知 能 と 複 雑 系 128－17
（２００２．５．２４）

免疫システムを用いた関連文書収集方法の提案

織　田　　充†

過去注目した文書を抗体とする免疫ネットワークを構成し，新たに注目するキーワードあるいは文書を基に得られる検索結果から，多様性を考慮した関連文書を収集する方法を提案する．

Relevant Document Gathering based on the Immune System

Mitsuru Oda†

In this paper the method to collect the related documents in consideration of diversity from the retrieval result based on the use of the keyword and the document is proposed. To evaluate the relevancy of the document in the retrieval result to the document referred by a user, this method uses the immune network model that treats the document as an antibody.

1. はじめに

近年みられるインターネットの爆発的な普及により，インターネットを通じて多種多様かつ膨大な情報にアクセス可能になり，検索システムを用いることで，様々な分野の情報を安価に利用することが可能になった．一方でインターネットの爆発的な普及は，インターネット上の情報の利用者を一般層にまで拡大させている．このため，一般な利用者とインターネット上の多種多様な内容を持つ情報を結びつける支援は，さらに重要度を増している．本研究では，一般な利用者が情報収集過程において，検索システムを用い収集，参照した文書の持つ情報に影響されつつ，情報収集を繰り返すことに注目し，利用者が過去参照した文書の持つ情報内容に沿いつつ，情報内容の多様性を考慮した関連文書を，検索結果から抽出する方法を提案する．

一般な利用者は，情報収集の開始時点において，ある特定の興味に基づく明確な情報要求を持っているというより，むしろ漠然とした情報要求を持っているのが自然な状態である．このため一般な利用者は，検索過程で使用した検索キーワード，得られた検索結果，参照文書等に含まれる情報に触発され，自身の漠然とした情報要求を明確化，また新たな情報要求へ変化させることで，多段な検索を実行していると考えられる．また一般な利用者は，自身の求める情報の分野に不慣れである場合が多く，検索システムに対して，適切な検索要求を一時に与えられるとは限らず，さらに誤った検索要求をさえ与える場合も考えられるため，一般な利用者にとって，検索要求は参照文書を得るためのトリガーである．

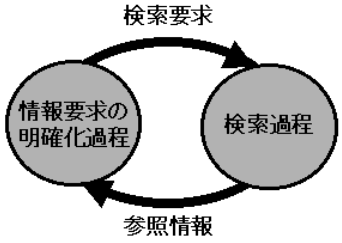

図 1　情報収集過程

これらが原因となり，一般な利用者が行う情報収集過程は，自身の情報要求を検索システムへ提示する検索要求に翻訳する過程である**情報要求の明確化過程**と，検索結果を基に文書参照し情報収集する過程である**検索過程**が，交互に出現しつつ進行する過程（図 1）となる．ネットサーフィンは，利用者が情報収集過程で検索を繰り返し行う中，遭遇する情報が与える影響を，自身の情報要求の変化へ積極的に利用した，発見的な情報収集スタイルである．

一方，近年検索システムが扱う文書量は膨大なものであり，利用者が的確な検索要求を与えられないと，検索結果に多量な適合文書が含まれてしまう場合がある．多くの検索システムでは，利用者がこれら多くの適合文書から参照文書を選択する負荷を軽減する目的

† 九州システム情報技術研究所
Institute of Systems & Information Technologies/ KYUSHU

から，検索結果に含まれる適合文書をランキングし，利用者に提示している．しかし，多くの利用者はランキング上位の適合文書だけを参照する傾向があり，下位に出現する適合文書に有益な情報が含まれていた場合，発見できない．この結果，利用者が情報収集過程で遭遇した情報に，利用者自身の情報要求自体が影響されることにより，インターネット上の多種多様な情報を含む文書を有効に利用できなくなる危険性がある．

