統計的手法を用いた自然言語データベースからの意味的検索および自動分類

大学入試センター 研究開発部　石岡 恒憲

要約：
潜在的な意味的検索をおこなう Deerwester(1990)の Latent Semantic Indexing 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、文書検索において有効であることを確認した。また重複を許す単語のクラスタリングを試みた。さらにクラスタリングについては、*k*-means 法の逐次繰返しと BIC による分割停止基準を用いることで，情報理論的に最適と考えられるクラスター数を自動的に決定するアルゴリズムを提示した。

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、google 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、文書分類などが挙げられる。

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/)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)。

さて、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 Indexing ”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(1995)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、国内においてはすぐには現われなかった。だが、石岡(1999)などの啓蒙もあり、LSI の優位性は我が国においても広く知られるようになった。事実、現在では多くの検索システムのエンジンに、この LSI の手法が用いられつつある。

そこで本稿の前半では、LSI 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紹介する(2 節)。また、毎日新聞の過去の記事を用いて、検索をおこなった結果を例示し、重複を許す単語クラスタリングの一提案を行なう(3 節)。技術的な詳細

は石岡(1999)を見られたい。

さてこのような検索技術を獲得した今、次に我々が早急に獲得すべき技術は、得られたデータをいかによく整理するかである。その代表的な方法の一つにクラスタリング（グループ分け）がある。しかし、大量データのクラスタリングに向く k-means 法や近年、盛んに使われている自己組織化マップ（コホネン・マップ; Kohonen,1996）は、分けるべきクラスター数を予め決めておくものである。データマイニングあるいは知識発見といった立場からは、取り扱うべき対象が幾つのクラスターに分割されるか、といったクラスター数それ自体が有益な知見になるため、クラスター数が自動的に設定される方法が望まれる。

そこで本稿の後半（4 節）では、クラスター数を自動決定するアルゴリズムについて紹介する。内容の詳細は石岡(2000)を見られたい。

2 単語の共起に基づく関連文書の検索

2.1 Latent Semantic Indexing

LSI は予め十分に多くの文書に出現する単語の頻度を表した $k \times d$ の行列 X （t は単語数、d は文書数）を特異値分解（たとえば柳井,1983 などを参照）

$$X = T_0 S_0 D_0'$$

することから始まる。T_0 および D_0 は、$T_0' T_0 = T_0 T_0' = I_t$ および $D_0' D_0 = D_0 D_0' = I_d$ を満たす直交行列である。ここで、I_t および I_d はそれぞれ t 次、d 次の単位行列である。また $0 \le d \le t$ とする。'は転置を示し、S_0 の対角要素は大きい順とする。

ここで行列 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、

$$\hat{X} = TSD' \qquad (1)$$

となり、$\hat{X}$ 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。Deerwester(1990)によれば、言語データの場合、経験的に k は 50 ～100 程度にすればよい。

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、(1)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。

さて行列 X は一般に巨大な疎行列(sparse matrix)となるが、このような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、Berry(1992)の SVDPACK が知られる。SVDPACK は http://www.netlib.org/svdpack/から入手でき、そこでは 8 通りのアルゴリズムが利用できる。

2.2 性能評価実験

毎日新聞の１年分の記事（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 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。）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。

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。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。

表１：SVDPACK による特異値分解の計算

		{2,055文書、	4,041単語}	{20,211文書、	44,833単語}
プログラム		CPU時間	所要メモリ	CPU時間	所要メモリ
		(秒)	(MB)	(秒)	(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
ブロック	bls1	174	14.2	2680	156
ランチョス法	bls2	82.3	10.2	1810	90.8

参考までながら、ここでいう単語とは IPA 品詞体形(THiMC097)でいう「名詞」のうち、一般（普通名詞）、固有名詞-一般（一般的な固有名詞）、固有名詞-組織（組織を表す名称、「通産省」など）、固有名詞-地域-一般（国名以外の地名）、固有名詞-地域-国（国名）、サ変接続（格要素をとり、後ろに「する」、「できる」などが後接できるもの、「悪化」、「下取り」など）、形容動詞語幹（いわゆる形容動詞の語幹で、「な」の前に現われるもの、「健康」、「安易」など）とした。

2.3 疑似文書を用いた検索

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

$$d_q = x_q' TS^{-1} \tag{2}$$

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 = diag ($\mu_1, \mu_2, \ldots, \eta_k$) としたとき、S^{-1} = diag$(1/\mu_1, 1/\mu_2, \ldots, 1/\eta_k)$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:

$$r(d_q, d_c) = \frac{(d_q, d_c)}{\|d_q\| \|d_c\|} \tag{3}$$

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(3)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、(2)式と、比較する文書の数だけの(3)式の計算をすればよい。(3)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。

