特集論文

自己組織化ネットワークによる動的クラスタの可視化編纂

Compilation to Visualize the Dynamic Clusters by the Adapted Self-Organizing Network

福井 健一 Ken-ichi Fukui
大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
fukui@ai.sanken.osaka-u.ac.jp, http://www.ai.sanken.osaka-u.ac.jp

斉藤 和巳 Kazumi Saito
静岡県立大学 経営情報学部
School of Administration and Informatics, University of Shizuoka
k-saito@u-shizuoka-ken.ac.jp

木村 昌弘 Masahiro Kimura
龍谷大学 理工学部
Department of Electronics and Informatics, Ryukoku University
kimura@rins.ryukoku.ac.jp

沼尾 正行 Numao Masayuki
大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
numao@ai.sanken.osaka-u.ac.jp

keywords: Self-Organizing Map, cluster dynamics, visualization, topic transition

Summary

We have been developing a neural network-based approach for visual *information compilation*. We have extended the Self-Organizing Map (SOM) model by introducing a sequencing weight function into the neuron topology, called Sequence-based SOM (SbSOM). SbSOM visualizes the dynamics of various clusters such as their generation or extinction, convergence or divergence, and merging or division. By utilizing the neuron topology and the neighborhood function of SOM, SbSOM can mitigate the problems associated to the conventional sliding-window method. We clarified a target problem class of SbSOM and confirmed the basic properties of this proposed method using a two-dimensional simulated sequential dataset. Moreover, our experiment using a dataset of real-world news articles indicates that topic transition can indeed be seen from the acquired map. Visualization of cluster sequential changes aids in the comprehension of such phenomena which come useful in various domains such as fault diagnosis and medical check-up, among others.

1. はじめに

世の中の様々な出来事や物理現象は時間と共に変化するため，時系列を考慮して現象の主要な変化とその成分を捉えることは重要である．例えば，ニュース記事などから自動抽出したトピックの生成・消滅などのトピック系譜，様々なセンサーデータから得られるプラントなどの状態変化，医療情報から健康状態の変化など，様々挙げられる．このような時々刻々と変化する現象の全貌を可視化することは，現象の理解を助け，具体的な応用ではプラントなどの故障診断や、医療情報からの健康診断などの一助となる重要な技術であると考える．

一方，近年，雑多な情報を知的な処理により理解を容易にし，情報へのアクセスを支援する情報編纂（Information Compilation）が提唱されている [加藤 06]．情報編纂には様々な要素技術が必要とされるが，本研究は，文書データに限定されない汎用性のある学習手法をベースとする，機械学習に基づく情報可視化の基盤技術と言える．ここで，大規模データを大まかに捉えるためにはクラスタリングする必要があるが，そのクラスタの時系列変化（クラスタダイナミクス）を可視化する方法として，著者らは自己組織化マップ (Self-Organizing Map:SOM)[Kohonen 95] 学習モデルを拡張した Sequence-based SOM (SbSOM) を提案している [Fukui 05]．SOM は教師なし競合型のニューラルネットワーク学習のひとつであり，クラスタリングと低次元（通常 2 次元）への射影を同時に得ることができるため，視覚的データマイニング手法としても知られている．SbSOM では，通常 SOM の予め定義される可視化層トポロジーの近傍に空間的近傍が現れる性質を活かし，トポロジー上にデータの系列（すなわち時間）に依存する重みを導入することでトポロジーの一方向に時間的な意味を持たせる．従来の単純なウィ

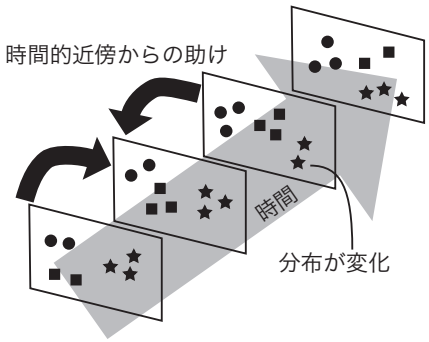


図 1 時間的近傍の助けを借りたクラスタリングの概念図.

ンドウ方式と比べて，2 章で述べる問題点を軽減できるメリットがある．

我々はこれまでに，新聞記事系列データ [Fukui 05]，肝炎患者の血液検査値から成る医療系列データ [Fukui 06]，固体電池の損傷計測値である Acoustic Emission(AE) データ [福井 07, Fukui 07] に SbSOM を適用し有用性を実証してきた．新聞記事データではトピック系譜，医療データでは薬の効果有り/無しでの患者クラスタの変化の様子を，AE データでは固体電池の破壊ダイナミクスを可視化し，それぞれ現象として解釈できることを確認した．

本稿では，それらを包括する形でまとめ，対象とする問題クラスおよび SbSOM の特性を明らかにする．まずひとつめの実験では，2 次元の人工系列データを用意し，SbSOM によるクラスタダイナミクス可視化の基本的性質を確認する．さらに，実際の新聞記事データから抽出したトピック推移の可視化について再考・詳細な分析を行い，トピックの派生，多様化と収束などのトピック推移が読み取れることを示す．人工データによるデータの特性毎の効果および，それらに基づくトピック推移の考察は著者らの先行研究 [Fukui 05] にはなかった新たな知見である．これにより，自己組織化ネットワークに基づく機械学習による可視化編纂の可能性を探る．

2. 従来の手法の問題点と提案法のコンセプト

従来のクラスタリング・分類問題は一般に静的なデータを対象としており，クラスタの時間的な変化を陽に捉えることはできない．単純な方法として，一括して分類したデータに対して時間変化を追って行くことや，スライディングウィンドウ方式でクラスタを生成していく方法が考えられる．しかし，前者はデータがどのクラスタに属するか時系列でみることはできてもクラスタ自体の変化を追うことは不可能である．後者はクラスタの変化をある程度追うことができるが，適切なウィンドウ幅の設定，必然的にウィンドウ内のサンプル数の減少，またウィンドウ間でのクラスタの対応付け問題がある．

これらの問題に対して，本研究ではウィンドウ方式をベースとし，時間的近傍からの助けを借りたクラスタリングを考える (図 1)．ここで，SOM の可視化層ニューロンには予めトポロジーが定義されており，近傍関数により (特徴) 空間的近傍がトポロジー上の近傍になるような学習が行われる．これにより空間的に緩やかなクラスタリングを可能にしている．この特性に着目し，SbSOM では可視化層トポロジー上に順序付け重み関数を導入することで，時間軸方向にも緩やかなクラスタリングを行う．これにより，リジッドなウィンドウ方式と比べて，適切なウィンドウ幅の設定やサンプル数の減少問題は緩和される．また近傍関数によりトポロジー上の近傍に時空間的近傍が現れるため，クラスタの対応付け問題はなくなる．

