北海道大学工学部研究報告
第151号（平成2年）

Bulletin of the Faculty of Engineering
Hokkaido University. No. 151 (1990)

Design and Implementation for Scientific Article Data Base

Tatsuki SAITO*

(Received March 31, 1990)

Abstract

Scientific article information data base system ANGEL was designed based on the data base management system ADABAS. It consists of an interactive input and retrieval system, intelligent input system and a system for the generation of a relation matrix. These systems were implemented by NATURAL of ADABAS. ANGEL is available for other scientific article data base systems. Initially, ANGEL was developed as an article data base relating to the sculptural surface generation theory of CAD (Computer Aided Design)/CAM (Computer Aided Manufacturing).

ANGEL is now open for public usage and is used at Hokkaido University Computer Center. In this paper, the design concept of the system and input system of bibliographic items are described, and several results obtained from the system are discussed. The input object of ANGEL is scientific articles written in English.

1. Introduction

Considering valid use of a scientific article data base, as well as use of bibliographic information, it is important to use relational information among articles effectually in order to reproduce higher intellectual information, and it is the motivation of this paper. Thereupon, the necessity of a system to comprehend a research trend was examined and it was investigated what information makes the essence of research clear. Attention was paid to the various relations among scientific articles. The relations are expressed by a labeled -graph, in which an article is represented as a point and a relation between articles is represented as a line. The aim of this paper is to develop the methodology for analyzing characteristics of the graph.

A citation relation among articles is an important information of relations. Accordingly, it is inputted with information of cited place because it expresses the significance of citation.

A keyword relation also is one of the necessary information to represent the commonness of the content or the theme. If there are many common keywords between two articles, then these articles are interpreted to be treated with a similar theme or the same

* Department of Precision Engineering, Faculty of Engineering, Hokkaido University, Sapporo 060, Japan

theme. These information among articles are the objective relations of information that are not influenced by the person to examine.

Relational graphs of articles are made out from these various information using this data base system, and some relational matrices are created from these graphs. The matrices are similarity matrices which have real values. After setting a mathematical criterion according to a purpose of processing, they are unified into one similarity matrix finally in order to clarify the property of the relational structure by cluster analysis. Thus configuration of the system is the target of this paper. In consequence, the system is designed to attach importance to an effectual input of these information.

The data base system ANGEL consists of keyboard subsystem, bibliographic item extraction subsystem, retrieval subsystem, display subsystem and storage subsystem in the entire composition. There are some shared modules which are entities of subsystems. These subsystems were implemented in NATURAL on ADABAS.

2. Constitution of the Data Base ANGEL

The constitution of the data base ANGEL is shown in Fig. 1. Bibliographic information of scientific articles is put into the data base system by manual operation directly or by optical character reader automatically through recording media like a floppy disk device. The most outside frame (thick lined) means the data base management system ADABAS that is the relational model type for main frame computer. The chained lined frame means the data base system ANGEL which consists of four parts. The first part (upper-left) is the interactive input and retrieval system, and the second part (upper-right) is the intelligent input system. The third part (middle) is the data storage (file). The fourth part (lower) is the generating system of relation matrix among scientific articles in the data base. The interactive input and retrieval system, and the intelligent input system are constructed respectively by four subsystems, that is, keyboard subsystem, display subsystem, retrieval subsystem and storage subsystem. The keyboard subsystem of the interactive input and retrieval system is for interactive manual input or interactive retrieval from VDT (visual display terminal).

On the other hand, the bibliographic item extraction subsystem of the intelligent input system is for automatic input from a recorded file. The display subsystem is for displaying processed results in the interactive input and retrieval system or in the intelligent input system. In the intelligent input system, the processed results are displayed in order to verify them.

The retrieval subsystem is to check whether processing data overlaps recorded data or not, and the subsystem of the interactive input and retrieval system enables to retrieve by author name, title, keyword et cetera. The storage subsystem is to store not only bibliographic data but also the logging data of processing date, processing type (store, update, delete), user-ID, the last article number or the last article-ID to be accessed. An example of the logging data is shown in Fig. 2.

