3K3-4

自己組織化ネットワークによるクラスタ変遷の可視化

Visualization of Cluster Transition utilizing Self-Organizing Network

福井 健一[*1] 斉藤 和巳[*2] 木村 昌弘[*3] 沼尾 正行[*1]
Ken-ichi Fukui Kazumi Saito Masahiro Kimura Masayuki Numao

[*1]大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka Univerisity

[*2]静岡県立大学 経営情報学部
School of Administration and Informatics, University of Shizuoka

[*3]龍谷大学 理工学部
Department of Electronics and Informatics, Ryukoku University

We have proposed neural-network based visualization approach, called Sequence-based SOM (Self-Organizing Map) that visualizes transition of dynamic clusters by introducing the sequencing weight function onto the neuron topology. This approach mitigates the problems with a sliding window-based method. This paper presents the topic's separation and merger using a real news articles data set. Moreover, we can conclude this visualization result is consistent among subjects. Visualization of cluster transition 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変遷，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，医療情報からの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．

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．従来の単純なウィンドウ方式と比べて，ウィンドウ幅の設定が不要，ウィンドウ内のサンプル数減少の軽減，またウィンドウ間のクラスタの対応付け問題がなくなるメリットがある．

本稿では，SbSOM によるクラスタ変遷の可視化応用例として，新聞記事データから抽出したトピックにおいて，SbSOM 学習結果に後処理を行うことで，関連トピックの派生と融合が抽出でき，その様子が可視化されることを示す．これにより，自己組織化ネットワークに基づく可視化編纂の可能性を探る．

連絡先: 福井健一，大阪大学産業科学研究所，大阪府茨木市美穂ヶ丘 8-1，Tel:06-6879-8426，fukui@ai.sanken.osaka-u.ac.jp

2. Sequence-based SOM

Sequence-based SOM のアルゴリズムは通常の SOM と同様であるが，勝者ニューロン決定の距離定義に時間的距離を考慮するために，順序重み付け関数を導入している点が異なる．

2.1 アルゴリズム

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．

以下に，一般によく用いられている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，次式のガウス関数がよ

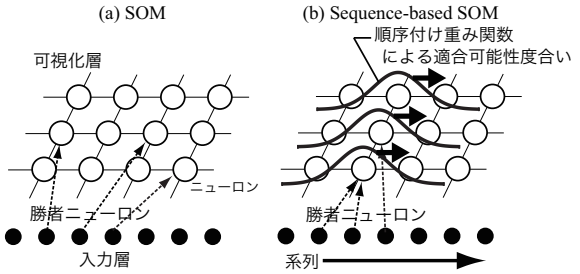

図 1: 勝者ニューロン選択の違い．(a) SOM は空間的距離のみにより決定される, (b) Sequence-based SOM では，順序付け重み関数の下で空間的距離により決定される．順序付け重み関数は入力データの系列に従って可視化層トポロジー上の一方向に移動していく．

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．

2.2 順序付け重み関数

通常 SOM では，(特徴) 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ことで，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．

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

$\psi(n, \xi_j)$ 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．

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，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*1 ことで両者のバランスを取る．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^* の割合に位置している．ここで，$\xi^* = \max_j \xi_j$ とする．その差の絶対値を $\epsilon = |\xi_j/\xi^* \text{-} n/N|$ とおく．順序付け重み関数は次式により与える．

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

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 1 参照)．また，$w = 0$，すなわち $\psi(n, \xi_j) = 1$ のとき，通常 SOM になる．

2.3 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・空間的近傍の意味合いを持つことになる (図 2)．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．

*1 これにより緩い時間となっていることに注意する．

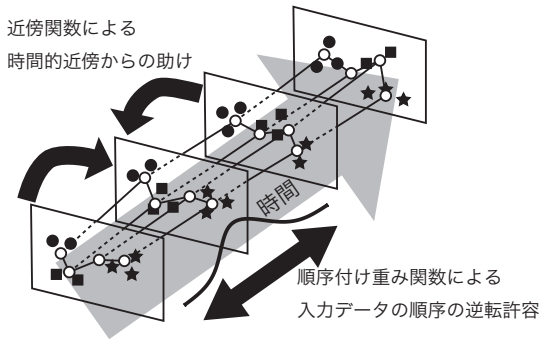

図 2: 特徴空間とニューロンの可視化層トポロジーの関係の概念図．○がニューロンを表している．Sequence-based SOM では時間的・空間的近傍がトポロジー上の近傍になる．ただし，ψ_{exp} では時間は図のようなウィンドウ方式とはなっていない．

3. 応用例：ニューストピック変遷の可視化

3.1 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（トピック成分量と考えられる）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

$$x_{n,i} = \begin{cases} z_{n,l} & (z_{n,l} > 0), \\ 0 & (z_{n,l} \leq 0). \end{cases} \quad (3)$$

$i = 2l$ のとき

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

本実験では，$L = 10$ すなわち，全 20 トピック抽出した (SbSOM への入力は 20 次元)．また，SbSOM のニューロンの可視化層トポロジーは，24×20 の六角格子とし，順序付けパラメータ $w = 200.0$ とした．

