フルテキストサーチにおけるフィールド検索の実験

7E－1

野上謙一[1]　中本幸夫[1]　森谷精得[2]　安高みわ[3]　内海正樹[3]　田野崎康雄[4]
[1]東芝コンピュータエンジニアリング（株）　[2]東芝アドバンストシステム（株）
[3]（株）東芝　青梅工場　[4]（株）東芝　マルチメディア技術研究所

1．はじめに

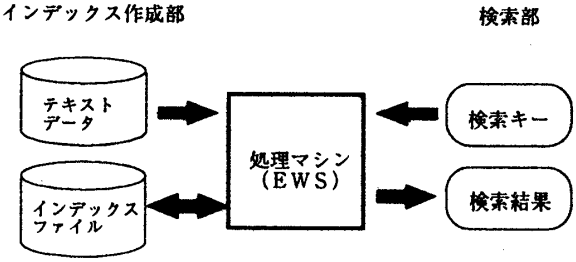
コンピュータ、ワープロなどの浸透により、文書の多くは電子化されることが当然のようになってきている。そして、電子化された文書の多くは、学校や企業等でデータベース化され、そのデータベース中から、ユーザの要求している文書をフルテキストサーチにより検索を行っている。しかし、従来のフルテキストサーチでは、ユーザが入力した検索キーワードを含む文書を抽出するため、ユーザが意図していない文書も抽出してしまうのが現状であった。そこで、フルテキストサーチによる検索が普及するに従って、ユーザはより精度の高い検索を求めるようになってきた。

そこで、我々は文書における「表題」や「まえがき」などの文書の構造を示す範囲（以下、フィールドと称する）に注目し、ユーザが任意にフィールドを指定することで、ユーザが本当に意図している文書を検索することを可能にするフィールド検索方式を開発した。

本稿では、フルテキストサーチにおけるフィールド検索の実現方式と、技術文書を対象とした評価実験について報告する。

2．フルテキストサーチの概要

従来のフルテキストサーチにおける検索処理の概要を図１に示す。処理部は大きく２つに分かれている。第１は、インデックス作成部で、テキストデータからフルテキストサーチにおけるインデックスを自動作成する。第２は、検索部で、ユーザが任意に指定した検索キーワードを含む文書をインデックスを参照することにより、高速検索を行う。

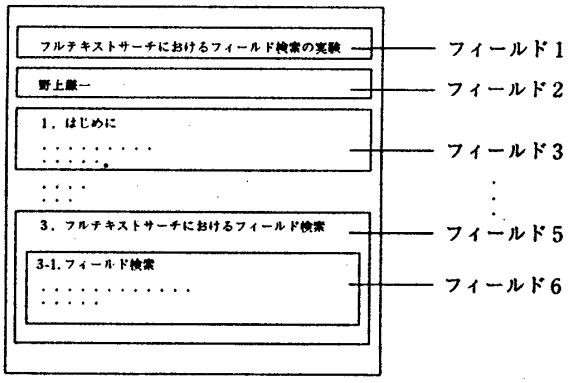


【図１】処理の概要

3．フルテキストサーチにおけるフィールド検索

3-1. フィールド検索

フルテキストサーチでは、ユーザの指定した検索キーワードを含む文書を抽出するため、ユーザの意図している文書とは異なる文書も抽出してしまうことがある。そこで、図２に示すように、「はじめに」など、文書の構造を示す特定のフィールド内にユーザの指定した検索キーワードが含まれている文書のみを抽出するように、フィールド指定が可能な検索方式が求められていた。



【図２】フィールドの例

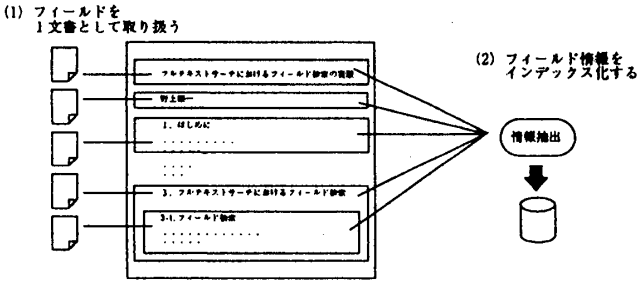
図２に示すような、文書の「表題」「著者名」「はじめに」などの文書構造を示すフィールドを指定して検索できるフィールド検索を実現する方式を開発した。

