電子図書館のための効率的な文書検索
— 検索／提示のための文書構造化と抄録生成 —

住田 一男, 酒井 哲也, 小野 顕司, 三池 誠司
(株) 東芝 研究開発センター
〒210 川崎市幸区小向東芝町 1
Tel: 044-549-2240, Fax: 044-520-1308
E-Mail: {sumita,tets1,ono,miike}@eel.rdc.toshiba.co.jp

概要

電子図書館での情報検索の観点から、内容に基づく検索と対話的な文書提示を目的とした文書の構造化と抄録生成について提案する。電子図書館では、内容に基づいた検索を簡単かつ効率的に行えることが求められる。文書は、表層的 (タイトル, 章節, 段落など) あるいは意味的 (背景, 目的, 結論など) レベルで様々な情報を含んでいる。全文文書を対象に、これら情報を自動的に構造化し、内容に基づく検索と対話的な抄録生成を行い、効率的な検索を具体化する。

キーワード

電子図書館, 文書検索, 文書提示, 文書構造, 自動抄録, 抄録生成

Effective Document Retrieval for Digital Library — Document Structure Analysis and Automatic Abstract Generation —

Kazuo Sumita, Tetsuya Sakai, Kenji Ono and Seiji Miike
Research and Development Center, Toshiba corp.
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki-shi, 210, JAPAN
Phone: +81-44-549-2240, Fax: +81-44-520-1308
E-Mail: {sumita,tets1,ono,miike}@eel.rdc.toshiba.co.jp

Abstract

This paper suggests that document structure extraction and abstract generation are fundamental functions of a digital library. A digital library should have an effective document retrieval facility. In a document, several information axises can be defined: a title, section, paragraph, background, purpose, conclusion, etc. A document structure represents such information axes. The document structure is automatically extracted from a full-text, and is used for generating abstracts to realize effective document retrieval.

Keywords

Digital library, Document retrieval, Document structure, Document presentation, Automatic abstract generation

1. はじめに

インターネットの普及にともない、世界中のどこからでも様々な情報にアクセスできるようになってきた。米国ゴア副大統領の情報スーパハイウェイ構想に歩調を合わせて、電子図書館に関するプロジェクトが数多く始動しており、電子図書館の実現に対する期待が高まりつつある。

『電子図書館は、従来の図書館が提供してきた収集、目録作成、検索、SDI などのサービスを再現/模擬/拡張するために必要な内容物およびソフトウェアとともに、計算/データ記憶/通信の機器を適切に組み合わせたシステムである。従来の電子図書館の基本的なサービスを提供するとともに、良く知られたディジタル的な情報蓄積、検索、並びに通信の長所を生かさなければならない [1]。』また、その利用者は、図書館員から小学生にまでわたるため [2]、専門家以外の人の利用を前提としたシステムでなければならない。

膨大な情報から所望の情報を的確に探し出すため、特に情報検索の高度化への期待が大きい。本稿では、情報検索に的を絞り、効率的な情報検索のための枠組について検討する。

2. 電子図書館における効率的な検索

検索の過程を大きく分類すると、利用者が検索条件を入力し検索を行う検索フェーズ、検索した結果を利用者に提示する情報提示フェーズの 2 つのフェーズに分割できる。これらの 2 つの面から効率化を考える必要がある。

検索フェーズの効率化のためには、利用者の意図に沿った検索が精度よく行われることが最も重要である。単純な boolean 検索では、十分な検索精度が得られず、検索ノイズが多くなることが指摘されている [3]。このため、従来より、単語の統計情報を用いるなど、検索精度を向上させる研究が多数行われてきた。しかし、検索精度を劇的に向上させる手法はいまだに実現していない。

このような現実の下では、利用者は、検索と情報提示を繰り返しながら、検索条件を段階的に修正し、所望の情報に絞り込んでいかざるを得ない。そこで、利用者とシステムの対話過程を、円滑に進められることが重要なポイントとなる。特に、電子図書館においては、専門家 (図書館員やサーチャーなど) 以外の様々な人々が利用者となる。効率的な検索を実現するためには、効果的な情報提示が重要な研究課題であるといえる。

