研究論文　3-6

情報検索における ベクトル空間モデルの応用

大谷紀子[1]

1　はじめに

インターネットやパソコンの普及に伴ない、今日では様々な情報がオンライン、或いは電子記憶媒体を介して入手できるようになった。インターネット上のホームページでは、企業情報や施設紹介などの定常的なストック情報だけでなく、ニュースや新製品情報など次々と更新されるフロー情報が公開されている。個人宛情報は電子メールという形式での配信が主流となり、情報の１対１、１対多、多対多のやり取りのいずれもが、よりスムーズに行なえるようになった。しかし、その弊害として膨大な量の情報が氾濫し、必要な情報の入手が困難という「情報洪水」の問題が生じている。

現在、ホームページからの情報収集を支援するサービスとして、Google、goo、Infoseek などのロボット型サーチエンジンや、Yahoo!に代表されるディレクトリ型サーチエンジンが普及している。

ロボット型サーチエンジンでは、Crawler、Robot、Worm などと呼ばれるプログラムにより、インターネット上に公開されているホームページを自動収集し、検索用インデックスを作成する。検索時にはインデックスに対して全文検索を行ない、入力された検索キーワード、およびそれと同等の語を含むホームページのリストを提示する。情報収集を自動で行なうため、多数のホームページが検索対象として網羅される。しかし、検索対象が多いことに加えて、検索キーワードが主題となっていないページや、スパムページをも検索結果に含むため、結果は膨大な数に及ぶ。検索結果は、検索キーワードとページの記載事項との合致度や、ページの重要度によりランキングされているものの、大量の結果から有用なページを探し出す作業は非常に労力を要する。

ディレクトリ型サーチエンジンでは、人手により登録されたホームページが検索対象となるため、検索キーワードと無関係なページが提示されることはないが、ロボット型サーチエンジンと比較すると、非常に限定された対象からの検索しか行なえない。登録処理に人手を介する点も短所の１つといえる。

電子メールに関しては、差出人や宛先、件名、本文中の単語などで受信フォルダを振り分ける機能をメールソフトに持たせることで、大量メールの管理を支援している。しかし、振り分けの基準は表層的な情報のみであるので、ユーザの要求を満足する支援とはいえない。

電子化文書を最大限に活用するためには、単なるキーワード検索ではなく、文書の内容を把握した上での検索が重要となる。本稿では、文書および語の意味の数量的表現技法としてベクトル空間モデルを取り上げ、大量の電子化文書からの効率的な検索を目的とした応用事例を紹介する。

2　ベクトル空間モデル

ベクトル空間モデル（Vector Space Model）は Salton ら[7]により提案された技法であり、情報検索分野で幅広く利用されている。出現単語に基づいて文書あるいは文章を１つのベクトルで表現し、ベクトルの向きによって内容を判断する点が特徴である。本節では、ベクトル空間モデルの処理方法のうち、tf・idf 法と単語間共起に基づく方法について概説する。

2.1　tf・idf 法

tf・idf 法では、ある文書における出現頻度が高く、すべての文書のうち特定の文書に偏在する単語が、その文書の特徴を表す単語であると見なす。こ

1　武蔵工業大学環境情報学部講師

のような単語を有効語と呼ぶ。有効語を軸として文書をベクトル表現することで、文書の内容がベクトルの向きとして示される。すべての文書ベクトルで形成される空間は、文書の意味内容を反映したベクトル空間となる。

2.1.1 有効語抽出

有効語としては、文書中の単語のうち重要度の高い語が選択される。一般に、名詞や形容動詞の語幹、サ変動詞の語幹が有効語の候補とされる。形態素解析により抜き出した該当品詞の単語について重要度を算出し、その値の高い方から順に、あらかじめ定められた個数の単語を有効語とする。

ある文書における単語の出現頻度を tf（term frequency）、単語が出現する文書の割合を idf(inverse document frequency）と呼び、両者の積をその文書における単語の重要度とする。文書 D における単語 T の出現頻度を $tfreq(T, D)$、単語 T を含む文書数を $dfreq(T)$、全文書数を M とすると、単語 T の文書 D における重要度 $w(T, D)$ は以下のように定義される。

$$w(T, D) = tf(T, D) \cdot idf(T)$$

$$tf(T, D) = tfreq(T, D) \tag{1}$$

$$idf(T) = \log \frac{M}{dfreq(T)} \tag{2}$$

tf は文書長の影響を受けやすいため、以下のような正規化が有効と考えられる。ここで、$len(D)$ を文書 D の文字数、$tnum(D)$ を文書 D に含まれる単語数とする。