3-2. フィールド検索の実現方式

構造化された文書の各フィールドを１文書として取り扱うことにより、フィールド検索を実現することが可能である（図３(1)）。しかし、この方式を用いた場合、複

An Experiment of Full-Text Retrieval for Structured Documents
Ken'ichi NOGAMI[1], Yukio NAKAMOTO[1], Kiyonori MORIYA[2], Miwa ATAKA[3], Masaki UTSUMI[3], Yasuo TANOSAKI[4]
[1]Toshiba Computer Engineering Corp.
[2]Toshiba Advanced System Corp.
[3]Ome Works, Toshiba Corp.
[4]Multimedia Engineering Lab., Toshiba Corp.

数の文書が１文書として取り扱われるため、文書管理が複雑になる。そのため、インデックス作成時や検索時の演算速度等に問題が生じてしまう。そこで、検索対象文書中から、フィールドに関する情報（フィールドの区切り等）のみを抽出し、インデックス化する（以下、フィールドインデックスと称する）。そして、このフィールドインデックスと従来の検索方式のインデックスとを併用することにより、フィールド検索を実現した（図３(2)）。



【図３】フィールド検索の実現方法

また、検索キーワード単位にフィールドの指定が可能な方式を採用しているため、下記に示すようなフィールド間の論理演算が可能である。

（「表題：検索」OR「表題：フルテキストサーチ」）
AND「まえがき：フィールド」

４．実験

4-1.実験対象

今回、検索対象文書として実験に用いた技術文書（東芝レビュー）は、あらかじめ決められた文書構造で記述されているため、実験対象として採用した。
実験に用いた技術文書のデータおよびフィールドの種別を表１に示す。

なお、フィールドの種別は、検索対象文書の文書構造のうち、固有のフィールドを選択している。

また、今回設定したフィールドは、入れ子（フィールド内にフィールドが出現している状態）も含めている。

4-2.評価および考察

実際に、フィールド検索に必要なインデックスを作成し、検索を行った。

【表１】検索対象文書のデータ

検索対象文書	東芝レビュー 517文書（約2,300頁）
テキスト容量	約 5.9 MB
分割フィールド数	8
フィールド種別	表題・著者名・ｱﾌﾞｽﾄﾗｸﾄ まえがき・本文１ あとがき・参考文献 本文２（まえがき、あとがきを除く）

(1) インデックス容量

インデックス容量は、フィールドインデックスを作成するため、従来のインデックス容量よりも増加する。しかし、実際の増加容量は 10%程度であった。小容量のフィールドインデックスを、従来の検索方式で作成されたインデックスに追加することにより、フィールド検索を可能にすることができた。

(2) インデックス作成時間

インデックス作成時間は、フィールドインデックスを作成するため、従来のインデックス作成時間よりも時間を要してしまう。しかし、前述した通り、フィールドインデックスは、従来のインデックスの10%程度と小容量であるため、ほぼ同等の時間でインデックスを作成することができた。

(3) 検索性能

検索方式は、従来の検索方式と同一方式であるため、従来通りの高速な検索が可能となっている。また、検索キーワード単位にフィールド指定を行うことが可能なため、細かな検索式を作成することができる。これにより、従来の検索方式よりも、よりユーザの意図した文書を抽出してくる精度の高い検索を行うことが可能となった。

これらのことから、今回実験を行ったフィールド検索が有効な方式であると考えられる。

５．おわりに

今回、フルテキストサーチにおけるフィールド検索方式を開発した。評価実験では、これまでの全文検索を行うためのインデックスに、わずか 10%程度の小容量のフィールドインデックスを追加することで、フィールド検索を実現することが可能となった。また、検索方式は従来と同一方式のため、高速検索が可能となっている。これらのことより、本フィールド検索方式の有効性を実証できた。

今後、技術文書とは異なる構造を持つ文書やＳＧＭＬ対応文書などを対象に、さらに実験を行っていきたい。