データベース・システム 31－2
（1982. 7. 15）

情報検索システムORIONにおける漢字処理

池上 信男
（日立製作所ソフトウェア工場）

1. はじめに

これまでの情報検索システムでは、取扱うデータのほとんどが英文であり、日本語の処理はカタカナ止まりであって、漢字データを扱うシステムは極めて特殊であると見られがちであった。最近、漢字端末や漢字プリンタ等の漢字用ハードウェアの普及、及びオペレーティングシステム、ユティリティ、言語による漢字サポート等、漢字ソフトウェアの普及にともない、漢字による情報検索システムの開発は必須のものとなった。日立製作所では、1979年12月から情報検索システムORION（Online Retriever of Information）を提供している。当システムでは英数カナデータのみをサポートしていたが、上記の要求にこたえて、1981年6月から日立漢字情報システムの一環として、ORION漢字支援の提供を開始した。現在、ORIONのユーザ数は約50、ORION漢字支援のユーザ数は約30であり、大学、研究所、官公庁、民間企業で利用されている。

2. システムの概要

2.1 ORIONの特長

ORIONは、文献（文章）データ、数値データを中心とする情報を蓄積、維持し、それを対話的に効率よく取出すことができる。また、検索結果を加工、編集しレポートを作成することもできる。特にORION漢字支援が加わったことにより、漢字を含む日本語データが容易に扱えるようになり、適用範囲は一層拡大された。ORIONの機能は、データベースの情報を検索利用するエンドユーザ向けと、データベースを作成管理するシステム管理者向けとの二つに大別できる。それぞれの側から見た機能の特長を図1に示す。

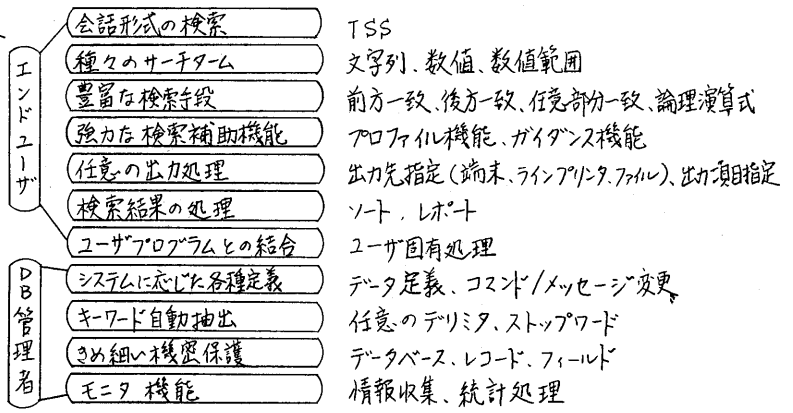

図1 ORIONの特長

2.2 ORIONの構成

ORIONの構成を図2に示す。ここで、検索機能と検索補助機能は、TSS、バッチの両モードで動作し、それ以外の機能は、バッチモードでのみ動作する。以下に、図中の主要ファイルを説明する。

（1）メッセージファイル

ORIONのメッセージ、及びコマンドを含む文法上の予約語を格納している。

（2）テーブルファイル

データ定義言語で指定された、データベースのファイル、レコードの定義情報、機密保護情報、索引づけオプション、及びその他種々のオプション指定情報を格

納している。

（3）ヘッドファイル

情報レコード自体を格納している。キーはアクセション番号（Accession number）つまり、各レコードに付けたユニークな番号である。

（4）インデクスファイル

インデクスターム（キーワード）をキーとし、ヘッドファイル内のレコードへのポインタ（アクセション番号の列＝ポスティング）を持っている。

（5）レンジファイル

レンジタームと各レンジタームごとの実際の数値データを格納している。ここで、レンジタームとは、数値範囲とその数値の種類を表わす文字列から構成されるインデクスタームの1種である。

