



Title	学術文献情報データベースのための知的入力システム
Author(s)	斉藤, たつき; Saito, Tatsuki
Citation	北海道大學工学部研究報告, 140, 101-108
Issue Date	1988-05-30
Doc URL	<a href="https://hdl.handle.net/2115/42091">https://hdl.handle.net/2115/42091</a>
Type	departmental bulletin paper
File Information	140_101-108.pdf



## 学術文献情報データベースのための知的入力システム

齊藤たつき

(昭和 62 年 12 月 26 日受理)

### **An Intelligent Input System for Scientific Article Information Data Base System**

Tatsuki SAITO

(Received December 26, 1987)

#### **Abstract**

It is described to develop a scientific article information database system for an highly intellectual utilization of scientific information in this paper. In particular, an importance of an information input system is discussed in detail. Two input system were designed and implemented. The first is an interactive type. The second is an intellectual batch type. The latter input system enables to extract every field from standardized bibliographic format by production system. Those input systems were implemented by NATURAL on DBMS ADABAS. In order to aim at efficient input processing or in order to avoid misinput, a common feature of both information input systems is to input another field data after retrieving an author name of a registered article or after retrieving a title of a registered article. There is a fair prospect to realize a completely automatic data input system for scientific article information by using an printed character type OCR.

#### 1. はじめに

学術文献情報のもつ種々の関係を構造化しその関係から、より高度な情報を再生産することを目的としたシステムを構築するための学術情報データベースシステム ANGEL を設計し製作した。ANGEL は当初、CAD (Computer Aided Design)/CAM (Computer Aided Manufacturing)関連の自由曲面創成理論に関する文献データベースとして開発したものであるが、汎用の学術文献データベースシステムとして利用できるように設計してあるので、他の文献データベースシステムとしても利用可能である。ANGEL は現在、全国共同利用施設である北海道大学大型計算機センターで一般に公開、使用されている。本報告ではデータベースに文献情報を入力する作業をできるだけ機械化することを目標にした知的入力システムを中心に学術情報文献データベースシステムについて述べる。なお、ANGEL は英語文献を処理対象にしている。

#### 2. 学術情報文献データベースシステム

ここで学術文献情報データベースシステム ANGEL の基本概念について述べる。学術文献情報データベースの有効利用を考えたとき、単に書誌情報の利用にとどまらずより高次の情報をそこ

から再生産することができるのではないかというのがこの研究の動機である。すなわち、研究動向を把握するにはいかなるシステムが必要なのか、そして研究の本質を明らかにするにはどのような情報を手がかりにして追究したらよいのかということを検討した。そこで学術文献間に存在するいろいろな関係を着目した。そして、文献を点、文献間の関係を線とするグラフとして表現しそのグラフの性質を明かにする方法論を開発することが目標である。関係にはいろいろある。たとえば、文献間の引用関係があるが、これはその文献の著者がその研究に関係あるものとして挙げたものであり重要な関係情報の一つである。そして、この引用関係の情報はその分野での研究相互間の関連性を把握するうえで重要な情報の一つである。そのため ANGEL では文献の引用、被引用関係は重要な情報の一つであるのでデータ入力時に指定するようになっている。キーワード関係も文献の内容、テーマ等の共通性を知るための情報として必要なものの一つである。すなわち二つの文献間の共通キーワード数が多いほど二つの文献は共通したテーマを扱っているとみなせるからである。また、文献間に著者の重なりがあれば、同一または、似た内容を扱った文献と解釈してもよからう。こうした文献間の情報はその文献を解釈または評価する人に影響されない客観的な情報量である。本データベースシステムを利用してこれらの複数の情報から文献間の関係グラフを作成し、こうしたグラフよりいろいろな関係行列をつくる。これらの行列は実数値の類似度行列であり、目的に応じた数学的基準を設定して最終的に一つの類似度行列にしたものをクラスタリングし、その特徴を明らかにするシステムの構築を目指している。そのため、これらの情報をいかに効率的に入力するかに重点を置いてシステムを設計した。

