知 能 と 複 雑 系 128−21
（２００２．５．２４）

WWW 文書集合族からの時系列的話題情報の抽出・可視化手法の提案

高間 康史

東京都立科学技術大学電子システム工学科

〒 191-0065 東京都日野市旭が丘 6-6
E-mail. ytakama@cc.tmit.ac.jp

Abstract:
WWW から得られる時系列的関連を持った文書集合族から話題の流れに関する情報を抽出・可視化する手法について提案する．オンラインニュースを始め，Web 上に公開されている情報は，新規の話題や流行に関する情報を多く含んでおり，これらをユーザに提示することができればビジネスチャンスなどにもつながる事が期待できる．提案手法は，免疫ネットワークモデルに基づくランドマーク抽出により文書集合毎の可視化を行い，文書集合間の時系列的関連性は，免疫記憶細胞モデルを導入することにより考慮する．抽出されたランドマークで表される話題を強調したキーワードマップを生成するために，免疫ネットワーク・メタファをバネモデルに組み入れる手法についても提案する．9 月 17 日〜9 月 21 日の間に公開されたオンラインニュース記事を対象として実験を行った結果，全文書集合を通じてランドマークの対応付けによる話題の流れが抽出可能であること，および免疫ネットワーク・メタファによりそれらを強調したキーワードマップを安定して生成可能であることを示す.

Proposal of Topic Stream Visualization from Sequence of WWW Document Sets

Yasufumi Takama

Tokyo Metropolitan Institute of Technology

Asahigaoka 6-6, Hino, Tokyo 191-0065 Japan
E-mail. ytakama@cc.tmit.ac.jp

Abstract:
A Web information extraction/visualization method based on the document set-wise processing is proposed to find the topic stream from a sequence of document sets. Although the hugeness of the Web as well as its dynamic nature is burden for the users, it will also bring them a chance for business and research if they can notice the trends or movement of the real world from the Web. The proposed method extracts the landmark keywords from each document set. Immune network model is employed to calculate the activation values of keywords, while the property of memory cell is used to find the topical relation among document sets. The immune network metaphor is also proposed to obtain the keyword map that emphasizes the topic distribution based on the landmarks. Experimental results with using a sequence of online news-article sets show the proposed method can find the topic stream through the sequence , as well as obtain the keyword map emphasizing the landmarks.

1 はじめに

検索結果やオンラインニュース記事集合など，時系列的な関連を持つ文書集合族が Web 上から多く入手可能であることに着目し，これらから話題の流れを可視化する手法について提案する．オンラインニュースを始め，Web 上に公開されている情報は，新規の話題や流行に関する情報を多く含んでおり，これらをユーザに提示することができればビジネスチャンスなどにもつながる事が期待できる．また，ユーザにとって未知の領域についての情報を収集する場合，検索エンジンを用いた情報検索プロセスを複数回繰り返すのが通常であるが [3]，これらの検索結果間の対応を明示することは，ユーザによる新たな知識の理解，概念体系の構築にも効果的であると考える．話題の検出・トラッキングに関しては，TDT(Topic Detection and Tracking) プロジェクト [1] などにおいても研究されているが，これらがイベントの同定に主眼を置いているのに対し，本研究では関連話題をゆるくつないだ形で話題の流れを抽出することを目指す点で異なる．

提案手法は，文書集合毎に可視化を行う際に，過去の可視化結果との対応付けを考慮することで時系列性を考慮する．文書集合毎の可視化は，話題分布をキーワードの空間配置により表現するキーワードマップと文書クラスタリング [4, 15] を同時に考慮するために，主要話題に関連しつつ，互いに共起しないキーワード集合を免疫ネットワークモデル [5] に基づいて抽出することにより行う [12]．文書集合間の対応付けは，既使用のランドマークを免疫記憶細胞と見なすことにより実現する．

情報可視化システムにおいては，システムの提示情報をユーザが理解する際の手がかりとなる**メタファ**が重要である [8]．これまでに，構造（階層）化された文書データベースに対しては Book メタファ [8] や入れ子構造に基づくもの [7] などが提案されているが，キーワードマップおける理解の手がかりは基本的にキーワード間の距離情報だけであり，ユーザの主観やキーワードマップ表示に対する習熟度に大きく依存する．特に，一般ユーザが利用する Web インタフェースに適用するには，可読性の更なる向上が重要であると考える．