3. SOM 学習アルゴリズム

まず，提案法のベースとなる Kohonen の自己組織化マップ (Self-Organizing Map: SOM) について概説する．入力データを N 個の V 次元ベクトル $\mathbf{x}_n = (x_{n,1}, \cdots, x_{n,V})$, $(n = 1, \cdots, N)$ とする．SOM は入力層と可視化層の 2 層から成り，可視化層は通常，予めトポロジーの定義された低次元 (多くの場合 2 次元) 格子状に配置された M 個のニューロン (ノード) 群で構成される．

第 j ニューロンの可視化層の位置を $\mathbf{r}_j = (\xi_j, \eta_j)$ とする．各ニューロンには入力データと同じ次元の参照ベクトル $\mathbf{m}_j$ が割り当てられ，SOM での学習は，ニューロンの参照ベクトルを入力データに近づけるように収束するまで更新する．その際，可視化層における近傍ニューロンからの影響を受けることで，特徴空間での位相関係を可視化層においてできるだけ保存する．大規模なデータに対して SOM は，各入力データに対する勝者ニューロンによりクラスタリングされ，さらに可視化層トポロジーにおいてクラスタ間の類似度を反映したマップを生成する．

以下に，一般によく用いられているバッチ処理かつ学習パラメータの減少戦略を取る，SOM 学習アルゴリズムを示す．

Step 1. 参照ベクトル $\{\mathbf{m}_1, \cdots, \mathbf{m}_M\}$ を初期化する．

Step 2. 勝者ニューロン $\{c(\mathbf{x}_1), \cdots, c(\mathbf{x}_N)\}$ を次式により求める．

$$c(\mathbf{x}_n) = \arg\min_j ||\mathbf{x}_n - \mathbf{m}_j||.$$

Step 3. 勝者ニューロン $\{c(\mathbf{x}_1), \cdots, c(\mathbf{x}_N)\}$ が前回と変わらなければ終了．

Step 4. 参照ベクトル $\{\mathbf{m}_1, \cdots, \mathbf{m}_M\}$ を次式により更新する．

$$\mathbf{m}_j^{new} = \mathbf{m}_j + h_{c(\mathbf{x}),j}[\mathbf{x}_n - \mathbf{m}_j].$$

ここで，$h_{c(\mathbf{x}),j}$ は近傍関数であり，勝者近傍の更新の大きさを調節する．近傍関数には，次式のガウス

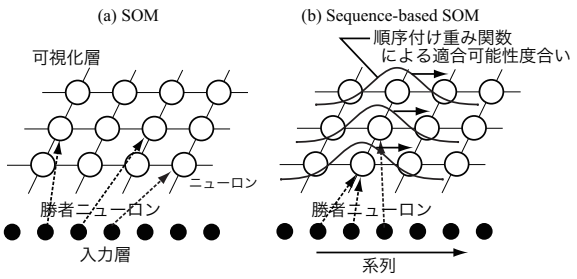


図 2 勝者ニューロン選択の違い．(a) SOM は空間的距離のみにより決定される，(b) Sequence-based SOM では，順序付け重み関数の下で空間的距離により決定される．順序付け重み関数は入力データの系列に従って可視化層トポロジー上の一方向に移動していく．

関数がよく用いられる．

$$h_{c(\mathbf{x}),j} = \alpha \exp\Big(-\frac{\|\mathbf{r}_j - \mathbf{r}_{c(\mathbf{x})}\|^2}{2\sigma^2}\Big).$$

Step 5. 近傍関数の学習パラメータ α および σ を数回の反復毎に減少させる．step 2 に戻る．

4. Sequence-based SOM

4·1 順序付け重み関数の導入

通常 SOM では，空間的距離のみによって勝者ニューロンを決定 (その結果，クラスタリング) していた．これに対して SbSOM では，入力データの系列に従って可視化層トポロジー上に重みを与える．これにより，時間的距離と空間的距離の両方を考慮して勝者ニューロンを決定する．具体的には，トポロジーの ξ 方向を時間方向にする場合，SbSOM では勝者ニューロン選択における距離定義を以下のように修正する．

$$c(\mathbf{x}_n) = \arg\min_j \psi(n,\xi_j)||\mathbf{x}_n - \mathbf{m}_j||. \qquad (1)$$

$\psi(n,\xi_j)$ は可視化層トポロジー上に，入力データの系列に応じた重み付けをする関数である．この順序付け重み関数により，可視化層トポロジー上に時間の概念*1を導入する．例えば，空間的距離が同じ場合，順序付け重み関数の値が小さい方の参照ベクトルが勝者として選ばれることになる．式 (1) は，時間的距離と空間的距離の和で与えることも考えられるが，規格化が困難なため積で与えた．

当然ながら時間的距離を導入することで空間分解能は相対的に低下するため，両者のバランスを取れるようにすることを考える．ある入力データと参照ベクトルとの空間的距離が近いとき，実際の入力系列に対してトポロジー上に現れるデータ順序の逆転を許容するように順序関数を与える．n 番目のデータは n/N の割合に位置し，時系列を導入するトポロジー上の一方向のニューロン ξ_j はその方向に ξ_j/ξ_M の割合に位置している．その差の絶対値を $\epsilon = |\xi_j/\xi_M\text{-}n/N|$ とおく．

順序の逆転を許容する場合，次式により与える．

$$\psi_{exp}(n,\xi_j) = e^{w\epsilon}. \qquad (2)$$

ここで，$w \geq 0$ は系列の順序関係の勝者ニューロン選択への影響力を調節するパラメータである．w が大きくなるほど，可視化層トポロジー上の一方向に対して系列順序の入れ替わりを抑止する効果を持つ．直感的には，データの系列に合わせて重み関数を軸の一方向にスライドすることで，可視化層上に系列の順序関係を実現している (図 2 参照)．また，$w = 0$，すなわち $\psi(n,\xi_j) = 1$ のとき，通常 SOM になる．

順序の逆転を許容しない場合，次式により与える [Fukui 05]．

$$\psi_{strict}(n,\xi_j) = \begin{cases} 1 & if \quad \epsilon < \frac{1}{2K}, \\ \infty & otherwise. \end{cases} \qquad (3)$$

ここで，K は ξ 方向のニューロン数を表す．

式 (2) において十分に w を大きく取れば逆転が起こらなくなるため式 (3) と等価である．しかし，式 (3) のように窓関数として与えた方が，式 (1) においてウィンドウ内に含まれる参照ベクトルとの比較のみで済むので，計算コストの点から有利である．

