追加検索語候補提示に関する一検討

1N−6

井上 孝史　杉崎 正之　早川 和広　田中 一男

NTTヒューマンインタフェース研究所

1 はじめに

テキスト検索システムでは、最初に指定した検索条件による一回の検索でユーザの要求が満たされないことが多々ある。その場合検索語を追加したり変更したりすることによって検索条件を変えて再検索することが良く行なわれるが、何を検索語として追加すればよいかを知る手がかりがなく、試行錯誤で繰り返し行なわれるため、効率がよくない。

本稿では、検索対象のテキストデータベースの中の語の共起関係を調べ、元の検索語と連想関係にある語を検索語の候補として提示する方法について報告する。

2 語の共起関係の強さに基づく候補の提示

語の共起関係の抽出

検索対象のテキストデータベース全体から、共起関係の強い語のペアをあらかじめ抽出しておく。二つの語 x,y の共起の強さ (以下共起度と呼ぶ) としては相互情報量を用いており、次の式で計算する。

$$\text{共起度} = I(x,y) = log_2 \frac{P_d(x,y)}{P(x)P(y)}$$

ここで、$P(x), P(y)$ はそれぞれ x, y の出現確率であり、$P_d(x,y)$ は x, y が d 語の幅の窓の中 (付属語は除く) に同時に出現する確率である。

候補の提示

検索時には、検索結果とともに、元の検索語と強い共起関係にあるものを共起度の強いものから順に、追加検索語の候補として提示する。ユーザはその中から適当と思われるものを選択し、再検索する。

3 実験

前節で述べた方法に基づき、我々が開発している全文検索システムに、検索語候補を提示する機構を組み込んで実験を行なっている。実験では、テキストデータベースとして、ホームページ検索サービス NTT DIRECTORY のデータ (Web ページ紹介文、約 50MB) を用いた。テキストを形態素解析し、すべての自立語を抽出したあと、頻度の低いもの (30 回以下) を除いた。その結果残った語は約 1 万語で、3 語以内に同時に出現するものを共起関係抽出の対象とした。追加検索語の候補としては、元の検索語との共起度が 3 以上のもののうち、上位 20 個程度を提示している。図 1 に抽出した共起関係に基づく検索語候補提示の例を示す。この図は「ロック」という語で検索したあと、再検索するための追加検索語の候補を提示しているところである。候補として、「ヘビー」、「パンク」、「洋楽」、「北欧」などが提示されており、いずれかの語を選ぶことで、より細かいジャンルに関係するテキストが検索できると予想される。また、検索語の候補は現在のところ図のようにリストとして表示しているだけだが、それぞれの語を指定するとどの程度結果が絞り込めるか等の情報を、より視覚的にユーザに提示する方法についても研究を行なっている。

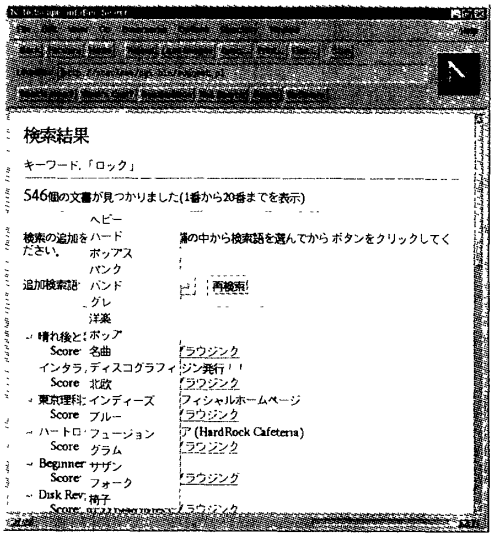

図 1: 追加検索語候補提示の例

4 まとめ

語の共起関係に基づき検索語の候補を提示する方法について報告した。今後は、(1) 実システムでの有効性の検証、(2) より効果的な検索語候補の提示方法、について研究を進める予定である。

A Study on Term Suggestion for Text Retrieval
Takafumi INOUE, Masayuki SUGIZAKI,
Kazuhiro HAYAKAWA, and Kazuo TANAKA
NTT Human Interface Laboratories