人文科学とコンピュータ　8－5
（1991. 3. 8）

文書データベースの設計と構築

桶谷猪久夫
帝国女子短期大学

既存のデータベース処理、情報検索システムは分類された特性データを対象とするが、本システムは書籍などの文書（テキスト情報）のような特別に加工されていない一次情報を対象とする。

我々は文書データの情報構造に着目し、その物理的な操作対象であるページという物理的な構成と、章、節、項などの論理的な構成を利用者の操作対象とする。文書データベースの機能は、目次検索機能、通覧機能と索引機能からなり、特に効率のよい検索にとって重要な索引機能について述べる。キーワードの抽出において、不要語除去方式を使用した自動キーワード抽出方式と形態素解析の有効性について報告する。

Design and Development of the Document Database

Ikuo OKETANI
Teikoku Women's Junior College

Existing database and information retrieval systems in general are built upon classified data, but this system handles the primary information in the form of documents such as books (text information).

We examine the information structure of the document data, and have two types of operational clues in the document database, which are (1)the physical composition, such as the individual pages which separate the document physically and (2)the logical composition, such as chapters and paragraphs.

The functions of the document database consist of content retrieval, browsing and indexing functions. We describe in detail the indexing function required for effective retrieval.

We describe two automatic keyword extraction methods; one is by morpheme analysis and the other is by the tool based on stop-word removal method.

1．はじめに

近年の半導体技術を中心にしたハードウェア技術の進歩は、安価で高性能なマイクロプロセッサを組み込んだワードプロセッサやパーソナルコンピュータなどのＯＡ機器を急速に普及させている。また、これらはその利用の簡便さから情報発生場所に設置され、日常的に発生するデータを計算機処理可能な形態で蓄積している。

情報化社会といわれる現在、これら種々の種類・形態で蓄積されたデータを効率よく管理し、有効な情報を迅速・正確に活用できる環境を提供することは、最も基本的な要求になってきている。

データベースはその基盤であり、種々な分野で活用されている。しかし、今日利用可能なデーターベースの多くは、個々の項目の下に分類・整理され意味付けられたデータを対象としている。しかし、オフィス文書、書籍などのように複合データを持ちデータ構造が定め難く、また物理的な構造によって制約された文書を直接の対象としていない。

最近、大容量、高密度の記憶装置の開発により情報発生場所で入力され、格納された文書ファイルを汎用コンピュータに転送し、集中管理を行う電子ファイリングシステムがある。しかし、電子ファイリングシステムは文書単位での保管、検索が中心である。

文書データベースは、目的・用途に合わせ特定情報（文献二次情報）を抽出・加工して操作対象とするように作成される文献情報検索と異なり、文書（テキスト情報）が加工されずに文献一次情報をそのまま対象とする。

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具体的に検討し、それに基づいて設計・構築した文章データベースの各機能について、特に効率的な検索にとって重要な索引機能について述べる。ここで示す文書データベースは民法要覧であり、民法に関する講義テキストで、体系的にまとめられたものである。

2．文書データベースの概要

2－1　文書データベースの構造

文書に関するデータでは、文書内容そのものである文献一次情報とデータを解析し簡潔な形で表現し直した文献二次情報がデータベースの対象となる。文献検索システムにおいては、直接対象とされるデータは入力データそのものではなく、表題名、著者名、主題等の文献を特徴づける属性を設定し、これに各文献ごとの値を対応させた形で扱い、質問に応じる際は、必要な属性の値についてのみ合致具合いを判断していくいう方法を採用している[1]。

文書データベースは文章（テキスト情報）が何ら加工されない文献一次情報として直接操作の対象とするため、現在のデータベース処理・情報検索と異なり、その枠組みは明確に存在しない。

そのため文書データベースでは文書構造の枠組みを、我々が書籍などの文書を利用する状況に基づいて規定し実現する[2][3]。

書籍などの文書構造は、物理的構造として紙面からの制約であるページの集まりとして捉えることができ、一方、論理的構造として章、節、項、段落などの論理的な単位から構成される。これらの論理的構成は階層的な構造によって表現される。さらに、我々は目次や索引を介して物理的な操作単位であるページを参照する。目次、索引は文献二次情報の転置ファイルに相当するが、ページという概念はそれ自体が物理的制約からのものであり、論理的に独立なデータ単位でないので、文献二次情報の枠組みでは操作対象にはならない。しかし、文書データベースにおいては、章、節、項などの論理的な単位で参照されるが、実際のアクセスは物理的な操作単位であるページを介して行われるように機能しなければならない。従って文書データベースは、論理的な構成と物理的な構成に作用する検索機能と物理的構成に作用する通覧機能を備え、これらは連携して機能する必要がある。

