知能と複雑系 117－5
(1999. 7. 22)

独立話題分析

－独立性最大化による特徴的話題の抽出－

篠原靖志
(財) 電力中央研究所　情報研究所
〒201 東京都狛江市岩戸北２－１１－１
Tel:03-3480-2111 E-mail: sinohara@criepi.denken.or.jp

あらまし
文書データベースにおける特徴的話題を見つけ出すことは、文書データの整理・検索・要約などにおいて重要な役割を果たす。本稿では、文書データベース中での特徴的話題の組を、その話題を特徴づける単語の生起が互いに独立な話題の組として定義し、その抽出方法として、数量化Ⅲ類型独立成分分析を提案する。さらに、実験により、数量化Ⅲ類の持つ情報圧縮性を組み合わせることで、独立性の高い文書・単語へのグループ化が行えることを示す。

キーワード　独立成分分析、話題抽出、文書検索、数量化Ⅲ類

Independent Topic Analysis

－ Extraction of Characteristic Topics by maximization of Independence

Yasusi Sinohara
Communication and Information Research Laboratory
Central Research Institute of Electric Power Industry
Iwado-Kita 2-11-1, Komae, Tokyo 210 Japan
Tel. 03-3480-2111　E-mail: sinohara@criepi.denken.or.jp

Abstract
Topic plays important role in organizing/retrieving/summarizing documents in a document database. Especially, the topics characterizing groups of documents in the database are useful. We define these characteristic topics as independent topics and propose the mothod called "Dual Scaling Type Indpendent Component Analysis" to find them. We also show the method find the independent groups of documents and words characterized by the found topics combining the reduction of dimensionality by dual scaling.

key words　Indpendent Component Analysis, Topic Extraction, Text Retrieval, Dual Scaling

1 はじめに

文書整理，文書検索，データベースの内容要約など，大量の文書データに対して必要となる各種の処理においては，文書データベースで特徴的な話題を見つけて，あらかじめ，文書をグループ化することが重要な役割を果たす．このような特徴的話題に基づく文書のグループ化は，文書の有効な整理手法の一つであり，また，見つけ出された特徴的話題は，文書データベースの内容要約の役割を果たすであろう．さらに，検索要求と関連の高い特徴的話題を見つけることで，検索キーワードの字面ではなく，意味的な関連を持つ文書の検索を行うことができる．

本稿では，文書データベースでの「特徴的話題群」を，互いに無関係な「独立」な話題の組みとして捉える．そして，データベース中の各文書の内容をこれらの独立した話題の組み合わせによって表現することを試みる．このための諸要請の検討を行い，従来の主成分分析の拡張ではなく，数量化III類の拡張となる数量化III類型独立成分分析を提案する．さらに，表現に使用する話題の数に制限を加えることで，各話題の一般性を向上させることにより，文書のグループ形成がなされることを示す．

2 ベクトルモデルと話題分解

文書検索におけるベクトルモデル [5] では，一般に，文書や検索要求を単語などを基底としたベクトルで表現する．例えば，文書 d_i 中での各単語 w_j の重要度 a_{ij} の並びが，単語を基底とした文書ベクトル $\vec{d_i}$ となる

$\vec{d_i} = (a_{i1}, ..., a_{im}) \cdot (\vec{w_1}, ..., \vec{w_m})^t$

(ただし，X^t は X の転置を指す)．

さらに，各文書 d_i 中での文書データベース中の各単語 w_j の重要度 a_{ij} を並べた，文書数×総単語数の行列 A は，文章データベース中の全文書を，単語を基底として行列表現したものとなる．

$(\vec{d_1}, .., \vec{d_n})^t = A \cdot (\vec{w_1}, ..., \vec{w_m})^t$

文書中の単語の重要度は，用語頻度に比例した非負の値が割り当てられ，文書中で使用されない単語に対しては、重要度 0 を与えることが一般的である．本論文でも、このベクトルモデルによって、話題や文書内容を表現する．

話題 t_i での各単語 w_j の重要度 x_{ij} の並びが，単語を基底とした話題ベクトル $\vec{t_i}$ となる．

$\vec{t_i} = (x_{i1}, ..., x_{im}) \cdot (\vec{w_1}, ..., \vec{w_m})^t$

さらに，各話題 t_i 中での文書データベース中の各単語 w_j の重要度 x_{ij} を並べた，話題数×総単語数の行列 X は，話題群の単語を基底とした行列表現となる．