したがって，一般な利用者が行う情報収集過程を支援するシステムは，情報収集過程において行われる一連の情報要求の明確化過程，検索過程から，利用者の情報要求の変化に沿った適合文書の提示を行う必要があり，かつ常に利用者が偏った情報内容を持つ文書を参照し続けることによる機会損失を回避するために，ランキング上位に出現する適合文書群が，その情報内容に関して多用性を持つことが必要である．さらに，その多様性は情報の利用者の興味の方向，すなわち情報要求に対しての多様性である必要がある．したがって，利用者の情報要求に適合性がよく，かつ情報内容の多様性が保持されるランキングを行えることが重要である．

このような一般な利用者が行う情報収集過程を支援を実現するために，

- 検索過程におけるキーワード入力，あるいは文書参照といった利用者の行動履歴から，現時点における利用者の情報要求の推定，
- 推定された情報要求に適合し，かつ全体として文書の情報内容が偏らない多様性を備えた文書群を選択するための適合文書評価

を行う機能を備えた支援システムが必要である．

2. 従来技術の問題点

2.1 従来情報収集過程支援方法の問題点

Taylor[6] による分類によれば，情報収集過程における利用者の情報要求の状態として，下記４段階が挙げられている．

(1) **直観的要求**：現状に不満であることは認識しているが，具体的な表現に言語化し説明できない状態，
(2) **意識された要求**：問題を意識できるが，うまく言語化できない状態，
(3) **形式化された要求**：問題を具体的に言語化可能な状態，
(4) **調整済みの要求**：問題解決に必要な情報源を同定できるほど問題が具体化されている状態

漠然とした情報要求しか持たない利用者の状態とは，第１段階と第２段階の中間にある利用者に対応している．情報要求の明確化過程は，Taylor の情報要求の分類における第１段から第４段へ至る過程に対応する．

従来検索過程支援として，代表的な検索質問拡張（query expansion）および適合性フィードバック（relevance feedback）に注目してみる．

検索質問拡張は，利用者が検索システムに提示する検索要求に対して，検索キーワードを補足するなどすることで，検索要求を自動的に補完する手法である．Taylor の分類に従えば，検索質問拡張は，第３段階にある利用者を第４段階に上げるための支援であり，第３段階まで上げる支援は行なわれていない．

また，適合性フィードバックは，利用者が自身の検索要求に対する文書の適合性判定を行い，検索システムはその結果を基に次点での検索式を再構成することを繰り返し，段階的に適合文書候補を絞り込みを行うことで，情報収集過程を支援するものである．したがって，適合性フィードバックでは，利用者の段階的な情報要求の明確化を行なっている．しかし，適合性フィードバックで想定されている利用者は，検索システムが提示する適合文書が自身の情報要求に対して適合するか否かという適合性判定ができることを前提しているため，参照情報により利用者の情報要求そのものが影響され，変化する可能性は考慮されていない．すなわち適合性フィードバックは，検索質問拡張と同様に，Taylor の分類での第３段階にある利用者を第４段階に上げるための支援であり，同じく第３段階まで上げる支援は行なわれていない．さらに Taylor が指摘した情報要求の第１段階あるいは第２段階にある利用者である，漠然とした情報要求しか持っていない利用者に対して，適合性フィードバックを用いた支援を行うと，一貫した適合性判定を行なわない可能性があるため，結果として誤った方向に誘導してしまう危険性がある．

2.2 従来適合文書評価方法の問題点

次に検索結果に現れる文書評価方法であるランキング手法の観点から，従来技術の問題点を述べる．

tf-idf を利用したランキング手法では，検索キーワードの文書中での出現頻度により評価されランキングがなされ，それ以外の文書の持つ情報内容に依存していない．したがって検索結果に現れる適合文書の持つ情報内容に関しては，なんら言及しない点で中立的なランキング手法である．これに対して，Google 等の検索システムで採用されている PageRank[4] や HITS[5]

など文書間参照性に基づく適合度評価によるランキングの場合，適合文書に関心度が高い文書が含まれていると特定の情報に関する興味を共有する多数の情報提供者，すなわち強力なコミュニティの存在により特定の内容を持つ文書が上位にランキングされたならば，その他の情報を有する文書は下位にランキングされてしまい，上位に出現する適合文書の情報内容が偏る可能性が大きい．したがって情報要求を明確化していない一般の利用者に対するランキング手法としては注意が必要である．

