

テキストを含む文献情報システムの 汎用DBMSによる構築と検討

北海道大学工学部 前 田 隆

本稿では、特定の研究分野における文献情報システムを、提供されているデータベース作成支援システムおよび汎用DBMSの資源を利用して構築した経験について記述している。この中で、機械可読データの作成の問題、作成されたデータ元にして具体的にデータベースを構築するプロセス、構築されたデータベースへのアクセスの問題等について検討されている。データベース構築の段階では、使用した支援システムは汎用型DBMSの運用に必要とされる作業のかなりの部分を自動化することにより、データベース管理者ではない一般の利用者に対して、基本的に有効であった。また、構築されたデータベースへのアクセスに関しては、一般的な検索利用については当然ながら特に問題はないが、複雑なテキスト・サーチ等には対応できないという困難があった。文献情報、より一般的にテキスト情報への有効なアクセスの問題は今後益々重要性を増してくると考えられるので、種々の高度のテキスト・サーチおよびテキスト処理法の開発が望まれる。

CONSTRUCTING A DOCUMENT INFORMATION SYSTEM USING A GENERALIZED DBMS

Takashi MAEDA

Department of Engineering Science, Hokkaido University
Sapporo 060, Japan

In this paper, an experience is described for constructing a document information system for a field of research, using a generalized DBMS under a supporting system for database construction at Hokkaido University Computing Center. The supporting system may be evaluated to be effective even for non-professionals in constructing a database, by automating many processes required for making the DBMS to work. Access method, though restricted to its subsystem for information retrieval in this paper, is practical to common use as a matter of course except for some sophisticated text search queries. Then, it is stressed that advanced effective functions for text search and processing would be requested in even the domain of DBMS for their increasing importance as information search methods for text information in many applications.

1. はじめに

情報科学の分野における慣習に従えば、情報システムには、通常、次の三つのタイプが区別される：データベース管理システム、文献（書誌）検索システム、および質問応答システム [Sa1]。ここで、データベース管理システムはその応用の拡大と共にデータ構造も単純値だけでなく、次第に複合的な構造を持つデータを取り扱うように発展してきており [Pis]、また元々概念的かつ知識型情報を対象としている文献検索及び質問応答システムにおいても、単なるキーワード検索以上の内容的検索への志向が様々な形で展開されてきている [Sa2]。質問応答システムは、その機能的水準を別にすればすべての情報システムはこの範疇に入るものではあるが、ここでの含蓄は問題解決型のシステムという点に基本的性格があり、今日的な表現ではいわゆる推論機能を備えた知識ベースということになる。これらは相互に影響を与え合いながら、扱う対象を構成している種々の情報種の相違と特徴を統合する形での取扱いが要請されている。すなわち、マルチ・メディア・データベースおよび知識型データベースへの指向がその様な傾向を表現している。

一方、こうした一般的、公共的なデータベースの開発・作成とは別に、あるいはこれらのデータベースの利用結果の蓄積とその再利用などを含めて、一定の研究環境における種々の情報を適当なデータベースとして確保し、有効かつ適切に利用したいという要求も根強く存在する [Ma3]。研究室レベルにおける、文献をはじめとする学術情報を手軽にデータベースに投入し、また手軽に利用できるというような要求に見合う形の柔軟な DBMS、およびその使い勝手の良いインターフェースの開発はまだ必ずしも充分とは言えない。他方、DBMSにおける文書処理機能の増強の試み [Sto]、また文書の作成とその内容的なチェックを自然言語などに関するデータベースを用いて支援しようとするアプローチが見られるようになってきた [Hei,Tak]。しかし、文献等の実際のデータの収集からその機械可読化、データベースの構成、有効なアクセス法の確立等は、仮に汎用的なデータベース作成支援システムを利用したとしても、それぞれ大変な努力と時間を必要としているのが現状である [Ma4]。情報システムにおけるデータへのアクセスには基本的に二つのタイプを区別できる：要求が具体的で明確であるもの、例えばある特定の著者の文献などのように検索の結果が要求に対して合致しているかどうか明確なもの。もう一つは要求そのものがあまり明確に表現しにくいもの、例えば主題検索のように結果が要求とどの程度合致しているかを客観化しきたいもの。前者のような要求への処理は特に問題はない。後者のようなアクセスは本質的に主観的であり、試行錯誤的である。これは蓄積しているデータそのものがまた主観性を有する場合は一層複雑となり、文献検索はこの典型的な一例である。これに対して、現段階では、(1) 索引づけされたキーワードないしはキーフレーズによる方法、また更にこれを補助的にシソーラス等により再現率を高める方法、(2) テキスト・サーチによる方法、の二つが代表的である。この中で前者は多くの文献検索において採られている手法であるが、後者のアプローチとその可能性は、システムの利用者が直接興味を抱いた用語ないし文字列そのものを含む対象へのアクセスという点で極めて重要であるといえよう [Sa1]。

