階層型自己組織化マップとユーザ嗜好反映を取り入れた検索システム

小室　達也[†]　青野　雅樹[‡]

[†]豊橋技術科学大学　情報工学専攻　〒441-8580 愛知県豊橋市天白町雲雀ヶ丘 1-1
[‡]豊橋技術科学大学　情報工学系　〒441-8580 愛知県豊橋市天白町雲雀ヶ丘 1-1
E-mail:　†komuro@ kde.ics.tut.ac.jp,　‡aono@ics.tut.ac.jp

あらまし　多量なデータの中から必要とする情報を検索する際において，ユーザが結果を求める際の検索にかかる検索時間や，列挙された検索結果の中から必要とするデータを確認していく作業などに負担がかかる場合がある．このような問題に対して本研究では，多量なデータの中から必要なデータを効率的に得る一手法として，階層的自己組織化マップを用いる手法を提案し，データを階層的に分類し，各層を 2 次元のマップとしてユーザに視覚的にデータを提示するシステムの開発を行ってきた．また，ユーザがデータを検索する際の問い合わせ結果を用いてユーザの嗜好を反映させ，ユーザに沿ったマップ上のデータの構成を再構成させる機構を提案する．本報告では，使用するデータとして特許データに焦点を当て，階層型自己組織化マップを用いた特許データ検索及び，ユーザ嗜好を反映させた後のマップ上のデータの配置の変化の効果について述べる．

キーワード　自己組織化マップ，嗜好反映，情報検索・可視化

Similarity Search using Hierarchal Self-Organizing Map with Users' Interests

Tatsuya Komuro [†]　and　Masaki AONO[‡]

†‡Department of Information and Computer Sciences, Toyohashi University of Technology

1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi-ken, 441-8580 Japan
E-mail:　†komuro@ kde.ics.tut.ac.jp,　‡aono@ics.tut.ac.jp

Abstract　With the spread of the Internet in recent years, the amount of information on the Web has increased explosively. Many Internet users tend to spend many times on trying to find the most relevant item from the search result of Web search engines in the form of "long sorted list". In this paper, we have made the visualization with the SOM special in several ways. First, we have offered two different levels of SOMs depending upon the size of data to visualize, i.e. a single-layered SOM and a Hierarchical SOM called GHSOM (Growing Hierarchical Self-Organizing Map). Second, after initial view is shown, users can modify the view of the SOM to their likings iteratively until they become satisfied. In order to verify the effectiveness of proposed methods, we have carried out experiments of our methods for visualizing the search result. We also conducted experiments how users' interests are reflected to the modified SOM to see if they are better than the initial SOM that system provides to the users. From these preliminary experiments, we have demonstrated that our system is flexible enough to change its view to the users' likings to some extents.

Keyword　Self-Organizing Map, Users' Interests, data visualization

1. 緒言

1.1. 研究背景

WEB検索やインターネット上の通販サイトでの商品の検索，あるいは特許文献検索など，多量にあるデータの中から必要とする情報をユーザが検索する場面において，検索結果の提示方法は通常，WEB検索エンジンなどに見られるリスト形式で結果がユーザに提示される．この際，ユーザが検索結果としてリスト状の形式で提示されたデータの中から目的としたデータを視覚的に確認し，目的としたデータが現れるまで次々と探索を行っていく作業など，ユーザにとって負担となる場合がある．このようなリスト形式でのデータ提

示方法の問題に対して，今までに自己組織化マップ（Self-Organization Map：SOM）[1]やツリーマップ(treemap)[2]などを用いて，マップ状でのデータ提示方法が提案・開発されてきた．

本研究は，マップ状で検索結果の提示方法を実現するための手段としてSOMを用いた多次元の特徴を持つデータの提示・検索システムの開発に焦点を当ててきた．

1.2. 研究目的

SOMは，多次元の特徴を持つデータを低次元のマップに非線形に写像し，クラスタリングあるいは，データ群の可視化を行う手法の一つで様々な分野に応用されている．しかし，SOMで多量なデータを扱う場合，データ量に応じて適切なマップサイズに変更しなくてはならず，また，マップ上に表示されたデータの分布をユーザが把握しきれず，可視化の効果が薄れるといった問題が生じる場合がある．

そこで本研究では，多量なデータを階層的に分類し，検索結果の可視化を行う階層的なSOMを用いたシステムを提案する．本システムでは，階層型のSOMとして成長型階層自己組織化マップ(The Growing Hierarchical Self-Organizing Map：GHSOM)[3]を用いることにより，多量なデータを階層的に分類し，2次元マップ上で表示させるように可視化を行う．

