オンライン漢字情報検索プログラム
—DORIS－21—

池田 幸雄　中村 仁之輔
（日本電信電話公社 横須賀電気通信研究所）

1. まえがき

文献情報，特許情報，新聞記事情報など大量の情報の中から必要とする情報をオンラインで迅速に検索し，利用者に提供する情報検索サービスが，我国でも急速に発達し，定着しはじめた。公社では，このような情報検索システムを構築するツールとして1978年にオンライン情報検索プログラムDORIS－2*を開発した。[1]以来，DORIS－2は多くの情報検索システムの構築に適用されてきたが，適用領域の拡大により，機能拡充の要求も高まってきた。

DORIS－21は，この要求に応えるため，DORIS－2をベースに漢字処理機能を中心とする機能拡充を図ったものであり，現在，公衆システムDEMOS－Eでサービスに供されている。

本稿では，DORIS－21の主要な拡張機能である以下の項目の実現方法について報告する。

① 従来の英・数・カナ文字（以下ANKと呼ぶ）系コードと漢字系コードが混在するコード系混在データベースの制御方式
② 漢字データをデータベースに効率よく格納する漢字データ圧縮格納方式
③ 端末への入力促進メッセージや質問手順を利用者が自由に定義できる会話手順定義言語

2. 設計方針

DORIS－21の機能拡充に当っては，以下の基本方針をとった。

① ANK系コードをベースとするTSS**配下で動作し，従来のANK系プログラムと共存できること。
② ANK系データベースを対象としたDORIS－2の機能を完全に包含し上方向互換を保証すること。
③ 読解性に優れている反面，入力が煩わしいという漢字の特徴を考慮し，マンマシンインタフェースの向上を図ること。
④ 漢字コード，キャラクタ集合の拡張法等については，情報交換性を考慮しJISに極力準拠した方式を採用すること。

3. システム概要

DORIS－21はCODASYL仕様[2]をベースに，インバーテッドインデックス機能を追加した検索向きのデータベースモニタを中核として構成されており，大規模データベースの構築が可能である（図1参照）。検索機能としては，文献情報向きの論理検索，定量検索，シソーラス参照検索，一般的情報向きとしてリレーショナルデータベース・モデルに基づいた関係検索機能があり，目的に応じ選択できる。また，TSS上で情報検索サービスを容易に構築できるよう，オンライン情報収集機能や利用者面サービス管理機能等を具備している。

* Dendenkosha Online system for Retrieval of Information and Storage
** Time Sharing System

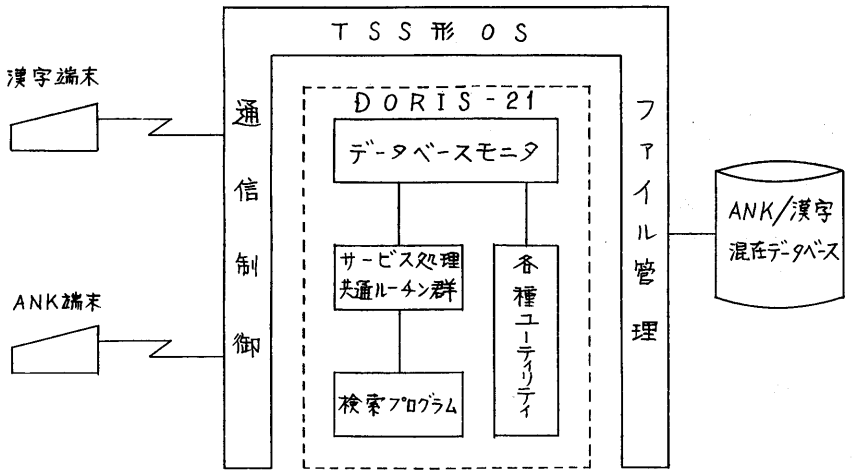


図1　オンライン漢字情報検索システムの構成

4. 漢字処理機能の導入範囲

DORIS-21は，従来のANK系データベースにおけるANKデータの操作と同レベルの機能を漢字データについても実現できることを目標とし，以下の機能を実現している。