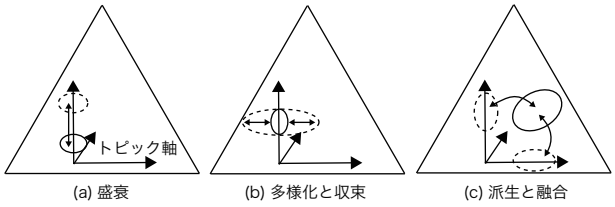

図 3: トピック変遷と特徴空間中でのクラスタの変化との関係概念図．トピックの (a) 盛衰はトピック軸方向の変化，(b) 多様化と収束はトピック軸に垂直方向の変化，(c) 派生と融合は複数のトピック成分の重ね合わせに対応している．円で囲まれた部分が記事が多く分布する範囲を示している．

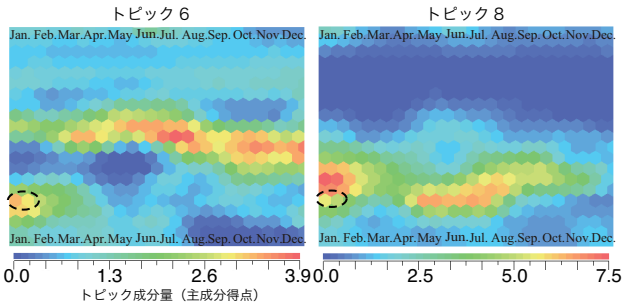

図 4: トピック変遷の例．トピックの盛衰はノード値 (色) により，多様性は縦方向の広がりにより，またトピックの融合は同一ノード上で複数のトピック成分を持つ部分 (図中では点線部分) により表現されている．

3.2 トピック変遷と SbSOM の特性

我々は人工データを用いた実験により SbSOM の特性を明らかにしており，次の 3 つのクラスタ変遷 (ここではトピック変遷) を可視化できることを確認している [Fukui 07].

(a) 盛衰 主成分得点はトピック成分の含有量であるため，トピック軸方向の変化 (図 3(a)) は，トピックの盛衰を表していると言える．トピック毎の成分マップ (図 4) 上では，ノードの値，すなわち色によって表される．

(b) 多様化と収束 トピック軸と垂直方向の変化 (図 3(b)) は含まれる単語の種類の多様性を表している．マップ上ではホットな領域の縦方向の広がりに現れる (図 4).

(c) 派生と融合 複数のトピック成分の重ね合わせ (図 3(c)) は，トピックの融合と言える．マップ上ではノード (クラスタ) はサブトピックの最小単位である．同一ノードで複数のトピック成分マップ上で成分量が多い部分は，メインのトピックを両方含むような融合したサブトピックであると考えられる (図 4).

3.3 後処理

本稿では複数のトピック成分が重なり合っているトピックの融合部分に着目し，閾値を設定することで関連トピックを判別し，それら関連トピックの合成マップを生成した．

3.3.1 関連トピックの判定

同一ノード上で複数のトピック成分が閾値 θ_g 以上であれば，それらトピックはそのノードが示すサブトピックにおいて関連していると考えられる．関連トピックの判定とそれらの合成

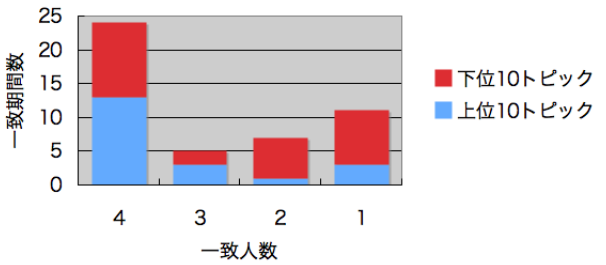

図 5: 被験者実験による一致ホット期間数．

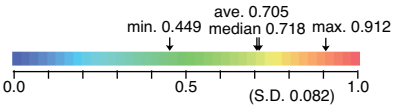

図 6: 被験者実験によるガイド閾値の決定．

マップ (new) の k 番目のノード値 $m_{k,new}$ は，参照ベクトルの要素 $m_{k,i}$ と $m_{k,j}$ を用いて次式により与えた．

$$\forall\, i,j\ (j>i)\ \text{s.t.}\ \exists k\ m'_{i,k}>\theta_g \text{ and } m'_{j,k}>\theta_g$$
$$m_{k,new}=m_{k,i}+m_{k,j}. \qquad (5)$$

ここで，m' は $m'\in[0,1]$ になるように $\max_k m_{i,k}$ で除して正規化した値である．

3.3.2 ガイド閾値の決定

関連トピック判定の閾値は被験者実験により決定した (被験者は大学院生 4 名)．図 4 に示すような 20 枚のトピック毎のマップを見て，トピック毎に盛り上がっているホットだと思う期間と判断した閾値を答えてもらった．

ここで，この閾値は “実際にホットな期間” のしきいとは異なることに注意しておく．人がマップを見た時に “ホットだと思う色” を示す目安となる閾値であるので，ガイド閾値と呼ぶことにする．実際にホットな期間が分かれば，この閾値により示された色をその期間に合うように調節すれば良い．この閾値は同じカラーセット*2を用いている限りにおいて有効である．