本稿では、やや古典的ながら、特定の研究分野における文献情報を対象とし、汎用型データベース管理システムを利用した、文献情報システムの構築の経験におけるいくつかの問題点を検討する。ここでは、単純に筆者らの研究に関連する主要な情報源である文献集合をデータベース化することにより、適切かつ効率的な利用の手段を整備することが目的であり、従ってデータベースの作成者がまた利用者である。一般的に DBMS が想定しているデータベース管理者は専門家であるが、この場合はいわば素人である利用者がその管理者の役割を行うのである。このような状況において、できるだけ既存の計算機環境における資源を有効に利用することにより不必要な労力を避けるために、本学大型計算機センターで提供されている汎用 DBMS による方法が採用された。

データベース作成にあたって、検討する基本点は以下の三点である：

- (1) データベースの作成者が専門的な管理者でない場合、機械可読データからデータベースを構築する際、汎用 DBMS および関連するユーティリティーの利用は充分効率的かどうか、
- (2) 構築されたデータベースへのアクセス手段は、想定される利用の可能性に対して、充分柔軟であるかどうか、
- (3) 以上を踏まえて、今後に望まれる要請点・課題は何か。

以下において、第2章ではデータベースの内容をなす文献集合とその情報構造およびこれを踏まえたシステムの基本設計について述べられ、第3章では本学大型計算機センターで提供されている汎用 DBMS によるデータベース構築の実際について記述される。第4章は作成されたデータベースの、特に文献抄録に対する高度のテキスト・サーチによる情報検索とそ

の意義について議論される。最後に、今後への課題等が第5章においてまとめられる。

2. 文献情報システムの設計

2.1 文献データと情報構造

ここで対象とする文献情報は、第2.1表にまとめられているように、「人工知能および情報検索」(Artificial Intelligence and Information Retrieval)の分野における基本的な文献集合を集積しようとするものであり、これを基礎とする情報システム(Information System)ということで、データベース名はAIRISと名づけられている。文献データを構成する項目としては書誌情報と共に、内容に関する知見を表現している抄録を含めている。

学術文献の一般的な情報構造を一意に定めることは不可能に近いが、ここで取り扱う範囲での典型的な文献情報は次のような項目(属性)により、これを特徴づけることができよう：

文献' := [文献コード、標題*, 著者#(姓、名、住所*), 書誌(雑誌名、巻数、
号数、頁、発行年)、分類コード#, 鍵語#, 抄録*, 言語、所在*,
参考#(参考文献*)]、

ここで、*, #, {} はそれぞれテキスト情報、集合ないしは多値情報、および複合の情報という属性を示している。

2.2 非正規形モデル

前節で特徴づけられた典型的な文献の情報構造は、関係モデルとの関連で、いわゆる非(第一)正規形関係(Non First Normal Form: NF2)モデルの考え方と適合するところが多い[Sch]。すなわち、非正規形の関係は、これを構成する各属性の定義域が単純値だけでなく、集合値や関係そのものをも許すことによって得られる一般的な関係形式であり、次のように定義される：

$$R \subset C_1 \times C_2 \times \dots \times C_n,$$

ここで、 C_i ($i=1, \dots, n$) は基本(属性)定義域の集合 $D = \{D_1, \dots, D_k\}$ のべき集合の要素であり、複合(属性)定義域と呼ばれる。各 C_i , D_j ($j=1, \dots, k$) の要素はそれぞれ単純値、複合値等と呼ばれる。複合値としては、集合、組、(順序つき)リスト、多重集合などを取ることができて、これらにより、テキストやベクトル、マトリクスを含むデータ構造を取り扱うことが出来る。また、この様な複合属性に対応して関係代数も拡張されており、更にNF2に固有のネスト(nest)、アンネスト(unnest)、グループ化(grouping)等の高度の操作も用意されている。NF2モデルはデータが有する意味を自然な形式で表現すると共に、関連した情報を複数の関係に分割しなくてよい、データ(項目)間の階層関係を反映できるなどの利点があり、これらを生かしたデータベースの設計や利用者にはわかりやすい操作やインターフェースを構成できる点も大きい利点である[Kam, Pis]。

