情 報 学 基 礎 61－2
自 然 言 語 処 理 142－2
（2001． 3． 5）

ブーリアンクエリのベクトル展開

川谷隆彦

日本ヒューレット・パッカード㈱ ヒューレット・パッカード研究所

takahiko_kawatani@hp.com

本報告の目的は検索方法としてのベクトル空間モデルが有する索引語間の関係を記述できないという欠点を解消することにある。所与のクエリに用いられた各索引語の有無を成分とするベクトルを索引ベクトルとし、所与のクエリと適合する文書がとりうる全ての索引ベクトルの集合と、被検索文書の索引ベクトルとの間で SVSM 類似度を求めてクエリに対する適合の度合いとする基本的な手法、及び基本的な方法での望ましくない結果をフィードバックさせる手法を提案する。AND、OR、NOT などの関係で結ばれた 8 語から成る幾つかのブーリアンクエリに対して模擬実験を行い、全ての例で適合文書と非適合文書を完全に分離することができた。

Vector Expansion of a Boolean Query

Takahiko KAWATANI

Hewlett-Packard Labs Japan, Hewlett-Packard Japan

takahiko_kawatani@hp.com

To cover the drawback in the vector space model that it can not describe the relationship between query terms, this paper proposes a fundamental method and a learning one. Let us consider a vector whose components represent whether each query term in a given query is extracted as an index word. In the former, SVSM-similarity between the vector of a document and a set of the vectors of all documents relevant to the query is obtained as the degree of relevance. In the latter, the undesirable results on the former method are feedbacked. Through simple experiments using some Boolean queries in which query terms are described using AND, OR and NOT, it was confirmed that relevant documents can be perfectly separated from irrelevant ones.

1.まえがき

近年の情報流通の増大に伴い、情報検索技術の重要性はますます高まっている。情報検索技術としては、これまで、ブーリアンモデル、拡張ブーリアンモデル、ファジィモデル、ベクトル空間モデル、確率モデル、ネットワークモデルなどが知られているが[1][2][3][4]、これらの中でブーリアンモデルは最も古典的かつ基本的な方法であり、ベクトル空間モデルは最も一般的なものとなっている。ブーリアンモデルと比べ、ベクトル空間モデルは一般に

(A)索引語の重み付けの容易性

(B)検索結果のランキングの容易性

(C)検索結果のフィードバックの容易性
において優れ、
(D)索引語間の関係の記述能力
において問題があると言われている。検索モデルとしては、両者の長所を併せ持つようなものが望ましい。そのようなモデルは現在まで知られてないように思われる。本報告は、ベクトル空間モデルにおいて索引語間の関係が反映されるよう改善し、ベクトル空間モデルの利点は生かしつつ、ブーリアンモデル並みの適合／非適合文書の分離能力を実現することを目的とする。

本報告では、問題を以下のように設定する。先ず、N 個の索引語から成るブーリアンモデルによるクエリを $Q(w_1,\ldots,w_N)$（以下 Q と略す）とする。また、文書 i の索引ベクトル（以下、単にベクトルともいう）を $\mathbf{f}_i=(b_1,\ldots,b_N)$で表す。$b_n$ は w_n が文書 i に用意された索引語集合の中に存在するか否かを示す 2 値の変数である。問題は、Q を N 次元の複数のベクトルに変換し、各文書が Q と適合するか否かの判断をベクトル $\mathbf{f}_i$ と変換されたベクトルとの照合によって可能にすることである。

この問題を解くための考え方は以下の通りである。先ず、N 個の索引語が与えられたとき、索引ベクトルは理論上 2^N-1 通り存在することになるが、これらのうち、Q と適合する文書が有する索引ベクトルを適合索引ベクトル、もしくは適合索引と呼ぶこととし、存在しうる全ての適合索引ベクトルの集合をΩ_1、その補集合をΩ_0とする。このようにすると、あらゆる文書のベクトルはΩ_0もしくはΩ_1に含まれることになる。従って、任意の文書 i がクエリに適合するか否かは、文書 i のベクトル $\mathbf{f}_i$ がΩ_1、Ω_0の何れに含まれるかにより決まることになる。本報告では、$\mathbf{f}_i$ とベクトル集合Ω_1との間の類似度を求め、類似度が一定値を越えていれば $\mathbf{f}_i$ はΩ_1 に含まれると判断することとする。さらに、類似度の尺度としては SVSM 類似度を用いることとする。SVSM 類似度は筆者が先に提案した文ベクトル集合モデル（Sentence Vector Set Model : SVSM）[5]に基づくものであり、Ω_1 に含まれる全ベクトルの平方和行列（後述）の固有ベクトルと固有値を用いる事により $\mathbf{f}_i$ とΩ_1 に含まれる全てのベクトルとの照合を実質的に実現するものである。

