覚え書

単語の共起に基づく関連文書検索、算法と検索事例

大学入試センター 研究開発部 情報処理研究部門 石岡 恒憲
株式会社リコー ソフトウェア研究所 亀田 雅之

Document retrieval based on words' cooccurrences, the algorithm and its application

Tsunenori Ishioka
Masayuki Kameda

要 旨　異なった文書に同時に現われる単語に着目することにより、潜在的な意味的検索をおこなう Deerwester(1990) の Latent Semantic Analysis を日本語の比較的大規模な文書集合に対して適用した。その中で、大型疎行列における特異値分解アルゴリズムの比較検討を行ない、日本語文書検索に適した方法を見つけた。これを実際の新聞記事で試し、文書検索、および関連語表示において有効であることの見通しを得た。また実装する上での工夫として、関連文書検索においては、文書の大きさによる基準化が必要なことがわかった。さらに、重複を許す単語のクラスタリングを試みた。

1　はじめに

近年、急速に関心の高まってきているデータマイニングの分野において、その適用分野の一つである文書マイニングは、goo や AltaVista など検索エンジン利用の普及と伴って、コンピュータの非専門家にとってもとりわけ注目の高いところである。文書マイニングは、膨大な電子化文書から興味深い知見を得るもので、その応用としては、文書/ウェブ検索、関連語検索、文書分類なとが挙げられる。

文書マイニングでは、扱うデータ量が膨大である上に、実用に耐えうる速度での応答が求められる。このため、その解決方法の一つとして、著者らは単語の共起に基づいた検索アルゴリズムに注目してきた。「単語の共起」とは、同一の文書/文に複数の単語が同時に出現することをいう。事実、この方法は、パターンマッチによる方法 (lexical searching technique) に比べ、検索効率が 30% 良いという報告がある (Letsche(1997))。

従来、単語の共起に着目した「文書マイニング」には大別して 2 つのアプローチがあったと思われる。

一つは、入力キーワードを含む文書集合に成立する相関ルールを求め、そのルールに基づき検索をおこなうものである。発見すべき知識は、どのような単語同士が共起しやすいかである。Apriori(Agrawal(1992)) は、その代表的なアルゴリズムである。「問答」(河野 (1996);http://www.kuamp.kyoto-u.ac.jp/labs/inforcom/mondou/)

Key words: singular value decomposition, latent semantic analysis, clustering

も、ユーザの検索要求の近傍のキーワード空間における相関ルールを関連語の知識として提示するものである。

もう一つのアプローチは、入力キーワード/問い合わせ文と検索対象文書に現われる単語との共起の度合いによって、より適切と考えられる文書を検索するものである (Gravano(1995) など)。基本的な考え方は、検索要求ベクトルに類似したベクトルをもつ文書は、適切な文書であると判断するもので、一般にはベクトル空間モデル (vector-space model) と呼ばれる。

その際に、単語の重み付けがしばしば行なわれるが、その方法として、単一文書中で出現する頻度 (within-story term frequency) に応じて重みを与える Luhn(1957) の tf 法と、その単語が出現する文書数の逆数 (inverse document frequency) に応じて重みを与える (すなわちさまざまな文書に出現するありふれた単語の重みを低くする) Jones(1972) の idf 法とを組み合わせた tf·idf モデル、もしくはその派生が用いられることが多い (これらの要約については Allan(1998) など)。

統計的色彩が強い方法としては、Deerwester(1990) によって提案された "Latent Semantic Analysis" がある。これは、共起の頻度を示す単語-文書行列を特異値分解 (たとえば柳井 (1983) など) することにより、文書の潜在的意味構造を抽出するものである。得られた意味空間において、互いに関連した文書や単語は近接するように構成される。この方法も結果的にはベクトル空間モデルの一つであるが、共起という一種のパターンマッチを間接的に用いているために、「入力キーワードを全く含まないが意味的に近い」文書をも選ぶこともできるようになる。たとえば、"結婚" という語を入力キーワードにして、『"結婚" という語を含まないけれども "披露宴" や "新婚旅行" といった "結婚" に関連の深い単語を含む文書」を検索することが可能となる。特に日本語の場合、表記のゆれ (「インターフェース」と「インタフェイス」など) の問題があり、パターンマッチによらない検索は、英語に比べ更に有利であると考えられる。

Latent Semantic Indexing (以下 LSI と略す) が英語の文献検索において有効なことは TREC(Text REtrieval Conference) などで既に主張されていることで、Berry(1995a,1995b) には、単語-文書行列に単語/文書の追加があった場合の特異値分解行列の更新方式について述べられている。また Letsche(1997) には、より効率的な LSI の実装について述べている。

しかしながら、日本語文書の場合は、単語を分かち書きしないという特徴のため単語の切り出しが難しいことと、一般に特異値分解は巨大なメモリ空間を必要とする (最も標準的と考えられる Numerical recipes (Press(1986)) にある特異値分解アルゴリズムでは、実行時間とメモリの関係から、ワークステーションでは事実上、実行不可能である) という理由のため、データ数が数千を越える大きさの問題に対して LSI を適用した事例は、国内においてはほとんどないものと思われる。

本稿では、この Deerwester(1990) による "Latent Semantic Analysis" の方法を下敷きにし、Berry(1992) による疎行列に適した特異値分解パッケージを用いることで、十分に大きい実データに対して、受容可能な応答を提供する実装について述べる。ある全国紙の過去の記事を用いて、検索、および関連語の表示をおこなった結果を例示し、重複を許す単語分類の一提案を行なう。

