじめとして，電子化された大量の文書が利用可能となっており，利用者が必要な情報を効率よく見つけ出すための検索技術の重要性がますます高まっている．特に，現在 WWW の検索エンジンなどでは，全文検索技術が広く用いられている．しかし全文検索では，検索語そのものを含む文書しか検索できないため，検索要求の主題に合致した文書でも，検索語そのものが含まれていなければ検索できない．

これに対し，検索語が含まれていない文書も検索可能とすることにより，検索精度を向上させる手法がいくつか提案されている．たとえば，文書内単語共起を利用して検索語を拡張（追加）する検索手法 [1] がある．一度目の検索の後，その検索結果の上位の文書中から重要な単語を抽出し，拡張検索語として再度検索を行う．この方法は，検索語と同じ文書に出現する単語は検索語との関連度が高いという前提のもとに検索語を拡張している．しかし，表層的な単語の出現情報だけを用いているため，検索語とあまり関連のない単語が拡張検索語として追加さ

れ，逆に検索精度が低下することがある．

これに対し，意味的な情報を利用して検索を行う手法が検討されている．芥子ら [2, 3] は，単語の意味情報を 266 次元のベクトルによって表現している．基本となる単語には人手によってベクトルを付与し，その他の単語については文書内単語共起を利用して自動的にベクトルを付与している．また，[3] では，単語の意味情報を用いた検索と全文検索のスコアを一定の比重で加算することにより，検索精度を向上させている．

本論文では，各単語に対して意味ベクトルを付与し，これを参照することで意味情報に基づく検索（以下，意味検索とよぶ）を行う．また，[3] のように意味検索と全文検索のスコアを一定の比重で加算するのではなく，意味分類辞書中で検索語が属する意味分類の階層によって，意味検索と全文検索のスコアの比重を変化させる．BMIR-J2 を用いた実験により，その有効性を評価する．

2 単語への意味ベクトルの付与

単語の意味ベクトルは，意味素性を基底とし，各意味素性との関連度の大小を属性値として与えることにより，単語の意味をベクトルの形で表現したものである．たとえば,「公害」の意味ベクトルを（産業，環境，通信，損失，利益）などの意味素性を用いて以下のように表現する．

公害 = (産業:0.2, 環境:0.15, 通信:0, 損失:0.2, 利益:0, ···)

本章では，まず意味素性の決定方法について説明したあと，単語への意味ベクトルの付与方法について述べる．

2.1 意味素性の決定

意味ベクトルの意味素性は単語の意味の表現要素となるため，意味素性を選択するには単語の概念や意味の体系化が必要である．すでに体系化されたものとして「分類語彙表」[4] 「角川類語新辞典」[5] などの意味分類辞書がある．本論文では，これら既存の意味分類辞書の意味分類をそのまま意味素性として用いる．意味分類辞書の意味分類は，図 1のように階層的な構造になっている（図 1では 3 階層）．

第 1 階層	第 2 階層	第 3 階層（収録単語）
社会	施設	住居 (住まい，住宅，社宅，···)
		役所 (官庁，県庁，郵便局，···)
		···
	報道	発表 (公表，披露，広報，···)
		編集 (監修，特集，新編，···)
		···
	···	
···		

図 1: 意味分類辞書の階層的な意味分類の例

意味素性の粒度は，検索精度に大きな関係があると考えられるため，意味分類の階層を限定して意味素性群を用意した．たとえば，図 1において第 2 階層のすべての意味分類（施設，報道，···）を意味素性とする．階層ごとの意味素性を用いて意味ベクトルを付与し，意味情報に基づく検索実験を行ったところ，意味素性の数が多いほど高い検索精度が得られた．そこで，もっとも意味素性の数が多い角川類語新辞典の第 4 階層 (一部第 3 階層を含む) による 2810 分類を基本の意味素性とした．

しかし，この基本意味素性では，まだ意味分類の意味の範囲が広い場合がある．たとえば,「住宅」と「マンション」のように意味の広い単語と意味の狭い単語が同じ分類に属しているので,「マンション」で検索すると「住宅」を含む文書が検索されてしまうことがある．そのため，検索漏れは減少するが，不要な文書も検索されるので適合率が低下する．そこで，角川類語新辞典の単語に記されている語義文を利用し，一つの意味分類をその同義語と下位範疇語の二つに自動的に分割した．たとえば,「住居」という意味分類に含まれる単語を,「住居」の同義語である「住宅」「住まい」などからなる分類と，その下位範疇語である「マンション」「社宅」などからなる分類の二つに分割した．