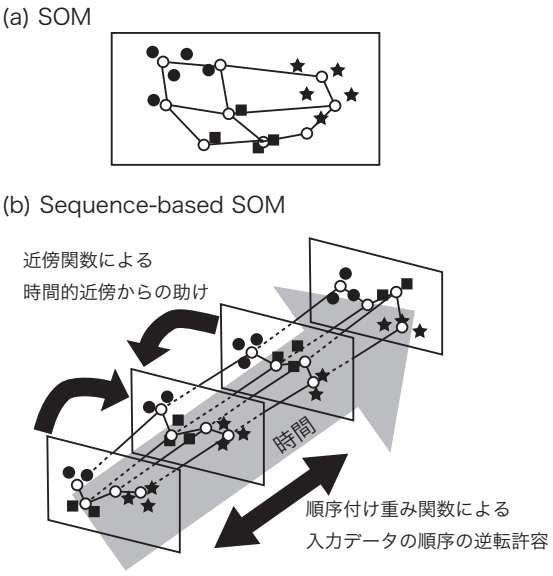


図 3 特徴空間とニューロンの可視化層トポロジーの関係の概念図．○がニューロンを表している．(a) SOM では空間的近傍がトポロジー上の近傍になる． (b) Sequence-based SOM では時間的・空間的近傍がトポロジー上の近傍になる．ただし，ψ_{exp} では時間は図のようにウィンドウ方式とはなっていない．

4·2 Sequence-based SOM における近傍関数の役割

通常の SOM には近傍関数 $h_{c(\mathbf{x}),j}$ が定義されており，空間的近傍からの影響を受けて参照ベクトルは更新される．これに対して SbSOM での近傍関数は，時間的・空間

*1 場合によっては実際の入力系列に対して，トポロジー上に現れるデータの順序が入れ替わることを許容する緩い時間となっている．

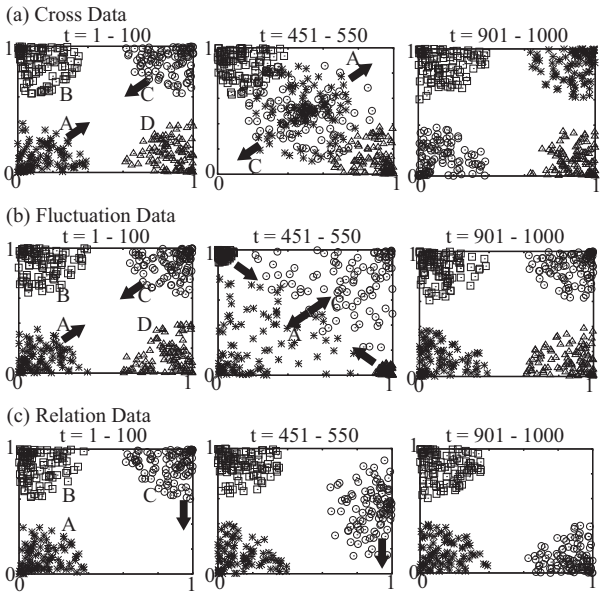


図 4 3 種類の 2 次元人工系列データ．図は一定期間のデータをプロットしている．

的近傍の意味合いを持つことになる (図 3)．これによりウィンドウ方式でのサンプル数減少問題に対して，前後のデータの助けを借りながらクラスタリングできる．また，SbSOM では，時間的・空間的近傍がトポロジー上の近傍になるため，ウィンドウ間のクラスタの対応関係も自己組織的に取ることができる利点がある．

5. 実験 1: 人工データを用いた実験

実験 1 では，人工的に生成した 2 次元系列データを用いて，クラスタの融合・分離，拡大・縮小，クラスタ間の関係の変化が SbSOM マップ上にどのように現れるか基本的性質を確認した．

5·1 人工データセット

まず，実験に使用した人工データセットについて説明する．人工データセットは，2 次元の系列データであり，領域を $x \in [0,1]$，$y \in [0,1]$ とした．データの生成手順について説明する．まず，4 クラス（または 3 クラス）のデータ生成関数の初期中心位置を領域の四つ角（または 3 つの角）に配置した．生成関数は，中心と半径をパラメータとして持つ円形領域内にランダムにデータを生成する．そして，各クラスから順番にデータを 1000 点 (t=1-1000) (つまり 4 クラスの場合，1 点目はクラス 1，2 点目はクラス 2，3 点目はクラス 3，4 点目はクラス 4，5 点目はクラス 1... 以下続く，各クラス 250 点)，生成関数の中心位置と半径を時間と共に変化させて生成した．このようにして生成した次の 3 種類の人工データセットを用意した．

(1) Cross Data クラスタの融合と分離を表し，対角線上のクラスタ A と C が中央で融合し交差する (図 4(a))．

(2) Fluctuation Data クラスタ領域の拡大と縮小を表し，まずクラスタ A と C の領域が時間と共に増加し，B と D の領域は減少する．その後，初期状態に戻る (図 4(b))．

(3) Relation Data クラスタ間の関係の変化を表す．このデータセットは 3 クラスから成り，初期状態はクラスタ C の中心は $(1,1)$ に位置しクラスタ B に近いが，時間と共に $(1,0)$ に移動しクラスタ A に近くなる (図 4(c))．

5·2 人工データに対する結果

人工データを用いた可視化結果を以下に示す．SbSOM の可視化層トポロジーは 20×15 の直交格子に設定し，順序付け重み関数は ψ_{exp} を用い，そのパラメータは $w = 200.0$ とした．マップ右方向が系列方向を表している．また，各クラスタ (ニューロン) の代表クラスは，クラスタに属するデータのクラスの多数決により決定した．図には代表クラスのみ表示してある．

(a) クラスタの融合と分離 マップ左側，系列の始めの方は 4 つのクラスタは分かれているが，中盤ではクラスタ A と C が混じり合って現れている．実際の入力データは t=500 で A と C は完全に混じり合っているが，系列の前後の影響である程度つながったクラスタとして現れている．そして，マップ右側，終盤では A と C は再び分離していく様子がマップ上に現れている (図 5(a))．

(b) クラスタ領域の拡大と縮小 マップ中盤で，クラス A・C を代表とするクラスタが増加し，クラス B・D を代表とするクラスタは減少している (図 5(b))．そして，終盤では初期状態に戻っていく様子が現れており，クラスタ領域の拡大と縮小がマップから確認できる．

(c) クラスタの関係の変化 最初はクラスタ B と C が隣り合っているが，中盤からクラスタ B に隣り合う C が減っていき，クラスタ A に隣り合う C が増えていく様子が現れている (図 5(c))．徐々にクラスタ間の位置関係が変化していく様子が確認できる．

6. 実験 2: 新聞記事データを用いた実験

次に，実際の新聞記事から自動抽出したトピック群を用いて，SbSOM による可視化結果からトピック推移が読み取れることを確認し，実際の出来事との対応関係を例示する．また，順序付けパラメータを変化させた時の分類性能に及ぼす影響を定量的に評価し，さらに視覚的な変化を確認した．

