Vol.2011-IFAT-102 No.3
Vol.2011-DD-80 No.3
2011/3/28

ブログ記事のクラスター構造と経時変化の可視化

石　川　雅　弘[†1]

ウェブ上にはブログをはじめとして一般ユーザにより生産された大量のテキストデータが蓄積されており，その量は今後も継続的に増加すると考えられる．我々はタイムスタンプ付きテキスト集合のクラスター構造とその経時変化を可視化するための手法を提案してきたが，そこでは文書ベクトルの次元削減と潜在意味処理を行なうために全データを一括して処理する必要があった．そのため，ブログなど大量のテキストが生産される続ける漸増的環境に適用するには問題があった．本稿では，ブログのような漸増的動的環境下でも，潜在意味処理を伴なったクラスタリングと可視化を効率良く行うための、文書ベクトル生成手法を提案し，例として収集したブログ記事集合への適用結果を示す．

Visualizing Cluster Structures of Blog Articles and Their Changes over Time

Masahiro ISHIKAWA[†1]

Nowadays, huge amount of user generated texts is produced and accumulated on the web. They will be continuously increased in the future too. We have proposed a method for visualizing cluster structures of time-stamped texts and their changes over time. However, in the method, the whole dataset had to be processed at a time for dimension reduction of document vectors and incorporating latent semantics of words. Thus, the method have some problem in dynamic incremental environments, such as blogs, to apply. In this paper, a new method for document vector generation which can enable efficient text clustering and visualization in dynamic environments is proposed. As an example, the proposed method is applied to blog articles to demonstrate its effectiveness.

†1 つくば国際大学
Tsukuba International University

1. はじめに

ブログや SNS などのソーシャルメディアは，既存マスメディアと異なり誰でも情報発信者となることができ，個々人の直接的意見表明や結び付きを支える道具として成長してきた．日記的なものからリンクやトラックバックによるつながり，あるいはマイクロブログなどのリアルタイム性の高いものなど，その形態も進化を続けている．また画像・音声の利用などマルチメディア性も高まってきた．しかし人々が意見を表明し他人とコミュニケートする際に最も重要な表現メディアは依然として文字による文章であり，それはこれからも変わらないであろう．PC だけではなく携帯機器にも拡がったソーシャルメディアの浸透により，ウェブ上に蓄積されるテキストはこれからも増加し続けると考えられる．このようにして個人個人が自由な立場で記述した大量のテキストには，その時々の各々の関心や意見，あるいは社会情勢が反映されていると考えられ，それらを分析することで伝統的・形式的なアンケートでは得るのが難しかった情報をも獲得できる可能性がある．このような考えから，例えばブログ記事の市場分析への利用などが試みられている．大量のデータを分析する際には，その概観を把握し分析者をナビゲートするためにクラスタリングが行なわれる事が多い．しかし，例えば市場分析者にとっては，クラスタリング結果が数値的に示されただけではそれを有効に活用するのは難しい．各分野の専門家がそれを有効に活用するためには，クラスタリング結果の可視化が重要である．また，ブログのように時間と共に刻々と増え続けるデータの場合，クラスター構造もまた変化していると考えられ，その変化をも可視化することが望まれる．

このような観点から，時系列テキスト集合のクラスター構造とその経時変化を可視化するために SOM(Self-Organizing Map)[1] を利用したクラスタリング・可視化手法を提案してきた．[2][3] しかしそこでは，文書ベクトルの次元削減と語の潜在意味処理のために Random Projection[4] と LSI(Latent Semantic Indexing)[5] を適用しており，LSI の実行において全データを一括処理する必要があった．そのためブログのような漸増的環境下で継続的に適用するには問題が残っていた．

本稿では，漸増的環境下でより効率的にクラスタリングと可視化を行なうための，Random Indexing[6][7] を応用した文書のベクトル表現手法を提案する．これにより Random Projection と LSI の組み合わせで行なっていた，比較的低次元のベクトルによる文書の表現と潜在意味処理を可能とし，同時に漸増的環境に対応した．可視化については論文 3) で提案した Time-Arrayed SOM をそのまま用いており，本稿の焦点は新たに導入した文書ベ