この様な形式的取扱いを、前節の文献情報と対応させると、*, #, {} 等の記号を付した部分が複合属性であり、残りが単純属性である。また、データ型としては、*印はリスト、#印は集合、{} は組である。この様な対応と解釈により、NF2モデルで構成されている諸操作に対応するデータ処理を文献情報に対しても適用することが出来る。本稿では、データモデルおよびこれに基づく諸操作の形式化等に関してこれ以上立ち入らないが、理論的な研究として重要である。

2.3 設計方針

上記の議論および文献情報における各属性の性質、また利用する汎用DBMSを含むデータベース作成支援システムにおける制約を考慮し、次の四点がシステム設計の基本方針として採用された。

- (1) データベース作成支援システムの利用： データベースの作成から検索・利用のすべての過程を、基本的には、提供されている計算機資源を運用する立場で行う。
- (2) 文献情報の適切な分割： 基本的には各文献を単位とするレコード構成において生ずる冗長性の排除が目的であるが、次の点と調和を図る。
- (3) データの意味の保存： (2)の分割に伴う各レコードの意味の破壊やデータの分散を避け、データの自然な意味の単位を保持する。
- (4) 柔軟なアクセス機能： 複合的・非定型の情報への柔軟かつ有効なアクセス機能を重視する。

3. 文献データベースの構築

3. 1 入力データとファイル構成

文献データから機械可読データ作成の作業は、文献コピーから編集することなしに直接入力され、各データ項目への分離・写像は別の処理プログラムにより実行された。これらの中で、不定長かつ非定型なデータ項目は標題、著者の住所、および抄録テキスト等である。著者に関する部分は、連名の場合、それぞれ姓、名、住所の各項目が抽出され、これらの組として展開された。また、参考文献は重要な情報項目ではあるが、技術的な理由により、今回のデータ作成作業からは除外された。

2. 3の設計方針に基づき、また後述するように使用したDBMSにおけるレコード長に関する制限もあり、レコードとしては書誌事項と抄録テキスト部分を分離することにした。結局、その他の事情も考慮して、ファイル構成としてはこれらのレコード形式を基にして二個となった。すなわち、書誌事項および抄録テキストの二つのファイルであり、これらは文献コードにより結合されている。尚、抄録のない文献が若干あったため、結果的に自然な分割であった。2. 1の「文献」に対して、重要項目についてのキーワード検索が可能のように必要な属性（フィールド）を追加し、結局以下の様に構成された：

文献：= [文献コード、標題*、標題キー#、著者#（姓、名、住所、住所キー#）、
書誌（雑誌名、雑誌名キー#、巻数、号数、頁、発行年）、分類コード#、
鍵語#、所在*、所在キー#]

抄録：= [文献コード、抄録*]。

3. 2 汎用DBMS (ADABAS) の利用

本学大型計算機センターで提供されているデータベース作成支援システムの下では、データベースの構築支援を可能とする汎用型のシステムとして、情報検索型のORIONとデータベース管理型のADABASの二つがある[Ama, Moc]。ADABASにはサブシステムとして、汎用型の情報検索システムであるSOARが含まれている。AIRISデータベースは基本的には情報検索システムではあるが、抄録テキストを含む情報の探索や処理などの高度利用、またインターフェースに関してもいづれ検討する計画があるので、より柔軟な操作言語を有するADABASシステムを使用することにした。しかし、現時点ではこの中のSOARサブシステムを利用してのデータベース構築と検索利用の範囲に議論を絞ることにする。

