潜在意味分析の原理と数理

―女児向けコミック雑誌の意味構造の変遷を題材として―

フェリス女学院大学文学部
高田明典

1. はじめに

おおむね 9 歳から 14 歳程度の女子児童・女子生徒における価値観の形成に関して，少女向けコミック・少女漫画は少なからぬ役割を果たしていると考えられる．筆者はこれまで，物語構造分析の手法を用いてアニメ・小説・CM・コミック・テレビゲームなどの訴求構造分析を行ってきたが，これまでの分析手法のもとでは，多数の物語に共有されている価値観を包括的にとらえることは，不可能とは言えないまでも容易ではない．また，大量のテキストから物語構造を抽出するという作業は，テキストが数十万文字となるような場合には現実的ではない．本研究においては，テキスト分析の手法の一つである潜在意味分析を用いることによって，包括的な「意味構造」の抽出を通して，そこに表現されている価値観の推定を試みた．分析対象としては，現在発刊されている主要な女児向けコミック雑誌とし，その文字表現にのみ着目した分析を行った．また、潜在意味分析という手法をテキスト分析に適用する場合、その理論的背景の理解が欠かせない。本論においては、主成分分析との比較を通して、潜在意味分析において前提とされている原理と数理に関して説明しつつ、その手法の利点および問題点を探る。

2. 分析方法

2．1. 潜在意味分析

分析にあたっては，主として潜在意味分析を用いる．この手法においては，物語の流れを捨象して単語の構成のみを分析対象にすることから，いくつかの利点が生じる．その一つには，今回対象としたコミック雑誌のように，（連載途中のため）物語として完結していないものでも分析対象とすることができることがあげられる．また，物語の話素の抽出に際しては，比較的よく訓練を受けた分析者が必要であるが，潜在意味分析ではそのような作業が必要なく，単にテキストを入力し，必要な前処理を施すという作業のみで分析にかけることができるという簡便性も特長の一つとしてあげることができる．また，分析手法に際しての客観性をあまりに求めることは，必ずしも当該分野の研究の質の向上にはつながるものではないと考えるが，物語構造分析において指摘されがちな分析手法の客観性に関しても，ある程度担保されると考えられる．

2．2. 潜在意味分析の原理

潜在意味分析は，対象とする文の単位（文オブジェクト）にどのような単語が含まれているかを示す「文オブジェクト－単語共起行列」に対して特異値分解を施し，次元縮約を行うことによってテキストの主たる意味構造を把握する主成分分析的手法である．

今、いくつかの文があったとする。それぞれの文は、何らかの「意味」を持っているはずである。すると、文１の意味をs_1、文２の意味をs_2というように、s_iなる文の「意味」を想定することができる。

今、文１が四つの単語w_1～w_4から成っていると考えると、

$$s_1 : w_1, w_2, w_3, w_4$$

と表現できる。これは、文１の意味（文１を発した人が想定していた意味）から、その 4 つの単語が発生したということを示している。また、s_2（文２の意味）に対応する単語は、文１とは異なり、

$$s_2 : w_1, w_5, w_6, w_7$$

などとなっているとする。このとき、s_2 に何らかの処理が施されて単語 $w_1 \sim w_4$ が発生したと考えると、

$$[w_1, w_5, w_6, w_7] = g(s_2)$$

(式1)

なる関数 $g()$ を想定することができる。この関数は、人間の言語処理の仕組みを示している。たとえば、s_2 がある欲求や意図の表現であったとき、

$$[\text{私}, \text{おいしい}, \text{柿}, \text{食べたい}] = g(s_2)$$

(式2)

のように、何らかの処理 $g()$ によって、欲求や意図である s_2 が 4 つの単語を発生させたと考える。もちろん本来は、語順や助詞・助動詞なども発生させられているが、ここでは捨象して説明している。

さらに、複数の文（もしくは文書）の存在を考える。たとえば、「おいしい」という単語が文 1、文 2、文 4、文 9 で使用されているという例である。

$$[\cdots, \text{おいしい}, \cdots] = g(s_1)$$
$$[\text{私}, \text{おいしい}, \text{柿}, \text{食べたい}] = g(s_2)$$
$$[\cdots, \text{おいしい}, \cdots] = g(s_4)$$
$$[\cdots, \text{おいしい}, \cdots] = g(s_9)$$

(式3)

「おいしい」という単語以外の並びは、それぞれの文において様々であることが想定されるが、それらの文の意図・意味を表現したもののうちの一つの要素として「おいしい」という単語を位置づけることができる。今、単語「おいしい」を w_5 と置き、その単語が表現されるための元となる意味・意図を m_5 と置くと、この状況を、新たに導入する関数 $f^{\circ}()$、$h^{\circ}()$ によって以下のように示すことができる。

$$m_5 = f^{\circ}(s_1, s_2, s_4, s_9)$$

(式4)

$$w_5 = h^{\circ}(m_5)$$

(式5)

	文1	文2	文3	…
単語1	1	0	1	…
単語2	0	1	0	…
単語3	0	1	1	…
⋮	⋮	⋮	⋮	

表 1

つまり、ある単語の意味・意図を、その単語表現単体で発生させられているものであると考えるのではなく、ある文の意味・意図から、単語を発生させる上で必要とされる意味・意図が発生させられ、その意味・意図の表現として、ある単語表記が行われていると考える。

ここで問題となるのは、(式 4)である。文 i を O_i と置くと、その意味・意図との関係は、前述のように、以下のごとく表現される。

$$O_i = g(s_i)$$

(式6)

観察しうるのは、O_i のみであるので、ここで $g()$ の逆関数 $g^{-1}()$ を考える。つまり、

$$s_i = g^{-1}(O_i)$$

(式7)

となる。したがって、(式 4)は、

$$m_5 = f^{\circ}(g^{-1}(O_1), g^{-1}(O_2), g^{-1}(O_4), g^{-1}(O_9),)$$

(式8)

と表現される。つまり、

$$m_j = f^{\circ}(g^{-1}(O_i)) \quad (i = 1,2,\cdots n), (j = 1,2,\cdots p)$$

(式9)