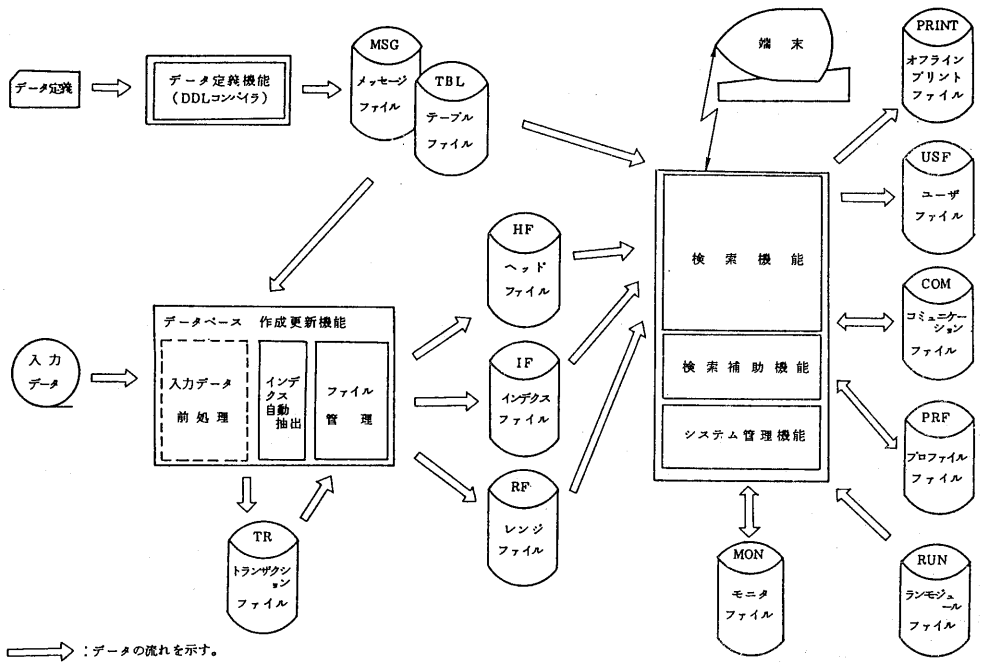

図2　ORIONの構成

2.3　会話コマンド一覧

検索機能、検索補助機能とコマンドの対応を表1、表2に示す。

表1　検索機能一覧

機能		コマンド	意味
会話制御	検索開始	CALL (TSS)	ORIONを起動する。
	検索終了	QUIT	ORIONを終了する。
	データベース変更	RISTART	対象データベースを変更する。
	コマンド処理中止	割込みキー	コマンド処理を中止する。
	オプション設定	SET	実行時に各種オプションを変更する。

表1のつづき

機能		コマンド	意味
検索方式	インデックスサーチ	FIND	インデックスファイルを使って検索する.
	レンジサーチ	FIND	数値範囲指定で検索する。
	シーケンシャルサーチ	SCAN	ヘッドファイルを各種指定条件で検索する.
特殊サーチ	ユニバース	SET UNIV	指定した部分集合内で検索を進める.
	ハイアラキー	SET HIER	指定条件で集合を絞っていく.
	エクスパンド	EXPAND	指定した部分集合を別の条件で分類.
検索結果出力	端末出力	DISPLAY	検索結果を端末に出力する.
	オフラインプリント	PRINT	中央のラインプリンタに出力する.
	アドレス指定	ADDRESS	送り先住所をプリントする.
	表題指定	TITLE	検索結果に表題をプリントする.
ソート		SORTO	検索結果を指定条件でソートする.

表2　検索補助機能一覧

機能		コマンド	意味
プロファイル	作成	−MAKE	検索コマンド群を作成し、登録する.
	編集	−EDIT	登録されたコマンド群を修正する.
	実行	−EXECUTE	登録されたコマンド群を実行する.
	可変パラメタ	〔value〕	検索コマンドの一部を可変パラメタとする.
ガイダンス	隣接ターム表示	LOOK	インデクスターム不一致時隣接タームを表示する.
	ターム一覧表示	LOOK	指定に従ってインデクスタームの一覧表を表示する.
	入力再表示	LIST	前に入力したコマンドを再表示する.
	説明表示	?	コマンドの一覧や説明を表示する.
	時間表示	CLOCK	経過時間、CPU時間を表示する.
レポート	レポート言語		SET、GET、IF等23種のステートメント.
	システム開始	REPORT	レポートシステムを起動する.
	プログラム作成	MAKE	レポート言語でプログラムを作成する。
	編集	EDIT	プログラムを修正する.
	コンパイル	COMPILE	プログラムをコンパイルし登録する.
	実行	EXECUTE	登録されたレポートプログラムを実行する.
ユーザプログラム呼出		RUN	ユーザ作成プログラムを呼出し実行する.

