7R-2　リレーショナル・データベース・システムを用いた文書データベースの開発

桶谷猪久夫（帝国女子短期大学），渡辺豊英（名古屋大学・工学部）
北川善太郎（京都大学・法学部）

1．はじめに

データベースは情報活動の基盤として様々な分野で開発され利用されている。しかし、これらの多くは十分に整理、分類された特性データであり、オフィス文書、書籍などのテキスト文書はほとんど実現されていない。

現状ではファイリング・システムの下で画像データとして蓄積することが唯一の試みであった。文書データベースは既存データベース（たとえば、文献二次情報など）と比べて、必ずしも明確な枠組みがあるわけでなく、その性格を異にする。

本稿では、文書データベースを既存のデータベース・システムの枠組みに制約されずに、我々が書籍を利用する状況に基づいて規定し、リレーショナル・データベース・システムの下に実現する方式を検討する。我々のアプローチではリレーショナル・データベースは文書データのファイル格納構造として利用し、文書データベースはこの格納構造の上に仮想構造として実現する。

2．文書データベースの概要

文書構造の枠組みを、我々が利用する書籍の機能に従って構成する。[1)] 文書データベースは操作ビューとして、論理的構成と物理的構成を提供する。書籍などの文書は物理的レコードであるページと、また論理的なレコードである章、節などから構成され、文書データベースはこのような視野を操作しなければならない。

既存のデータベース管理システム、情報検索システムでは特性データから成る論理レコードだけを扱い、上記の要請を満たしていない。従って、文書データベースを既存のデータベース・システムの下に適切に開発することは容易ではない。文書データベースを既存のデータベース・システムの下に構築するには、ファイルの格納構造として利用する方法が有効である。すなわち、既存データベース・システム（リレーショナル・データベース・システム）の上に構築される文書データ管理システムは図1のような構成になる。

3．文書データベースのデータ表現

文書データベースをリレーショナル・データベース・システムの下で実現する場合、物理的構成と論理的構成を操作ビューとして実現しなければならない。このために、リレーショナル・データベースは文書データを物理イメージで格納し、論理的に階層的なデータ・セグメント間の関係を保持できるように構築する。

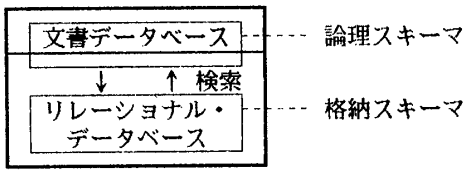

図1　文書データベース管理システムの構成

書籍の利用方法から文書データベースを規定すると、

(1)文書は物理的なページで構成され、ページは直接利用者が操作するオブジェクトである。

(2)文書には目次、索引という特定の文書内容に位置付ける機構が必要である。これらはどちらもページを指示する。

(1)は文書の物理的な操作対象を、また(2)は文書の論理的な操作対象を定めている。

リレーショナル・データベース上に文書データベースを写像する場合、リレーショナル・データベースのデータ表現の最小単位のカラムをどのように利用するかが問題であり、行を対応させた。すなわち、格納形式としてのカラムは行データ、行属性、及びその他の属性値で構成し、属性によって行を意味付けることにより、ボトム・アップに物理的構成、論理的構成を組み上げることを可能にする。

以下に、文書の格納形式として設計したテーブルのカラム属性とその意味をまとめたものを示す。

```
CREATE DATABASE LIS          ：データベース名
CREATE TABLE LIS.DOC         ：本文テーブル名
 ( TEXT    NCHAR(127),       ：本文の行
   TEXT-P  CHAR(8),          ：本文レベル指示
                               章・節などのレベルを格納
   CONT-P  CHAR(8),          ：目次レベル指示
                               章・節などのレベルを格納
                               するが、表題に関する部分
                               のみ有効値を設定。
   PAGE    INTEGER,          ：ページ指示
   ROW-NO  INTEGER,          ：各行のページ内行番号
```

Development of a document database on the relational database system
Ikuo OKETANI[1],Toyohide WATANABE[2],Zentaro KITAGAWA[3]
1:Teikoku Women's Junior College
2:Nagoya university, 3:Kyoto University

```
BLOCK-P   CHAR(1) )   ：図表指示
                      論理的に１つの単位として
                      扱う行を指定。[G]:図、
                      [T]:表、[L]:文字列域
```

このようなデータの表現の下に、各属性値により図２の対象オブジェクトとして構成可能となる。すなわち、応用プログラムによって、リレーショナル・データベースの格納形式をデータの集団が意味付けられた操作オブジェクトまで抽象化することができる。このような組のデータを作成することは、入力データに一定の規則を与えれば容易に自動化できる。