ANGEL は全体の構成としては、情報収集サブシステム、情報検索サブシステム、情報蓄積サブシステム、表示サブシステム等からなっている。それらの構成要素である各モジュールの中には共用しているものもある。なお、リレーショナル型 DBMS である ADABAS 上に NATURAL でインプリメントした。

### 3. 文献情報入力システム

#### 3.1 学術文献情報の収集

学術文献情報の収集の仕方として、データベースに入力する方式に、市販のデータベースを購入し、その決められたフォーマットにしたがって機械的にデータ項目を抽出しデータとして使用する方式があり一般的に用いられている。文献そのものから印刷文字を自動的に読み取った後、手動で各項目を切り出す方式もある。項目の自動抽出が可能になり後者の方式が実用化されるまでは、やはりマニュアル入力に頼らざるを得ないのが現状である。研究者データベースではできるだけ詳細な情報を必要とする。そのため前者の方式のように、流通している文献データベースのみからのデータでは十分な情報が得られないということもあり、オリジナルペーパーから直接、必要な情報を入力することが必須となる。しかし、研究者にとってデータ入力に多くのコストをかけることはなかなかできないのが実情である。研究者みずからデータを入力するか、あるいはせいぜいアルバイトに頼ることになる。そのため、われわれのような非職業的キーオペレータがデータ入力作業にあたって効率的に入力できるようなシステムの必要性が生ずる。こうしたことからデータ入力にあたってキーを打つ回数をできるだけ少なくして操作を効率化する必要がある。データベースに登録済みの情報を利用する2つのシステムを作成した。1番目のものが対話型入力システムで、2番目のものが一括型知的入力システムである。

#### 3.2 対話型入力システム

ここでは一括型知的入力システムとも共通する文献情報入力システムの基本設計概念と、対話

型入力システムに特有な特徴について述べる。設計に際してつぎの点を重視した。

- 1 入力作業効率の向上。
- 2 誤入力率の低下。
- 3 登録済みのデータの有効利用。

1は職業的キーオペレータでない利用者にとって特に効果が期待できることである。2は職業的キーオペレータでさえ入力ミスが3%程度あるといわれているので、大量のデータを入力する場合は、如何に誤りが少ない方式をとるかが重要な課題となる。こうしたことを踏まえて3の結論に達した。

研究者用学術文献情報データベースの入力処理あるいは、検索処理を考察すると、文献間の種々の関係を扱うために、親文献も引用文献である子文献もともに重要であり見かけ上は同等に扱わなければならない。そのため類似のテーマをもった研究の引用文献として同じ文献が何度も出てくるので登録文献数の増加にともない、同一著者の文献出現頻度が高くなる傾向が顕著である。そのため引用文献の場合も含めて、入力データと登録済みデータとの重複の有無をチェックする必要が生ずる。そして、重複チェックは第一著者のファミリーネームを第一検索キーとするのが信頼性および能率の面で有効である。検索項目としては、著者名、標題、文献番号、文献ID、分類番号、雑誌名、巻、号、キーワード、とした。必要に応じて、ファーストネームやミドルネームあるいは、標題、その他の書誌情報との論理積をとって検索条件にすることもできるようにしてある。こうしておくことで簡単な検索処理にもそのまま使用できる。著者のフルネーム、所属等はそれらの情報が正確に登録された文献のみ格納し、同一著者の新規の文献を登録する場合は、それらの情報は実際は格納しないで、正確なデータの格納されている参照先の文献IDのみ格納することもできる。こうすることによって、データの重複を防止することができ、記憶領域の無駄を省くと同時に、データ間の整合性を確立できる。むしろ、システムとしては検索画面で必要なデータが表示された時点で登録モードに移行できるようにしてあるので、登録済みの必要なデータを取り込んで新規に登録可能である。なお、所属が途中で変わった場合は、その時点で新規に登録することもできる。ところで、対話入力中は画面を送った後にミスに気付くこともあるので、データベースに記録する前に修正できるように画面の後退ができるようにした。