3. 漢字サポート

3.1　ORION漢字支援の特長

ORIONと組合せて、漢字データを含む情報検索システムを実現するため、ORION漢字支援を開発した. ORION漢字支援は、入力の簡便さと日本語出力の見やすさに重点を置き、次のように配慮している。

(1) 検索条件の入力

カナ入力と漢字入力の両方が可能である。漢字入力については、カナ入力に比べ操作性は劣るが、検索時の適中率を高めるために必要と考えた.

(2) 検索結果の出力

原文(日本語データ)をそのまま漢字端末やセンタの漢字プリンタに出力する

だけでなく、フィールドの見出しも日本語が使える。さらに、レポート機能を利用すれば、簡単な表形式出力、ユーザファイルへの出力もできる。

（3）日本語メッセージ

対話処理の際、ＯＲＩＯＮから検索者へのメッセージを日本語で表示する。ＯＲＩＯＮとＯＲＩＯＮ漢字支援を組合せた場合、漢字データベースと既存データベースの両方を使い分けることができる。メッセージファイルは、日本語と英語の２種類が用意され、検索セション開始時に、どちらを使うか指定できる。この方式は、将来各国語メッセージをサポートする道を開いている。

3.2　ＯＲＩＯＮ漢字支援のサポート項目及び実行例

ＯＲＩＯＮ漢字支援のサポート項目を表３に示す。

表3　ＯＲＩＯＮ漢字支援のサポート項目

機能分類	サポート項目
日本語メッセージ出力	ＯＲＩＯＮ会話処理中のメッセージ
データ定義機能	(1)フィールド属性として漢字の指定 (2)出力時見出しの１６進指定 (3)カナ漢字対応づけ検索用索引づけフィールドの指定 (4)漢字直接入力検索用索引づけフィールドの指定 (5)ユーザ定義メッセージの１６進指定
データベース作成更新機能	(1)入力データ前処理プログラムによる入力（漢字）データのリスト出力 (2)インデクスファイル作成プログラムによる漢字インデクスターム及びカナ漢字インデクスタームのリスト出力
検索機能・検索補助機能	(1)カナ漢字対応づけ検索（ＬＯＯＫコマンド） (2)漢字直接入力検索（ＦＩＮＤ、ＳＣＡＮコマンド）。 (3)漢字フィールド値による分類、表示及びインデクスサーチ（ＥＸＰＡＮＤコマンド） (4)漢字データ出力（ＤＩＳＰＬＡ、ＰＲＩＮＴコマンド） (5)漢字フィールドの編集出力及びヘッダの１６進指定（ＲＥＰＯＲＴコマンド）
システム管理機能	ユティリティによるカナ漢字対応づけタームの表示

