6Z－1

InfoBeeフルテキスト検索システムにおける インデキシング方式の検討

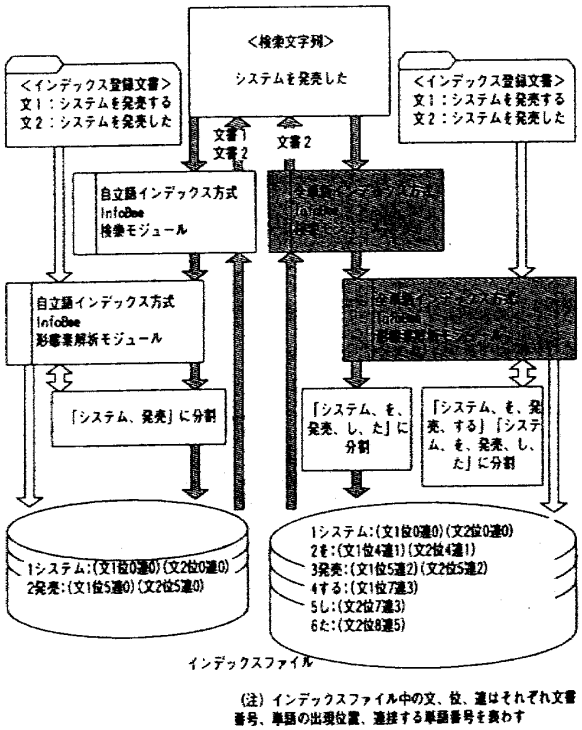
稲垣 博人† 渡邊 貴之†† 橋谷 隆幸†† 田中 一男†
†NTTヒューマンインタフェース研究所
††NTTインテリジェントテクノロジ（株）

1 はじめに

企業、あるいは個人におけるPCの普及、インターネット利用者の驚異的な増大により電子文書などの流通量が増大している。これに伴いそれらの電子化データを高速に検索できるサービスへの要求が高まってきており、実際国内外においてさまざまな検索サービスが開発されている。こうした中、InfoBeeフルテキスト検索システムはインターネットホームページの大規模な電子文書に対応した日本語検索エンジンとして開発され、実際に運用されている。InfoBeeフルテキスト検索システムは形態素解析を用いたインデックス型の検索エンジンであるが、現在は自立語のみをインデックス化し、このインデックスファイルを元に検索を行う。この方式では自立語以外の助詞などの単語を伴う特殊表現などを検索する事が難しい。先に述べたとおり検索対象となる電子化データの総量は増加の一途をたどっており、望む情報をより的確に示すノイズの少ない検索エンジンへのユーザ要求が高まってきている。本論文では自立語以外の単語もインデックス化し検索を行うエンジンを作成し、全単語インデックス方式による検索特性の検討を行う。

2 インデックス検索方式

InfoBeeフルテキスト検索システムでは、入力された文字列を形態素と呼ばれる単語単位に文字列をわけ、インデックスファイルを作成すると同時に検索文字列の分割を行う。形態素の切り分けにはInfoBee/TCという形態素解析モジュールを用い、切り出された各単語にそれぞれユニークなID番号を付与する。インデックスファイルはこのID番号と出現回数、出現位置を元に作成され保存される。検索時には、入力された検索文字列を同様の手法で形態素に分割し、先に作成したインデックスファイルから該当する文字列を検索する。さらに、例えば"日本国憲法"などのように複数の語からなる複合語については各単語の連接関係の情報もインデックスファイルに書き出している。



(注) インデックスファイル中の文、位、連はそれぞれ文書番号、単語の出現位置、連接する単語番号を表わす

図1 .異なる2インデックス方式による検索状態

通常のInfoBeeフルテキスト検索システムでは、切り出した形態素のうち、それ自体で意味を持つ名詞や動詞などの自立語と呼ばれる単語のみを用いて、自立語インデックス検索方式により検索を行っている。自立語インデックス検索方式では検索の難しい、助詞などの自立語以外の語や記号なども検索時に利用できるように、インデックス化する電子文書に出現する全単語、

A Study on Indexing Method of InfoBee Full Text Search Engine
†Hirohito INAKGAKI, ††Takayuki WATANABE, ††Takayuki HASHIYA, †Kazuo TANAKA
†NTT Human Interface Labs.
††NTT Intelligent Technology Co.,Ltd.