書籍の利用方法から文書データベースを規定すると、

(1) 文書データベースは物理的な構成と論理的な構成の操作ビューの下に実現され、それらは連携して作用する。
(2) 物理的構成は順序関係を有した一定の大きさのページの集まりで構成され、ページは直接利用者が操作するオブジェクトである。
(3) 論理的構成は文書の階層構造により定められる。
(4) 文書データベースは目次、索引などの検索機能の下に該当ページを位置付け、かつそのページ域を参照する機能を提供する。[4]

既存データベースに比べて、物理的操作単位であるページの下に検索・参照機能を実現しなければならない。この構造によって実現される文書データベースの

構造を図１に示す。

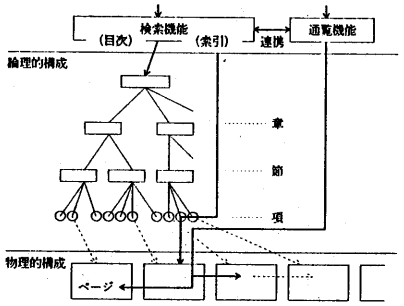

図１　文書データベースの構造

２－２　「民法要覧」の概要と文書構造

データベース化した文書は、図２で示す「民法要覧」で、民法に関する要点をまとめた講義テキストであり、債権法、民法総則、物権法、親族法・相続法の４文冊からなる。各々は書籍として200～300ページに相当する[5]。これらの文書は、研究者、研究補助者により手元のパーソナルコンピュータを使用しフルテキストとして入力、蓄積された。文書の情報構造は、講義テキストとしての性質から前節で述べた論理的な構造としての章、節、項などが整然と形成されている特徴がある。

現在、法律情報システムの対象となる法律情報については、これまで法令と判例に関する情報が中心として取り扱われ、今日まで数多くのデータベースが構築されてきた。しかし、それらは文書データベース（文献一次情報）の枠組みではなく、情報検索システム（文献二次情報）であった。法律情報システムの多様な役割を考えると、より広範な種々の情報をその範囲として把握することが必要である[6]。

法律情報がデータベース化され有効に活用された場合の利用用途は、

(1) 法律事務：法令・判例情報、著書・論文の検索
(2) 法学研究：法学研究・著作活動
(3) 法学教育：研究、法律問題の論理的構造を習得

の３つの領域で機能することが要求される。また、一般的に法律情報の正確な理解、解決の多様性に対応するためには、該当箇所のみでなくその前後関係、さらに全文の参照を必要とする。ここに述べる文書データベースの枠組みと機能は、これらの要求に少なからず応えることができる。

1.　第一部　債権序論
1.1　債権法の概要
1.11　民法典の構成
1.111　概要
「第一章　総則」「第二章　契約」「第三章　事務管理」「第四章　不当利得」「第五章　不法行為」。
第二章以下は、債権の発生原因（→　4．参照）であり、第一章は、各発生原因から生じた債権の総則である。第一章は、そのほとんどが「第二章　契約」に関連している。
1.112　債権総則（民３９９～５２０）
債権の発生から消滅にいたる法律問題がその対象であるが、必ずしもそのすべてを規定していない。たとえば、債権の発生一般については、とくに規定がない。個々的には、債権の発生原因について、第二章以下が詳細に定めている。
「債権の目的」は、債権者が債務者より要求しうる給付（例えば、物の引渡、金銭の支払）について定め、「債権の消滅」は、契約でいえば契約内容が債務の履行・弁済により実現され、債権が消滅していくまでの、消滅の過程と消滅の結果にかかわる諸問題を扱う。
「債権の効力」には性質の異なるものが入っているが、債務の不履行がその主要問題である。また、債権が例外的に債務者以外の者にもその効力を及ぼす場合があり（債権者取消権・債権者代位権）、これらも「債権の効力」で規定されている。
「多数当事者の債権関係」では、複数の債務者や保証人の存在によって債権内容の実現が担保される諸関係を対象としている。いわゆる人的担保と称されている不可分債務、連帯債務、保証債務がそうである。