抽出されたランドマークおよび関連キーワードで構成されるキーワードのかたまりは，文書集合中の話題を表現するものであり，これをキーワードマップ上で明確に可視化するために，免疫ネットワーク・メタファを導入する手法についても提案する．キーワードマップ作成の代表的手法であるバネモデル [13] を基に，ランドマーク関連のバネ長やバネ定数を調整することにより，ランドマークによって表現される話題分布を強調した配置が安定して得ることができる．

9 月 17 日〜9 月 21 日の間に公開されたオンラインニュース記事を対象として実験を行った結果，全文書集合を通じた話題の流れが発見可能であること，およびランドマークを強調した配置が安定して得られることを示す．

2 免疫ネットワーク・メタファに基づく情報可視化

2.1 提案アルゴリズムの概要

本稿では，キーワードマップ読解の手がかりとなるランドマークとしての性質と，文書クラスタ識別子としての性質を共に満たすキーワードを抽出するために，免疫ネットワークモデルの活性伝播機構を採用する [12]．本稿で提案する手法では，あるキーワードを共有する文書をクラスタとみなす．このキーワードをクラスタ識別子と呼ぶ．

提案手法では，キーワードを抗体，文書を抗原と見なすことにより，免疫ネットワークモデル (式 (5)–(9)) に基づいてキーワードの活性値を計算し，高活性化したものをランドマークとして抽出する．具体的な処理手順は以下の通りである．ここで，ネットワークの定常状態とは，高活性化するキーワード集合が一定となった状態とする．

1. 文書集合から，出現文書数 DF が TH_2 以上のキーワードを抽出．出現文書集合が等しいキーワードは一つにまとめる．
2. キーワード間接続強度 (J^b_{ij}) を決定.
3. キーワード・文書間接続強度 (J^g_{ij}) を決定.
4. キーワード，文書の活性値計算 X_i, A_i をネットワークが定常状態になるまで繰り返す.

ステップ 2,3 において，キーワードおよび文書間の接続強度は以下の様に定義する．

キーワード間接続強度（J^b_{ij}）

$$\textbf{強接続}\,(SC) \quad \ldots \quad CDF_{ij} \geq TH_2 \tag{1}$$
$$\textbf{弱接続}\,(WC) \quad \ldots \quad 1 \leq CDF_{ij} < TH_2 \tag{2}$$

キーワード・文書間接続強度（J^g_{ij}）

$$\textbf{強接続}\,(SC) \quad \ldots \quad TF_{ij} \geq TH_1 \tag{3}$$
$$\textbf{弱接続}\,(WC) \quad \ldots \quad 1 \leq TF_{ij} < TH_1 \tag{4}$$

ここで，CDF_{ij} はキーワード i, j が共起する文書数，TF_{ij} は文書 j 中のキーワード i の出現頻度，SC, WC はそれぞれ，強接続，弱接続の強度を表す．上記条件を満たさない場合オブジェクト間には接続関係はないものとする．

また，ステップ 4 の活性値計算には，本研究では以下に示す数理モデルを使用する [2, 6, 9].

$$\frac{dX_i}{dt} = s + X_i(f(h^b_i) - k_b) \tag{5}$$
$$h^b_i = \sum_j J^b_{ij} X_j + \sum_j J^g_{ij} A_j \tag{6}$$
$$\frac{dA_i}{dt} = (r - k_g h^g_i) X_i \tag{7}$$

$$h_i^g = \sum_j J_{ji}^g X_j \quad (8)$$

$$f(h) = p\frac{h}{(h+\theta_1)}\frac{\theta_2}{(h+\theta_2)} \quad (9)$$

ここで，X_i が抗体（キーワード）濃度，A_i は抗原（文書）濃度をそれぞれ表す (初期濃度 $X_i(0), A_i(0)$)．s は抗体の補充率，r は抗原の再生率，k_b, k_g はそれぞれ，抗体，抗原の死滅率である．h_i^b, h_i^g は field と呼ばれ，認識可能な抗原，抗体からの影響は式 (9) より，field の対数を横軸とするベル型の関数により定義される．J_{ij}^b は，抗体 i, j 間の接続強度，J_{ij}^g は抗体 i と抗原 j 間の接続強度を表す．

免疫ネットワークモデルの持つ非線形性により，共起キーワード同士は活性化しあって話題に対応したキーワードの塊を形成すると同時に，活性値が一定以上大きくなると互いに抑制しあうことにより，互いに共起しないキーワードの集合が最終的に高活性化する事が期待できる．