となるが、ここで、$f^{\circ}(g^{-1}(\ ))$ なる関数を新たに $f()$ と置き、

$$m_j = f(O_i) \quad (i = 1,2,\cdots n), (j = 1,2,\cdots p)$$

(式10)

とする。ここで、この関数 $f()$ を明らかにするこ

とが果たして可能であるかという問題に逢着する。

O_i において単語 w_1 が使われていたら値として 1 を与え、使われていない場合は 0 を与えるとすると、上記における m_1 を次のように考えることができる。

$$m_1 = f(O_1 = 1, O_2 = 0, \cdots, O_8 = 1, \cdots O_{30} = 0)$$

(式11)

これは、表 1 に示したように、ある語が使用されていればその語を示す位置に 1 を置き、使用されていなければ 0 を置くということに換言できる。

単語は、行列表示されることとなり、たとえば前述の例の場合、

$$(1,0,0,0,1,1,1,0,0,\cdots,0)$$

(式12)

となる。このデータ形式が、潜在意味分析において用いられる共起行列である。

関数 $f()$ に求められる性質は、$O_1 \sim O_n$ までの n 個の文に基づいて得られた結果によって、$w_1 \sim w_p$ までの p 個の単語を、「最もよく識別できること」であると考える。つまり、関数 $f()$ の本来的な内容や処理のアルゴリズムを考えるのではなく、その関数に求められる性質を考えるということに等しい。この考え方は、潜在意味分析だけではなく、主成分分析や因子分析、もしくは双対尺度法と呼ばれる対応分析・数量化Ⅲ類列の分析においても同様に設定される基本的な立脚点であるが、ここにも論理的な置き換えがあるので注意が必要である。ただし、これは分析の動機を考えれば当然のことである。$s_1 \sim s_n$ および $m_1 \sim m_p$ なる「意味」を知ることが分析の動機であることから、「最もよく識別ができる」「最もはっきりとした線引きができる」ことを中心に据えるということである。

2．3．主成分分析の数理

ここでさらに、上記 2.1. で述べた原理がどのように具現化されているかについての直観的な理解を得るため、まず主成分分析の数理に関して説明する。後に述べるが、潜在意味分析で用いられる特異値分解は、主成分分析での固有値問題と密接な関連を有している。

前項での「最もよく識別できる」とは、統計的な意味においては「分散を大きくする」という処理に置換することができる。端的に言うならば、分解能が高い関数を想定することである。図 1 に示したのは、3 つのデータ点(1, 1)，(1. 5, 2)、(2, 3)を、上から（垂直方向から：①）見た場合と、横から（水平方向から：②）それぞれ見た場合の、識別力の違いについてである。①の場合、データ点の間隔は 0. 5 となっており、②の場合は 1. 0 となっている。つまりこの場合は、②の方向から見たほうが、「データの見晴らしが良い」、つまり、分解能が高いということになる。

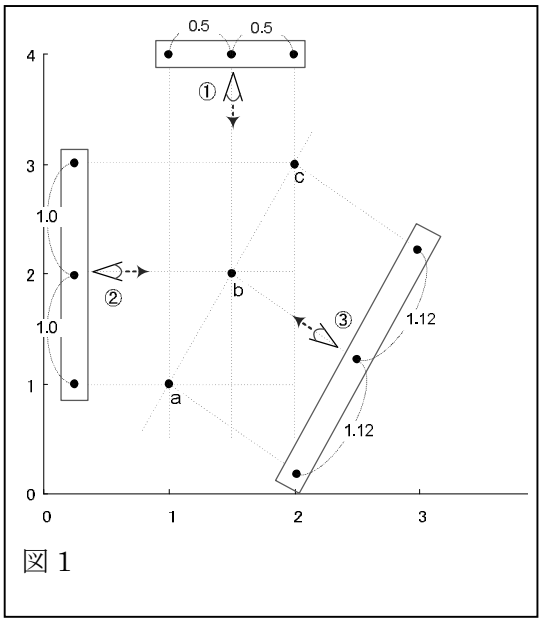

図 1

図 1 に示したように、この例では、③の方向からデータを見ることが、最も分解能が高い状態となる。

ここでは単に「～の方向から見る」という表現をしているが、これは、データ分布の重心を中心として回転させることによって実現される。図 2 と図 3 にその描像を示した。図 2 は、各データ点の座標値から x の平均と y の平均をそれぞれ減じることに

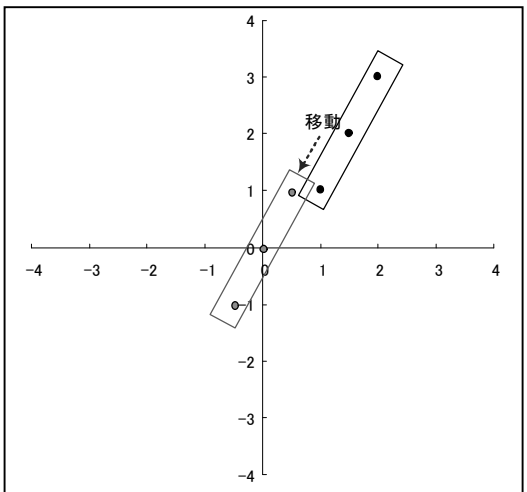

図 2

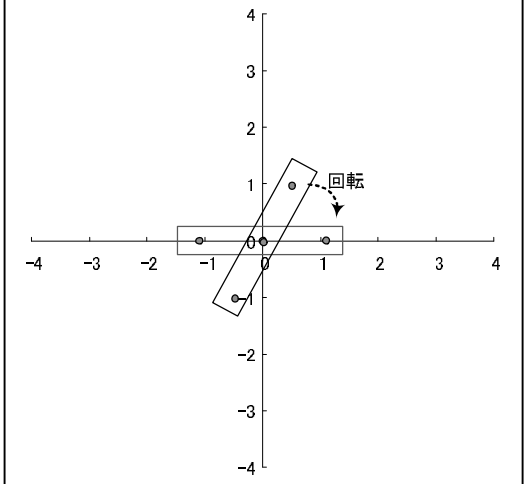

図 3

よって実現される。この処理によって、データ分布の重心が原点となる。図 3 は、データ分布の中心とされた原点まわりに、x 軸方向の分散が最大となるように回転させている状態の模式図である。図 3 のように回転させ、x 軸方向の分散を計算することが、図 1 において「③の方向から見る」ということの具体的な実現方法である。

