

文書データベースの枠組みに関する考察

7R-1

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

1. はじめに

大容量、高密度のファイル媒体の開発は、膨大なデータを計
算機処理可能な形態で蓄積することを可能にした。情報化社会
と呼ばれる今日、種々様々なデータや膨大なデータを効率良く
管理し、必要に応じて迅速・正確に利用できることが要求され
ている。データベースはその基盤であり、様々な分野で活用さ
れている。しかし、今日構築されたデータベースの多くは個々
に分類・整理された論理的な特性データであり、物理的な構成
によって制約された文書を直接の対象としていない。情報活動
にとって、文書は必要不可欠であり、その枠組みは少なからず
現存のデータベース処理、情報検索とは異なる。

本稿では、書籍などの文書をデータベース化するための枠組
みを考察する。考察の視点は今日のデータベースの枠組みで扱
い得る文献二次情報と対比させて、その性質を言及し、文書デ
ータベースの構造・機能を検討することである。

2. 文書の情報構造

文書に関するデータでは、文書の特定情報である文献二次情
報と、文書内容そのものである文献一次情報が、データベース
の対象となる。今日のデータベース処理、情報検索では個々に
分類された特性データの文献二次情報を対象としているが、非
分類のデータ列である文献一次情報を対処していない。文献二
次情報は図書館によって利用法・構造が明らかにされたが、
文献一次情報は未だ十分に検討されていない。我々が書籍など
を利用する場合、現存のデータベース処理、情報検索の機能に
必ずしも一致していない。表1に、図書館などの文献二次情報
と書籍などの文献一次情報の特徴を整理した。文献一次情報は
構築されていないために、構築・利用に対しても、今日の
データベース管理システム、情報検索システムの枠組みに適合
していない。今日のデータベースは特性データの扱いに重点を
置いている。文献一次情報のデータベース化には、我々が書籍
などを利用する様を実現する必要がある。

特性データを基本とするデータベースでは、論理的なレコー
ド、データ項目の下に構成されるが、書籍では物理的なページ
が基本となる。目次や索引などの指示対象はページであり、文
献二次情報におけるキーワード(アクセス・ポイント)の指示対
象のレコードとは異なる。図1にこの機構を表す。目次、索引
は文献二次情報の逆ファイルと同様に機能しているが、ページ
は論理的に独立なデータ単位ではないので、論理レコードを基
本とする文献二次情報の枠組みでは操作対象とはならない。ペ
ージを直接操作できる機構が必要となる。すなわち、文献一
次情報では検索機能と通覧機能が必要である。

文献一次情報は章、節、項などの論理的な単位で参照される
が、実際のアクセスはページという物理的な単位で操作される。
従って、データベースとして二重構造を実現できなければならない。図2に、文献一次情報のデータベース構造を示した。検
索機能は論理的な構成と物理的な構成に、また通覧機能は物理

表1 文献二次情報と文献一次情報

Table with 3 columns: 文献二次情報, 文献一次情報, and comparison rows for data units, relationships, roles, search methods, search procedures, and other features.

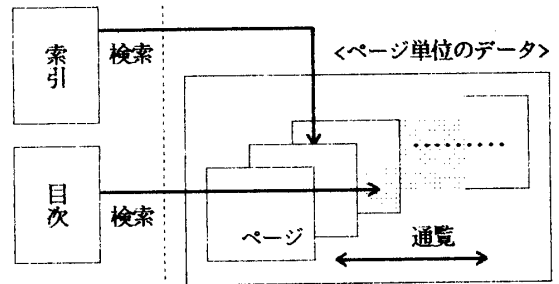


図1 文献一次情報へのアクセス機構

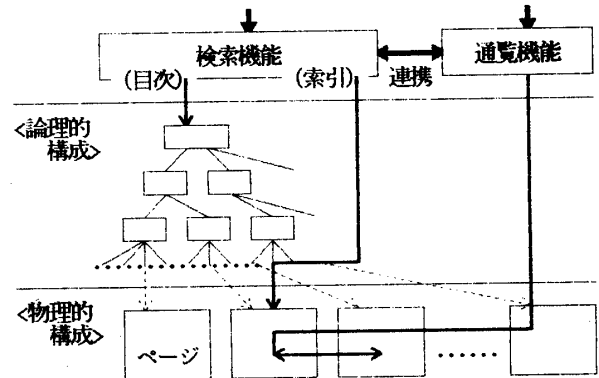


図2 文書データベースの構造

的な構成に作用できなければならない。もちろん、検索機能と通覧機能は同一の操作ビューの下で実現され、連携している必要がある。

3. 文書データベースの構造

文書データベースでは、従来のデータベースにおける物理構造と論理構造の視点とは異なった立場で、文書データの物理的構成と論理的構成を実現しなければならない。物理構造と論理構造の関係では、利用者の操作対象としての論理構造、システムの操作対象としての物理構造の枠組みが実現されてきたが、文書データベースでは物理的構成と論理的構成は利用者の視野であり、操作対象である。もちろん、従来の枠組みでは物理的構成と論理的構成は論理構造に属する。

文書データベースを規定するために以下の条件を設定する。

〔条件1〕 文書データベースは物理的構成と論理的構成の利用者ビューからデータを操作できなければならない。

〔条件2〕 物理的構成は物理的に制約された大きさのページを単位とする。このページは相互に順序関係を有している。すなわち、ページは順序関係を有した物理的なレコードである。