以下、**2.**では本報告のベースとなる SVSM 類似度について概要を紹介する。**3.**では先ず基本的な手法として学習に依らない方法の概要と実験結果について述べる。**4.**では、**3.**における基本的な方法において望ましくない結果を与えるベクトルを平方和行列にフィードバックし精度を更に向上させる手法の概要と実験結果について述べる。**5.**では、ベクトル空間モデルが本来有する利点が損なわれていないかどうかの考察、高速化の可能性などについて述べる。**6.**で纏めを行う。

2. 文ベクトル集合モデル

本章では、SVSM 類似度とそのベースとなる文ベクトル集合モデルを紹介する。文ベクトル集合モデルでは、文書を構成する各文ごとにベクトルを構成し、その集合として文書を表す。本章で云う“文ベクトル”は前章の“索引ベクトル”に対応する。

2.1 SVSM 類似度

文ベクトル集合がそれぞれ$\{\mathbf{d}_1,\ldots,\mathbf{d}_M\}$、$\{\mathbf{t}_1,\ldots,\mathbf{t}_J\}$で与えられる文書 D、T を考える。SVSM 類似度は、文書 D、T 間の類似度 r をそれぞれの文ベクトルの全ての組み合わせについて求められる内積の 2 乗和をもとに以下のように定義される。

$$r=\left(\frac{\sum_{m=1}^{M}\sum_{j=1}^{J}(\mathbf{d}_m{}^T\mathbf{t}_j)^2}{\sqrt{\sum_{m=1}^{M}\sum_{j=1}^{M}(\mathbf{d}_m{}^T\mathbf{d}_j)^2\sum_{m=1}^{J}\sum_{j=1}^{J}(\mathbf{t}_m{}^T\mathbf{t}_j)^2}}\right)^{1/2} \tag{1}$$

T は転置を表す。ところで式(1)のカッコ内の分子は以下のように変形できる。

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{j=1}^{J}(\mathbf{d}_m{}^T\mathbf{t}_j)^2\\&=\sum_{m=1}^{M}\sum_{j=1}^{J}\mathbf{t}_j{}^T\mathbf{d}_m\mathbf{d}_m{}^T\mathbf{t}_j\end{aligned} \tag{2}$$

ここで、

$$S=\sum_{m=1}^{M}\mathbf{d}_m\mathbf{d}_m{}^T \tag{3}$$

で定義される行列 S を平方和行列と呼ぶこととす

る。Sは単語間の共起の程度を表す行列である。SのランクをR_D、そのk次の固有値、固有ベクトルを$\lambda_k(\geq \lambda_{k+1})$、$\phi_k$とする。とすると、Sは

$$S=\sum_{k=1}^{R_D}\lambda_k\ \phi_k\phi_k^T \qquad (4)$$

のように展開されるので、これを式(2)に代入すると、

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{j=1}^{J}\mathbf{t}_j^T\mathbf{d}_m\mathbf{d}_m^T\mathbf{t}_j\\&=\sum_{k=1}^{R_D}\sum_{j=1}^{J}\lambda_k(\phi_k^T\mathbf{t}_j)^2\end{aligned} \qquad (5)$$

のようになる。さらに、文書Tにおける平方和行列のランクをR_T、k次の固有値、固有ベクトルを$\gamma_k(\geq\gamma_{k+1})$、$\tau_k$とすると、

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{j=1}^{J}(\mathbf{d}_m^T\mathbf{t}_j)^2\\&=\sum_{m=1}^{R_D}\sum_{j=1}^{R_T}\lambda_m\gamma_j(\phi_m^T\tau_j)^2\end{aligned} \qquad (6)$$