以下、実行例を示す。なお、下線部分はユーザ入力を示す。

（1）ＬＯＯＫコマンドの例

（例1）カナ漢字インデクスタームのカナ部分の後方が「オオコク」となっているものを表示し、タグ指定により検索結果集合を作成する。（カナ漢字対応づけ検索における後

（例1）　3／LOOK　GF：＊オオコク　　（下線部はユーザ入力）

	件数	タターム
A	3	GF：オオコク　王国
B	1	GF：リッケンオオコク　立憲王国

前方／後方一致タームのおわりです。
タグあるいは，コマンドを入力して下さい。
3／A
3　件で検索結果集合　3　を作成。

方一致検索：カナ入力）

（例2）漢字インデクスタームの後方が、「共和国」となっているものを表示する。（漢字後方一致タームの表示：漢字入力）

(例2)

```
4/LOOK  GFK:*〔共和国〕
```

下線部はユーザ入力
〔〕内は漢字データを示す。

	件数	ターム
A	5	GFK：社会主義共和国
B	1	GFK：立憲共和国
C	4	GFK：民主共和国
D	7	GFK：連邦共和国
E	8	GFK：人民共和国

前方/後方一致タームのおわりです。
タグあるいは．コマンドを入力して下さい。

(2) ＳＣＡＮコマンド、ＤＩＳＰＬＡＹコマンドの例

（例3）ＦＩＮＤコマンドで全レコード（全世界の国）の集合を作成し、これに対して、１７番フィールド（対日関係のフィールド）中で「友好関係」または「親日的」という文字列をＳＣＡＮコマンドで探し、その国名を表示する。さらに、トンガ王国（９レコード目）の内容を表示する。

(例3)

```
4/ FIND NM:*
 NM:
 100  ターム以上あります。
 *     166    4/  NM: ( 166  ターム 連結)

   22/ SCAN 4 #17 *友好関係* OR #17 *親日的*

 *       7   22/  SCAN  4   #17 *友好関係*
 *       2   23/  SCAN  4   #17 *親日的*
 *       9   24/  SCAN  4   #17 *友好関係* OR #17 *親日的*
   25/ DISPLAY 1 FOR ALL
    1. 国名              ベルギー王国
    1. 国名              イギリス
    1. 国名              スイス連邦
    1. 国名              ドイツ連邦共和国
    1. 国名              イタリア共和国
    1. 国名              スペイン王国
    1. 国名              ネパール王国
    1. 国名              パキスタン
    1. 国名              トンガ王国
   25/ DISPLAY ALL FOR 9
    1. 国名              トンガ王国
    2. 国名（英語）      TONGA
    3. 地域              大洋州
    4. 面積              671.0 平方キロ
    5. 人口              9.6 万人
    6. 人口密度          130.0 人／平方キロ
    7. 首都              ヌクアロファ
    9. 言語              英語，トンガ語
   10. 宗教              キリスト教
   11. 政体              リッケンクンシュコク 立憲君主国
   12. 元首              タウファアハウ・トウポウ４世国王
   13. 加盟国際機関      アジア太平洋経済社会委員会。アジア開発銀行。ＷＨＯ
       。英連邦。南太平洋諸国会議
   14. 通貨              パアンガ
   15. ＧＮＰ            0.4 億ドル
   16. 主生産物          コプラ，バナナ，野菜
   17. 対日関係          トゥポウ国王は親日的で，トンガ力士が来日するなど，
       関係は全般的に良好。最近では無償援助（食糧など），青年海外協力体によ
       る技術協力が行われた。また７６年には「日本人村」も開村。日本での
       人国査証発給事務などは在火英国大中使館が代行している。
   18. 寸評              北部の山島，中南部のサンゴ礁など大小１５０の島々
       からなっているため，輸送手段が開発に不可欠だが，多くは港湾施設がな
       いのがネック。また人口増加を抑えるため家族計画が導入され，出生率は
       ６０年体の３％代から最近では１％台まで低下している。
   19. 主要語句          イギリスレンポウ 英国連邦/シカツノウキヨウコク 自活農業国
       /トンカリキシ トンガ力士/ニホンジンムラ 日本人村
```

(3) EXPANDコマンドの例

(例4) EXPANDコマンドにより、AREA(地域)フィールドのタームが何種類あるかを示し、各タームについて検索結果集合を作成する。

```
(例4) 1/ FIND NM:*
NM:
100  ターム以上あります。
*      166     1/  NM:( 166  タ ー ム  連 結 )
    2/ EXPAND AREA COUNT="地域あります。" MODE=S ECHO
*       20     2/  AR:アジア
*       45     3/  AR:アフリカ
*        1     4/  AR:ソ連
*       24     5/  AR:ヨーロッパ
*       11     6/  AR:大洋州
*       17     7/  AR:中央アメリカ
*       25     8/  AR:中東
*        8     9/  AR:東欧
*       12    10/  AR:南アメリカ
*        3    11/  AR:北アメリカ

     10 地域あります。
```