3. 1で述べた「文献」、「抄録」の各データをADABASデータとして実現するために、次のような対応により形式的な調節を行った：すなわち、#印のついた集合値はADABASのマルチプル、*印のついたテキスト情報は適当な長さには区切り、これらのリスト（実際はマルチプルとしている）、および著者項目はピリオディック、書誌事項はグループ、等の各フィールドとしてフォーマット間の対応を取っている。この中で、特にテキストについてはフィールド値の上限が253バイトに制限されているための措置であり、またレコード全体としても現時点で約7200バイトという制約がある。

SOARシステムはそれ自身のユーティリティとして、機械可読データからADABASデータベースへの変換を行うSOASTA、およびキーワード自動抽出機能を実現するSOARKW等を有しており、これらの一連のプロセスは、基本

的に、支援システムにおいて主としてパラメータ指定により、順次メニューに従って誘導的に実行していくように構成されている。第3.1図はこれらのプロセスのうち、セグメント形式の書誌データのADABASデータベースへの変換の実行例の一部を示している。図において、(1)の部分でSOARシステムが保有するユーティリティの選択が行われている。この段階で、上述のSOASTA、SOARKW、あるいはSOARのいずれかを選択する。ここで、SOARは会話型コマンド言語であり、次の章で議論する。(2)の部分は文献データのADABASデータベースにおけるフォーマットであり、左から二列目のID、TI等は書誌情報の各属性に対応するフィールド名である。3.1で追加された標題キーなどのフィールドは対応する各項目から、SOARKWを用いて索引づけされている。データへのアクセスにはこれらのフィールド名を用いなければならない。

4. 文献データへのアクセス

4.1 利用者言語

AIRISはADABASデータベースとして作成されているので、上述のSOARコマンド言語をはじめとして、ADABASデータ操作言語(DML)、サブシステムであるNATURAL言語による方法、および親言語としてのCOBOL、PL/I等を用いたプログラムによる方法など種々のアクセス法が可能である。

本稿では、独自のインターフェースについての検討はせず、提供されている範囲での資源によるデータアクセスの可能性を検討するものとする。その能力を越える処理については他の方法を考慮するとして、会話型コマンド言語SOARによって種々の検索を試みる。SOARはキー項目(デスクリプタ指定)の転置索引による高速検索と検索結果の集合操作にその特徴を持つ[S01]。コマンドには検索や操作による集合の作成、集合からのデータの読み込み(出力)、データの更新、集合の保管、利用状態などの表示、ファイルその他の制御等が含まれており、通常の文献検索等における利用に関する機能は基本的に具備されていると考えられる。但し、利用者の問合わせ(質問式など)の内容を保管するコマンドがないのは不便である。

4.2 文献データの検索とテキスト探索

SOAR会話コマンド言語[S02]は詳細な検索のための種々のデータ照合方式を備えており、また検索結果の表示に対する豊富な編集機能などを有しているため、想定している範囲でのきめの細かいデータの探索とその獲得を可能としている。例えば、データの照合については、必要ならば、(1)照合時に文字コードの変換、指定された句切り文字、無視文字などを利用することができ、かつ照合するべき長さを指定して、実行することが可能であり、(2)照合すべき文字列の表現として、前方・後方の切断の指定、位置演算子として文字の内容/存在を問わないという、いわゆるワイルドカードの指定も可能であり、更に、(3)以上の指定を含んだ語(文字列)に対して、語と語の距離を照合条件に含めることができる。すなわち、テキストなどの語の列に対して、照合条件として複数の語とそれらの間の許容距離(現在、7語以内)を指定することにより、文脈的な探索を可能にすることを意図しているものである。

以上の照合方式を用いて、種々の探索の要求をコマンドを利用してどの様の実現できるかについて例題を通じて検討する

- (1) 著者名がSaltonであり、キーワードにvectorかmodelを含む文献の標題と書誌事項を求めよ：これは通常のキーワード検索である。

```
¥FIND @EX-1 TK = VECTOR, MODEL * LN = Salton.  
¥READ EX-1 TI1-3% / JO, UL, NO, YR, PG.
```

- (2) 標題にDATA BASE SYSTEMという語句を(この並びで)含む文献：

a. ¥FIND @EX-2 TK = DATA * TK = BASE * TK = SYSTEM.