また，本システムではユーザの嗜好データを用いることで，個人の好みを反映させたマップを構成するシステムを提案する．

本報告では、使用するデータとして特許データに焦点を当て，GHSOMを用いた特許データ検索及び，ユーザ嗜好を反映させた後のマップ上のデータ配置の変化の効果について述べる．

2. 関連研究

SOMを応用し，文書群をクラスタリングする試みがKohonenらのグループにより行われており，WEBSOM[4]と呼ばれるシステムなどがある．

また，データ群を階層的に分類し,可視化する代表的な方法としては，treemapを用いた可視化が有名であり，newsmap[5]のようなニュース記事を視覚的に提示するシステムがある．同様に，階層型の分類と可視化手法に関して，階層の末端のデータ数が1000件を越える場合にSOMを用いて多量なデータの分類を行い，データ検索に用いる方法[6]も提案されている．

本提案手法は，階層型の分類という点で従来手法と類似するが，ユーザの嗜好に合わせてマップ上のデータの配置を設定できる点に特色がある．

3. 提案システム

本研究が提案するシステムの構成図を図 1 に示す．システムはブラウザ上でユーザとシステムとのやり取りを行う「ユーザインターフェース部」(図中A)，GHSOMの計算を行う「GHSOM計算部」(図中B)，分類するデータを蓄積する「データサーバ部」(図中C)，検索結果がユーザの嗜好に合うように学習を行う「ユーザ嗜好反映部」(図中D)から構成される．

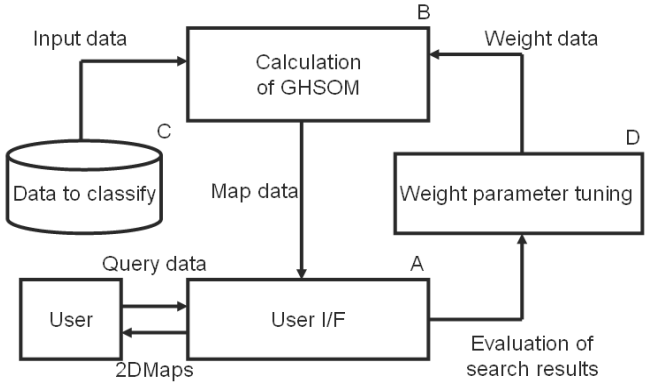

図 1. システム構成図

3.1. ユーザインターフェース部

ユーザインターフェース部ではユーザからのクエリを受け取る作業や，GHSOM 計算部によって出力されたマップデータの表示を行う．

図 2. ユーザインターフェース画面

3.2. GHSOM 計算部

GHSOM は 1 層で表現される通常の SOM に比べ，多階層のマップ群を形成することができる(図 3). そのため，多量なデータを分類及び可視化する際において有益なものとなる．さらに，GHSOM は各層のマップサイズも入力されたデータ群に対して自己的に変更することが可能である．GHSOM 計算部では，与えられたデータ群の階層分類を行う．GHSOM のアルゴリズムを以下に記す．このアルゴリズムでは，①から⑦までのステップをマップの階層拡張が終了するまでを繰り返し実行する．

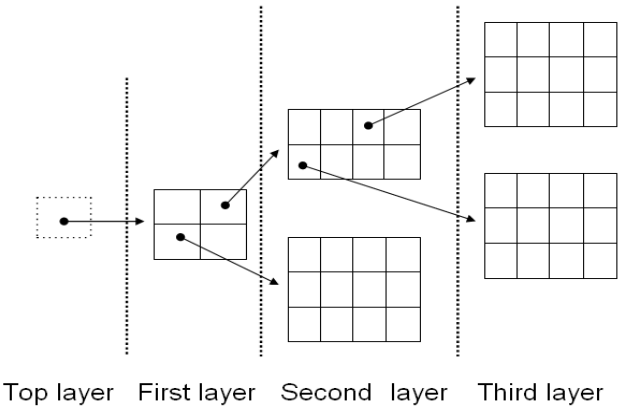

図 3. GHSOM のイメージ

① 第 0 層の量子化誤差（各ユニットの値と各データとの距離）の平均である MQE_0 を求める．

② 初期サイズのマップとして，2×2 のマップ m,k（m は階層レベル，k は階層中のマップ数）を生成する．

③ 入力データセットを通常の SOM のアルゴリズムに基づいてマップ m,k に投影する．

④ マップ m,k の各ノード i とそのノード i に割り当てられた入力データ間の量子化誤差 qe_i を求め，その平均である $MQE_{m,k}$ を求める．