Vol.2011-IFAT-102 No.3
Vol.2011-DD-80 No.3
2011/3/28

クトルの生成手法である．最後に，提案手法の適用例として，ブログ記事集合の可視化例を示す．

2. 問題設定

2.1 目　　標

本研究の目標は，ブログのように継続的に生産され続けるタイムスタンプの付随した文書を対象とした，以下の要件を満たすクラスタリング・可視化手法を提案することである．

(1)　クラスタ構造の全体と部分の関係およびその経時変化の可視化
(2)　誤字・脱字・同義語などを含めた潜在意味の取り込み
(3)　単語の意味の経時変化の反映
(4)　スケーラビリティの確保
(5)　動的環境への適応

以下でそれぞれの意図を説明する．なお，クラスタリングと可視化については既に論文 3) で提案した手法をそのまま用いるため，本稿の焦点はその入力となる文書ベクトルの生成手法の提案にある．

文書検索や文書クラスタリングでは，文書を高次元の数値ベクトルとして表現し，ベクトル間に定義した類似度や距離で文書同士の類似性を測るベクトル空間モデルを用いるのが一般的である．[9] しかし高次元空間では，次元数と共に指数的に拡大するデータ空間の容積に対して，実データの分布は疎であり，空間全体を意識するよりも実データの分布のみに集中した方が効果的な可視化が行えると考えられる．データの分布を学習する多様体学習を利用した可視化も同様の考えに基づくと考えられる．この場合，各クラスターの位置は，データ分布全体，あるいは他のクラスターとの相対的なものとして示す事が望ましく，これが要件 (1) の意図である．

要件 (2) は，単語の表層形が異なっても意味的に類似した語同士には類似性を認めるためである．職業記者が統制下で作成し校閲を経てリリースされるマスメディアのテキストとは異なり，ソーシャルメディア上のテキストは様々な背景の一般個人が自由な環境で作成する．したがって多様な同義語，誤字・脱字，当て字，俗語・隠語など，表記のゆれが大きいと考えられる．このような場合，表層形の異なる語や，場合によっては誤字・脱字，当て字であっても，意味の類似した単語間には類似度を設定するのが望ましい．また，マイクロブログなどでは最大で 140 文字などの文字数制限のあるものもある．極端に短かい文章では，表層形の比較だけでは，実際の意味は同じであっても類似性無しと判定されてしまう可能性が高まると考えられ，語の潜在意味処理の重要性は一層大きいと考えられる．

単語の意味自体も，流行や新たな語義の獲得などにより時間と共に変化するであろう．例えば “twitter” や「つぶやき」という単語はマイクロブログの Twitter 登場以前から存在しているが，Twitter 普及前後の使用では内包自体が変化した可能性がある．また,「首相」と言ってもその時々で指し示される個人は異なるように，同じ語の出現でも類似性を認めるべき語が変化することがあり得る．ソーシャルメディアという非統制下のメディアを長いスパンで捉えようとした時，このような単語の意味の変化にも対応できることが望ましく，これが要件 (3) の意図である．

少数の職業記者から多数の一般個人に書き手が拡がった事で，生産されるテキストの量も爆発的に増加した．ウェブ上には既に新聞一紙の数百年分に相当するブログ記事が蓄積されていると見られ，その膨大なデータを扱うにはスケーラビリティの確保が必要である．

そして，ニュース記事やブログ記事は，日々増加し続けている．過去から現在までの全ての記事のクラスター分析を終了したとしても，また明日には新たなテキストが生産されており，それらを含めた分析が必要となる．このような環境下では，そのたびに全データを一括して処理する必要がある手法では継続的な適用は困難である．