のように変形される。従って、式(1)の類似度は

$$r=\left(\frac{\sum_{m=1}^{R_D}\sum_{j=1}^{J}\lambda_m(\phi_m^T\mathbf{t}_j)^2}{\sqrt{\sum_{m=1}^{R_D}\lambda_m^2\sum_{m=1}^{J}\sum_{j=1}^{J}(\mathbf{t}_m^T\mathbf{t}_j)^2}}\right)^{1/2} \qquad (7)$$

もしくは

$$r=\left(\frac{\sum_{m=1}^{R_D}\sum_{j=1}^{R_T}\lambda_m\gamma_j(\phi_m^T\tau_j)^2}{\sqrt{\sum_{m=1}^{R_D}\lambda_m^2\sum_{j=1}^{R_T}\gamma_j^2}}\right)^{1/2} \qquad (8)$$

のように書き直される。

式(1)、(7)、(8)は同じ文書に対して同一の結果を与えるものであり、決して異なる定義ではない。何れにせよ、SVSM類似度は両方の文書の全ての文の組み合わせにより求められるので、非常に正確な類似度の尺度となっている。

2.2 固有値、固有ベクトルの解釈

ここで平方和行列Sの固有値、固有ベクトルの性質と解釈について述べておきたい。

(1) 各固有ベクトルは各単語の頻度の線形結合で表現されるベクトルなので、それ自身概念を表す。ϕ_1は文ベクトル集合をただ1つのベクトルで近似したときの2乗誤差を最小にする軸、言い換えれば各文ベクトルを射影したときの射影値の2乗和を最大にする軸であるので、ϕ_1の方向は各文に最も共通する概念を表すと云える。固有ベクトルは文書固有に決まるので、ϕ_1は文書の概念を最も良く代表する固有概念を表すことになる。ϕ_2はϕ_1と直交するという条件のもとで射影値の2乗和を最大にする軸であるので、その方向は2番目に文書を代表する固有概念ということになる。3次以降も同様である。k次の固有概念を持つ文をここではk次の固有文と呼ぶ。

(2) 固有値λ_kは各文ベクトルのϕ_kへの射影値の2乗和そのものである。ここで、各文はそのベクトルのノルムの2乗で与えられるエネルギーを持つものとする。また、文書は文エネルギーの総和で与えられるエネルギーを持つものとする。とすると、固有値λ_kは文書がϕ_k方向に持つエネルギーということになる。従って、固有値λ_kはk次の固有文のエネルギーを表すと解釈でき、k次の固有文のベクトルは$\sqrt{\lambda_k}\ \phi_k$で与えられることになる。

(3) $\sum_{k=1}^{R}\lambda_k$ は文書本来のエネルギーと等しいので[5]、$\lambda_k/\sum_{k=1}^{R}\lambda_k$ はk次の固有文のエネルギーが文書全体のエネルギーに対して占める割合を示す。これは、k次の固有文の文書全体に対する代表度とみなすことができる。

(4) 一般に高次になるほど固有値の値は小さくなるのでL+1次以降は無視し、L次までの固有値、固有ベクトルで近似することができる。1〜L次の固有ベクトルが張る空間をL次元の概念部分空間と呼ぶ。$\sum_{k=1}^{L}\lambda_k/\sum_{k=1}^{R}\lambda_k$ は文書の最も重要なL個の固有文の文書全体に占めるエネルギーの割合を示し、L次元概念部分空間の代表度とみなすことができる。この値はLの値を具体的に決定するときの目安とすることができる。

(5) 以上のことから、式(7)は文書Dの固有文と文書Tの文ベクトルとの照合、式(8)は固有文同士の照合により類似度を求めていることが分かる。さらに、文書D、TをL_D、L_T次元の概念部分空間で表した場合には文書間類似度は、式(7)、(8)の分

子のR_D、R_TをL_D、L_Tにより置き換えることにより与えられる。このようにすると、さ程精度を落とさずに式(7)、(8)の計算量を削減することが可能となる。