(1) モニタ／ユーティリティの実現機能
- (i) スキーマ／サブスキーマ定義時の漢字データ属性の指定
- (ii) 漢字コードを意識したデータ操作（比較，転送）
- (iii) 漢字データの圧縮格納

(2) 検索プログラム／サービス処理共通ルーチンの実現機能
- (i) 漢字端末からの漢字データ入力（カナ漢字変換機能等）
- (ii) 漢字データとANK系データを意識した書式編集
- (iii) 入力促進メッセージやエラーメッセージ等の漢字化

5. 実現方式

5.1 コード系混在データベースの制御方式[3]

従来のANK系データは1バイト／文字のコードで表現されるが，漢字データは文字種が多く2バイト／文字で表現される。「JIS-C-6226 情報交換用漢字符号系」では漢字キャラクタ集合には，ANK文字や特殊記号も含まれており，文字列データは全て2バイト／文字の漢字系コードで表現することもできる。しかし，データベース中に格納されるデータ項目には，ANK系文字のみで表現されるものも多い。データベースへの格納効率や伝送効率を考慮すると，これらは，従来どおり，ANK系コードで表現するのが望ましい。コード系が混在するデータを扱うためには，システムの各所でコード系を識別して処理することが必要となる。そこで，DORIS-21では，図2に示す方法によりコード系混在データベースの制御を可能とし，ANK系機能との融合を実現した。

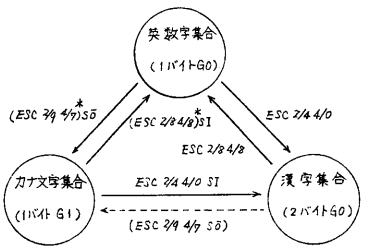
(1) 通信回線上のキャラクタ集合の切替え方法

漢字端末から入力される質問文や検索結果データなど通信回線上の伝送テキストは，ANK系と漢字系コードが混在したものとなり，両者を識別する手段が必

要となる。複数キャラクタ集合の切替え方法について規定している「ＪＩＳ－Ｃ－６２２８」では漢字集合（２バイトＧ０集合），英数字集合（１バイトＧ０集合），カナ文字集合が独立したキャラクタ集合として定義されており，各キャラクタ集合を指定するため，それぞれ３桁のエスケープ・シーケンスが定められている。ＤＯＲＩＳ－２１は，ＡＮＫ系ベースのＴＳＳ配下で動作し，既存のＡＮＫ系プログラムとの融合を図るため，図２に示すようにカナ文字集合に対するエスケープ・シーケンスは省略し，１バイトＧ０，Ｇ１集合は，従来どおり，シフトイン／シフトアウト符号（ＳＩ／ＳＯ符号）のみで切替える方式を採用した。

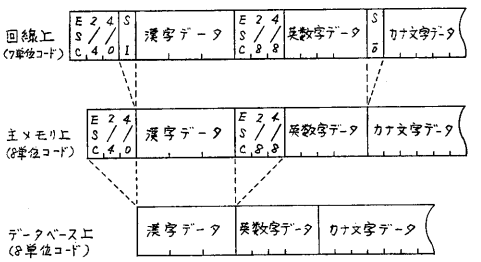
(2) データベース内の識別方式

データベースに格納されるデータは，データベース定義時にデータ項目単位でその表現形式が属性として定義される。ＤＯＲＩＳ－２１では，従来の数値データ属性，文字列データ属性の他に新たに漢字データ属性を追加し，識別可能とした。検索プログラムはディレクトリの属性情報を参照しながら，データ項目の属性に応じた処理ができる。このためデータベース格納時には，格納効率を向上させるため，エスケープ・シーケンスは削除し，回線への出力データ編集時に検索プログラムで付加する方式をとった。なお，１データ項目中にＡＮＫ系／漢字系コードの混在を許すと，インラインにエスケープ・シーケンスを記述する必要があり，比較操作や書式編集処理も複雑となる上，格納効率向上効果も小さいため１データ項目内の混在は禁止している。