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。

３ 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。

さて、改めて(1)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、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、10 個まで取り上げた。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)

以下、省略。

no.=1のクラスターには新聞記事を代表する典型的な用語が含まれ、no.=2のクラスターには国内政治に関する用語が含まれること、no.= 3 のクラスターは死亡公告に関する用語で構成されていること、no.=4のクラスターはアメリカ関連、no.=5のクラスターはオリンピック関連の用語であることなどがわかる。

しかしながら、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のも事実である。その理由としては、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るが、ここではその作業は含まれていないためであろうと思われる。

4 クラスター数を自動決定する *k*-means アルゴリズムの拡張

4.1 *k*-means 法

非階層的なクラスタリング方法の1つである *k*-means 法は、クラスター数を *k*、標本の大きさを *N* としたとき O(*kN*)の計算量で済むために、自己組織化マップ(self-organizing map)とともに、特に大規模データのクラスタリングにしばしば用いられている。ことに近年のデータマイニング研究の隆盛にともない、*k*-means の高速化の手法が精力的に開発されている。しかしデータマイニングツールとしての見地からいえば、予めクラスター数を定めなければならないことは、より大きな制約であると考えられる。

そこで本節では、情報量規準の一つである BIC(Bayesian Information Criterion; Schwarz,1978)を用いることで、十分に小さなクラスター分割から始めて、各サブクラスターにおいて、分割が妥当と判断されるまで2分割を繰り返すアルゴリズムを支持し、これが有効に機能することを示す。

4.2 *x*-means 法

x-means 法は *k*-means におけるクラスター数 *k* が未知であることに由来する。*x*-means のアルゴリズムそのものはきわめて単純で、始めに十分に小さなクラスター分割から始めて、各サブクラスターについて2分割が適当であると判断される限り、分割を繰り返すものである。そのアルゴリズムは、以下のように要約される。

0. 解析すべきデータとして *n* 個の *p* 次元データを用意する。

1. 十分に小さなクラスター数の初期値 k_0（特に指定しなければ2）を定める。

2. $k = k_0$ として *k*-means を適用する。分割後のクラスターを $C_1, C_2, \ldots, C_{k_0}$ とする。

3. $i = 1,2,\ldots,k_0$とし、手順 4〜9 を繰り返す。
4. クラスターC_iに対して $k = 2$ として k-means を適用する。分割後のクラスターを C_i^1, C_i^2とする。
5. C_i に含まれるデータ x_iに p変量正規分布

$$f(\theta_i; x) = (2\pi)^{-p/2}|V_i|^{-1/2}\exp\left[-\frac{1}{2}(x-\mu_i)^t V_i^{-1}(x-\mu_i)\right]$$

を仮定し、そのときの BIC を以下により計算する:

$$\text{BIC} = -2\ \log\ L(\hat{\theta}_i; x_i \in C_i) + q\ \log\ n_i$$

ここに$\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$は、$p$変量正規分布の最尤推定値とする; μ_iは p次の平均値ベクトル、V_iは $p \times p$の分散・共分散行列である qはパラメータ空間の次元数で、V_iの共分散を無視すれば（0 と置けば）$q = 2p$である。共分散を無視しなければ、$q = p(p+3)/2$である。
x_iはクラスター C_iに含まれる p次元データとし、n_iは C_i に含まれるデータする。Lは尤度関数で$L(\cdot) = \prod f(\cdot)$である。

6. C_i^1, C_i^2のそれぞれに対して、パラメータθ_i^1, θ_i^2をもつ p変量正規分布を仮定し、2 分割モデルにおいてデータの従う確率密度を

$$x_i \sim \alpha_i[f(\theta_i^1; x)]^{\delta_i}[f(\theta_i^2; x)]^{1-\delta_i}\,\mathrm{d}x \qquad (4)$$

とおく。ここで

$$\delta_i = \begin{cases} 1, & x_i \in C_i^1 \\ 0, & x_i \in C_i^2 \end{cases}$$

とする。（x_iはC_i^1かC_i^2のいずれか一方に必ず含まれる。）またα_iは、(4)式を確率密度とするための基準化定数である$(1/2 \le \alpha_i \le 1)$。ここではα_iの近似として、$\alpha_i = 0.5/K(\beta_i)$により計算する。ここで、$K(\cdot)$は標準正規分布の下側確率とし、$\beta_i$は $f(\theta_i^1; x_i)$ と $f(\theta_i^2; x_i)$の分離の程度を示す指標で

$$\beta_i = \sqrt{\frac{\|\mu_i - \mu_2\|^2}{|V_1| + |V_2|}}$$

で示すものとする。この 2 分割モデルにおける BIC を以下により計算する:

$$\mathrm{BIC}' = -2 \log L(\hat{\theta}_i'; x_i \in C_i) + q' \log n_i$$

ここに $\hat{\theta}_i' = [\hat{\theta}_i^1, \hat{\theta}_i^2]$ は、2 つの p 変量正規分布の最尤推定値である;共分散を無視すれば、$q' = 2 \times 2p = 4p$ となる。共分散を無視しなければ、$q' = 2q = p(p+3)$ である。