$(\vec{t_1}, .., \vec{t_s})^t = X \cdot (\vec{w_1}, ..., \vec{w_m})^t$

もう一つの文書の表現として，単語ではなく，話題を基底とした表現も有り得る．すなわち，各文書 d_i における各話題 t_j の重要度 b_{ij} の並びにより表現する．各文書での各話題の重要度 b_{ij} を要素とする行列 B がこの行列表現となる．

$(\vec{d_1}, ..., \vec{d_n})^t = B \cdot (\vec{t_1}, .., \vec{t_s})^t$

上記行列 A，B，X には以下の関係が成立する．

$$A = B \cdot X \tag{1}$$

各文書中の単語重要度の行列 A の文書での話題重要度の行列 B と話題の単語重要度の行列 X の積への分解となっているので、この表現を文書データベースの「話題分解」と呼ぶ．

3 独立話題分析における諸要請

文書データベース中の文書の単語の出現を単語生起と呼ぶ．同一文書中の同一単語でも，異なる位置に出現すれば異なる単語生起である．単語生起の話題 t の重要度は，使用する単語 w により定まり $X(t,w)$ と記す．$\Pr(\exists w, X(t,w) = v) = \text{if } X(t,w) = v, \Pr(w), \text{else } 0$ などが成立する．この時，独立話題分析で要請する諸条件について述べる．

3.1 話題の独立性の最大化

独立話題分析で，最も本質的要請は，話題分解の基底となる話題が独立であるという要請である．話題の独立性は、以下により定義する．

定義 1 話題の組 $t_1, .., t_{n_t}$ が「独立」であるとは，下記を満たすことである．

$$\begin{aligned} &\Pr \quad (\exists w, \wedge_{i=1}^{n_t} X(t_i, w) = b_i) \\ &\quad = \prod_{i=1}^{n_t} \Pr(\exists w_i, X(t_i, w_i) = b_i) \end{aligned} \tag{2}$$

話題の組が独立であるとは、ある単語生起における各話題の重要度が指定された組み合わせである確率が，各話題で指定された重要度を持つ確率の積に等しいことを意味する．従って，各話題の重要度の確率は，他の話題の重要に依存しない．

独立性を持つ話題を得るメリットは，大きく 2 つある．第 1 に，他の話題との関連を考える必要がない自律的で特徴的な話題として，各話題を扱うことが可能となる．これにより，単語や文書の良いグルーピングが得られる可能性がある．

第 2 に，話題分解の式 $A = B \cdot X$ から，文書データベース中での，文書と単語の関連の強さである重要度の

確率分布を，下記により簡便に推測することができる．

$$\begin{aligned}\Pr(\exists w, \wedge_{i=1}^{n_d} A(d_i,w) &= a_i) = \\ \prod_{i=1}^{n_t} \Pr(\exists w, X(t_i,w) &= x_i)/\det(B) \qquad (3)\\ \text{ただし,} [a_1,...,a_{n_d}] = &\ B \cdot [x_1,..,,x_{n_t}]^t\end{aligned}$$

このような文書と単語の重要度の分布の効率的なモデル化は，例えば，文書検索においては、検索精度の向上につながると期待される．

しかし、上記の独立性の条件を完全に満たす話題の組は必ずしも存在しない．そこで、独立性の程度を示す指標の最大化を図る．独立性の指標としては，独立性の定義式 (2) の左辺と右辺の両分布の乖離を示す Kullback 情報量 KL を取り，この最小化を図る．

$$\begin{aligned}KL &\equiv \int \Pr(\exists w, \wedge_{i=1}^{n_t} X(t_i,o) = v_i) \\ &\quad \cdot \log \frac{\Pr(\exists w, \wedge_{i=1}^{n_t} X(t_i,o) = v_i)}{\prod_{i=1}^{n_t} \Pr(X(t_i,o) = v_i)} dv \\ &= H(\exists w, \wedge_{i=1}^{n_t} X(t_i,s) = v_i) \\ &\quad - \sum_{i=1}^{n_t} H(\exists w, X(t_i,s) = v_i) \qquad (4)\\ \text{ただし,} \quad & H(x) = \int \Pr(x)\log(\Pr(x))dv\end{aligned}$$

