自然言語処理 140－5
（2000．11．21）

文ベクトル集合モデルによるテキスト処理

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

本報告はベクトル空間モデルの改善に関するものであり、新しいモデルとして文ベクトル集合モデルを提案する。本モデルは文書を文ベクトルの集合で表現するもので、その単語-文行列に対して特異値分解を施すことにより文ベクトル集合を互いに直交する固有文で展開する。固有文は文書固有に決まるもので、文書固有の概念とエネルギーにより定義される。その結果文書の中心的な概念を定量的に把握することが可能になる。さらに、本モデルの応用として、文書間類似度、文の重要度、質問文に対する文毎の関連度について新しい尺度を提案する。文書間類似度は 2 文書間の全ての文の組み合わせに対する照合を、文の重要度は各文と文書の固有概念との照合を、質問文に対する文毎の関連度は質問文と固有概念の照合をもとにそれぞれ行う。

Text Processing Based on Sentence Vector Set Model

Takahiko KAWATANI

Hewlett-Packard Labs Japan, Hewlett-Packard Japan

takahiko_kawatani@hp.com

This paper is concerned with improvement of the vector space model, and proposes the sentence vector set model as a new model. In this model, a document is treated as a set of sentence vectors. By applying singular value decomposition to the term-sentence matrix obtained from the vector set, the document vector set is expanded to mutually orthogonal eigenSentences. Each eigenSentence which is particular to the document has its own concept and energy. Consequently, it becomes able to analyze central concepts of the document quantitatively. As an application of the model, this paper proposes new measures of document similarity, sentence importance and relevance between each sentence and a query.

1. まえがき

ベクトル空間モデル VSM[1]は文書に表れる単語の頻度、又はその重み付けされた値を要素とする文書ベクトルにより文書を表現するものであり、文書ベクトルは文書の概念を表すとされている。VSM は文書の検索、分類をはじめとしてテキスト処理に広く用いられているが、以下のような問題点も存在する。

(1) 各単語が独立に扱われるため[2]、単語間の共起関係がベクトルに反映されない。その結果、
(2) 文書ベクトル自体が表す概念自体にも曖昧性が生ずる。例えば、単語 a、b、c、d を考えた場合、文中で a-b、c-d の組み合わせで用いられ

た文書と、a-c、b-d の組み合わせで用いられた文書とは異なった概念を表すと考えられるが、文書ベクトル上では区別できない。

(3) 文書が表す概念には広がりが存在すると考えられるが、単一のベクトルで広がりを表すことは不可能である。

本報告ではこのような問題に対処するために文ベクトル集合モデル（Sentence Vector Set Model : SVSM）を提案する。SVSM では、文書を構成する文毎にベクトル（文ベクトル）を求め、その集合として文書を表す。これにより、上記(2)、(3)の問題は解消される。しかし、この表現のみでは、文書が全体として表す概念は明確にできない。そこで、この文ベクトル集合に対して求められる単語-文行列に対して特異値分解（SVD）を施し、文ベクトルの平方和行列（後述）の固有値、固有ベクトルによって文ベクトル集合を表現する。この処理は、従来文書ベクトルの集合に対して行われてきた潜在的意味解析（Latent Semantic Analysis）[3][4]を文ベクトル集合に対して適用したものである。処理としては原点を基点とする KL 展開[5][6]と等価である。このようにして求められる固有ベクトルには単語間の共起関係が自ずと反映されるようになり、上記(1)の問題は解消される。固有ベクトルは文ベクトル集合の主要成分、即ち文書の主要な概念を表し、固有ベクトルの張る空間は文書固有の概念空間と見なすことができる。また、個有値は対応する固有概念の重みを与える。この結果、SVSM では入力文書が表す複数の固有概念を定量的に把握することができ、文書間の類似度を的確に求めることができる他、色々な応用が可能となる。

以下、**2.**では、SVSM の具体的方法として、文書の定義、固有ベクトルの求め方、固有ベクトルや固有値の解釈について述べる。**3.**では簡単な実験により、単語間の共起と求められる固有値、固有ベクトルの関係を示す。**4.**では SVSM の応用として、文書間の類似度、文書内の各文の重要度、質問文に対する文毎の関連度のそれぞれについて新しい尺度の求め方を述べる。