全文字を処理する全単語インデックス検索方式のモジュールを用意した。これは、形態素の切り出しまでは通常の InfoBee フルテキスト検索モジュールと同等のモジュールを利用し、インデックスファイルの書き出し時、および検索キーワードの切り分け時に語幹と活用語尾などを含む全ての字句が検索対象となるようにした。

3 実験方法 実験結果

使用マシン

SUN Ultra2 (CPU : UltraSPARC Ⅱ 300MHz　メモリ : 1 Gbyte)

測定対象

日刊工業新聞 1993 年 3 月～同年 7 月分記事

原文サイズ : 約 64 Mbyte

記事件数 : 約 17,000 件

実験方法

測定対象である文書中から 300 文字程度の文を紙面 1 月当り 2 件、計 10 件を無作為に抽出した。この 10 件の文を検索文字列として、文後方から 1 文字（2byte）づつ切り落とし、プログラムによって連続検索させた。検索は同一の検索文字列を用いて、自立語インデックス検索方式と全単語インデックス検索方式とで交互に行い、さらにこの連続検索を約 1 分間のインターバルを置いて 2 回、さらにこのループを 5 回、計 10 回行った。

	自立語インデックス検索方式	全単語インデックス検索方式
インデックスファイルサイズ	約 51 Mbyte	約 93 Mbyte
インデックス登録のべ単語総数	約 1,200 万語	約 2,300 万語

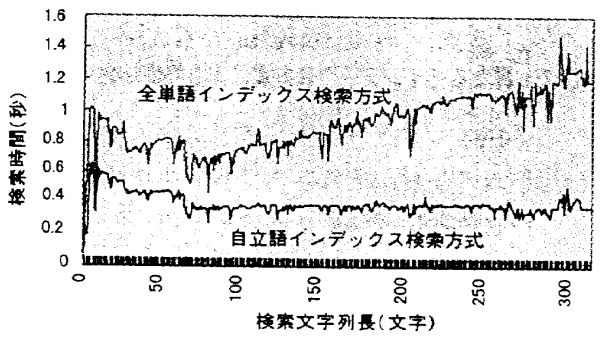


図 2．検索文字列長と検索時間

4 インデックス型検索方式の評価

4－1 検索時間

図 2 の結果に現れているように全単語インデックス方式の検索では検索文字列の長さが増加するほど検索時間がかかっている事が分かる。これは主にインデックス登録単語の増加に伴う探索空間の広がりに起因しているものと考えられる。

また検索文字列長が短い時に検索時間が一時的に増大している個所があるが、これは検索手法によるものであると考える。InfoBee では検索を高速化するために、検索文字列を形態素に分割した後、その形態素のうちインデックスファイル上でもっとも登録回数の少ない語から検索を開始するようになっている。このように検索開始ポイントの空間を狭める事で検索の高速化を図っているのだが、短い検索文字列の場合、探索空間の狭い語の出現割合が減少するために、ある文字列長までは一時的に検索時間が増加するものと考えられる。

4－2 検索ノイズ

フルテキスト検索エンジンのニーズを考えた時、そのひとつに、ユーザが望む情報をより少ない検索条件で一意に決定できる検索エンジンがある。本検索方式を用いれば、検索文字列が的確な形態素単位に分割された場合、一意に検索したい情報を決定することができる。また、今回の実験方法のように、正解文書を 1 件とし、正解文書中の文字列を検索入力とするような検索実験では、自立語インデックス検索方式と比較し、半分以下の検索入力で正解文書を探索することができた。

5 まとめ

日本語の全文検索エンジンで、最適な検索結果を出すものや、ユーザの検索要求を完全に満たすようなシステムは未だ構築されていない。しかし、検索システムの利用用途は確実に増えており、またそれを利用するユーザのニーズも多様化してきている。今回実験に用いた全単語インデックス検索方式の検索エンジンは、ニーズに応じて多様化していくであろう検索エンジンの一つになると考える。

参考文献

[1]田中 : InfoBee 検索エンジンを用いたディレクトリ検索サービス, NTT 技術ジャーナル, 8 ,pp24-27, 1996.