後述するように正規直交性の要請の下では，話題分解の基底 X は、回転 Rot の自由度を除いて一意に定まる．この時、$\det(Rot) = 1$ であるから上記の第 1 項は，Rot に依存しない定数となる．従って， 式 (4) の最小化は，第 2 項 $\sum H(...)$ を最大とする回転 Rot を求める問題となる．

この考え方は，信号処理分野で提案されている「独立成分分析」[1, 2, 3, 4] と呼ばれる手法と同一である．Comon [2] らは Edgeworth 展開を用いて、第 2 項の要素 $H(y_i = v_i)$ を近似して，以下の最大化を提案している．

$$\sum_{i=1}^{n_t} \{\frac{K_{iii}^2}{2\cdot 3!} + \frac{K_{iiii}^2}{2\cdot 4!} - \frac{7\cdot k_{iii}^4}{2\cdot 4!} - \frac{K_{iii}^2 K_{iiii}}{8}\} \qquad (5)$$

K_{iii} は 3 次のキュムラントを，K_{iiii} は，4 次のキュムラントを指す．式 (5) は， 偏り (K_{iii}) が小さく、分布の裾野が重い (K_{iiii}が大きい) 場合に， $H(y_i = v_i)$ を最大化する．偏りを無視できる場合には、4 次のキュムラントの平方和の最大化を行えば良い．甘利ら [3] は，Gram-Chalier 展開を用いて $H(y_i = v_i)$ を近似し，類似の式を得ている．ただし、甘利らの近似では，$H(y_i = v_i)$ は正規分布に近い分布で最大化される．tanh 型の関数 [1] による場合も甘利らと類似の特性を持つ．このような独立性最大化の目的関数をコントラスト関数と呼ぶが、どのようなコントラスト関数が良いかはデータの分布特性による．

3.2 正規直交性

話題ベクトル の内積、長さ、平均、共分散を下記で定義する．

定義 2 話題ベクトル t_i, t_j の内積

$$< t_i, t_j > \equiv \int b_i b_j \quad \Pr(\quad \exists w, X(t_i,w) = b_i \wedge X(t_j,w) = b_j) db_i db_j$$

定義 3 話題ベクトルの長さ $|t| \equiv \sqrt{< t,t >}$
話題ベクトルの長さを話題の強度とも呼ぶ．

定義 4 話題ベクトルの平均
$\overline{t_i} \equiv \int b \cdot \Pr(\exists w, X(t_i,w) = b) db$

定義 5 話題ベクトルの共分散
$cov(t_i,t_j) \equiv < t_i - \overline{t_i}, t_j - \overline{t_j} >$

定義 6 話題ベクトルの分散
$var(t) \equiv cov(t,t) = < t - \overline{t}, t - \overline{t} >$

各話題が独立であれば無相関，すなわち，内積 $< t_i - \overline{t_i}, t_j - \overline{t_j} > = 0$ である．従って，独立する話題ベクトルは，話題ベクトルの原点ではなく，各話題ベクトルの平均 $(\overline{t_1},...,\overline{t_s})$ で直交する．

前述したように独立性は必ずしも成立しないので、最低限の要請として、中心化した話題 $(t_i - \overline{t_i})$ の直交性を要請する．また、その強度 (長さ $|t_i - \overline{t_i}|$) を規格化するために各話題の分散 $= 1$ を要請する．以上の 2 つの要請により，中心化した話題は正規直交性を満たす．$\Pr(w)$ を対角成分とする対角行列を P_w，単位行列を I_{n_t} とするとこの条件は下記により表される．

$$(X - \overline{X}) \cdot P_w \cdot (X - \overline{X})^t = I_{n_t} \qquad (6)$$

3.3 話題/文書の類似度の定義

ある文書 d が，特定の独立話題成分 t_jのみからなる時，すなわち，$\vec{d} = k \cdot \vec{t_j}$ と表される時，文書 d は，独立話題成分 t_j に属すると呼ぶことにする．ある検索要求に適合する文書を検索する場合，自然な要請として，異なる独立話題成分に属する 2 つの文書は，互いに話題 (内容) が依存しない無関係な文書であるのだから，(独立話題ベクトルの直交点付近にある場合を除いて) 類似の文書とはならないことが望ましい．

文書検索においては，ベクトルで表現される文書や検索要求の類似度をベクトル間の角度 $cos^{-1}(\frac{<t_1,t_2>}{|t_1|\cdot|t_2|})$ により定義する場合が多い．しかし，独立話題成分の直交点

