6B−2−04

Web上の時系列文書集合族を対象とした情報可視化システム*

高間　康史[1,2†]　廣田 薫[1‡]

[1]東京工業大学 大学院総合理工学研究科[§]
[2]科学技術振興事業団[¶]

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．

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に基づいて抽出することにより行う[4]．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．

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す．

*Information Visualization Designed for Sequence of Document Sets on Web
†Yasufumi Takama
‡Kaoru Hirota
§Inderdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
¶PREST, Japan Science and Technology Corporation

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[4]．具体的には，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する．

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}）

$$\text{強接続}\ (SC)\ldots CDF_{ij} \geq TH_2 \qquad (1)$$

$$\text{弱接続}\ (WC)\ldots 1 \leq CDF_{ij} < TH_2 \qquad (2)$$

キーワード・文書間接続強度（J_{ij}^g）

$$\text{強接続}\,(SC)\ldots TF_{ij} \geq TH_1 \quad (3)$$

$$\text{弱接続}\,(WC)\ldots 1 \leq TF_{ij} < TH_1 \quad (4)$$

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, 3].

$$\frac{dX_i}{dt} = s + X_i(f(h_i^b) - k_b) \quad (5)$$

$$h_i^b = \sum_j J_{ij}^b X_j + \sum_j J_{ij}^g A_j \quad (6)$$

$$\frac{dA_i}{dt} = (r - k_g h_i^g) X_i \quad (7)$$

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

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

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 [4].

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 [5]．本稿では，(1) を採用して免疫記憶細胞モデルを導入する．

3 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[1] 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は表 1 に示す．

[1] http://yahoo.co.jp/

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	

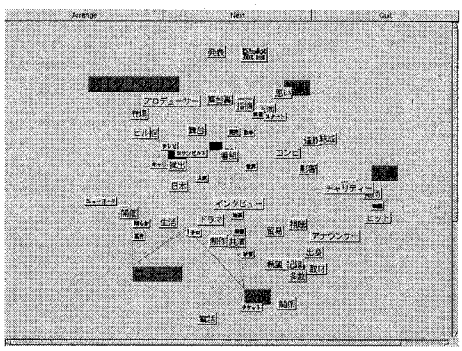

図 1: 9/20 のキーワードマップ（免疫記憶細胞あり）

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[2]．また，図 1 は 9 月 20 日に公開されたオンラインニュース記事集合から作成したキーワードマップを示す．大きいフォントで色つきのキーワードはランドマーク，色つきだがフォントの小さいものはかつてランドマークとして使用されたものを示す．

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

[2]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 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

強接続のキーワード）を示している．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の直後であり，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，免疫記憶細胞モデルを導入した場合には，同時多発テロ関係の記事を分割する際に，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，

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	

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．

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基づいており，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．

今後は，ユーザにとってより可読性の高い可視化インタフェースの考察・実装を行うとともに，評価方法の検討を行う予定である．特に，評価方法に関しては TDT(Topic Detection and Tracking) プロジェクト [1] などとの関連も調査する予定である．

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

[4] 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.

[5] Y. Takama and K. Hirota, “Consideration of Memory Cell for Immune Network-based Plastic Clustering method,” 2nd Int’l Conf. on Intelligent Technologies (InTech’2001), pp. 409–414, 2001.