2.2 単語への意味ベクトルの付与

検索対象文書中の各単語に対し，意味ベクトルを付与する方法について説明する．

2.2.1 基本単語への付与

まず，基本単語として，検索対象文書中の単語のうち角川類語新辞典に収録されている単語に意味ベ

クトルを付与する．前節で決定した意味素性（意味分類）を用いて，その単語が属する分類の属性値を1，それ以外の属性値を0とする．たとえば，意味分類「学校」に属する単語「小学校」の意味ベクトルは,「学校」の属性値を1，その他の意味素性の属性値を0とする．ただし,「菓子」のように上位の階層の意味分類に属する単語については,「和菓子」「洋菓子」などその下位の意味分類に対応する意味素性にも属性値を付与した．また，複数の意味分類に属する単語の意味ベクトルについては，その長さを1に正規化した．

2.2.2 全単語への付与

次に，検索対象文書中のすべての単語に対して意味ベクトルを付与する．文書中の単語共起を利用し，検索対象文書中で単語 w_i の周辺に出現する基本単語の意味ベクトルから単語 w_i の意味ベクトルを算出する．具体的には以下の手順で意味ベクトルを付与する．

1. 「茶筌」[6] を用いて，検索対象文書に対して形態素解析を行い，解析結果から名詞と未知語を抽出する．
2. すべての検索対象文書中で，単語 w_i が出現する文書に含まれる基本単語の集合を $C = (b_1, \cdots, b_m)$ とする．
3. C 中の基本単語 b_j の意味ベクトル $\overrightarrow{B_j} = (b_{j1}, \cdots, b_{jn})$ をもとに，単語 w_i の意味ベクトル $\overrightarrow{W_i}$ を次式により求める．

$$\overrightarrow{W_i} = \frac{\overrightarrow{W_i'}}{|\overrightarrow{W_i'}|} \quad (1)$$

$$\overrightarrow{W_i'} = \sum_{j=1}^{m} TF_{j,C} IDF_j (FIDF_1 b_{j1}, \cdots, FIDF_n b_{jn}) \quad (2)$$

ここで，$TF_{j,C}$, IDF_j, $FIDF_k$ はそれぞれ以下の式によって求められる．IDF_j は基本単語 b_j の重要度を，$FIDF_k$ は意味素性 f_k の重要度をそれぞれ表している．また，n は意味ベクトルの次元数，すなわち意味素性の数である．

$$TF_{j,C} = \log_2(C \text{ における } b_j \text{の頻度} + 1)$$

$$IDF_j = \log_2 \frac{\text{検索対象文書数}}{b_j \text{の出現する文書数}} + 1$$

$$FIDF_k = \log_2(\text{全基本単語数}/\text{意味素性 } f_k \text{ に属する基本単語数}) + 1$$

3 検索手法

意味検索の検索手法を述べたあと，本論文の提案手法である意味検索と全文検索の統合について説明する．

3.1 意味検索手法

意味検索の検索手法は，単語の意味ベクトルに基づくベクトル空間法を用いる．単語の $TF \cdot IDF$ に基づくベクトル空間法と同様，検索対象文書と検索要求に意味ベクトルを付与し，両者のベクトルの類似度として余弦（$cosine$）を計算することにより，文書の検索スコアとする．

具体的には，まず検索対象文書 d_i に対して「茶筌」を用いて形態素解析を行い，解析結果から，名詞と未知語を抽出する．抽出した単語群 $(dw_1, \cdots, dw_m)$ の意味ベクトルから，以下の式によって検索対象文書あるいは検索要求の意味ベクトル $\overrightarrow{D_i}$ を求める．

$$\overrightarrow{D_i} = \frac{\overrightarrow{D_i'}}{|\overrightarrow{D_i'}|} \quad (3)$$

$$\overrightarrow{D_i'} = \sum_{j=1}^{m} TF_{ji} IDF_i (FIDF_1 dw_{j1}, \cdots, FIDF_n dw_{jn}) \quad (4)$$

