2J3-NFC2-6in

Wikipedia SOM

Wikipedia SOM

中山浩太郎
Kotaro Nakayama

東京大学 知の構造化センター
Center for Knowledge Structuring, The University of Tokyo

In our previous research work, we proposed a self organizing document map algorithm named MIGSOM, inspired by neuronal migration. MIGSOM has two notable characteristics; the stable clustering performance for large scale data and the interactive cluster visualization functionality. In this paper, we introduce a case study of MIGSOM that we aim to analyze Wikipedia's link data to uncover the structure of Wikipedia in both globally and locally.

1. はじめに

Web ブラウザを経由して誰でも編集可能なオンライン百科事典「Wikipedia」は，半構造化されたデータ構造を持ち，幅広い分野に高い網羅性を持つなどの特徴を持つことから，人工知能，自然言語処理，Web マイニングをはじめとする各種の研究分野で，コーパスとして活用されてきた．Wikipedia 上に公開される情報は日々増加しており，全ての言語を合計すると，1,800 万以上の記事が存在する．この結果，どの分野にどの程度の情報が存在し，分野同士がどうつながっているのか，といったような Wikipedia の全体像を把握することが困難になっている．Wikipedia をコーパスとして利用する研究においては，データの特性に応じてアルゴリズムを設計することが多いため，どのような記事集合がどれほどあり，どのようなクラスタがあるのか，クラスタ間の関係はどうなっているのかなど，全体を俯瞰することが重要である．また，Wikipedia の閲覧や編集など一般ユーザとして関わる場合にも，全体を俯瞰することは，不足している情報を把握することや分野同士の関係性を調べるといった用途において重要であると考えられる．

本研究では，神経細胞移動に着想を得た自己組織化マップアルゴリズムの「MIGSOM」[Nakayama 11] を Wikipedia に適用し，全体情報を俯瞰する方法を提案する．MIGSOM は，大規模な疎データを可視化し，文書マップを作成する技術である．MIGSOM には二つの特徴がある．一つは大規模なデータに適用した時にも安定したクラスタリング性能が期待できる点である．もう一方の特徴は，ズーム機能を利用したクラスタ解析が可能な点である．これにより，大局的なクラスタと局所的なクラスタの解析が可能になった．

本論文は以下のとおり構成される．まず，関連研究を説明した後，MIGSOM の概要について述べる．次に Wikipedia に MIGSOM を適用する方法について解説し，実際にクラスタリング分析を行った結果について考察する．最後にまとめとして，全体の考察と今後の展開について述べる．

2. 関連研究

Kohonen の自己組織化マップ（SOM: Self Organizing Map）[Kohonen 98] は，多次元データを低次元データに写像するために利用される教師無し学習手法である．Kohonen の SOM は多くの SOM 研究の原型であり，今日に至るまで様々なアルゴリズム改良が施されたが，基礎的な手順はほぼ同様である．一般的に Kohonen の SOM はスケーラビリティの高いアルゴリズムとして知られているが，これは学習に際して一回の反復（イテレーション）で利用されるレコードは一つだけであり，入力データ数が多くなっても，使用するメインメモリの要領が一定なためである．しかし，扱うデータが密ベクトルの場合は問題ないが，疎ベクトル集合が入力として与えられた場合，ベクトル情報が反復の度に大きくなるという問題がある．また，マップのサイズが十分に小さくなければ，大量のメモリを必要としてしまう．Kohonen の SOM ではマップサイズを自由に指定できるため，必ずしも大規模なデータに対して大きなマップが必要ではない．極論だが，数百万件の入力レコードに対して，2 行 x2 列（4 ノード）のマップを利用して学習することも可能である．しかし，文書をマップ上に配置して解析するための**ドキュメントマップ**を作成するには，**データ数に応じた十分な大きさのマップが必要**となる．これは，大規模データを扱う場合，マップ領域が十分に大きくなければ一つのノードに多くのレコードが重複してマッピングされてしまう上に，十分な解析精度が得られないためである．