図２　民法要覧の文書例

３．文書データデースの設計と機能

３－１　文書データベース・システム

既存のデータベース管理システム、情報検索システムは、特性データを基本とし論理的なレコード、データ項目の下に構築される。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。従って、実現法として既存のデータベース・システムを低位のデータ管理システムとして利用し、その上に文書データベース・システムを機能的に構成する。既存のデータベースとして、簡便で論理的なインタフェースを備え、柔軟なデータ構造表現を許し、拡張性に富むリレーショナル・データベースを利用した。特に、応用プログラムからのデータベース・アクセスのインタフェースを備えていることが、既存データベースを格納構造として利用させ、文書データベースを各種機能の下に仮想構造として実現させることを容易にする。

３－２　データ表現

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、どのように規定し利用するかが問題になる。これを解決するために、カラムを文書の行に対応させた。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され、各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。

文書データベースにおいては、その文書の内容を最も的確に表現する単語を検索情報として、効率的な検索を行う索引検索機能を必要とする。本システムは索引検索機能として従来のキーワード付与で対処し、文書データベースの構築の章で述べる各種の方法で索引用ファイルを作成する。

図3に、文書の格納形式として設計した本文テーブル、索引テーブルのデータベース定義における各々のカラム属性とその意味を示す。

本文テーブルのカラム属性

```
CREATE  DATABASE   KLIS1          ：データベース名（債権法）
CREATE  TABLE   KLIS1.DOC         ：本文テーブル名
(TEXT   VARCHAR(84) NOT NULL,     ：本文の行データ
 TEXTP  CHAR(12),                 ：本文のレベル指示：章・節
                                    などのレベル属性を格納
 CONTP  CHAR(12),                 ：目次レベル指示
          ：表題に関する部分のみ章・節などのレベルを格納
 PAGE   INTEGER,                  ：ページ指示
 PAGE   INTEGER,                  ：ページ指示
 ROW    INTEGER,                  ：各行のページ内番号
                                    表、画像：バイト単位（数）
 BLOCK  CHAR(3)  )                ：図表指示
     ：論理的に１つの単位として扱う行を指定。［Gnn］：図、
     nnは図の範囲（行数），［Tnn］：表，［Lnn］：文字列域
```

索引テーブルのカラム属性

```
CREATE  TABLE   KLIS1.IND         ：索引テーブル名（債権法）
(WORD   VARCHAR(22) NOT NULL,     ：キーワードとして抽出した語
 KWORD  VARCHAR(40) NOT NULL,     ：抽出語のカタカナ表示
 TIND   CHAR(12),                 ：本文表の本文レベル指示
 PIND   INTEGER,                  ：本文表のページ指示に対応
 PNOT   CHAR(2)  )                ：本文表に現れないキーワード
                  指示、研究者が指示、または参照画面で入力
```

図3　カラム属性とその意味

3－2　文書データベースの機能

文書データベースは格納構造としてリレーショナル・データベースを利用し、その上に仮想構造として実現され、各機能は格納されたデータとその属性情報により、利用者は有効な文書操作を提供する。つまり、前述の本文テーブルに格納された行データを制御用データのカラムの属性値により、応用プログラムはページを単位とする物理的構成と、章、節、項などを階層構造の要素とする論理的構成として組み上げ、利用者に各種操作機能として提供する。本文テーブルから目次検索機能、通覧機能を索引テーブルから索引検索機能、通覧機能が実現する。既存データベースの格納構造を利用し、その上に仮想構造としての文書データベースを構築する時、その各機能、特に文書操作コマンドを応用プログラムで実現する必要がある。