2　特異値分解モデル

記号表

記号	用例、定義	説明
A	$A = a_{ij}; i = 1, \cdots, p; j = 1, \cdots, q$	$p \times q$ 行列
$'$	A'	転置行列
	x'	転置ベクトル
$^{-1}$	A^{-1}	逆行列
I, I_n	$I_n = (\delta_{ij}; i, j = 1, \cdots, n)$	$n \times n$ 単位行列
$\| \ \|$	$\|x\| = (\sum_{i=1}^{n} x_i^2)^{1/2}$	ユークリッド・ノルム
diag	$\mathrm{diag}(a) = \mathrm{diag}(a_1, \cdots, a_n)$	ベクトル a の成分を対角要素とする対角行列
trace	$\mathrm{trace}(A)$	トレース

2.1　単語-文書データ

もととなるデータ:

単語-文書行列を $t \times d$ の行列 X と置く。t は単語数、d は文書数である。

行列 X の各要素を x_{ij} としたとき、

$$x_{ij} = \begin{cases} 1 & \text{単語 } i(1 \leq i \leq t) \text{ が文書 } j(1 \leq j \leq d) \text{ に含まれているとき} \\ 0 & \text{含まれないとき} \end{cases} \tag{1}$$

とおく。単語 i が文書 j に複数回、出現したときに x_{ij} に 1 を越える値を与えることも考えられるが、文書が新聞記事のような場合では文書中における単語頻度が、その単語の重要度と必ずしも見なすことができないので、そのようにはしていない。例えば、重要な既出単語が、文書中では「同なにがし」という形で表現されることは多い。

文書に含まれる単語は、形態素解析ツールにより取得する。本実験では、著者の一人が開発した QJP (亀田 (1995)) の形態素解析機能を利用した。QJP は高速 (10KB テキスト/秒、Pentium 133 MHz)、軽量 (5,000 語、100KB 未満の辞書、300KB 未満の動作メモリ) の日本語解析系 (日本語形態素解析および構文解析系) である。

特異値分解:

行列 X の特異値分解は、$T_0'T_0 = T_0T_0' = I_t$ および $D_0'D_0 = D_0D_0' = I_d$ を満たす直交行列 T_0 および D_0 を用いて以下の通り示される:

$$X = T_0 S_0 D_0' \tag{2}$$

ここで I_t および I_d はそれぞれ t 次、d 次の単位行列であり、T_0 は $t \times m$ 行列，S_0 は $m \times m$ の正方対角行列 (対角要素以外全て 0)，D_0' は $m \times d$ 行列である。また $0 \leq t \leq d$ とする。$'$ は転置を示す。S_0 の対角要素は大きい順とする。

幾つかの特異値分解ルーチンでは、S_0 の対角要素が大きい順であることを保証しないので、注意を要する。

次元低減:

行列 S_0 の対角要素を k 番目までとり、これを新たな行列 S とする。それに応じて、T_0 および D_0 も k 列までを抜き出し、これを新たな行列 T および D とする。このとき、

$$\widehat{X} = TSD' \tag{3}$$

となり、$\widehat{X}$ は X の近似となる。ここで T は $t \times k$ 行列 , S は $k \times k$ の正方対角行列 , D' は $k \times d$ 行列である。

経験的に k は、言語データの場合、$50 \sim 100$ 程度にすると良いようである。

特異値分解は幾つかの多変量解析手法の基本となるもので、$X'X$ の固有値問題は主成分分析に相当し、(3) 式において、TS は主成分得点、D' は主成分の係数を表わす。

また因子分析においては、T は共通性を 1 としたときの因子得点、SD' はその因子負荷行列に対応している。

2.2 特異値分解モデルから得られる基本的な比較量

行列 X の 2 つの行ベクトルのベクトル積は、その対象となる 2 つの単語が、全文書にわたってどの程度、共起の似たパターンを持っているかを示す量となる。したがって行列 $\widehat{X}\widehat{X'}$ は、全ての単語-単語のベクトル積をその要素とする $t \times t$ の正方対称行列となる。ここで、S は対角行列で、D は直交行列にあることに注意すれば、単語-単語間の近さ行列として、

$$\widehat{X}\widehat{X'} = TS^2T' \tag{4}$$

を容易に導くことができる。この式は行列 $\widehat{X}\widehat{X'}$ の i 行 j 列要素が、行列 TS の i 行と j 行とのベクトル積をとることで得られることを示している。ここで TS 空間は T (単語) 空間を特異値の大きさに比例して拡大/縮小したものになっている。

同じようにして、文書-文書間の近さ行列は、

$$\widehat{X'}\widehat{X} = DS^2D' \tag{5}$$

なる $d \times d$ の正方対称行列で与えられることがわかる。 (同じように DS 空間は D (文書) 空間を拡大/縮小したものである。)

一方、単語-文書間の近さは、$\widehat{X}$ の個々の要素で示すことができると考えられる。したがって単語-文書間の近さ行列を再度ここに示せば、

$$\widehat{X} = TSD' \tag{6}$$

となる。$\widehat{X}$ の i 行 j 列要素は、行列 $TS^{1/2}$ の i 行と行列 $DS^{1/2}$ の j 行とのベクトル積によって得られ、$TS^{1/2}$ 空間と $DS^{1/2}$ 空間との近さの度合いを示している。

これら統計量に関連する数学的な裏付けについては Papadimitriou(1998) に詳しい。

3　疎行列に対する特異値分解

単語-文書行列は一般に巨大な疎行列 (sparse matrix) となる。巨大な疎行列に対する特異値分解のためのソフトウェア・パッケージとしては、SVDPACK(Berry(1992)) が知られ、部分空間 (subspace) 反復法やランチョス (Lanczos) 法が利用できる。

Numerical recipes (Press(1986)) にもその算譜のある Householder 変換によって 2 重対角化にする方法は、もとの行列を逐次変換する方法であるから、疎行列に対してこれを適用すると、変換のたびに新たな非ゼロ要素が生成されて、疎であるという性質が失われてしまう。疎行列には適さない方法である。