〔条件3〕 論理的構成は概念的に一つの纏まったデータの論理的なレコードであり、構造的に各論理的なレコードを捉えることができる。

〔条件4〕 物理的構成と論理的構成は一つのデータに対して異なった操作ビューを実現する。従って、総てのデータ操作に対して補完的な関係にある。

以上の条件により、ページで構成された物理的構成に、論理的構成の構造化視野が形成され、利用者は論理的に関係を有したデータの集合として、また物理的に定形化されたデータの固まりとしてデータを操作できる。従って、検索時にはこれらの構成を介した多様な操作が可能であるが、更新時には論理的な変更操作と物理的な変更操作が相互に作用しなければならない。そのためには、論理的構成と物理的構成の対応関係が必要であり、従来の概念に従えば論理的構成が論理構造に、物理的構成が格納構造に対応すると思われる。

条件1から条件5に従うと、文書を画像データとして格納し、それにコード・データのガイド情報を付帯させたファイリング・システムは文書データベースではない。このシステムは物理的構成だけを提供している。文書データベースの枠組みをより現実的にするために、ページに対して行の概念を導入する。

〔条件5〕 ページは一つ以上の行から構成される。行は物理的な概念であり、論理的構成に直接関与しない。

条件5は便宜的であり、文書データベースを現状の枠組みの下でより実現し易くするが、文書データベースの範囲をより狭くする。たとえば、一般に書籍などで頁に図・表が挿入された文書は対象外となる。すなわち、行の下にページ内の総てのデータが指示可能でなければならないので、旧来のライン・プリンタで作成可能な文書様式のみを定義している。条件5の下で、画像形態の文書を扱う場合には、データ属性(バイト単位、ビット単位、ベクトル単位など)によって別々の部分文書として指示しなければならない。文書データベースのスキーマ定義を概略図3のように考え得る。

4. 文書データベースの機能

文書データベースを物理的構成と論理的構成から規定したが、データ操作もこれらに有効に機能しなければならない。ここでは、検索時の機能を検討する。文書(文献一次情報)は第2節で述べたように、目次・索引検索機能の下に通覧機能が必要である。我々の枠組みでは、通覧機能は物理的構成によって、目次検索機能は論理的構成によって実現される。また、索引検索機

```
define NAME-1 ;
  segment NAME-2 ;
    terminal SYMBOL-1 ;
    space SYMBOL-2 ;
    page NUMBER-1 ;
    line NUMBER-2 ;
    key NOTATION-1 ;
    type TYPE-1 ;
  segment NAME-3 ;
  .....
end ;
```

図3 文書データベースのスキーマ定義例

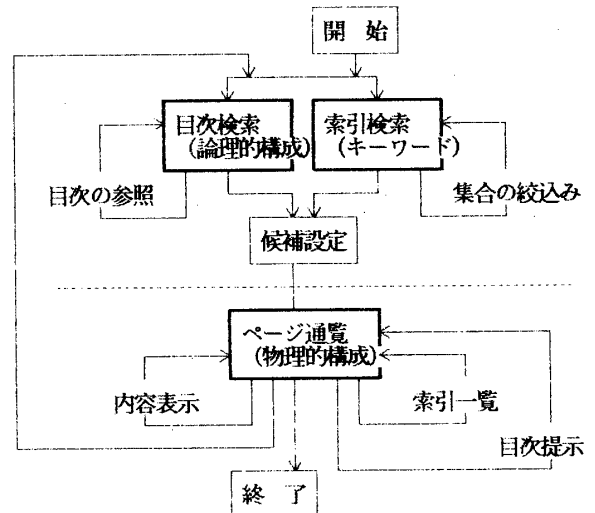


図4 文書データベースの機能

能は従来のキーワード検索機構で実現可能である。

目次データを文書と別に用意すれば、目次検索は容易に実現されるが、本文に含まれる章、節などの表題を利用すれば文書の内容検索となる。すなわち、目次検索では章、節などのデータ・セグメント(論理的なレコード)のキー項目として表題データを構成するように、論理的構成を操作する必要がある。一方、通覧は目次検索・索引検索の下に位置付けられたページに対して参照域を定め、参照制御する。この場合、単にページ指示だけでなく、参照域(物理的なページ、論理的なレコード)の目次や索引(キーワード)の表示も必要である。我々が文書を利用するときの自由度が必要である。たとえば、図4のような状態遷移が可能でなければならない。

5. おわりに

本稿では、文書データベースの構造とその機能について考察したが、より詳しく検討しなければならない機構も多い。しかし、本稿における枠組みだけでも、特性データの扱いを主体とした現状のデータベース処理、情報検索の機構では充分に対処できない。より適切な枠組みの下に文書データベース・システムが開発されることが望まれる。更新機能とその制御法は本稿では触れられていないが、エディタがテキストを管理するような機構の導入が必要となる。しかし、それだけでは不十分である。また、本稿では条件5を加えて述べたが、条件5を排除した場合の検討が必要であり、物理的構成をより仮想化して捉える必要がある。これらは今後の検討課題である。

謝辞 日頃からご教授・ご鞭撻いただいている名古屋大学工学部・吉田雄二教授、稲垣康善教授、鳥脇純一郎教授、及び中京大学・福村晃夫教授に感謝します。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。