と原点とは一致しないので，ベクトル角による類似性の定義では，異なる独立話題成分に属する文書でも類似文書と見なされてしまう可能性が高い．

直交点を原点にとりなおして，中心化した検索要求や文書ベクトル $(d_i - \overline{d_i})$ に対してベクトル角を求めるならば，上記の問題は解決するように見えるが，これにも問題がある．すなわち，ベクトル角により類似度を定義する場合，検索要求や文書ベクトルの各単語に割り当てられた重要度が一定倍になっても，ベクトル角は不変である．すなわち、この類似度の定義では，単語重要度の比が文書や検索要求の内容を決定する．逆に、このような関係が成立するように、重要度が与えているとも言える．しかし，原点を直交点に変更することにより，変更前の重要度の比が同一であっても類似度が異なるという事態が生じて，当初の重要度の意味が失われてしまう．

そこで，上記の要請を満たすには，話題や文書間の類似度を，ベクトル角以外の形で定義する必要がある．

ベクトル間の距離は原点移動によって不変であるので，本稿では，類似度をベクトル間の距離により定義することとする．そして、この類似度の定義と整合的に重要度を与える手法として数量化 III 類型独立成分分析を提案する．

4 独立話題分析の定式化–数量化 III 類型独立成分分析

文書/話題/検索要求の類似度をベクトル間の距離により定義する数量化 III 類型の独立話題分析を提案する．

n_oを文書データベース中の延べ単語数，Fを文書 d 中の単語 wの出現確率 (頻度/n_o) を要素とする文書数×単語数の行列，P_d を文書中の単語生起数/n_oを対角成分とする対角行列，P_w を文書データベース中の単語の出現確率 (頻度/n_o) を対角成分とする対角行列とする．

今，n_t個の話題に対応する n_t次元の空間を考える．そして，単語 w に，各話題 t_iの重要度 $X(t_i, w)$ を並べた n_t次元の座標 $(X(t_1, w), ..., X(t_{n_t}, w))$ を，文書 d に，各話題 t_iでの重要度 $Y(t_i, d)$ を並べた n_t次元の座標 $(Y(t_i, d), ..., Y(t_{n_t}, d))$ を割り当てる．ただし，単語の重要度 X は、各話題について正規直交化させる (式 (7))．

各単語生起が生じる文書の座標と，使用単語の座標の距離の二乗平均は，文書配置と単語配置の誤差と解釈できる．そこで、話題についての正規直交化条件

$$X \cdot P_w \cdot X^t = I_{n_t} \tag{7}$$

の下で，全単語生起での平均二乗誤差

$$\begin{aligned} \mathrm{mse} &= \mathrm{trace}(Y \cdot P_d \cdot Y^t) + \mathrm{trace}(X \cdot P_w \cdot X^t) \\ &- 2\mathrm{trace}(Y^t \cdot F \cdot X) \end{aligned} \tag{8}$$

Step1. (数量化 III 類) $A = P_d^{-1} \cdot F \cdot P_w^{-1}$ の
一般化特異値分解 U，S，V を求める．ただし，
$U \cdot P_d \cdot U^t = V \cdot P_w \cdot V^t = I_{n_t}$
かつ，第 1 特異成分は捨てる．

Step2. (初期化) 対称行列 Rot_0 を任意に定める．

Step3. (回転行列化)
$Rot_i = (Rot_i \cdot Rot_i^t)^{-1/2} \cdot Rot_i$

Step4. (終了判定) 回転行列 Rot_{i-1} と回転行列 Rot_i の間の角度が一定値より小さい場合に Step7. (終了) へ．

Step5. (話題成分) $X = Rot \cdot V^t$

Step6. (回転行列の更新)
$Rot = (X.^3 \cdot P_w \cdot V^t) - 3 \cdot Rot.$
Step 3 (回転行列化) へ．