SOM はデータの全体像を把握するための可視化ツールとして有効だが，大規模データを扱うには上記の問題を解決しなければならない．そのため，大規模な疎データを扱うために Lagus らの WebSOM[Lagus 04] に代表される次元圧縮を利用した手法が主に研究されてきた．WebSOM は，文書単語行列などの大規模疎行列を，LSI（Latent Semantic Indexing）[Deerwester 90] やランダムマッピングなどの手法を用いて次元圧縮し，小規模密ベクトルに変換してから解析する方法を提案している．これらの研究によって次元圧縮の有効性は十分に証明されてきたが，SOM 自体を大規模な疎行列の解析に適用させる研究は少なかった．

次に，Wikipedia と可視化に関する関連研究を議論する．Wikipedia のデータを可視化する研究やシステムは，本研究以外にもいくつか存在する．「Wikipedia Visualizations」[*1] では，Wikipedia の記事の間の類似度を計測し，似た記事を近くに配置して表示するシステムを公開している．本システムの詳細なアルゴリズムは論文などで公開されていないため，その有効性を評価することは困難であるが，膨大化する Wikipedia

連絡先: 中山浩太郎，東京大学知の構造化センター，東京都文京区本郷 7-3-1，03-5841-0462，03-5841-0454，nakayama@cks.u-tokyo.ac.jp

∗1 http://scimaps.org/maps/wikipedia/

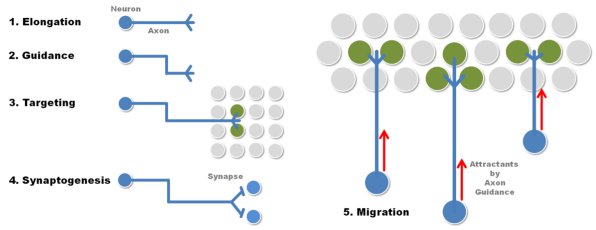

図 1: Neuronal Migration (Translocation)

の全体像を把握することの重要性とその方法を示唆したシステムとして評価されている．

「History Flow Visualizations」[Viégas 04] は Wikipedia における編集履歴を可視化する研究である．編集履歴を「大量削除」や「攻撃的コピー」,「偽コピー」などに分類し，その経緯を時系列で可視化する手法である．その他にも，編集パターンを可視化する「Chromogram」[Wattenberg 07] や，時系列を追って言語毎の記事数を可視化した「Language Development on Wikipedia」*2などが存在する．

上述のとおり，Wikipedia データを利用した可視化の研究はいくつか存在する．その目的や利用するデータは様々であるが，これらのシステムや研究の多くにおいて共通しているのは，膨大化する Wikipedia の全体像を把握するために可視化が有効であるという認識である．

3. MIGSOM

本節では MIGSOM のアルゴリズムを詳述する．アルゴリズムの説明に先立ち，MIGSOM の基礎モデルとして神経細胞移動について説明する．

3.1 神経細胞移動

脊髄や脳などから構成される中枢神経系は，情報の分析・判断・決定など，知的活動を担当する部位である．中枢神経系が持つ効率的・合理的な情報処理能力を解明することは，長い間多くの研究者の興味の対象であり，活発に研究が進められてきた．その中でも，神経細胞が最適な場所を見つけるために移動する現象「**神経細胞移動（Neuronal Migration）**」の重要性が近年活発に議論されている．

神経細胞移動では，神経細胞が自身の軸索を様々な場所に伸ばし，移動するべき場所を見つけた後に，軸索を縮めて移動する．神経細胞移動の詳細な手順を図 1 に示す．

まず，ニューロンは生成された後，軸索を延ばし始める．次に少しずつ軸索で周辺を探索をしながら，軸索を延ばす方向を決める．そして，定着する場所を決めた後に，周辺ニューロンとの間にシナプス（結合部）を生成する．最後に軸索を縮めることで，ニューロンが移動する．