ここで本稿の目的では、特異値、および特異ベクトルの全てが必要ではないことに注意する。大きい方から 50 ないし 100 個の特異値、および特異ベクトルがわかればよい。このような場合には、ランチョス法では 3 重対角化を途中で打ち切ってよく、ここで得られた小さな 3 重対角行列の特異値が、もとの行列の特異値のよい近似になっていることが知られている (Saad(1980))。部分空間反復法では、必要とする特異値の数よりも少し大きな数の次元をもつ部分空間で、特異値、および特異ベクトルのよい近似が得られる。

以下の 3.1~3.4に、比較検討のために取り上げた特異値分解の諸方法を示す。より詳しくは Berry(1992) を見られたい。なお SVDPACK は、http://www.netlib.org/svdpack/から入手でき、sis1, sis2 等のプログラム名は SVDPACK における略記である。

また、行列 X の特異値問題は、対称行列

$$\begin{pmatrix} O & X \\ X' & O \end{pmatrix},\ X'X \tag{7}$$

の固有値問題と同値であり、以下に示す特異値問題を解くアルゴリズムは、本質的に (7) 式の (疎な) 対称行列に対する固有値問題を解くアルゴリスムに過ぎないことに注意されたい。

3.1　部分空間反復法 (sis1, sis2)

部分空間反復法は、巨大な疎行列に対する特異値問題を解く最も単純なアルゴリズムである。Parlett(1980) が述べたように、これは古典的なベキ乗法 (power method) のブロック化と見なすことができる。部分空間反復法は Bauer(1957) にあり、$(t+d) \times (t+d)$ の行列

$$B = \begin{pmatrix} O & X \\ X' & O \end{pmatrix} \tag{8}$$

を用いて、以下を更新する。

$$Z_j = B^j Z_0 \tag{9}$$

ここで $Z_0 = [z_1, z_2, \cdots, z_s]$ は $(t+d) \times s$ である。

もし列ベクトル z_i が (ベキ乗法でなされるように) 分離して基準化されるならば、それら (列ベクトル) は、行列 B の優越固有ベクトル (絶対値最大の固有値に対応する固有ベクトル) に収束してゆく。したがって、行列 Z_j は漸次、列ごとの線形独立性を失ってゆく。行列 B の大きな p 個の特異

値ペアを近似するために、Bauer(1957) は、各ステップにおいて修正 Gram-Scmidt プロシージャを用い z_i を互いに直交にすれば、それらの間の線形独立性が保たれることを示した。しかしながら、z_i の B の特異ベクトルに対する収束はわずか 1 次に過ぎない。

SVDPACK では部分空間反復に、洗練された Rutishauser(1970) の ritzit プログラム (部分空間反復に、さらに Rayleigh-Ritz プロシージャと Chebyshev 多項式を経た高速化を行なっている) を使用している。

(8) 式で示された行列 B の特異値問題を解くプログラムを sis1、$B = X'X$ に対するそれを sis2 と略記する。

3.2 トレース最小化法 (tms1, tms2)

部分空間 (subspace) 法のもう一つの方法は、Sameh(1982) に述べられている一般的な固有値問題

$$Hx = \lambda Gx \tag{10}$$

におけるトレース最小化アルゴリズムに基づくものである。ここで H, G は対称行列で、G はさらに正定とする。

$t \times d$ の行列 X の特異値分解をおこなうために、行列 H を以下に定義される行列 $\widetilde{B}$ に置き換える (tms1):

$$\widetilde{B} = \begin{pmatrix} \gamma I & X \\ X' & \gamma I \end{pmatrix} \tag{11}$$

ここで、γ は、$\widetilde{B}$ が正定となるように定められる。あるいは $H = X'X$ とおく (tms2)。

このとき $G = I_{t+d}$ ($H = X'X$ のときは $G = I_d$) とおけば、Courant-Fischer 定理 (Wilkinson(1965)) により、$Y'Y = I_p$ を満たす $(t+d) \times p$ 行列 Y が存在して、

$$\min_{Y} \ \mathrm{trace}(Y'\widetilde{B}Y) = p\gamma - \sum_{i=1}^{p} \sigma_i \tag{12}$$

が、成り立つから、この最小化問題を解くことで、行列 X の特異値を得ることができる (Sameh(1982))。ここで、σ_i は、行列 X の特異値で $\lambda_i = \gamma \pm \sigma_i$ は、行列 $\tilde{B}$ の固有値である。また $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$ である。

3.3 ランチョス法 (las1, las2)

ランダムに定めた基準化ベクトルに、行列 B を掛ける演算を繰り返すことによって、B と相似な３重対角行列 T の要素を計算する方法である。

(8) 式で示された行列 B を用いるプログラムを las1、$B = X'X$ に対するそれを las2 と略記する。

3.4 ブロック (block) ランチョス法 (bls1, bls2)

部分空間反復法は固有値問題に対するベキ乗法のブロック化であるので、ベクトルランチョス法による繰り返しに対しても、同じようなブロック化を考えることができる。

この方法をブロック・ランチョス法と呼び、(8) 式で示された行列 B を用いるプログラムを bls1、$B = X'X$ に対するそれを bls2 と略記する。

3.5 性能評価実験

ある全国紙の 1 年分 (1994 年) の記事 (毎日新聞の CD_ROM を利用) から、先頭 2,055 文書と、そこに出現する頻度 4 以上の 4,041 単語 (名詞) を用いて、SVDPACK で用意されている $4 \times 2 = 8$ 通りの方法に対して、特異値分解に要した CPU 時間、ならびに所要メモリをまとめたのが表 1である。