(注) *：JISで定義されているが省略
←---：JISでは許されているが禁止

(1) キャラクタ集合の切替え法（回線上）



(注) データベースではディレクトリでデータ項目単位に属性(漢字/英数カナ)を管理しており，端末への出力時エスケープシーケンスを付加する。

(2) コード系混在データの表現例

図２ コード系混在データベース制御方式

(3) システム・メッセージの切替え

検索プログラムから端末に出力する促進メッセージやエラー・メッセージなどは，マンマシン・インタフェース向上の面から漢字化するのが望ましい。しかし，ＡＮＫ系端末からＡＮＫ系データベースを検索する利用者に対しては，漢字メッセージは出力できない。このため，システム・メッセージはＡＮＫ系，漢字系の２系統用意し，使用端末種別を識別し，切替えて出力する方式を採用した。

(4) 漢字データの入力形式

ＤＯＲＩＳ－２１での漢字データの入力は，大量の場合は，事前に作成した入力ファイル（ＩＳＯ　２７０９形式）から一括入力可能であるが，小量の更新データや質問文中の漢字キーワード等の入力は通常，端末より会話的に入力する。後者の場合，簡易な入力方法の提供が重

表１ 漢字データの入力形式

項番	形式	入力例	変換後
1	漢字直接入力	$図書検索$	図書検索
2	カナ漢字変換入力	%トショ ケンサク%	図書検索
3	区点入力	[3162 2981 2401 2687]	図書検索

要となる。漢字入力に関する利用者インタフェースは，漢字端末側で機能分担するのが望ましいが，現時点では端末価格の問題があり，センタ側で補完する必要がある。ＤＯＲＩＳ－２１では，漢字キーボード等による漢字直接入力機能を持たない簡易漢字端末の接続を可能とするため，表１に示すようなカナ漢字変換機能，外字入力用としてＪＩＳの区点入力機能を実現した。

カナ漢字変換では変換用辞書が必要である。あらゆる分野の用語が網羅されたシステム辞書を提供できるのが理想であるが，現在そのような辞書は存在しない。このため，ＤＯＲＩＳ－２１では，システム辞書と利用者定義辞書を切替えて利用できる機能を実現し，一般用語は前者で，分野毎の専門用語は後者で対処することとした。

5.2 漢字データ圧縮格納方式[4][5]

情報検索サービスでは，大量の情報を蓄積するため，ファイル料金がシステム運用経費の大部分をしめており，経済化のためにはファイル容量の削減対策が不可欠である。ＤＯＲＩＳ－２１では各種のデータ圧縮技術を実現しているが，ここでは，漢字データの圧縮格納の実現方式について述べる。

表2　漢字データ圧縮方式の比較

比較項目	シフトコード方式	コンパクトコード方式	ハフマンコード方式	FVCC方式
ファイル削減率	25％	20％	45％	40％
処理オーバヘッド*	1.05	1.10	1.08	1.08
コード変換テーブル	不用	7ページ	100ページ	6ページ
汎用性	有り	無し**	有り	有り
総合評価	不適	不適	不適（英数データ加味）	適

（注）＊　非圧縮時の処理ステップ数を１とした値
＊＊　バイナリデータが混在する場合は適用できない

(1)　圧縮方式の比較

従来，提案されている文字列データに対する代表的な符号化圧縮方式には以下のものがある。

①　シフトコード方式

シフトコードを利用し，各文字をより短いビット列の固定長符号で表現する方式

②　コンパクトコード方式

文字列の出現頻度を調べて，出現頻度の高い文字列を符号表中の未使用コードに割り当て，１文字で表現する方式

③　ハフマンコード方式

文字単位の出現頻度を調べて，出現頻度の高い文字は短いビット列を，低い文字には順次長いビット列を割り当て，可変長符号で文字を表現する方式

各方式を代表的な文献データに適用した結果は表２に示すとおりである。ファイル削減率ではハフマンコード方式が最も優れているが，コード変換テーブル・サイズが大きくなり，メモリ使用量に制限のある場合適用できない。そこで，ＤＯＲＩＳ－２１では，ハフマンコード方式の長所を活かし，テーブル・サイズをコンパクトにおさえ，文字種の多い漢字データへの適用を可能にしたＦＶＣＣ方式*を考案した。