3.2 MIGSOM の基本原理

MIGSOM は，神経細胞移動に着想を得た SOM アルゴリズムである．Kohonen の SOM と同様，反復処理による教師無し学習手法であるが，データの表現方法と学習対象の点で大きく異なる．Kohonen の SOM では，マップ上の各ノードが独自のベクトルを持ち，その値を変更（ベクトル修正）していくことで学習が行われる．この際，マップ上の一つのノードが必ずしも一つの入力レコードに対応するわけではない．これとは対照的に，MIGSOM では，**マップ上のノードが入力レコードに対応**する．ノード上に配置された入力レコードを**ニューロン**と呼び，ニューロンがマップ上を移動することでマップ全体の学習が行われる．また，入力レコードが割り当てられないノードには，ランダムに生成されたベクトルを持つ**グリア細胞**が配置される．グリア細胞は，ニューロンの間を埋める補助的なベクトルであり，ニューロンの移動をガイドする役割を果たす．ニューロンの移動方向は，ランダムに周辺に軸索を伸ばし，自身のベクトルと類似しているノードが多く集まる方向を発見することで決定される．

MIGSOM では，まず全ての入力レコードをニューロンとしてマップ上のノードにランダムに配置する．次に，空いているノード上にグリア細胞としてランダムなベクトルを生成する．そして，次に示す処理を反復することで学習する．

Algorithm $train()$:

1 Randomly select g from G
2 $\vec{m_g}=\vec{O}$ #Initialize by null vector
3 $N = GaussianSelection(g)$
4 **for each** $n \in N$
5 $dist = distance(n, g)$ #Distance in map
6 $power = tanh(dist)$
7 $\vec{m_g}=\vec{m_g} + U(n, g) \cdot power \cdot Sim(n, g)$
8 if $|\vec{m_g}| > t$
9 $Translocate(g, \vec{m_g})$

- G: マップ上の全ノード集合
- $\vec{m_g}$: g の移動ベクトル
- $GaussianSelection(g)$: ノード g 周辺のノードをガウシアン分布に従ってランダムに選択する関数
- $distance(n, g)$: ノード n とノード g のユークリッド距離
- $tanh(dist)$: 距離 dist の逆正接
- $U(n, g)$: g から n への単位ベクトル
- $Sim(n, g)$: n と g が持つベクトルのコサイン類似度
- $Translocate(g, \vec{m_g})$: 移動ルーチン．もし $|\vec{m_g}|$（$\vec{m_g}$ のノルム）が閾値 t より大きい場合，g はステップ 9 で $\vec{m_g}$ 方向に移動（Translocate）

本アルゴリズムでは，まずマップ上の全ノード集合 G からランダムにノード g を一つ選択し，訓練用ニューロン（グリア細胞）として採用する．次に，g がどの方向に移動するかを示す $\vec{m_g}$ の値をゼロベクトルで初期化する．そして，g 周辺のノードからガウシアン分布に従ってランダムにノードを選択する．つまり，マップ上の距離が近いニューロン（やグリア細胞）ほど高確率に選択される．$GaussianSel(g)$ は，ノード g 周辺のノードをガウシアン分布に従って選択する関数である．

ランダムに選択された周辺ノード集合 N の各ノード n に対し，g と n が持つベクトルの類似度とマップ上の距離を求める．そして，類似度の高いノードが存在した場合，類似度に応じてそのノードの方向へと移動ベクトル $\vec{m_g}$ を修正する．最後に，移動ベクトルのノルム $|\vec{m_g}|$ を求め，閾値以上であれば，移動する．この時，ニューロンが移動できる範囲は g（起点）の周辺 8 ノードのみとした．また，移動に伴い，移動先のニューロン（かグリア細胞）は g に移動する．つまり，移動とは厳密には移動先のニューロンと位置を入れ替えることである．

MIGSOM の性能を検証するために，ランダムに疎行列を生成し，データサイズ，次元数，データの密度をそれぞれ変更

*2 http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/language-development-on-wikipedia

しながら，精度とメモリ使用量の両面を評価する実験を行った [中山 10]．その結果，MIGSOM は通常の SOM に比べて疎なデータに対する解析では精度とメモリ効率の両面で従来手法より良い成績を残した．特に，データサイズが大規模になった場合，従来手法では著しい精度低下が見られたのに対し，MIGSOM ではデータサイズの増加の影響を受けにくく，精度低下を抑制できていることがわかった．