2. 1表 収録文献の内容

雑誌・会議録名	収録年	文献数
AI Journal	1970-1984	280
Proc. of IJCAI	1969-1984	1324*
Proc. of AAAI	1980-1984	360
小計		1964
Inform. Process. & Manage. (Inform. Stor. Retr.)	1963-1984	641
J. Of. ASIS	1970-1984	798
The J. of Docum.	1970-1984	289
小計		1728
合計		3692 (2082)*

HADASOAR

```
>>PLEASE SELECT FUNCTION AS 'SOASTA' 'SOARKW' OR 'SOAR'
> FUNCTION ? .....SOASTA ..... (1)
EX 'SYSS.XDBA.CLIST(SOASTA)'
```

```
>>PLEASE ENTER FILENAME ( 3 - 8 CHAR )
> FNAME ? ...
&
& DICTIONARY UTILITY<UTIL03> STARTED AT 16:37:51 ON 86-
&
```

UTILITY RUN SUCCESSFUL

```
.....
: STEP 01. UTIL03 ENDED :
: CONDITION CODE = 0 :
: .....
```

```
>>PLEASE ENTER INPUT DSN IN FULL NAME
>
>>+ + + IF ERROR OCCURD IN UTIL03, ENTER EMPTY RETURN + + + <<
> INPUTDSN ? ...
```

```
>>PLEASE ENTER KARTE DSN IN FULL NAME
>
> KARTE ? ...KARTE
```

```
>>PLEASE ENTER SYSOUT-CLASS OF SOASTA LIST
>> OUTCLASS(A/K/L/H/D/T) ?...
SOASTA SEGMENT RECID=(01,1),FLDID=(2,3),SUBFLDID=(74,1),OCCUR=(76,2),
SOASTA VALUE=(6,68),LRECL=80,BLKSIZE=6160,SEQ=(78,3)
01.ID.008.A.MU.DE IDE
01.TI.068.A.MU.NU TIT
01.TK.020.A.MU.NU.DE TIK
01.AU.PE AUT
02.LH.015.A.MU.DE AUT-L
02.FN.015.A.MU.DE AUT-F
02.AD.068.A.MU.NU AUT-A
02.OK.020.A.MU.NU.DE ADK
01.BI BIB ..... (2)
02.JO.068.A.MU JOU
02.JK.020.A.MU.NU.DE JOK
02.VL.006.A.MU.DE VOL
02.NO.006.A.MU.DE NUM
02.YR.004.A.MU.DE YEA
02.PG.010.A.MU.DE PAG
01.CL.010.A.MU.DE CLA
01.KW.020.A.MU.NU.DE KEY
01.LO.068.A.MU LDC
01.LK.020.A.MU.NU.DE LDK
```

INPUT = 00004245 OUTPUT = 00000290

NORMAL END

```
& SOASTA JOB ENDED AT 16:38:58 ON 86-
& INPUT DSN=' '
& SOAOUT DSN=' '
.....
```

ADAMAH NORMAL END

```
.....
: STEP 02. ADAMAH ENDED :
: CONDITION CODE = 0 :
: .....
```

3. 1図 ADABASデータベース作成過程 (一部)

この場合、語順は無視されるので意図した検索になるとは限らない。FIND文ではこのような検索はできず、次のようなSEARCH文となる。

b. ¥SEARCH @EX-2 BIBL TI, 1D = DATA/<+1>/BASE/<+1>/SYSTEM.

- (3) 抄録テキストにおいて DATA BASE, DATA-BASE, DATABASE のいずれかを含み、その後方にQUERY (語尾変化を許す) を含む文献:

¥SEARCH @EX-3 BIBL AB, 1D = DATA/<+1>/BASE/<+7>/QUER/.