3. 提案手法（基本的な方法）

3.1 方法の概要

1.で述べたように、文書 i がクエリ Q と適合するか否かの決定は $\mathbf{f}_i$ とベクトル集合Ω_1 との間の類似度r_iに基づいて行う。提案手法では、**2.1** における文書 D の文集合は適合索引ベクトル集合Ω_1に、文書 T は$\mathbf{f}_i$に対応するものとして、類似度としては式(7)を採用することとする。従って、類似度は

$$r_i = \left[\frac{\sum_{m=1}^{L} \lambda_m (\boldsymbol{\phi}_m{}^T \mathbf{f}_i)^2}{\sqrt{\sum_{m=1}^{R} \lambda_m{}^2} \|\mathbf{f}_i\|^2} \right]^{1/2} \qquad (9)$$

で与えられる。ここで、λ_m、ϕ_mは

$$S = \sum_{\mathbf{f} \in \Omega_1} \mathbf{f}\mathbf{f}^T \qquad (10)$$

で定義される平方和行列 S の m 次の固有値、固有ベクトルである。また、L は部分空間の次数であり、最大値は行列 S のランク R となる。そして、αを閾値として、$r_i > \alpha$であれば文書 i はクエリ Q と適合すると判断する。

3.2 実験方法

Q として暫くの間、8 索引語から成る

$$Q = (w_1 \text{ OR } w_2 \text{ OR } w_3 \text{ OR } w_4) \text{ AND } (w_5 \text{ OR } w_6) \text{ AND } w_7 \text{ AND } w_8 \qquad (11)$$

を考える。この Q に対する適合索引数は 45 である。評価は再現率と精度の重みを等しくした F 値を用いた。2^N-1 通りの索引ベクトルの中で、適合索引ベクトル数を A、適合と判断された索引ベクトル数を B、B の内での適合索引ベクトル数を C とすると、再現率は C/A、精度は C/B で定義される。本実験では存在しうる全ての索引ベクトルを用いて評価するので適合索引ベクトルのみを適合と判断できれば、実際の検索の場においても適合文書と非適合文書とを完全に分離できたことになる。また、本手法の結果は、余弦類似度を用いた通常のベクトル空間モデルの結果と比較する。この場合、ベクトルとしては、Ω_1に含まれる全ベクトルの平均ベクトル、及び全成分を 1.0 としたベクトルの 2 種類を用いた。前者の方法を VSM1、後者を VSM2 と呼ぶこととする。

3.3 実験結果

表 1 に本手法(SVSM)、VSM1、VSM2 における最適なαの下での F 値の比較を、表 2 に各次数の固有ベクトルの各索引語に対する係数、及び VSM1 で用いた平均ベクトル（$\mathbf{f}_{ave}$）の各成分の値を、表 3 に各次数の固有値と代表度をそれぞれ

表 1　F 値の比較

	SVSM	VSM1	VSM2
F値	87.06%	87.06%	53.62%

表 3　各固有値と代表度

次数	1	2	3	4	5	6	7	8
固有値	184.87	15.00	12.00	12.00	12.00	7.12	3.01	0.00
代表度	75.15	6.10	4.88	4.88	4.88	2.90	1.22	0.00

表 2　固有ベクトルと平均ベクトルの各語に対する係数

	各語に対する係数							
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
ϕ_1	0.27	0.27	0.27	0.27	0.34	0.34	0.49	0.49
ϕ_2	0.00	0.00	0.00	0.00	0.71	-0.71	0.00	0.00
ϕ_3	0.79	-0.21	-0.58	0.00	0.00	0.00	0.00	0.00
ϕ_4	-0.21	0.79	-0.58	0.00	0.00	0.00	0.00	0.00
ϕ_5	-0.29	-0.29	-0.29	0.87	0.00	0.00	0.00	0.00
ϕ_6	-4.00	-4.00	-4.00	-4.00	0.39	0.39	0.18	0.18
ϕ_7	-0.13	-0.13	-0.13	-0.13	-0.48	-0.48	0.18	0.18
$\mathbf{f}_{ave}$	0.53	0.53	0.53	0.53	0.67	0.67	1.00	1.00

示す。表3に示されるように、この場合には0でない固有値は8個ではなく、7個である。これは、式(11)のQの場合、Qと適合する索引ではw_7とw_8は常に共起し、そのためにΩ_1に含まれる全ベクトルの平方和行列の7番目の行（列）ベクトルと8番目の行（列）ベクトルは全く同じになったことが起因して、ランクが8でなく7となったためである。そのため、表2においても0でない固有値に対応する固有ベクトルの係数を示している。また、表1のSVSMにおいてF値を求めるときにはL=7としている。

