

情報獲得支援システム *DualNAVI*の実装

1 J-2

西岡 真吾 高野 明彦
日立製作所基礎研究所

澤田 瑞穂 山崎 幹夫
日立東北ソフトウェア

藤井 泰文
日立デジタル平凡社

1 はじめに

近年の電子出版やインターネットなどのめざましい発達により、膨大な量の情報が電子的な形態で提供される様になり、また個人が容易にそういった情報にアクセスすることができる様になった。ところで、ディレクトリサービスやハイパーリンク等に見られる様に、互いに関連する情報を共に提供することにより情報の価値を高めることが可能である。情報提供者は、自分が出す情報と他の情報提供者による情報との関係により価値を高めつつ、自分が出す情報そのものは自分の財産として大事に運用することが望まれる。この様な問題への一つのアプローチとして我々は *DualNAVI* 検索モデルを提唱している。本稿では *DualNAVI* 検索モデルの概要とサーバの実装について述べる。

2 *DualNAVI* 検索モデルの特徴

本節では *DualNAVI* 検索モデルの特徴について簡単に述べる。*DualNAVI* 検索モデルはクライアントとサーバから構成されており、以下の様な特徴を持っている。

DualNAVI クライアントは、文章、文書の一部、キーワード、検索された文書などから、それに類似した文書を検索する（以下、連想検索）ことが可能である。この様に *DualNAVI* 検索モデルでは多様な検索キーの与えかたが可能であり、検索意図を柔軟に表現することができるためユーザは状況に応じて適切な検索方法を使うことができる。さらに、*DualNAVI* は分散管理された異なるデータベースを統一的に探索するインタフェースを提供するため、*DualNAVI* は検索結果のタイトル一覧とともに、それら検索結果を特徴付ける単語（以後、特徴語）を関連

度でランク付けし、グラフ表示することにより検索結果の概要を表示する。タイトル一覧と特徴語グラフは並置され、それによって効果的な検索ガイドと関連度フィードバックを支援する。さらに *DualNAVI* はタイトル一覧と特徴語グラフのクロスチェック機能を提供する。これにより、検索結果の詳細な分析が可能となる。

DualNAVI サーバは通常の連想検索に加えて、分散管理された異なるデータベース間での連想検索を行うことができる。この機能により、異種データベース間の相互運用性を保証している。さらに、異種データベース間での連想検索の際には必要最小限の情報のみ（オリジナルの文書は不要）が転送されるため、*DualNAVI* サーバが管理しているコンテンツの知的所有権を保護することが可能である。また、検索サーバと本文サーバは分離しており、検索サーバは連想検索を行うための必要最小限の情報しか必要としないため、コンテンツの保全と柔軟なアクセス管理を行うことができる。図1に *DualNAVI* 検索モデルの概要を、図2に *DualNAVI* クライアントのインタフェース画面を示す。

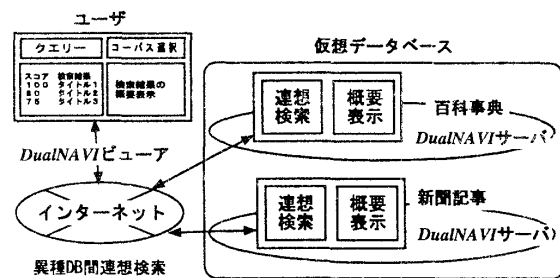


図1: *DualNAVI* 検索モデルの概要

The Design and the Implementation of the *DualNAVI* Information Retrieval Interface, by Shingo Nishioka, Takano Akihiko (ARL, Hitachi, Ltd.), Mizuho Sawada, Mikio Yamazaki (Hitachi Tohoku Software Co.), and Yasufumi Fujii (Hitachi Digital Heibonsha Co.).

