第1回 MYCOM 資料
2000年5月15日,16日

情報検索作業を通じた話題分布構造可視化の提案

高間 康史　　廣田 薫

東京工業大学大学院総合理工学研究科
{takama, hirota}@hrt.dis.titech.ac.jp

Abstract

検索対象領域における話題分布構造を可視化する手法を提案する．提案手法はクエリーをノードとするネットワークを，検索作業履歴，および検索結果文書中の話題分布に応じて漸進的に構築していく．免疫ネットワークとのアナロジーによる提案手法の実現可能性についても議論し，予備実験結果より免疫システムの持つ多様性が利用可能であることを示す．

1　はじめに

ユーザが情報検索作業を通じて，検索対象領域に関する知識を漸進的に獲得する過程を支援するシステムの開発を目標とし，検索対象領域における話題分布構造を可視化する手法を提案する．提案する手法は，クエリーをノードとするネットワーク (クエリーネットワーク) を，検索作業を通じて漸進的に構築していくものであり，作業履歴，検索結果文書中の話題分布の両者を考慮してネットワーク状態を動的に更新する．本稿では，クエリーネットワークの提案および概要を説明するとともに，免疫ネットワークを用いたモデル化の可能性を，予備実験の結果も踏まえて考察する．

クエリーネットワークの概要については 2節，免疫ネットワークモデルを用いたクエリーネットワークのモデル化については 3節でそれぞれ述べる．提案手法に関する予備実験として，キーワード抽出を行った結果については 4節にまとめる．

2　クエリーネットワーク：話題分布構造可視化手法の概要

WWW 空間の爆発的増大，および WWW ブラウザなどの普及により，我々の知的作業は，WWW 情報検索から始まる様になったと言っても過言ではない．ユーザは，あらかじめ明確な検索要求を持つことなしに検索を行うことが通常であり，むしろ検索作業を繰り返し，試行錯誤することによって，検索対象領域に関する知識を得，そこに存在する話題，有効なキーワードなどを把握しながら検索要求を明確化していくと言える．従来からの文献検索研究では，特定ページの発見が主目的であり，検索作業を通じて漸進的に獲得される，対象領域全体に渡る情報は副産物としてあまり注意が払われていなかった．この様な副産物的な情報に着目し，これをユーザに提示し，対象領域に関する知識獲得の支援につなげようというのが本研究の目的である．特に，WWW 空間は膨大かつ動的であり，新しい話題，流行などに敏感であるため，情報検索作業を通じた知識獲得効果は非常に高い事が期待できる．

本研究で提案するクエリーネットワークは，数個のキーワードからなるクエリーをノードとし，共有キーワードの有無や検索可能文書集合の重なりといった相関関係に基づいてノード間を接続する事により構築する．ここで，話題とはユーザの検索要求 (クエリー) に対応するような，数個のキーワードで表現できるものと定義する．この仮定より，以下の手順に従い，情報検索作業を通じて漸進的に構築されるクエリーネットワークの状態 (構造および各ノードの活性状態) は，検索対象領域における話題分布構造を表現できることが期待できる．

1. 検索文書集合からキーワードを抽出する．

2. 抽出されたキーワードから新規クエリーを生成し，既存のクエリーネットワークに追加する．

3. 各クエリーの活性度を計算する．現在の検索結果に関連したクエリーほど強く活性化する様にする．

4. 活性度の低いクエリーをネットワークから除去することにより，ネットワークサイズの爆発を押さえ，可読性を維持する．

5. ユーザはネットワークからクエリーを選択し，サーチエンジンへ送ることにより，新たな検索プロセスを開始する (1. へ戻る)．

3 免疫ネットワークに基づくクエリーネットワークのモデル化

提案したクエリーネットワークを用いて，情報検索空間における話題分布構造を可視化するためには，(1) 漸進的かつ柔軟な組織化，(2) ユーザの検索作業履歴の保持，(3) 話題分布における多様性の維持，の３特性を兼ね備える必要がある．(1) は，ユーザによる情報検索作業 (クエリー生成) と，検索結果からの知識獲得作業はスパイラルを描く事を考慮したものである．既存の情報検索に関する研究では，検索作業を通じてユーザの検索要求が明確化されていく過程については着目しているものの，知識獲得，情報可視化の観点からこの様なインタラクションを通じた漸進性を強く意識した研究は少ない．例えば Grouper[5] では，検索結果にクラスタリングを適用する事により，検索結果における話題分布を考慮していると言えるが，クラスタリングは各検索プロセス毎に独立であり，漸進的構造化は考慮されない．