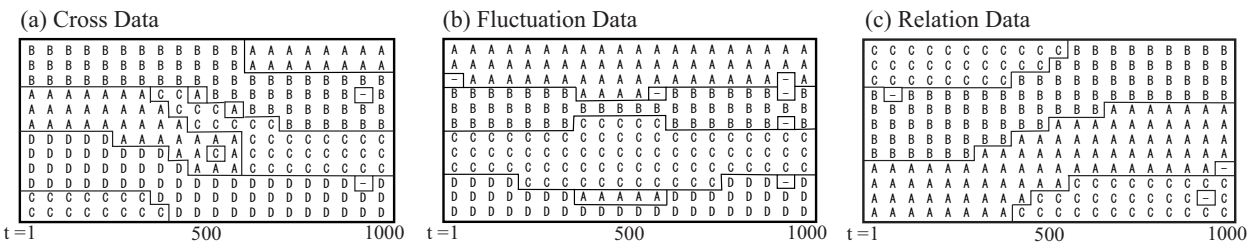


図 5 2 次元人工系列データの時系列可視化結果.

6·1 新聞記事データセットと前処理

新聞記事は，1993 年 1 月から 12 月の毎日新聞国際記事欄を用いた．この期間の記事数は，5,824 記事, 予め定めた Stop Word を除いた異なる単語の総数は，24,661 語であった．各記事は有意な単語を基底とした単語の重みベクトル (Bag-of-Words: BoW) で表現される．重みベクトルには，記事内の単語出現頻度 (term frequency) と，単語の特殊性を表す重み (inverse document frequency) の積 (tf-idf) を用いた．

これら記事群から，単語のベクトル空間内でトピックはある特定の方向に分布していると考え，主成分分析によりトピック軸 (主成分軸に相当) を抽出した [Kimura 05]．トピック軸は大まかな概念を表していると考えられる．主成分軸の重心からプラス方向とマイナス方向はそれぞれ別のトピックを表していると考えられるため，SbSOM への入力ベクトル $\mathbf{x}_n$ はプラス方向とマイナス方向で別の特徴とした．主成分数を L，$z_{n,l}$ を第 n 番目の記事の第 l 主成分得点（トピック成分と考えられる）とすると，$\mathbf{x}_n$ の第 i 要素は以下のように与えた．

$i = 2l - 1 (l = 1, \cdots, L)$ のとき

$$x_{n,i} = \begin{cases} z_{n,l} & (z_{n,l} > 0), \\ 0 & (z_{n,l} \leq 0). \end{cases} \tag{4}$$

$i = 2l$ のとき

$$x_{n,i} = \begin{cases} 0 & (z_{n,l} > 0), \\ |z_{n,l}| & (z_{n,l} \leq 0). \end{cases} \tag{5}$$

本実験では $L = 10$ とした．つまり SbSOM への入力ベクトルは 20 次元であり，第 1 から第 20 トピック成分とした．また，各記事はトピック成分が最大となる主成分番号（プラス・マイナス方向を区別）により推定トピックラベルを付与した．第 n 番目の記事のトピックラベル l_n は次式で与えられる．

$$l_n = \arg\max_i x_{n,i} \tag{6}$$

6·2 新聞記事データに対する結果

SbSOM による時系列可視化結果を図 6 および図 7・図 8 に示す．図 6 の横軸は時間軸[*2]を表しており，上下に目安となる月を記した．図中の数字は各クラスタ（サブトピックの最小単位に相当）の代表トピックを表しており，式 (6) により推定したトピックラベルの多数決で決定した．以降，このマップを全体マップと呼ぶことにする．

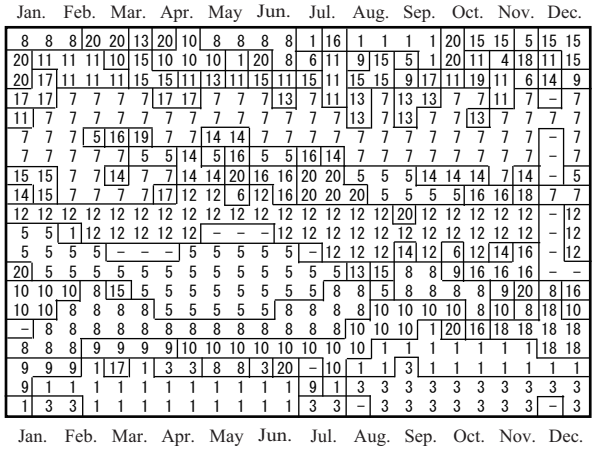


図 6 Sequence-based SOM による毎日新聞記事から抽出したトピックの時系列可視化結果（全体マップ）．数字は多数決で決定した各クラスタの代表トピック番号，“-” はそのノードに分類された記事がなかったことを示している．

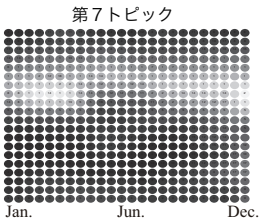


図 7 トピックの盛り上がりと衰退の例（成分マップ）．明るいほど，そのサブトピック内に該当トピック成分を多く含んでいる，つまりそのトピックのホット度合いを示している．

また図 7 と図 8 はトピック成分（参照ベクトルの要素）の分布を示しており，そのサブトピック内でのホット度合いに相当する．これを成分マップと呼ぶことにする．全体マップの数字ノードと成分マップの濃淡ノードは一対一に対応しており，色は明るいほどホット度合いが高いことを示している．これら 2 種類のマップを合わせてトピック推移マップと呼ぶことにする．

ニューロンのトポロジーは，24×20 の直交格子とし，SbSOM の順序付け重み関数は ψ_{exp} を用い，そのパラメータは $w = 150.0$ に設定した．

*2 順序重み付け関数により記事の順番は空間的距離が優先されて前後する可能性があり，また月により多少記事数の増減があるため厳密な意味での時間軸とはなっていない．

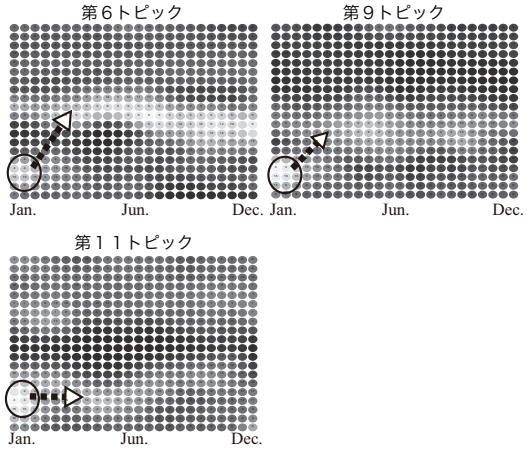


図 8 派生の例（成分マップ）．1 月に共通してホットになっているサブトピックから別々のサブトピックに派生していく様子が現れている．

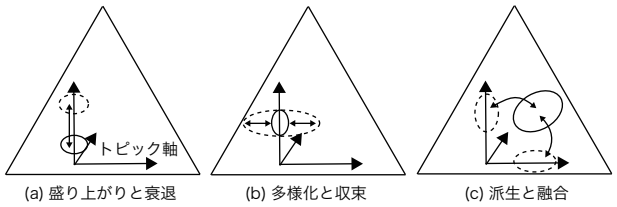


図 9 トピック推移と特徴空間中でのクラスタの変化との対応関係．(a) 盛り上がりと衰退はトピック軸方向の変化，(b) 多様化と収束はトピック軸に垂直方向の変化，(c) 派生と融合は複数のトピック成分の重ね合わせに対応している．円で囲まれた部分が記事が多く分布する範囲を示している．