Step7. (終了) $Y = Rot \cdot U^t$

図 1: 数量化 III 類型独立成分分析アルゴリズム

が最小の配置の中で，話題の独立性が最大となる話題の組みを求めるのが，数量化 III 類型独立話題分析である．

式 (7) と (8) から，X，Yは，回転 Rot の自由度を除いて一意に決定される．

$$\begin{aligned} X &= Rot \cdot V^t \\ Y &= X \cdot F^t \cdot P_d^{-1} = Rot \cdot U^t \\ \text{ただし，} & \quad P_d^{1/2} U, S, P_w^{1/2} V \text{は、} \\ & \quad H \text{の階数 } n_t \text{の特異値分解とする．} \\ H &= P_d^{1/2} \cdot A \cdot P_w^{1/2} \\ A &= P_d^{-1} \cdot F \cdot P_w^{-1} - 1_{n_d}^t \cdot 1_{n_w} / n_o \\ Rot &\text{ は 任意の回転行列 } (Rot^t \cdot Rot = I_{n_t}) \end{aligned}$$

回転行列 Rot が単位行列 I_{n_t} の時は，X，Yは数量化 III 類の結果に一致する．座標系を回転しても各点の距離が不変であり，平均二乗誤差 $\mathrm{mse} = n_t - \mathrm{trace}(S) - 1$ となる．

Aでは，自明な特異値 1 に対する特異値ベクトル 1_{n_d}，1_{n_w}を除いている．これにより，$\overline{X} = 0$ となり，直交点と原点が一致する．従って、正規直交化条件 (7) は式 (6) に等しい．上記の解を，話題分解 $A = B \cdot X$としてみるならば、$B = Y^t \cdot (Rot \cdot S \cdot Rot^t)$ となっている (一般の

独立成分分析では，$B = U \cdot S \cdot Rot$ となる)．

話題の独立性を最大とする、すなわち、式 (4) を最小とする回転 Rot を決定するには、コントラスト関数の最大化を図れば良い．ただし，コントラスト関数の中で使用されるキュムラントの定義などは、計量行列 P_w による加重を持つ．例えば、4 次のキュムラント K_{iiii}は下記となる．

$$\sum_{\mathrm{w}} \mathrm{X(t,w)}^4 \cdot \mathrm{P_w(w,w)} - 3(\sum_{\mathrm{w}} \mathrm{X(t,w)}^2 \cdot \mathrm{P_w(w,w)})^2$$

このような拡張をすることで、独立成分分析で使用されるほとんどのアルゴリズムが適用可能である．4 次キュムラントの平方和を最大化する場合は，JADE [4] など高速なアルゴリズムが存在する．より広いクラスに適用可能で高速なアルゴリズムとしては，FPICA [1] が，さらに一般的アルゴリズムとしては、自然勾配法によるアルゴリズム [3] などがある．一例として，4 次キュムラントの平方和の最大化を行う場合の FPICA アルゴリズムを拡張したものを使用した数量化 III 類型独立話題分析アルゴリズムを図 1に示す．

なお、数量化 III 類型独立話題分析結果に基づく検索では，指定された検索要求ベクトルを分散 = 1 に正規化した後，そのベクトルから一定距離内にある文書ベクトルを返せば良い．

5 評価実験

5.1 実験目的

本章では，提案した数量化 III 類型独立話題分析について，実験 I では，数量化 III 類との特性の比較を行う．実験 II では，数量化 III 類に基づいて情報圧縮した空間で数量化 III 類型独立成分分析を行い，独立成分分析の持つ特徴的話題の抽出能力，空間特性の発見能力について調べる．なお，独立成分分析でのコントラスト関数としては、tanh 型などと実験による対比の結果、式 (4) に対してより小さい値を与えることから，4 次のキュムラントの 2 乗和を使用し，図 1に示したアルゴリズムを用いた．

5.2 実験対象データ

毎日新聞 93 年東京版朝刊 1 月分の新聞記事 4426 本を対象とした．以下の手順で，分析を行うための出現確率行列 F を作成した．

1. 形態素解析ソフト「すもも」を使用して各記事の形態素解析を行い，数字，数詞，年月日などを除く名詞を抜き出す．

2. 抜き出した名詞の内，23 本〜3000 本の記事に出現する 2055 語を抽出する．

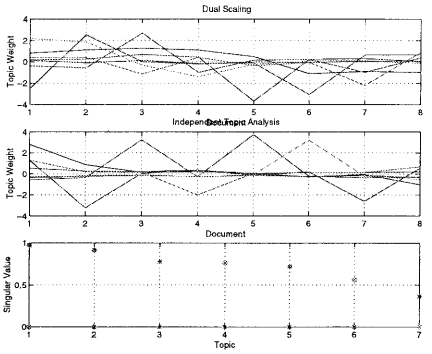

図 2: 数量化 III 類と独立話題分析の比較

3. 文書×単語の出現確率行列 F を作成

5.3 実験 I:数量化 III 類と独立話題分析の特性比較

実験 I では，数量化 III 類と独立話題分析の特性の違いを調べる．この比較のために，表 1 に示すカンボジア和平に関する記事 5 本と，皇太子御婚約に関する記事 3 本を抽出した．上記 F からこの 8 本についての部分行列を正規化し，実験用の新たな F とした．2 本以上の記事に出現しない単語は除いている．この新たな F に対して，数量化 III 類と独立話題分析を行った．