$$tf(T, D) = \frac{1.0 + \log(tfreq(T, D))}{\log(len(D))} \tag{3}$$

$$tf(T, D) = \frac{\log(tfreq(T, D))}{\log(tnum(D))} \tag{4}$$

$$tf(T, D) = \frac{\log(tfreq(T, D) + 1)}{\log(tnum(D))} \tag{5}$$

idf の算出には下記の式が用いられることもある。

$$idf(T) = \log \frac{M}{dfreq(T)} + 1 \tag{6}$$

$$idf(T) = \log \frac{M - dfreq(T)}{dfreq(T)} \tag{7}$$

2.1.2 文書ベクトルと類似度判定

２文書に同一単語が多く出現する場合、文書の内容が類似していると考えられる。有効語の出現状況に基づいて文書をベクトル表現することで、内容の近い文書のベクトル同士は近くに、内容のかけ離れた文書のベクトル同士は遠くに位置するよう、ベクトル空間が形成される。N 個の有効語 $V_1, V_2, \cdots, V_N$ が抽出されたとき、文書 D を以下のような N 次元ベクトル $\vec{d}$ として表現する。

$$\vec{d} = (w(V_1, D), w(V_2, D), \cdots, w(V_N, D))$$

２つの文書 D_i と D_j の類似度 $sim(D_i, D_j)$ は、各文書を表す文書ベクトル $\vec{d}_i$ と $\vec{d}_j$ のなす角の余弦値により求められる。

$$sim(D_i, D_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\| \vec{d}_i \| \cdot \| \vec{d}_j \|}$$

類似度は０以上１以下の実数値で、値が大きいほど２つの文書は類似しているといえる。ある文書に類似した文書を検索する場合は、類似度の閾値 α をあらかじめ定めておき、文書間の類似度が α を超えるような文書を検索結果とする。検索要求を文で指定する場合には、文書ベクトル生成と同じ要領で検索要求文をベクトル表現し、各文書との類似度を算出して、類似度が α を超える文書を検索結果とする。

2.2 単語間共起に基づく方法

各文書が独立に存在する場合は前述の tf・idf 法が利用されるが、文書の属するカテゴリが指定されている場合には、カテゴリ内における単語の共起状況を考慮したベクトル空間が有効であり[12]、文書

分類システムで利用されている[3,4,5,6,10]。

2.2.1 有効語抽出と有効語ベクトル

単語間共起に基づく方法では、カテゴリの特徴を表す単語、つまり特定のカテゴリに偏って出現する単語を有効語と見なす。各文書が L 個のカテゴリ $C_1, C_2, \cdots, C_L$ に分類されているとき、カテゴリ C_i 内の文書のうち単語 T を含む文書数を $dfreq(T, C_i)$、カテゴリ C_i の文書数を $dnum(C_i)$ とすると、単語 T のカテゴリ C_i への帰属度 $bel(T, C_i)$ は以下の式で求められる。

$$bel(T, C_i) = \frac{dfreq(T, C_i)}{dnum(C_i)}$$

帰属度 $bel(T, C_i)$ を正規化して、エントロピー $ent(T)$ を算出する。

$$bel(T, C_i) \leftarrow \frac{bel(T, C_i)}{\sum_{i=1}^{L} bel(T, C_i)}$$

$$ent(T) = -\sum_{i=1}^{L} bel(T, C_i) \log_2 bel(T, C_i)$$

エントロピー $ent(T)$ から局在度 $par(T)$、単語 T を含む文書の割合から信頼度係数 $rel(T)$ を求め、両者の積を T の重要度 $w(T)$ とする。ここで γ は任意の定数とする。

$$par(T) = 1 - ent(T)$$

$$rel(T) = \frac{1 - \gamma^{\frac{dfreq(T)}{M}}}{1 + \gamma^{\frac{dfreq(T)}{M}}}$$

$$w(T) = par(T) \cdot rel(T)$$

tf·idf 法と同様に、重要度の高い N 個の単語 $V_1, V_2, \cdots, V_N$ を有効語とする。これらの有効語には、同義語や類義語、対義語が混在すると考えられる。また、有効語間の意味的な類似性はそれぞれ異なっているため、有効語そのものを軸としてベクトルを生成するのは望ましくない。単語の意味的類似度をベクトル空間に反映させるために、有効語を共起係数に基づくベクトルで表現する。有効語 V_i と有効語 V_j の両方を含む文書数を $dfreq(V_i, V_j)$、有効語 V_i を含む文書数を $dfreq(V_i)$ とすると、有効語 V_i と有効語 V_j との共起係数 $c(V_i, V_j)$ は次のように定義される。