### 3.3 対話型入力システムの詳細

見易く、かつ入力ミスを低く抑えるために入力画面はメニュー方式とし、一度入力した情報を保存しできるだけ再利用するようにした。画面設計にあたってNATURALのスクリーン・エディタのMAP機能を活用した。これは自由に画面のレイアウトができる便利な機能である。図1にANGEL起動画面を示す。ここでアクセスするファイル名、パスワード、処理選択のパラメータを入力する。この起動画面に続いて親文献かどうかを入力する画面が表示され、親文献の場合は“Y”を、そうでない場合は“送信”のみを押すと図2の検索画面が表示される。ここでは検索項目として著者名、標題を入力する。引用文献の項目に同一文献が多数回出現するので、

(1)同一データの重複登録を回避する。

(2)登録済みのデータの誤りのチェックする。

などのためにまず検索処理を行う。検索条件は著者名のSURNAME(姓)のみ、またはフルネーム、標題のみ、あるいはそれらの論理積をとったものでもよい。標題は著者名に比較して長いことと、引用の場合不正確なことがあるので入力しない方が効率的である。この画面で必要事項を入力後“送信”を押すと指定した条件に従って検索処理をする。該当文献が既に登録済みであれば図3のような既登録データの内容が表示される。同一著者のデータが複数ある場合は“送信”

```

          ##          ##          ##          #####          #####          ##
          ###         ##          ##          ##          ##          ##
          ##          ##          ##          ##          ##          ##
WELCOME TO DATABASE ##          #          ##          #####          ##
          #####          ##          #          ##          ##          ##
          ##          #          ##          ##          ##          ##
          ##          #          ##          ##          ##          ##
          ##          #          ##          ##          ##          ##
*****
*                                     [MPSU20Y] *
*      FILE-NAME,  PASSWORD AND  PROCESS-SELECTION INPUT MENU *
*
*      USER-ID :  A10118 *
*-----*
* |                                     | *
* |   WHAT ARE FILE-NAME              (+FILENAME) CAD          | *
* |   AND PASSWORD TO BE ACCESSED ? (+PASSWD)                 | *
* | SELECT PROCESS (+SELPRO1) ( SEARCH, REGISTER, UPDATE=     , LIST=L, END=E) | *
* |-----*
*
*****

```

図1 ANGEL 起動画面

```

*****
*                                     [MPSR5-10] *
* RETRIEVAL MENUE OF REGISTERED ARTICLES FOR FILE : CAD *
*
*
*      LAST LOGON-UID : X10044      DATE :          14:03:16.6 *
*      LAST ARTNO = 12 *
*-----*
* ( RETRIEVAL CONDITION : OR==> , AND==>A, CONTINUE==>C, QUIT==>Q ) *
*-----*
*      (SURNAME)              (FIRST, MIDDLE NAME) *
* ( ) AUTHOR: COONS *
* *
* ( ) TITLE: *
* *
* *
*****

```

図2 検索画面

```

*****
THE FOLLOWING ARTICLE ALREADY REGISTERED !! [MPDS5E]
*****
ARTICLE-NO 1          ARTICLE-ID COONS(1974)
=====
TITLE :  SUREACE PATCHES AND B-SPLINE CURVES
-----
AUTHOR 1 COONS          , S. A.          AFFIL. NO. 01
AFFIL. NO. 1 : AFFILIATION SYRACUSE UNIV.
=====
JOURNAL COMPUTER AIDED GEOMETRIC DESIGN
VOLUME          NUMBER
PAGE-BEGIN 1          PAGE-END 16
YEAR 1974          MONTH 0
-----
PUBLISHER ACADEMIC PRESS
CLASS-NO
( NEXT ARTICLE ==> , NEW ARTICLE TO BE REGISTERED BY THE SAME AUTHOR ==>N,
  UPDATE OR DISPLAY OF CURRENT ARTICLE==>U, DELETE==>D, QUIT==>Q )

```

図3 既登録データ

を押すと次々に該当文献データを表示する。表示途中で誤りを見つけたら、“U”をキーインして正しいデータを入力できる。検索中の著者の新しい文献を登録する場合は、“N”をキーインすると既に入力したデータはそのまま生き、未入力事項のデータのみをキーインすればよい。このようにデータ入力の際に検索処理をすることは、データベース作成の初期段階では効果は少ないが、データ量が多くなる実用期ではその効果が期待できる。特に研究者用のデータベースでは、比較的限られた範囲の文献を扱う性質上、同一データが重複して出現することになり、分野にもよる