表1　文書データベース操作コマンド

文書操作コマンド	機　能　概　要
SELECT PAGE nnnn	利用者が明示的にページ指定を行い、その該当（候補）ページを検索・表示
SELECT CONTENTS n‥‥n SELECT CONTENTS 文字列	目次の検索 nn ：12，第１章第２節の表題の検索 nn ：12%，第１章第２節とそれ以降の検索 文字列：目次の内容検索
SELECT KW　文字列 SELECT KW　@ローマ字	索引検索　：文字列部分一致、 ローマ字カタカナ変換（@印）
SELECT TEXT 文字列	内容検索　：文書データの文字列部分一致
BEFORE　n ,　AFTER　n	前後ページの連結　n :ページ数、省略時は１
NEXT　n	次該当（候補）ページへの連結
INSERT テーブル指定 （カラムの並び） VALUE（定数の並び） UPDATE , DELETE	INSERT:テーブルに指定された行を挿入する UPDATE,DELETE ：補助的使用（運用管理者） （注）現バージョンでは、テーブル指定は索引テーブルのみ使用可
SELECT　‥‥　OUT[n] ‥‥　AND ‥‥　OR	OUT省略：最初の該当（候補）ページ表示 OUT1　：該当（候補）ページ一覧の表示 OUT2　：該当（候補）ページ一覧とページ毎の目次表示 OUT3　：該当（候補）ページ一覧と目次、関連キーワード一覧の表示 AND,OR ：条件式（論理演算子）

文書データベース・システムが提供する文書操作コマンドの概要を表１に示す。

各コマンドは検索効率や汎用性を考慮しリレーショナル・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、比較条件式（比較、論理演算子）は集合の絞り込みのためサポートされる。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と文書データへのアクセス機構を図４に示す。

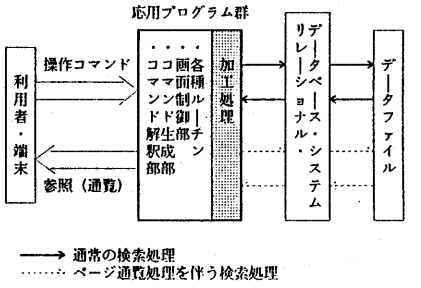

図4　文書データベースへのアクセス機構

文書操作コマンドは応用プログラムによって解釈され該当するリレーショナル操作コマンドに変換され発行される。当然この関係は１対１ではなく要求に応じて加工されたり、複数のリレーショナル操作コマンドとして生成される。つまり、検索は文書操作コマンドとして要求され、応用プログラムのコマンド解釈部で

リレーショナル・データベース操作コマンドとして変換され、データ構造である物理レコードに対し行われる。その結果は応用プログラム内の指定領域に渡され、参照域としてのページを再構成し通覧機能を構築する。

通覧機能は目次検索・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る。さらに、前後のページを連結する機能と次該当（候補）参照域への連結機能が必要である。

さらに、我々が文書を利用するときの自由度が必要であり、該当（候補）ページ指示の表示だけでなく、ページ毎の目次と関連キーワード一覧表示が行えるサブコマンドを用意し、利用者が参照する物理的なページの絞り込みを支援する。文書データベースの目次検索、索引検索、通覧機能の概要を図５に示す。

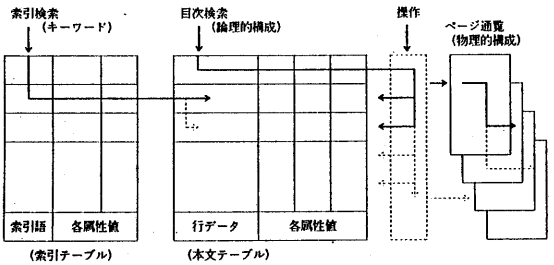

図５　文書データベースの検索機能

４．文書データベースの構築

４－１　構築概要

（１）文書データの入力

（２）文書データの変換・編集

（３）文書ファイルの転送

（４）文書データの構造化

（５）キーワード抽出

ここでの自動抽出方式については、索引ファイルの構築の節で詳しく述べる。

（６）データベースの作成

（７）応用プログラムの設計と開発

今後の発展を考慮し、モジュール構造として設計・開発した。検索・通覧機能については、前に述べたのでそれ以外の機構・機能について以下に述べる。

(1) コマンド解釈・生成部

文書操作コマンドを解釈し、対応するリレーショナル・データベース操作コマンドに変換する。文書データは日本語を主体に英数字・記号が混在する複合データから構成されている。現状では、汎用計算機と接続されるターミナルの全てが日本語入力機能を備えていなく、またその機能も十分ではない。そのため、索引機能における索引データのカラムは漢字属性とカタカナ属性を持っている。応用プログラムはローマ字カタカナ変換ルーチンを備え、ローマ字入力を許している。