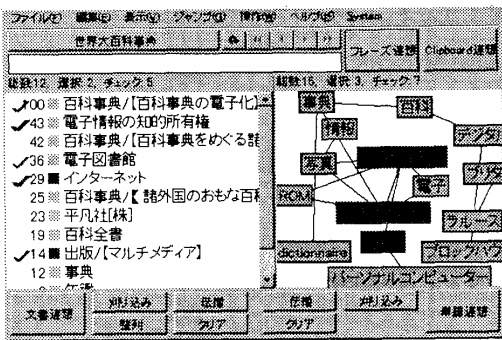


図 2: DualNAVIのインタフェース画面

3 DualNAVIサーバの実装

本節では DualNAVIサーバの実装について簡単に述べる。DualNAVIサーバの主要な機能として、1. フレーズ検索 (与えられた文章からの連想検索)、2. 検索結果の概要の作成 (特徴語の抽出とグラフ表示)、3. 連想検索 (与えられた文書からの連想検索)、がある。これらはそれぞれ 1. 単語 → 文書、2. 文書 → 単語、3. 文書 → 文書という連想検索を行うことに相当しており、DualNAVIサーバはこれらの検索を高速かつ効率的に行うことが求められる。そのため、DualNAVIサーバは管理している文書については予め形態素解析を行い、タームインデックスを作成している。多くの場合、このインデックスはスパースであるため、各文書に出現する単語とその頻度 (tf) を引くためのインデックス-A と、その逆に、各単語についてその語が出現する文書とその頻度を引くためのインデックス-W の2つのインデックスを準備している。さらに、DualNAVIサーバは内部的には文書、単語とも整数を割りあてて使っているため、インデックス-A とインデックス-W は同一のコードで操作することができる。これにより、1. および 2. のタイプの連想検索を高速に行うことが可能となる。

3. の連想検索の実現方法であるが、DualNAVIではこの検索を、連想元の要約の作成と要約からのフレーズ連想という2つのステップにより行っている。このように、一旦要約を作成することで、与えられた文書中の検索にとって重要な語が抽出される。その一方で、偶然含まれる検索意図にそぐわない語は複数の文書に現れる可能性が低く、その結果要約には現れにくくなる。このことにより、高い精度の連想検索が可能となっている。DualNAVIでは、連想検索のために 2. と同じアルゴリズムを用い 200 単語からなる要約を作成することで、取りこぼしが少なくかつノイズにも強い連想検索を実現している。

また、要約としての単語リストを作ってから検索を行

うため、あるコーパスで与えられた文書の要約を作り、それを用いて他のコーパスの文書を検索することにより、異なるコーパスにまたがる連想検索が可能となる。なお、現在の DualNAVIサーバの実装はコーパスごとに単語の分布が同じであると仮定している。さらに、要約を他のサーバに転送することにより、異なるサーバが管理しているコーパスを検索することも可能となる。要約は十分小さいためネットワークを通して高速に転送することが容易である。またオリジナルの文書を送ることなく連想検索が可能であるため、ある程度の知的所有権の保護も同時に行うことができる。(コーパスに特化した形態素解析や語の分布に関する知識などが含まれるため、完全に保護できる訳ではない。)

4 おわりに

DualNAVIクライアントは常に検索結果の要約を表示する。要約と検索結果の関係は対話的な分析手段により詳細に調べることができるため、ユーザが検索結果を理解し次の検索に役立てることができる。さらに複数の検索手段を目的に応じて使い分けることで、より効率の良い検索を行うことができる。

また、DualNAVIサーバは分散管理された異なるデータベース間での連想検索を行うことが可能である。そのため、情報の保全と柔軟なアクセス管理を行いながら、他データベースとの有機的結合を可能とし、情報の価値を高めることができる一方で、ユーザはどのデータベースにアクセスすべきかを意識せずに効率的な検索を行うことが可能となる。

謝辞

サーバの設置、運用等で協力していただいた (株) 日立国際ビジネスの三村、神庭両氏に感謝します。

参考文献

- [1] 西岡, 他. 文献検索支援インタフェース DualNAVI, インタラクティブシステムとソフトウェア V, pp. 43-48, 1997.
- [2] Iwayama, M. Cluster-Based Text Categorization: A Comparison of Category Search Strategies, SIGIR'95, pp. 273-280, 1997.
- [3] 丹羽. 動的な共起解析を用いた対話的文書検索支援, 情報処理学会研究報告, 96-NL-115, Vol. 96, No. 87, pp. 99-106, 1996.