検索時は，検索要求 q_i の意味ベクトル $\overrightarrow{Q_i}$ を検索対象文書 d_j の意味ベクトルと同様の方法で求める．$\overrightarrow{Q_i}$ と $\overrightarrow{D_j}$ の類似度 $sim_m(q_i, d_j)$ を以下の式で求め，意味検索スコアとする．

$$sim_m(q_i, d_j) = \frac{\overrightarrow{Q_i} \cdot \overrightarrow{D_j}}{|\overrightarrow{Q_i}||\overrightarrow{D_j}|} \quad (5)$$

3.2 意味検索と全文検索の統合

同じ意味分類に属している基本単語には同じ意味ベクトルが付与される．そのため，文書内単語共起を利用して全単語に意味ベクトルを付与した場合で

も，同じ意味分類に属している単語は意味ベクトルが近い．意味分類は数千個しかないため，意味情報だけを用いて検索を行った場合，単語の細かい意味の違いを表現できず，検索精度が低くなることがある．そこで，意味検索と全文検索を統合した手法を提案する．

本論文では，意味の狭い単語の検索には全文検索が適しており，意味の広い単語の検索には意味検索が適していることが多いことに着目した．たとえば，「住宅」とその下位範疇語である「マンション」にはお互いに近い意味ベクトルが付与されている．そのため，意味の狭い単語である「マンション」で検索すると，「マンション」を含む文書だけではなく「住宅」を含む文書も検索され，精度が低下することがある．逆に意味の広い単語である「住宅」で検索すると，「住宅」を含む文書だけではなく「マンション」を含む文書も検索され，精度の向上につながる．

また，角川類語新辞典に収録されていない単語は，単語の表層的な出現情報だけをもとに意味ベクトルを付与しているため，意味検索を行うと精度が低くなることが多い．

そこで，検索要求を形態素解析して抽出した各検索語について，意味分類辞書に収録されていないか，意味分類辞書において最下位の階層の意味分類に属する単語であるときは，全文検索のスコアの比重を大きくし，それよりも上位の階層の意味分類に属するときは意味検索のスコアの比重を大きくして全体の検索スコアを計算する．

まず，全文検索の検索スコアは単語の$TF \cdot IDF$に基づくベクトル空間法 [7] を用いて算出する．検索要求 q_i と検索対象文書 d_j の単語ベクトルに基づく類似度 $sim_f(q_i, d_j)$ を以下の式によって求める．

$$sim_f(q_i, d_j) = \frac{\sum_{k=1}^{n}(tq_{ki} \times td_{kj})}{\sqrt{\sum_{k=1}^{n} tq_{ki}^2 \times \sum_{k=1}^{n} td_{kj}^2}} \quad (6)$$

ここで，tq_{ki}, td_{kj} はそれぞれ，検索要求 q_i, 文書 d_j 中の単語 w_k の出現頻度に基づく重要度を表しており，以下の式で示される．

$$tq_{ki} = \log_2(q_i\text{中の } w_k\text{の出現頻度} + 1)$$

$$td_{kj} = \log_2(d_j\text{中の } w_k\text{の出現頻度} + 1) \times (\log_2 \frac{\text{検索対象文書数}}{w_k\text{の出現する文書数}} + 1)$$

検索要求 q_i から抽出された検索語を $(wq_1, \cdots, wq_n)$ とすると，意味検索と全文検索を統合した検索スコア $sim_i(q_i, d_j)$ は以下の式によって示される．

$$sim_i(q_i, d_j) = \sum_{k=1}^{n}(\alpha_k sim_f(wq_k, d_j) + (1 - \alpha_k) sim_m(wq_k, d_j)) \quad (7)$$

ここで，$\alpha_k (0 \leq \alpha_k \leq 1)$ は検索語 wq_k に対する全文検索スコアの比重であり，以下の式によって表される．

$$\alpha_k = \begin{cases} \alpha_{wide} & wq_k\text{が上位の意味分類に属するとき} \\ \alpha_{narr} & wq_k\text{が最下位の意味分類に属するとき} \end{cases} \quad (8)$$