表1から分かるように、本手法の結果はVSM2よりは良好であるが、VSM1とは全く同等であり、期待に反したものになっている。この理由は以下のように考えられる。表2から分かるように、1次固有ベクトルの各係数と平均ベクトルの各成分とは殆ど正比例の関係にある。事実、1次固有ベクトルと平均ベクトルとの余弦類似度を求めてみたところ、0.9998という高い値が得られた。一方、2次以上の固有ベクトルは色々な方向を向いている。これらの事実は、1次固有ベクトルはΩ_1に含まれる全ベクトルの直流分を表し、索引語空間におけるΩ_1中の索引ベクトルの分布の微細構造は2次以上の固有ベクトルに反映されていることを示している。また、表3に眼を転じると、1次の固有値は代表度が75.15%と非常に高いことが分かる。さらに、式(8)のSVSM類似度の定義では、固有ベクトルと索引ベクトルの内積の2乗について各次数の固有値を重みとして和を求める格好になっている。そのため、1次の固有ベクトルに比べ2次以上の固有ベクトルの類似度に対する寄与が小さくなり、分布の微細構造が類似度に反映されなくなった結果、本手法でのF値はVSM1と変わらなくなってしまったと考えられる。これを確認するために、一定値β以上の固有値は次数に関係なくβにするというクリッピング処理を行って、F値を求めてみた。図1にF値とβの関係を示す。図1から分かるように、βの値が8〜20の範囲でF値は100%に達する。βの値を8にすることは、表3から分かるように、1次〜5次の固有値は全て8とすることを意味する。この結果から、2次以上の固有ベクトルを適度に強調することが性能を向上させるうえで非常に有効なことが確認できた。

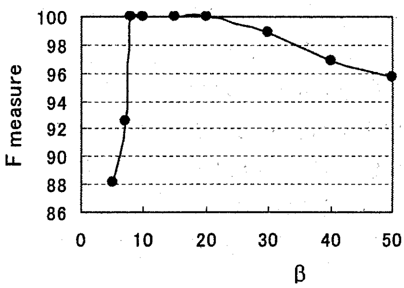

図1　固有値をクリッピングした時のF値

3.4　他のクエリーに対する結果

式(11)以外のクエリーに対する結果を表4に示す。表4でSVSM0は固有値のクリッピングを行わない場合のF値、SVSM1は行った場合の最高のF値を示す。Q1はクエリが一部入れ子となっており、w_1はw_2、w_3、w_4を用いて言い換えられた格好になっている。Q2とQ3はORの関係にある索引語につき出現する語数に制限を設けたものである。Q2とQ3に現れるCは$b_1+b_2+b_3+b_4$、Dはb_5+b_6を表している。C<3はw_1〜w_4のうち3個以上現れてはならない、C>2は3個以上現れなく

表4　他のクエリに対する結果

	クエリ	SVSM0	VSM1	SVSM1	適合索引数
Q1	(w_1 OR (w_2 AND w_3 AND w_4)) AND (w_5 OR w_6) AND w_7 AND w_8	86.8	86.8	100.0	27
Q2	(w_1 OR w_2 OR w_3 OR w_4) AND (w_5 OR w_6) AND w_7 AND w_8 AND C<3 AND D=1	60.6	59.7	100.0	20
Q3	(w_1 OR w_2 OR w_3 OR w_4) AND (w_5 OR w_6) AND w_7 AND w_8 AND C>2 AND D=1	76.9	76.9	100.0	8
Q4	(w_1 OR w_2 OR w_3 OR w_4) AND (NOT(w_5 OR w_6)) AND w_7 AND w_8	78.6	78.6	100.0	15
Q5	(w_1 OR w_2 OR w_3 OR w_4) AND ((NOTw_5) OR w_6) AND w_7 AND w_8	86.3	85.4	94.3	45
Q6	(w_1 OR w_2 OR w_3 OR (NOTw_4)) AND (w_5 OR w_6) AND w_7 AND w_8	90.5	90.5	96.8	45
Q7	(w_1 AND w_2) OR (w_3 AND w_4) OR (w_5 AND w_6) OR (w_7 AND w_8)	94.5	87.0	100.0	175
Q8	w_1 OR (w_2 AND w_3 AND w_4) OR (w_5 AND w_6) OR (w_7 AND w_8)	94.1	92.4	99.0	193