2. 文ベクトル集合モデル（SVSM）

2.1 SVSM における文書の表現

M 個の文から成り、現れる単語集合が $\{w_1,\ldots,w_N\}$ で与えられる文書 D を考える。文書 D の文 m のベクトルを

$$\mathbf{d}_m=(f_{m1},\ldots,f_{mN}) \qquad (1)$$

により表すこととする。ここで、f_{mn} は単語 w_n の文 m における出現頻度である。さらに、$\mathbf{d}_m$ を列ベクトルとする行列、即、単語-文行列を A とする。

前述のように、SVSM では A に対して SVD を施すものであり、A が次式のように展開されたとする。

$$A = U\Sigma V^t \qquad (2)$$

ここで t は転置を表す。U の列ベクトルは AA^t で定義される行列 $S=(S_{ij})$ の固有ベクトルで与えられる。S は次式によっても定義される。

$$S = \sum_{m=1}^{M} \mathbf{d}_m {}^t\mathbf{d}_m \qquad (3)$$

式(3)から $S_{ij} = \sum_{m=1}^{M} f_{mi} f_{mj}$ となることからも分かるように、S は単語間の共起の程度を表す行列である。ここでは S を平方和行列と呼び、ランクを R、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。さらに、Σ は $\sqrt{\lambda_k}$ を要素とする対角行列であり、V は A^tA の固有ベクトルを列ベクトルとする行列である。文ベクトル $\mathbf{d}_m$ の固有ベクトルの張る空間への射影（図 1 参照）を

$$\mathbf{z}_m = U^t\mathbf{d}_m \qquad (4)$$

とし、V^t の列ベクトルを $\mathbf{v}_m$ とすると、

$$\mathbf{z}_m = \Sigma\mathbf{v}_m \qquad (5)$$

の関係がある。

2.2 文の概念とエネルギー

ここで、次節以降の説明を容易にするため、文ベクトルの属性として、概念、エネルギーの 2 つについて説明しておく。例として、3 次元の文ベクトル A(1,1,2)と B(3,3,6)を考える。ベクトル B は A をスカラー倍したものであるが、これらのベク

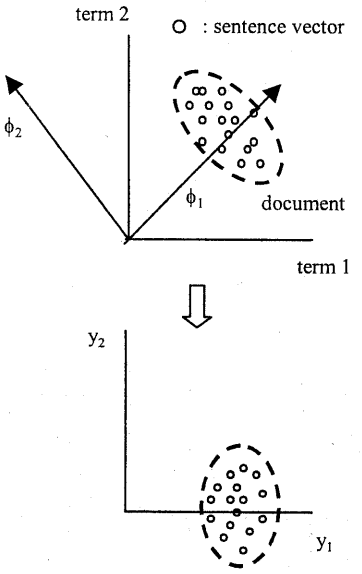

図 1　単語空間から固有ベクトル空間への射影（$y_n=\phi_n{}^t\mathbf{d}$）。

トルでは何が同じで何が異なるのであろうか？ここでは、これらのベクトルの表す概念は同じで、文の持つエネルギーが異なるとみる。即、どのような単語がどのような比率で文に現れるかが文の概念を決定すると考え、ベクトルの方向が概念を表すと見なす。また、文のエネルギーは文ベクトルのノルムの2乗で与えられると考える。そこで、文 m のエネルギー$E(\mathbf{d}_m)$を

$$E(\mathbf{d}_m)=\sum_{n=1}^{N}\|f_{mn}\|^2 \qquad (6)$$

により定義する。また、単語 w_n のエネルギー$E(w_n)$を次式のように定義する。

$$E(w_n)=\sum_{m=1}^{M}\|f_{mn}\|^2 \qquad (7)$$

さらに、文書 D の全エネルギーを E(D)とすると

$$E(D)=\sum_{m=1}^{M}E(\mathbf{d}_m)=\sum_{n=1}^{N}E(w_n) \qquad (8)$$

なる関係式が成り立ち、文書エネルギーは文エネルギーの総和、単語エネルギーの総和と等しくなる。同一単語が同一文に2回以上現れないならば、文エネルギーは各文に現れた単語の数、単語エネルギーは各単語が現れた文の数となる。

