SGML/HTML 文書 DB におけるテキスト格納方式の提案

2Q－5

波内 みさ　鶴岡 邦敏
NEC C&C 研究所

1 はじめに

SGML 文書などの構造化文書をデータベース (DB) に格納し、その構造情報を使って文書を検索するための手法が、現在までに数多く提案されている。

これらのうち、文書をその構成要素 (element) ごとに分割し、文書部品として DB 中に管理する手法がある。この手法により、各 element (文書部品) を単位として、他の構造化文書で再利用することができる。しかし、一つの文書中のすべての element に対応して生成された文書部品がすべて再利用されるとは限らず、逆に部品化によるオーバヘッドが生じる場合がある。

本稿では、SGML 文書中の各 element の役割に着目した DB への格納方法を提案し、ユーザ指定により必要な element のみを部品化するための機構を示す。

2 文書分解アプローチとその問題点

SGML 文書を DB で管理する手法の代表的なものとして、文書を element 毎に分解して DB に格納する「文書分解アプローチ」と、文書本体は DB の外部の文書検索システムに格納するが、DB に新たなデータ型を導入し、他の DB データとの統一的なアクセス手法を提供する「文書データ型導入アプローチ」の 2 つが知られている [1]。

このうち、文書分解アプローチを利用して各 element をオブジェクト指向データベース (OODB) で管理する方式が提案され、いくつかの SGML 文書 DB 製品でもこの方式が取り入れられている。この場合、各 element をどのようなオブジェクト (クラス) で管理するかによって、文書部品を定義する DB スキーマにいくつかの選択肢が存在するが [2]、いずれもすべての element を部品化することによって以下の問題が生ずる。

1. 部品として再利用される可能性がほとんどない element もオブジェクトして格納されてしまう
2. DTD (Document Type Definition) において、その内容が「テキスト」と定義されている element に対してのみ、テキスト・データを保持したオブジェクトを生成すると、元文書全体あるいは文書の指定された一部分を取り出す場合に、上記テキスト・オブジェクトを集めて再合成しなければならない

問題 1は、例えば文書の章や節の題名や箇条書の一項目などの、比較的小規模のテキストを扱う element が文書部品として DB の管理対象となることを指す。文書部品は、一般に、その親部品、子部品へのポインタや更新履歴などの再利用のための情報を持つが、再利用の可能性がほとんどない element に対しては、これらの情報は不要である。

問題 2を回避する手段の一つに、各文書部品に、対応する element とそれが包含するすべての element のテキストを重複して格納する手段がある。しかし、element のネストが深い場合にはその重複部の冗長さは増し、部品化されていることによる簡便さと格納容量とのトレードオフが生じることになる。

これらの問題は、SGML が様々な目的の element をその DTD に自由に定義可能であることに起因する。ある DTD に則る SGML 文書のどの element に再利用される可能性があり、オブジェクトとして部品化し、テキストを (重複してでも) 格納しておく価値があるかは、一般に、システムが自動的に判断することは不可能である。このため、すべての DTD に対してシステムが最適な格納方法を与えることはできない。

これに対して我々は、各 element の文書中での役割を考慮した 3 種類の格納方式を用意し、扱う SGML 文書の特性によってユーザが格納方式を選択可能な文書 DB システムを提案する。

以下、提案する格納方式とその指定方法について述べる。

3 OODB における SGML 文書 element の格納方式

3.1 文書部品化方式

SGML 文書中の element には、特定の値や叙述部を管理するものの他に、章、節の題名など他の element の補助情報として利用されるものや、箇条書きの宣言など他の element の構造を示すためのものなどがある。これら element の用途の違いは、文書あるいは特定の文書部品の検索において利用することができる。

A Proposal for Text Storing Strategy on an SGML/HTML Document Database

Misa NAMIUCHI (nami@swl.cl.nec.co.jp)
Kunitoshi TSURUOKA (tsuruoka@swl.cl.nec.co.jp)
C&C Research Laboratories, NEC Corporation