4. Wikipedia SOM

実データとして英語版 Wikipedia[*3]へ MIGSOM を適用し，可視化した事例を以下に紹介する．Wikipedia は Web ブラウザを利用して誰でも利用可能なオンライン百科事典であり，その規模や網羅性，実用性から，Web マイニングや自然言語処理，情報検索などの研究において基盤リソース（コーパス）として広く利用されている．Wikipedia の記事は，一つのエンティティ（概念）に対応しており，ハイパーリンクによる密なリンク構造を持つ．

本事例では，2009 年 4 月の英語版 Wikipedia（記事数 300 万程度）をデータセットとして利用した．まず，各記事を，記事内に出現するリンクでベクトル化し，300 万行× 300 万列の隣接行列を作成した．次に，被リンク数が 10 件以下の記事など，ノイズデータを除外した後に，可視化で利用できる記事だけ（画像が含まれる記事だけ）抽出した．その結果，約 11 万 8 千件の記事が残った（つまり，11 万 8 千行× 300 万列の隣接行列を作成）．

次に，サイズが 500 × 300，15 万ノードのマップを作成し，記事をランダムに配置した．その後，記事が割り当てられなかった 3 万 2 千個のノードにグリア細胞（ランダムに生成したベクトル）を配置し，MIGSOM で学習した．

なお，実行結果の解析を容易にするために，予め記事データを K-Means クラスタリングによって 10 のクラスタに分類し，クラスタ毎に色分けした．解析の様子を図 2 に示す．ここに示しているのは，Xeon 2.2GHz のマシンで解析開始直後から 48 時間後までの様子である．まず，初期状態（左上）では全てのノード（記事）がバラバラに配置されているが，1 時間経過（右上）すると徐々に小さなクラスタが発生してきていることがわかる．24 時間経過（左下）すると，さらに明瞭に大きなクラスタが発生し，類似している記事が近くに配置されていることがわかる．データサイズが大きいために精度を計測することはできなかったが，24 時間の解析結果と 48 時間の解析結果は目視ではそれほど大きな違いが無いように見える．

5. クラスタリング分析

次に，本解析結果を基に，各記事に含まれる画像を表示してズームして閲覧できるシステム「Wikipedia SOM Visualizer」を開発し，詳細に検証した結果を図 3 と図 4 に示す．Wikipedia SOM Visualizer は，各記事に含まれる画像をマップ上の記事の上に表示し，Ajax 技術を使ってインタラクティブな拡大縮小機能を実装することで，シームレスな分析を可能とした．

本システムを利用して SOM の実行結果を解析すると，「文化・スポーツ」「ポップ・ミュージック」「地理情報」などの明瞭なクラスタができているのがわかる．次に，地理情報の箇所をズームしてみると，さらにアメリカの国道に関する記事が集まった「道」に関するサブクラスタや，有名な観光場所が集まった「場所」といったサブクラスタが存在することがわかる．また，意外な知見としては，「化学」に関するクラスタと，銃などの「兵器」に関するクラスタが密に交差しているということが判明した．境界領域を調べてみると，新型の火薬など新しい化学の技術を利用した兵器に関する情報が双方の中間ハブとしての役割を果たしているように見える．