今、2 次元平面の場合、移動前の座標を(x, y)、移動後の座標を(x', y')と置くと、角度θによる座標の回転移動は、

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

(式13)

で示される。説明の単純化のため、

$$\cos\theta = a_1$$
$$-\sin\theta = a_2$$

(式14)

と置く。また、重心の原点への移動はすでに行われているものとする。すると、

$$\begin{cases} x' = a_1 x + a_2 y \\ y' = -a_2 x + a_1 y \end{cases}$$

(式15)

となるが、今、第 1 主成分としてのx軸方向の分散だけが問題となるので、y'については無視できる。

ちなみに、(式 15)を合成変量の算出であると見ることもできる。そのとき、個々の回転量が合成変量における各項の結合係数となる。

すでに重心の原点移動が済んでいると仮定しているので、i番目のデータの原点移動前のx値をx_{ii}°、平均値を$\bar{x}$と置き、また同じくyについて原点移動前のy値をy_{ii}°、平均値を$\bar{y}$と置くと、

$$x = x_i^{\circ} - \bar{x}$$
$$y = y_i^{\circ} - \bar{y}$$

(式16)

である。

xの分散s_x^2は、データの偏差平方和をデータ数nで除したものであるので、以下の式で示される（ここでは標本分散を例にするが、不偏分散を考える場合には$n-1$で除す。特に、潜在意味分析を想定する場合、n は圧倒的に大きな数となるので、これは詳細に考慮すべき問題ではないといえる。ただし、潜在意味分析においては、不偏分散が用いられる。ここでは式を単純にするため、標本分散で説明する）。

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{\circ} - \bar{x}\right)^2$$

(式17)

同じく、yの分散s_y^2は

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\circ} - \bar{y}\right)^2$$

(式18)

である。

今、x' の分散は、

$$s_{x'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i'^{\circ} - \bar{x}'\right)^2$$

(式19)

であり、本来

$$s_{x'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(a_1\left(x_i^{\circ} - \bar{x}\right) + a_2\left(y_i^{\circ} - \bar{y}\right)\right)^2$$

(式20)

として計算されるが、ここでは重心の原点移動が済んでいると仮定しているので、(式 15)により、

$$s_{x'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(a_1 x_i + a_2 y_i\right)^2$$

(式21)

と単純化することができる。これを計算し、

$$s_{x'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(a_1^2 x_i^2 + 2a_1 a_2 x_i y_i + a_2^2 y_i^2\right)$$

$$= a_1^2\frac{1}{n}\sum x_i^2 + 2a_1 a_2\frac{1}{n}\sum x_i y_i + a_2^2\frac{1}{n}\sum y_i^2$$

(式22)

が得られる。これに、(式 17)、(式 18)、(式 16)を代入すると、

$$s_{x'}^2 = a_1^2 s_x^2 + 2a_1 a_2 s_{xy} + a_2^2 s_y^2$$

(式23)

となる。ここで、s_{xy} は、x と y の共分散である。(式 14)から、

$$a_1^2 + a_2^2 = 1$$

(式24)

であるので、問題は、この束縛条件下において分散 $s_{x'}^2$ を最大化するということになる。これはラグランジュの未定乗数法によって解けることが知られている。すなわち、

$$L = a_1^2 s_x^2 + 2a_1 a_2 s_{xy} + a_2^2 s_y^2 - \lambda(a_1^2 + a_2^2 - 1)$$

(式25)

を a、b で偏微分した二つの式の値がともに 0 になるとき、最大値が得られる。つまり、

$$\frac{\partial L}{\partial a_1} = 2a_1 s_x^2 + 2a_2 s_{xy} + 2\lambda a_1 = 0$$

$$\frac{\partial L}{\partial a_2} = 2a_1 s_{xy} + 2a_2 s_y^2 + 2\lambda a_2 = 0$$

(式26)

となる。この連立方程式を解くことが、想定している分散の最大値を得ることに等しい。ここで、(式 25)を眺めると、この連立方程式は、行列を用いて以下のように表せることがわかる。

$$2\begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + 2\lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$

(式27)

両辺を 2 で除し、移項すると、

$$\begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

(式28)

が得られる。この(式 28)は、左辺の左側の行列が共分散行列となっている。すなわち、共分散行列の固有値問題を解くことが、回転後の主軸方向の分散の最大値を得ることと相同な処理であることがわかる。またそのとき、得られた固有ベクトルが、主成分分析における結合係数となっている。

2．4．特異値分解と固有値問題

今、2 次元のデータ列 A を、

$$A = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_i & y_i \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$$

(式29)

とおくと、共分散行列 coV とは、

$$coV = \frac{1}{n}A^T A$$

(式30)

で示される。ここで A^T は、A の転置行列である。つまり、

$$\frac{1}{n}\begin{pmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_n \\ y_1 & y_2 & \cdots & y_i & \cdots & y_n \end{pmatrix}\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_i & y_i \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$$

$$=\frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x\ y \\ \sum_{i=1}^{n} x\ y & \sum_{i=1}^{n} y_i^2 \end{pmatrix}$$

(式31)

ここで、(式 16)の仮定を使い、データ列 A がすでに、平均偏差となっているとすると、

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x\ y \\ \sum_{i=1}^{n} x\ y & \sum_{i=1}^{n} y_i^2 \end{pmatrix}$$

$$=\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\left(x_i^\circ-\bar{x}\right)^2 & \frac{1}{n}\sum_{i=1}^{n}\left(x_i^\circ-\bar{x}\right)\left(y_i^\circ-\bar{y}\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left(x_i^\circ-\bar{x}\right)\left(y_i^\circ-\bar{y}\right) & \frac{1}{n}\sum_{i=1}^{n}\left(y_i^\circ-\bar{y}\right)^2 \end{pmatrix}$$

$$=\begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

(式32)