(2)　ＦＶＣＣ方式

ハフマンコード方式では，全ての文字に可変長コードを割り当てるため，漢字データのように字種が非常に多い場合（通常３０００～８０００字程度必要），圧縮および復号用コード変換テーブルは膨大となる。一方，実際に使用されてい

＊ Data Compression Method with Fixed Variable Length Coded Characters

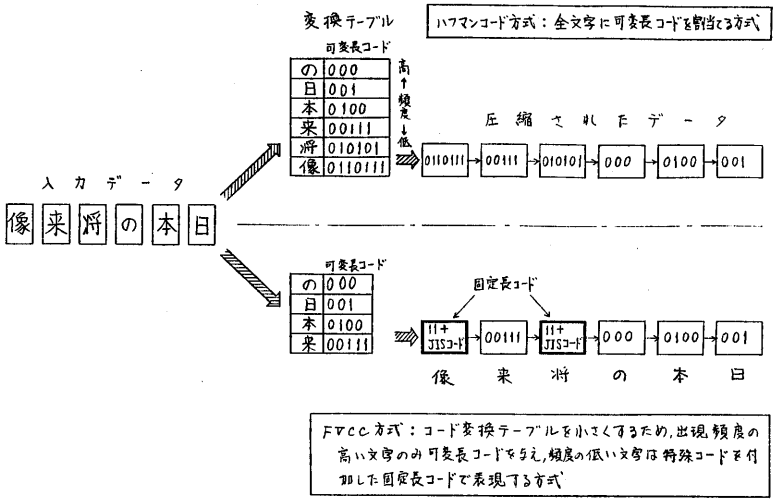


図3　FVCC方式による漢字データ圧縮格納の概念

3文献および新聞の記事情報データベースを対象に文字の出現頻度の統計をとったところ，頻度の高い数百文字が全体の出現頻度の95%以上をしめていることが明らかとなった。FVCC方式は，この特性を利用して，出現頻度の高いものから順に一定数の文字のみに可変長コードを割り当て，その他の文字は復号用コード変換テーブルが不要な固定長コードで表現する方式である．可変長コードにはハフマンコードを割り当て，固定長コードは可変長コードと混在していても正常な復号処理が行なえるよう，縮退を示す1個のハフマンコードを元のJIS漢字コードに付加したもので表現する（図3参照）。

本方式の圧縮効果は可変長コード化する文字数に依存し，数百文字を可変長コード化する場合，ハフマンコード方式より数パーセント低下するが，コード変換テーブル・サイズは小さくてすみ漢字データの圧縮に対しては実用的な方式である。DORIS-21では，さらに以下に述べるコード変換テーブルの構成方法を考案し，メモリ制約の厳しいTSS上でも実現可能なテーブル・サイズを得ることができた。

(3)　圧縮用コード変換テーブル構成法

テーブル構成方法については，次の3方式が考えられる（図4参照）。いずれの方式も圧縮方式はFVCC方式を前提とし，テーブルへの登録文字数は出現頻度の高い一定数に限定することとする。

①　バイナリサーチ方式

2バイトのJIS漢字コードと対応するハフマンコードおよびハフマンコード長の3項目から成るエントリをJIS漢字コードを数値とみなし昇順にテーブルに格納しておく。対応コードのサーチは，テーブルの中央のエントリから始めて，順次テーブルを2分割しながら範囲を絞り，一致するJIS漢字コードがみつかるまで続ける。