(2) に関して，検索履歴はユーザにとっても明示的に保持される方が好ましいとの考えに基づき，本研究ではクエリーを基本単位としてネットワークを構築する手法を採用する．また，従来研究のように，ユーザの興味に合致した特定ページの発見が目的である場合には (3) は重要ではないが，話題分布構造可視化においては非常に重要な特性である．

生体におけるマルチエージェントシステムと言われる免疫システムは，これらの特性を全て備えているため，本研究では免疫ネットワークモデルによりクエリーネットワークをモデル化する事を提案する．すなわち，クエリーを抗体 (B-Cell)，検索結果文書を抗原と対応づける事によりモデル化を行う．

3.1 免疫ネットワークモデル

免疫ネットワークモデルは，免疫システムのうち，特に抗体 (B-Cell) 間の相互作用をモデル化したものであり，1970 年代に Jerne[2] によって提案されて以降，数理生物学の分野などでそのダイナミクスの解析が行われている [3]．数理モデルで一般に使用される，抗原および抗体の濃度更新に関する微分方程式を以下に記す．ここで，X_i, A_i はそれぞれ抗体，抗原の濃度 (クエリーの活性値に相当) であり，s は抗体の補充率，r は抗原の再生率，d, k 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 (5) 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 間の親和度，J_{ij}^g は抗体 i と抗原 j 間の親和度を表す．抗体間の相互作用 (式 (5)) は，接続している抗体からの影響が一定量を越えると，促進作用から抑制作用に変わるという性質を持っており，これは免疫ネットワークが多様性を持つ上で重要な役割を果たしていると考えられている．

$$\frac{dX_i}{dt} = s + X_i(f(h_i^b) - d), \tag{1}$$

$$h_i^b = \sum_j J_{ij}^b X_j + \sum_j J_{ij}^g A_j, \tag{2}$$

$$\frac{dA_i}{dt} = (r - kf(h_i^g))X_i, \tag{3}$$

$$h_i^g = \sum_j J_{ji}^g X_j, \tag{4}$$

$$f(h) \quad = \quad p\frac{h}{(h+\theta_1)}\frac{\theta_2}{(h+\theta_2)}, \tag{5}$$

4 キーワード抽出に関する予備実験

3節で紹介した免疫ネットワークモデルを，本研究対象である情報空間に応用し，漸進的構築，履歴保持，多様性などの性質が得られることを検証するために，様々な予備実験を行っている [4]．ここでは，検索文書からのキーワード抽出に関して予備実験を行った結果について報告する．キーワード抽出は一般に，検索文書から予め抽出されたキーワードそれぞれについて，その重要度を評価する事により行うが，キーワードの重要度を評価する指標として従来主に用いられてきた TFIDF[1] では，キーワード間の出現文書パターンに関する相関関係を一切考慮していない．2節で述べたように，検索結果文書中からのキーワード抽出は，クエリーネットワーク構築手順の第一ステップとして非常に重要であり，特に話題分布を考慮してキーワードを検索結果文書集合からまんべんなく抽出するためには，キーワード間の相関関係に基づく多様性の考慮が必要である．

実験は，第 15 回ファジィシステムシンポジウム予稿集からアブストラクトを 11 編選びだし，以下の手順で行った．

1. 出現文書数 (DF) が 3 以上のキーワードを抽出し，ノードとする．ここで，出現文書パターンが一致するキーワードは一つにまとめる．
2. ノード間接続 (J^b_{ij}) を設定する．
   - 強接続 . . . ノード i, j の共起文書数が n 以上
   - 弱接続 . . . ノード i, j の共起文書数が 1 以上 3 未満
3. ノード・文書間接続 (J^g_{ij}) を設定する．
   - 強接続 . . . 文書 j 中におけるノード i の出現回数 (TF) が 3 以上
   - 弱接続 . . . 文書 j 中におけるノード i の出現回数が 1 以上 3 未満
4. ノード，文書の活性値 X_i, A_i を式 (1)〜(5) に従って更新する．ノード，文書の初期値はそれぞれ 10, 100, 000 とする．また，各式中のパラメータ設定は次の通りである：$s = 10$, $d = 0.4$, $r = 0.01$, $k = 0.0001$, $\theta_1 = 1,000$,$\theta_2 = 1000,000$, $p = 1.0$.