2.2 本稿の課題

ベクトル空間モデルでは，文書は単語の異なり総数を次元数とする超高次元ベクトルで表現される．そのため，後段の処理の空間および時間コストが増大し，扱うデータ量が大きい場合には実行が困難になることがあり，文書ベクトルの次元削減が必要となることが多い．

論文 3) で示したテキストコレクションのクラスタリング・可視化手法では，全てのデータを一括処理できるという前提で Random Projection と LSI を適用することで次元の削減を行なった．また，LSI の適用により潜在意味処理も実現していた．しかし LSI の適用には全データの一括処理が必要であるため，漸増的環境への適用，すなわち要件 (5) の実現が難しいという問題が残った．

本稿の課題は，その問題を解決し，全ての要件を満たすような文書ベクトル生成手法を提案することである．

3. 提案手法

3.1 Random Indexing

本稿で提案する文書ベクトル生成手法は Random Indexing に基いているため，ここで例を用いて簡単に説明する．

Vol.2011-IFAT-102 No.3
Vol.2011-DD-80 No.3
2011/3/28

Random Indexing は，単語の意味の分布仮説に基づく単語ベクトル生成手法である[6]．ある単語の意味は，その出現文脈に現れる他の単語群により決定される．例として，次のような文章があり各語には語の索引ベクトルと呼ばれるランダムなベクトルが割り当てられているとする．

私　は　旅行　に　行き　たい

この時，各語の前後 k 語の範囲をその語の文脈とし，文脈中の各語の索引ベクトルを合計することでその単語のこの出現における文脈ベクトルを得る．例えば「旅行」の前後 2 語を文脈とすると,「私」,「は」,「に」,「行き」の索引ベクトルの合計が「旅行」のこの出現における文脈ベクトルである．全文書における全出現の文脈ベクトルを合計することで，その単語のベクトル表現が得られる．結果として，類似した文脈に出現する語は類似したベクトル表現を得る．ここで索引ベクトルの次元を単語の異なり総数より小さな値とすることで次元削減が実現される．なお，ここで用いるランダムベクトルは一定の条件を満たす必要があり[7]，本稿では論文 10) で提案された手法を用いている．

3.2 Random Indexing を利用した文書ベクトルの生成

提案手法では，モデル化と処理を単純化するため，各語の文脈をその語が含まれる文書全体とする．すなわち，単語 i の文書 j における出現数 $d_{i,j}$ を要素とする行列を $\mathbf{D}$，各列ベクトル $\mathbf{r}_i$ が単語 i の索引ベクトルである行列を $\mathbf{R}$ とすると，文脈ベクトルは下式で与えられる．

$$\mathbf{C} = \mathbf{RD}$$

行列 $\mathbf{C}$ の各列 $\mathbf{c}_i$ が，文書 i に出現する全ての語に共通する，文書 i における文脈ベクトルである．s_j を文書 j に含まれる単語数，$\mathbf{G}$ を以下のような要素 $g_{i,j}$ を持つ行列とする．

$$g_{i,j} = \begin{cases} \frac{1}{s_j} & \text{(単語 i が文書 j に出現している場合)} \\ 0 & \text{(それ以外)} \end{cases}$$

この時，最終的な単語ベクトルは，下式で与えられる．

$$\mathbf{W} = \boldsymbol{N}(\mathbf{RDG}^{\mathrm{T}})$$

ただし $\boldsymbol{N}(\cdot)$ は行列の列ベクトルを長さ 1 に正規化する演算子とする．

行列 $\mathbf{W}$ の列ベクトル $\mathbf{w}_i$ が，単語 i の出現文脈に基いた単語ベクトルであり，単語が出現する文脈を重み付きで合計したものである．類似した文脈で出現する単語同士は，例え同じ文書に出現することはなくとも，類似したベクトルとなる事が期待できる．例えば正しい表記とありがちな誤記は，同じ文書の中に同時に出現することは少ないかもしれないが，出現するのは類似した文脈であろう．そのような場合でも類似性を見出す事ができるため，誤字・脱字だけでなく当て字や隠語の場合にも適切な類似度を設定できると期待できる．

