日本語解析を用いたフルテキストサーチの実験

4B−4

中本　幸夫*　田野崎　康雄**　岩井　勇**
*:東芝ｺﾝﾋﾟｭｰﾀｴﾝｼﾞﾆｱﾘﾝｸﾞ(株)　**:(株)東芝 情報処理機器・技術研究所

1．はじめに

文書の電子化が進められ大容量のデータベースが構築されている。ワープロやＯＣＲの普及により、大量の文書テキストがコード化されるにつれフルテキストサーチによる文書検索が求められている。しかし、単純に全データベースをなめて検索する方式では、データ容量が増すにつれて検索速度が遅くなり、また、不要データも多く検索されてしまうという問題点がある。

そこで我々は、日本語解析を用い、意味のまとまりである単語に注目し、データ容量に左右されず高速に検索する方式を開発した。この解析では、同義語や類語処理も行う。

本稿では、フルテキストサーチによる高速文書検索方式の概要と、技術文書を対象とした実験システムの評価結果について述べる。

2．概要

本検索方式の処理の概要を図１に示す。処理部は大きく２つの部分に分かれている。第１はインデックス作成部でテキストデータを辞書を用いて日本語解析し、文書中から全ての単語を抽出しフルテキストサーチのためのインデックスを自動作成する。第２は検索部でインデックスを用いてユーザの自由なキーワードによるフルテキストサーチを高速処理する。

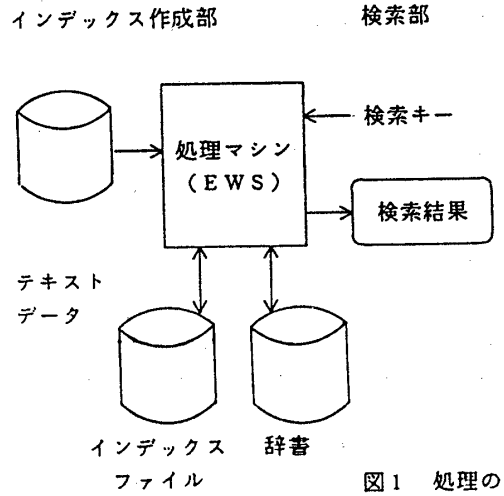

図１　処理の概要

3．日本語解析を用いたフルテキストサーチ

3-1　インデックス作成部

フルテキストサーチに用いるインデックスは、テキストデータを日本語解析し、文章中から単語を自動抽出し単語単位のインデックスファイルを自動作成する。日本語解析は以下に述べるように各種辞書を用い単語単位の解析を行う。

（１）日本語形態素解析処理

形態素解析は、約１０万語の解析辞書を用いて検索対象文書の単語切りを行う処理である。切り出された単語には品詞情報が付加され、検索キーとして有効な品詞、たとえば「名詞」がつけられた単語を検索用インデックスとして抽出する。（図２）

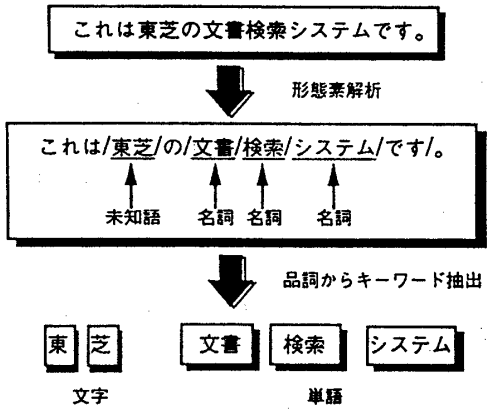

図２　形態素解析処理

（２）異表記処理

「コンピュータ」と「コンピューター」のように同一意味で表記が異なる語は異表記辞書を用いて、代表語に置き換えてインデックスファイルを作成する。

（３）複合語処理

文書中には複合語が多く含まれている。複合語処理は、形態素解析処理で抽出された単語をさらに意味単位に分割する処理である。これにより、例えば、検索対象文書中の「文書を検索する」と検索キーの「文書検索」が図３に示すようにマッチし、検索漏れを防ぐことができる。

An experiment of Full-Text Retrieval by Japanese Analysis Method
Yukio NAKAMOTO, Yasuo TANOSAKI
*　Toshiba Computer Engineering Corp.
**　Information Systems Engineering Lab. Toshiba Corp.