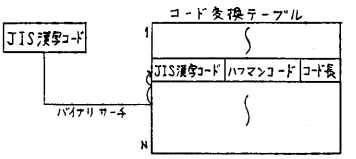
②　直接アドレス方式

ハフマンコードとその長さの2項目から成るエントリを，それに対応するJIS漢字コードの値を持つ番地に格納しておく。対応コードはJIS漢字

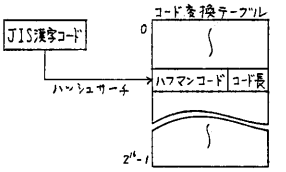
コードの値で直接見つけることができる。

③ 間接アドレス方式

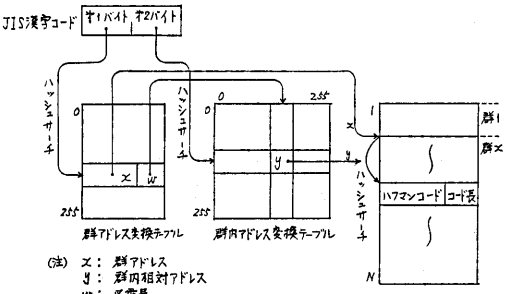
ハフマンコードとその長さの2項目から成るエントリを，予め対応するJIS漢字コードの第1バイト目の値（区）で群分けし，変換テーブルとして登録しておき，そのエントリへ高速にアクセスするための2つのマッピング・テーブル（群アドレス変換テーブル，群内アドレス変換テーブル）を持たせる。マッピング・テーブルはJIS漢字コードの第1バイト目の値より，変換テーブルの群開始番地が，2バイト目の値（点）より，群内相対番地が得られるよう構成する。対応コードは，3回のハッシュ・サーチで見つけることができる。



(1) バイナリサーチ方式



(2) 直接アドレス方式



(注) x：群アドレス
y：群内相対アドレス
w：区番号
N：出現頻度の高いものN字

(3) 間接アドレス方式

図4 圧縮用コード変換テーブルの構成方式

これらの3方式を変換テーブルに登録する文字数をパラメータとして，テーブル・サイズ，処理速度について比較すると図5のようになる。

バイナリサーチ方式はテーブル・サイズは小さいが処理速度が極端に遅く，直接アドレス方式は，処理速度は速いが，テーブル・サイズは極端に大きくなる。間接アドレス方式は処理速度，テーブル・サイズのバランスがとれており実用的な方式であるため，DORIS-21では本方式を採用した。

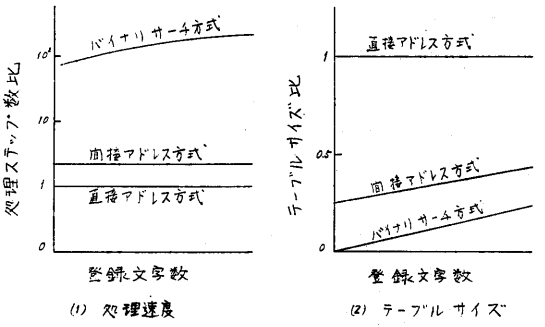


図5 コード変換テーブル構成方式の評価

(4) 復号用コード変換テーブル構成法

復号用テーブルの構成方法としては，次の2方式が考えられる。

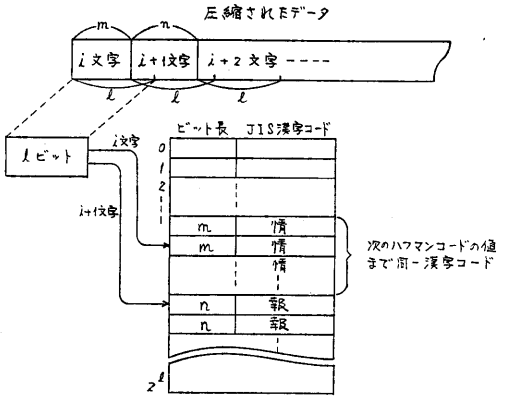
① バイナリサーチ方式

ハフマンコードとその長さおよび対応するJIS漢字コードの3項目から成るエントリをハフマンコードの値の昇順にテーブルに格納しておく。対応するJIS漢字コードは，ハフマンコードをキーとして，圧縮用コード変換テーブルのバイナリサーチ方式と同様の手法でサーチする。

② 直接アドレス方式

図6に示すようにJIS漢字コードと対応するハフマンコードの長さの2項目から成るエントリを2^{ℓ}（ℓは最大ハフマンコード長）個持つテーブルを用意し，長さn（$n \leq \ell$）のハフマンコードについては，先頭nビットが

