「文書群を特徴づける単語群の抽出」を用いた
直観的インタラクションモデル

1. まえがき

双対型情報検索インタフェースは内容類似度に基づく高速連想検索と、検索結果からの特徴単語の自動抽出による鳥瞰図表示（特徴語グラフ）を用いて、対話性の高いテキストアクセス環境を提供する。
特徴語は検索結果における単語の頻度と共起解析により動的に構成され、検索結果の概観を与えると共に、絞り込みなどに利用することができる。
タイトルリストと特徴語グラフは並置されることにより、キーワード検索と連想検索を本質的に結びつける「橋」として機能する。

2. 検索結果と特徴語の並列表示を利用したインタフェース

インターネットに代表される肥大化する一方の情報空間に押しつぶされることなく、有用な情報を的確に取り出し、創造的な仕事に生かせる環境作りの基本として、双対型インターフェースという対話性に富んだ文書アクセス手段を開発した。
双対型インターフェースはふたつの「双対性」によって特徴づけられ、名前の由来ともなっている。一つは検索結果の表示に関する双対ビューであり、もう一つは、検索要求の表現に関する双対で、これは通常のキーワード検索と、文書間類似度に基づく連想検索の両方を組み合わせながら使えることを意味する。

双対ビューでは、検索結果を表示する際に、通常のタイトルのスコア順リストと並べて検索結果全体を鳥瞰できるような特徴語グラフというビューを提示する。特徴語グラフは、検索された文書群における単語の頻度とそれらの共起解析により動的に構成される、検索結果の概観を与えると共に、絞り込みなどに利用することができる。双対ビューにより、利用者はタイトルに現れた具体な情報と、全体の要約を同時に見ることができるため、検索作業の各段階でより的確に状況判断をしながら求める情報へ漸近的にアクセスすることができる。
一方、検索要求の双対については利用者の検索要求が具体的な場合にはキーワード検索、漠然とした例に基づく場合には連想検索というように使い分けられる利点がある。

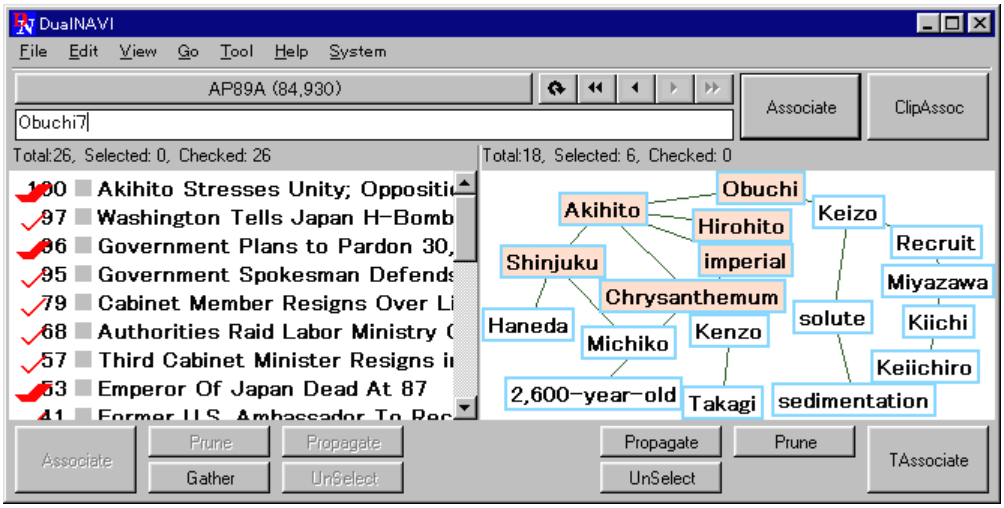

図 1. 双対型インターフェースの外観

3. 特徴語の表示を利用した適合性フィードバック

キーワード検索と連想検索という異なるタイプの検索は、それぞれに有効な手段であるが、双対ビューのもとではじめて有機的に結ばれ、相乗効果を発揮する。図 2 はそのしくみを示したものである。
スタートは状況に応じて任意タイプの検索が使える。その結果、タイトルリスト中に興味深い文書を発見した場合には、それらを選択して連想検索の方でフィードバックをかけることができる。

