5F－3

統合型文書データベースにおける オブジェクトと文書内容の相互参照機構の構築*

絹谷 弘子　吉川 正俊　加藤 弘之†

奈良先端科学技術大学院大学　情報科学研究科‡

1. はじめに

日々蓄積される電子化文書の再利用の観点から「ある日の記事中の'合併'という文字列に関連する会社について」の情報を獲得するには，記事内容と会社データとの相互参照機構を持った統合型文書データベースを検索することで可能となる．データベースシステムによる構造化文書の管理において，内容と構造の両方に関する問合せが可能となる文書データモデルとして**Paratext**モデルを提案してきた[YIU96), 加藤 97b), 加藤 97a)]．Paratextモデルでは，文書中の文字列が表す意味とデータベースオブジェクトとを対応させておくことで，データベースオブジェクト間の計算に基づく文書の操作や文書処理に基づくデータベースオブジェクトの操作が可能となる．本稿ではデータベースオブジェクトと文書内容の間に相互参照機構を保持する統合型文書データベース「**Paratext**データベース」のアーキテクチャと相互参照機構について提案し，その実装の概要を述べる．

2. Paratextモデルを実現するアーキテクチャ

2.1 Paratextモデルにおけるリンク

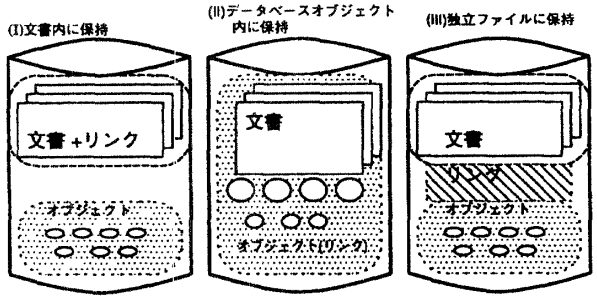

図1 リンク情報の格納場所

Paratextモデルで扱うリンクとは(1)文書中の文字列（文書中での開始位置，終了位置，部分文字列）「**表記層**」と(2)その文字列に対応付けされたデータベースオブジェクト識別子や原子型データ値の集合「**参照層**」との相互参照関係である．このリンク情報の格納場所として(I)文書オブジェクト内，(II)データベースオブジェクト内，(III)独立ファイル内の三通りが考えられる（図1）．さらにそれぞれに関して(a)文書がプレーンテキストの場合，(b)文書が構造化文書（SGML）でデータベースオブジェクト識別子を指定する特定のタグ<dbobject id=doc1.#01>を定義して，そのid属性を利用して文字列識別子を指定する場合，(c)文書が構造化文書でタグ<dbobject oid=#1234>に添加部分要素のoid属性を利用してリンク先のオブジェクト識別子(OID)を指定する場合が考えられる．(c)については，文書文字列内に表記層と参照層の両方の情報を持つことができる．さらに独立ファイルを持つ場合(α)OIDとidとの対応付け，(β)OIDと文字列出現位置との対応付けが考えられる．

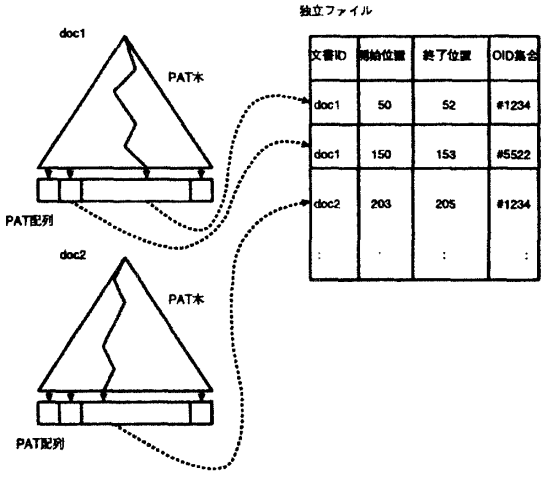

図2 PAT配列を利用した索引

2.2 相互参照機構の実装方法

以上のすべての組み合わせを検討した結果，Paratextデータベース実現の方法としてまず独立リンクファイルを持ち，文書とデータベースオブジェクトにはリンク情報を保持しない場合について 次の二通りを選択した．これらは既存のプレーンテキストと既存のデータベース間での相互参照機構構築に適用できる．

1. 表記層の索引としてPAT配列[GBYS92)]での実装（図2）
   文書中の文字列に対してPAT配列中の位置から，文書中の位置とともに独立ファイル中のリンク先のOIDの集合が得られる．逆に独立ファイルをOIDをキーとして探索すれば 参照層の索引となり，OIDに対してリンクされている表記層が得られる．

*A cross-reference mechaism between database objects and document context for integrated document databases.