となる。前項において、データの分布の 1 次元的な分散を最大にするためにデータ列を回転させるという処理が、共分散行列の固有値問題を解くことと同じであることを見たが、2 次元データの場合、元のデータ列 A を使って表現すると以下のように表せることがわかる。

$$\frac{1}{n}A^T A\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}=\lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

(式33)

両辺に n をかけて、$n\lambda=\lambda'$ と置くと、

$$A^T A\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}=\lambda'\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

(式34)

ここで、2 次元データの場合、2 組の固有値と固有ベクトルが求められる。そのうち最大の固有値を λ_1' とおき、もう一方を λ_2' とおく。

また、最大の固有値に対応する固有ベクトルを α_1 と置き、もう一つを α_2 と置き、さらにそれぞれに対して結合係数（固有ベクトルの要素）の添え字を振りなおして、以下のように表す。

$$\alpha_1=\begin{pmatrix} a_{1,1} \\ a_{1,2} \end{pmatrix}$$

$$\alpha_2=\begin{pmatrix} a_{2,1} \\ a_{2,2} \end{pmatrix}$$

(式35)

すると、(式 34)は以下のように置き換えられる。

$$A^T A\alpha_1=\lambda_1'\alpha_1$$

(式36)

ここで、

$$D=\begin{pmatrix} \sqrt{\lambda_1'} & 0 \\ 0 & \sqrt{\lambda_2'} \end{pmatrix}$$

(式37)

なる対角行列をおき、また、

$$U=\begin{pmatrix} \frac{1}{\sqrt{\lambda_1'}}A\alpha_1 & \frac{1}{\sqrt{\lambda_2'}}A\alpha_2 \end{pmatrix}$$

(式38)

とおくと、

$$UD=\begin{pmatrix} A\alpha_1 & A\alpha_2 \end{pmatrix}$$

(式39)

となる。さらに、

$$P=\begin{pmatrix} \alpha_1 & \alpha_2 \end{pmatrix}=\begin{pmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \end{pmatrix}$$

(式40)

なる固有ベクトルの行列を置くと、

$$AP=A\begin{pmatrix} \alpha_1 & \alpha_2 \end{pmatrix}=\begin{pmatrix} A\alpha_1 & A\alpha_2 \end{pmatrix}$$

(式41)

となることから、この(式 41)を(式 39)と比較することによって、

$$AP=UD$$

(式42)

が得られる。

このとき、P は正規直交行列であり、

$$PP^T = I$$

(式43)

である（I は単位行列を示す）。したがって、

$$A = APP^T$$

(式44)

と表せることから、これに(式 42)を代入すると、

$$A = UDP^T$$

(式45)

が得られる。データ行列 A を、(式 37)、(式 38)、を用いて、(式 45)に示した 3 つの行列に分解することを、行列の特異値分解と呼ぶ。単純化して説明するならば、行列の特異値分解とは、共分散行列の固有値問題を元のデータ行列を用いて表すことによっても示すことができるということになる。

ちなみに、ここまで 2 次元データのみを例としているが、次元数が増えてもこれらの行列のサイズが大きくなるだけで原理は変わらない。

(式 37)の対角要素が特異値であり、(式 39)の個々の要素は左特異ベクトル、(式 40)の P の個々の要素は右特異ベクトルと呼ばれる。

2．5. 文化現象に対しての潜在意味分析の意義

本論で後述する分析例は、女児向けコミック雑誌のテキスト分析への潜在意味分析の応用事例である。このような応用を行う場合、潜在意味分析において本来想定されている原理とはいくぶん異なる前提が用いられる。

潜在意味分析においては、文の意味・意図から、単語の意味・意図を推定するという流れが基本となるが、これを文化現象を示すテキスト分析に応用する場合には、文の意味や意図を問題とするのではなく、その文が持つ訴求力が問題とされる。つまり、ある時代において広く受容され、人気を得たテキストには、何らかの訴求力が存在していると考える。ただし、訴求力には様々なものがあると想定されるので、それらを訴求因と呼ぶ。一つのテキストには、複数の訴求因の表現形であり、単語列として表現されている。それらは、一つのテキストにおいては薄くしか見ることができないが、多数のテキストを分析することによって抽出することが可能であると考えるのが、基本的立脚点である。

もちろん、多変量解析を発見的手法として使用する際の問題点と同じく、そのような訴求因が発見できない場合も少なくない。テキストマイニングとは、テキスト分析のうちで発見的手法をとるものに対しての呼称であるが、どのように優れた分析手法を用いても、意義のある構造を発見できない場合もある。それは「テキストマイニング」という比喩的呼称にも良く示されている。「鉱山土壌からの有用な物質の精製」においては、そもそもその土壌に有用な物質が含まれていないのであれば、どのように高機能な精製機械を以てしても、よい結果を得ることはできない。

3. 分析

3.1.分析対象

1979 年から 2009 年にかけて発刊された女子児童・女子生徒向けコミック雑誌を約 10 年ごとに選択した.2009 年分の選択にあたっては発行部数を参照し、その多いものを選んだ。過去発行分に関しては、発行部数を確認することが困難であったため、2009 年分として選択した雑誌の中から選択した。ただし、「Betsucomi ベツコミ」に関しては、その前身である「少女コミック」を選択した。1979 年・1989 年・1999 年に関しては原則として 3 冊づつ選択, 2009 年に関しては 7 冊を選択し,合計 15 誌が対象とされた.

(1979 年) 3 誌 24 タイトル

週刊少女フレンド　1979 年第 18 号

別冊マーガレット　1979 年 4 月号

週刊少女コミック　1979 年第 4 号

(1989 年) 2 誌 16 タイトル

花とゆめ 1989 年 2 号
別冊マーガレット 1989 年 2 月号

(1999 年) 3 誌 25 タイトル
花とゆめ　1999 年 2 号
月刊少女フレンド　1999 年 11 月号
少女コミック 1999 年 1 月 20 日号

(2009 年) 7 誌 90 タイトル
ちゃお　2009 年 12 月号
Betsucomi (ベツコミ) 2009 年 12 月号
別冊花とゆめ　2009 年 12 月号
マーガレット 2009 年 11/20 号
別冊 マーガレット 2009 年 12 月号
なかよし 2009 年 12 月号
りぼん 2009 年 12 月号

入力した作品数は計 155 タイトル, 合計 398671 文字／25905 行であった.