従来の文献検索システム (例えば、JOIS や PATOLIS など) では、あらかじめキーワードづけされた文書に対して、単語やその論理式を入力して検索する。検索結果の表示は、検索された文書集合の文書数だけである。タイトルを表示するにも、さらに何らかのコマンド入力を行わなけらばならない。このような検索インタフェースでは、自分の入力した検索条件が正しいか否かを直観することが難しく、まったく意に反した検索をしてしまう可能性も大きい。また、入力した単語に対応するキーワードがそもそも設定されていない場合もあり、キーワードとして存在しないのか、検索されるべき文書が存在しないのかが利用者に明示されない。

一方、近年商用化が進みつつある全文検索システムでは、あらかじめキーワードづけされていない文書を対象に検索を行う。単語やその論理式で検索条件を入力し、また、各文書のファイル名やタイトルを検索結果として表示するシステムが多い。検索結果の件数だけを表示するのに比べ、どのような検索がなされたかを利用者が見ることができ、検索条件の修正に役立つ。対話的な検索を指向しているといえる。しかし、このような直接的な表示では、大量の文書が検索された場合、すべてのタイトルを見渡すことができない。一覧性に問題がある。

情報検索を効率的に行うためには、得られた検索結果を利用者がよく把握できるインタフェースが重要である。検索結果の情報量が膨大になる可能性がある一方で、人間が一度に見られる情報量には限りがある。

そこで、情報検索の結果得られた情報をすべて提示するのではなく、その全体量に応じて情報を圧縮し、等身大にして提示することが望ましい。

また、検索の範囲が絞り込まれていくにつれ、大まかな情報の要求から詳細な情報の要求へと段階が進んで行く。このため、情報提示においても、詳しさを段階的に変更できる仕組みが必要となる。

以上まとめると、効率的な検索結果の提示のためには、以下の2点が必要となる。

- 様々な階層での一覧性の確保
- より抽象度の高い情報から詳細な情報までの段階的提示

これらの情報提示は、表層的な処理だけでは十分な効果が得られないと考えられる。そこで、文書の構造化によりこの実現を図る。

3. 文書集合と文書の構造化

検索結果は、検索された文書集合と、その文書集合を構成する各文書の2つの階層からなる。これらのそれぞれの階層について、一覧性のある情報提示を可能にするには、それぞれの階層における情報の構造化が必要である。

3.1 文書集合の構造化

検索結果の文書集合の構成は、検索条件によって動的に変わる。検索条件との類似度、他の文書との関係、などが情報提示の軸となると考えられる。

一般に、各文書と検索条件との間に何らかの類似度を定義することにより、その類似度にしたがって文書を順序づけられる。一定の類似度内に納まる文書を1つのクラスタとしてまとめられるので、この情報が文書集合の提示に利用可能である。

類似度以外の情報提示軸として、いくつかの軸が考えられる。例えば、文書を作成した時間やその文書のカテゴリなどを軸として、三次元的に文書提示を行う提示インタフェースが提案されている [4]。また、各文書に情報提示軸に相当するスロットを設け、そこにあらかじめ入力された情報にしたがって、順序づけて表示するブラウジングインタフェースも考えられている [5]。これらの提示インタフェースでは、あらかじめ設定されている情報軸でしか表示できない。このため、図書館の書架を眺めながら書籍を探すような検索には適しているが、利用者が検索条件を設定して文献を探すような能動的な検索には向かない。

出来事を記述することが中心であるニュース記事では、出来事の5W1Hが情報提示の軸となりうる。このような観点から、情報検索と提示の試みがなされている [6]。

技術論文などの文献検索を想定すると、記述されている内容は出来事ではなく、事実とその事実から導き出される論理的な記述が中心となる。背景や目的、特徴、結論などがその典型的な記述項目であるといえる。そこで、どの記述項目の観点で類似しているかを、構造化して提示することにより、直観的な文書集合の提示が可能である。上記の意味的な軸 (意味役割と呼ぶ) 以外にも、タイトルや著者、参考文献といった形式的な情報も利用できることはいうまでもない。

このように文書間の関係を構造化するためには、文書の意味役割 (目的，話題，特徴，結論) や、形式的な情報 (タイトル，著者，参考文献) などを抽出する必要がある。

3.2 文書の構造化

文書中は、技術論文の場合、タイトル、著者名、概要、参考文献などの項目とともに、章や節などの本文から構成される。さらに、各章や節の本文は、段落や文などがその構成要素となっている。各文書を提示する場合には、これらが軸となる。