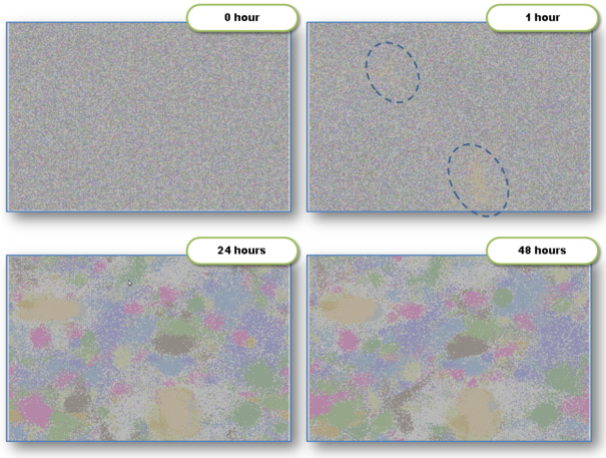

図 2: MIGSOM for Wikipedia Data

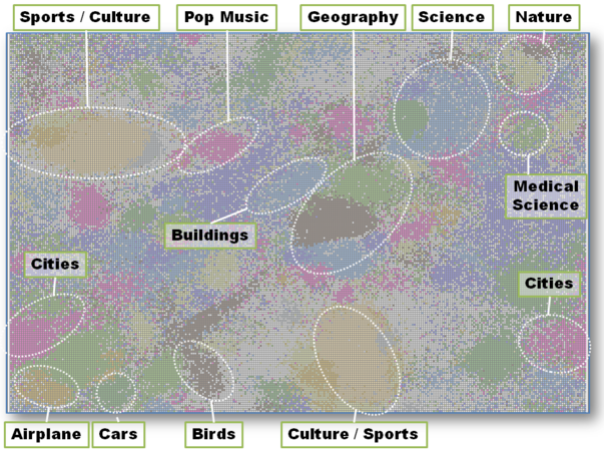

図 3: Discovered Clusters

クラスタの境界に着目して解析を進めると，より興味深い知見が得られる．図 5 に，境界面がはっきりしているクラスタの例として「大学」クラスタの様子を示す．大学クラスタは，その上部の文化クラスタなどとは明確に分離されていることがわかる．これは，局所的に密なリンク構造を持っている記事集合が存在する場合，よく出現するパターンである．次に境界面がはっきりしていないクラスタの例として「道」クラスタの様子を図 6 に示す．道クラスタは，その上部に場所クラスタや建築物クラスタといった，互いに密なリンクを持つクラスタを保有し，その境界面ははっきりとわかれていない．これは，クラスタ間で相互にリンクが存在する場合などに良く出現するパターンである．

このように，MIGSOM は単に大量情報をクラスタリングするだけでなく，クラスタ間の関係性などを俯瞰する用途に利用できる．これにより，記事集合の特性に応じた解析手法の設計などに貢献できる可能性がある．また，不足している情報を把握することや分野同士の関係性を調べることにより，閲覧や編集など，ソーシャルメディアに関わるユーザを支援する技術として利用できる可能性がある．