さらに、1 年分の記事に現われる 400,776 単語のうち、頻度 10 以上の 44,883 単語と 20,211 文書を用いた場合の CPU 時間、ならびに所要メモリも併記する。(一部の方法については、メモリの制限、および時間がかかり過ぎるなどの理由から、実行できていない。これを表中—で示す。)

表 1: SVDPACK による特異値分解の計算

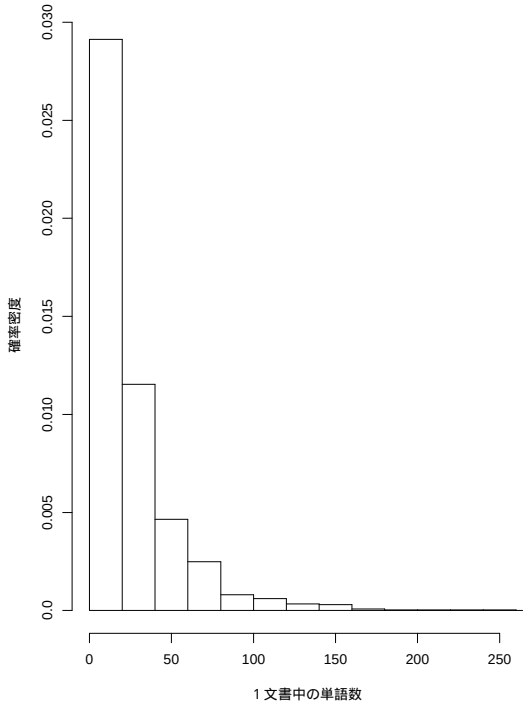
プログラム		{2,055 文書、4,041 単語}		{20,211 文書、44,883 単語}	
		CPU 時間 (秒)	所要メモリ (MB)	CPU 時間 (秒)	所要メモリ (MB)
部分空間	sis1	208	44.7	—	—
反復法	sis2	97.6	15.2	1780	148
トレース	tms1	989	9.8	—	—
最小化法	tms2	663	5.1	—	—
ランチョス法	las1	61.7	26.1	—	—
	las2	9.53	12.7	134	53.0
ブロック	bls1	174	14.2	2680	156
ランチョス法	bls2	82.3	10.2	1810	90.8

一つの文書に含まれる単語数は、平均で 49.5 であり、その標準偏差は 48.3 である。さらに頻度 10 以上の出現頻度のもつ単語に限れば、一つの文書に含まれるそれらの単語の数は、平均で 42.7 であり、その標準偏差は 42.4 である。参考までに、1 年分の記事において、1 文書中に現われる単語数の分布と、単語出現頻度; 着目した単語の出現文書数の分布を図 1に示す。

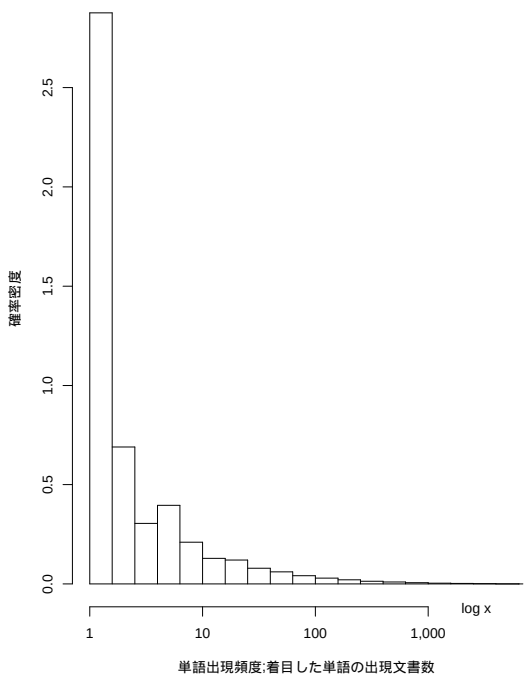
必要な特異値の数は 50 とし、これに 10 を加えた数 (すなわち 60) まで 3 重対角化を行なうようにパラメータを設定した。また必要な特異値の精度は 10^{-6} とした。2,055 文書の場合、非ゼロ要素数は 52,909 で、その出現密度は 0.64%である。20,211 文書の場合、非ゼロ要素数は 862,054 で、その出現密度は 0.095%である。

実行時におけるマシン環境は以下の通りである。

OS: SunOS 5.5.1 (Solaris 2.5.1)
アークテクチャ: sun4u sparc SUNW, Ultra-1



(a) 1 文書中に現われる単語数の分布



(b) 単語出現頻度 (着目した単語の出現文書数) の分布

図 1: 1 年分の記事に現われる単語/文書

C コンパイラ: GNU gcc 2.7.2.3
コンパイル ・オプション: –O

これより、この程度の大きさ、およびこのような (日本語文書における単語-文書の共起行列のような) データ構造における特異値分解では、las2 が速く実行できることがわかる。また bls{1,2} は las{1,2} の少なくとも実行時間における改良にはなっていないことがわかる。英語文書における Berry(1992) の結果とは、必要とする特異値の数、および取り扱うデータの大きさが異なるので単純な比較は難しいが、las2 が最速であることは違いがないようである。

ちなみに最初に読み込む疎行列データのデータ格納形式には、Duff(1989) にある Harwell-Boeing sparse matrix format を用いる。これは、疎行列の行数、列数、非ゼロ要素数、列ごとに読み込んだ場合の各列先頭時点における非ゼロ要素の累積数+1、各列における非ゼロ要素の行番号 (1 から始まる) と非ゼロ要素の値そのもの (整数、あるいは実数) を順に記述するものである。疎行列に対してデータを効率よく格納できるので、ディスクの節約、ならびにデータ読み込み時間の大幅な低減をはかることができる。