3.2.分析手順

3.2.1.文オブジェクトの抽出

分析対象とするテキストは，コミックの本編中の文字列のみとし，また，原則としてセリフのみを分析対象とした．つまり，状況の説明やト書きに該当する部分はすべて捨象し，原則として「吹出し」の中に記述されているもののみを対象とした.ただし，吹出しの中に記載されておらず，図中にあるものであっても，明らかに登場人物の内言や内心の表現であると思われるものや，セリフであることが明らかであるものに関しては，分析対象とした．一つの作品を一つの文オブジェクトとし，合計 155 の文オブジェクトが構成された.

3.2.2.文オブジェクトー単語の共起行列の作成

前項で示した方法によって抽出された文オブジェクトのデータ列から単語を抽出し，文オブジェクトー単語の共起行列を作成した.

形態素解析は MeCab を用いて行い，文オブジェクトー単語の共起行列の作成は RMeCab を用いて行った．抽出する単語は，名詞・形容詞とし，動詞・形容動詞・助詞・助動詞・数詞・接続詞・記号は分析対象からは除外した．また、1 文字のひらがな、1 文字のカタカナ、英単語（英語表記のもの）もすべて除外した。

共起行列の作成にあたっては，2 つ以上のタイトルに共起している単語のみを対象とし，重み付けは行わず，一つのタイトル中に同じ単語が何度出現しても得点を 1 とした.ただし,分析の内容に応じて，IDF などの大域的重み付けや，文書の長さによる正規化重み付けを使用しているが，本稿で後述する分析においては重み付けを行っていない.

3.2.3.潜在意味分析

前項の手順で作成された文オブジェクトー単語の共起行列を特異値分解し，単語の主成分得点および文オブジェクトの主成分を得た．特異値分解は統計処理ソフトウェアＲを用いて行われた。

3.2.4.頻度分析

抽出された文オブジェクトのデータ列から単語を抽出し，年ごとの単語の頻度を計測した．形態素解析は MeCab を用いて行い，頻度の計測は RMeCab を用いて行った．抽出する単語は，名詞・形容詞とし，動詞・形容動詞・助詞・助動詞・数詞・接続詞などは分析対象からは除外した.

5. 結果および考察

5.1.潜在意味分析

ここでは，名詞のみを分析対象として行った潜在意味分析の結果のみを例示した．4.2 で述べたように重み付けは行わなかった．また，また，90 タイト

ルのうち，60%（54 タイトル）を越えるものに出現する「それ」「ここ」「何」「人」など 18 個の単語は共起行列から除外した．

第１軸において主成分得点の絶対値が大きかった単語は，「好き」「みんな」「たち」「一緒」「中」（いずれも負値）などであり，第２軸では「わけ」「教室」「ウソ」「ムリ」「気持ち」「クラス」「告白」（いずれも正値），「君」「街」「世界」「危険」「問題」「言葉」（いずれも負値）などが見られた．

これらの結果から，分析対象としたコミックの意味構造として「孤独―凝集」「私的世界－公的世界」を抽出した．

これら二つの軸によってコミック（文書オブジェクト）を配置した例を図 1 に示す．図中，H_，B_などの記号列は以下の意味を示す．

C_ ：ちゃお　2009 年 12 月号
B_ ：Betsucomi（ベツコミ）2009 年 12 月号
H_ ：別冊花とゆめ　2009 年 12 月号
M_ ：マーガレット 2009 年 11/20 号
BM_：別冊 マーガレット 2009 年 12 月号
N_ ：なかよし 2009 年 12 月号
R_ ：りぼん 2009 年 12 月号

また，タイトル名の記載は冒頭 2 文字に省略されている．

5.2.頻度分析

1979 年の少女向けコミック誌 3 誌において出現頻度の高かった名詞の 1 位から 4 位は「こと（1.63%）」「人(1.19%)」「あたし(0.99%)」「ちゃん(0.92%)」（いずれもかっこ内は，1979 年度分析対象テキストの総名詞数 12006 に対しての比率）であった．ここでは，人称表現に着目し，分析結果の一部を示す．

まず，1 人称に関して，特に女性が多く使用していると推測される 1 人称の名詞の変遷を図１に示した．「わたし」が使用されなくなり，「私」に代わっていく様子が見られるが，これは他の単語でも同様に見られる漢字化傾向である．全体として 1 人称の出現頻度は 1979 年から順に，196(1.6%)，167(1.8%)，226(1.6%)，755(2.0%)と上昇しており，1979 年と 2009 年の間での出現頻度の差は統計的に有意である（比率検定，$df = 1$，$p < 0.01$)．少女向けコミック誌においては，女性が自分のことを１人称代名詞で呼ぶ機会が多くなっているということができる．また，2 人称についての同様の分析を図２に示した．ここにおいては，2 人称の出現頻度が有意に低くなっている（比率検定，$df = 1$，$p < 0.001$)．さらに，図３に示したように，「みんな」「いっしょ」「～たち」「～達」などといった人の凝集状態を示す単語（凝集性単語）の出現頻度は，特徴的な変遷を示している．これらの単語は 1979 年から，177(1.47%)，35(0.38%)，138(0.99%)，468(1.26%)（かっこ内は，各年度の名詞出現総数に対しての比率）というように，1989 年のテキストとでは急激に出現頻度が低くなり，そののち回復するように上昇している．1979 年と 2009 年の比率の差は，統計的に有意であるとは言えないが（比率検定，$df = 1$，$p = 0.085$），それ以外の組み合わせ(1979-1989, 1979-1999, 1989-1999, 1999-20009）はすべて有意である．