$$c(V_i, V_j) = \frac{dfreq(V_i, V_j)}{dfreq(V_i)}$$

共起係数を成分とする N 次元ベクトル $\vec{v_i}$ で有効語 V_i を表す。

$$\vec{v_i} = (c(V_i, V_1), c(V_i, V_2), \cdots, c(V_i, V_N))$$

以上のように生成された有効語ベクトルの次元数は有効語数 N と等しい。有効語ベクトルの保持領域およびベクトルに関わる計算時間はベクトルの次元数に依存するため、N の値を大きく設定すると、メモリ不足や、実時間での実行が不可能といった問題が生じる。情報量を減少させずにベクトルの次元を圧縮するには、有効語のうち、互いにかけ離れた意味を持つ語を基底語として選択し、基底語を軸として有効語ベクトルを算出すればよい。基底語選択のアルゴリズムを以下に示す。

1. 重要度 $w(V_i)$ が最も高い有効語 V_i を基底語 B_1 とする。
2. K 個の基底語 $B_1, B_2, \cdots, B_K$ が既に選択されているとき、次式で定義される値 $w'(V_j)$ が最も高くなる有効語 V_j を B_{K+1} とする。
   $$w'(V_j) = w(V_j) \cdot min_l \{1 - c(V_j, B_l)\} \quad (l = 1, 2, \cdots, K)$$
3. $w'(V_j)$ が重要度の最低値を下回らない限り、2. を繰り返し、基底語を選択する。

N' 個の基底語 $B_1, B_2, \cdots, B_{N'}$ が選ばれたとき、有効語 V_i は次のように N' 次元ベクトル $\vec{v_i}$ で表現される。

$$\vec{v_i} = (c(V_i, B_1), c(V_i, B_2), \cdots, c(V_i, B_{N'}))$$

2.2.2 文書ベクトルとカテゴリの判定

文書 D を表すベクトル $\vec{d}$ は文書 D に含まれる有効語のベクトルの和で求められる。

$$\vec{d} = \sum_{i=1}^{N} tfreq(V_i, D) \cdot \vec{v_i}$$

有効語ベクトルを足し合わせるときに、文書内での出現位置や文章中における言語的役割などにより重みをつけることで、より内容を反映したベクトルを作成することができる。

文書の内容を反映したベクトル空間が生成されると、同一カテゴリに属する文書のベクトルは互いに近くに存在する。$dnum(C)$ 個の文書を含むカテゴリ C のベクトル $\vec{a}$ は、各文書のベクトル $\vec{d}_1, \vec{d}_2, \cdots, \vec{d}_{dnum(C)}$ の平均で表される。

$$\vec{a} = \frac{1}{dunm(C)} \sum_{i=1}^{dnum(C)} \vec{d_i}$$

ある文書 D がカテゴリ C に属するか否かは、両者の類似度と閾値 α との大小関係により決定する。類似度 $sim(D, C)$ は、文書ベクトル $\vec{d}$ とカテゴリベクトル $\vec{a}$ とのなす角の余弦値で表す。

$$sim(D, C) = \frac{\vec{d} \cdot \vec{a}}{\|\vec{d}\| \cdot \|\vec{a}\|} \geq \alpha \Rightarrow D \in C$$

$$sim(D, C) = \frac{\vec{d} \cdot \vec{a}}{\|\vec{d}\| \cdot \|\vec{a}\|} < \alpha \Rightarrow D \notin C$$

2.3 評価指標

検索システムや分類システムは、再現率（recall）と適合率（precision）により評価される。

$$\text{再現率} = \frac{\text{結果文書のうちの正解文書数}}{\text{結果とすべき文書数}}$$

$$\text{適合率} = \frac{\text{結果文書のうちの正解文書数}}{\text{結果として返した文書数}}$$

再現率は検索漏れの少なさを表し、適合率は検索ノイズの少なさを表す。理想的なベクトル空間が生成された場合には、結果とすべき文書とそれ以外の文書の文書ベクトルは離れて位置するため、再現率と適合率を共に１とするような閾値 α が存在する。しかし、一般に再現率と適合率はトレードオフの関係にあり、閾値 α の変動によって互いに逆の増減を示す。高い再現率を維持した状態で適合率を向上できるようなベクトル空間の生成が本技術の研究目標であるといえる。