最後に，下式で文書の意味ベクトルを得る．

$$\mathbf{X} = \boldsymbol{N}(\mathbf{WH})$$

ただし $\mathbf{H}$ の各要素は下式で与えられる．

$$h_{i,j} = \begin{cases} 1 & \text{(単語 i が文書 j に出現している場合)} \\ 0 & \text{(それ以外)} \end{cases}$$

行列 $\mathbf{X}$ の列ベクトル $\mathbf{x}_i$ が，語の潜在意味を取り込んだ文書 i の意味ベクトルである．これらのベクトルをクラスタリングする事で，語の潜在意味を反映した文書のクラスタリングが実現される．

3.3 漸増的環境への対応

隣接した期間 1，2 に作成された文書群に対応する行列をそれぞれ $\mathbf{D}_1$，$\mathbf{G}_1$,$\mathbf{H}_1$ および $\mathbf{D}_2$，$\mathbf{G}_2$,$\mathbf{H}_2$ とする．この時両期間合わせた行列は，それぞれ

$$\mathbf{D}_{1+2} = (\mathbf{D}_1 \ \mathbf{D}_2)$$
$$\mathbf{G}_{1+2} = (\mathbf{G}_1 \ \mathbf{G}_2)$$
$$\mathbf{H}_{1+2} = (\mathbf{H}_1 \ \mathbf{H}_2)$$

である．期間 1 に出現した単語に対応するランダム行列 $\mathbf{R}_1$ に期間 2 で新たに出現した単語に対応する列を追加したものを $\mathbf{R}_{1+2}$ とする．また各個別の期間の行列にそれぞれ $\mathbf{R}_{1+2}$ に対応するように要素が全て 0 の行を追加した行列を $\mathbf{D}_1$，$\mathbf{G}_1$,$\mathbf{H}_1$ および $\mathbf{D}_2$，$\mathbf{G}_2$,$\mathbf{H}_2$ とする．このとき期間 2 における文書 $\mathbf{D}_2$ の文脈行列は

$$\mathbf{W}_2 = \boldsymbol{N}(\mathbf{R}_{1+2}\bar{\mathbf{D}}_1\bar{\mathbf{G}}_1^{\mathrm{T}} + \mathbf{R}_{1+2}\bar{\mathbf{D}}_2\bar{\mathbf{G}}_2^{\mathrm{T}})$$
$$\mathbf{X}_2 = \boldsymbol{N}(\mathbf{W}_2\bar{\mathbf{H}}_2)$$

である．従って文書が増加した際には増加分に応じた計算だけで最新の情報に更新することができる．

3.4 クラスタリングと可視化

文書の意味ベクトルからなる行列が得られた後のクラスタリングと可視化については，論文 3) で提案した “Time-Arrayed SOM” をそのまま利用する．Time-Arrayed SOM(TaSOM) は，では，データをタイムスタンプに従ってデータセットを分割し，各期間ごとに二次元 SOM でクラスタリングを行なう．その後それらの全クラスターのセントロイドを一次元環状 SOM でクラスタリングすることで，各クラスターに隣接関係に基づいた角度と色を割り

Vol.2011-IFAT-102 No.3
Vol.2011-DD-80 No.3
2011/3/28

当て，それらを利用してクラスター構造の可視化を実現している．詳細は可視化例で述べる．

4. 可視化例

4.1 対象データ

実験データとするため，goo ブログ[8] からブログ記事を収集した．収集対象としたのは，2010 年 5 月 9 日の 24 時間中に更新のあった 26969 のユニークブロガーで，過去に遡って全てのブログ記事を収集することとした．そのうち収集できなかった記事やタイムスタンプに異常のある記事を含むブロガーを除外し，25217 ブロガーの全記事を収集できた．今回の実験に用いたのは，このうち 2007 年 12 月 1 日〜2008 年 1 月 31 日のタイムスタンプを持つブログ記事計 390,798 件である．キーワードの異なり総数は 428,235 であった．