上記の分析のみから推定しうることはそう多くはないが，2009 年の少女向けコミック雑誌の潜在意味分析に見られたような「私的－公的」という対立関係や「孤独－凝集」に対しての価値観の変化が示唆される．また，その他，形容詞，動詞などの出現頻度の変遷を見ることによっても，同様のことが示唆されるとともに，恋愛に対しての価値観や距離感の変化を見ることができる．

４．結果および考察

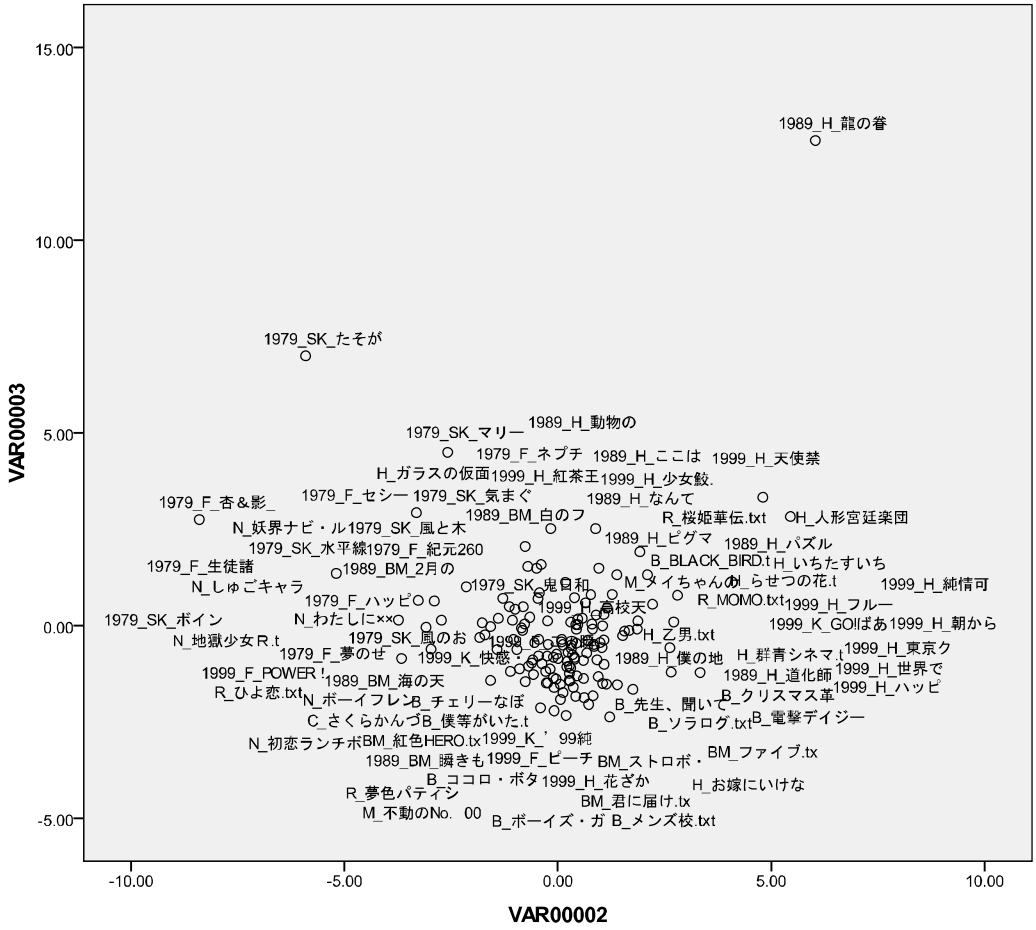

図 4　第 2 主成分ー第 3 主成分の変量プロット

4. 結果および考察

4.1.潜在意味分析

名詞および形容詞を分析対象として行った潜在意味分析の結果を示す.