図 2に，両者の分析結果を示す．図 2 上段が数量化 III 類による結果を，図 2 中段が独立話題分析による結果を示す．各図は，縦軸を話題における文書の規格化された重要度に，横軸を文書 1 〜 8 に対応させ，一つの話題の各文書での重要度を線で結んでいる．図 2 下段は各話題の強さ (特異値の大きさ) を示す．

文書と話題の関係について着目すると，数量化 III 類では，その第 1 話題成分が正であることが，カンボジア和平と，負であることが皇太子御婚約と対応していることが分かる．

すなわち，数量化 III 類は，データベース中の文書に共通する話題を抽出している．これは，行列 X の階数 r の特異値分解 U，S，V が，指定した階数 r の行列の内で最良の近似 ($X - U \cdot S \cdot V'$ のフロベニウスノルム最小) となるものを与えるという情報圧縮性を持つためである．

一方，独立話題分析では，各独立話題成分が，ほぼ 1 対 1 で各文書に対応することが分かる．各記事は，他の記事にはあまり依存せずに書かれたものであり，これらが独立した話題成分として抽出されるのは当然といえる．この時，各話題/記事で高い重要度を与えられた単語は，その記事で特徴的に現れる単語群である．このことから，数量化 III 類が，記事間の共通性の高い話題を抽出した

表 1: 対象記事リスト

1	93/01/01	ポル・ポト派、ＵＮＴＡＣ陣地を砲撃
2	93/01/09	ＵＮＴＡＣの選挙監視員ら５人をポル・ポト派が一時拘束－－カンボジア
3	93/01/17	カンボジア大統領選の実施時期、「ポル・ポト派が受け入れ」－－タイ外相
4	93/01/19	ポル・ポト派の選挙参加、１０日以内に決断求める－－ＵＮＴＡＣスポークスマン言明
5	93/01/27	「公平」条件に大統領選挙参加も－－ポル・ポト派が声明
6	93/01/20	［特集］皇太子さま、ご婚約　皇室会議は７回目
7	93/01/20	［特集］皇太子さま、ご婚約　自然体でゆったりと　寛仁さまのメッセージ
8	93/01/20	［特集］皇太子さま、ご婚約　３５年前も華やかに

のに比べて，独立話題分析は，独立性の高い特徴的話題を抽出する能力に優れていることがわかる．

なお，文書数が 8 本であるのに対して，上記で抽出されている話題の本数は，7 本である．これは，自明な特異値 1 に対する話題を除いているためである．この結果，残る 1 個の隠れた話題は，原点の周りに集中している．独立話題分析では記事 8 が，どの独立話題成分も持たない記事として特徴づけられている．

5.4 実験 II:情報圧縮した空間でのグループの発見

前節では，数量化 III 類と独立話題分析の特性の違いを見た．そこで、両者の特性を活かした使い方として，数量化 III 類により 文書データベースにおける共通性の高い n 個の話題 (値の大きい特異値 n 個に対応する) を抽出して得られた文書×単語の空間に対して，独立成分分析を行い、その空間で特徴的な話題を分析することが考えられる．