⑤ マップ m,k の平均量子化誤差 $MQE_{m,k}$ とその上位層のマップ $m\text{-}1,k$ が式 1 の関係を満たしていた場合，⑥のマップの拡張ステップに進む．そうでなければ，⑦の階層化判定ステップに進む．

$$MQE_{m,k} >= \tau_2 \cdot MQE_{m-1,k} \qquad (1)$$

⑥ マップの拡張規則に沿ってマップの拡張を行う．ここで，マップは必ずしも正方形の形にはならない．マップ拡張終了後，③の SOM の計算ステップに戻る．

⑦ ノード i の量子化誤差 qe_i と第 0 層のマップの平均量子化誤差 MQE_0 が式 2 の関係を満たしていた場合，マップの拡張を行い，②の初期マップ生成ステップに戻る．そうでなければノード i からは新たなマップは生成されない．

$$qe_i >= \tau_1 \cdot MQE_0 \qquad (2)$$

3.3. ユーザ嗜好反映部

提案する検索システムでは提示された検索結果に関してユーザの嗜好を反映させる．嗜好の反映方法は以下の通りである．

① ユーザはシステムにクエリを与える．

② システムは与えられたクエリをもとに，検索対象となるデータ群の中から最も類似しているデータを計算によって求め，そのデータをユーザに提示する．

③ ユーザはシステムによって提示されたデータの詳細なデータを確認し，そこで自分が望むデータであった場合はそれをマッチデータとしてシステムに返し終了．そうでない場合は，提示されたデータの近傍のデータを確認していき，マッチデータとなるデータが見つかるまでマップの探索を繰り返し行い，見つかった場合，④のステップへ進む．

④ システムは③でユーザより返されたマッチデータを受け取り，それを教師データとし，各階層に存在するマップ内の構成をユーザの嗜好に沿ったものに変化させる．このマップ上のデータの配置の変化を行う方法として 3 層のバックプロパゲーションを用いる．具体的にはバックプロパゲーションを用いて，ユーザが選んだマッチデータを教師データ，システムによって提示されたデータを入力データとして学習を行い，学習して変換された特徴ベクトルを出力する． なお，今回用いるバックプロパゲーションの学習方法は，入力層，中間層，出力層の 3 層のニューラルネットワークを用いた n 入力 n 出力の学習である．

⑤ ④による学習後，システムはマップを再作成し，ユーザに再度提示を行う．これを繰り返し行うことによって，ユーザの嗜好を反映させ，ユーザに特化したマップに変化させていくことが可能となる．

4. システム検証

本研究では開発したシステム，GHSOM によるデータ分類及びその効果の検証，並びに，ユーザ嗜好反映によるマップ上のデータの配置の変化についての検証を行った．

4.1. 検証に使用するデータ

検証に使用するデータとして，NTCIR-5 に含まれる 2002 年の特許文書，約 35 万件のうち 1000 件の文書を使用した．この特許文書から検証で使用するための文書ベクトルを作成する．文書ベクトルの作成方法は以下の通りである．

① 特許文書内に含まれるタグなどを除去し，文書の整形を行う．

② Java 用の形態素解析器「Sen」を使用し，①で整形した特許文書を形態素解析する．その結果が，

「名詞」及び「未知語」及び,「連続した名詞」となった単語で構成される TF 値を求める.

③ DF 値を作成し，DF 値 3 以下と 500 以上のものを削除する.

④ ②及び③で求めた TF 値，DF 値を用いて TF-IDF を計算する.

上記の作成方法をもとに，文書ベクトルを作成し，今回,次元数が 7680 次元となる文書ベクトルが作成された.

また，文書ベクトルの次元数及び，データ件数を考慮した結果, SOM の計算量の増加やクエリと文書ベクトルデータとの類似度計算の部分での計算量増加が考えられるため，文書ベクトルの次元削減を行った.

具体的には，文書ベクトルの次元削減手法としてよく用いられる LSI(Latent Semantic Indexing)[7]を用いて，作成した 7680 次元の文書ベクトルを 50 次元の文書ベクトルへと変換した．以後の検証ではこの次元削減後の文書ベクトルを使用する.

4.2. 検証方法

① GHSOM を用い，特許文書データ 1000 件を投影し，階層化されたマップ群を表示させ，データの階層化分類の妥当性の検証を行う．GHSOM のパラメータとして τ_1=0.008, τ_2=0.75 とし，各階層における SOM はバッチ型を用い，学習回数を 1000 回として計算を行った.