該ハフマンコードに一致するℓビットの2進数値をアドレスとする$2^{\ell-n}$個のエントリに対応するJIS漢字コードとハフマンコード長を重複して格納しておく。復号処理時は，圧縮されたデータからℓビット切り出し，それを2進数値で表現したアドレスとみなし，対応するエントリに直接アクセスし，元のJIS漢字コードを得る。



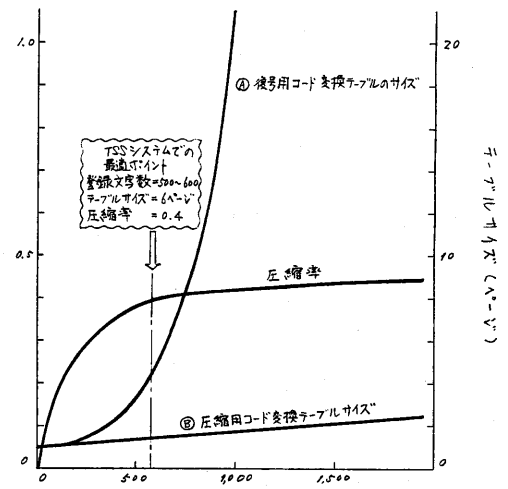
(注) ℓ：最大ハフマンコード長

図6 復号用コード変換テーブル(直接アドレス方式)

この2方式を比較すると，バイナリサーチ方式は，圧縮用コード変換テーブルと同様にコード変換テーブルは小さくてすむ反面，処理速度は格段に遅くなる。復号時のオーバーヘッドは，検索の応答時間に与える影響が大きいこと，直接アドレス方式でも，ハフマンコード化する文字数を一定値におさえるFVCC方式の場合，テーブル・サイズが許容範囲に収まることを考慮し，DORIS-21では，処理速度の速い直接アドレス方式を採用した。

(5) FVCC方式の評価

上記のコード変換テーブルの構成法を採用したFVCC方式を代表的な文献データに適用した時の圧縮効果とテーブル・サイズの関係を図7に示す。圧縮効果は変換テーブルに登録する文字数に比例して大きくなるが，一定の文字数のところから飽和する。コード変換用テーブル・サイズは登録文字数に比例して大きくなるが，復号用テーブルは，一定文字数のところで急激に増大する。DORIS-21の走行するTSSシステムでは，圧縮／復号用コード変換テーブル・サイズの和が数ページで，圧縮効果の増加率が減少する登録文字数500～600が最適なポイントとなっている。



(注1) 圧縮率 = 削減された下ファイル容量 / 圧縮前のファイル容量

(注2) 本方式で必要なテーブルサイズ = Ⓐ + Ⓑ

図7 圧縮効果とテーブルサイズの関係

なお，本方式の圧縮・復号処理による処理ステップ数の増加は，標準的な使用パターンで約8%程度となっている。

5.3 会話手順定義言語[6]

DORIS-21では，エンドユーザ言語の1つとして，リレーショナル・データベース・モデルに基づいた関係検索機能を実現しており，端末から簡易なコ

マンド形式の質問文を入力することにより種々の検索が行なえるようになっている。しかし，本機能は汎用的である反面，全データ項目が検索キーになり得るため，条件式ではデータベース定義時に指定したデータ項目名（ドメイン名）を指定する必要があるとか，インデックスの作成されていないデータ項目を検索キーに指定した場合，レスポンスタイムが極端に長くなる場合がある等，利用者にある程度データベースの構造を意識させざるを得ないという問題があった。そこで，関係検索機能の長所を活かし，かつ利用者インタフェースを更に改善することを目的として，以下に述べる会話手順定義言語（ＵＤＬ）＊を開発した。

