知識指向文書管理基盤の開発(3)
文書管理ミドルウェア DocumentBroker における構造化文書管理方式

5P－9

里 佳史† 松本 正義† 岡本 卓哉† 高橋 亨† 内角 真‡
(株)日立製作所 †システム開発本部 ‡ソフトウェア事業部

1. はじめに

近年、大量の電子化文書を構造化文書形式で格納し、高度な検索や再利用を行いたいというニーズが高まっている。そのため、筆者らはDocumentBroker[1]における構造化文書の管理機能を開発した。本稿ではまず構造化文書管理の中核となる概念 SGML 文書について説明し、次に、概念 SGML 文書においてエンティティ構成と論理構造とを管理するための仕組みを説明する。

2. 構造化文書の管理

2.1 概念 SGML 文書

DocumentBroker で扱う SGML 文書の例を図 1 に示す。

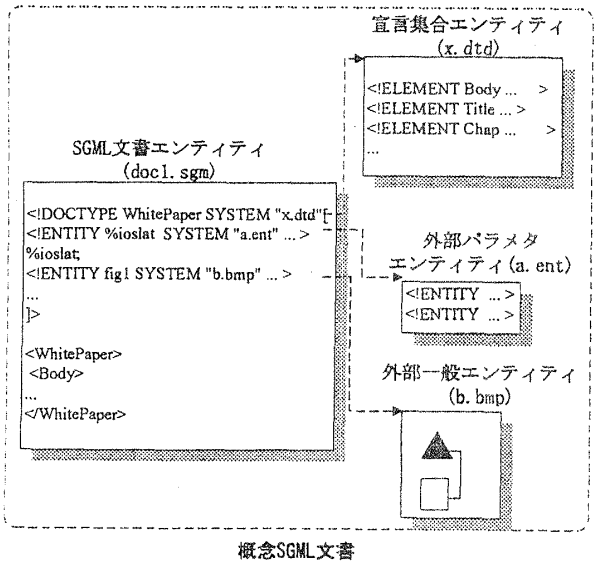

図1 概念 SGML 文書

一つの SGML 文書は、一般に複数のエンティティ(ファイルシステム上で一つのファイルに該当するようなデータ)から構成される。エンティティには、SGML 文書エンティティ(内部 DTD と、文書の内容データである文書インスタンスからなる)、宣言集合エンティティ(外部 DTD)、外部パラメタエンティティ(DTD から参照される宣言データ)および外部一般エンティティ(文書インスタンスから参照されるテキストやデータ)がある。一個の SGML 文書エンティティと 0 個以上の他のエンティティからなる SGML 文書を、本稿では「概念 SGML 文書」と呼ぶ。

DocumentBroker では、個々のエンティティをそれぞれ一つの文書オブジェクト、つまりDocVersion[1]として独立に管理しつつ、複合文書としての SGML 文書全体をも一個の文書オブジェクトとして管理するために、概念 SGML 文書の考え方を導入した。概念 SGML 文書は、具体的には文書の内容データを表す ContentTransfer[1]の位置に、ConceptualSgmlDocument オブジェクトを持つ DocVersion オブジェクトとして表現される。ConceptualSgmlDocument は、以下に説明する通り、概念 SGML 文書のエンティティ構成と論理構造の両方を統一的に管理する。

2.2 概念 SGML 文書のエンティティ構成の管理

図 2 に、概念 SGML 文書のエンティティ構成を管理するモデルの概要を示す。図 2 において最上位に位置する DocVersion は概念 SGML 文書自体に相当する DocVersion である。これに対して、下に並んでいる DocVersion は個々のエンティティに該当する。概念 SGML 文書に相当する DocVersion

---

Management of Structured Documents on the Enterprise Document Management System "Document Broker".
Yoshifumi SATO†, Masayoshi MATSUMOTO†, Takuya OKAMOTO†, Toru TAKAHASHI†, Makoto UCHIKADO‡
†Business & Information Systems Development Division, Hitachi,Ltd. ‡Software Division, Hitachi,Ltd.