従って，高活性化キーワードを**ランドマーク**とすることにより，キーワードマップ上の話題の分布を理解する手がかりとなると同時に，このキーワードを含む文書単位でクラスタリングを行った場合，クラスタ間のオーバラップを避けることができる．

実際にオンラインニュース記事集合に適用した結果，生成クラスタの話題に関する結束性，ランドマークの品質に関しては，アンケート結果より，k-means クラスタリングと同等かそれ以上の評価が得られている [12]．

2.2 免疫記憶細胞モデルの導入

2.1 節で提案したアルゴリズムは，単独の文書集合に適用される．この手法を適用して，時系列的な関連を持つ文書集合族から話題の流れを発見するためには，現在の文書集合を可視化する際に，過去の文書集合から抽出・可視化された話題と類似するものがあれば優先的に抽出・可視化する必要がある．これは，一度ランドマークとして抽出されたキーワードは以降の文書集合において優先的に高活性化する様に優先権を与えることにより実現できる．

実際の免疫システムでは，一度体内に侵入した抗原については免疫記憶細胞が生成され，二度目以降の抗原提示で迅速に反応可能である（二次反応）事に着目し，本稿では，ランドマークとして抽出されたキーワードを以降の処理で免疫記憶細胞と見なす事により，上述の優先権を与える．

免疫細胞モデルについては，(1) 通常細胞よりも低い k_b を与える，あるいは (2) 式 (9) で θ_1 を小さく，θ_2 を大きくする，などにより実現可能であり，実験の結果，通常細胞と比較して 6-14 倍，高活性化しやすくなる事が示されている [11]．本稿では，(1) を採用して免疫記憶細胞モデルを導入する．

2.3 免疫ネットワーク・メタファに基づくキーワード配置アルゴリズム

文書集合中に含まれる話題分布構造を可視化し，ユーザに提示する手法として，集合中から抽出したキーワードを，文書中における共起関係などに基づき，類似性・関連性の高いほど二次元空間上で近くに配置する**キーワードマップ**が用いられることが多い [13, 14]．キーワードマップはテキストマイニングや発想支援システム [13] などでよく利用され，有効性が示されているが，以下のような問題点が存在する．

- 読解の手がかりがキーワード間距離情報のみであり，ユーザの主観やキーワードマップ表示に関する習熟度に大きく依存する．
- 多次元空間を二次元空間に写像するため，(多次元空間上での）正しい距離関係を必ずしも反映できない．
- キーワードマップでよく用いられるバネモデル [13] や多次元尺度構成法 [10] では局所最適な配置を求めるため，実行の度に異なる配置が得られることがある．

本研究では，免疫ネットワーク・メタファをキーワードマップに導入し，ランドマークに対応する話題を強調した配置を行うことを提案する．ここでは，バネモデル [13] を改良することにより免疫ネットワーク・メタファを導入する．

まず，通常のバネモデルを用いたキーワード配置は以下のようにして実現できる．

- 接続関係のあるキーワード間にバネを設定する．
- バネ定数は全て等しくする．
- 強接続のバネ長は弱接続のものより短くする．

さらに，以下の設定を行うことにより免疫ネットワーク・メタファを導入する．

- ランドマークに接続しているバネのバネ定数を大きくする．
- ランドマーク間のバネ長を長く設定する．

ランドマーク間の距離を大きくとることにより，ランドマークを中心としたキーワードのかたまりを分離して表示することが可能となる．また，ランドマーク周辺のバネ定数を強くすることにより，ランドマークを中心としたキーワードのかたまりを他よりも優先して形成することが期待できる．さらに，局所最適な配置が得られると言うバネモデルの特徴に関しても，ランドマーク周辺の配置にバイアスを加える形になるため，得られる配置のばらつきが少なくなることも期待できる．

3 評価実験結果

3.1 時系列文書集合のクラスタリング結果

前節で提案した情報可視化手法を用いてオンラインニュース記事集合を時系列的に処理した結果について示す．実験には，Yahoo! Japan News (http://yahoo.co.jp/) の「エンターテインメント」カテゴリにおいて，2001 年 9 月 17 日から 21 日の間に公開されたオンラインニュース記事を用いている．実験に使用したパラメータを表 1 に示す．