The generating system of the relation is to set up the initial relation matrix from a

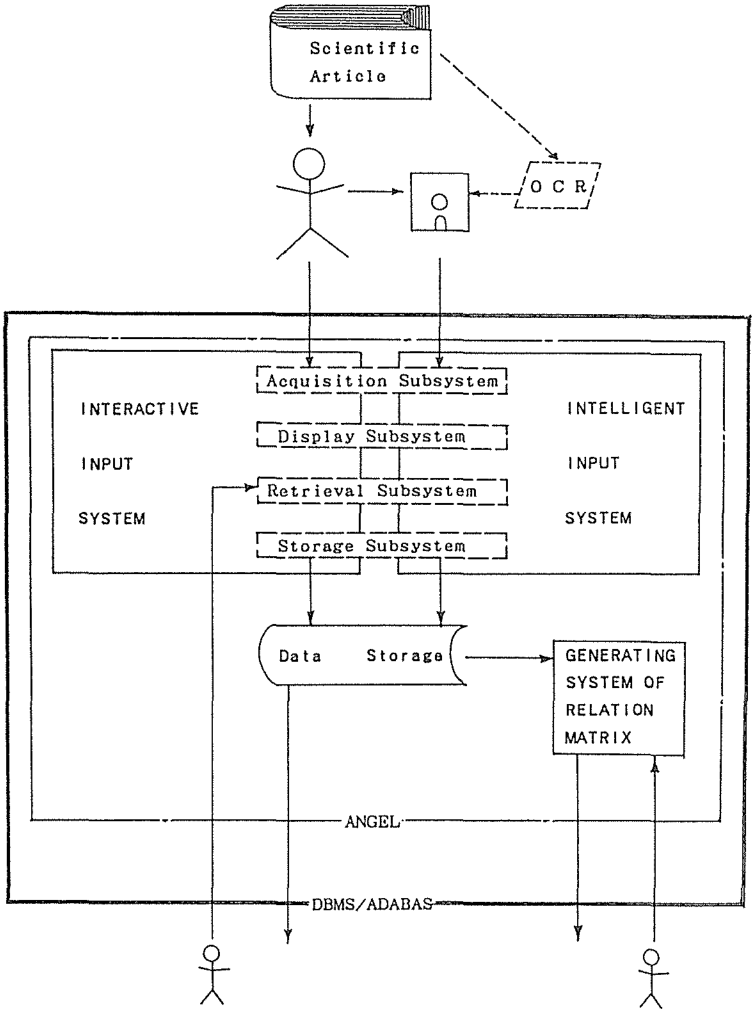


Fig. 1 The System constitution of the data base ANGEL.

specified relation and to output it into file.

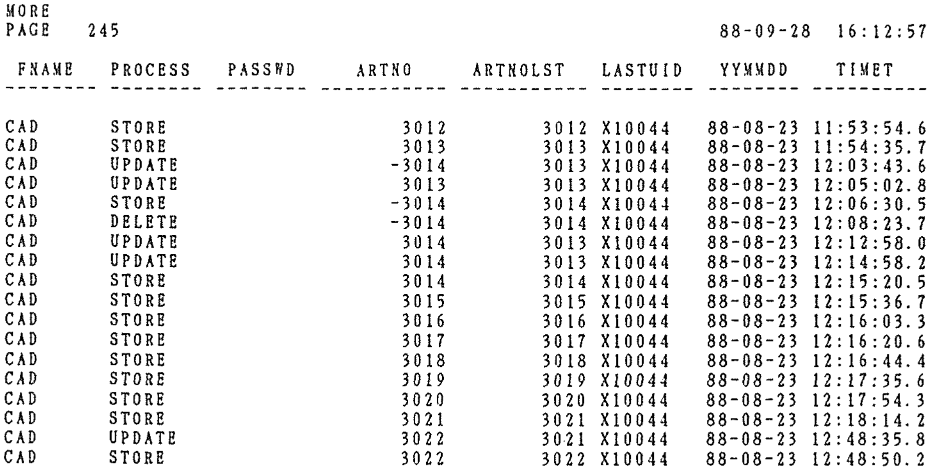
3. Scientific Article Information Input System of the Data Base