4 疑似文書を用いた検索

著者らが取り扱う新聞記事のような巨大な言語集体は、一般にコーパス (corpus) と呼ばれる。コーパスを用いて達成すべきわれわれの課題は、以下の 2 つである。

- このコーパスから意図する文書/単語を検索すること (4節)。
- コーパスを代表する単語群を自動的に分類/抽出すること (5節)。これができれば、ジャンルごとの基本辞書が自動的に作成でき、将来的にはシソーラスの自動生成への期待がもてると考えられる。

4.1 類似文書の検索

問い合わせに用いる疑似文書 q のデータは、t 次元の単語ベクトル x_q で表現することができ、これを用いて、文書空間 D の行に対応する $1 \times k$ の文書ベクトル

$$d_q = x_q' T S^{-1} \tag{13}$$

を導くことができる。

ここで T は $t \times k$ 行列、S は $k \times k$ 正方対角行列である。$'$ は転置を、$^{-1}$ は逆行列を示す。$S = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_k)$ としたとき、$S^{-1} = \mathrm{diag}(1/\mu_1, 1/\mu_2, \ldots, 1/\mu_k)$ である。

ここで疑似文書 d_q (k 次元ベクトル) に対し、比較の対象とする文書を d_c (k 次元ベクトル) とすれば、両文書の相関係数 $r(d_q, d_c)$ は、両文書がなす角の余弦で与えられる:

$$r(d_q, d_c) = \frac{(d_q, d_c)}{\|d_q\| \, \|d_c\|} \tag{14}$$

これより、疑似文書 d_q に近い文書を、近さの順に提示することが可能となる。なお (14) 式の右辺分子の括弧は、内積を示す。

ここで疑似文書 d_q、および比較の対象とする文書 d_c は、いずれも k 次元に次元低減されていることに注意しなければならない。行列 T, S, D は、検索前に予め用意しておくことができるので、検索時には、(13) 式と、比較する文書の数だけの (14) 式の計算をすればよい。(14) 式分母の $\|d_c\|$ も予め用意しておくことができ、$\|d_q\|$ の計算もわずか 1 回で済む。大量の文書検索では、リアルタイムでの応答が求められるので、検索時に必要な計算量を減らす工夫は、きわめて大切である。

もちろん、問い合わせは、文書の形である必要はない。一つ、あるいは複数のキーワードの並びであってよい。このような場合、それらキーワード (単語) を含む疑似文書を仮定し、その疑似文書 d_q に近い文書を、近さの順に提示する。

例として「政治」、「改革」、「政治改革」という 3 つの単語をキーワードにして {20,211 文書、44,883 単語} から検索した結果、得られた相関係数の高い上位 3 文書を以下に示す。

最初の # に続く番号は、システム内部での文書番号を示し、その後の括弧の中身は相関係数を示す。たとえば、文書番号 17252 に対する相関係数は 0.999 である。

> # 17252(0.999) # 掲載面: ２面 （０２ ）# その他: ’９４．３．１１ 朝刊 ２頁 写図無 (全１１４文字）# 記事題: 政治改革法、きょう公布 公職選挙法改正など４日に成立した政治改革法が１１日付の官報で公布される。４法のうち衆院選挙区画定審議会設置法改正は、８日に閣議決定された同法施行令と合わせて１１日に施行され、小選挙区３００の区割り作成に向けた動きが本格化する。
>
> # 5285(0.998) # 掲載面: ２面 （０２ ）# その他: ’９４．１．２５ 朝刊 ２頁 写図無 (全１６７文字）# 記事題: 民社党３５年、「記念日の集い」 結党３５年目を迎えた民社党の「記念日の集い」が２４日、東京・虎ノ門の同党本部で開かれた。政治改革関連法案をめぐる与野党折衝のあおりで、出席したのは神田厚院内総務ら国会議員４人と党職員だけという寂しい集会だったが、大内啓伍委員長 (厚相) は「民社党が提起した労働運動や真の議会制民主主義、政策が実現されていることを誇りに思う」と述べた。
>
> # 5896(0.997) # 掲載面: 社説 （０５ ）# その他: ’９４．１．２７ 朝刊 ５頁 写図無 (全２６４文字）# 記事題: [みんなの広場] 採決は分離できないのか＝無職・松浦清治 ６６ (千葉県四街道市） 六年越しの政治改革関連法案が土壇場で否決された。しかし、財界人、識者の反応に比べ、街の声は意外とクールだった。それは国民の望んだ政治改革は腐敗防止という、単純明快なものであったのに、小選挙区制度を含む四法案一括という形に変質して、期待を裏切ったことへのしらけもあったのではないだろうか。
> ところで、関連法案だから一括審議も分かるが、採決は分離できないのだろうか。だれがどの法案に反対、賛成したか国民に分かる。今回の政治改革関連法案もすべてが否定されず、改革の足掛かりぐらいは残せたのではないかと思っている。

なお実装にあたっては、文書の大きさによる基準化をおこなった。大きな文書ほど単語共起が起きやすいので、1 文書中に現われる単語数で共起頻度を割ったものを、もとの単語-文書行列に与えてい

る。上の結果は、この基準化データに対する結果である。1 文書中に現われる単語数の平行根で割るのも、経験的に良い基準化である。

4.2　関連語の検索

疑似文書がキーワード (単語) の並びであった場合に、その疑似文書に関係の深い単語を提示することも重要である。これは一般に “関連語表示” と呼ばれるもので、検索結果があまり的を得ていなかったり、あるいは検索結果が多すぎてさらに絞り込む必要があったときに、関連語を表示することで検索を支援することが考えられる。