が保持する ConceptualSgmlDocument は、概念 SGML 文書のエンティティ構成を管理する ConfigurationTable オブジェクトをそのメンバとして保持する。ConfigurationTable は、各エンティティを指し示すレコードの集合であり、各レコードには次のような情報が記述される。

- EntityName:エンティティのエンティティ名
- Role:エンティティの役割(文書エンティティ、宣言集合エンティティなど)を示すコード
- Head: エンティティデータを保持する DocVersion
- SystemID:エンティティのシステム識別子
- PublicID:エンティティの公開識別子
- VTMode:エンティティの最新バージョンに追従するか否かを示すモード(構成管理機能[2]参照)

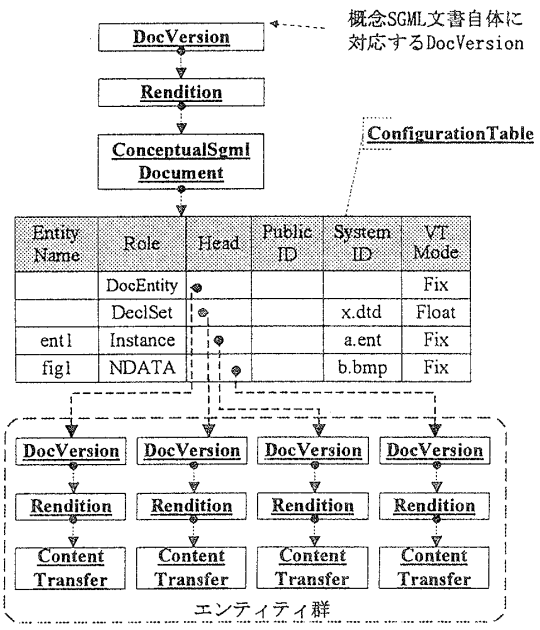

図2　エンティティ構成の管理

2.3 概念 SGML 文書の論理構造の管理

これらのエンティティの集合を一つの概念 SGML 文書として解釈した解析結果を管理するために、ConceptualSgmlDocument は SgmlInterpretation オブジェクトをメンバとして保持する(図 3)。SgmlInterpretation は、DTD 情報を管理する DeclSet オブジェクトと、文書インスタンスの論理構造を管理する DocInstance オブジェクトとをメンバとして保持する。文書インスタンスの論理構造は、Element(論理要素)や Cdata(内容データ)などのノードオブジェクトのツリー構造によって表現される。DocumentBroker はこれらのノードからなるツリー構造をアクセスするナビゲーション機能、及びタグ名を指定した Element ノードの検索機能を提供する。

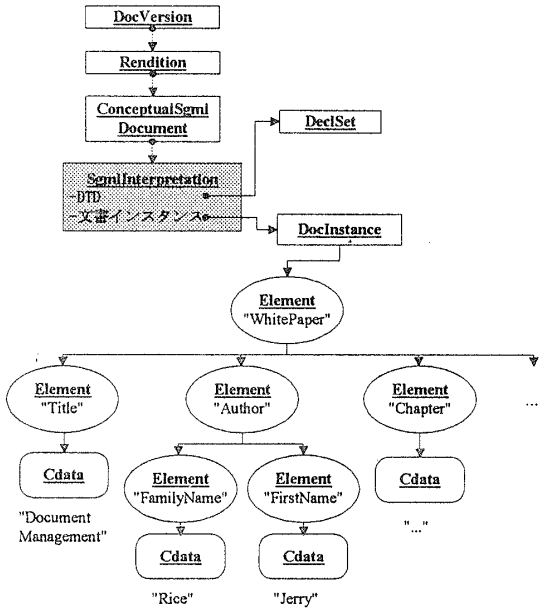

図 3　論理構造の管理

3. まとめ

文書管理ミドルウェア DocumentBroker における構造化文書管理機能を開発した。本機能は、SGML 文書を構成する個々のエンティティおよび概念的な SGML 文書全体を、それぞれ文書オブジェクトとして取り扱い、文書のエンティティ構成及び論理構造にアクセスする手段を提供する。今後は XML 文書への対応を進めていく予定である。

参考文献

[1] 三原、他: “文書管理ミドルウェア DocumentBroker のシステムアーキテクチャ”, 情報処理学会第 59 回全国大会 5P-07

[2] 青山、他: “文書管理ミドルウェア DocumentBroker における文書管理モデル”, 情報処理学会第 59 回全国大会 5P-08