このように従来のランキング手法では，検索結果に現れる適合文書の有する情報内容間の関連性は適合度評価に反映せず，適合文書の多様性が適合度評価に考慮されているとは言えない．

3. 免疫ネットワークモデルを用いた文書評価

3.1 文書ネットワークによる情報要求のモデル化

ここでは本研究における利用者の情報要求を推定する手法の概要を述べる．

本研究では，利用者の情報要求を推定するために，利用者の情報収集過程で出現する参照文書群に注目する．情報収集過程の現時点における参照文書は，現時点における利用者の情報要求に照らし合わせ，利用者が有益な情報を含むと判断した文書と見なすことができる．情報収集過程に現れた一連の参照文書のうち，現時点における参照文書に類似な文書は，現時点における情報要求に対して有益な情報を含む可能性が高く，逆に非類似な文書はその可能性が低い．このことを用い，いま文書集合で，それに含まれる文書間が情報内容の類似性で関連付られているものを**文書ネットワーク**と呼ぶ．また参照文書から構成される文書ネットワークを，特に**参照文書ネットワーク**と呼ぶ．参照文書ネットワークを構成する各参照文書に対して，現時点における情報要求に対する有益さの程度を表す値（以下，評価値）を付値する．この参照文書ネットワークに対して，現時点における参照文書を追加する際，現時点での参照文書に対して類似な参照文書の評価値の値を上げ（活性化），また非類似な参照文書に関してはその評価値を下げる（非活性化）という操作（活性伝播）を行うことで得られる参照文書ネットワークを，現時点での利用者の情報要求を近似モデルとして扱う（図 2）．これはちょうど利用者に対して，どの文書がどの程度良いかを示してもらうことで，利用者がどのような内容の文書を良いと判定しているかを推定していることに対応する．

ただし，現時点における情報要求に対して，現時点での参照文書は有益な情報を含むと考えられるが，不要な情報もまた含む場合がある．したがって単純に類似性を推移的に扱って，参照文書の評価値を決定するのは危険である．また，次点での情報要求は現時点のものとは異なるものになる可能性もある．したがって前記活性伝播により，現時点での参照文書に類似な他の参照文書の活性が著しく上がったり，逆に非類似な参照文書の活性が著しく下がるといった過度な推定により，推定可能な情報要求の多様性が失われるのは好ましくない．参照文書ネットワークでの参照文書の評価値を決定する活性伝播の方法は，これらことが避けられることが必要である．

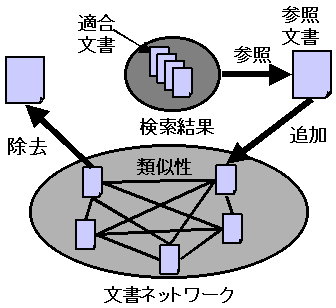

図 2 文書ネットワーク

本研究では，文書ネットワークにおける活性伝播の方法として，生体の免疫システムにおける抗体間の相互作用の数学的モデルである免疫ネットワークモデルを用いる．

3.2 免疫ネットワーク

生体が有する免疫システムは，遭遇した抗原の情報を記憶（免疫記憶）し，以降同じ抗原を検知すると素早い抗体産出（免疫応答）を行うことができる．Jerne[8)9)10)11)] が提案した免疫ネットワーク（イディオタイプネットワーク）説によれば，個々の抗体は他の抗体に対する抗原として機能し，その結果，抗体間の刺激・抑制による相互作用による抗体ネットワークが構成される．この抗体ネットワーク上で，抗体間の相互作用による自己フィードバックが生じ，その結果，抗体ネットワークがある平衡状態に達する（図 3）．この抗体ネットワークの平衡状態により，免疫記憶が実現されているとしている．新たな抗原と遭遇により，抗体ネットワークの平衡状態が動的に変化し，侵入した抗原が消滅した後においても，変化した平衡状態が維持されることで，免疫記憶の更新がなされる．この際，抗体間の刺激・抑制による相互作用により，高活性あるいは低活性な抗体の増加が抑制されるように調整される．この結果，免疫システムが特定の抗原に対し