ここでは、概念そのものはエネルギーを持たず、その概念を有する文がエネルギーを持つとの解釈に立つ。

2.3　固有値、固有ベクトルの性質と解釈

上述のように求められた固有値、固有ベクトルの性質と解釈について述べ、固有概念、固有文の概念を導入する。

(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。従って、固有値λ_k は k 次の固有文のエネルギーを表すものと解釈できる。従って、k 次の固有文のベクトルは $\sqrt{\lambda_k}\,\phi_k$ で与えられる。

(3) 平方和行列 S は

$$S=\sum_{k=1}^{R}\lambda_k\,\phi_k\phi_k{}^t \qquad (9)$$

と展開できる。これは文ベクトルの集合が互いに直交する固有文で展開されたことを示す。

(4) 一般的な公式から、$tr(S)=\sum_{n=1}^{N}\|f_{mn}\|^2=\sum_{k=1}^{R}\lambda_k$ が成り立つ。$\sum_{k=1}^{R}\lambda_k$ は式(4)による変換後の文エネルギーの総和を表し、式(7)(8)

から $\sum_{n=1}^{N}\|f_{mn}\|^2$ はもともとの文書のエネルギーと等しいから、式(4)による変換に際して文書のエネルギーは保存されることが分かる。

(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。

(6) 一般に高次になるほど固有値の値は小さくなるので式(9)において 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。

(7) $\mathbf{d}_m$ を L 次元概念部分空間に射影することにより、文ベクトルを L 次元に圧縮することができる。

3. 固有値、固有ベクトルの観察

3.1 模擬実験

本節では簡単な模擬実験により、単語間の共起に対して個有値、児湯ベクトルどのように求められるかを示す。今、単語 1～4 が現れる文書を考え、図 2 のような平方和行列が得られたとする。対角要素は式(7)で与えられる単語エネルギーを示し、単語 1 が最も大きくなっている。a、b、c はパラメータで、a は単語 1、2 間、b は単語 2、3 間、c は単語 3、4 間の共起の程度を与える。a、b、c に適当な値を与えた平方和行列から求めた固有値、固有ベクトルを表 1 に示す。表 1 では、n を次数として n 次の固有値 λ_n、n 次の固有ベクトルの各単語に対する係数、ϕ_{n1}、ϕ_{n2}、ϕ_{n3}、ϕ_{n4} を示す。この結果から以下が言える。

(1) a=b=c=0 の時は、4 つの単語は共起しないこと

$$\begin{bmatrix} 10 & a & 0 & 0 \\ a & 9 & b & 0 \\ 0 & b & 8 & c \\ 0 & 0 & c & 7 \end{bmatrix}$$

図 2　想定する平方和行列

表 1　図 2 の平方和行列に対する固有値、固有ベクトル

a,b,c	n	λ_n	ϕ_{n1}	ϕ_{n2}	ϕ_{n3}	ϕ_{n4}
a=0	1	10.00	1.00	0.00	0.00	0.00
b=0	2	9.00	0.00	1.00	0.00	0.00
c=0	3	8.00	0.00	0.00	1.00	0.00
	4	7.00	0.00	0.00	0.00	1.00
a=0	1	10.00	1.00	0.00	0.00	0.00
b=0	2	9.00	0.00	1.00	0.00	0.00
c=1	3	8.62	0.00	0.00	-0.85	-0.53
	4	6.38	0.00	0.00	-0.53	0.85
a=0	1	10.54	0.00	0.00	0.76	0.65
b=0	2	10.00	1.00	0.00	0.00	0.00
c=3	3	9.00	0.00	1.00	0.00	0.00
	4	4.46	0.00	0.00	-0.65	0.76
a=3	1	12.54	-0.76	-0.65	0.00	0.00
b=0	2	10.54	0.00	0.00	0.76	0.65
c=3	3	6.46	-0.65	0.76	0.00	0.00
	4	4.46	0.00	0.00	-0.65	0.76
a=3	1	12.68	0.73	0.65	0.21	0.11
b=1	2	10.51	-0.27	-0.05	0.73	0.63
c=3	3	6.50	0.63	-0.73	-0.05	0.27
	4	4.32	-0.11	0.21	-0.65	0.73