(1) エンドユーザ・インタフェース実現方式の比較

エンドユーザ・インタフェースの実現方式としては次の４方式が考えられる。

① 業務プログラム作成方式

適用業務毎にデータベース操作言語（ＤＭＬ）＊＊と一般のプログラミング言語を使用して検索プログラム全体を作成する方式

② 入出力部オウンコーディング方式

汎用的な検索プログラムを作成し，個々の適用業務毎に入出力処理部をオウンコーディングする方式

③ 標準コマンド提供方式

汎用的な検索プログラムを作成し，標準的なコマンドのみを提供する方式

④ マクロ言語方式

汎用的な検索プログラムの検索言語をベース言語として，マクロ定義言語により，適用業務毎の利用者インタフェースを設定する方式

表3 検索用エンドユーザ言語実現方式の比較

比較項目	業務的プログラム作成方式	入出力部オウンコーディング方式	標準コマンド提供方式	マクロ言語方式
ユーザインタフェースの多様性	最良	良	劣る	良
ニーズへの適応性	遅い	やや遅い	既製品の範囲で対処	速い
高度な機能の実現	制限無し	やや制限あり	制限有り	やや制限有り
プログラム開発期間	大	中	プログラム作成無し	小
総合評価	可	良	良	優

これらの方式の比較結果を表３に示す。ＤＯＲＩＳ－２１では多様な利用者のニーズに速やかに対処できることが重要であり，マクロ言語方式と入出力部オウンコーディング方式を併用することにした。

(2) ＵＤＬ言語仕様

マクロ言語方式の思想に基づき開発した会話手順定義言語ＵＤＬの言語仕様を表４に，ＵＤＬを使用した会話手順の定義例および実行例を図８に示す。

ＵＤＬの特徴は，以下のとおりである。

① 関係検索用標準コマンドをベース言語とする一種のマクロ定義言語であり，端末から入力されたパラメータ値がマクロ骨格を構成する標準コマンド言語

表4 会話手順定義言語ＵＤＬの機能

分類	文名	機能
宣言文	＊NAME（＊N）	質問文ブロックの定義を行う。
	＊PARAMETER（＊P）	質問文ブロックのパラメータの省略時値を定義する。
実行文	＊COMMENT（＊C）	指定されたメッセージを端末に出力する。
	＊REQUEST（＊R）	指定された促進メッセージを端末に出力し，パラメータの入力を要求する。
	＊JUMP（＊J）	無条件あるいは端末からの入力に従い指定された質問文ブロックに制御を移す。

(注) （ ）内は省略形

＊ End-User Interface Definition Language

＊＊ Database Manipulation Language

に展開されて実行される。

② 端末への入力促進メッセージや案内情報を漢字かな混りの日本語文で自由に設定でき，メニュー選択方式のインタフェースが容易に実現できる。

(3) UDL処理系

UDLの処理概念を図9に示す。UDLインタプリタは検索プログラムからテキスト要求を受付けると会話手順登録ファイルから，指定された会話手順テキストを1レコードづつ読み込み，該レコードが表4に示したUDL文であれば，解釈，実行し，端末へメッセージを送信したり，入力パラメータを受信したりする。UDL文でない場合，ベース言語の仕様に従ったテキストとみなし，入力パラメータのうめ込みなど展開処理を行った後，要求元へテキストを渡す。検索プログラムは，自分の仕様に従い構文解析，実行を行ないながら次のテキストをUDLインタプリタに要求し，順次処理を実行する。

UDLインタプリタの処理と検索プログラムで実行されるベース言語の処理は独立であり，UDLインタプリタは各種のベース言語をもつ検索プログラムから共通に利用できる構成となっている。

(4) UDLの効果

CODASYLタイプのデータベース操作言語（DML）が使用できる拡張COBOLを用いた場合と，関係検索言語とUDLを用いて同一機能を実現する場合のステップ数を比較すると後者は約1／10〜1／20と小さく，UDLによりプログラム作成工数が大幅に削減できることが確認できた。

```
•N STA
•C
•C §×××  どの検索を行いますか。番号で答えて下さい。  ×××§
•C
•C § [1]   [XX（情報検索‥）] に関する文献を借りた人は？§
•C § [2]   [ZZ] 氏（編）の借りた本は？§
•C § [3]  昭和YY年MM月DD日に借りた人は？§
•C § [4]  昭和YY年MM月DD日に返却した人は？§
•N RPLY
•C
•R §番号は？§   [1/2/3/4]   §……  §
•J 1=BLK1,2=BLK2,3=BLK3,4=BLK4,END=STOP,E=STOP,ERR1
•N BLK1,BLK1
•C
•C §キーワードを入れて下さい。   [例§  ジョウホウケンサク]
•C
•R §    ……   §
•P C1=LSI
@GET B,K,M                                          ┐
@FOR B.BCODE=K.BCODE AND K.NCODE=M.NCODE            │---- 検索プログラム
@FOR EXIST(B.KW1)="$C1$"                            │     の標準コマンド言語
@PRI M.NAME2,M.SECTION,M.PHONE,K.LDATE,B.TITLE3;,,40┘
/RUN
                    ⋮
```