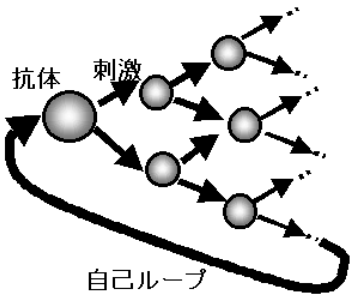

図 3 抗体ネットワーク

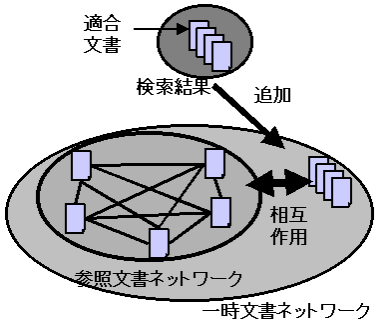

図 4 文書ネットワークによる適合文書評価

て過敏になること防ぎ，また，免疫応答に寄与していない抗体を消滅させること（免疫忘却）が可能になる．

Farmer[12)13)] が提案した抗体の濃度変化を表す微分方程式を以下に示す．

$$\frac{dx_i}{dt} = c\,(A_1 x_i - k_1 A_2 x_i + A_3 x_i) - k_2 x_i \quad (1)$$

$$A_1 = \sum_{j=1}^{N} m_{ji} x_j \quad (2)$$

$$A_2 = \sum_{j=1}^{N} m_{ij} x_j \quad (3)$$

$$A_3 = \sum_{j=1}^{M} m'_{ji} y_j \quad (4)$$

ただし上式において，x_i は抗体 i の濃度，y_j は抗原 j の濃度，N は抗体の総数，M は抗原の総数，m_{ij} は抗体 i を抗原とした場合の抗体 j との反応の強度，m'_{ij} は抗原 i と抗体 j との反応の強度，c は衝突による抗体生産の比率，k_1 は刺激と抑制の差に対応する係数，k_2 は抗体が自然消滅する比率を表す係数である．

3.3 免疫ネットワークを応用した文書ネットワークの構築

本研究では，文書を抗体に，文書間類似度を抗体間反応強度に，文書評価値を抗体濃度に対比させることで，抗体ネットワークの枠組みを文書ネットワークへ応用する．文書ネットワークには，抗原に対応する対象がなく，文書評価値の変化をもたらすのは，抗体に対応する文書のみである．このため，Farmer が示した式から抗原が関与する項を除いた，次式を文書評価値の計算に用いる．

$$\frac{dx_i}{dt} = c\,(A_1 x_i - k_1 A_2 x_i) - k_2 x_i \quad (5)$$

ただし，x_i は文書 i の評価値，N は文書ネットワークに出現する文書の総数，m_{ij} は文書 i と文書 j との類似度とし，文書 i, j に出現する単語を次元，出現頻度を値とするベクトル間の内積相関値で与える．したがって文書間類似度には対称性（$m_{ij} = m_{ji}$）が成立する．

以下に文書ネットワークでの文書評価値の計算手順を示す．

(1) 文書ネットワークに追加される文書それぞれに対する評価値として，初期評価値を与える，
(2) 文書ネットワークに出現する文書，および文書ネットワークに追加される文書それぞれに対して，文書中に出現する単語を次元，頻度をその値とするベクトルを生成する，
(3) 各ベクトル間の内積相関値を計算し，対応する文書間の類似度とする，
(4) 文書ネットワークに出現する文書，および文書ネットワークに追加される文書それぞれに対する評価値を，式 (2), (3), (5) に従って更新する，
(5) 更新後の評価値が閾値以下の文書を，文書ネットワークから除去する．