を意味する。表 1 では n 次の固有ベクトルの係数は単語 n が 1 で他は 0 である。これは、各固有ベクトルは各単語の軸そのものであり、固有文の概念は各単語の概念そのものであることを示す。各固有値は平方和行列の対角要素の値と等しい。即、各固有文のエネルギーは各単語エネルギーと等しく、各単語の頻度で決まる。

(2) a=b=0、c=1 の時には、単語 3、4 間に単語の共起関係が存在する。その結果、3 次と 4 次の固有文の概念は単語 3、4 の組み合わせで決まるようになる。従来のベクトル空間モデルでは

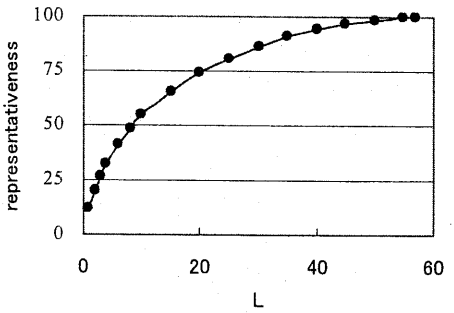

図3　実文書における代表度の例

単語の共起関係を表現することは出来なかったが、SVSM では固有文の概念に自然な形で単語の共起関係が反映される。また、3 次の固有文のエネルギーは a=b=c=0 の時のそれに比べ、大きくなる。a=b=0、c=3 の時には、1次の固有ベクトルはϕ_{13}、ϕ_{14}が 0 でない値を持っており、単語 3、4 の組み合わせで概念が決まる固有文のエネルギーが最も大きくなることが分かる。これは固有文のエネルギーは単語の頻度と単語間の共起の両方から決まることを示す。

(3) a=3、b=0、c=3 の時には、単語 1-2 間、単語 3-4 間で単語は共起し、単語 1、2 と 3、4 の間では共起しない。その結果行列 S は部分行列に分解される。このような場合、固有文の概念は単語 1、2 の組み合わせ、もしくは単語 3、4 の組み合わせで決まる。しかし、a=3、b=1、c=3 となって、単語 2-3 間に共起が存在するようになると、固有文の概念には全ての単語が拘るようになる。単語 1 は単語 3、4 とは直接的な共起関係はなく、高次の共起関係[7]にあると云えるが、このような高次の共起関係も固有文の概念に影響を与えることが分かる。実際の文書においても単語間の直接的な共起だけでなく、高次の共起の連鎖が概念の形成に重要な役割を果たしていると考えられる。

3.2　実文書における固有値の観察

文数 58、総単語数 1100 の英文ニュース記事から名詞を抽出して 146 次元の文ベクトルを構成した場合の概念部分空間の代表度と次元数 L の関係を図 3 に示す。この場合の平方和行列のランクは 58 であり、L の最大値も 58 となる。代表度 50%、75%を達成する次元数はそれぞれ 8、20 であり、大幅な次元圧縮が可能なことを暗示している。また、1 次の固有文の代表度は 12.2%であるが、この文書の平均文ベクトルの代表度（各文の平均ベクトルへの射影値の 2 乗和の文書エネルギーに対する割合）は 10.4%であった。1 次の固有文は平均ベクトルを上回る代表度を有することが確認された。1 次の固有文の代表度は文書が長くなるにつれ低下する傾向にある。

4.　SVSM の応用

本章では、SVSM の応用として文書間類似度、及び文の重要度について新しい尺度を提案する。

4.1　文書間類似度

4.1.1　定義

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。また、文書 D の平方和行列を S_D、そのランクを R_D、k 次の固有値、固有ベクトルを$\lambda_k(\geq\lambda_{k+1})$、$\phi_k$とし、文書 T については、それぞれ S_T、R_T、γ_k、τ_k とする。ここでは、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する。

$$r=\left(\frac{\sum_{m=1}^{M}\sum_{j=1}^{J}(\mathbf{d}_m{}^t\mathbf{t}_j)^2}{\sqrt{\sum_{m=1}^{M}\sum_{j=1}^{M}(\mathbf{d}_m{}^t\mathbf{d}_j)^2\sum_{m=1}^{J}\sum_{j=1}^{J}(\mathbf{t}_m{}^t\mathbf{t}_j)^2}}\right)^{1/2} \tag{10}$$