(1) 会話手順の定義

```
×××  どの検索を行いますか。番号で答えて下さい。  ×××

[1]   [XX（情報検索‥）] に関する文献を借りた人は？
[2]   [ZZ] 氏（編）の借りた本は？
[3]  昭和YY年MM月DD日に借りた人は？
[4]  昭和YY年MM月DD日に返却した人は？

番号は？  [1/2/3/4]  ……  1

キーワードを入れて下さい。   [例  ジョウホウケンサク]

   ……  ヒカリツウシン

通番 氏名（漢字）          研究室名            電話番号   貸出日
     和文標題

0001 東  修               井上研究室             3190     550409
     アトランタ光ファイバシステム実験   光通信システム用GaAlAsレーザ送信
     器

検索を続けますか？  [YES(Y)/NO(N)]  ……  Y
                    ⋮
```

(2) 実行結果（XXXXは利用者の入力を示す）

図8 会話手順定義言語の使用例

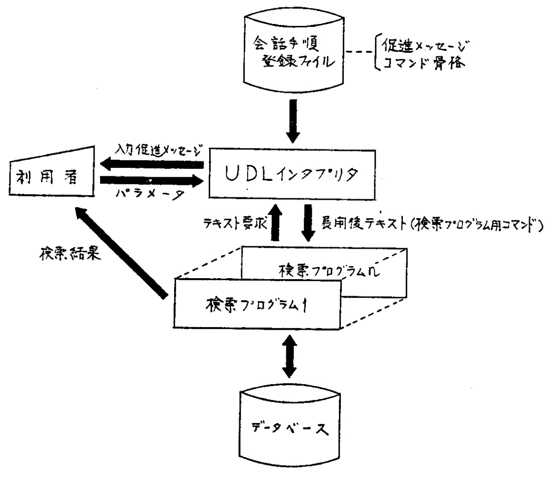


図9 会話手順登録言語UDL処理概念

6. むすび

ＴＳＳ上の情報検索プログラムの適用領域拡大を目的とし，漢字処理を主体とする機能拡充を行ったＤＯＲＩＳ－２１の実現方法について述べた。

本実用化で確立した主要技術は以下のとおりである。

① ＡＮＫ系と漢字系コードが混在するコード系混在データベースの制御方式

② 可変長コードと固定長コードを併用し，漢字データを効率よく格納するデータ圧縮格納方式

③ 端末への入力促進メッセージや質問手順を利用者が自由に定義できる会話手順定義言語の実現方式

本実用化により，ＴＳＳ上で，漢字情報検索サービスが容易に構築できるようになった。

＜参考文献＞

〔1〕 長峯・北村他： 情報検索プログラム（ＤＯＲＩＳ－２１），第１７回データベース管理システム研究会，１９８０

〔2〕 ＣＯＤＡＳＹＬ： ＣＯＤＡＳＹＬ Data Description Languages Committee. Journal of Development，１９７８

〔3〕 池田： 英カナ文字システムでの漢字処理機能実現法，情処学会全大，１９７９

〔4〕 杉山・中村他： 日本語データの効率的な格納方法，情処学会全大，１９７９

〔5〕 杉山・長峯： 日本語データのハフマンコード圧縮アルゴリズムについて，信学会情報システム部門全大，１９７９

〔6〕 池田・杉山他： 日本語情報検索用ＥＵＬの一実現法，信学会全大，１９８１