*3 http://wikipedia.org

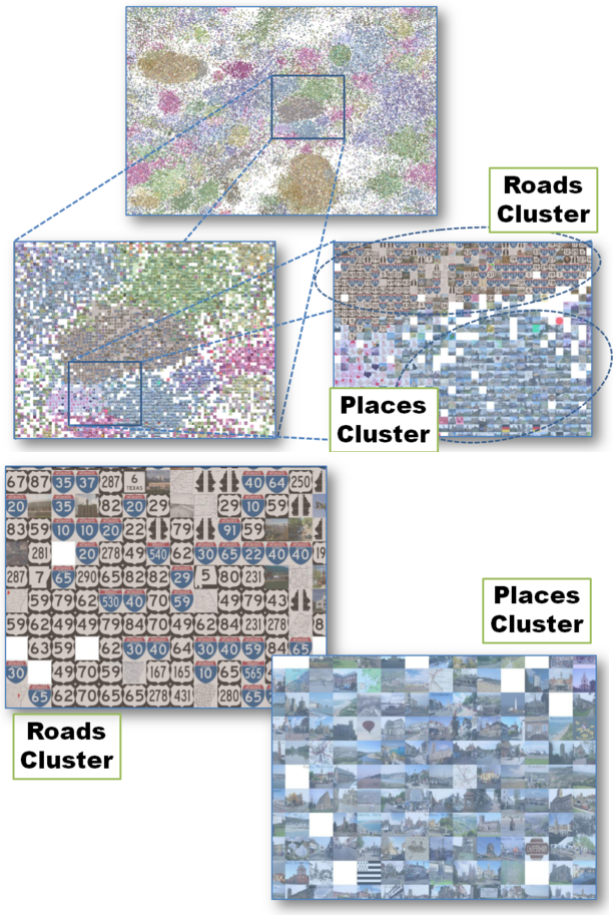

図 4: Interactive Zooming Interface

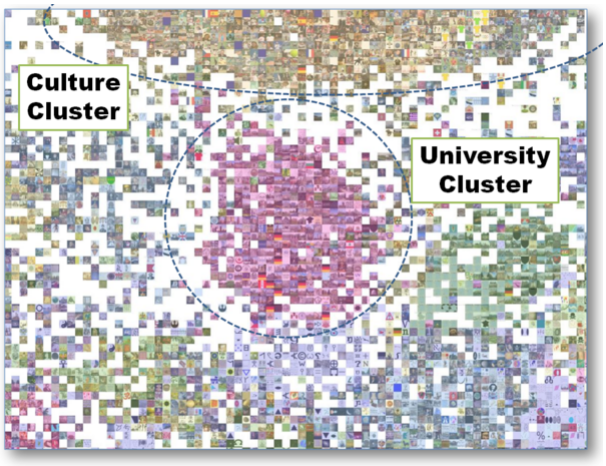

図 5: Clearly distinguished clusters

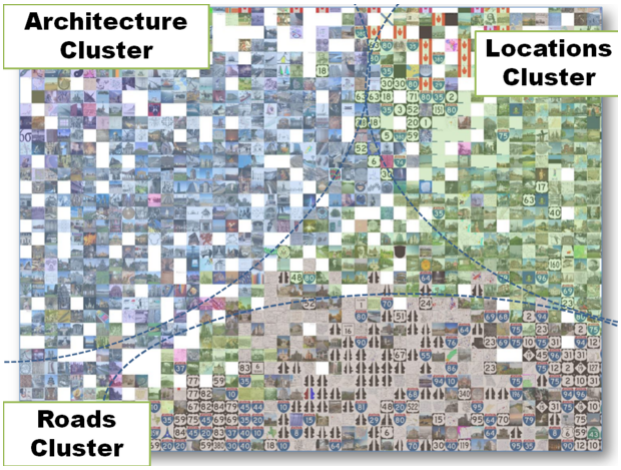

図 6: Unclear clusters

6. まとめ

本稿では，神経細胞移動に着想を得た自己組織化マップアルゴリズム「MIGSOM」を利用し，Wikipedia の記事データを可視化する手法を提案した．解析結果から，境界面が明確なクラスタや，他のクラスタと密に交差しているクラスタなど，様々なクラスタが存在することを発見し，さらにサブクラスタとの関係なども解析できることを確認した．

当面の技術的課題は解析時間である．マップが十分収束するまでに数十時間必要としているので，アルゴリズムの改良により，計算時間を削減することを目指したい．

参考文献

[Deerwester 90] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407 (1990)

[Kohonen 98] Kohonen, T.: The self-organizing map, *Neurocomputing*, Vol. 21, No. 1-3, pp. 1–6 (1998)

[Lagus 04] Lagus, K., Kaski, S., and Kohonen, T.: Mining massive document collections by the WEBSOM method, *Information Science*, Vol. 163, No. 1-3, pp. 135–156 (2004)

[Nakayama 11] Nakayama, K. and Matsuo, Y.: A self organizing document map algorithm for large scale hyperlinked data inspired by neuronal migration, in *Proc. of International World Wide Web Conference (WWW), Companion Volume*, pp. 95–96 (2011)

[Viégas 04] Viégas, F. B., Wattenberg, M., and Dave, K.: Studying cooperation and conflict between authors with *istory flow* visualizations, in *Proc. of ACM CHI Conference (CHI)*, pp. 575–582 (2004)

[Wattenberg 07] Wattenberg, M., Viégas, F. B., and Hollenbach, K. J.: Visualizing Activity on Wikipedia with Chromograms, in *Proc. of Conference on Human-Computer Interaction (INTERACT)*, pp. 272–287 (2007)

[中山 10] 中山 浩太郎：Migsom:神経細胞移動モデルに基づく自己組織化マップ～大規模リンクドデータへの応用～, Web とデータベースに関するフォーラム (WebDB Forum 2010) (2010)