さて疑似文書は、t 次元の単語ベクトル x_q で表現することができることに注目すれば、疑似文書の座標をその疑似文書が含む単語群の中心 (centoroid) に定めるのは、きわめて妥当であると思われる。すなわち

$$t_q = \frac{1}{n_q} \sum_{t_i \in T} t_i \tag{15}$$

とする。ここで、n_q は、疑似文書 q が含む単語の数、t_i は行列 T の行ベクトルである。これが妥当であると理由としては、ユーザが専門用語を問い合わせのキーワードとして選びやすく、そのためベクトルとして互いに距離の近いものからなる集合を作ることが挙げられる。

このようにすれば、比較の対象とする単語を t_c (k 次元ベクトル) とするとき、両単語の相関係数は、(14) 式と同じように $r(t_q, t_c)$ として与えられる。これより疑似文書に関係の深い単語をその近さの順で提示することができるようになる。

4.1節と同じように、例として 「政治」、 「改革」、 「政治改革」という 3 つの単語をキーワードにして {20,211 文書、44,883 単語 } から検索した結果、得られた関連の大きい上位 30 単語は以下の通りである。単語の後の括弧の中身は、相関係数を示す。

> 政治改革 (0.997)、改革 (0.996)、政治 (0.996)、法案 (0.994)、参院否決 (0.989)、政治改革関連 (0.986)、関連法案 (0.985)、関連法 (0.985)、政治改革関連法案 0.984)、政治改革関連法 (0.984)、政治改革法案 (0.984)、改革法案 (0.984)、法案採決 (0.978)、否決 (0.977)、野党折衝 0.977)、関連 (0.977)、法案成立 (0.976)、党籍 (0.972)、与野党折衝 (0.972)、成立 (0.970)、社会党首脳 (0.970)、政治改革関連法案採決 (0.969)、関連四法案 (0.968)、議席配分 (0.965)、大選挙区 (0.963)、不成立 (0.961)、参院政治改革特別委 (0.961)、小選挙区制 0.960)、改革法 (0.960)、票読み (0.960)、中選挙区制 (0.959)

実際のアプリケーションでは、問い合わせに用いたキーワードそれ自身は、除いて表示する必要があろうが、概ね、良好な結果が得られていることがわかる。

5　単語重複を許すクラスタリング

単語あるいは文書を予め定められたカテゴリーに分類する方法は、通常は重複を許していないという点で、我々の要求しているものではない。重複は許されていなければならない。言語では、一つの

単語が複数の異なった意味を有する場合があり、たとえば 「勝手」には、 「勝手きまま」の意味と 「勝手口 (台所)」の意味とがあるからである。

もっとも自然言語で書かれた文書を予め定められたカテゴリーに分類する方法は、Text categorization と呼ばれ、ここ数年、統計的なアプローチが積極的に試みられている。たとえば、回帰モデル (Yang(1994)) や、最近傍分類法 (Creecy(1992))、ベイジアン分類法 (Tzeras(1993))、決定木 (Fuhr(1991))、ルール学習アルゴリズム (Apte(1994))、ニューラル・ネットワーク (Wiener(1995))、オンライン学習法 (Cohen(1996)) などがある。しかしながら、これらは計算量的な問題のためか、排他的なクラスタリングであって、重複を許すクラスタリングになってはいない。本稿と同じ LSI を用いたクラスタリングには Schütze(1997) があるが、これにしても次元低減した文書ベクトル間の空間距離を群平均法でクラスタリングするもので、クラスタリングそれ自体は排他的なものである。

さて、改めて (3) 式を見ると、T は共通性を 1 としたときの因子得点である。このことに着目すれば、各因子ごとに因子得点の絶対値の大きい単語をまとめることで、重複を許す単語のクラスタリングを考えることができることに気づく。各因子は、新聞記事を分類したときのカテゴリーに相当すると考えられる。

そこで 50 個の因子からクラスターを形成してみたのが以下の結果である。紙面の都合で、20 個まで取り上げた。各因子ごとに、(i) 因子得点の絶対値が最大となる単語を探し、(ii) それとの比率の絶対値が 0.3 以上で、かつ因子得点の符号が同じものをまとめた。したがって各クラスターにおける個数は不定となる。クラスター番号 (cluster no.) の後の括弧中の数字は、特異値である。単語の後の括弧中の数字は、因子得点である。

cluster no. = 1 (69.5)

問題 (0.209)、細川護熙首相 (0.190)、自民党 (0.185)、可能性 (0.159)、首相 (0.155)、社会党 (0.151)、米国 (0.150)、考え (0.141)、動き (0.135)、政府 (0.134)、意見 (0.125)、連立与党 (0.122)、姿勢 (0.118)、必要 (0.117)、与党 (0.110)、方針 (0.097)、立場 (0.096)、言葉 (0.092)、見方 (0.090)、責任 (0.089)、政治改革関連法案 (0.087)、影響 (0.087)、最大 (0.086)、細川首相 (0.084)、関係 (0.084)、政治 (0.083)、理由 (0.078)、国会 (0.077)、最後 (0.076)、企業 (0.074)、批判 (0.074)、内容 (0.074)、法案 (0.073)、政治改革 (0.072)、合意 (0.067)、事態 (0.067)、連立政権 (0.065)、記事 (0.064)、意味 (0.063)、判断 (0.063)

cluster no. = 2 (46.9)

自民党 (-0.238)、細川護熙首相 (-0.228)、社会党 (-0.205)、首相 (-0.166)、連立与党 (-0.164)、与党 (-0.149)、政治改革関連法案 (-0.129)、考え (-0.099)、法案 (-0.087)、政治改革 (-0.082)、参院 (-0.077)、細川首相 (-0.075)、新生党 (-0.074)

cluster no. = 3 (39.3)