これらの記事をタイムスタンプの日付に従い，一日分を一つのデータセットとし，合計 62 のデータセットを作成した．索引ベクトルの次元は 5000 とし，128 のクラスターを作成した．

文書からの単語の抽出は形態素解析器[11] を適用して行なった．一部の例外を除き，形態素解析により名詞とされたものを抽出した．

4.2 適用結果

4.2.1 面積チャート (area chart)

Time-Arrayed SOM では，クラスター間の隣接関係を学習し，類似した隣接クラスター間には類似度に応じて類似した角度と色を割り当てる．その色を用いることで，クラスター構造の経時変化を面積チャートとして可視化することができる．

図 1 は，62 日間のクラスター構造の変化を示す面チャートである．上図は縦軸にクラスターの実サイズを，下図では相対サイズを用いている．クラスターの積み上げ順は，クラスタリングによって得られた隣接関係によって決定されている．したがって一つのクラスターだけではなく，類似したクラスター群を単位としてクラスターの衰亡を観察することができる．また，クラスタリングに用いた SOM は環状トポロジーであるため，図の最下辺クラスターと最上辺クラスターは隣接している．

図から，12 月 24 日と 1 月 1 日を中心とした二箇所で青色系クラスターが増大していることが分かる． クラスター内の記事を人手で確認したところ，12 月 24 日のクラスターの大多数はクリスマスに関する話題，1 月 1 日のクラスターは正月に関する話題で占められていた．

4.2.2 極チャート (polar chart)

面積チャートはクラスターサイズを積み上げる事で得られた．Time-Arrayed SOM で導入した極チャートでは，クラスターサイズを用いたヒストグラムにより可視化を実現している．ただしここでは極座標を用いており，各クラスターを表すバーはクラスタリングによって与えられた角度に置かれる．隣接クラスター間距離が小さいほど与えられる角度の差も小さくなるため，クラスターの分離度を観察することができる．

図 2 に示すのは，12 月 22 日〜27 日の極チャートである．面積チャートからも確認できたように，この図からも 24, 25 の両日，青色系 (斜め左下の 220 °付近) のクリスマスクラスターが拡大していることが分かる．

4.2.3 SOM マップの 3D 表示

図 3 に示すのは，62 枚の SOM マップを 3D グラフィックスとして表示したものである．SOM のセルは円筒として描画されており，左端が 12 月 1 日，右端が 1 月 31 日のマップである． TaSOM では，i 番目の SOM マップと $i+1$ 番目の SOM マップ間で，類似したクラスターは類似した位置に構成されるよう工夫されているため，複数の日をまたがるクラスターも三次元の塊として確認できる．また，セルの透明度や輝度などを調整することで，特定のクラスターのみを強調することができるため，インタラクティブに特徴的なクラスターを発見するのに役立つ．

5. まとめと今後の課題

本稿では，ブログのように，テキストデータが漸増的に生成される環境下でテキストデータを効率的にクラスタリング・可視化するための，文書ベクトルの生成手法を提案した．提案手法では，テキストデータや総単語数の増加によっても文書ベクトルの次元が変化しないため，大量のテキストデータの効率的な処理が可能である．また，提案手法を実際のブログ記事集合に適用し，他稿で示したクラスタリング・可視化手法と組み合わせることで効率的に可視化が実現できる事を示した．

しかし，文書のベクトル表現の妥当性のチェックという意味では未だ十分ではない．今後は生成された文書ベクトルの妥当性とクラスタリング結果の数値的な検証を行ないたい．

また，可視化において，クラスタの分布やサイズを示すだけではその内容は把握できない．各クラスターの内容を簡潔に示すラベルの提示など，クラスター内容を可視化する手法を検討する必要がある．