(4) プロファイルの実行例

(例5) 地域が南アメリカで、宗教がプロテスタントの国を、FINDコマンドの論理演算により検索し、その国名を表示するようにプロファイル(TEST1)を実行する。漢字の可変パラメタ「地域は?」と「宗教は?」を用いている。

```
(例5) 16/ -TEST1
   次のパラメタを入力して下さい。
   地域は? 南アメリカ
   宗教は? プロテスタント
     16/ FIND AR:南アメリカ AND RL:プロテスタント
*        12    16/  AR:南アメリカ
*        28    17/  RL:プロテスタント
*         2    18/  AR:南アメリカ AND RL:プロテスタント
     19/ DISPLAY 1 FOR ALL
      1. 国名              アルゼンチン
      1. 国名              ブラジル
```

4. 漢字サポート上の考慮点

ORION漢字支援の開発時に考慮した項目を以下に述べる。

(1) データ列の切換えコード

日立漢字情報システムでは、従来の英数カナデータつまりEBCDIK(Extended Binary Coded Decimal Interchange Kana)コードと漢字データつまりKEIS(Kanji Processing Extended Information System)コードを同一データ列上で区別するためのコードとして、コード開始機能キャラクタ(2バイト)を使用している。ORIONでは、この切換コードの値をプログラムでは一切含まず、テーブルファイル(データベース単位)に持っている。これにより、2バイトコードを必要とする外国語サポートを含む将来への見通しがある。

(2) 漢字端末や漢字プリンタへの表示

データ列中にEBCDIKコードとKEISコードが混在していると、それらを区別するための機能キャラクタが必要となる。機能キャラクタは、漢字端末や漢字プリンタ用制御コードであり、表示上の長さに入らないので、プログラム中のデータ領域と出力の際の行幅との間で長さの調整が必要であった。ORIONでは端末を画面モードでなく行モードのみでサポートしている。しかし、これが幸いして漢字端末と漢字センタプリンタを共通仕様で実現できた。

(3) データ定義言語における漢字サポート

DISPLAYコマンドやPRINTコマンドで検索データを表示する際のフィールド見出し、漢字フィールドから索引語抽出するための区切り文字等は、漢字を利用できることが望ましい。これらについては、テーブルファイル作成は何度も行う作業ではないし、このような指定を1度指定した後、ほとんど変更する

ことはない。そこで、データ定義言語内では漢字コードを16進数字として指定する仕様とした。

5. 将来の課題

(1) カナ漢字変換

検索条件として漢字を直接指定できるだけでなく、検索質問をプロファイルに登録あるいは修正、さらにレポート言語を使用してレポートプログラムを作成し、レポートコマンドによりレポートファイルに登録あるいは修正するために、ORION専用エディタ機能を利用する際にも漢字を直接指定することができる。この機能を利用するには漢字鍵盤つきの端末が必要となり、デバイスの制約があり、さらに操作面でも素人には扱いにくい。これを解決する方法として、カナ漢字変換があるが、これには、ソフトウェアつまり変換プログラムによる方式とハードウェアつまり日本語ワードプロセサ(端末)による方式が考えられる。性能及びプログラム開発コストの面から、日本語ワードプロセサを使用する方がよいと考えるが、カナ漢字変換プログラムの利用も含めて検討していく。ともあれ、種々の分野のデータベースにおいて出現する用語は多様であり、カナ漢字変換辞書をいかにして作成するかが当面の大きな問題である。これについても、日本語インデクスタームについては、現状ユーザが文章中から選んでおり、これのみに限れば辞書の作成は可能であり、カナ漢字変換実現の第1段階として検討する。

(2) 日本語文自動索引

日本語文からのインデクスタームの自動抽出については、以前から実験、試作が行われており、特定分野においては、ある程度成果を収めている。当面の課題として、以下の問題があると考える。

(a) 要語辞書、不要語辞書およびこれらの分野別例外辞書の整備の労力。

(b) 校正過程での自動抽出結果の校正と校正結果の抽出方式や辞書への反映。