†Hiroko KINUTANI, Masatoshi YOSHIKAWA, Hiroyuki KATO

‡Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

文字列	出現回数	文書ID	開始位置	OID集合
big bang	2	doc1 doc8	304 20	#1234,#2200 #1234
MOF	1	doc1	50	#2200
NNN	8	doc2 doc3	203 :	nil :

図 3 転置ファイルの拡張利用

2. 表記層の索引として転置ファイルの拡張利用での実装（図３）
   従来の転置ファイルを拡張し，リンク先 OID の指定のある文字列については無条件に転置ファイルに追加し，さらに指定されている OID 集合も保持する．逆にこの転置ファイルを OID をキーとして探索すれば 参照層の索引となり，OID に対してリンクされている表記層が得られる．索引構造は B+木で実装している．

これらは，容易に構造化文書に対しても利用可能であるが，さらに文書の構造を利用できる．例えば特定のタグ内に含まれる文字列についての正規表現による問合せが可能となる．また，文書型定義 (DTD) の論理的な木構造を利用した問合せ結果の表示ができる．しかし，表記層としてタグも含んでいるため，問合せ文字列に対し 格納されている文字列にはタグが挟まれている可能性もあり，タグに関する処理が必要となる．

3. 実験システム概要

統合型文書データベースへのユーザインタフェースには，データベースインタフェース (SQL 等) だけではなく，文書に対する従来の文書処理機能を利用するためのファイルインタフェース (エディタ等) も必要である．この条件を満たす統合型文書データベースの実験システムの構築をウィスコンシン大学で開発された SHORE(Scalable Heterogeneous Object Repository)[Wis96]上で行なっている．SHORE はオブジェクト指向データベースシステム技術（永続オブジェクト，抽象データ型，継承，トランザクション管理，同時実行制御，リカバリ等）とファイルシステム技術（階層型名前空間，ファイル，ディレクトリ，リンク等）の統合システムで，SHORE サーバが通信制御グループとして実行され，Paratext データベースは，SHORE のひとつのアプリケーションクライアントに位置する（図４）．

SHORE システムの基本クラスである text クラスのインスタンスとして文書ファイルを定義することによって，その文書ファイルは，SHORE データベース中

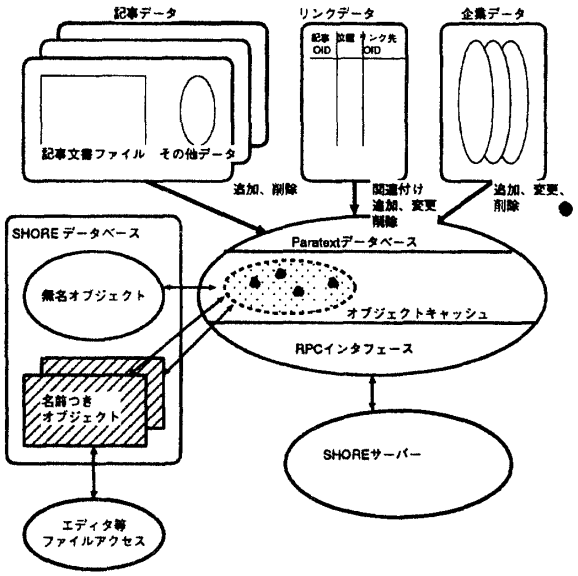

図 4 Paratext データベースのアーキテクチャ

の名前空間では，名前付きオブジェクトとなり UNIX のディレクトリ内のファイルと同様のアクセスが可能となる．Paratext データベースでは，Paratext クラスの属性としてこの text クラスを持つ．文書ファイルを含むデータクラスは、Paratext クラスのサブクラスとして定義するが，記事データクラスはそのひとつである．企業データクラスと記事の内容との相互参照は索引とリンクデータクラスを介して行なっている．

4. おわりに

本稿では，Paratext モデルにおける表記層と参照層の相互参照機構の構築とその実験システム Paratext データベース実装について述べた．今後は，構造化文書についても同様の実験システムを構築する予定である．

参 考 文 献

[GBYS92] G. H. Gonnet, et al., New indices for text: Pat trees and pat arrays. In W. B. Frakes et al., editors, *Information Retrieval - Data Structures & Algorithms* –, chapter 5, pp. 66–82. Prentice-Hall, 1992.

[Wis96] Univ. Wisconsin. Shore project home page. http://www.cs.wisc.edu/shore/, Mar 1996.

[YIU96] M. Yoshikawa, O. Ichikawa, and S. Uemura. Amalgamating sgml documents and databases. In *Proc. of the 5th International Conference on Extending Database Technology (EDBT'96), Lecture Notes in Computer Science, No. 1057, Springer-Verlag*, pp. 259–274, March 1996.

[加藤 97a] 加藤弘之, 吉川正俊, 絹谷弘子. A database system integrating structured documents and objects. 情報処理学会第 55 回全国大会, 5F-04, September 1997.

[加藤 97b] 加藤弘之, 吉川正俊, 植村俊亮. Querying structured documents with object links. 情報処理学会研究報告, 97-DBS-113. 情報処理学会, July 1997.