病院 (0.577)、喪主 (0.552)、告別式 (0.541)

cluster no. = 4 (37.0)

米国 (0.456)、政府 (0.176)

cluster no. = 5 (34.1)

選手 (0.272)、米国 (0.206)、最後 (0.177)、チーム (0.161)、優勝 (0.152)、期待 (0.133)、試合 (0.115)、動き (0.113)、大会 (0.110)、金メダル (0.099)、メダル (0.095)、練習 (0.094)、ボール (0.093)、今季 (0.087)、トップ (0.087)、言葉 (0.085)、目標 (0.085)、自信 (0.084)、細川護熙首相 (0.082)、気持ち (0.082)

cluster no. = 6 (32.1)
可能性 (0.532)、調べ (0.170)

cluster no. = 7 (31.1)
細川護熙首相 (0.345)、首相 (0.307)、女性 (0.163)、調べ (0.123)、事件 (0.117)、考え (0.116)、可能性 (0.114)、遺体 (0.109)、武村正義官房長官 (0.109)

cluster no. = 8 (30.5)
問題 (0.309)、米国 (0.300)、子供 (0.150)、自民党 (0.104)、言葉 (0.098)

cluster no. = 9 (29.5)
問題 (-0.621)

cluster no. = 10 (28.4)
可能性 (0.542)、子供 (0.219)、問題 (0.176)

cluster no. = 11 (27.5)
政府 (0.445)、問題 (0.192)、女性 (0.168)

cluster no. = 12 (26.8)
方針 (0.340)、子供 (0.237)、影響 (0.210)、意見 (0.201)、理由 (0.171)、米国 (0.138)、記事 (0.125)、姿勢 (0.115)、動き (0.110)

cluster no. = 13 (26.7)
問題 (0.381)、動き (0.231)、女性 (0.180)、細川護熙首相 (0.180)、影響 (0.168)、米国 (0.162)

cluster no. = 14 (26.1)
動き (0.293)、必要 (0.198)、意見 (0.173)、社会党 (0.166)、遺体 (0.122)、大阪 (0.119)、上田容疑者 (0.107)、捜査本部 (0.098)、調べ (0.098)、考え (0.098)、共同捜査本部 (0.097)、立場 (0.092)、高橋 (0.092)

cluster no. = 15 (25.5)
動き (0.497)、政府 (0.395)

cluster no. = 16 (25.3)
影響 (-0.403)、言葉 (-0.349)、最後 (-0.200)、社会党 (-0.132)、大蔵省 (-0.128)

cluster no. = 17 (25.2)
子供 (0.505)、可能性 (0.169)

cluster no. = 18 (25.0)

女性 (-0.550)

cluster no. = 19 (24.8)

影響 (-0.320)、女性 (-0.215)、理由 (-0.170)、米国 (-0.162)、ロシア (-0.143)、政治 (-0.133)、政府 (-0.120)、必要 (-0.116)、首相 (-0.099)

cluster no. = 20 (24.6)

米国 (0.383)、社会党 (0.152)、可能性 (0.138)、意見 (0.130)

以下、省略。

この結果をみるに、どうも重複の単語が多すぎ、また似たような特性を持つ因子が散見され、十分なクラスタリングができていないように思われる。

4節のように文書の大きさによる基準化、あるいは単語の出現頻度による基準化も試みてみたが、結果はこれより芳しいものではなかった。そもそもクラスタリングとは、クラスター内における各データ間の距離分散を小さくしつつ、クラスター間の距離分散を最大にすることであり、ここではその作業は含まれていない。その意味で、我々がおこなった方法はクラスラリングいう名に値しないし、結果もその限界を示したものかもしれない。その半面、全く捨てさるほどのものでもなく、実用には今後の検討が必要であると思われる。

6 まとめ

Deerwester(1990) にある Latent Semantic Analysis を日本語の比較的大規模な文書集合に対して適用した。その中で、大型疎行列における特異値分解アルゴリズムの比較検討を行ない、本稿の目的である単語の共起を示す日本語の単語-文書行列に適用する際に適した方法がわかった。これを実際の新聞記事で試し、文書検索、および関連語表示において有効であることの見通しを得た。また実用レベルに落とし込むための工夫として、関連文書検索においては、文書の大きさによる基準化が必要なことがわかった。さらに、重複を許す単語のクラスタリングを試みた。

検索精度は、一般に (英語の場合)、再現率 (正解文書を漏れなく検索できる能力指数) と適合率 (正解文書のみを検索できる能力指数) の両方で評価されるが、日本語の場合は、表記のゆれ (「インターフェース」と 「インタフェイス」など) や異体字 (「斉藤」、 「斎藤」、 「齋藤」など)、さらに翻訳語としての同義語 (「コンピュータ」と 「電子計算機」など) の問題がある。どのようにして検索精度を測定すべきか、という問題を含めて、検索精度の評価は今後の課題となろう。また Latent Semantic Analysis は、従来、知識工学的な側面が強かったデータマイニングの分野に、統計学的な面からアプローチを試みるものであり、統計学者の貢献が期待されるところである。

なお、著者等が、この研究を行なっている途中で、我が国においても、ジャストシステムが意味的に類似した文書を抽出するタイプの検索システム ConceptBase Search を 1997 年 12 月に販売した。使用されている技術の詳細については不明であるが、パンフレットや解説記事 (星野 (1997), 野村 (1999)) に散りばめられているキーワードからは、ベクトル空間モデルであること、ならびに単語間の相関度に着目した 「意味」を抽出するのが特徴であるとのことである。

謝辞

本稿の改訂にあたり、レフェリーおよび編集理事の白幡慎吾先生より、多くの適切なご意見をいただきました。大学入試センター ・研究開発部の柳井晴夫教授、ならびに清水留三郎教授には、草稿の段階から、丁寧に査読いただき、有益なご意見、ご批判をいただきました形態素解析結果の加工には、株式会社リコー ・ソフトウェア研究所の林 寛子氏が作成したツールを使用しました。
ここに記して、厚くお礼申しあげます。

参 考 文 献

Allan, J., Carbonell, J., Doddington,G., Yamron,J. and Yang, Y.(1998):Topic Detection and Tracking Pilot Study Final Report, *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February.