以下に，トピック推移マップから抽象レベルで読み取れること，また実際の出来事とトピック推移マップとの対応についての 2 つの観点から具体例を挙げて説明する．

§1 マップから抽象レベルで読み取れること

トピックの生成と消滅 トピック推移マップにおいて，時空間的近傍がトポロジー上の近傍に現れる SbSOM の特性により，近い日付の近い内容の記事が時間軸方向に連なって現れている．この連なりの開始と終端はトピックの生成と消滅を表しており，例えば，全体マップからは第 3 トピックは 7 月初旬から出始めており，また第 5 トピックは 7 月終わりに消滅していることが見て取れる．

トピックの盛り上がりと衰退 あるトピック成分の量が多ければ，現在，そのトピックに関して盛り上がっていると考えられる (図 9(a))．全体マップからは読み取れず，成分マップを参照する．例えば，図 7 では，ホット度合いは 2 月中旬辺りで高くなり，一旦低くなり，再度 12 月に高くなっており，トピックの盛り上がりと衰退に対応していると考えられる．

トピックの多様化と収束 トピックの多様性は，トピック軸回りの記事群の広がりに対応していると考えられ (図 9(b))，人工データの Fluctuation Data に相当する．つまり，トピック推移マップ中の縦方向 (空間的近傍の意味を持つ方向) への広がりが，そのトピックの多様性に相当している．例えば，全体マップにおいて，第 5 トピックは 4 月中旬から 6 月中旬にかけて縦方向に広がり，つまり多様化し，その後収束している．

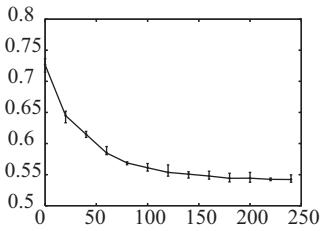


図 10 順序付けパラメータを変化させた時の MiP 値．10 回実行した平均値，最大値および最小値を記してある．

トピックの派生と融合 トピックの派生と融合は，複数のトピックがくっついたり離れたりすることを指す．これは，サブトピックの持つ複数のトピック成分の重ね合わせによって表現される (図 9(c))．人工データでは Cross Data に対応している．例えば，図 8 では，いずれのトピックも 1 月は近いサブトピックを持っているが，その後別々のサブトピックに分かれていくことが見て取れる．融合はこの逆である．

§2 実際の出来事との対応

次に，トピック推移マップと実際の出来事との対応について述べる．式 (6) により分類した記事群に対して，人手によりおおまかにトピックタイトルを付与した．トピックタイトルは著者ら以外の 2 名に付与してもらい，それらをまとめた．タイトル付与の方針は，分類された記事群から主要な出来事を抜き出し，それらを包含するタイトルを付けてもらった．また，複数のタイトルを並列させることも許容した．(付録 A にタイトルと抜き出した主要な出来事の一部を記載）斉藤らによるトピック抽出に関する認知科学的実験 [斉藤 05] では，表記ゆれ程度で実現できており，本実験のタイトルも頑健なものであると考えられる．ただし，主成分分析の性質上，下位のトピックほど焦点のぼけたトピックとなるため，少数のタイトルで表すのは難しくなる．それ故に，タイトルの信頼性は低くなるため，今回の考察では上位 10 トピック (5 主成分）程度を考察の対象とした．

このように作成したタイトルと主要な出来事を元に，トピック推移マップを読み解く．

トピックの盛り上がりと衰退 図 7 の第 7 トピック (中国情勢) について，3 月 27 日に新国家主席が選出されており，また 12 月 26 日は毛沢東誕生百周年記念であった．これらのサブトピックに関してそれぞれ盛り上がっていたと読み取れる．

トピックの多様化と収束 第 5 トピックはカンボジア総選挙に関するトピック (5 月 23 日〜25 日総選挙) で

あり，全体マップを見ると多様化しているが，総選挙後に新政権や新憲法のサブトピックが続き 7 月頃収束し消滅している．

トピックの派生と融合 第 9 トピックは米国の軍事・外交に関するトピックを含んでいる．主な出来事として 1 月 20 日にクリントン大統領が就任しており，図 8 の第 9 トピック，1 月辺りのホット度合いが高くなっているのが分かる．第 6 トピックの北朝鮮核問題は 6 月 2 日に米朝高官会談，第 11 のイラク問題は 1 月 23 日に米軍のイラク空爆があり，いづれのサブトピックもクリントン大統領就任のサブトピックから派生したと読み取れる．

6·3 順序付けパラメータの影響

§1 分類性能

クラスタ代表ラベルに対するクラスタ純度を Micro Averaged Precision(MiP) により評価した．MiP は次式により与えられる．

$$\mathrm{MiP} = \frac{\sum_j \#\{n : \mathbf{x}_n \in C_j, l_n = L_j\}}{N} \quad . \qquad (7)$$

ここで，C_j は第 j クラスタ (ニューロンに対応)，L_j は第 j クラスタの代表ラベル，l_n は $\mathbf{x}_n$ のラベルを表し，$\#$ は集合の要素数を表すものとする．完全に分類できていれば，すなわちどのクラスタにおいても一種類のラベルを持つデータのみ存在するならば，MiP 値は 1 となる．

順序付けパラメータ w を変化させて MiP 値を評価した (図 10)．$w = 0$ で通常 SOM となり，w が増加するに従って入力データの順序の入れ替わり許容が弱まっていく，すなわち勝者ニューロン選択の際に順序による制約が強まる．

図 10 より MiP 値は $w = 0$ の時，最も高く，w が増加するに従って急速に減少していくが，およそ $w = 200$ 以降ほぼ変化がなくなった．本実験の人工データのような単純な低次元データではない場合，可視化層への系列順序付けと分類性能はトレードオフの関係にある．実際の使用では，アプリケーションに応じて，どの程度系列の逆転を許容して空間的な解像度を得るかによって，w を調節することになる．