表 1: 実験に使用したパラメータ

Parameter	s	r	k_g	p
Value	10	0.01	10^{-4}	0.06
Parameter	TH_1	TH_2	$X_i(0)$	$A_i(0)$
Value	3	3	10	10^5
Parameter	θ_1	θ_2	SC	WC
Value	10^3	10^6	1.0	10^{-3}
Parameter	k_b(normal)		k_b(memory)	
Value	0.4		0.3	

表 2: 生成クラスタのランドマーク・関連キーワード（記憶細胞なし）

日付	ランドマーク	関連キーワード
9/17	公演	同時，テロ
	招待客，祝福	ダンス，タレント
	深田，恭子	バラエティー，タレント，撮影
	銀座	東京
9/18	コメディー	制作，米国
	♠♡ **公演**	日本，コンサート
	♠ 会見	都内，東京
	♠ 寄付	支援，テロ
9/19	危険	事件
	♠ 大手	出版
	大阪	容疑，所持，壱成，取締，逮捕，麻薬，いしだ，違反，大麻，捜査，拘置，請求，延長，公判，地裁，タレント
9/20	♠ 結婚	スタート
	発売	人気
	♠ ニュース	同時，事件，テロ
	説明	同時，テロ
	キャリー	被害，人気，同時，テレビ，♡ **寄付**，テロ，ロサンゼルス
9/21	社会，アニメ	
	番号	放送
	新作	公開，映画，クリス，テロ，ロック
	♠♡ **公演**	出演

このオンラインニュース記事を，同じ日に公開された記事集合毎に，提案アルゴリズムで処理を行った結果を表 2 と 3 に示す．表 2 は免疫記憶細胞モデルを利用せず，各日付毎に独立して処理した場合，表 3 は免疫記憶細胞モデルを利用したランドマークに活性化優先権を与えた場合について，生成された各クラスタのランドマークおよび関連キーワード（ランドマークと強接続のキーワード）を示している．初期集合である 9 月 17 日では免疫記憶細胞が存在しないので，表 3 では省略している．

表 3: 生成クラスタのランドマーク・関連キーワード（記憶細胞あり）

日付	ランドマーク	関連キーワード
9/18	公開	映画
	♠♡ **公演**	日本，コンサート
	♠ 会見	都内，東京
	♠ 寄付	支援，テロ
	多発	同時，テロ
9/19	♡ **多発**	ニューヨーク，同時，事件，テロ，未通し
	♠ 大手	出版
	♡ **公演**	事件，発表，ホール
	拘置，請求，延長，公判，地裁	容疑，所持，壱成，取締，逮捕，麻薬，いしだ，違反，大麻，大阪
9/20	♠ 結婚	スタート
	ドイツ	ベルリン
	♠ ニュース	同時，事件，テロ
	♡ **公演**	被害，同時，チケット，米国，テロ
	写真	
9/21	♡ **ニュース**	日本
	都内	発表
	チャリティー	歌手，同時，中枢，発表，テロ
	♠♡ **公演**	出演
	主人公	

表中で，免疫記憶細胞の有無によらず，両実験で同様に生成されたクラスタのランドマークについては，♠ マークを付与している．また，ランドマークとして使用されたキーワードが再び（ランドマークあるいは関連キーワードとして）現れた場合には ♡ で記している．

これより，ランドマークを免疫記憶細胞とすることにより，次回以降のクラスタリングの際に再びランドマークとして抽出されやすくなっていることがわかる．

実験結果の中で,「公演」が全ての記事集合からランドマークとして抽出されている．本実験で用いたニュース記事が公開された期間は，米国での同時多発テロ事件直後のため，エンターテインメントカテゴリにおいても関連記事が多数存在しており，その中には，公演の延期やチャリティー公演に関する記事も比較的多かった．提案手法では，多様な話題を発見するために，サイズの大きなクラスタの生成は抑制され，複数のクラスタに分割される傾向にある．そのため，同時多発テロ関係の記事を分割する際に，免疫記憶細胞モデルを導入した場合には一度ランドマークとして抽出された「公演」の観点が再利用される事により，文書集合族を通じた話題の流れの一つをとらえることができたと考える．

また，9 月 20 日において両実験により「ニュース」をランドマークとするクラスタが生成されているが，このクラスタに含まれる記事は，同時多発テロ事件を契機に人々がニュースに注目している事を表す，興