ステップ 1で生成されたノード数は 22 であり，ステップ 4 を 500 回行った後で各ノード (キーワード) の活性値を比較した．主なノードの活性状態，他のノード・文書との接続関係，および出現文書パターンを表 1に記す．図 1に示す通り，各クエリーの活性値は周期的に変動するものの，その状態は高，中，低の 3 種類に明確に分類可能であり，表 1にはこれについて記している．

11 文書与えた場合に高活性化するノードは「増加」,「獲得」,「分析」の 3 つであり，それぞれ異なる文書をカバーしている事がわかる．反対に「データ，クラスタリング」の活性値が低いのは，強接続数が多すぎる事，特に「分析」と強接続しているためである．そこで,「分析」,「データ，クラスタリング」の少なくとも一方を含む 2 文書 (1, 9) を除いて同様の実験を行ったところ，両ノードの関係が逆転し,「データ，クラスタリング」が高活性化し，反対に「分析」の活性度が低下することが確認できた．以上より，キーワード間の出現文書パターンの重複を考慮して，検索文書をバランス良くカバーするキーワード集合が活性化することがわかる．

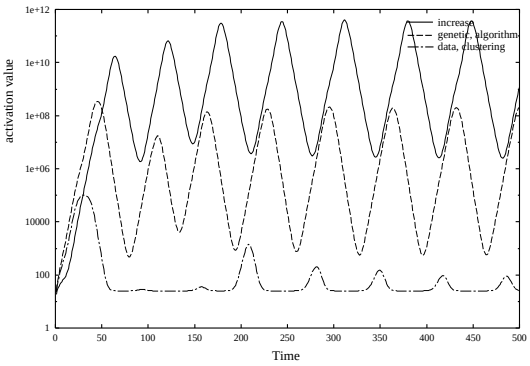

図 1: 活性値の時間変化

表 1: 活性値の比較（強接続は,「抗体との接続数」/「抗原との接続数」を表す.）

ノード	出現文書パターン	11 文書		9 文書	
		強接続	活性値	強接続	活性値
増加	3,5,6	2/0	高	2/0	中
獲得	0,2,4	2/3	高	2/3	高
分析	1,7,8,10	5/3	高	0/2	中
データ, クラスタリング	1,6,7,8,9	6/5	低	2/3	高

5 おわりに

情報検索作業を通じた話題分布構造可視化手法として，クエリーネットワークを提案し，その概要について述べた．また，免疫ネットワークモデルを用いてモデル化することも提案し，キーワード抽出に関する予備実験を通じて，免疫システムの持つ特性の一つである，多様性を考慮した組織化が行えることを示した．今後は，クエリーネットワーク全体の実装を進め，残る二つの特性 — 漸進的構築，履歴保持 — についても実現，検証を進めていく予定である．また，免疫システムの機能の解明は医学，数理生物学の分野においても進行中の課題であり，工学的応用も十分に進んでいない状況にある．話題分布構造可視化だけでなく，他の工学的用途へも広く適用可能な免疫モデルを提案する事も目的の一つである．

参考文献

[1] E. Bloedorn, I. Mani, and T. R. MacMillan, Machine Learning of User Profiles: Representational Issues, AAAI96, Vol. 1, pp. 433/438, 1996.

[2] N. K. Jerne, The Immune system, Sci. Am. Vol. 229, pp. 52–60, 1973.

[3] A. U. Neumann and G. Weisbuch, Dynamics and Topology of Idiotypic Networks, Bull. of Math. Biology, Vol. 54, No. 5, pp. 699/726, 1992.

[4] 高間, 廣田, 免疫ネットワークモデルを用いた話題分布構造可視化の提案, 第 14 回ファジィ・ワークショップ in 那須, pp. 20/23, 2000.

[5] O. Zamir and O. Etzioni, Grouper: A Dynamic Clustering Interface to Web Search Results, 8th World Wide Web Conference,
http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html, 1999.