検索結果は，$sim_i(q_i, d_j)$ の大きい順に文書をランキングして出力する．

4 評価実験

本論文で提案した，意味検索と全文検索の統合手法の検索精度を評価した．

4.1 評価対象

検索用テストコレクション BMIR-J2 を利用して検索精度を評価した．BMIR-J2 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作業により，毎日新聞 CD-ROM'94 データ版を基に構築したものである．毎日新聞 1994 年の記事 5,080 件からなる検索対象文書，60 件の検索要求，正解判定結果から構成されている．本評価では，検索要求のうち 50 件の標準セットを利用した．正解判定結果は A，B，C の 3 つのランクが付与されており，本評価ではランク A（検索要求を主題とする記事）とランク B（検索要求の内容が少しでも記述されている記事）の文書を正解とした．

4.2 単語の意味ベクトル

角川類語新辞典を利用して5,126個の意味素性を決定し，これを用いて，BMIR-J2の新聞記事5,080件中に含まれる単語約24,000語に対して意味ベクトルを付与した．なお，基本単語の数は約13,000語であった．この意味ベクトルを参照し，意味検索を行った．

4.3 評価方法

検索要求は「高速道路の建設」のようなフレーズになっているので，「茶筌」[6]を用いて形態素解析を行い，名詞と未知語の単語を抽出し，それらを検索語として検索を行った．

評価尺度としては，平均適合率 (average precision) を用いた．また，再現率 (recall) が $10\%, 20\%, \cdots, 100\%$ のときの適合率 (precision) を測定した．

α_{wide}, α_{narr} の2つのパラメータを調整することにより，以下の3通りの検索手法の検索精度を測定した．

1. 手法A

   $\alpha_{wide} = \alpha_{narr} = 1$ とし，全文検索のみを利用した．

2. 手法B

   $\alpha_{wide} = \alpha_{narr}$ とし，意味検索と全文検索のスコアを一定の比重で加算した．パラメータを変化させて評価した．

3. 手法C

   $\alpha_{narr} \neq \alpha_{wide}$ とし，検索語が属する意味分類によって意味検索と全文検索のスコアの比重を変更する．ただし，α_{narr} は手法Bでもっとも精度が高かったときの値を採用し，α_{wide} を変化させて評価した．

4.4 評価結果および考察

手法A,Bの実験結果を表1に示す．カッコ内の値は $\alpha_{wide} = \alpha_{narr}$ の値である．全文検索に加えて意味検索も利用した手法Bは，全文検索のみを利用した手法Aと比較して平均適合率が大きく向上している．また，手法Bの中では $\alpha_{wide} = \alpha_{narr} = 0.7$ の時に平均適合率がもっとも高い．

表1: 手法A,Bの実験結果

再現率 (%)	適合率 (%)			
	手法A (1.0)	手法B (0.8)	手法B (0.7)	手法B (0.6)
0	64.9	67.6	69.2	67.5
10	49.0	57.8	58.4	56.9
20	44.0	51.8	52.9	52.0
30	41.5	47.9	48.1	48.0
40	38.8	45.3	45.6	44.8
50	37.3	42.5	42.7	42.2
60	33.1	37.9	38.4	37.8
70	29.5	33.6	33.7	33.4
80	27.3	29.7	29.9	29.9
90	16.5	20.9	20.7	20.9
100	10.7	13.3	12.7	13.2
平均適合率	33.3	39.2	39.8	38.9

表2: 手法Cの実験結果 ($\alpha_{narr} = 0.7$)

再現率 (%)	適合率 (%)			
	手法C (0.4)	手法C (0.5)	手法C (0.6)	手法B ($\alpha_{wide} = 0.7$)
0	72.1	73.1	72.0	69.2
10	59.7	60.6	60.3	58.4
20	54.2	54.4	54.5	52.9
30	49.5	49.6	49.2	48.1
40	45.6	46.1	46.3	45.6
50	42.4	43.1	42.9	42.7
60	39.2	39.7	39.2	38.4
70	34.3	34.2	34.1	33.7
80	28.9	30.1	29.9	29.9
90	19.5	20.3	20.6	20.7
100	11.1	12.0	12.4	12.7
平均適合率	39.9	40.6	40.5	39.8

次に，$\alpha_{narr} = 0.7$ に固定し，α_{wide} だけを変化させた手法Cの実験結果を表2に示す．$\alpha_{wide} = 0.5$ のときにもっとも高い平均適合率を示した．