が同じ著者が何度も出てくる場合が少なくない。そのため入力システムの著者検索処理が効果的である。なお、著者フィールドはキー指定することによってインバーテッドファイルが作成されるので、登録件数の増加に伴って検索速度が極端に悪くなるようなことは起きない。図2の検索画面で、最初のパラメータに“C”を指定すると、図4のような検索画面の続きが表示されるので、文献番号(複数の場合は、開始値、終了値を指定する)、文献ID、分類番号、雑誌名、巻、号、キーワード等を入力すればそれらの検索処理を行う。左端の( )に“A”を指定するとその項目との論理積をとったものを検索条件にセットする。

```
*****
*                                     [MPSR5-11]*
* RETRIEVAL MENUE OF REGISTERED ARTICLES FOR FILE : CAD
* *****
* LAST LOGON-UID : X10044      DATE : 88-01-05 14:03:16.6
* LAST ARTNO = 12
* =====
* ( RETRIEVAL CONDITION : OR=> .AND=>A )
* .....
* ( ) ARTICLE-NO = 1          - 20
* ( ) ARTICLE-ID =
* ( ) CLASS-NO =
* ( ) JOURNAL:
*
* ( ) VOLUME:
* ( ) NUMBER:
* ( ) KEYWORD:
* *****
```

図4 検索画面

### 3.4 一括型知的入力システム

3.3で述べた対話型入力方式は大型コンピュータでしか稼働しないADABASを起動してオンラインTSSで対話的に入力するために、システムのサービスタイムしか利用できないという難点がある。また、小型のコンピュータ・システムをできるだけ利用してデータを作成したいということもあり、デリミッタによって各項目を区切ったソース・データをあらかじめオフラインで作成しておき、それらを識別して各項目を抽出する入力システムについてつぎに述べる。このシステムはある標準的な雑誌の引用文献の記法に従って書かれた書誌事項であれば、そのまま自動的に項目の抽出が可能である。すなわち、つぎの要件を満たしている記法のものであれば処理可能である。

(1) AUTHOR/AFFILIATION: “TITLE”, JOURNAL, VOLUME, NUMBER, PAGE-PAGE, (YEAR-MONTH), PUBLISHER, 「KEYWORD, # CLASS NUMBER, \* REMARKS, (CONTENT)の順になっていること。

(2) “TITLE”以外の項目は省略されていても構わない。

以上の程度のデリミッタの付加であれば、多少の前処理によって印刷文字を読み取って自動的に項目を抽出するシステムが実現可能なものとなろう。この入力システムはプロダクション・システムを用いてNATURALでインプリメントした。プロダクション・システムはRule部, Data Base部, PSI (Production System Interpreter)の3つの部分からできている。

1) ルール部は順序対 LHS → RHS で表現される。

ここでLHS (Left Hand Side)は左側規則, 左辺, あるいは条件部と呼ばれ, RHS (Right Hand Side)は右側規則, 右辺, あるいは実行部と呼ばれる。これはつぎのようにも表現できる。IF{(LHS)=T} (→) THEN DO {(RHS)}

すなわち、条件部 LHS が真であれば、実行部 RHS を実行するという意味である。

2) データベース部は処理対象のストリングを蓄えておく部分で、Working Memory と呼ばれることもある。ルールの中で LHS を真にするものがあれば、そのときの RHS によって当該ストリングが書換えられて別の状態を作ることから、Production System (生成システム) と呼ばれる。

3) PSI (ルール適用部) は PS 全体をコントロールしている部分で、システム内部に蓄えられているルールとデータベース内の処理対象ストリングとのマッチングをとり、LHS が真なら RHS を適用する。一般的にルールは複数存在するので、LHS が偽であればつぎのルールを調べる。このような繰り返し処理を recognize-act cycle (認知-実行サイクル) という。