味深いものであった．これに対し，免疫記憶細胞モデルを利用した場合に 9 月 21 日のニュース記事集合から抽出された「ニュース」をランドマークとするクラスタは，話題としての結束性が低いものであった．9 月 20 日においては,「同時多発テロ事件」の「サブ話題」として,「ニュース」に関するクラスタに意味があったが，9 月 21 日ではその様な上位話題が存在しないにも関わらず，無理にクラスタを生成してしまったものと考える．これを防ぐためには，各文書からキーワードを抽出する際に，話題を反映したフレーズ単位で抽出するなどの工夫が必要と考えている．

3.2 免疫ネットワーク・メタファに基づくキーワード配置実験

免疫ネットワーク・メタファの導入が，キーワードマップの配置に与える効果について実験した結果について示す．パラメータは以下のように設定する．

バネの長さ:
強接続：弱接続：ランドマーク間 = 1 : 8 : 50

バネ定数: ランドマークに接続されたバネは通常のものの 5 倍

ここで検証したいのは以下の点である．

1. ランドマーク間の距離が十分にとられているか．
2. ランドマーク・関連キーワードによるキーワードのかたまりが常に安定して得られるか．

これらを評価するために，3.1 節で使用した 9/17 から 9/21 までのニュース記事集合それぞれについて，異なる初期配置（ランダムに決定）から 5 回バネモデルによる配置を実行した際の，キーワード間の距離の平均値（AVG），標準偏差（STD），両者の比率（STD/AVG）について求めた．バネモデルによる座標計算は，収束具合などを考慮して，1,500 回に実験的に決定した．表 4 は接続のある全ノード間の距離について，表 5 はランドマーク間距離について，表 6 はランドマークと関連キーワード間の距離についての実験結果について，それぞれまとめたものである．IM は免疫ネットワーク・メタファを導入した場合，NML は通常のバネモデルに基づいて配置を行った場合の結果である．

これらの結果より，全ノード間については両手法に差はそれほど見られないが，ランドマーク間については免疫ネットワーク・メタファを導入することにより，通常のバネモデルよりも 2 倍程度，距離を大きく保てていることがわかる．また，STD/AVG の値より，免疫ネットワーク・メタファを導入することにより，ランドマーク・関連キーワードによるキーワードのかたまりに関して，実行毎の配置のばらつきが少ないことがわかる．

免疫ネットワーク・メタファを導入したキーワードマップの例として，9/19 の文書集合に免疫記憶細胞を用いて提案手法を適用した結果を図 1 に示す．図において，白い矩形のものがランドマーク，濃い色の

表 4: 全ノード間距離の標本平均，標本偏差，比率

Date	Type	AVG	STD	STD/AVG
0917	IM	113.40	10.197	0.13101
	NML	104.53	10.867	0.15652
0918	IM	83.043	21.093	0.28056
	NML	67.083	23.748	0.37127
0919	IM	108.11	2.9122	0.04172
	NML	98.100	15.412	0.22634
0920	IM	101.06	16.309	0.21441
	NML	92.156	16.492	0.23857
0921	IM	78.040	23.330	0.32437
	NML	73.733	21.876	0.32757

表 5: ランドマーク間距離の標本平均，標本偏差，比率

Date	Type	AVG	STD	STD/AVG
0917	IM	224.23	22.487	0.09997
	NML	104.66	21.248	0.23679
0918	IM	194.10	45.859	0.24082
	NML	68.552	32.709	0.56259
0919	IM	231.44	0.29962	0.00136
	NML	129.17	8.0741	0.08742
0920	IM	209.03	32.641	0.16548
	NML	103.53	33.100	0.39004
0921	IM	190.53	46.334	0.24419
	NML	68.912	42.187	0.64712

表 6: ランドマーク・関連キーワード間距離の標本平均，標本偏差，比率

Date	Type	AVG	STD	STD/AVG
0917	IM	93.138	7.8363	0.08980
	NML	46.737	7.2387	0.21191
0918	IM	85.904	7.5271	0.10911
	NML	43.424	17.717	0.41925
0919	IM	71.130	0.57815	0.01145
	NML	55.835	5.3415	0.13120
0920	IM	105.25	6.5842	0.07204
	NML	70.125	7.7490	0.16809
0921	IM	78.379	10.167	0.13507
	NML	33.471	10.117	0.31009