7. BIC > BIC′ならば、2 分割モデルをより好ましいと判断し、2 分割を継続すべく $C_i \leftarrow C_i^1$ とする。C_i^2 については、p 次元データ、クラスターの重心、対数尤度と BIC を保持し、これらをスタックに積む。手順 4 へ。
8. BIC ≤ BIC'ならば、2 分割しないモデルをより好ましいと判断し、C_i^1 についての 2 分割を停止する。(手順 7 で作成された) スタックからデータを取り出し $C_i \leftarrow C_i^2$ とし、手順 4 へ。スタックが空なら次の手順へ。
9. C_i における 2 分割が全て終了。 手順 4〜8 で作成された 2 分割のクラスターが C_i 内で一意になるようにデータの属するクラスター番号を振りなおす。
10. はじめに k_0 分割したクラスター全てについて 2 分割が終了。全データに対してその属するクラスター番号が一意になるように番号を振りなおす。
11. 全データの属するクラスター番号、および各クラスターの重心、各クラスターに含まれるデータ数を出力する[終了]

モデル選択規準として提案されている多くの情報量規準の中から BIC を用いるのは、(ⅰ) BIC がその導出過程で、指数型分布族の選択を考えていること (正規分布は指数型分布族に含まれる)、(ⅱ) 分布間の距離に基づくのではなく、モデルの事後確率を比較していること、の理由による。これより *x*-means の適用については BIC が最適であると考えた。なお、このプログラムのソースコードは、http://www.rd.dnc.ac.jp/~tunenori/xmeans.html より入手できる。

5 おわりに

現在、われわれは、電子メールやフルテキストによる新聞記事、Web 上で入手されたアンケート、あるいは Web ページそれ自体など、膨大な自然言語で書かれた電子媒体が容易に入手できる環境にある。また様々な検索エンジンも実用に供されている。このような状況を鑑みれば、我々が早急に獲得すべき技術は、大量の自然言語データを(ⅰ)いかによく検索/抽出し、(ⅱ)その検索結果をいかに整理するか、そして(ⅲ)その結果を情報の質という

面からどのように定量的に評価するのか、である。本研究は、このうち(ⅰ)と(ⅱ)を満足すべき一つの試案である。

今後は、(ⅱ)と(ⅲ)が研究の主たるパラダイムになると考えている。たとえば、情報検索では、検索精度は、一般に再現率（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ことが多かったわけだが、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）の問題があり、この問題の解決なくしては正しい評価はできないであろう。また、再現率と適合率の両方を一次元尺度に要約した van Rijsbergen の *F* 尺度や 11 点平均精度が、通常の統計学で用いる 2×2 分割表における評価指標、たとえば四分相関係数やファイ係数とどのような関係にあるかを調査することも興味深いテーマである。

なお、本研究に対しては、平成 12 年度 HITOCC 応募テーマとして多大なるご援助をいただきました。この場をお借りして、関係各位に厚くお礼申し上げます。

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, 560-573.

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

Berry, M.W., Dumais,S.T., and O'Brien, G.W.(1995): Using Linear Algebra for Intelligent Information Retrieval, SIAM Review, Society for Industrial and Applied Mathematics, **37**(4), 573-595.

Berry, M.W.(1992):Large scale singular value computations, *International Journal of Supercomputer Applications*, **6**(1), 13-49.

Cohen, W.W. and Singer,Y.(1996):Context-sensitive learning methods for text categorization, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development on in Information Retrieval (SIGIR'96)*, 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*, **35**, 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, **41**(7), 391-407.

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*, 606-623.

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

石岡 恒憲、亀田 雅之(1999):単語の共起に基づく関連文書検索、算法と検索事例、応用統計学, **28**(2), 107-121.

石岡 恒憲(2000):クラスター数を自動決定する k-means アルゴリズムの拡張について、応用統計学, **29**(3), 141-149.

Jones, K.S. (1972):A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, **28**(1), 11-21.

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*, 27--34, Tokyo.

Kohonen, T. (1996): Self-Organizing Maps, Springer-Verlag Berlin Heidelberg New York.

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

Luhn, H.P.(1957):A Statistical Approach to Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, **1**(4), 307-317.

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

Schwarz, G. (1978): Estimating the Dimension of a Model, *Ann. Statist.* **6**(2), 461-464.

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

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

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)*, 253-277.