なお、上で述べたタイプの PS を antecedent-driven (前提部駆動型) PS あるいは forward chaining (前向き) PS といい、RHS が実行されるための LHS を求めるタイプの PS を consequent-driven (結果部駆動型) PS あるいは backward chaining (後向き) PS という。本システムは前提部駆動型 PS である。以下に本システムの PS の一部として、著者名および所属、標題の抽出のルールおよび文献 ID の自動付番のルールと実際の処理例を示す。

[著者名, 所属] の抽出

R1 LHS1 : が存在する。RHS1 その左側を著者情報データとする。

R2 LHS2 : が存在しない。RHS2 著者名を NONE とする。

R3 LHS3 : 著者情報データに / が存在する。RHS3 その左側を著者名とし、右側を所属名とする。

R4 LHS4 : 著者名に, が存在する。RHS4 その左側を姓とし、右側を名とする。

図 5 に処理例を示す。

```
J.J. FLORENTIN/DEP. OF COMPUTER SCIENCE:A.J. SAMMES/PROJECT MANAGEMENT WAVELL:
"SYSTEMS WITH STATE TE-SET"
CJ,18-2,135-139.(1975-5)
(O A..1 I..2 FINITE-STATE MODELS OF PARTS OFCOMPUTER SYSTEMS.2.1 TRANSITION ERRO
RS IN FINITE-STATE MODELS.2.2 STATE RESETTING IN FINITE-STATE MODELS.2.3 THEORET
ICAL SYNTHESIS OF FULLY SYNCHRONISING FINITE-STATE MODELS.3 APPLICATION OF STATE
RESETTING,3.1 THE WIPDOS SYSTEM,3.2 FORWARD SUCCESSOR TREE FOR WIPDOS)
*S1
@
```

```
=====
      DISPLAY SCREEN OF EXTRACTED DATA READ                                [MPST1-0A]
LAST LOGON-UID: X10044      DATE: 87-12-25 TIME: 14:03:16.6 ARTNOLST: 12
=====
ARTICLE-NO 13
-----
              (SURNAME)                      (FIRST,MIDDLE NAME)
AUTHOR 1  FLORENTIN                          .      J. J.          AFFIL. NO.  01
AUTHOR 2  SAMMES                             .      A. J.          AFFIL. NO.  02
AUTHOR 3  .                                  .              AFFIL. NO.
AUTHOR 4  .                                  .              AFFIL. NO.
AUTHOR 5  .                                  .              AFFIL. NO.

AFFIL. NO. 1 : AFFILIATION  DEP.OF COMPUTER SCIENCE
AFFIL. NO. 2 : AFFILIATION  PROJECT MANAGEMENT WAVELL
```

図 5 「著者名, 所属」の処理例

[標題] の抽出

R5 LHS5": が 2 個存在する。RHS5 その中に囲まれた文字列を標題とする。

R6 LHS6": が存在しない。RHS6 NONE TITLE の警告を出す。

図 6 に処理例を示す。

[文献番号, 文献 ID] の付番

```

*****
*                                     [MPSR5-13]*
* RETRIEVAL MENUE OF REGISTERED ARTICLES FOR FILE : CAD
*
*****
* LAST LOGON-UID : X10044      DATE : 87-12-25 14:03:16.6
* ARTNO = 13                LAST ARTNO = 12
* ARTID =
*-----*
* ( RETRIEVAL CONDITION : OR=> ,AND=>A,CONTINUE=>C,QUIT=>Q,REPLACE=>P )
*-----*
* (SURNAME) (FIRST,MIDDLE NAME)
* ( ) AUTHOR: FLORENTIN , J.J.
*
* ( A ) TITLE: SYSTEMS WITH STATE TE-SET
*-----*