4. 結果と考察

4.1 被験者実験結果

図 5 に被験者がホットであると答えた期間の一致人数毎の数を示す．開始と終わりがそれぞれ 1ヶ月以内のずれであれば一致していると見なした．3 名以上一致の割合は，全 20 トピック中で 61%(4 名一致 51%)，また主成分分析の特性上，焦点のぼやけたトピックが得られてしまう下位 10 トピックを除いた上位 10 トピックに限っては，80%(4 名一致 61%) に達した．この結果は，多くの人の判断が一致する可視化結果が得られていることを示している．

次に，判断した閾値の統計量を図 6 に示す．トピック毎に成分量が $[0,1]$ になるように正規化を行っている．図より，ガイド閾値 $\theta_g=0.7$ とした．標準偏差は 0.082 と小さく，ここでも多くの被験者の一致が見られた．また，0.7 辺りから暖色系の黄色になっていることは，その判断の裏付けとなる．

4.2 トピックの派生と融合

式 (5) により判定した関連トピックとその期間の一部を表 1 に示す．関連トピックは全 41 組，54 期間抽出された．

*2 本実験では Mindware 社の SOMine により描画している

表 1: 上位 10 トピック間の関連トピックとその期間 ($\theta_g = 0.7$)

関連トピック	期間
T2,T3	8 月〜10 月中旬, 11 月中旬〜12 月中旬
T4,T5	3 月中旬〜5 月中旬
T4,T9	5 月
T5,T9	5 月
T6,T8	1 月中旬
T6,T9	1 月
T8,T9	1 月中旬
T8,T10	1 月, 5 月〜5 月中旬, 6 月中旬〜8 月中旬

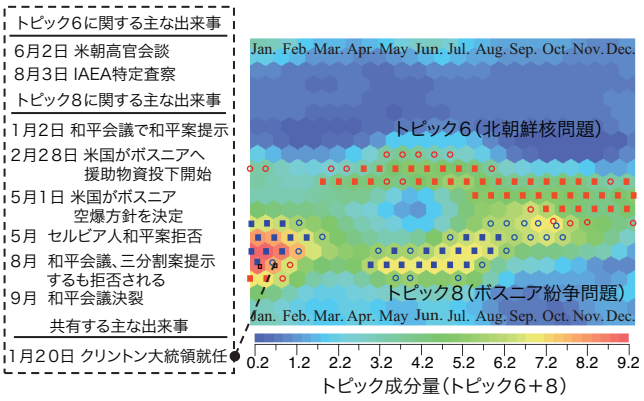

図 7: トピックの派生と融合の例．トピック 6 と 8 の合成マップと主な出来事．それぞれのトピックで■は $\theta_g = 0.7$，○は $\theta_g = 0.6$ 以上のホットなサブトピックを示している．

また，一例として図 4 のトピック 6 と 8 の合成マップを図 7 に示す．1 月・2 月に共通するサブトピックからそれぞれ派生し，10 月頃にまた融合する様子が見て取れる．共有する出来事として，1 月 20 日にクリントン大統領が就任している．トピック 6 は「北朝鮮核問題」に関する記事が多く，特にそれに対する米国の動向に関する記事が含まれていた．一方，トピック 8 は「ボスニア紛争問題」に関する記事が多く，トピック 8 もまた米国のボスニア紛争への軍事介入に関する記事が含まれていた．このマップからもクリントン政権の両問題への関与がうかがえる．これら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．

5. 関連研究

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，これらはクラスタ変遷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．

Swan ら [Swan 02] 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する TimeMines を提案している．しかし，TimeMines は文書データに特化しているため汎用性があるとは言えない．それに対して我々の手法は，汎用性のある機械学習法に基づいている．

また，長谷川ら [Hasegawa 07] は 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 を提案している．T-Scroll では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．

6. おわりに

本稿では新聞記事から抽出したトピックの変遷を例に，自己組織化ネットワークを基にした汎用性のある学習手法による可視化編纂の可能性を示した．SbSOM 学習結果において，人がトピックのホット期間を判別する際にガイドとなる閾値を設定することで関連トピックを判定し，ひとつのマップ上にトピックの派生や融合が可視化されることを確認した．また，ガイド閾値を決定する被験者実験の結果は，多くの人々の判断が一致するマップが生成できていることも示している．

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．また利用者の観点からは，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により行われた．

参考文献

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

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

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

[Fukui 07] 福井 健一, 斉藤 和巳, 木村 昌弘, 沼尾 正行: クラスタのダイナミクスを可視化する Sequence-based SOM に関する一考察, 人工知能学会 第 4 回データマイニングと統計数理研究会, on CD-ROM (2007).

[Barreto 01] G. A. Barreto and A. F. R. Arajo: 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).

[Swan 02] R. Swan and D. Jensen: TimeMines: Constructing Timelines with Statistical Models of Word Usage, *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 73-80 (2000).

[Hasegawa 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).