タイトルや章や節の構造は、行頭に位置する章節番号や、空白などのインデント情報を手掛かりにして解析する。文書構造を解析することにより、タイトルだけを表示したりというような様々な詳細レベルでの表示が可能になる。

一方、各章節の文章自体について、詳細度を変えた提示を可能にするためには、文章の抄録生成が求められる。日本語の場合、文と文の間の関係を接続詞や文末の表現などで明示する場合が多く、これらの表層的な手掛かりを用いて重要文を選択することができる。接続詞や文末の表現などを手掛かりに、文章を構造化し [7]、抄録の自動生成を行う [8]。

接続詞や文末表現など文間の接続的な関係を明示する表現を、34 種類の関係 (修辞関係と呼ぶ) に抽象化/分類した。文書構造の解析では、この修辞関係を各文ごとに抽出する。そして、抽出した修辞関係から、論旨のまとまり具合を解析し構造化を行う (詳細は文献 [7])。

一方、意味役割の抽出は、「目的」や「結論」などを表現する言語表現を手掛かりにして、各文がどの意味役割に相当するかを抽出する。また、意味役割が明示されていない文についても、文書構造を参照し前後の文から情報を補う処理も行っている (詳細は文献 [9])。

4. 対話型検索システム BREVIDOC

対話型検索システム BREVIDOC(a broadcatching system with an essence viewer for retrieved documents) において文書集合と文書の 2 階層での抄録提示を実現した [10]。それぞれの抄録から、より詳しい情報への対話的なアクセスが可能である。

図 1 に文書集合の表示例を示す。検索条件と文書との類似度を求め、その類似度ごとに文書集合として表示している。各文書集合には、代表となる文書のタイトルと、また、その文書集合に属する文書の件数を表示している。

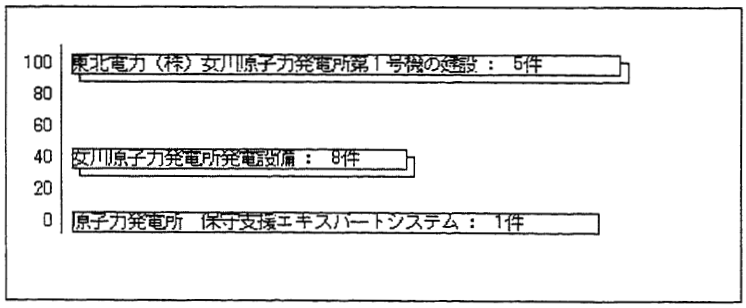

図 1 文書集合の表示 1

文書集合の表示は、単に検索結果を文書集合として提示するだけでなく、より詳しい情報を提示するための指示メニューとして機能する。図 1 では 3 つの文書集合が表示されているが、その 1 つをマウスでクリックすることにより、その文書集合に属する文書のタイトルのリストを表示できる。

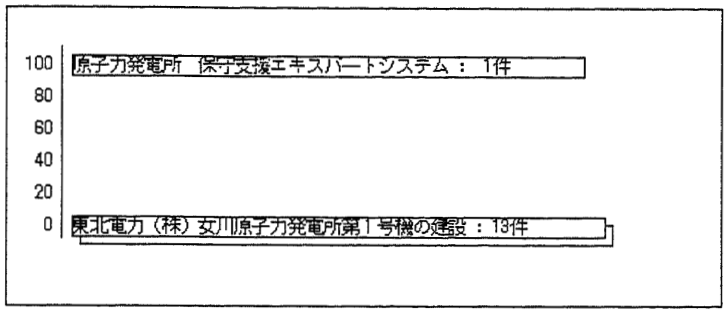

図2 文書集合の表示2 (表示の軸を目的から結論に変更)

図2は、情報の軸を変更した場合の文書集合の表示例である。図1では、目的の軸を表示の軸としているが、この図では、結論の軸を表示の軸としている。

文書の提示例を図3に示す。右側には原文を、左側にはその原文から自動生成した抄録を表示している。抄録では、タイトル、著者名、原文中の要約の抄録、章節見出しを表示している。例で示した技術論文の場合、原文を1つのウィンドウに一度に表示することはできない。しかし、抄録表示では、文書の内容を一覧できるようになる。