§2 視覚的変化

次に，図 11 と図 12 にパラメータを変化させた際の視覚的変化を示す．初期値によって得られる局所解の違いを排除するため，同じ初期値に対してパラメータを変えて実行した．どちらのマップでも $w = 100$ 辺りまでは視覚的にはほとんど変化が見られないが，$w = 50$ 以下では徐々に空間的距離が強く効いてきて，全体マップでは横に伸びていた同じ代表トピックのクラスタが縦方向に集まり始めてくる．$w = 0$ で通常 SOM になるため，横方向に時間的な意味はなくなり空間的距離のみを反映したマップとなる．図より，パラメータを変化させて空間的距離と時間的距離のバランスを取りながら，先に述べたような評価が同様に可能である．

7. 関 連 研 究

最後に，本研究が対象とする問題クラスとそれに関連する研究について述べる．

7·1 対象とする問題クラス

本研究では，次の 2 つの特徴を持つデータを対象としている．

(1) ベクトルオブジェクト 各オブジェクトは，多次元の特徴ベクトルとして表現される．ベクトルオブジェクトは，ある時点での対象物の特徴のことである．例えば，本稿の新聞記事データの場合は各記事から得られる単語頻度ベクトルに相当する．

(2) オブジェクト系列 オブジェクトデータは順次到着し，系列は時間的な順序関係を表している．データセットは単一のオブジェクト系列[*3]として得られる．

ある一定期間のデータは分布を成し，いくらかクラスタ (文書群ならばトピックに相当) が現れるが，その分布範囲やクラスタ間の関係が時間と共に変化していく動的データを対象としている．例えば，別々に発生した 2 つのトピックが融合したり派生したりする，などである．本研究の目的は，潜在するクラスの生成関数を求めるのではなく，観測されたデータのクラスタやその互いの関係の変化を視覚的に捉えることである．すなわち，系列を考慮したクラスタリングと可視化を行う必要がある．よって，本研究では次の 2 つのタスクを対象としている．

(3) 系列を考慮したクラスタリング ウィンドウ方式での適切なウィンドウ幅の設定，サンプル数の減少問題，クラスタ間の対応付け問題に対処したクラスタリング.

(4) クラスタ変化の可視化 クラスタの融合・分離，拡大·縮小など系列変化を反映した 2 次元マップの生成.

これらの特徴を持つ問題クラスに対して，我々は Sb-SOM を提案しているが，上記の 4 点を同時に扱っている先行研究はほとんどない．次小節以下に，関連研究について述べる．ここで，便宜的に系列を考慮しない静的データに対するクラスタリングと可視化をそれぞれ (3)', (4)' とする．

7·2 射 影 法

ベクトルデータを 2 次元平面へ射影する古典的方法として，多次元尺度構成法 (MDS)[Kruskal 78] や，Sammon's Mapping[Sammon 69]，もしくは主成分分析を用いて主

*3 従来よく研究されている時系列解析が対象とする時系列データは，ひとつのオブジェクトの時間変化であり，本研究では複数のオブジェクトが時系列で入力される点が異なっていることに注意しておく．

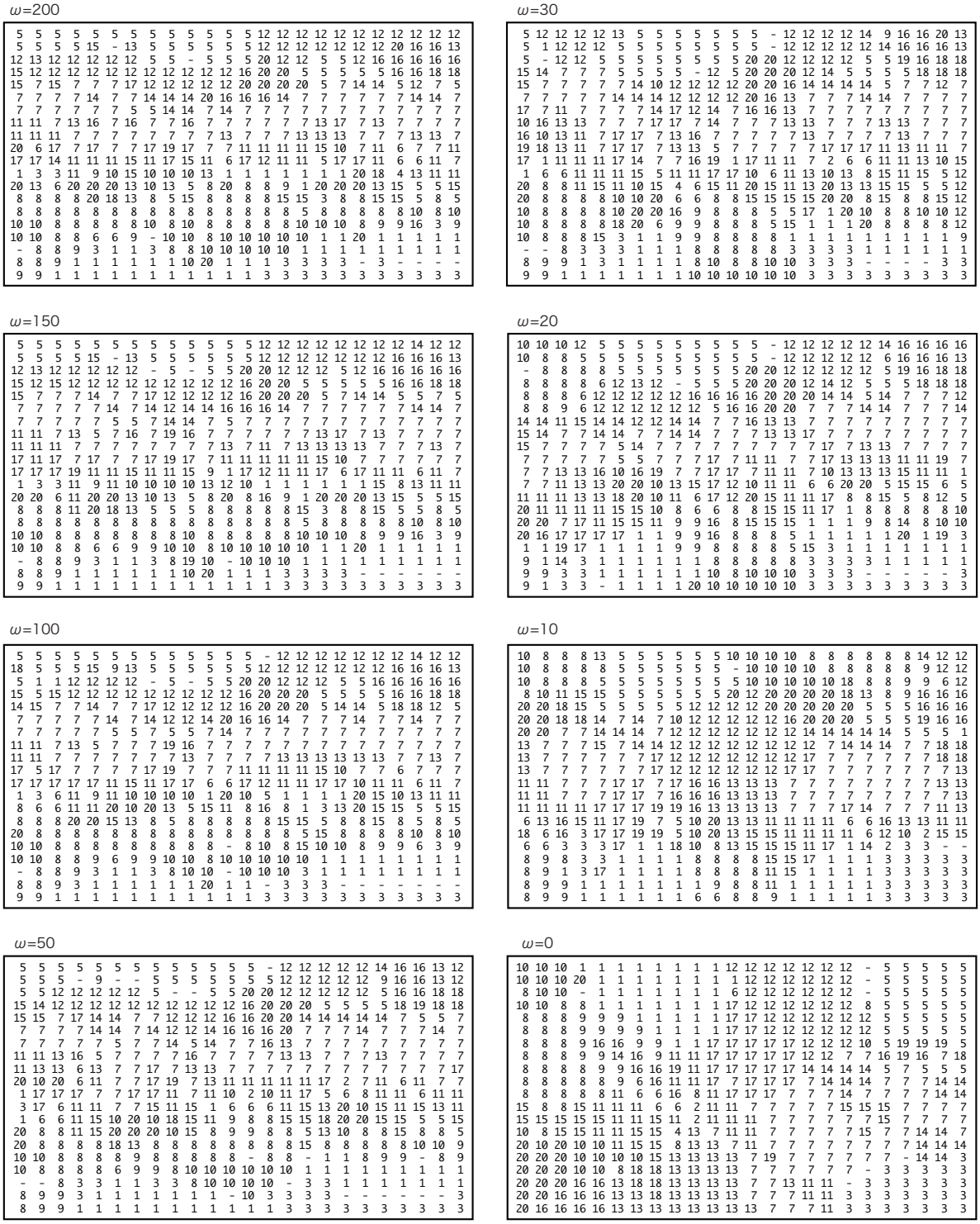


図 11 順序付けパラメータを変化させた時の視覚的変化 (全体マップ).

