4J－11

文字成分表型全文検索システムの SGML文書検索への拡張

藤井洋一　今村誠　高山泰博　森口修　鈴木克志
三菱電機(株)　情報技術総合研究所

1．はじめに

文書の電子化が進み、大量の電子化文書を計算機によって扱う必要性が増すに従い、電子化された文書内容を扱うための全文検索方式がさまざまに提案されてきた。全文検索は、電子化された大量文書中から、任意の文字列をキーワードとして検索することができる検索方式であり、我々はファイリングシステムの検索機能として、全文検索機能を製品化している。

一方、文書の構造自体を電子化された文書中に持たせた構造化文書がある。構造化された文書は、構造によって文書の意味が定義されるため、より細かい意味付けが文書中でなされている。その中で、SGML (Standard Generalized Markup Language)規格[1]に従った構造化文書が、最近注目されている。SGML文書は、文書型定義(DTD: Document Type Definition)が明確に定義された上で、文書が作成されるので、計算機で扱いやすい構造化文書である。全文検索を構造化文書に対応することで、検索ノイズを減少させ効率良い検索が可能となる。

今回、全文検索の一方式である、我々の文字成分表型全文検索システムをSGML文書検索に対応出来るよう拡張したものを試作したのでその方式及び、問題点を考察する。

2．文字成分表型全文検索

文字成分表型全文検索システムは、登録文書に対して、文書中に出現する文字の情報を文字単位にビット情報として文字成分表として格納する。検索時には、文字成分表を用いることで検索条件を満たす可能性のある文書を前もってフィルタリングして絞り込み、実際の文書内検索を行う対象を少なくして、システム全体の性能を向上させ、高速な検索を実現する方法である。(図１)

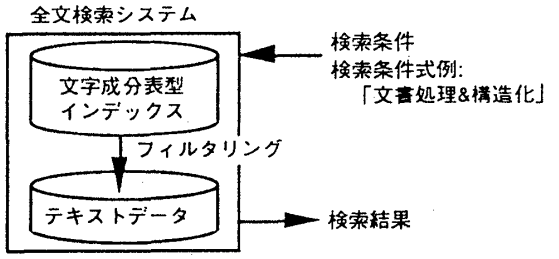

図１　全文検索方式

3．SGML文書への対応

全文検索では検索キーワード（文字列）のAND/ORによって検索条件を設定するが、SGML文書では、文書中に付与されるタグ情報が重要な意味を持ち、検索条件としても、タグを利用した検索が柔軟に行えることがポイントとなる。

(1)インデックス情報の追加

SGML文書のタグの位置情報をインデックスとして検索用に作成し、文字成分表による絞り込みを行った後で利用する方式を取ることとした。

(2)検索条件式の記述方式

従来の全文検索による検索では、文書中にキーワードが存在するかどうかを評価し、AND/ORによって検索条件式全体を評価することによって、検索結果かどうかを判定する。しかし文書のタグを利用した検索を行う場合は、全文検索による検索条件式を単純にタグによる制限を記述できるよ

Extension of A Signature Based Full-text Retrieval System for SGML Documents
Youichi FUJII, Makoto IMAMURA,
Yasuhiro TAKAYAMA, Osamu MORIGUCHI,
Katsushi SUZUKI
Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa 247, Japan

うに拡張した[2]のでは、「<章>の<タイトル>が"はじめに"でその<章>の中に"文書処理"を含む文書を検索する」といった複雑な検索条件を記述することが困難である。そこで、文書中の位置情報の集合を部分検索条件の検索結果として処理し、検索結果かどうかは、集合が空かどうかで判定する方式に変更した。(図2)

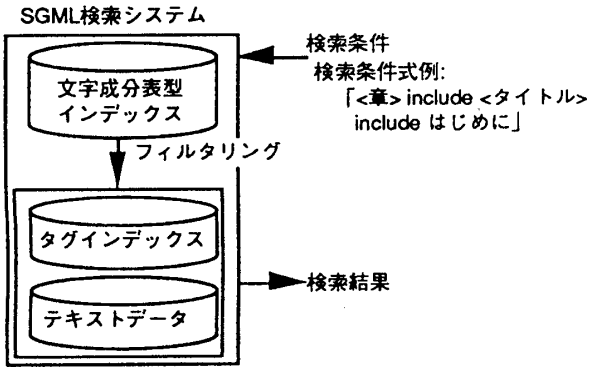

図2　ＳＧＭＬ文書検索方式

4．問題点と考察

文字成分型のインデックスを用いた全文検索をSGMLのタグを利用した検索を可能としたことで、基本的なSGML文書の検索機能は実現することが出来た。今回、検索はWWWクライアントのMosaic等を利用したので、検索条件式は直接入力する方式を取ったが、複数のDTDで記述されたSGML文書が格納された中から検索しようとする場合は、文書型定義がわからなければ、検索を実行することもままならない。

そこで、SGML文書を登録する時点でDTDの情報も格納し、そのDTDの情報を表示して、ユーザに検索するための情報を提供し、検索式の作成を助けることが必要不可欠と考える。

(1)ビジュアル表示による検索式の作成

DTDを木構造表示[3]させ、その木構造上で検索条件式を作成できるようにすることによって、文書の構造を十分に把握していないユーザにも詳細な検索条件の設定を可能とする。(図3)

(2)その他の機能

DTDを木構造表示した検索条件の設定環境を提供しても、実際にはSGMLのタグ名が長さの制限を受ける[1]ため省略表記され、タグ名の意味が不明瞭であったり、同じタグ名が別の構造上で出現する場合があるため、以下の機能が必要と考える。

・DTD中のコメントを参照表示する機能
・特定のタグのパスを別名で定義しておき参照できる機能

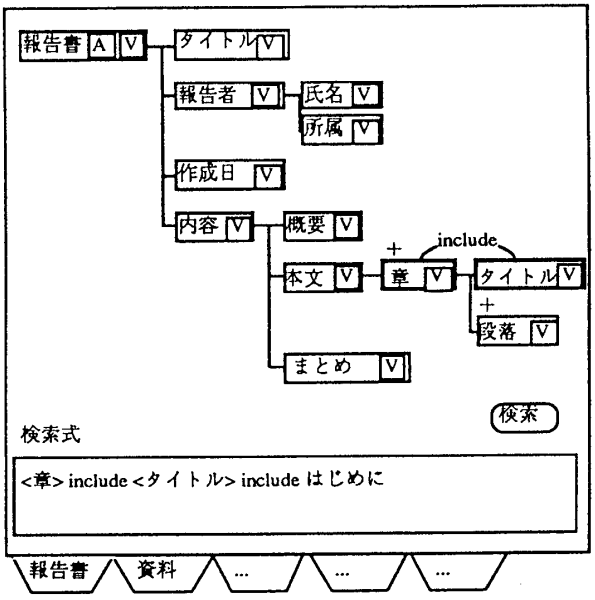

図3　検索式作成の例

5．おわりに

今回、SGML文書検索のための基本的な機能を実現し、問題点を考察した。今後は、考察で述べたユーザインタフェースの実現を含めた、より使い易いSGML文書検索システムへの展開を考える予定である。

参考文献

[1] JIS X 4151「文書記述言語SGML」(1992)
[2] 野上謙一他「フルテキストサーチにおけるフィールド検索の実験」情処51回全大 7E-1(1995)
[3] 原正一郎他「文書構造に基づく全文データベース検索用利用者インタフェースの試作」信学会秋季大会D-54(1993)