手法C($\alpha_{wide} = 0.5$) の平均適合率は，手法Aより7.3%，手法B($\alpha_{narr} = 0.7$) より0.8%向上している．手法Cは特に，再現率が低いときの適合率が高く，再現率が0%のときの適合率73.1%は手法Bの69.2%より3.9%向上している．また，検索要求50件のうち，上位の階層の意味分類に属する検索語を含む検索要求は30件あり，手法Cを手法Bと比較して，平均適合率が向上した検索要求は20件，低下した検索要求は10件あった．

また，手法B，手法Cの平均適合率に大きな差が見られた検索要求を表3に示す．上段が手法Bの

精度が高かった検索要求，下段が手法Cの精度が高かった検索要求である．なお，上位の意味分類に属する検索語は [販売] のように [ ] で囲んである．

手法Bと手法Cを比較すると，全検索要求の平均としては手法Cの方が高い精度を示したが，手法Bの方が精度が高かった検索要求もあった．たとえば,「マンション，販売」という検索語では「販売」が上位の意味分類に属している．出力文書と正解文書を比較，分析したところ，手法Cでは「販売」よりも下位の意味分類に属する「発売」という単語を含む文書のスコアが高くなったために検索精度が向上したことが分かった.「菓子，メーカー」という検索語も同様に,「菓子」が上位の意味分類に属しており，その下位分類の単語である「チョコレート」「ヨーグルト」などを含む文書のスコアが高く，精度が向上した．

一方,「赤字, 国債, 発行」という検索語では「発行」が上位の意味分類に属しており，手法Cは手法Bに比べ「発行」を含む文書のスコアが低くなっていた．しかし，実際には，正解文書は「発行」を含む文書が多かったため，手法Cの精度が低下していた．また,「女性，雇用，問題」という検索語においても同様に，手法Cでは「女性」を含む文書のスコアが低くなるが，正解文書には「女性」を含む文書が多かったため手法Cの精度が低下した．

つまり，検索語が上位の意味分類に属するとき，その下位分類の単語が正解文書に含まれることが多ければ，手法Cの検索精度は向上する．しかし，逆に検索語そのものを含む正解文書が多い場合もあり，そのようなときは検索精度が低下する．このように，検索語が上位の意味分類に属するとき，その検索語には意味検索が適していることが多いが，逆に全文検索の方が適している場合もあることが分かった．

5 おわりに

本論文では，意味分類辞書および文書内単語共起を利用して，単語に意味ベクトルを付与し，これを参照することで意味情報に基づく検索を実現した．また，検索語が属する意味分類の階層によって，意味検索と全文検索のスコアの比重を決定し，両者の検索スコアを加算するという統合検索手法を提案した．検索用テストコレクションBMIR-J2を用いた評価実験によって，従来手法と比較して検索精度が向上することを示した．今後は意味検索に適した検索語の性質をさらに分析し，検索語についての条件を追加することによって，さらに検索精度を向上させる手法を検討していきたい．

表 3: 平均適合率の差が大きい検索要求

検索語	手法B	手法C
赤字, 国債, [発行]	54.5	46.8
[女性], 雇用, [問題]	42.3	38.7
[菓子], メーカー	58.9	68.6
[製造], 現地, 法人	18.7	22.0
[政党], 献金	69.2	75.2
マンション, [販売]	89.1	92.8
材料, [設備], 現地, 調達	21.1	38.3
[業績], [不振], 責任, [経営]	7.9	17.6

本研究は放送・通信機構（TAO）の平成11年度委託研究として実施したものである．また,（株）角川書店より,「角川類語新辞典」の研究利用の許可を頂いた．この場を借りて感謝いたします．

参考文献

[1] J. Xu and W. B. Croft. Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 4–11, 1996.

[2] 芥子育雄, 乾隆夫, 石鞍謙一郎. 大規模データベースからの連想検索. 信学技報 AI92-99, pp. 73–80, 1993.

[3] 芥子育雄, 黒武者健一, 河村晃好. 連想検索を用いたIREX-IRの結果について. IREXワークショップ予稿集, pp. 75–79, 1999.

[4] 国立国語研究所編. 分類語彙表. 大日本図書, 1994.

[5] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[6] 松本裕治ほか. 日本語形態素解析システム『茶筌』version 2.0 使用説明書. Information Science Technical Report NAIST-IS-TR99012, Nara Institute of Science and Technology, 1997.

[7] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.