(2) ユーザ・インタフェースの強化

●画面制御部

通覧機能は物理的構成であるページを単位として参照・表示を行い、またページ指示の表示のみでなく、目次や関連するキーワード一覧表示を行う機能を備えている。また、要求に応じて文書操作コマンドが入力される必要がある。そのため、ディスプレイ装置はデータ操作の結果を利用者に適合した形で出力・表示するマンマシン・インタフェースを備えていることが要求される。画面制御部はこれらの機能を、フルスクリーン型ディスプレイ端末で達成する。応用プログラムは画面単位（ページ単位）の入出力操作を行い、画面を表示域と文書操作コマンド入力域に分割し管理する。これらの機能が使用できるかどうかは、応用プログラムの初期化時に端末の種別・特性を調べるサポートプログラムが発行され、応用プログラム内の端末属性テーブルと照合することによって決定される。

●メニュー方式による文書データベースの選択

現在、下記のように４つの文書データベースが構築され、要求する文書データベースがメニュー方式で選択できるようになっている。

```
WHICH  DATABASE  (1,2,3,4)
    1 -------  債権法
    2 -------  民法総則
    3 -------  物権法
    4 -------  親族・相続法
```

●コマンドヘルプ機能のサポート

４－２　索引用ファイルの構築

書籍などのテキスト情報は特別に加工されずに、そのまま文献１次情報の形式で格納されている。テキスト情報は意味内容検索的な機能を要求されるが、現時点では本システムは文献内容の検索に対してキーワード付与で対処している。

キーワードは文献の内容を的確に表現し、また他と明確に識別される必要がある。たとえば、債権法における「債権」というキーワードは文献の内容を表現しているが、この場合、通覧機能における物理的構成で

あるほとんどのページに該当し、識別の能力をほとんど発揮しない。適切なキーワード抽出は、データベースの効率的な検討にとって重要な課題であり、またその作業は文献数や量（ページ数）の増加により膨大な労力を必要とする。

以下、当システム構築において採用したキーワード自動（半自動）方式について述べる。

（1）不要語除去方式によるキーワード自動抽出

日本語を中心としたテキスト情報を入力し、品詞単位に空白を挿入した分かち書き文を作成する分かち書き処理を行い、その中から不要語削除処理を行いキーワードの対象となる単語を抽出する。

これらの処理には各種辞書として語彙辞書、分かち書き辞書、不要語辞書が使用される。

ここでのキーワード抽出の方法として、開発の行程を考慮し、専門用語辞書、不要語辞書の構築は行わずに基本辞書と分かち書き辞書の下に、パラグラフ単位でのキーワード抽出を行った。組合せキーワードに対しては、最小語基として1を最大語基を3に設定した。たとえば、「損害賠償責任」は、「損害」、「賠償」、「責任」、「損害賠償」、「賠償責任」の6個のキーワード候補を抽出する。また、検索効率と入力の容易さを考えカタカナ読みの抽出と、パラグラフ内の抽出キーワード頻度を通知する。

キーワード頻度の通知は、語の出現頻度による適切なキーワード抽出の技法を利用するためである。文献中に高頻度で、または低頻度に出現する語は、文献内容を識別するキーワードとして向かないという法則（Zipfの第1、第2法則）を適用する。たとえば、高頻度に出現する「債権」はキーワードとして適さないことは前にも述べた。このキーワード頻度の通知は、機械的処理に向いており文献全体の統計的解析から、より適切なキーワード選択の指標になる。

抽出されたキーワード候補の例を図6に示す。

文献「債権法」は、パラグラフ数2,756から33,235個のキーワードの対象となる単語が抽出された。キーワード自動抽出の評価指標は、

適合率＝（抽出キーワード中の適切なキーワード数）
／（抽出したキーワード数）

再現率＝（抽出キーワード中の適切なキーワード数）
／（適切なキーワード数）

で一般的に求められる。現在、研究者が検討中である

債務不履行を理由とする損害賠償請求権や契約解除による現状回復請求権は契約効果でなしに法定効果である。

168*1 債務 不履行 を 理由 と する 損害 賠償 請求権 や 契約 解除 に よる 現状 回復 請求権 は 契約 効果 で なし に 法定 効果 で ある 。 *2