すべてのタイトルを表示しているので把握しにくいが、発行年による偏差を見ることができる。以下に、発行年によって分類した変量プロットを示す（図 5

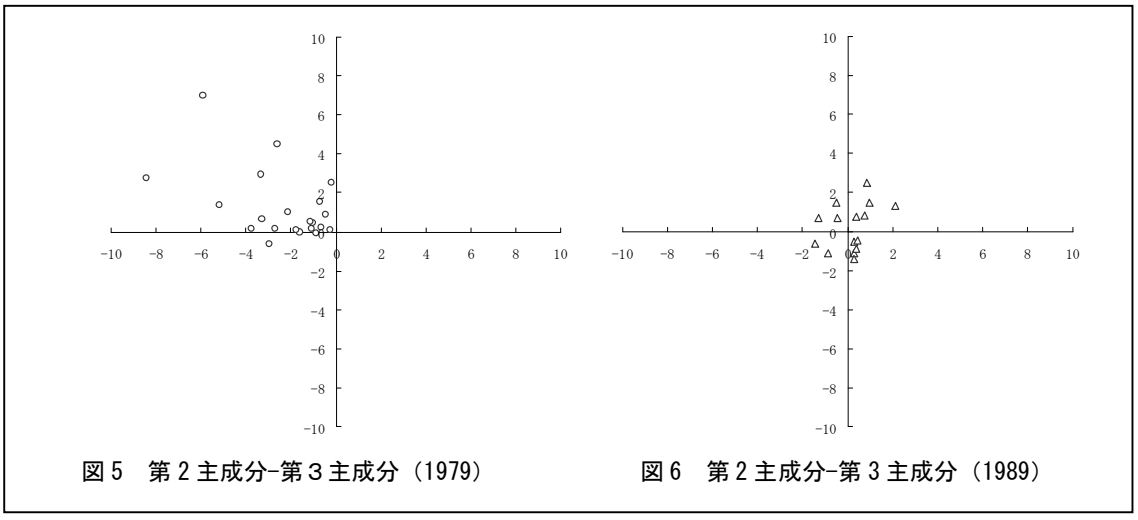

図 5　第 2 主成分-第３主成分（1979）

図 6　第 2 主成分-第 3 主成分（1989）

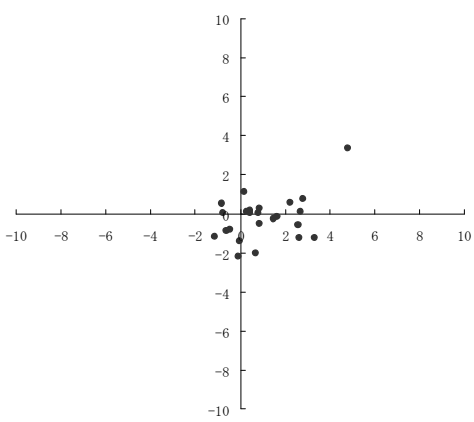

図 7　第 2 主成分-第 3 主成分（1999）

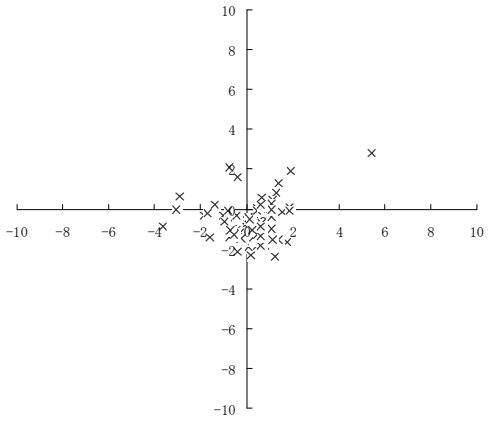

図 8　第 2 主成分-第 3 主成分（2009）

～図 8）。そこに見られるように、第２主成分（横軸の値）は、負値から正値への移動が見られる。ただし、1999 年まで正値方向に振れていた値は、2009 年に原点近傍に戻している。

第３主成分値（縦軸の値）についても同様の移動が見られるが、1999 年と 2009 年の間にはほとんど差異が見られない。

第１主成分に関しては、特徴的な変化は見られなかった。経年によって特徴的に変化しているのは、ここで示した第２主成分と第３主成分である。

第２主成分と第３主成分による単語の主成分得点の一部を以下に示す（表 2、表 3）。しかしながら、これらの主成分得点のリストからは、それらの主成分軸が何を示しているのかは判然としない。

4.2.頻度分析

1979 年の少女向けコミック誌３誌において出現頻

順位	単語	第２主成分得点	単語	第３主成分得点
1	俺	2.53	わたし	2.04
2	私	2.50	さま	1.89
3	達	2.43	昔	1.73
4	誰	2.24	力	1.58
5	お前	2.22	ぼく	1.57
6	様	2.16	とき	1.54
7	奴	2.06	心	1.47
8	一緒	1.88	身	1.44
9	今	1.79	体	1.41
10	君	1.68	若い	1.40
11	事	1.60	者	1.32
12	他	1.59	確か	1.29
13	方	1.58	命	1.29
14	大変	1.57	本	1.27
15	物	1.52	風	1.27
16	確か	1.51	娘	1.26
17	絶対	1.47	ひとつ	1.26
18	バカ	1.45	名	1.26
19	前	1.40	闇	1.26
20	者	1.34	ところ	1.25
21	所	1.32	きり	1.23
22	僕	1.28	旅	1.22
23	良い	1.26	はず	1.20
24	皆	1.25	使い	1.16
25	後	1.21	あなた	1.16
26	無事	1.20	きれい	1.16
27	大丈夫	1.20	ども	1.15
28	命	1.15	きみ	1.15
29	何	1.14	ごろ	1.13
30	欲しい	1.13	暗い	1.13

表 2　単語の主成分得点（得点上位）

順位	単語	第２主成分得点	単語	第３主成分得点
3525	みたい	−1.16	フツー	−0.91
3526	たち	−1.17	人	−0.91
3527	学校	−1.18	みたい	−0.91
3528	うそ	−1.19	こっち	−0.93
3529	オレ	−1.19	気持ち	−1.01
3530	さま	−1.19	女子	−1.01
3531	やっ	−1.24	マジ	−1.04
3532	おもしろい	−1.25	ハイ	−1.05
3533	おまえ	−1.28	彼氏	−1.08
3534	はやい	−1.31	楽しい	−1.11
3535	なに	−1.33	ムリ	−1.12
3536	家庭	−1.36	お前	−1.13
3537	くん	−1.39	ホント	−1.15
3538	教室	−1.40	とこ	−1.16
3539	わるい	−1.40	さん	−1.16
3540	あと	−1.41	コレ	−1.20
3541	男の子	−1.44	先生	−1.26
3542	ほう	−1.44	友達	−1.35
3543	いつか	−1.46	あんた	−1.39
3544	ころ	−1.64	オレ	−1.39
3545	ぼく	−1.64	一緒	−1.48
3546	友だち	−1.65	そっ	−1.51
3547	あたし	−1.70	もん	−1.55
3548	わたし	−1.79	くん	−1.69
3549	とき	−1.88	今日	−1.72
3550	きょう	−1.98	かわいい	−1.88
3551	いま	−2.12	俺	−1.91
3552	いっしょ	−2.15	ちょ	−1.94
3553	だれ	−2.21	ちゃん	−2.07
3554	きみ	−2.34	好き	−2.22

表 3　単語の主成分得点（得点下位）

度の高かった名詞の 1 位から 4 位は「こと（1.63%)」「人(1.19%)」「あたし(0.99%)」「ちゃん(0.92%)」（いずれもかっこ内は，1979 年度分析対象テキストの総名詞数 12006 に対しての比率）であった．ここでは，人称表現に着目し，分析結果の一部を示す．まず，1 人称に関して，特に女性が多く使用していると推測される 1 人称の名詞の変遷を図 1 に示した．「わたし」が使用されなくなり，「私」に代わっていく様子が見られるが，これは他の単語でも同様に見られる漢字化傾向である．全体として 1 人称の出現頻度は 1979 年から順に，196(1.6%)，167(1.8%)，226(1.6%)，755(2.0%)と上昇しており，1979 年と 2009 年の間での出現頻度の差は統計的に有意である（比率検定，$df = 1$，$p < 0.01$)．少女向けコミック誌においては，女性が自分のことを 1 人称代名詞で呼ぶ機会が多くなっているということができる．また，2 人称についての同様の分析を図 2 に示した．ここにおいては，2 人称の出現頻度が有意に低くなっている（比率検定，$df = 1$，$p < 0.001$)．さらに，図 3 に示したように，「みんな」「いっしょ」「～たち」「～達」などといった人の凝集状態を示す単語（凝集性単語）の出現頻度は，特徴的な変遷を示している．これらの単語は 1979 年から，177(1.47%)，35(0.38%)，138(0.99%)，468(1.26%)（かっこ内は，各年度の名詞出現総数に対しての比率）というように，1989 年のテキストとでは急激に出現頻度が低くなり，そののち回復するように上昇している．1979 年と 2009 年の比率の差は，統計的に有意であるとは言えないが（比率検定，$df = 1$，$p = 0.085$），それ以外の組み合わせ(1979-1989, 1979-1999, 1989-1999, 1999-20009）はすべて有意である．