② ユーザ嗜好反映によるデータの配置の変化についての検証を行う．具体的には，ユーザがクエリとなる文書を与え，それに対してシステムが提示するデータをマッチデータであるかないかの 2 択で評価し，学習を行う．その後，マップ上のデータの配置がどのように変化しているかを検証する. 尚, 今回は単層の SOM を用いて検証を行う.

4.3. 検証結果

4.3.1. 検証結果 1

図 5 は検証方法①の GHSOM によって生成されたマップである．データセットは全 5 階層のマップ群で構成され, GHSOM によるデータの階層化が確認できた．さらに，各マップの詳細なデータを確認していくと図 5 の様に，データナンバー・02001565（内容：レーザ加工装置）とデータナンバー・02001566（内容：レーザ加工装置及びレーザ加工方法）と，同じような特許データが集約していることを確認することができ，GHSOM による特許データの階層化分類の妥当性が確認できた.

図 5. GHSOM による特許データ分類結果の一部

4.3.2. 検証結果 2

実際に本研究が開発した単層の SOM を用いたシステムを使用し，学習を行った結果について述べる．方法として，各クラスタに属しているデータが正しいクラスタに属するよう,ユーザ嗜好の反映を合計 20 回行い，その後，マップの再作成を行い，マップ上のデータの配置の構成を変化させた後のマップ上の各クラスタの分類誤り率を，そのクラスタに付与されているラベルを元に式（3）より求め，初期のマップの分類誤り率との比較を行う．尚，今回は 20 回の試行の内，データがインクジェットに関する特許データ，レーザー加工に関する特許データ，及び，浄化装置に関する特許データに焦点を当て，試行を行った.

表 1 は，任意に抽出した 5 つのクラスタの初期マップ及び, 20 回学習後のマップにおける分類誤り率の表である．この表より，No.4 の「レンズ，パターン」のクラスタを除く各クラスタにおいて，分類誤り率が改善し,ユーザの意図したマップに近づけることができ，学習の効果を確認することができた.

$$\text{分類誤り率} = \frac{\text{あるクラスタに誤って分類された文書数}}{\text{あるクラスタに分類された特許文書数}} \quad (3)$$

5. 結言

成長型階層自己組織化マップを用い，特許データの階層化分類及び可視化手法を取り入れた検索システムを提案した．また，マップ上のデータの配置をユーザの嗜好によって変化させる方法を提案した．検証の結果，特許データを階層的に分類し視覚化することが確認できた．また，ユーザ嗜好反映機構により，マップ上のクラスタの分類誤り率の改善を確認した.

今後の課題として，ユーザ嗜好反映方法の開発及び,

表１ 学習前と学習後の分類誤り率

		分類誤り率（%）	
No	ラベル	初期 マップ	学習後の マップ
1	インク インクジェット	0.794	0.368
2	インク インクタンク	0.200	0.091
3	インク ノズル	0.385	0.111
4	レンズ パターン	0.455	0.474
5	フィルター ポンプ	0.389	0.369

具体的な検証を行う必要があると考えられる．また，今回はユーザ嗜好反映を単層 SOM で試みたが，GHSOM での実装も試みる．

謝辞

NTCIR テストコレクションは国立情報学研究所の許諾を得て使用させていただきました．また本研究は科研費基盤（C）20500090 の支援を受けて遂行しました．

参考文献

[1] T.Kohonen, Self-Organizing Maps, Springer, 1996

[2] B.Johnson and B.Shneiderman, “Tree-Maps:A Space-Filling Approach to the Visualization of Hierarchical Information Structure,” In *Proc. IEEE Visualization ’91*, pp.284-291, 1991.

[3] A. Rauber, D. Merkl, and M. Dittenbach:, “The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data”, IEEE Transactions on Neural Networks, IEEE. , 2002

[4] Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996b). Newsgroup exploration with WEBSOM method and browsing interface, Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

[5] Newsmap:http://marumushi.com/apps/newsmap/newsmap.cfm

[6] 澁谷慧一郎，遠山元道．"自己組織化マップを利用する分類済み階層の自動設定"，DEWS2007，2007

[7] S.C.Deerwester, S.T.Dumais, T.K.Landauer, G.W.Furnas, and R.A.Harshman, “Indexing by latent semantic analysis”, Journal of the American Society for Information Science, Vol.41, no.6, pp.391-407, October 1990.

[8] 徳高平蔵，岸田悟，藤村喜久郎，自己組織化マップ-理論・設計・応用，マーク M.ヴァン・フッレ著，徳高平蔵，藤村喜久郎　監訳，海文堂，2001