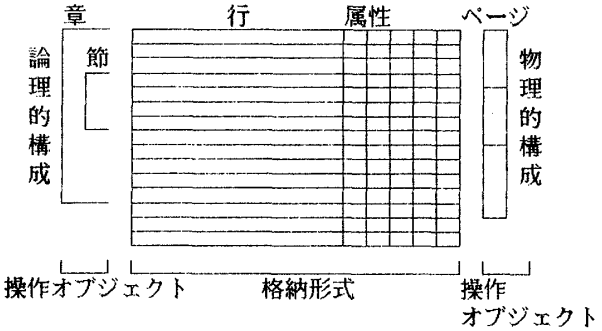

図２　格納形式と操作オブジェクト

以上のデータ表現により、目次検索、ページ通覧に対して実現可能であるが、索引検索に対しては従来のキーワード付与で対処する必要があり、別に設定するか、索引語自動抽出機能を開発するかなどの方法がある。現状は日本語の索引語自動抽出が困難なため、別にデータを用意する方法を採っている。もちろん、先のページ通覧機能の下で対話的に索引語を指示することにより、半自動化して索引テーブルを作成する。以下に、索引テーブルのカラム属性とその意味を示す。

```
CREATE TABLE LIS.IND    ：索引テーブル名
( WORD    NCHAR(20),    ：キーワードとして抽出した語
  TEXT-P  CHAR(8),      ：本文表の本文レベル指示
  PAGE    INTEGER )     ：本文表のページ指示に対応
```

４．文書データベースの実現例

リレーショナル・データベースの下に格納したデータを文書データベースとして実現するには、応用プログラムで文書操作の機能を作成する必要がある。少なくとも、既存データベース・システムが提供する操作コマンドは、文書操作に対しては低レベルの機能である。

文書操作コマンドは基本的には目次検索、索引検索、通覧であり、現在表１に示すコマンドを用意している。基本的にはリレーショナル・データベース操作コマンドの機能に準じて作成される。当然、検索条件式（比較演算子、論理演算子など）をサポートしている。これらのコマンドは応用プログラムによって、リレーショナル・データベースの操作コマンドを制御して機能する。たとえば、通覧機能では検索で位置付けられたページに対して、前後のページを連結する制御が必要である。

表１　文書操作コマンド

SELECT PAGE nnn	該当ページを検索し表示
SELECT CONTENTS n…n	目次の検索　nn:12,第一章第二節の表題の検索　nn:12%, 第一章第二節とそれ以降の検索
SELECT KW ” 文字列”	索引検索
SELECT TEXT ” 文字列”	内容検索
BEFORE,NEXT	前後のページの連結
SELECT … OUT[n]	省略：ページ単位表示 OUT1: ページ表示, OUT2: ページ、目次表示, OUT3: ページ、目次、キーワード一覧表示

（注）省略形はコマンド、オペランドの前の３文字

文書操作コマンドとリレーショナル・データベースの操作コマンドとの関係については以下のようになる。

（例）内容（文字列）検索し表示する。

〈文書操作コマンド〉

```
SELECT TEXT '担保物件'
```

〈リレーショナル・データベース操作コマンド〉

```
SELECT * FROM DOC WHERE TEXT '担保物件'
    ↓
SELECT CONT-P,TEXT FROM DOC
       WHERE PAGE-C = [    ]
```

（注）[　　]：前の検索結果のページ指示を使用。

（注）複数発行されるかも知れない。また文書操作コマンドBEFORE,AFTERも可。

５．おわりに

本稿では、文書データベースを既存のリレーショナル・データベース・システムの上に実現したので、その概要を示した。必ずしも十分に実現できたわけではないが、文書データベースが有する特徴にだけ注目して検討した。操作機能として更新処理については検討していないが、今回の試みの枠組みで実現するよりも、文書データベース・システムのアーキテクチャの下で検討することが重要と思われる。また、文書の活用は単に文書データベース・システムとして位置付けられるのではなく、分散環境の情報システム全体で検討する必要があり、これも今後の課題である。

現在、我々は京都大学大型計算機センターのAIM/RDBの下に本稿で述べた文書データベースを開発している。

謝辞　日頃からご教授・ご鞭撻をいただいている帝国女子短期大学・佐野四郎教授に感謝します。また、富士通京都支店・今井恒雄氏、岡本匡人氏に感謝します。

参考文献

(1)渡辺他：「文書データベースの枠組みに関する考察」
情報処理学会第３８回全国大会講演論文集