文書 D を固有文に展開すると式(10)の括弧内の分子は

$$\sum_{m=1}^{M}\sum_{j=1}^{J}(\mathbf{d}_m{}^t\mathbf{t}_j)^2$$

$$= \sum_{m=1}^{M}\sum_{j=1}^{J} \mathbf{t}_j{}^t\mathbf{d}_m \mathbf{d}_m{}^t\mathbf{t}_j$$
$$= \sum_{j=1}^{J} \mathbf{t}_j{}^t S_D \mathbf{t}_j \qquad (11)$$
$$= \sum_{m=1}^{R_D}\sum_{j=1}^{J} \lambda_m (\boldsymbol{\phi}_m{}^t\mathbf{t}_j)^2$$

となり、さらに文書Tを展開すると、

$$\sum_{m=1}^{M}\sum_{j=1}^{J} (\mathbf{d}_m{}^t\mathbf{t}_j)^2$$

$$= \sum_{m=1}^{R_D}\sum_{j=1}^{R_T} \lambda_m \gamma_j (\boldsymbol{\phi}_m{}^t\boldsymbol{\tau}_j)^2 \qquad (12)$$

のように書くことができる。これらの式の導出にあたっては式(3)、(9)の関係を用いている。従って、類似度は、式(11)、(12)に対応して、それぞれ

$$r = \left(\frac{\sum_{m=1}^{R_D}\sum_{j=1}^{J} \lambda_m (\boldsymbol{\phi}_m{}^t\mathbf{t}_j)^2}{\sqrt{\sum_{m=1}^{R_D} \lambda_m{}^2 \sum_{m=1}^{J}\sum_{j=1}^{J} (\mathbf{t}_m{}^t\mathbf{t}_j)^2}} \right)^{1/2} \qquad (13)$$

$$r = \left(\frac{\sum_{m=1}^{R_D}\sum_{j=1}^{R_T} \lambda_m \gamma_j (\boldsymbol{\phi}_m{}^t\boldsymbol{\tau}_j)^2}{\sqrt{\sum_{m=1}^{R_D} \lambda_m{}^2 \sum_{j=1}^{R_T} \gamma_j{}^2}} \right)^{1/2} \qquad (14)$$

のように定義することができる。この類似度をSVSM類似度と呼ぶこととする。

4.1.2 SVSM類似度の解釈

文書Tが1つの文 $\mathbf{t}$ で与えられる場合は、$\bar{\mathbf{t}} = \mathbf{t}/\|\mathbf{t}\|$ により単位ベクトル化すると、式(10)は

$$r = \left(\frac{\sum_{m=1}^{M} (\mathbf{d}_m{}^t\bar{\mathbf{t}})^2}{\sqrt{\sum_{m=1}^{M}\sum_{j=1}^{M} (\mathbf{d}_m{}^t\mathbf{d}_j)^2}} \right)^{1/2} \qquad (15)$$

のようになる。式(15)における $(\mathbf{d}_m{}^t\bar{\mathbf{t}})^2$ は文ベクトルmの $\mathbf{t}$ 方向のエネルギーを表し、式(15)の括弧内の分子全体は文書Dの $\mathbf{t}$ 方向のエネルギーを表す。云いかえれば、この値は文書Dと文 $\mathbf{t}$ とが概念としてどれだけ共通するかをエネルギーで表したものである。一方、式(10)の分子は、文書Dの $\mathbf{t}_j$ 方向のエネルギーを $\mathbf{t}_j$ のエネルギーを重みとして文書Tの全文ベクトルについて加重和を求めたものである。これも文書Tと文書Dとが共通して持つ概念に応じた量になっている。このように、SVSM類似度では、両文書の共通する概念の程度が正確に反映されるので、従来のような各文書に1つ用意されたベクトル間の余弦類似度を用いる場合に比べ、より自然で正確な定義になる。また、両方の文書の全ての文の概念が類似度に反映されるので、文書の概念の広がりも自ずと類似度に反映される。