そこで，以下の 2 ステップの実験を行った．

1. 情報圧縮: 4426 本 × 2055 単語から，数量化 III 類により，特異値が大きいほうから 250 成分を話題として取り出す．

2. 独立話題成分: この 250 成分について，独立話題分析を行う．

図 3に，数量化 III 類の結果得られた話題成分の内，第 3 成分の単語の重要度 $X(3,w)$ を横軸に，第 4 成分の単語の重要度 $X(4,w)$ を縦軸に各単語をプロットしたものを示す．これを見ると，主に 3 つの直線成分が混ざったものに見受けられる．これらの 3 つの直線は，独立話題分析の結果得られた第 106 独立話題成分，第 154 独立話題成分，第 172 独立話題成分に対応している．第 106 独立話題成分は，「全豪」「オープン」「テニス」の順に強い重要度を与えられている．第 154 独立話題成分は，「名義」「実勢価格」「宅地」「預貯金」の順に強い重要度を与えられている．第 172 独立話題成分は，「不全」「呼吸」「告別式」「死去」の順に強い重要度が与えられている．図 4に，独立成分分析の結果得られた第 106 成分と，第 154 独立成分での単語の重要度を横軸，縦軸に取り，各単語をプロットした結果を示す．この結果，各独立成分に対応する軸上に単語が直線状に並び，さらに，比較的少数の単語が強い重要度を与えられ，その単語は他の独立話題成分では，大きな値を持たないことが分かる．他の独立成分についても同様に，各単語が各軸上か，原点周辺に並び，軸から離れた点には，ほとんど単語がプロットされない．

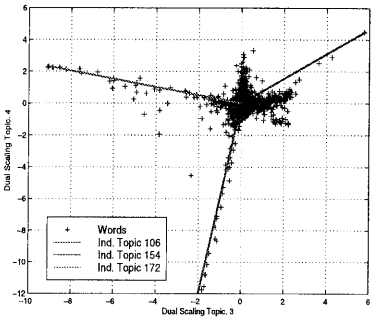

図 3: 数量化 III 類の単語重要度のプロット

さらに，図 5に数量化 III 類の結果得られた話題成分，第 3 成分の文書の重要度 $Y(3,d)$ を横軸に，第 4 成分の文書の重要度 $Y(4,d)$ を縦軸に取って，各文書をプロットしたものを示す．図 6に，独立成分分析の結果得られた第 106 成分と，第 154 独立成分での文書の重要度を横軸，縦軸に取り，各文書をプロットした結果を示す．これらの文書のプロットは，単語のプロットと同様なものとなっている．各文書は，出現する単語との平均距離が最小となるように配置したので，自然な結果である．独立話題分析の結果では，各独立話題成分に対応する軸上に文書が集中していることから，文書のグループ分けがなされていることがわかる．これは，共通性の抽出による情報圧縮で (陰で) 行われたグループ化であり，その構造が，独立成分分析によって明らかになった．

以上の実験から，独立話題分析は，主成分分析/数量

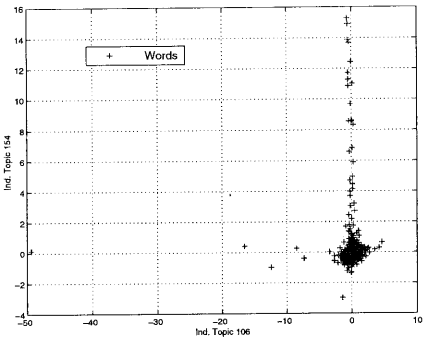

図 4: 独立話題分析の単語重要度のプロット

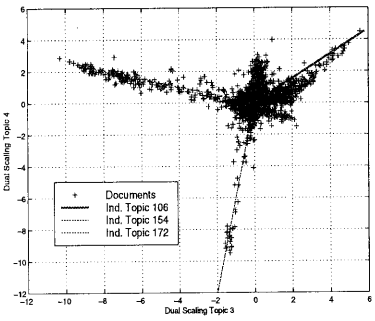

図 5: 数量化 III 類の文書重要度のプロット

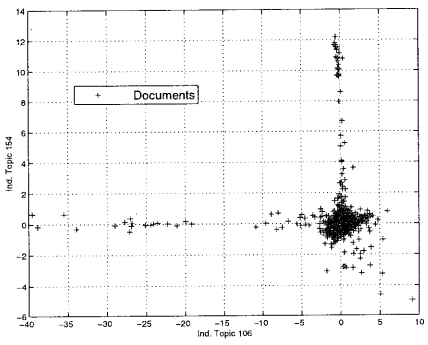

図 6: 独立話題分析の文書重要度のプロット

化 III 類によって，情報圧縮された話題空間のグループ構造を明確にすることができることが示された．さらに，得られた各独立話題成分は，比較的少数の単語により特徴づけるものであった．このような空間構造は文書検索の効率化に利用可能であろう．また，各文書における単位長当たりの強度が強い話題成分に対して，その話題を特徴づける単語を多く含む文を抽出することで，文書の特徴的内容の要約などを行うことも可能となろう．