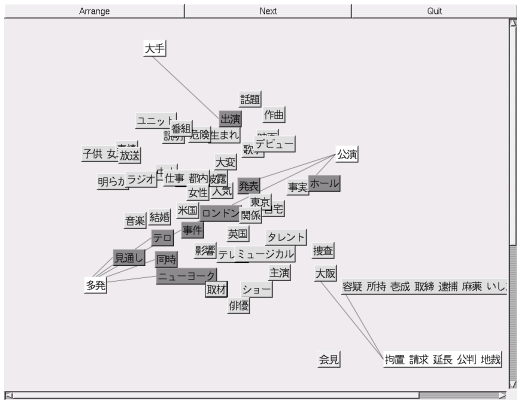

図 1: 9/19 のキーワードマップ（免疫記憶細胞あり）

ものが関連キーワードを表す．エッジはランドマークとの強接続のみ示している．これより，関連キーワードの多くはテロに関するものであり，多数の文書に出現する影響で中央に寄った配置となっているものの，各ランドマーク間の距離を大きくとることにより，それぞれが表す話題が明確になっていることがわかる．

4 まとめ

検索結果やオンラインニュース記事集合など，時系列的な関連を持つ文書集合族が Web 上から多く入手可能であることに着目し，これらから話題の流れを可視化する手法について提案した．提案手法は，文書集合毎に可視化を行う際に，過去の可視化結果との対応付けを考慮することで時系列性を考慮する．文書集合毎の可視化は免疫ネットワークモデルに基づくランドマーク抽出によって行い，免疫記憶細胞モデルを導入することにより時系列的関連を考慮している．9 月 17 日〜9 月 21 日の間に公開されたオンラインニュース記事を対象として実験を行った結果，全文書集合を通じて類似話題が発見可能であることを示した．

また，免疫ネットワーク・メタファを用いたキーワードマップ作成に関しても，従来バネモデルよりもランドマークを強調したキーワードマップを安定して作成可能であることを実験により示した．

今後は，実ユーザを対象とした，可視化インタフェースの評価を行い，提案手法の有効性を検討していく予定である．

参 考 文 献

[1] J. Allan, R. Papka, V. Lavrenko, “On-line New Event Detection and Tracking,” Proc. 21st annual Int’l ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 37–45, 1998.

[2] R.W. Anderson, A. U. Neumann, A. S. Perelson, “A Cayley Tree Immune Network Model with Antibody Dynamics,” Bulletin of Mathematical Biology, Vol. 55, No. 6, pp. 1091–1131, 1993.

[3] C. Cole, “Interaction with an Enabling Information Retrieval System: Modeling the User’s Decoding and Encoding Operations,” Journal of the American Society for Information Science, Vol. 51, No. 5, pp. 417–426, 2000.

[4] M. A. Hearst and J. O. Pedersen, “Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results,” SIGIR’96, pp. 76–84, 1996.

[5] N. K. Jerne, “The Immune System,” Sci. Am., Vol. 229, pp. 52–60, 1973.

[6] A. U. Neumann and G. Weisbuch, “Dynamics and Topology of Idiotypic Networks,” Bulletin of Mathematical Biology, Vol. 54, No. 5, pp. 699–726, 1992.

[7] J. Rekimoto and M. Gree, “The Information Cube: Using Transparency in 3D Information Visualization,” Proc. 3rd Annual Workshop on Info. Tech. & Sys. (WITS’93), pp. 125–132, 1993.

[8] 柴山，ブラウザのための可視化とナビゲーション支援, 人工知能学会誌, Vol. 16, No. 4, pp. 509–514.

[9] B. Sulzer et al., “Memory in Idiotypic Networks Due to Competition Between Proliferation and Differentiation,” Bulletin of Mathematical Biology, Vol. 55, No. 6, pp. 1133–1182, 1993.

[10] 角，堀，大須賀，テキストオブジェクトを空間配置することによる思考支援システム, 人工知能学会誌, Vol. 9, No. 1, pp. 140–147, 1994.

[11] Y. Takama and K. Hirota, ”Consideration of Memory Cell for Immune Network-based Plastic Clustering method,” InTech’2001, 409–414, 2001.

[12] 高間，廣田，WWW 上の情報収集／可視化のための免疫ネットワークを用いたクラスタリング，第 46 回人工知能基礎論研究会資料, pp. 61–66, 2001.

[13] 高杉，國藤，スプリングモデルを用いたアイデア触発のための思考支援システムの開発，人工知能学会誌, Vol. 14, No. 3, pp. 495–503, 1999.

[14] ビジュアルテキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 226–232, 2001.

[15] O. Zamir and O. Etzioni, “Grouper: A Dynamic Clustering Interface to Web Search Results,” Proc. 8th Int’l WWW Conference, 1999.