本手法の応用可能性としては，例えば文書から抽出する単語として評価や感想を述べる際

Vol.2011-IFAT-102 No.3
Vol.2011-DD-80 No.3
2011/3/28

に用いられる形容詞などを用いることで，製品の評価の変遷などを可視化できる可能性がある．可視化により評価の傾向が大きく変化した時点を特定できれば，実社会での出来事と照らし合わせてその要因を特定できる可能性もあり，有効な市場分析手段となり得る．

参 考 文 献

1) Teuvo Kohonen, *Self-Organizing Maps, Third Edition*, Springer-Verlag, 2001.
2) 石川 雅弘, クラスター構造の経時変化を可視化するための *Time-Arrayed SOM* の提案, 情報処理学会 第 72 回全国大会講演論文集, 2010.
3) 石川 雅弘, 時系列テキストコレクションの可視化, 電子情報通信学会，第 17 回 Web インテリジェンスとインタラクション研究会 (WI2-17), WI2-2010-31, 2010.
4) Bingham, Ella and Mannila, Heikki, *Random projection in dimensionality reduction: applications to image and text data*, In Proc. of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining, pp245–250, 2001.
5) Christos H. Papadimitriou et al., *Latent Semantic Indexing: A Probabilistic Analysis*, In Proc.of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp159–168, 1998.
6) Sahlgren, Magnus, *An Introduction to Random Indexing*, In Proc. of the Methods and Applicatons of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 2005.
7) Kanerva P, Kristofersson J, Holst A, *Random indexing of text samples for latent semantic analysis*, In Proc. of the 22nd Annual Conference of the Cognitive Science Society, p.1036, 2000.
8) goo ブログ, `http://blog.goo.ne.jp/`.
9) Christopher D. Manning, Prabhakar Raghavan, Hinrich Sch tze, *Introduction to Information Retrieval*, Cambridge University Press; Anniversary, 2008.
10) Dimitris Achlioptas, *Database-friendly Random Projections*, In Proc. of the 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp274–281, 2001.
11) MeCab: Yet Another Part-of-Speech and Morphological Analyzer, `http://mecab.sourceforge.net/`.

