絞り込み検索語候補の抽出に関する一検討

3Y－3

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。我々は、再検索を支援するために、元の検索条件と関連の強い語を、絞り込み等のための検索語の候補として提示する方法について研究を進めている。本稿では、実用的なシステムの中への適用に向けて、検索結果の文書集合の中から短時間で候補となる語を抽出する方法について報告する。

2 絞り込みの候補となる語

絞り込み検索等において追加する語の候補として、元の検索結果で得られた文書集合の中での頻度が、テキストデータベース全体の中での頻度に比べて相対的に大きい語（以下、相対頻度の大きい語と呼ぶ）を採用する。相対頻度が高い語は、検索結果の文書集合に特徴的な語であり、元の検索条件との関連が強い語であると考えられるからである。

3 抽出における問題点

上記の様な語を実際の検索時に抽出するためには、検索結果の文書集合中に出現するすべての語を取りだし、各語の、テキストデータベース全体での出現頻度と、結果文書集合中での出現頻度を求め、その比を計算する必要がある。前者については、あらかじめ求めておけばよいが、後者の頻度は検索が行なわれるたびに数えなければならない。検索結果の文書が多くなってくると、計算時間が増加し、実用的な時間でユーザにレスポンスを返すことができなくなるという問題がある。

そこで本研究では、これを近似的に解決するために、検索結果の文書から適当にサンプリングして得た文書集合中での頻度を求めることにした。サンプリングの方法としては次の二つの方法を考えた。なお、検索結果は $tf \cdot idf$ などの一般的なやり方でランクづけされているものを考えている。

(1) ランキングの上位 N 件（決まった数）を抜き出す

(2) ランキングの中で等間隔に N 件を抜き出す

4 実験

前節で述べた方法でサンプリングを行なった時に、サンプリングしない場合と比べて、抽出される語がどの程度一致するかについて実験を行なった。テキストデータベースとしては、Webページ紹介文（文書数約10万）を用いた。あらかじめテキストを形態素解析し、すべての自立語を抽出してテキストデータベース全体での頻度を計算しておく。

検索結果の文書数が1000–1500程度になる検索条件を100個選び、それぞれの結果の文書集合から先に述べた方法で絞り込み検索の候補となる語を抽出し、相対頻度の値でソートした。前節の2つの方法でサンプリングした場合に抽出される語が、サンプリングしない場合に抽出される語と一致する度合の平均を表1に示す。なお、サンプル数 N として10、100、1000の場合を、また抽出された語のうち相対頻度の高い上位20語での一致度と上位50個での一致度を調べた。

	上位サンプリング		等間隔サンプリング	
N	上位20語	上位50語	上位20語	上位50語
10	9.8 %	13.0%	5.5%	10.0%
100	32.4%	28.5%	31.6%	24.6%
200	43.4 %	40.4%	43.3%	39.0%
500	59.1 %	56.1%	64.5%	59.4%
1000	81.7%	79.7%	85.8%	84.5%

表 1: 抽出された語の一致度

5 おわりに

絞り込み検索のための候補となる語を検索結果の文書の中から実用的な時間で抽出するために、文書をサンプリングする方法を提案し、サンプリングしない場合とどの程度抽出される語が一致するかを調べる実験を行なった。今後は、実システムでの有効性の検証を行なう予定である。

A Study on Term Extraction for Query Refinement
Takafumi INOUE, Masayuki SUGIZAKI,
Kazuhiro HAYAKAWA, and Kazuo TANAKA
NTT Human Interface Laboratories