------- (20)*3 ----------	----------
1610 -- (1) 債務 *4	ｻｲﾑ *5
1611 -- (1) 債務不履行	ｻｲﾑﾌﾘｺｳ
1612 -- (1) 不履行	ﾌﾘｺｳ
1613 -- (1) 理由	ﾘﾕｳ
1614 -- (1) 損害	ｿﾝｶﾞｲ
1615 -- (1) 損害賠償	ｿﾝｶﾞｲﾊﾞｲｼｮｳ
1616 -- (1) 損害賠償請求権	ｿﾝｶﾞｲﾊﾞｲｼｮｳｾｲｷｭｳｹﾝ
1617 -- (1) 賠償	ﾊﾞｲｼｮｳ
1618 -- (1) 賠償請求権	ﾊﾞｲｼｮｳｾｲｷｭｳｹﾝ
1619 -- (2) 請求権	ｾｲｷｭｳｹﾝ
1620 -- (2) 契約	ｹｲﾔｸ
1621 -- (1) 契約解除	ｹｲﾔｸｶｲｼﾞｮ
1622 -- (1) 解除	ｶｲｼﾞｮ
・	
・	
・	

（備考）*1：パラグラフ番号　*2：分かち書き文　*3：抽出キーワード数
*4：抽出キーワード　*5：抽出キーワードのカタカナ読み
：左端の数字はキーワードの合計数、括弧内の数字は出現頻度

図6　抽出キーワードの例

が、基本辞書のみを使用している、特別な不要語処理を行っていない、専門辞書が完備されていないなどの理由で、再現率はともかく適合率は低いと思われる。

索引用ファイルの構築には、簡単なプログラムで抽出したキーワード候補を画面に表示し、選択指示（登録／廃棄）をする。

（2）形態素解析を利用したキーワード抽出

入力された文章が、どのような形態素（文を構成する意味を持つ最小言語単位）から構成されているかを明らかにする処理が形態素解析である。

英語の文などは、単語と単語の間に空白を挿入し分かち書きされる。日本語の文は、膠着語であり単語単位で分かち書きされないため、形態素解析は容易でない。しかし、現在、単語と単語の間に空白を挿入しない膠着語の形態素解析が多く研究されている[7]。

手元のパーソナルコンピュータ（PC-9800シリーズ）を利用し、キーワード候補抽出のための形態素解析を自動的、短時間に行えればその効果はきわめて大きい。ここでは、荻野綱男氏の作成したカナ漢字変換用辞書を用いた日本語の形態素解析プログラムFIXSEGを利用した[8]。パーソナルコンピュータPC-9800VX21、ハードディスクベースで、文献「債権法」パラグラフ数2,756は60～70分で解析される。解析結果を図7に示す。

形態素解析の解析結果は、今後の言語情報を利用したキーワード自動抽出に利用できる可能性がある。

```
*
* 情報レベル＝詳細
*
入力＝債権の発生から消滅にいたる法律問題がその対象であるが、必ずしもそのすべてを規定していない。.
債権／名詞①　の／連体格助詞　発生／サ変名詞　から／格助詞　消滅／サ変名詞　に／格助詞　いた／ラ行五段①　る／ラ５終止連体　法律／名詞①　問題／名詞①　が／格助詞がを　その／連体詞②　対象／名詞①　で／断定助動連用　あ／補助動詞「ある」　る／ラ５終止連体　が／終止接続助詞　、／記号類　必ずしも／副詞②　その／連体詞②　すべて／名詞①　を／格助詞がを　規定／サ変名詞　し／サ変未然連用　て／連用接続助詞　い／「いる・みる」　な／助動ない　い／形容詞終止連体　／・
経過時間（秒）＝2
入力＝たとえば、債権の発生一般については、とくに規定がない。
たとえば／副詞④　、／記号類　債権／名詞①　の／連体格助詞　発生／サ変名詞　一般／名詞①　に／格助詞　ついては／副詞④　、／記号類
とくに／副詞②　規定／サ変名詞　が／格助詞がを
 解 no.1
な／形容詞③　い／形容詞終止連体
 解 no.2
な／ワ行五段①　い／ワ５連用１
```