また、式(13)、(14)は式(10)における文ベクトル同士の照合が、文ベクトルと固有文、あるいは固有文同士の照合に置き換えられたことを示す。合理的な変形と云える。2章で述べたように、文書は少数の固有文で近似することができる。文書D、Tをそれぞれ L_D、L_T 個の固有文で近似したとすると、類似度は式(13)、(14)で R_D を L_D、R_T を L_T に置き換えることにより得られる。この近似により精度をあまり落とさずに式(10)の計算の負担を軽減することが可能になる。

4.1.3 他の類似度尺度との関係

類似度として、式(13)のように入力特徴ベクトルと固有ベクトルとを照合して類似度を求める方法として、部分空間法[8]、複合類似度法[9]がパターン認識の分野で知られている。文書間の照合にあてはめて云うと、部分空間法は文ベクトルの概念部分空間への射影値を類似度として定義するもので固有値の情報は用いられていない。そのため文書の中心的な概念の間の量的な関係が反映されない。また、複合類似度法は、例えば文書Tが1つの文 $\mathbf{t}$ で与えられる場合、$\mathbf{t}$ と文書Dの各文との間で余弦類似度を求め、その2乗平均の平方根により定義するものである。そのため、例えば、文書Dが $\mathbf{t}$ をα倍（αはスカラー）したベクトル、及び $\mathbf{t}$ とは直交するベクトルの2つのベクトルで構成される場合、αが大きいほど文書TとDの間の類似度は大きいと云えるが、複合類似度ではαの値に拘らず常に $1/\sqrt{2}$ となってしまう。SVSM類似度ではこのようなことはない。

4.2 文の重要度

従来の文の重要度算出は文書中の各単語の出現頻度や文書内の位置情報などを用いて行っており[10]、各文と文書の中心概念との関係は必ずしも

明確でなかった。しかし、SVSMでは文ベクトルの集合として文書を記述するため、文書の概念と各文ベクトルの概念の照合から各文の重要度を求めることができる。文の重要度としては、次の2つの尺度が考えられる。

(1) 着目する文ベクトルを概念部分空間に射影し、その射影値の2乗を重要度とする方法。文書の中心的な概念とどれだけ共通する概念を持つかにより文の重要度が決まる。

(2) 着目する文ベクトルに全文ベクトルを射影したときの射影値の2乗和を重要度とする方法。これは文書が着目する文ベクトルの方向にどれだけのエネルギーを持つかにより重要度を定義することになる。

文書D中の文mの重要度pは、L次元の概念部分空間を用いたとして、(1)の場合、

$$p = \sum_{k=1}^{L} (\phi_k{}^t d_m)^2 \qquad (16)$$

(2)の場合、

$$p = \sum_{k=1}^{L} \lambda_k (\phi_k{}^t d_m)^2 \qquad (17)$$

でそれぞれ定義される。両者の差は固有値を重みとして用いるか否かの差である。

4.3 各文と質問文との関連度

情報検索において、ユーザの質問に適合した要約を動的に作成して欲しいというニーズがある[10]。この場合文毎に質問文との関連度を求める処理が有用となる。文毎の質問文との関連度は例えばベクトル間の余弦類似度を用いて定義することは可能であるが、この場合質問文と対象とする文との間で共通する単語がない限り関連度は0となる。実際問題としては共通する単語がなくとも、文書中で互いに共起する単語対の一方が質問文に、他方が対象とする文に含まれていれば、0でない関連度が求められるのが望ましい。SVSMでは、固有文には文書中の単語の共起関係が自然に反映されるため上記の要求を満たすことが可能であり、具体的な方法として以下の2つの方法を提案する。ここでは**q**を質問文ベクトル、**4.1**における文書Dを処理対象の文書とする。