3 応用事例

本節では、２節で概説した技術の応用事例として５つのシステムを紹介する。3.1節は電子メール、3.2節はフロー情報を掲載するホームページ、3.3節、3.4節、3.5節はサーチエンジンに登録されたホームページを対象とした情報収集支援システムである。

3.1 メールフィルタリング

現在、電子メールは様々な用途に利用されている。メーリングリスト経由で送付されるメールやダイレクトメールなどを含めると、その数は膨大なものになるが、必ずしもすべてのメールが必要とは限らない．新着メールの内容を踏まえ、必要なメールと不要なメールに自動分類するシステムは、我々のメール処理作業を簡易化すると考えられる。

受信メールの要・不要により、差出人は次の３種類に大別される。

【差出人 A】すべてのメールが必要な差出人

【差出人 B】必要なメールと不要なメールが混在する差出人

【差出人 C】すべてのメールが不要な差出人

差出人 A および差出人 C からのメールの分類は、メールのヘッダ情報を参照することで容易に実現できる。差出人 B からのメールの内容を把握し、要・不要を判断する場合に、ベクトル空間モデルが利用できる。

3.1.1 システムの概要

メールソフトAL-Mail32 Ver1.11上で動作するメールフィルタリングシステムを Java により構築した[1]。システムは学習部と判定部から構成される。

学習部では、あらかじめ要・不要の２カテゴリに

分類された既存メールを用いて、差出人 B からのメールを分類するためのベクトル空間を生成する。差出人 B の各人に対して、要・不要の２カテゴリの単語間共起に基づいて有効語を抽出し、有効語ベクトルとカテゴリベクトルを生成する。

判定部では新着メールの分類を行なう。まず、A～C のいずれの差出人であるかを判別し、A または C である場合にはそれぞれ「必要カテゴリ」、「不要カテゴリ」に分類する。学習データに存在しない差出人からのメールは必要と判定する。差出人 B である場合は、該当する有効語ベクトルを用いてメールをベクトル表現し、要・不要の２つのカテゴリベクトルとの類似度を算出して、より高い方のカテゴリへと分類する。有効語が１つも含まれないメールは「判定不能」とする。

メールソフトを利用しない時間に学習部を実行し、通常は判定部のみを稼動すると、実時間で処理が可能なフィルタリングが実現される。判定不能メール、および誤分類メールをユーザがその都度訂正し、有効語ベクトルとカテゴリベクトルを毎日更新することで、学習データの増加による分類精度の向上が期待される。また、ユーザの分類視点の時間変化にも追従することが可能となる。

3.1.2 評価実験

実際に送受信されたメールを用いて評価実験を行なった。大学生である６名の被験者 A～F が15名の差出人からのメールを10通ずつ、計150通を収集し、それぞれの分類視点で要・不要を判定する。150通のうち、100通を学習部で用いる既存データとし、判定部で残り50通を分類したときの再現率と適合率を表１に示す。

この結果、80%程度の再現率と適合率でメールフィルタリングが可能であることがわかる。６名の被験者のうち、被験者 A の適合率が極端に低くなっている。被験者 A の受信したメールには、他の被験者のメールと比較して次のようなメールが多く存在した。

・非常に短い文章のメール

表1. メールフィルタリング結果

被験者	A	B	C	D	E	F	平均
再現率[%]	76.9	61.9	73.3	100.0	76.0	89.5	79.6
適合率[%]	41.7	86.7	73.3	95.0	86.4	94.4	79.6

・主に数値で構成される株価情報のメール

両者とも本手法では有効語抽出が適切に行なえない種類のメールであり、特徴的な単語が選択できなかったために、必要カテゴリと不要カテゴリの差異を表現できなかった例といえる。さらに精度の高いフィルタリングシステムを構築するには、数値データや文章のパターンなどを考慮すべきである。

3.2 重要記事抽出

メールは個人宛に送付される情報であるが、不特定多数の人を対象として提供されるフロー情報に新聞記事がある。現在では、新聞社のホームページに記事が掲載され、最新のニュースを容易に閲覧できるが、重要な記事を一目で判別することは難しい。多忙時の重要ニュース確認には、重要記事の自動抽出システムが有用である。新聞記事もメールと同様に個人の視点で収集する場合が多く、ユーザの視点に基づく新聞記事自動分類システムが開発されているが[3,4,5,6,10]、ここでは一般的な視点で重要と判断される記事の抽出を目指す。

重要なニュースは多くの新聞社が記事にすると考えられる。複数の新聞社から収集した記事のうち、ベクトル空間モデルで内容が近いと判定された記事は同一ニュースを扱っているとし、すべての新聞社の記事として出現するニュースを重要と判断する。