図７　形態素解析の出力例（詳細形式）

解析結果はMS-DOSファイルの形式で保存され、ある一定の規則を与えたプログラムでキーワードとして作成される。たとえば、形態素解析結果から機能語（助詞、助動詞、区切り記号など）を除いた概ね名詞のみを、キーワードとする。さらに、名詞が連続している場合、それらを繋ぎ合わせて（語基数指定）組合せキーワードとして作成する。このようなキーワード候補の作成は簡単なプログラムで規則的に自動化され、そう難しいことではない。

（３）対話型キーワード指示

キーワード自動抽出方式は、あくまでも文書（文献）を構成する文字列の中に、その内容を適切に表現しうる単語があることが前提であり、それを効率よく自動的に抽出することである。つまり、文書を構成する文字列の表層上の特徴に着目し、探索・抽出する方式である。そのため、抽出されたキーワードは、文書（文献）内容を適切に表現している保証はなく、また文書の表層上の文字列に出現しないキーワードの必要性の問題もある。

これに対処するために、前述のページ通覧機能の下に、対話的に文書操作コマンドとして索引語指示ができる機能を備えている。

以上、索引用ファイルの構築における３つの方式を述べた。今回は不要語削除方式を利用し、簡単なプログラムで索引語を指示しながら索引用ファイルを構築した。さらに、ページ通覧の機能の下で、文書中に出現しない索引語の指示を行い、検索効率の向上を図っている。形態素解析を利用したキーワード抽出方式は、統語解析や意味解析との結合により、文献内容に合致したより適切なキーワード抽出の発展性がある。今後、出力結果や言語的特徴などの解析を行い実用化に向けて検討する。

４－３　文書データベースの使用例

以下では、「民法要覧」文書データベースの使用例について、特に、索引検索・ページ通覧を実行するときの検索手順と検索結果の画面表示について説明する。

(1) 索引検索を実行し、候補ページ一覧を画面表示

```
SELECT  KW  債権担保  OUT1
```

＜検索結果＞

```
***  （第 2 部）  ***
***     PAGE     ***
   4      12      152      155
 158     159      187      190
```

文書操作コマンドOUTが指定されていない場合、最初のページ４が候補参照域として画面に表示される。この例の場合、利用者の検索意図に合致すると思われる候補参照域のページ一覧が画面表示される。

(2) ページ４の画面表示

```
SELECT  PAGE  4
```

(3) ページの連結機能

文書の論理的構成（階層構造）から利用者の必要とする主題は、その内容を加味し付加されたキーワードの前後に存在する。しかし、システム設計は順序関係を有するページを、利用者が直接・間接に操作するオブジェクトとして提供している。そのため前後のページを連結する制御と次候補参照域への連結機能が必要である。これらの文書操作コマンドは,AFTER,BEFORE,NEXTとして提供される。

```
NEXT
```

検索条件の設定として、曖昧さをもつ検索要求にも対応するため、文字列の部分一致を採用している。また、法律情報は、問題群に対する学説・原則・主義の類のためのキーワードや人物→行為→判例・法令のようなキーワード間の階層構造を持っている[5]。しかし、格納構造として利用した既存システムは、データ項目での配列構造（マルチ・バリュー）を持つことを許し

<(2)の検索結果>

```
### (第 1 部)    4 PAGE ###
                第一部債権序論では、債権法の基本概念（債権や債務）を説
                明する。
                前述したように、債権総論の中心である契約法（第二部）に
                おいて債権総論の諸問題を論述する。（たとえば、債務不履
                行、債務の弁済、多数当事者の債権関係）。
                したがって、第二部は契約の意義、成立、種類等につづいて、
                契約債務、契約の履行、債権担保、債務不履行、損害賠償、    *
                契約の解除、債権譲渡、契約の終了を扱う。
                契約以外の債権関係について第三部で事務管理、不当利得、
                第四部で不法行為を説明する。
1.2             債権
1.21            意義
1.211           債権は、他人（債務者あるいは第三者）を介してある生活利
                益（法的財貨）を獲得することを目的とする権利である。
1.212           債権を債務者に給付行為を請求する権利とする説明もある。
                ただ、この説明は、第三者の弁済において問題に出会う。民
                法は、第三者が債務の弁済をなしうるものとしているが（民
                ４７４）、債権を債務者の給付行為を要求しうる権利とする
                と、第三者の弁済はその内容に入らないことになる。
1.213           債権と請求権
#####   INPUT  COMMAND   #####
: 文書操作コマンド入力領域
```