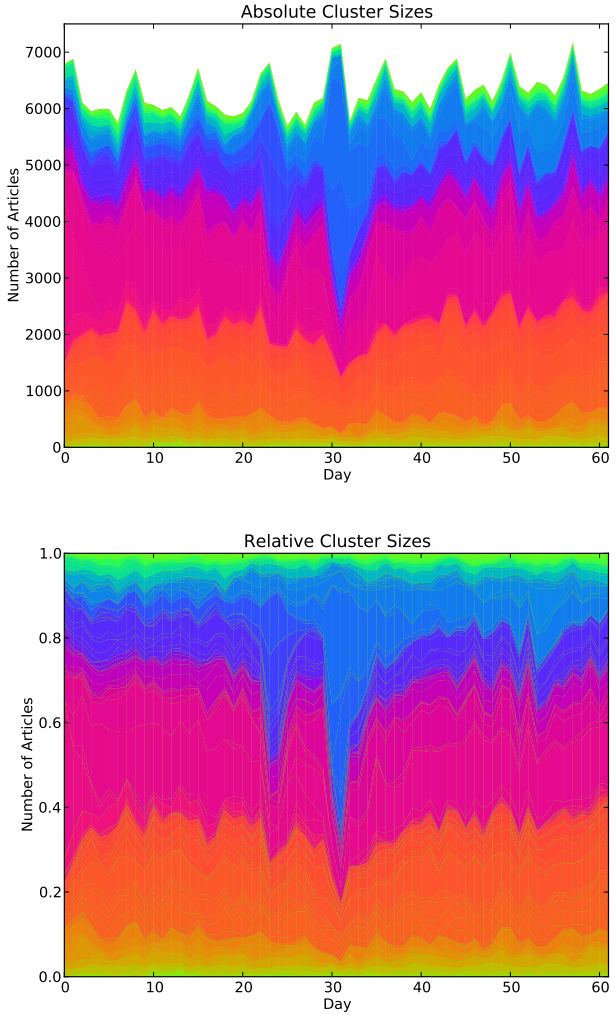

図 1 面積チャートによる可視化例 (2007 年 12 月 1〜2008 年 1 月 31 日): 横軸は 12/1 からはじまる日付，縦軸は各クラスターサイズ．上図は絶対サイズ，下図は相対サイズ
Fig. 1 Area chart visualization(from Dec. 1, 2007 to Jan. 31,2008)

Vol.2011-IFAT-102 No.3
Vol.2011-DD-80 No.3
2011/3/28

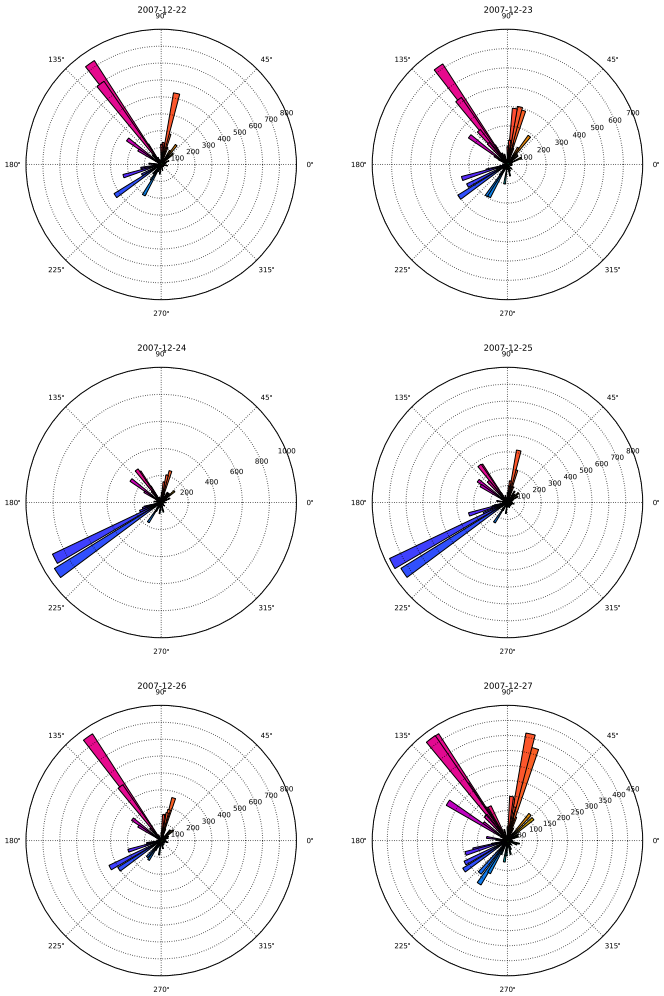

図 2 極チャートによる可視化例 (2007 年 12 月 22 日〜26 日): バーの長さはクラスターサイズ
Fig. 2 Polar chart visualization: (from Dec. 22 to 26)

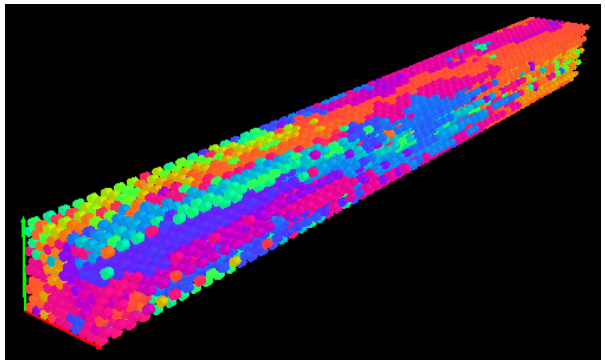

図 3 SOM マップの 3D 表示による可視化例
Fig. 3 3D visualization of SOM maps)