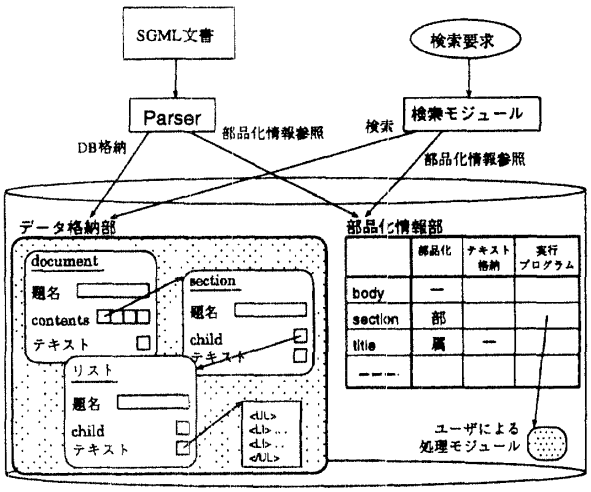

図 1: ユーザの部品化指定による SGML 文書 DB

ここでは、DB 中の文書および文書部品を検索する場合に、各 element がどのように利用されるかに着目し、以下の 3 種類の格納方法を提案する。

(1) 文書部品としてオブジェクトを生成して格納

(2) 親 element を管理するオブジェクト (親部品) の属性として格納

(3) 部品化せず、親 (先祖) 部品の一部として格納

(1) は、文書部品として検索頻度が高く、再利用される可能性が高い値・叙述部を管理する element に対しての適用が効果的である。対応するテキストは元文書から切り出し、オブジェクトに文書部品として格納する。

これに対して (2) は、(1) で格納する element の補助情報として使われる element の格納方法として効果的である。そのテキストを、(1) で生成した文書部品 (オブジェクト) の属性値として格納することにより、単体での利用頻度が低い部品の生成を抑制することができる。また、そのテキストを親部品を特定するためのキーとして利用することが可能となる。

(3) の格納方式は、箇条書きや表など、その集合体が一つの意味を持つとき、複数の element を一つの部品としてまとめて管理する場合に利用する。element の検索要求に対しては、その都度、親部品、先祖部品を解析し、そこから取り出して対応する。

3.2 利用者による部品化指定と部品オブジェクト生成方式

図 1に、提案方式を実現する SGML 文書 DB システムの構成を示す。

3.1節で述べた element の格納方式は、予めユーザが指定するものとする。ユーザは、DTD に定義された各 element 宣言に対応し、それぞれの DB への格納方法を部品化情報部に定義しておく。ここではさらに、ユーザによる処理モジュールを特定の element に関連付けて宣言しておくことによって、element の DB 格納直前にそれを実行し、テキストを AP が利用しやすい形に加工しておくことを可能とする。

SGML 文書を DB に登録する際、まずパーザが与えられた文書インスタンスを解析する。このときパーザは、部品化情報部を参照し、指定された方法で element を格納する。

検索時には、検索モジュールが部品化情報部を参照し、検索対象とそれに到達する過程の element の格納情報を使って、目的の element を得る。

4 HTML 文書への適用

WWW において広く利用されている HTML は SGML の一種とみなることができるので、上記格納方式を利用することができる。

しかし、実際の WWW での HTML の利用方法は、各 element(tag) に定められたレイアウト機能のみが注目されがちで、element の文書構造上の意味は必ずしも統一的ではない。よって、上記提案方式は、element の意味付けの統一が可能な企業内イントラネットなどにおいて効力を発揮できると考える。

5 まとめ

本稿では、OODB における SGML 文書の格納方法に関し、element の役割を考慮した格納方法を提案した。さらに、その格納方法による文書部品化をユーザによって指定可能とすることにより、より確実な文書部品 DB の構築を支援する。

提案方式による SGML 文書 DB は、現在、弊社のオブジェクト指向データベース *PERCIO*[3] 上にプロトタイプを開発中である。これにより、ユーザ指定による部品化の有効性を含め、実際の SGML 文書を利用した機能・性能評価を試みたい。

参考文献

[1] 吉川,「構造化文書とデータベース」, Proceedings of Advanced Database System Symposium '95, pp. 49-57, 情報処理学会, 1995.

[2] 波内,「OODB による SGML 文書データベースの設計」, 情処 DBS 研究会予稿集, 109-52, pp.311-316, July 1996.

[3] 鶴岡 他,「オブジェクト指向データベース管理システム PERCIO の開発と今後の課題」, 電子情報通信学会論文誌 D-I, Vol.J79-D-I No.10, pp.587-596, Oct. 1996.