てはならないという条件を示している。また、D=1 はw_5とw_6が排他的論理和の関係にあることを示している。Q_4〜Q_6はNOTが含まれるクエリである。Q7とQ8はORが主体のクエリである。表から分かるように、何れのクエリでも固有値のクリッピングの効果が現れており、Q5、Q6、Q8を除き、クリッピング後のF値は100%に達している。

4. 結果のフィードバック

本章では、前章で100%に達しなかったクエリにつき、結果をベクトル展開にフィードバックし、更に性能を向上させることを目指す。

4.1 考え方

提案手法では、平方和行列Sの固有ベクトルがクエリを展開したベクトルとして与えられる。従って、結果のフィードバックはSの修正によってなされる。**f**を望ましくない結果を与えるベクトルとし、式(9)の分子に対し、L=Rとして

$$\sum_{m=1}^{R}\lambda_m(\boldsymbol{\phi}_m{}^T\mathbf{f})^2+a(\mathbf{f}^T\mathbf{f})^2 \qquad (12)$$

のようにフィードバックを行ったとする。$(\mathbf{f}^T\mathbf{f})^2$は常に正なので、**f**を類似度の低かった適合索引とすると、aを正にとれば類似度の値は増加する。反対に、**f**を類似度の高かった非適合索引とすると、aを負にとれば類似度の値は減少するようになり、望ましい方向に類似度の値を変えることができる。$\sum_{m=1}^{R}\lambda_m(\boldsymbol{\phi}_m{}^T\mathbf{f})^2=\mathbf{f}^T\mathrm{S}\mathbf{f}$及び$(\mathbf{f}^T\mathbf{f})^2=\mathbf{f}^T(\mathbf{f}\mathbf{f}^T)\mathbf{f}$の関係を用いると、式(12)は

$$\mathbf{f}^T(\mathrm{S}+a\mathbf{f}\mathbf{f}^T)\mathbf{f} \qquad (13)$$

のように書き直され、結局、$\mathrm{S}+a\mathbf{f}\mathbf{f}^T$の固有値、固有ベクトルを式(9)に用いればよいことが分かる。上記は、結果が望ましくない結果が得られる度にフィードバックを行う例であるが、本報告では類似度の低かった適合索引のベクトル集合をΩ^+、類似度の高かった非適合索引のベクトル集合をΩ^-として、

$$\mathrm{S}'=\mathrm{S}+a\sum_{\mathbf{f}\in\Omega^+}\mathbf{f}\mathbf{f}^T-b\sum_{\mathbf{f}\in\Omega^-}\mathbf{f}\mathbf{f}^T \qquad (14)$$

の固有値、固有ベクトルを用いるようにする。ここでa、bはパラメータである。フィードバックのステップは以下の通りである。

ステップ1：**3.1**で述べた手法に従ってベクトル化を行う。また、Ω^+、Ω^-を空にしておく。

ステップ2：評価実験を行い、F値が収束していれば終了する。そうでなければ類似度の低かった適合索引ベクトルをΩ^+に、類似度の高かった非適合索引ベクトルをΩ^-に加える。

ステップ3：式(14)で定義される平方和行列を用いてベクトル化を実行。ステップ2へ。

4.2 実験結果

3.3における実験結果でF値が100%に達しなかったQ5、Q6、Q8につき、フィードバックを行って性能の改善を試みた。フィードバック用の索引ベクトルの選択は、適合索引の類似度の最小値をr_{min}、非適合索引の類似度の最大値をr_{max}として、類似度がr_{max}より小さい適合索引ベクトルをΩ^+に、類似度がr_{min}より大きい非適合索引ベクトルをΩ^-に追加することにより行った。式(14)でa=b=1.0としたときのフィードバック後のF値と繰り返し数との関係を図2に示す。何れのクエリもF値は100%に到達している。**3.**の結果と合わせ、提案手法により適合／非適合判定能力を保ったまま、ブーリアンクエリを複数のベクトルに展開できることが分かった。