最初の方法は、概念部分空間に射影されたベクトル同士の照合に基づくものである。**q**及び文$\mathbf{d}_m$の文書Dの L概念部分空間への射影を**y**、$\mathbf{z}_m$とすると、文mに対する関連度g_mは次のように定義される。

$$g_m = \mathbf{y}^t \mathbf{z}_m / \|\mathbf{q}\| \qquad (18)$$

式(18)は$\mathbf{z}_m$のノルムに比例する値を持つが、式(18)をさらに$\|\mathbf{z}_m\|$で正規化し$\mathbf{z}_m$のノルムに無関係な値をとるようにしてもよい。式(18)が単語を共有しない文の間の関連度を表す理由は以下の通りである。表1からも分かるように互いに共起する単語対に対して、単語の違いは低次の固有ベクトルには現れず、高次の固有ベクトルに反映される。そのため、L次元の概念部分空間に単語の違いが反映される高次の固有ベクトルが含まれなければ式(18)は大きな値をとり、文間の関連度を表すとみなされるからである。しかし、概念部分空間に単語の違いが反映される高次の固有ベクトルが含まれていれば、関連度の値は小さくなる。また、文ベクトルの固有ベクトルへの射影は負の値をとりうるため、式(18)では正の値が求められるとは限らない。

2番目は、文書Dの固有概念に沿う方向のエネルギーの照合を行う方法である。先ず、文書Dのk次の固有概念に対する重みs_kを以下のように定義する。

$$s_k = (\phi_k{}^t \mathbf{q})^2 / \|\mathbf{q}\|^2 \qquad (19)$$

s_kは質問文の全エネルギーに対してk次の固有概念方向のエネルギーの占める割合を示す。文mに対する関連度g_mを以下のように定義する。

$$g_m = \sum_{k=1}^{L} s_k (\phi_k{}^t \mathbf{d}_m)^2 \qquad (20)$$

g_mは文mのϕ_k方向のエネルギーのs_kを重みとする加重和となっている。従って重みの大きい固有概念への射影値が大きい文は関連度が大きくなる。また、式(20)において文mのエネルギーに無関係にするため$\|\mathbf{d}_m\|^2$で正規化してもよい。式(18)、

(19)のどちらが関連度として相応しいかは今後さらに検討が必要である。

5. まとめ

以上、文書表現の新しいモデルとして文ベクトル集合モデルSVSMを提案し、さらに文書間類似度、文の重要度、質問文と各文との間の関連度について新しい尺度の定義を試みた。SVSMでは文書中の全ての文の情報を用いるので、たとえ概念部分空間による近似が行われたにしても処理に用いられる文書本来の情報は従来に比べ格段に多くなっているはずである。そのため、文書間類似度は従来に比べ非常に正確になったものと思われる。また、文の重要度も文書の中心概念との比較のうえで決定されるので的確な尺度となっているものと考えられる。しかしながら、これらは現段階では提案のレベルに留まっており、実験的な確認が最大の急務である。さらに、SVSMでは文間の概念の関係も分かるようになるので、文を単位とする様々な処理に用いることができる筈である。SVSMの応用範囲を拡大していくことも重要な課題である。

参考文献

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

[2] G.Salton. *Automatic Text Processing*, Addison-Wesley, 1989.

[3] S.Deerwester, S.T.Dumais, G.W.Furnas, T.K.Landauer, and R.Harshman. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, **41**, pp.391-407, 1990.

[4] M.W.Berry, S.T.Dumais, and G.W.O'Brien. Using Linear Algebra for Intelligent Information Retrival. *SIAM Review*, **37**, 4, pp.573-595, 1995.

[5] K.Fukunaga. *Statistical Pattern Recognition*. Second Edition. Academic Presss, Inc. 1990.

[6] 石井健一郎, 上田修功, 前田英作, 村瀬洋. パターン認識, オーム社, 1998.

[7] 相澤彰子, 影浦峡. 学術文献の著者キーワードに基づく専門用語間の関連度計算とその応用. 情報処理学会自然言語処理研究報告, **99**, 3, pp.55-62, 1999.

[8] E.Oya, 小川英光, 佐藤誠訳. パターン認識と部分空間法, 産業図書, 1986.

[9] 飯島泰蔵. パターン認識, コロナ社, 1973.

[10] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理, **6**, 6, pp.1-26, 1999.