```

図6 「標題」の処理例

- R7 LHS7 文献番号が既登録でない。RHS7 最終文献番号に1を加える。
- R8 LHS8 発行年が存在する。RHS8 文献ID=SURNAME (YEAR)とする。
- R9 LHS9 発行年が存在しない。RHS9 文献ID=SURNAMEとする。
- R10 LHS10 文献IDが既登録のものと同じとする。RHS10 文献ID重複の警告を出す。図7に処理例を示す。

```

*****
THE FOLLOWING ARTICLE ALREADY REGISTERED AS BELOW !! [MPDS5E3]
*****
NEW ARTICLE-NO: 14                NEW ARTICLE-ID: FLORENTIN(1975)
OLD ARTICLE-NO: 13                OLD ARTICLE-ID: FLORENTIN(1975)
=====
N.TITLE: SYSTEMS WITH STATE TE-SET
O.TITLE: SYSTEMS WITH STATE TE-SET
(NEW)
(OLD)
-----
NEW AUTHOR 1: FLORENTIN , J.J.      AFFIL. NO. 1
OLD AUTHOR 1: FLORENTIN , J.J.      AFFIL. NO. 1
OLD AFFIL. NO. 1: DEP.OF COMPUTER SCIENCE
=====
O. JOURNAL: CJ
O. VOLUME: 18-2                O. NUMBER:
O. P-BEGIN: 135                O. P-END: 139
O. YEAR: 1975                  O. MONTH: 5
-----
( NEXT TO READ WORK FILE=> .END=>E,NEW ARTICLE TO BE REGISTERED BY THE SAME
  AUTHOR=>N,UPDATE OR DISPLAY OF REGISTERED ONE=>U,DELETE=>D,QUIT=>Q )

```

図7 「文献番号, 文献ID」の処理例

## 4. 結 論

一括型の入力システムはデータミスによる項目の誤抽出を防止するために各項目の抽出処理毎にその結果を画面に表示させて内容を確認しながら作業を進行するようにした。デリミッタの不適切な設定のため、あるいはデータそのものが最初から間違っていたため、データが正しく切り出されない場合は、その場で直ちに修正できるように配慮したため、この方が完全にバッチ処理をした後でチェックするよりも効率的と考えられるからである。本入力システムの方式は両方とも自動検索処理によって登録済みのデータと検証しながら登録処理をするものである。そして、本研究の結果をつぎのように要約できる。

1. プロダクション・システムによる書誌事項の抽出機能の実現。
2. 印刷文字 OCR との結合による完全自動化の可能性。
3. 入力作業負担の軽減化。

4. 小型コンピュータシステムでの前処理によるデータ入力コストの低減化。

## 5. おわりに

学術情報システムが本格的に利用できる環境が整備されつつあるが、書誌情報の検索利用は無論のこと、その利用をより高度なものにする意味でも本研究で述べたシステムも含めた知的利用方法の開発が急がれるところである。データベースシステムにおける最大のボトルネックはデータの収集と入力だといわれている。このことは知識処理システムにおいてもいえることなので、このような研究はますます重要になるものと考えられる。

## 参考文献

- 1) 斉藤たつき 他：引用情報の関連グラフによる研究動向把握の方法論開発とこれを用いた学術情報検索の効率化に関する研究，特定研究「情報システムの形成過程と学術情報の組織化」報告書，(昭 52)
- 2) 斉藤たつき 他：引用情報の関連グラフによる研究動向把握の方法論開発とこれを用いた学術情報検索の効率化に関する研究，特定研究「情報システムの形成過程と学術情報の組織化」報告書，(昭 53)
- 3) 斉藤たつき 他：引用情報の関連グラフによる研究動向把握の方法論開発とこれを用いた学術情報検索，特定研究「情報システムの形成過程と学術情報の組織化」報告書，(昭 54)
- 4) 斉藤たつき 他：引用情報の関連グラフによる研究動向把握の方法論開発とこれを用いた学術情報検索，特定研究「情報システムの形成過程と学術情報の組織化」総合報告書，(昭 54)
- 5) 斉藤たつき：学術文献情報知識ベースのための入力システムの設計，第 13 回日本動計量学会総会発表論文抄録集，(昭 60)
- 6) 斉藤たつき：学術文献用知的データベースシステム ANGEL～効率的対話入力方式の試みと引用関連モデル化のためのシステム構造～，北海道大学大型計算機センター・ニュース，19， 4，(昭 62)
- 7) 斉藤たつき：学術文献データベース用知的入力システム，情報処理学会第 35 回全国大会講演論文集，(昭 62)