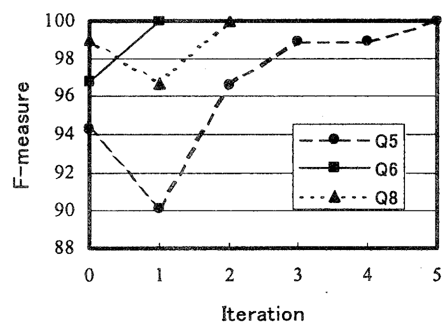

図2　フィードバック後のF値

5. 考察

5.1 ベクトル空間モデルの利点の維持

ベクトル空間モデルの利点については **1.** で述べた。ブーリアンクエリをベクトル化できたにしてもこれらの利点が失われるようでは意味がない。

先ず、クエリにおいてユーザが指定する特定索引語に重みが付与できるか否かを確認した。そのため、式(11)のクエリにおいて、w_1 に重みを与え、w_1 が存在する適合索引を検索結果の上位に集中させることを試みた。具体的には、式(9)に用いられる各固有ベクトルの w_1 に対する係数を c 倍することにより行った。c=1.0 の時、即、重み付けを行わないときには、検索結果の上位 13/25/33 個の索引中、w_1 が存在する索引数は 10/16/21 であったのに対し、c=1.1 とした場合には F 値は 100%で、上位 9/22/30 個の中、w_1 が存在する索引数は 9/21/24 であり、w_1 が存在する適合索引を検索結果の上位に集中できることが分かった。但し、c の選び方には注意が必要であり、大きくとると w_1 が存在する非適合索引が上位に存在するようになり、F 値が低下する。

次に、検索結果のランキングがどの程度の正確さで可能となるかを検討した。具体的には、式(11)のQを用いたとき、非適合索引を部分的な適合の程度に応じて正確にソートできているかどうかを観察することとした。式(11)のQは、(w_1 OR w_2 OR w_3 OR w_4)、(w_5 OR w_6)、w_7、w_8 の4つのサブクエリに分割できる。そこで、ひとつの索引ベクトルに対して満足されるサブクエリの数をAとして（Q との適合索引の場合 A=4)、255 種類の索引ベクトルを与えたとき、A の数が 4、3、2、1 となる索引ベクトルのグループ毎に類似度の平均、最大、最小を求めた。β の値を 2 次個有値 15.0 と等しくしたときの結果を表5に示す。表から分かるようにAが3のグループと2のグループとの間、2のグループと1のグループとの間ではやや重複が見られるが、各グループの平均は満足するサブクエリの数と比例しており、総じて望ましい傾向となっている。また、表6に各グループ毎に索引ベクトルと類似度の実例を示す。類似度は決して存在する索引語の数に比例している訳ではないことが分かる。

表5　類似度の分布

A	4	3	2	1
類似度平均値	0.679	0.621	0.580	0.560
類似度最大値	0.685	0.643	0.615	0.607
類似度最小値	0.667	0.593	0.541	0.383
索引数	45	108	82	20

表6　索引ベクトルと類似度の実例

A	索引ベクトル	類似度	A	索引ベクトル	類似度
4	11111111	0.684	2	00010100	0.615
	01110111	0.684		11110001	0.581
	00110111	0.682		00000101	0.570
	00010111	0.672		11111100	0.561
	00011111	0.667		00000011	0.541
3	01110011	0.643	1	00010000	0.607
	00000111	0.642		00000100	0.598
	11110011	0.641		11110000	0.547
	11111110	0.618		00001100	0.476
	00011101	0.593		00000001	0.383

また、式(14)による平方和行列の更新は望ましくない検索結果のベクトル生成過程へのフィードバックそのものであるから、実際の情報検索の場においても所謂適合性フィードバックを行うことは容易と考えられる。

以上から、本手法では本来ベクトル空間モデルが持つ利点は損なわれていないと云える。

5.2 高速化の可能性

本手法では処理量は従来のベクトル空間モデルに比べL倍であり、負担は大きくなる。そのため負担を軽減することが望まれる。

負担を軽減するための最も簡単な方法はL自体を極力小さくすることである。図3に式(11)のクエリに対するF値とLの関係を示す。この場合、β は 12.0 にとっており、1次から5次までの固有値を等しくしている。図から分かるようにF値はLが5以上であれば 100%に達している。また、Lが2であっても 97.8%となっており、100%に拘らずそこそこのF値が得られればよいということであれば、Lは2にしてもよい。なお、Lが3の時にF値が小さくなるのは以下の理由によると考えられる。即ち、表2から分かるように3〜5次の固