要な第一・二主成分で張られる平面に射影する方法が挙げられる．しかし，これらは全てデータひとつひとつの射影であり，クラスタリングによる汎化は行われないため，大規模なデータに適しているとは言い難い．また，これらの手法は通常，系列も考慮されないので，(1), (4)' に該当する．

7·3 系列データと可視化

(2) オブジェクト系列を扱い，(3)(4) 系列変化のクラスタリングと可視化を行っている研究として，Swan らの TimeMines[Swan 00] が挙げられる．Swan らは時系列文書群から χ^2 検定によりトピックを抽出・文書を分類し，トピック毎の出現期間とその強度を可視化している．しかし，Swan らの手法は文書データに特化しているため，(1) には該当せず汎用性があるとは言えない．

(1) から (4) に該当し汎用性のある研究に，K-means クラスタリングに時間的な逓減モデルを導入した T-Scroll[長谷川 07] が提案されているが，クラスタ間の対応付けはクラスタ間で共有するデータ数で閾値により判定しリンクで表している．適切な閾値の設定は難しい問題であり，また全てのクラスタ間で一定の閾値を用いてよいものなのかも不明である．それに対して本研究の手法は，閾値を用いることなく柔軟に対応できている．

7·4 SOM 学習モデルへの時間概念の導入

通常の SOM は，クラスタリングと可視化を同時に行うが時系列は考慮されないため，(1)，(3)'，(4)' に該当する．これに対して SOM の学習モデルに時間の概念を導入する研究に関しては様々な提案がなされている [Barreto 01]．しかし，これらはクラスタダイナミクスの可視化を目的としていなく，生理学的には短期・長期記憶のモデルとして導入されており，また時間伸縮 (Time Warping) や，系列パターンを学習するためのモデルであるため，本研究とは対象が異なる．

7·5 適用範囲の拡大

本稿では，入力は (1) ベクトルオブジェクトに限定されるが，近年研究が盛んに行われているカーネル法 [Bishop 06] を適用可能なカーネル SOM も開発されている [井口 05]．SbSOM も同様にカーネル化することが可能であり，これによりベクトルオブジェクトのみならず，グラフや文字列など様々な形式のデータに本手法は適用できると考えられる．

8. まとめと今後の課題

本稿では自己組織化ネットワークを基にした機械学習による可視化編纂の可能性を示した．動的なクラスタのダイナミクスを可視化する Sequence-based SOM について，対象とする問題クラスを明らかにし，2 次元の人工系列データおよび新聞記事データを用いて Sequence-based SOM の特性を確認した．本手法による時系列可視化結果からトピック推移が読み取れることを具体的に示した．本手法は文書データに限定されない学習手法を基にしているため適用範囲は広く，動的なクラスタの全貌を直感的に把握できるマップの自動編纂に有望な要素技術であると言える．

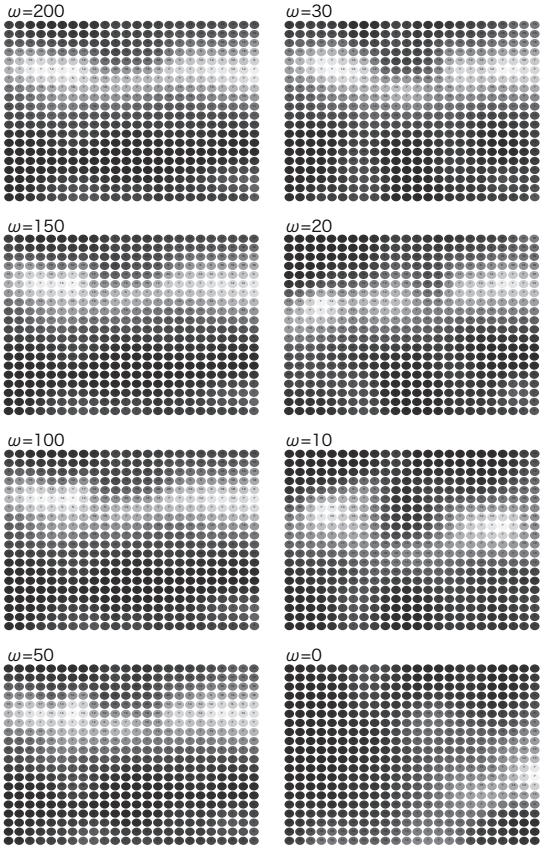


図 12 順序付けパラメータを変化させた時の視覚的変化 (成分マップ).

今後の展望としては，学習の観点からはカーネル化による適用範囲の拡大や，イベントとクラスタパターンとの関係の獲得（例えば，医療データであれば薬投与（イベント）と患者のクラスタパターンとの関係）が挙げられる．また，本稿での新聞記事の実験ではトピック推移マップの評価として，トピックに関する知識をある程度持っている状態ではあるがマップから抽象レベルで分かることと，具体的な出来事との対応関係について考察した．今後，トピックに関する知識をほとんど持たない人によるマップの読み取りと，トピックの具体的な出来事からのマップの解釈の 2 つの観点から被験者実験を行う必要がある．さらに，本稿では可視化までに留まっているが，ユーザとのインタラクションの仕方の検討，ユーザインターフェースを含めた総合評価などが考えられる．

謝　辞

本研究は文部科学省特別教育研究経費の補助を受けて行われた．

◇ 参 考 文 献 ◇

[Barreto 01] Barreto, G. and Arajo, A.: Time in Self-Organizing Maps: An Overview of Models, *International journal of Computer Research, Special Issue on Neural Networks*, Vol. 10, No. 2, pp. 139–179 (2001)

[Bishop 06] Bishop, C. M.: *Pattern Recognition and Machine Learning*, chapter 6, Springer-Verlag (2006)

[Fukui 05] Fukui, K., Saito, K., Kimura, M., and Numao, M.: Visualizing Dynamics of the Hot Topics Using Sequence-Based Self-Organizing Maps, *Lecture Notes in Artificial Intelligence*, Vol. 3684, pp. 745–751 (2005)

[Fukui 06] Fukui, K., Saito, K., Kimura, M., and Numao, M.: Visualization Architecture Based on SOM for Two-Class Sequential Data, *Lecture Notes in Artificial Intelligence*, Vol. 4252, pp. 929–936 (2006)

[Fukui 07] Fukui, K., Sato, K., Mizusaki, J., Saito, K., and Numao, M.: Combining Burst Extraction Method and Sequence-based SOM for Evaluation of Fracture Dynamics in Solid Oxide Fuel Cell, in *Proc. of The 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Vol. 2, pp. 193–196 (2007)