また、もし特徴語グラフの方に興味深い語を発見した場合には、それらを使って今度はキーワード検索でフィードバックをかけることができる。
また、双対ビューのそれ以外の有利な点として、クロスチェッキングのような小回りの効くインタラクション機能を実現しやすいということが挙げられる。例えば特徴語のいくつかを選択すると、それらを含む記事に強調マークがついて、上位に集めることができる。また逆に文書を選択すれば、それに含まれる特徴語群がマークされる。

4. 特徴語を視覚的に配置・表示する方法

ここでは特徴語グラフの生成方法についての概略を示す。詳細については（Niwa *et al.*, 1997）などを参考にして欲しい。3つの段階からなり、第1段階では特徴語の抽出を行なう。検索された文書群における各出現単語の頻度にもとづいて行われる。続いてグラフのリンク部分を作成するために、抽出された特徴語間の共起解析を行なう。その結果特に共起関係の強いペアにリンクを張る。最後は実際に表示をする際の各単語の座標を計算する。この場合、たて軸には検索結果文書群における文書頻度を取っている。上側ほどその文書頻度が高いもの、下に行くに従ってそれが低いものを配置することにより、全体として上位の話題のより詳細な話題の語が配置されるようになっている。横軸は重ならないように分散させているだけで特に意味はない。

特徴語の抽出には、出現各語(w)に以下のようなスコアを適用し、その上位のものを取っている。文書頻度とはその語を含む文書の数である。

(検索結果中の w の文書頻度) /

(データベース全体での w の文書頻度)

一般に(この尺度を含め)どんな尺度を用いても、高頻度の特徴語と低頻度の専門的な特徴語を安定的にバランス良く取ることは困難である。この困難を克服するため、頻度クラスを導入し、まず検索結果中の文書頻度に基づいて全体を分類する。その後、各頻度クラスから上記(あるいはその他)の尺度で上位の単語を取り出すことにより、バランスの取れた特徴語を抽出することができるようになっている。

リンク作成の段階では、前段階で抽出された各特徴語 X から リンクを張るべき特徴語 Y を X とY の共起頻度 *df(X&Y)*を用いて *df(X&Y) / df(Y)* という基準で選択し、*df(Y) > df(X)* を満たす Y の中からこの共起強度が最大となるものをリンク先とする。

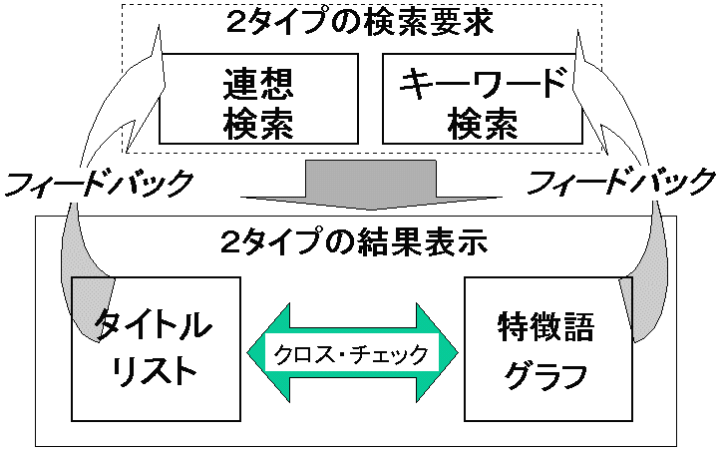

図2. 特徴語の表示を利用した適合性フィードバック

参考文献

Nishioka, S., Niwa, Y., Iwayama, M., and Takano, A. (1997). *DualNAVI*: An information retrieval interface. In *Proceedings of WISS'97* (pp. 43—48).

Niwa, Y., Nishioka, S., Iwayama, M., Takano, A., and Nitta, Y. (1997). Topic graph generation for query navigation: Use of frequency classes for topic extraction. In *Proceedings of NLPRS'97* (pp. 95—100).