4.3.潜在意味分析と頻度分析の結果からの考察

上記の分析のみから推定しうることはそう多くはないが，1979 年から 2009 年にかけて女児向けコミック雑誌に発生していた何らかの訴求因の変遷を、

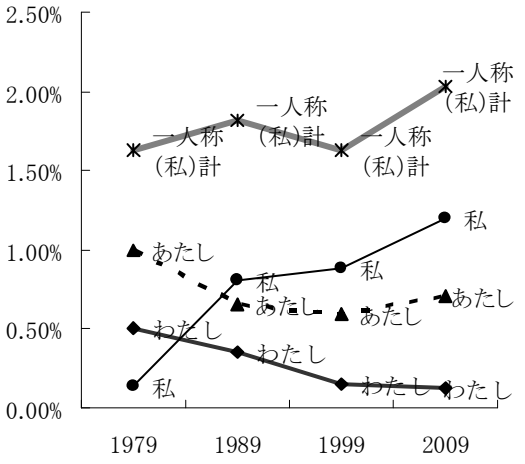

図 9　1 人称の出現頻度の変遷

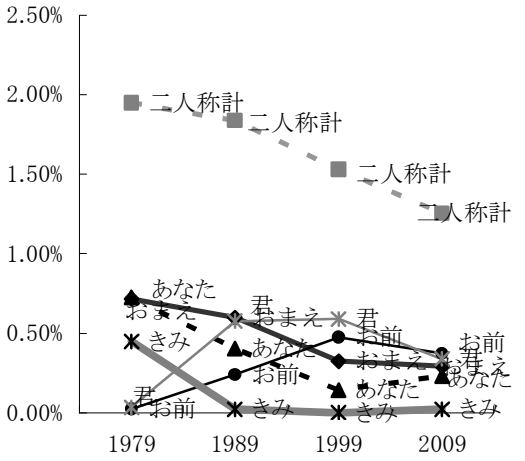

図 10　2 人称の出現頻度の変遷

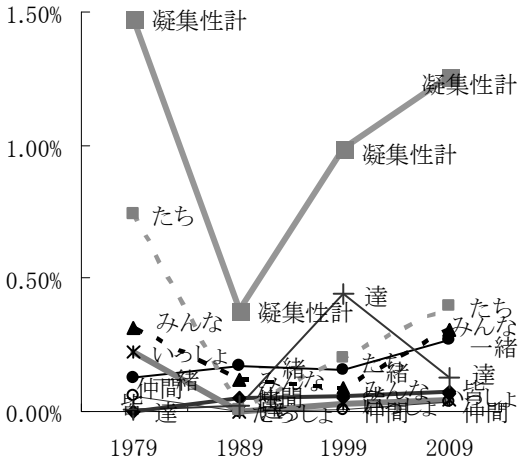

図 11　凝集性単語の出現頻度の変遷

頻度分析の結果によって解釈するならば、そこに存在しているのは「私的－公的」という対立関係や「孤独－凝集」に対しての価値観の変化であることが示唆される．特に、凝集性単語の使用の変遷は、潜在意味分析において抽出された第２主成分の変化に相同であり、「仲間」もしくは「小集団」との距離間や接し方、もしくはそれらにどの程度重きを置くかという価値観の変化が発生していたと考えることもできる。

5. おわりに

テキスト分析の領域において潜在意味分析は重要な分析手法の一つとなりつつあるが、その適用に関しては、まだ模索の状態が続いていると言える。潜在意味分析は、談話分析などのテキスト分析においても有用な手法であるが、本稿で述べたような文化的テキストの訴求構造分析手法として利用可能である。一方で、これまで標準的に用いられてきた主成分分析や因子分析、もしくは数量化Ⅲ類列に比べて、十分な理論的検討が行われているとは言いがたい状況が存在する。その原理と数理が十分に理解されれば、テキスト分析における主力の手法となると推定される。

注

[1]大木智世，高田明典：物語構造分析手法の RPG 分析への応用，多文化・共生コミュニケーション論叢，*1*，pp.45-77（2006）.

[2]水越詩緒莉，高田明典，林延哉：娯楽制作物の訴求構造抽出のための物語構造分析手法の提案，国際情報技術フォーラム（FIT2008）講演論文集，pp.3-343-3-344（2008）.

[3]高田明典：コンピュータゲームの心理学，芸術科学会誌，*1*，pp.66-74（2001）.

参考文献

石田基広：R によるテキストマイニング入門，森北出版(2008).

和多太樹，関隆宏，田中省作，廣川佐千男：単語の出現頻度に着目した病院評判情報の分析，情報処理学会研究報告．SLP，音声言語情報処理，*50*，pp.15-20（2005）.

豊田秀樹：データマイニング入門，東京図書(2008).

重久礼美，高田明典：映像作品の物語構造分析の自動化に関する一研究，多文化・共生コミュニケーション論叢，*2*，pp.11-22（2007）.

Landauer，T. K.，Foltz，P. W.，& Laham，D. ：Introduction to Latent Semantic Analysis，*Discourse Processes*，*25*，pp.259-284（1998）.

Deerweester,S.，Dumais,S.，Furnas,G,
Thomas,L.，Harshman,R.：Indexing by Latent Semantic Analysis，*Journal of the American Society for Information Science, 41, 1*，pp.391-407（1990）.