3.2.1 システムの概要

新聞社のホームページから新聞記事を自動的にダウンロードし、重要記事を抽出するシステムを Java により構築した[9]。各記事は式(2)と式(4)を用いた tf･idf 法によりベクトル表現される。全記事に含まれる単語の重要度を算出し、記事ごとに重要度の高い語を指定された個数ずつ選択して有効語とする。有効語をもとに生成された各記事ベクトルの類似度を求め、閾値 α 以上である場合に同一ニュースと判定する。すべての新聞社の記事となっている

図1　重要記事出力画面

ニュースを重要と判断し、該当記事を提示する。出力画面の例を図1に示す。

3.2.2　評価実験

日本経済新聞、朝日新聞、読売新聞のホームページ上に掲載されている国際・社会グループの記事のうち、2002年1月7日から9日までの3日分の記事を使用して実験を行なった。重要記事の正解データは、記事内容を人手で確認して作成した。各日の記事数および重要記事数は表2の通りである。

システムは重要記事の再現率と適合率により評価する。選択する有効語の個数と閾値 α を変化させて行なった予備実験で、再現率と適合率は有効語個数が10個になるとほぼ収束することがわかった。また、α を0から1まで0.05刻みで変化させると、0.3で再現率と適合率がほぼ等しくなった。予備実験の結果より、本実験では各記事から10個ずつ有効語を選択し、閾値を0.3として同一ニュースの判定を行なった。

実験で得られた再現率と適合率を表2に示す。日によって多少の高低はあるものの、高い精度で重要記事を抽出できたことがわかる。本システムにより提示される重要記事は全記事の35%程度であり、重要ニュースの確認作業における効果が期待される。

表2. 重要記事抽出結果

日付	1/7	1/8	1/9
記事数	121	127	131
重要記事数	45	36	41
再現率[%]	95.6	86.1	85.4
適合率[%]	97.7	91.2	89.7

3.3　検索結果分類

サーチエンジンでホームページを検索すると、あまりの検索結果の多さに愕然とする場合が多い。検索結果は一般的な有用度でランキングされているものの、ページタイトルと要約の羅列だけでは、どれが自分の探しているページなのか見当をつけることもできない。

検索キーワードとして一般的な語を入力した場合だけでなく、人名や地名など固有名詞を入力した場合にも同様である。例えば「大谷紀子」で検索すると、武蔵工業大学教員の大谷紀子だけでなく、漫画家、医者、陸上選手、芸術家などの大谷紀子さんに関するページも結果として返される。この場合、すべての「大谷紀子」について知りたいわけではなく、特定の「大谷紀子」に関する情報が得られればよいので、検索結果を各「大谷紀子」ごとに分けて表示することで、目的のページを探す作業が容易になる。

特定の人物に関するページでは、共通の話題が取り上げられる。同姓同名でも別人であれば、ページで扱われる話題は異なるはずである。ベクトル空間モデルにより検索結果ページをベクトル表現し、類似する話題のページに登場する人を同一人物と見なす。本手法は、人名以外の語をキーワードとする検索結果にも有効な方法である。サーチエンジンで得られた検索結果ページをクラスタリングすることにより、同一キーワードを含むページを話題や目的、種類等で分類して提示することが可能となる。

3.3.1　システムの概要

サーチエンジン goo[2] の検索結果を内容ごとに分類して提示するシステムを Java により構築した[11]。goo により得られた検索結果ページのうち、HTML ファイルを処理対象とする。本文から HTML タグを除去し、式(5)と式(6)を用いた tf·idf 法によりベクトル表現する。

各ページ間の類似度を算出し、類似度が閾値 α

```
n := 1;
C1 := {D1};
for (each i ∈ {2, 3, ···, M}) {
  flg := false;
  for (each j ∈ {1, 2, ···, n}) {
    if (sim (Di , Cj ) > α) {
      Cj := Cj ∪ {Di };
      flg := true;
      break;
    }
  }
  if (flg = false ) {
    n := n + 1;
    Cn := {Di };
  }
}
```

図２　クラスタリングアルゴリズム

以上であるページ同士を同一グループと判断する。クラスタリングアルゴリズムを図２に示す。ここで、検索結果ページを $D_1, D_2, \cdots, D_M$ とする。検索結果は、類似する話題のページを集めたグループごとに提示される。

3.3.2　評価実験

東武野田線の運河駅の時刻表を得ることを目的として、「東武野田線」「運河」「時刻表」という３つの単語を検索キーワードとして検索を行なったところ、25件の結果が C_1～C_4 の４グループに分類された。出力画面を図３に示す。