3.4 多様性を保持した関連文書抽出

本研究で提案する関連文書収集方法の概要は以下の通りである．

前記した文書ネットワークの構築方法に従い，利用者が情報収集過程において参照した一連の文書から，逐次文書ネットワークへ追加することにより構成された参照文書ネットワークに対して，検索結果に含まれる適合文書を追加し，同様に前記した文書ネットワークの構築方法に従い，一時文書ネットワークを更新する．この更新後の一時文書ネットワークにおける適合文書の評価値を現時点における検索要求に対する適合度とする（図 4）．

本研究で提案する関連文書収集方法の手順を以下に述べる．なお手順中のステップ番号は，手順を表す図 5 においてボックスで示された操作の番号に対応している．またステップ 1 は利用者が要求を発することで

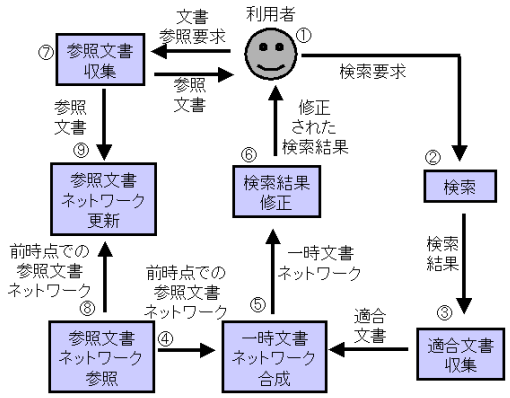

図 5 関連文書抽出手順

実行される処理である．

(1) 利用者の要求が検索要求であればステップ２へ，また文書参照要求であればステップ７へ移る，
(2) 利用者から送られた検索要求に従い検索を実行する，
(3) 検索エンジンから得られた検索結果に出現する適合文書を収集する，
(4) 一時文書ネットワークを現時点での参照文書ネットワークに置き換える，
(5) 検索結果に現れる適合文書を一時文書ネットワークに追加後，前記計算手順に従い文書評価値を計算し，一時文書ネットワークを更新する，
(6) 得られた一時文書ネットワークにおける，検索結果に現れる適合文書の評価値に従い，検索結果のランキングを修正し，その結果を利用者に提示し，ステップ１へ戻る，
(7) 利用者から送られた文書参照要求に従い，文書を収集し利用者に提示する，
(8) 現時点での参照文書ネットワークを参照する，
(9) 利用者の参照文書を参照文書ネットワークに追加し，前記計算手順に従い文書評価値を計算し，参照文書ネットワークを更新した後，ステップ１へ戻る，

上記手順に現れる処理ステップのうち，ステップ７，８，９が利用者の情報要求の推定処理に対応し，ステップ２から６が多様性を考慮したランキングによる関連文書収集支援に対応する処理である．

4. 情報収集支援システム

以上の考察を踏まえ，汎用検索エンジンが返す検索結果に含まれる適合文書を，利用者の参照文書群から推定される情報要求に基づき，多様性を考慮したランキングを行なう情報支援システムを設計した（図 6）．

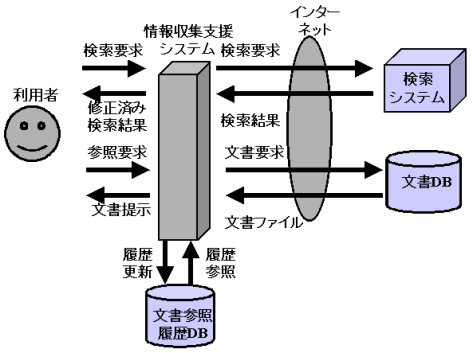

図 6 情報収集過程支援システム

この情報支援システム自体は，利用者側の計算機上で稼動する．これは既存の汎用検索システムを用いることで，多種多量な文書を利用できる点，利用する検索システムを切り替えることで，特定の検索システムの判定基準に従った適合文書群を利用しなくても済む点，また参照文書履歴という利用者の個人情報を利用するため，利用者のプライバシーに配慮したためである．

