

森 大二郎 杉崎 正之 田中 一男

NTT サイバーソリューション研究所

## 1 はじめに

フルテキスト検索システムでは、検索条件として文字列を入力し、この文字列を含んでいる文書を検索結果として出力するのが一般的である。しかし近年、フルテキスト検索システムのユーザ層が急激に拡大しており、また検索システムを使用する端末の形態が多様化しているため、文字列の入力作業そのものが障壁となる場合が少なくない。

本稿では、検索入力履歴内での単語の生起確率によって入力補完候補を決定し、ユーザの検索文字列入力を支援する方法について報告する。

## 2 フルテキスト検索における検索入力支援

Emacs 等のエディタや、いくつかの shell 環境では、引数を確定するのに必要なだけ文字列を入力すれば、確定した範囲で残りの文字列を決定する入力補完機能を持っている。しかし、フルテキスト検索システムにおいては、検索対象である文書に含まれる全単語が入力補完の対象となり、残りの文字列を一意に特定できる場合はごく少くなるため、同様の手法をそのまま適用することはできない。

入力された文字列から残りの文字列を少数に限定できない場合に、所定の条件によって絞り込んだ少数の候補をユーザに提示して選択させるという方法がある。絞り込む方法については、例えば、対象文書の中での出現頻度が高い単語を優先する方法が考えられるが、テキスト検索においては、対象文書の中で出現頻度の高い単語は、相対的に重要度が低い単語である場合が多いため[1]、検索条件としての単語を特定する手段としては好適でない。

そこで我々は、検索システムに与えられる入力文字列の履歴における単語の生起確率に着目して提示する単語の候補を求める方法を検討することとした。

## 3 検索入力履歴に基づく入力補完候補の提示

不特定多数のユーザから大量の検索要求を受けるテキスト検索システムにおいては、特定の単語が集中的に入力され、またそのような単語は時間と共に徐々に推移する傾向があることが分かっている[2]。

Word Completion in Full-Text Retrieval Using Query-Log  
Daijiro MORI, Masayuki SUGIZAKI,  
and Kazuo TANAKA  
{mori,sugizaki,tanaka}@aether.hil.ntt.co.jp  
NTT Cyber Solutions labs.

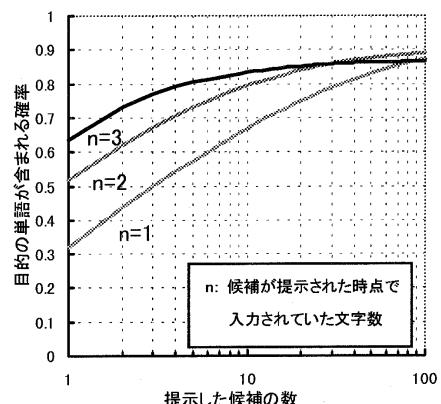


図1: 候補中に目的の単語が含まれる確率

この性質を利用して、一定の時間範囲内での検索入力履歴における単語の生起確率を求めておいて、ユーザの入力した文字列から限定される単語の中で、生起確率の大きい単語を候補として提示するシステムを試作した。

WWW の検索サイトにおける 24 時間分の検索履歴から単語の生起確率を算出し、これに基づいて、直後の 3 時間の検索履歴を対象として入力補完候補を提示した時のヒット率(候補中に目的の単語が含まれる確率)を図 1 に示す。2 文字入力した時点で 10 単語の候補を提示すれば、80% の確率で目的の単語が含まれていることになり、入力補完手段として実用的な精度を実現していると考えられる。

## 4 まとめ

テキスト検索システムにおいて、検索入力履歴内での単語の生起確率によって入力補完候補を提示する方法について報告し、WWW の検索サイトにおいて現実的な精度を実現できることを確認した。今後は、大量の検索要求を処理するフルテキスト検索システムにおいて、高速に候補を提示する手法について検討する予定である。

## 参考文献

- [1] G. Salton: Automatic Text Processing, Addison Wesley, 1989
- [2] 杉崎, 森, 田中: 検索入力の傾向分析によるフルテキスト検索の高速化, 第 60 回情処全大 3U-02, 2000