各グループのページには次のような情報が掲載されていた。

C_1：運河近辺を走る路線バスの時刻表

C_2：運河駅の電車の時刻表

C_3：地域情報、東京理科大学の情報等

C_4：運河以外の駅の電車の時刻表

有効語の出現頻度が低く、他のグループに属すると

図３　検索結果分類の出力画面

判断できないページが C_3 に集まった。本検索の目的に合致するグループは C_2 であり、探索対象が25件から６件に絞られたといえる。

次に、「CD-RW」をキーワードとして検索を行なったところ、30件の結果が C_1～C_4 の４グループに分類された。各グループの特徴は以下の通りである。

C_1：CD-RW ドライブ製品紹介

C_2：英語で書かれた情報

C_3：ドライバ、CD-RW 搭載 PC 等

C_4：CD-RW に関する知識・技術説明

前例と同様、C_3 に雑多な情報が分類された。本システムでは、処理対象を日本語に限定した手法を用いているため、英語で書かれたページの分類は難しいが、本結果のように英語のページのみを１グループにまとめることは可能である。サーチエンジンには検索結果を日本語のページに限定する機能も備わっているが、ほとんどが英語で書かれており、日本語を多少含むようなページは除外できない。日本語で書かれた情報を検索したい場合には、このようなページは探索対象外とするべきなので、本システムの分類が有用であると思われる。

3.4　絞込検索のための検索キーワード提示

3.3節では、検索結果は変更せず、提示方法を工

2　http : //www.goo.ne.jp/

夫することで、目的のページを探索する労力の軽減を図った。一方、検索キーワードを追加して再度検索を行ない、検索結果数を減少させることによっても、探索は簡易化されると考えられる。絞込検索の効果は、どのようなキーワードを追加するかによって決まる。最初の検索で入力した初期検索キーワードと類似した語を追加すると検索結果を絞り込むことはできず、意味のかけ離れた語を追加すると目的のページが漏れてしまう。

検索結果を適切に絞り込むキーワードとは、特定の検索結果文書においてのみ重要な語である。つまり、文書において出現頻度が高く、特定の文書に偏在する語が絞込検索に有効といえる。この特徴はtf・idf 法で選択された有効語の特徴と一致する。tf・idf 法による重要度の高い語を追加検索キーワード候補として提示することで、絞込検索における支援が可能となる。

3.4.1 システムの概要

サーチエンジン Google[3]で得られた検索結果に対し、絞込検索のための追加検索キーワード候補を提示するシステムを C 言語により構築した[2]。Google により得られた検索結果ページのうち、HTML ファイルを処理対象とする。HTML タグを除去した本文から名詞、形容動詞の語幹、サ変動詞の語幹を抽出し、重要度を式(3)と式(6)を用いた tf・idf 法により算出する。重要度の高い30語を追加すべき検索キーワードとして提示する。

3.4.2 評価実験

システムの有効性を示すために、男子大学生10名を被験者として実験を行なった。実験において、被験者は次の条件下で与えられた課題に解答することを要求される。

【条件１】Google を利用する。

【条件２】本システムで提示された追加検索キーワード候補を見ながら Google を利用する。

まず、条件１で解答を試み、次に条件２で解答を行なう。解答の制限時間は両条件で各10分とし、初期検索キーワードは１語に限定する。条件２で提示される追加検索キーワード候補は、条件１における初期検索キーワードの検索結果に対して生成されたものである。被験者には次の２つの課題を与える。

【課題１】日本の高速道路にあるトンネルの中で、全長が２番目に長いトンネルがある高速道路名を答えよ。

【課題２】神奈川県横須賀市にある小学校の数を答えよ。

実験の結果、いずれの課題の解答においても、条件２の下で絞込検索を行なったすべての被験者は、提示された追加検索キーワード候補のうちのいずれかの語を使用した。

課題１の解答に際し、各被験者の用いた初期検索キーワード、条件１での課題達成可否、条件２で絞込検索に使用したキーワード、条件２での課題達成可否をまとめたものが表３である。追加検索キーワードの括弧内の数字は、提示した30候補内での重要度の順位を示している。

条件１ではすべての被験者が正解を見つけることができなかったが、条件２では１名を除く被験者が絞込検索により正解にたどりつくことができた。絞込検索は追加検索キーワード候補のいずれかの語を選択して行なわれ、独自に考えた語を追加する被験者はいなかった。被験者 D は条件１での検索の困難さに嫌気がさし、絞込検索を行なわずに解答を放棄した。被験者 A と被験者 J は、条件２で最初に「tunnel」を追加検索キーワードとして選択したが、うまく絞り込むことができず、キーワードを変更して正解に到達した。