この場合、(+7)の部分はBASEと QUER/の語間距離が7以内であるものを検索する。現時点で7は上限値である。

- (4) 抄録テキストにおいて、ある語の集合P = {presents, consider, describe, ...} の要素のどれかを含み、同時にmodel, structureを含む文献:

この場合、集合Pをあらかじめ与えておいて検索する操作はできないが、定められている場合はその要素をすべて書き出すことにより、一応可能である。

- (5) (4)の検索の範囲を同一文中、あるいは部分テキストに制限した検索: 範囲の制限は、現状ではできない。

- (6) (4), (5) で集合Pが語句の集合である場合: (4)ができれば、一応可能。

- (7) (6)のような操作を複数個の集合に対して、また複数の部分テキストに対して検索する、等々: 基本的に困難。

これらの例あるいは検索要求は、通常の情報検索システムにおいては、それほど一般的でないかも知れないが、与えられた情報資源の範囲において可能なかぎり内容に立ち入った探索を行おうと考えると、それほど不自然な要求ではないであろう。特に、後半のいくつかは学術文献の抄録という比較的様式の定まったテキストに対する一つのアクセス法として有効であると考えられる。また、もしそうであるならば、逆にそれはテキスト情報に対する一種の構造的な分析法あるいは索引法の開発に応用できるといえよう [Ma2]。

5. おわりに

本稿では、特定の研究分野における文献情報システムを、提供されているデータベース作成支援システムおよび汎用DBMSの資源を利用して構築した経験について述べた。初期のデータ入力、すなわち機械可読データの作成には種々の実際的な問題があり、機械的な人力法の進歩が望まれる。次に、このデータを元にして具体的にデータベースを構築するプロセスに対して、支援システムは基本的に有効に稼働したといえる。しかし、支援システムが汎用DBMSの運用に必要とされる膨大な作業のかなりの部分を自動化してくれたとしても、データベース管理者ではない一般の利用者が使用する場合、やはり他の計算機利用と同等かそれ以上の労力と時間を要するものであるといえよう。

また、構築されたデータベースへのアクセスに関しては、一般的な利用法の場合は当然ながら特に問題はないが、別のアプローチをしようとするとなかなか難しい。文献情報、あるいはより一般的にテキスト情報への有効なアクセスの問題は今後益々重要性を増してくると考えられるので、4. 2で扱ったようなものを含めて、種々の高度のテキスト・サーチおよびテキスト処理法の開発が望まれる。

参考文献

- [Ama] Amano, K., T. Maeda: Database Management in Research Environment, Proc. 3rd Symp. on EFISS, 1985.
- [Hei] Heidorn, G.E., et al: The EPISTLE Text-Critiquing System, IBM Syst. J., 21(3), 1982.
- [Kam] 上林弥彦他: 関係データベースにおける意味制約を反映した非正規形の関係の設計問題、情報処理学会論文誌、第24巻第6号、昭和58年。
- [Ma1] Maeda, T., K. Amano: Document Information System in Scientific Research, Proc. 3rd Symp. on EFISS, 1985.
- [Ma2] Maeda, T.: An Approach toward Functional Text Structure Analysis of Scientific and Technical Document, Inform. Process. & Manage., 17(6), 1981.
- [Ma3] 前田 隆: 学術研究における文書情報システムの構成、北大工学部研究報告、第130号、昭和61年。
- [Ma4] 前田 隆: SOARによる情報学文献情報データベースAIRISの構築、北大大型計算機センター・ニュース、17(4), 1985.
- [Moc] 持田明野他: 北海道大学大型計算機センターにおけるデータベース運用管理システム、北大大型計算機センター・テクニカル・レポート、No.6, 1983.
- [Pis] Pistor, P., F. Anderson: Designing a Generalized NF2 Model with an SQL-type Language Interface, Proc. 12th Intl. Conf. VLDB, 1986.
- [Sa1] Salton, G.: Another Look at Automatic Text-Retrieval Systems, Comm. ACM, 29(7), 1986.
- [Sa2] Salton, G.: Some Characteristics of Future Information Systems, SIGIR Forum, 18(2-4), 1985.
- [Sch] Schek, H.-J, P. Pistor: Data Structure for an Integrated Data Base Management and Information Retrieval System, Proc. 8th Intl. Conf. VLDB, 1982.
- [S01] SOAR概説書、第一版第一刷、(株)ソフトウェア・エージェンシー、1984.
- [S02] SOARコマンド・リファレンス・マニュアル、第一版第一刷、(株)ソフトウェア・エージェンシー、1984.
- [Sto] Stonebraker, M., et al: Document Processing in a Relational Database System, ACM Trans. Office Inform. Syst., 1(2), 1983.
- [Tak] 武田浩一: 文書データベースシステムへのアプローチ、電子通信学会データ工学研究会資料、DE86-10, 1986.