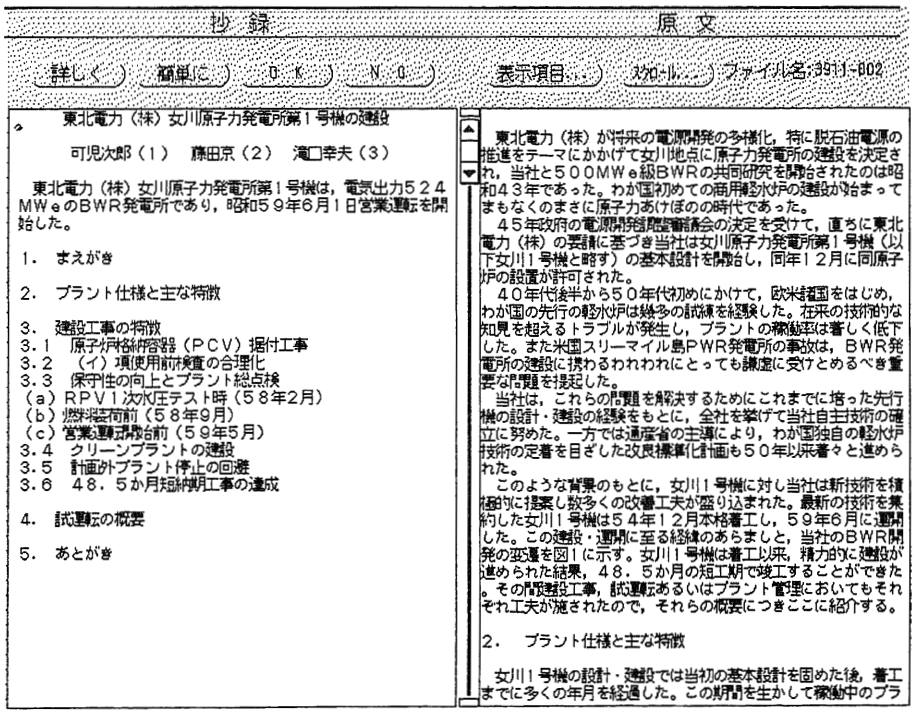

図3 文書の表示 (左:抄録, 右:原文)

システムは、章節見出しの表示を省略することも可能である。図4にその様子を示す。この例では、タイトル、著者名、抄録のみを表示している。図3で示した抄録表示に比べ、より概略的な抄録となっている。このレベルの抄録であれば、複数文書の抄録を同時に提示することも容易である。

文書集合の表示と同様に、抄録の表示ウィンドウは、より詳しい情報を提示するための指示メニューとし

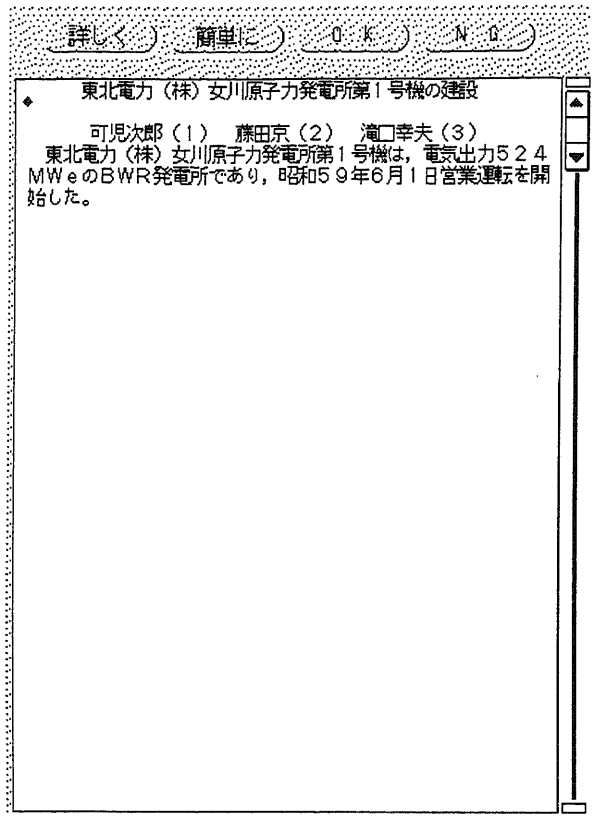

図 4 抄録の表示 1 (章節見出しの表示を省略)