（注）右余白の＊印はキーワードの存在する行を示している。

<(3)の検索結果>

```
### (第 1 部)   12 PAGE ###
1.25            債権の平等性
                債権の平等性とは、まず、同種の債権が同一目的について相
                前後して成立しても、その間にとくに優劣関係がないことを
                意味する。これは、債権発生における平等性である。この考
                え方は、自由競争の原理に対応したものといえる。更に、同
                一の目的に対して、複数の債権者が請求をしてきても、その
                間に優劣関係はない平等に扱われる。これは、債権の摑取力
                にも及んでいる。ある債権者が債務者所有の財産を差押えて
                も、その者が他の債権者に摑取力行使の点で優遇されない
                （この点は、差押債権者を、たとえば差押質権を与えて一般
                債権者より優遇する法制もある）
1.26            財産としての債権
                債権は、債権者が債務者の給付を要求する権利（対人権）で
                あるが、債権自身が、取引の対象となると、財産としての債
                権の側面が登場してくる。今日では、債権としての同一性を
                失うことなしに、債務者の変更が可能になり（債権譲渡）、
                債権が物と同様に侵害される（第三者による債権侵害）。：
                債権譲渡，債権を目的とする債権担保（債権質，代理弁済，    *
                受領），債権侵害，預金，証券　等
1.3             債務
#####   INPUT  COMMAND   #####
: 文書操作コマンド入力領域
```

（注）次該当（候補）として12ページが画面に表示される。

ていない。これに対処するために、階層構造を持つキーワードを別タブルとして格納し、論理演算子（AND, OR）、特にAND演算子で絞り込む必要がある。

5．おわりに

本稿では、文書データベースの枠組みを検討し、それに基づいて設計・構築した「民法要覧」文書データベースについて報告した。物理的構成と論理的構成という操作ビューの下に文書データベースを構築し、より自由度の高い検索機能（目次検索、索引検索）と通覧機能を提供する。

計算機可読の文書データ（テキストデータ）が大量に蓄積されている現在、これらのデータを効率よく管理し、必要に応じて迅速・正確に利用できる環境を提供するデータベース化と構築手順のツール化は有効な手段を提供する。

今後の課題として、更新機能がある。索引テーブルの更新は可能であるが、本文テーブルに対しては十分に検討されていない。これに対しては、物理的構成をより仮想化して捉える必要がある。また、索引機能の充実があげられる。キーワード抽出における、より一層の言語情報の解析手法や出現頻度の統計的手法の利用など、何らかの知的処理が望まれる。

本データベースは、現在、京都大学大型計算機センターのAIM/RDB（リレーショナル・データベース・システム）の下に実現されテスト運用されている。

謝　辞

日頃からご教授・ご鞭撻をいただいている京都大学大型センター・星野聰教授、法学部・北川善太郎教授、および特に設計についての指導・助言をいただいた名古屋大学・渡邊豊英助教授に感謝します。

参考文献

(1) 伊藤哲朗：『情報検索』昭晃堂,1986

(2) 渡邊・桶谷・北川：「文章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，pp.1088-1089，1989

(3) 桶谷・渡邊：「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」，情報処理学会第38回全国大会講演論文集，pp.1090-1091　,1989

(4) 桶谷・渡邊・北川：「民法要覧文書データベースの実現法」，電子情報通信学会春季全国大会講演論文集，1989

(5) 北川善太郎：「民法講義の要点」（物権法、民法総則、親族法・相続法），1988

(6) 財団法人比較法研究センター：『法律情報システムについての調査・分析（報告書）』,1988

(7) 田中：『自然言語解析の基礎』産業図書，1989

(8) 荻野：「カナ漢字変換用辞書を用いた日本語の形態素解析」テキストデータベース研究会資料1-4,90