Agrawal,R., Ghosh, S., Imielinski,T., Iyer,B. and Swami, A.(1992):An Interval Classifier for Database Mining Applications, *Proc. the 18th International Conference on Very Large Data Bases*, Vancouver, Canada, August, pp.560–573.

Apte, C., Damerau, F. and Weiss, S(1994): Towards language independent automated learning of text categorization models, in *Proceedings of the 17th Annual ACM/SIGIR conference*.

Bauer, F.L.(1957): Das Verfahren der Treppeniteration und verwandte zur Lösung algebraischer Eigenwertprobleme, *ZAMP*, 8, 214–235.

Berry, M.W., Dumais, S.T., and O'Brien, G.W.(1995a): Using Linear Algebra for Intelligent Information Retrieval, SIAM Review, Society for Industrial and Applied Mathematics, Vol. 37, No. 4, pp. 573–595.

Berry, M.W., Dumais, S.T. and Letsche, T.A.(1995b): Computational Methods for Intelligent Information Access. *Proceedings of Supercomputing'95*, San Diego, CA.

Berry, M.W.(1992): Large scale singular value computations, *International Journal of Supercomputer Applications*, Vol. 6, No. 1, pp.13–49.

Cohen, W.W. and Singer, Y(1996): Context-sensitive learning methods for text categorization, in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development on in Information Retrieval (SIGIR'96)*, pp.307–315.

Creecy, R.H., Masand, S.L., Smith, S.J. and Waltz, D.L.(1992): Trading mips and memory for knowledge engineering:classifying census returns on the connection machine, *Comm. ACM*, Vol.35, pp.48–63.

Deerwester,S., Dumais, S.T. Furnas, G.W. Landauer, T.K. and Harshman, R(1990): Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 7, pp.391–407.

Duff, I.S., Grimes, R.G. and Lewis, J.G.(1989): Sparse matrix test problem, *ACM Trans. Math. Software*, Vol. 15, pp.1–14.

Fuhr, N., Hartmanna, G., Schwantner, M. and Tzeras, K(1991): Air/x - a rule-based multistage indexing systems for large subject fields, *Proceedings of RIAO'91*, pp.606–623.

Gravano, L. and Garcia-Molina, H.(1995): Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies, *in Proc. of the 21st International Conference on Very Large Data Bases (VLDB'95)*, pp.78–89.

星野 友彦 (1997): 検索技術の新潮流 「コンセプト ・マイニング」疑似文書を掘り当てる新ソフト、日経コンピュータ、1997.12.8、pp.154–158.

Jones,K.S.(1972):A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–21.

亀田 雅之 (1995):軽量 ・高速な日本語解析ツール 簡易日本語解析系 QJP 」、言語処理学会、第 1 回年次大会、pp.349–352.

河野 浩之、長谷川 利治 (1996): WWW 情報空間における文書データマイニングを用いた知的検索システム, *Proc. of Advanced Database Symposium '96*, pp.27–34, Tokyo.

Letsche, T.A. and Berry, M.W.(1997): Large-Scale Information Retrieval with Latent Semantic Indexing. *Information Sciences - Applications*, Vol. 100, pp. 105–137.

Luhn,H.P.(1957):A Statistical Approach to Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, Vol.1, No.4, pp.307–317.

野村 直之 (1999):ConceptBase の言語処理と新しいソリューション、情報処理学会 自然言語処理研究会 129-1, pp.1–8.

Papadimitriou, C.H. , Raghavan, P., Tamaki, H., and Vempala, S. , Latent Semantic Indexing: A Probabilistic Analysis. *PODS* 1998: pp.159–168.

Parlett, B.(1980):*The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ.

Press, W.H.(1986): *Numerical recipes : the art of scientific computing*, Cambridge University Press.

Rutishauser, H.(1970):Simultaneous iteration method for symmetric matrices, *Numer. Math.* , Vol.16, pp.205–223.

Saad, Y.(1980): On the rates of convergence of the Lanczos and the block-Lanczos methods, *SIAM Journal of Numerical Analysis*, Vol. 17, pp.687–706.

Sameh, A.H. and Wisniewski(1982):A trace minimization algorithm for the generalized eigen value problem, *SIAM Journal of Numerical Analysis*, Vol.19, No.6, pp.1243–1259.

Schütze, H. and Silverstein, C.(1997): Projections for Efficient Document Clustering, *Proceedings of SIGIR*, pp.74–81.

Tzeras, K. and Hartman, S.(1993):Automatic indexing based on bayesian inference networks, in *Proc. 16th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pp.22–34.

Wiener, E., Pederson, J.O. and Weigend, A.S(1995): A neural network approach to topic spotting, in *Proceedings of the fourth Annual Symposium on Document Analysis and Information Retrieval (SIGIR'95)*,

Wilkinson, J.H.(1965): *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford.

柳井晴夫、竹内啓 (1983): 射影行列・一般逆行列・特異値分解、UP 応用数学選書 10、東京大学出版会.

Yang, Y. and Chute, C.G.(1994): An example-based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems(TOIS)*, pp.253–277.

著者連絡先 :〒 153-8501 東京都目黒区駒場 2-19-23
大学入試センター 研究開発部 情報処理研究部門
Phone: 03-3468-3311 E-mail: tunenori@rd.dnc.ac.jp