課題２の結果を表４に示す。各条件の達成可否に続く括弧内の数字は、正解の記述されたページが検索結果において何位にランキングされていたかを表している。

すべての被験者が両条件において課題を達成することができた。しかし、正解が掲載されているページの順位が条件１では15位または23位だったのに対

3　http : //www.google.com/

表3．追加検索キーワード提示の課題１における結果

被験者	初期検索キーワード	条件１	追加検索キーワード	条件２
A	トンネル	×	tunnel（３），日本（10）	○
B	高速道路	×	一覧（11）	○
C	高速道路	×	一覧（11）	○
D	高速道路	×	-	×
E	トンネル	×	日本（10）	○
F	トンネルの長さ	×	日本一（６）	○
G	高速道路	×	日本一（６）	○
H	高速道路	×	一覧（11）	○
I	トンネルの長さ	×	日本一（６）	○
J	トンネル	×	tunnel（３），日本一（６）	○

表4．追加検索キーワード提示の課題２における結果

被験者	初期検索キーワード	条件１	追加検索キーワード	条件２
A	横須賀市	○(23)	地域情報（４）	○(７)
B	横須賀市	○(23)	地域情報（４）	○(７)
C	横須賀市	○(23)	地域情報（４）	○(７)
D	神奈川県横須賀市	○(15)	高等学校（２）	○(６)
E	横須賀市	○(23)	地域情報（４）	○(７)
F	横須賀市	○(23)	地域情報（４）	○(７)
G	横須賀市	○(23)	地域情報（４）	○(７)
H	神奈川県横須賀市	○(15)	高等学校（２）	○(６)
I	横須賀市	○(23)	地域情報（４）	○(７)
J	横須賀市	○(23)	地域情報（４）	○(７)

し、条件２では６位または７位であった。順位が上位であるほど検索結果からの探索は容易であるので、システムが提示したキーワードを使って絞込検索をすることで、目的のページの探索が効率化されたといえる。

3.5 追加検索キーワード選択支援

3.4節のシステムでは、追加検索キーワード候補のみを提示するため、どのキーワードを追加するとどの程度絞り込めるのかがわからず、追加検索キーワードの選択は主観に頼らざるを得ない。また、絞込検索の結果が目的に沿っていない場合、別のキーワードを追加して絞込検索をやり直すことになるが、追加すべきキーワードを選択するための判断材料がまったくない。

前者の問題については、追加検索キーワード候補と共に重要度と絞込件数を提示すればよい。絞込件数とは当該キーワードを追加したときの検索結果数を表す。後者の問題は、各キーワード候補で絞り込まれるページ集合をカテゴリとみなし、カテゴリ間の類似度を提示することで解決される。閲覧中の絞込検索結果と、他のキーワードによる絞込検索結果の類似度を同時に参照することによって、再検索における追加検索キーワード選択をスムーズに行なうことができる。

3.5.1 システムの概要

サーチエンジン Excite[4]で得られた検索結果に対し、上記機能を追加した追加検索キーワード提示システムを Java により構築した[8]。追加検索キー

図４　追加検索キーワード提示画面

ワードは3.4節のシステム同様に決定し、重要度順に整列して、重要度と絞込件数と共に提示する。提示画面例を図４に示す。

検索結果ページは tf・idf 法によりベクトル表現される。各キーワード候補による絞込検索の結果ページをカテゴリとみなし、カテゴリベクトルを算出する。ユーザが提示されたキーワード候補を使って絞込検索を行なったとき、現在閲覧している絞込検索結果のカテゴリベクトルと、他のキーワードによる絞込検索結果のカテゴリベクトルの類似度を提示する。図４で「歌舞伎」を選択して絞込検索を行なったときの類似度提示画面を図５に示す。

3.5.2　評価実験

本システムの有効性を検証するための実験を大学生17名を被験者として行なった。本システムでは、検索対象に関連する情報はどのようなカテゴリに分類され、それらはどの程度類似しているのかが提示されるため、ユーザの検索対象に関する知識が少なく、理解が浅い場合に、本システムはより大きな効果を発揮する。

事前調査の結果、どの被験者も「NGO」という言葉は知っているが、詳細な知識は持ち合わせていないことがわかった。この結果を踏まえて、被験者に「NGO の動きについて調査せよ」という課題を与え、本システムおよび Excite を利用しながら課題に取り組むよう指示した。初期検索キーワードは