て機能する。例えば、ある節の見出しを指定することにより、その節の内容を表示することができる。この様子を図 5 に示す。図では、「まえがき」の部分を指定して、その抄録を表示した場合である。

5. おわりに

電子図書館を対象として、効率的な文書検索のための文書の構造化と抄録提示について検討した。BREVIDOC では、タイトルや章節構造など形式的な構造とともに、文と文の間の意味関係も抽出している。この構造に基づいて、文書集合表示、文書表示のそれぞれについて、一覧性と対話性を持たせた提示インタフェースを実現した。

システムを用いた評価実験により、検索の作業効率の向上を確認している [11]。しかし、文書集合の提示では、ある情報の軸に対する 1 次元表示を行っているにすぎない。2 次元的な表示を行うことにより、文書集合の全体像をより直観できる提示ができると考えられる。

参考文献

[1] Gladney, H.M. et al. : Digital Library: Gross Structure and Requirments: Report from a March. 1994 Workshop, Proc. Digital Libraries '94, pp.101-107.

[2] McKeown, K. et al. : The JANUS Digital Library. 1994 Workshop, Proc. Digital Libraries '94, pp.67-73.

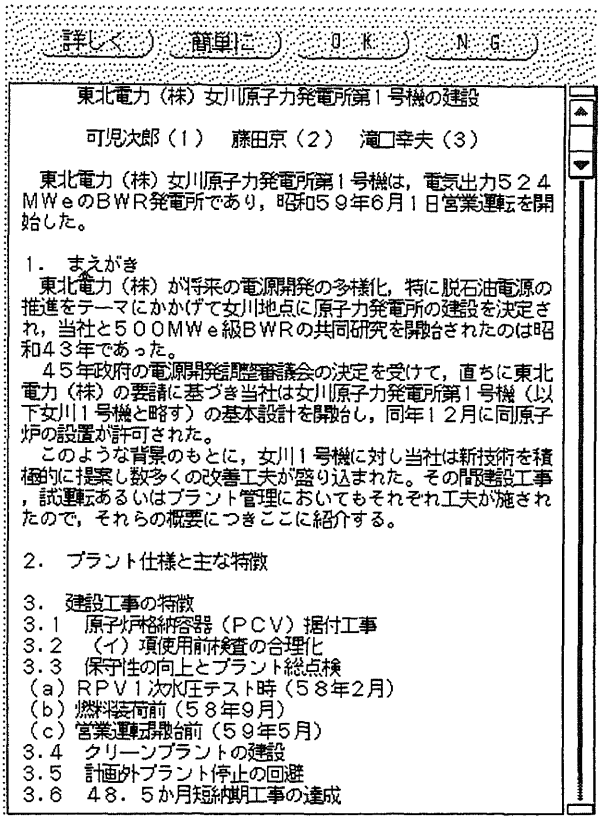

図 5 抄録の表示 2 (「まえがき」を指定して抄録を表示)

[3] Turtle, H. et al. : Evaluation of an Inference Network-Based Retrieval Model. ACM Trans. Information Systems, Vo.9, No.3, pp.187–222, 1991.

[4] Mackinglay, J.D. et al. : The Perspective Wall: Detail and Context Smoothly Integrated. Proc. SIGCHI'91, pp.173-179.

[5] 増田桂弘, 植田学, 石飛康宏 : フレーム関係軸モデルに基づく情報群の自動組織化と視覚的ブラウジング. 1994 年情報学シンポジウム, pp.219-228.

[6] 小澤英昭, 中川透 : 情報空間モデルに基づく情報検索システム. 情処研資 IM-92-53, pp.49-56, 1992.

[7] Sumita, K. et al. : Document Structure Extraction for Interactive Document Retrieval Systems. Proc. SIGDOC'93, pp.301-310.

[8] Ono, K. et al. : Abstract Generation Based on Rhetorical Structure Extraction. Proc. COLING '94, pp343–348.

[9] 三池誠司, 住田一男 : 文の意味役割解析に基づく全文検索. 情報研資 FI-34-3, pp.17-24, 1994.

[10] Miike, S. et al. : A Full-Text Retrieval System with a Dynamic Abstract Generation Function. Proc. SIGIR '94, pp.152-161.

[11] 酒井哲也, 三池誠司, 住田一男 : 文書検索システムの動的抄録提示インタフェースの評価. 情処研資 HI-57-7, pp49-54, 1994.