6 関連研究

特徴的話題を特徴づける単語の抽出で，関連する手法としては，篠原による ExtractRequest [9, 10] や福原らによる研究 [11] がある．ExtractRequest では，利用者が関連があると選択した文書群に共通に現れる単語を，それが現れる選択文書の組み合わせを縦軸に，データベース中の頻度を横軸にとって表示することで，データベース中の頻度が低く，関連文書に共通に現れる特徴的な用語を視覚的に利用者に選択させる．福原らは，クラスタリングによりグループ化した文書群の中で使用される単語の頻度分布を調べ，その単語の歪度 (平均 0，分散 1 の時の 3 次のキュムラント) と尖度 (平均 0，分散 1 の時の 4 次のキュムラント) の和 (cf. 式 (5)) を単語の識別力 G 値として，それを最小とする単語を各グループに特徴的単語として自動的に選択している．福原らの G 値は，歪度が大きく，尖度も大きい頻度分布を持つ単語，すなわち，少数の文書にしか出現しない単語で大きくなる．独立話題分析では，中心化された比較的対称性の高い単語の重要度の分布を考えるので，少数の文書にしか出現しない単語の条件は，尖度の平方和の最大化となり、G 値と類似の効果を持つ上に，特徴的話題 = 独立性の高い話題という明確な意味付けを与えることができる．

独立成分分析の応用例としては，左右 2 チャンネルのステレオ録音から，2 つの独立した音源の自動的分離 [12] や脳磁計の計測結果からの地磁気や家庭用電力 (50, 60Hz) の分離による雑音除去 [13]，画像の雑音除去などがある．これらは主成分分析の拡張となっているが、本稿での手法は数量化 III 類の拡張となっている．

7 まとめと今後の課題

著者は，文書内容の独立した話題への話題分解は，大規模データベースを対象とした文書整理，文書検索，文書要約などにおいて，効果的な表現と考える．本論文では，独立話題分析の考え方を示し、その特性を明らかにすることに重点を置いた．

今後の課題としては、文書要約や文書検索などへの適用効果の実証がある．

参考文献

[1] Hyvarinen, A.: Independent Component Analysis by Minimization of Mutual Information, No. Report A46, Department of Computer Science and Engineering, Helsinki University of Technology (1997).

[2] Comon, P.: Independent component analysis, a newconcept ?, *Signal Processing*, Elsevier, Vol. 36, No. 3, pp. 287-314 (1994)

[3] Amari, S., Cichocki, A. and Yang, H.H.: *A new learning algorithm for blind separation*, Advances in Neural Information Processing Systems 8, MIT Press, pp. 752–763, Cambridge MA (1996).

[4] Cardoso,J.F., and Souloumiac, A.: Blind beamforming for non Gaussian signals, *IEE Proceedings-F*, Vol.140, No.6, pp.362-370 (1993).

[5] Salton G. and McGill, M.: *Introduction to modern information retrieval*, McGraw-Hill, New York (1983)

[6] Deerwester,S., Dumais,S.T., Landauer,T.K., Furnas,G.W. and Harshman R.A.: *Indexing by latent semantic analysis*, Journal of the Society for Information Science, Vol. 41, No. 6, pp. 391–407 (1990)

[7] Dumais,S.T.: Latent Semantic Indexing (LSI) – TREC-3 Report, *Proc. of TREC-3*, NIST (1996).

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

[9] 篠原靖志: 文書検索システム ExtractRequest における用語分析マップによるフィードバックの評価, 情報処理学会研究報告 Vol. 98,No. 34(DBS-115 FI-49) (1998)

[10] 篠原靖志: ExtractRequest—利用者への情報開示に基づく検索要求抽出, 情報処理学会研究報告 Vol. 96,No. 21(HI-65 SLP-10) (1996)

[11] 福原友宏, 武田英明, 西田豊明: 統計情報と概念知識を用いたテキスト間の話題特定, 情報処理学会 知能と複雑系 研究報告, No. 115-1, pp. 1–8 (1999).

[12] 奥乃 博, 池田 思朗: 「Blind Source Separation による 2 話者同時発話認識」, 人工知能学会研究会資料,AI チャレンジ研究会 (第 1 回), pp. 1–6, (1998)

[13] 池田 思朗, 村田 昇: 「Independent Component Analysis を用いた MEG データの解析」 電子情報通信学会技術研究報告, NC98-28, pp. 29–36 (1998)