5. 関 連 研 究

情報収集過程で得られた情報により，情報要求自体が変化する可能性のある一般な利用者に対する支援方法で，免疫ネットワークを用いた手法が，高間ら [4] により提案されている．しかし高間らが提案する手法は，一連の検索過程で得られた検索結果群に含まれる文書群における話題分布を，それら文書群に現れるキーワードの多様性を考慮しクラスタリングし，利用者に対して提示することで，利用者の参照情報の選択を支援するものである．言い換えれば，提供される情報の多様性が，利用者に対して提示されるのである．

これに対して本研究は，利用者の持つ情報要求の多様性に注目し，検索結果に対して利用者の情報要求をバイアスとした多様性を考慮したフィルタリングを行うことで，多様性が考慮されたランキングに検索結果を再編するのに利用しているという違いがある．

6. ま と め

本研究では，情報要求が漠然とした利用者に対して，多様性を重視した文書評価ランキング手法を提案した．今後はシュミレーション，実データを用いた実験を通して，その有効性を検証する．

本研究では，情報要求が漠然とした一般な情報利用者に対する，検索システムを用いた情報収集支援を目

的に，過去参照した文書に関連する文書を，検索結果に出現する文書群から選択するための多様性を重視した文書評価方法を提案した．その本質的な難しさは，利用者自身の情報要求が漠然としている，言い換えれば，情報の有益無益の判定基準が曖昧である場合，支援システムが提示する推薦情報により，利用者自身の情報要求の方向性が偏向されてしまう危険性にある．これは推薦システム一般が持つジレンマである．本研究では，免疫システムの持つ多様性を保持しつつ，環境に適合する機能を応用することで，対処している．

また抗体ネットワークが生体が過去経験した抗原記憶であるように，参照文書ネットワークは，情報利用者の情報収集過程の記憶に対応している．したがって，受動免疫に見られる免疫獲得方法のように，ある個体が獲得した抗体を他の個体へ移行すること（移行抗体）で，経験していない抗原に対する免疫的経験を伝達することができるように，他利用者の参照文書ネットワークを流用することで，他利用者の情報収集過程における経験を，自身の情報収集過程支援において再生し利用することができる．本研究の枠組みを応用することで，利用者の情報収集の経験を共有した協調的検索に繋がると期待できる．

参 考 文 献

1) 徳永健伸: 情報検索と言語処理, 言語と計算５, 東京大学出版会 (1999).
2) 石田好輝 (編): 免疫型システムとその応用, コロナ社 (1998).
3) 伊庭斎志: 進化論的計算の方法, 東京大学出版会 (1999).
4) 高間康史, 廣田薫: 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, 人工知能学会研究会資料 SIG-FAI/KBS-J-10, pp. 61–66 (2001).
5) Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd: The PageRank Citation Ranking: Bringing Order to the Web, online manuscript, *http : //www − db.stanford.edu/ backrub/pageranksub.ps*.
6) J. Kleinberg: Authoritative sources in a hyperlinked enviroment, Proc. ACM-SIAM Symposium on Discrete Argorithems, (1998), *http : //www.cs.cornell.edu/home/kleinber/auth.ps*.
7) R. S. Taylor.: Question-negotiation and information seeking in libraries, College & Research Libraries, 29(3), pp. 178–194 (1968).
8) N. K. Jerne.: The Immune System, Sci. Am., Vol.51, No.5, pp. 52–60 (1973).
9) N. K. Jerne.: Towards the network theory of the immune system, Ann. Immunol. (Inst. Pasteur), 125C, 373389 (1974).
10) N. K. Jerne.: Idiotypic network and other preconceived ideas, Immunological Rev., 79, 524 (1984).
11) N. K. Jerne.: The generative grammer of the immune system, EMBO Journal, 4-4 (1985).
12) J. D. Farmer, N. H. Packard, and A. S. Perelson.: The Immune System Adaptation and Machine Learning, Physics, 42D, North Holland (1986).
13) J. D. Farmer.: A Rosetta Stone for Connectionism, Physics, 42D, North Holland (1990).