4　http : //www.excite.co.jp/

図５　カテゴリ間の類似度提示画面

「NGO」に限定し、 その後は自由に検索を行なって、被験者の満足する結果が得られた時点で終了とする。

作業後、次の３つの点について被験者の評価を調査した。

・追加検索キーワード候補の使用可能性
・提示内容の有効性
・システム全体の実用性

追加検索キーワード候補の使用可能性は、30個のキーワード候補のうち、使用したいと感じたキーワードの個数で表す。システム全体の実用性、提示内容の有効性については、1(とても低い)、2(やや低い)、3(普通)、4(やや高い)、5(とても高い）の５段階で評価する。各被験者の評価値の平均を表５に示す。

この結果、提示内容とシステム全体について標準値３を超える評価が得られた。提示したキーワードのうち18%程度が「使用したい」と感じるキーワードであった。すべてが同程度に使用したいと思う語であると、その中から実際に利用する語を選択するための負荷が生じる。少数に限定せず、選択の負荷が重くならない程度のキーワードに使用可能性が見出されるべきとすると、１/５程度という値は適切な値であると考える。

被験者の作業プロトコルには、最初は有用でない

表5. 追加検索キーワード提示に対する評価

キーワード候補	提示内容	システム
5.41	3.59	3.88

と感じていたキーワードに対して、類似度を基準として複数のカテゴリを閲覧する過程で有用性を見出し、最終段階で絞込検索に利用して目的の検索結果を得る様子が記録された。「絞込検索の結果が事前に確認でき、検索の無駄が省ける」と感想を述べる被験者もいたことからも、本システムの提示内容がユーザの検索作業を支援したことがわかる。

また、カテゴリ間の類似度を基準に検索キーワードを選択する際、閲覧したカテゴリとの類似度が極端に高いカテゴリや低いカテゴリに関しては迷うことなく選択する反面、類似度が0.4〜0.7のカテゴリに対しては選択に迷う姿が見られた。絞込検索をやり直すときには、「類似した結果が欲しい」「まったく異なる結果が欲しい」のいずれかの目的があるためと考えられる。現在のように類似度を数値で表示すると、類似関係の方向性を示すことはできないが、表現の視覚的効果を工夫することで、ユーザが微妙な目的の違いを反映できるシステムが実現すると思われる。

4　おわりに

ベクトル空間モデルを用いたさまざまな応用事例とその有効性を示す実験結果を紹介した。今後、我々を取り巻く電子化文書は間違いなく増加の一途を辿り、今以上に多種多様な情報が氾濫するであろう。大量の情報に埋もれることなく、その価値を最大限に引き出し、我々の生活を豊かにするために活用できるよう、さらなる情報検索技術の研究が必要と考える。

参考文献

［1］原　寛斉．ユーザー適応型メールフィルタリングシステムに関する研究，東京理科大学，2001.

［2］井上大輔．絞込検索のための検索語提示に関する研究．東京理科大学，2001.

［3］伊藤史朗，大谷紀子，柴田昇吾，上田隆也，池田裕治．フロー情報収集・活用のための知的検索システム fit（2）処理方式．情報処理学会第53回全国大会予稿集2T−9，1996.

［4］大谷紀子，伊藤史朗，柴田昇吾，上田隆也，池田裕治．フロー情報収集・活用のための知的検索システム fit（3）類似度判定．情報処理学会第53回全国大会予稿集2T−10，1996.

［5］大谷紀子，伊藤史朗，柴田昇吾，上田隆也，池田裕治．知的検索システム fit での類似度判定の改良．情報処理学会第55回全国大会予稿集5N−6，1997.

［6］N. Otani, F. Itoh, S. Shibata, T. Ueda, and Y. Ikeda. An Information Retrieval System based on Personal Viewpoints for Everyday Use. In *Proc. of KES '98*, pp.397-404, 1998.

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

［8］佐々木寛．絞込検索のための検索語選択支援に関する研究．東京理科大学，2002.

［9］佐藤　武．重要記事抽出システムに関する研究．東京理科大学，2002.

［10］上田隆也，大谷紀子，伊藤史朗，柴田昇吾，池田裕治．フロー情報収集・活用のための知的検索システム fit（1）コンセプト．情報処理学会第53回全国大会予稿集2T−8，1996.

［11］吉村大介．ロボット型サーチエンジンの検索結果の分類．東京理科大学，2001.

［12］湯浅夏樹，上田　徹，外川文雄．大量文書データ中の単語間共起を利用した文書分類．情報処理学会論文誌，Vol.36，No.8，pp.1819−1827，1995.