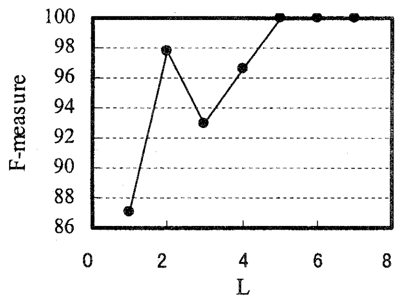

図３　部分空間の次元数とＦ値の関係

有ベクトルは w_1～w_4 により方向が決まっており、しかも固有値は等しくなっている。言い換えれば w_1～w_4 に関係する索引ベクトルの分布を正しく記述するためには 3～5 次の固有ベクトルの全てを必要とする。L=3 のときにＦ値が低下したのは、4 次、5 次の固有ベクトルを用いなかったことにより、却って索引ベクトルの分布の微細構造が歪められたためと考えられる。

２つ目の方法は、図３で 1～5 次の固有値を等しくしても高いＦ値が得られるという事実に基づく。いま、1 次から k 次までの固有値を λ_k としたとき、式(9)のカッコ内の分子は L=R として

$$\Sigma_{m=1}^{k}\lambda_k(\boldsymbol{\phi}_m^{T}\mathbf{f}_i)^2+\Sigma_{m=k+1}^{R}\lambda_m(\boldsymbol{\phi}_m^{T}\mathbf{f}_i)^2 \quad (15)$$

と書き直される。また、

$$\|\mathbf{f}_i\|^2=\Sigma_{m=1}^{N}(\boldsymbol{\phi}_m^{T}\mathbf{f}_i)^2 \quad (16)$$

の関係があるので、式(15)に代入すると、

$$\lambda_k\|\mathbf{f}_i\|^2+\Sigma_{m=k+1}^{N}(\lambda_m-\lambda_k)(\boldsymbol{\phi}_m^{T}\mathbf{f}_i)^2 \quad (17)$$

となる。m が R より大きいとき、λ_m=0 である。式(17)を式(9)に代入することにより、

$$r_i=\left[\frac{\lambda_k\|\mathbf{f}_i\|^2+\Sigma_{m=k+1}^{N}(\lambda_m-\lambda_k)(\boldsymbol{\phi}_m^{T}\mathbf{f}_i)^2}{\sqrt{\Sigma_{m=1}^{L}\lambda_m^{2}\|\mathbf{f}_i\|^2}}\right]^{1/2} \quad (18)$$

となる。式(18)から分かるように、固有ベクトルと索引ベクトルの内積の演算回数は R 回から N-k 回に減らすことができる。

5. まとめ

以上、検索手法としてのベクトル空間モデルについて、その長所は生かしつつ、索引語間の関係を記述できないという短所をカバーする手法について提案した。提案手法は、与えられたクエリに適合するであろう文書の索引ベクトル集合を網羅し、その集合と被検索文書の索引ベクトルとの間の SVSM 類似度を求めて被検索文書の適合の度合いとするものである。また、望ましくない結果をフィードバックさせる手法についても提案した。さらに、8 検索語から成るクエリについて簡単な模擬実験を行い、適合文書と非適合文書とを正確に分離できることを確認した。この事実は、SVSM 類似度の有用性を証明するものでもある。

また、見方を変え、$\mathbf{f}_i$ を文ベクトル、ベクトル集合 Ω_1 を文ベクトルの集合から成るひとつの文書と見ると、本報告で述べた処理は、ある文書がある与えられた文を含んでいるか否かをチェックするタスクであるとも解釈できる。このように見ると新たな応用が生まれる可能性がある。新たな応用を考えていくことも、実文書を用いた検索実験、*tf-idf* などで重み付けされた索引ベクトルに対する効果確認と並んで今後の重要な課題である。

参考文献

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

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] Robert M. Losee. *TEXT RETRIEVAL AND FILTERING*. Kluwer Academic Publishers(1998).

[4] 徳永健伸. 情報検索と言語処理. 東京大学出版会(1999).

[5] 川谷隆彦. 文ベクトル集合モデルによるテキスト処理. 情報処理学会自然言語処理研究報告, 2000-NL-140, pp.31-38(2000).