[Kimura 05] Kimura, M., Saito, K., and Ueda, N.: Multinominal PCA for extracting major latent topics from document streams, in *Proceedings of 2005 International Joint Conference on Neural Networks*, pp. 238–243 (2005)

[Kohonen 95] Kohonen, T.: *Self-Organizing Maps*, Springer-Verlag (1995)

[Kruskal 78] Kruskal, J. B. and Wish, M.: *Multidimensional Scaling, Quantitative Applications in the Social Sciences* (1978)

[Sammon 69] Sammon, J.: A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, Vol. c-18, pp. 401–409 (1969)

[Swan 00] Swan, R. and Jensen, D.: TimeMines: Constructing Timelines with Statistical Models of Word Usage, in *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 73–80 (2000)

[井口 05] 井口 亮, 宮本 定明：カーネル関数を利用した LVQ クラスタリングと SOM, 日本知能情報ファジィ学会誌, Vol. 17, No. 1, pp. 88–94 (2005)

[加藤 06] 加藤 恒昭, 松下 光範：情報編纂 (Information Compilation) の基盤技術, 第 20 回人工知能学会全国大会論文集 (2006)

[斉藤 05] 斉藤 和巳, 木村 昌弘, 上田 修功：文書トピックに関する認知科学的実験, 人工知能学会 第 69 回知識ベースシステム研究会資料, pp. 57–62 (2005)

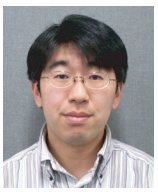
[長谷川 07] 長谷川 幹根, 石川 佳治：T-Scroll: 時間的トピックの推移をとらえる可視化システム, 日本データベース学会 Letters, Vol. 6, No. 1, pp. 149–152 (2007)

[福井 07] 福井 健一, 佐藤 一永, 水崎 純一郎, 斉藤 和巳, 沼尾 正行：固体酸化物燃料電池における破壊ダイナミクスの可視化法, 情報科学技術レターズ, Vol. 6, pp. 5–8 (2007)

〔担当委員：高間 康史〕

2007 年 11 月 30 日　受理

著 者 紹 介

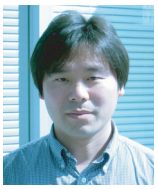


福井　健一（正会員）

2001 年 名古屋大学情報文化学部自然情報学科中退（飛び級進学）．2003 年 名古屋大学大学院人間情報学研究科物質・生命情報学専攻 修士課程修了．2005 年より大阪大学 産業科学研究所 新産業創造物質基盤技術研究センター特任助手．現在，同センター特任助教．機械学習，データマイニング等に興味を持つ．人工知能学会，情報処理学会会員．

斉藤　和巳（正会員）

1963 年生まれ．1985 年慶応義塾大学理工学部数理科学科卒業．同年日本電信電話株式会社入社．2007 年静岡県立大学経営情報学部教授．工学博士．学習アルゴリズムの研究に興味を持つ．情報処理学会，電子情報通信学会，日本神経回路学会各会員．



木村　昌弘（正会員）

1989 年大阪大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．現在，龍谷大学 理工学部電子情報学科 准教授．博士（理学）．ニューラルコンピュテーション，複雑系の数理モデリング及び数理解析，Web マイニングの研究に興味をもつ．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本数学会各会員．

沼尾　正行（正会員）

1982 年東京工業大学工学部電気電子工学科卒業．1987 年同大学院情報工学専攻博士課程修了．工学博士．東京工業大学大学院情報理工学研究科計算工学専攻助教授を経て，2003 年より大阪大学産業科学研究所教授．1989〜90 年スタンフォード大学 CSLI 客員研究員．人工知能，機械学習，関数型言語などの研究に従事．人工知能学会，日本認知科学会，日本ソフトウェア科学会，電子情報通信学会，AAAI，ACM 各会員．

◇ 付　　録 ◇

A. トピックタイトル一覧

表 A.1 人手によるトピックタイトルと主な出来事（一部）.

No.	記事数	トピック-主なサブトピック	主な出来事
1	453	ロシア情勢 -大統領と議会の対立	6 月 5 日〜 ロシア憲法会議 9 月 21 日 議会解体令
2	32	中東和平問題	9 月 13 日 オスロ合意
3	303	ロシア大統領解任事件 中東和平問題 -中東和平に対する各国の対応	10 月 4 日 モスクワ暴動
4	43	カンボジア総選挙 -ポルポト派の動向	2 月 4 日 鉄道爆破事件 3 月 4 日 ポルポト派が政府軍砲撃
5	471	カンボジア総選挙 -総選挙から連合政権設立まで	4 月 7 日〜5 月 19 日 選挙運動期間 5 月 23 日〜25 日 総選挙
6	148	北朝鮮核問題 -米軍の動向	6 月 10 日 米朝高官会談 8 月 3 日 IAEA 特定査察再開
7	721	中国情勢 -政権交代 -社会主義経済へ	3 月 15 日 憲法修正を発表 3 月 27 日 国家主席選出 12 月 26 日 毛沢東誕生百周年記念
8	452	ボスニア紛争問題 -和平交渉と米軍の介入 -ソマリア内戦	1 月 2 日 初の直接和平交渉開始 1 月 4 日 ソマリア和平会議開始
9	171	米国の軍事/外交	1 月 20 日 クリントン大統領就任 9 月 24 日 エリツィン大統領支持を表明
10	290	ボスニア紛争問題 -イスラム勢力の動向	6 月 18 日 イスラム勢力 3 分割案拒否 9 月 1 日 和平協議決裂
11	366	イラク問題 ボスニア紛争問題 -セルビア人勢力の動向	1 月 23 日 米軍のイラク攻撃 1 月 30 日 セルビア人勢力サラエボ砲撃
12	436	北朝鮮問題 -核開発問題	2 月 9 日 特別査察要請
13	345	日本と諸外国 -戦後補償問題	8 月 国連人権委員会
14	230	カンボジア新政権 中国情勢 パキスタン総選挙	9 月 24 日 新憲法発布 10 月 6 日 パキスタン総選挙
15	327	北朝鮮政権交代 中国政権交代	4 月 9 日 金正日国防委員長に
16	260	首脳会談と米国の動向 EC 統合	11 月 17 日 APEC 初の首脳会談開催 11 月 1 日 EC を元に EU が正式発足
17	253	中国の外交 -香港返還問題	6 月 21 日 香港返還をめぐる中英交渉 12 月 18 日 中台協議 台湾で初の開催
18	128	ロシア情勢 ソマリア情勢	12 月 12 日 新憲法国民投票
19	127	世界各国の選挙・投票	3 月 21 日 フランス総選挙 5 月 15 日 セルビア人住民投票
20	268	世界各国の会議・会談	6 月 21 日 EC 首脳会議 7 月 23 日 ASEAN 外相会議