キーワード検索における検索ナビゲータ

4S−2

富士 秀
(株)富士通研究所
fuji@flab.fujitsu.co.jp

1　はじめに

キーワード検索を支援するための検索ナビゲータを試作したのでこれについて述べる。

2　背景

近年になって多種多様な電子化文書が世の中に出回るようになってきた。その蓄積量は増加する一方だが、こうなると今度は文書の中から自分の必要とする情報を迅速かつ手軽に見つけ出す検索システムが必要となってくる。

現在までにいろいろな検索システムが考案されてきているが、現時点ではキーワードによる検索がもっとも実用的なものとなっている。しかしこのキーワード検索も一般検索者には使いにくく、実際には検索システムやデータベースに関する知識を持ったデータベースサーチャのような人でなければ効率的な検索ができない。

3　一般検索者の抱える問題点

一般検索者とデータベースサーチャの検索履歴を分析・比較した。検索履歴としては、社内で実際に運用されているデータベースの検索ログデータ約1ヵ月分を用いた（データの内容は社内SEの事例集）。この分析の結果、一般検索者は検索の際に次に挙げるような問題を抱えていることがわかった。

キーワード　一般検索者は、キーワードとしてどのようなものを入力すべきかということがわからない。また、入力したキーワードによってどのくらい効果的な検索ができたかわからない。

検索のシーケンス　一般検索者は、どのような演算子をどのようなタイミングで使って検索式を作るべきかわからない。

検索件数　一般検索者は、入力の結果として何件くらいの文書が検索されるかの予想がつかない。

検索項目　一般検索者は、どのような検索項目があり、どの検索項目をどのようなタイミングで使用すべきかわからない。（検索項目とは、例えば、文書の「日付」、「著者」、「公開番号」など）

4　検索ナビゲータ

以上述べたような問題を軽減することを目指して検索ナビゲータを試作した。一般検索者もサーチャに近い検索ができるよう設計したつもりである。

「キーワード」の問題は、絞り込み候補や拡張候補を提示することによって対処した。また、このような絞り込みや拡張のタイミングを提示することによって「シーケンス」の問題に対処した。キーワードの候補を提示する際に件数を示し、「件数」の問題に対処した。「検索項目」についてはこれからの課題となっている。

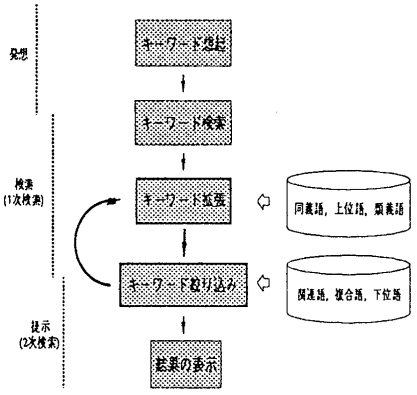

図1. 検索ナビゲータ

4.1　絞り込み候補の提示

得られた検索数を絞るためのキーワード候補を検索者に提示する。また、これらのキーワー

'Keyword Navigation for Keyword-Based Retrieval Systems
FUJI, Masaru
FUJITSU LABORATORIES LTD.

ド候補と一緒にそのキーワードを選んだ時に何件に絞られるかの件数の表示も行なう。

件数	キーワード
62	アドレスバス
34	データバス
5	コントロールバス
2	ビットハンドリング
2	長さカウンタ
2	書き込みタイミング

図2.「アドレスバス」に対する関連語

ナビゲータでは表示された関連語を直接次の入力として使えるので、検索者が求めるキーワードが表示された中にあれば、少ない操作で検索が進められる。なお、提示された候補の中に使いたいキーワードがない場合は、入力部から任意のキーワードを入力することができる。

4.2 拡張候補の提示

検索ナビゲータは、検索対象を拡張するためのシソーラスを持っている。この拡張候補は、ユーザの入力キーワードに対応して出力される。また、ユーザが関連語を選択した場合にも、関連語に拡張候補があればその拡張候補も表示するようにした。

4.3 検索シーケンスの提示

検索の再現率を保ちながら適合率を上げるには、論理和で検索数を十分に拡張しておいて、これに対して論理積をかけて絞り込んでゆくのが好ましい。また、この絞り込みのためのキーワード自身も絞り込みに先だって拡張できるだけ拡張しておかなければならない。つまり、拡張と絞り込みを繰り返しながら検索を進めることが望まれる。

ナビゲータでは、入力があるたびに絞り込みの画面、拡張の画面、検索式の画面などが適宜表示され、入力のためのカーソルがこれらの画面間を移動する。これにより検索者は検索式をあまり意識することなく、好ましい検索シーケンスをたどることができる。

絞り込みの画面や拡張の画面は検索式と連動しており、例えば拡張画面での入力は「論理和」として検索式に追加される。

5 提示データの作成

以上説明したように、ナビゲータでは各種の提示データを使用するが、これらの作成について以下に述べる。

5.1 関連語

関連語は、検索対象のデータベースの文書から自動的に抽出した。ある入力キーワードに対する関連語は、その入力キーワードと文書内で共起する単語群の中から自動的に選択を行なった。選択には、品詞情報、共起単語の頻度の統計情報などを使った。頻度情報は適当な正規化処理を施すことにより、意味的に中立なキーワードを排除した。

作成時間の関係上、関連語は検索に先だってバッチ的に作成することにした。試作版では、約8万件の特許抄録より約5万の関連語を抽出した。このとき、インバーテッドファイルに入っていても、全体頻度の低いキーワードに対しては関連語は作成しなかった。これは、低頻度のキーワードに対しては全ての一次情報を見ることができるので、提示の意味があまりないからである。

5.2 同義語（複合語・下位語）

同義語は、機械翻訳用のものを転用した。翻訳用辞書の概念記号をたどることによって、同義・類義語を得た。

6 まとめと課題

各種の提示機能によって、検索者はデータベースや検索状況に関するより多くの情報を得ることができるようになった。また、この情報を検索に使っていくことができるようになった。

ただし、提示されたデータが実際の検索にどのくらい有効かを判断するのは難しい。現状では、目視によって有効そうなキーワードが出るかどうかを判断しながら、提示データの抽出アルゴリズムを決めているが、もっと客観的かつ定量的な評価方法を考えてゆきたい。

7 文献

1. 富士 秀：「自然言語文書からの特徴キーワード抽出」、情処第43回全国大会 1H-8 (1991.10)
2. 下山 栄子、富士 秀、松井くにお：「サーチャのノウハウに見る検索インタフェース」、情処 III 研究会 41-2 (1992.03.02)
3. 松井 くにお、吉岡 誠：「情報通有におけるコンテンツと検索」、情処 情報メディア研究会 6-2 (1992.05.12)
4. 長尾 真、水谷 幹男、池田 浩之：「日本語文献における重要語の自動抽出」、情報処理 Vol.17 No.2 (1976.02)