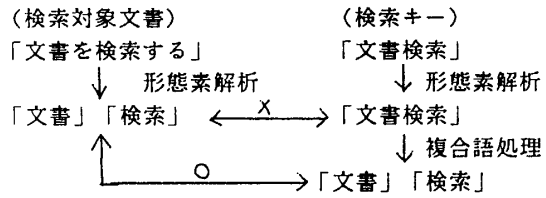

図3　複合語処理

（4）未知語処理

形態素解析で辞書にない語は未知語となる。未知語は文字単位に分解し、文字単位で検索できるインデックスを作成し、検索漏れを防いでいる。

3-2　検索処理部

インデックス作成部で作成したインデックスを用いて検索を行う。検索処理は、ユーザが指定したキーに対して、インデックス作成部と同一の日本語形態素解析、異表記、複合語、未知語の各処理を行う。そして、インデックスを参照して検索結果の出力を行う。

「パソコン」と「ＰＣ」のような同義語は同義語辞書に登録することにより、検索時にＯＲ条件で処理する。同義語の場合は、ユーザによって概念、使い方が異なることを考慮し、ユーザごとに登録、指定を行えるようにした。

単語単位に検索を行うことにより検索ノイズを減らし検索精度を向上させる効果がある。例えば、「導体」を検索キーとしても意味の異なる「半導体」は検索されない。（図4）

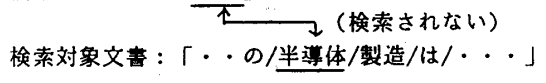

図4　単語単位の検索

4．実験

4-1　実験システム

技術文書（東芝レビュー）を検索対象文書とした検索システムを作成した。表1に本システムのデータ概要を示した。インデックス作成は、検索対象文書を形態素解析処理し、「名詞」の品詞情報が付加された単語は単語単位に、「名詞」以外の単語は文字単位に分解し、検索インデックスを作成した。

検索実験にはUNIXワークステーション(AS4470)を使用した。複数のキーからＡＮＤ／ＯＲ／ＮＯＴ検索を組み合わせて検索処理を行った。（図5）

4-2　評価及び考察

（1）インデックス作成

インデックス作成はバッチ処理でシステム構築時に行う。文書の追加、削除、修正時にはインデックスが自動更新される。インデックス容量は、テキスト容量の

表1　データ概要

検索対象文書	東芝レビュー３年間分 ５７０文書（約２５００頁）
テキスト容量	約　６．８Ｍバイト
インデックスに 登録した単語数	約　１１０００単語
インデックスに 登録した文字数	約　　３１００文字
インデックス容量	約　　１．１　Ｍバイト

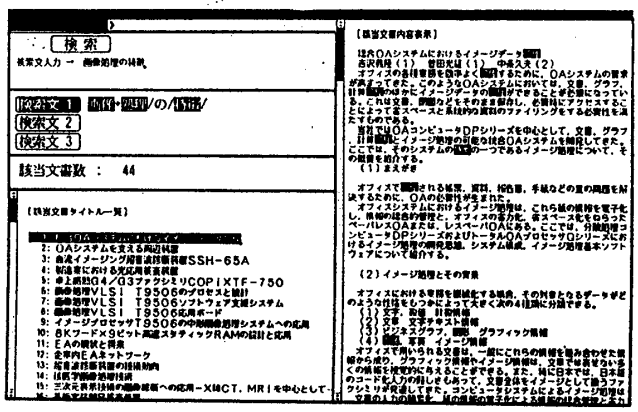

図5　実験システムの画面

約１／６と小さく抑えることができた。テキストデータを形態素解析処理するために現在は多少時間を要する。今後形態素解析の精度向上と高速化が必要である。

（2）検索速度

インデックスを用いることで１秒以下の検索が可能であった。本検索方式は文書件数が増加しても検索速度には影響を受けにくい方式である。今後は大容量文書での評価を続けていく。

5．おわりに

日本語解析を用いたフルテキストサーチによる文書検索を開発した。インデックス作成部ではテキストデータを日本語解析により全ての単語を抽出し、単語単位のインデックスファイルを自動生成する。検索部では自由なキーワードによる高速全文検索を行う。

実験システムでは技術文書を用い、高速検索、本方式の有効性を実証できた。

今後、大容量検索の実験とシソーラス辞書や同義語辞書などの辞書を拡張整備し、曖昧検索手法について検討を進めていく予定である。