3. 1 Acquisition of scientific article information

There is a choice to purchase a marketing article data base to acquire bibliographic items by extracting a data field mechanically according to the specified format, and such choice is popular in general. There is also a choice to input bibliographic items manually after printed strings of article is read automatically by an optical character reader. A present situation of data input into data base needs manual inputting until when the possibility of the automatic item extraction realizes and the latter choice is available practically.

The researcher data base needs as much as possible detailed information. Accordingly, sufficient information cannot be obtained only from the marketing bibliographic data base like the in former method, and then a system to input necessary information directly from an

Fig. 2 An example of the logging data in ANGEL.



```
MORE
PAGE   245                                                                  88-09-28  16:12:57

 FNAME    PROCESS    PASSWD      ARTNO       ARTNOLST    LASTUID   YYMMDD     TIMET
-------- --------- --------- ----------- ----------- --------- --------- ----------

CAD       STORE                    3012         3012 X10044    88-08-23 11:53:54.6
CAD       STORE                    3013         3013 X10044    88-08-23 11:54:35.7
CAD       UPDATE                  -3014         3013 X10044    88-08-23 12:03:43.6
CAD       UPDATE                   3013         3013 X10044    88-08-23 12:05:02.8
CAD       STORE                   -3014         3014 X10044    88-08-23 12:06:30.5
CAD       DELETE                  -3014         3014 X10044    88-08-23 12:08:23.7
CAD       UPDATE                   3014         3013 X10044    88-08-23 12:12:58.0
CAD       UPDATE                   3014         3013 X10044    88-08-23 12:14:58.2
CAD       STORE                    3014         3014 X10044    88-08-23 12:15:20.5
CAD       STORE                    3015         3015 X10044    88-08-23 12:15:36.7
CAD       STORE                    3016         3016 X10044    88-08-23 12:16:03.3
CAD       STORE                    3017         3017 X10044    88-08-23 12:16:20.6
CAD       STORE                    3018         3018 X10044    88-08-23 12:16:44.4
CAD       STORE                    3019         3019 X10044    88-08-23 12:17:35.6
CAD       STORE                    3020         3020 X10044    88-08-23 12:17:54.3
CAD       STORE                    3021         3021 X10044    88-08-23 12:18:14.2
CAD       UPDATE                   3022         3021 X10044    88-08-23 12:48:35.8
CAD       STORE                    3022         3022 X10044    88-08-23 12:48:50.2
```

original article is indispensable. A researcher cannot offord to the mounting many cost easily for inputting. Therefore, a researcher is required to input the data by himself, or the data is inputted by a part-time assistant. The system that enables to input data efficiently even if by nonprofessional key operator is necessary. For this reason, it is necessary to decrease the hitting frequency as well as possible at the time of data input in order to make key operation easy. Consequently, two systems using registered information in the data base were programed. The first system is interactive input and retrieval system, and the second is intelligent input system.

3. 2 Interactive input and retrieval system

When the input systems were designed, the following points were considered ;

1) Improvement of input working efficiency.

Especially, for a user who is not an professional key operator, to improve the efficiency of input work is effective.

2) Reduction of misinput rate.

In case of inputting abundent data, it is said generally that even professional key operators misinput at about 3 % rate. In consequence, the reduction of misinput is an important theme to acquire the information in order to minimmize errors. So data were processed on the basis of the next point.

3) Valid use of registered data.

To consider an input processing or a retrieval processing for the researcher database of scientific article, both a parent (citing) article and a child (cited) article are equally significant, then it is necessary to treat them as independent articles because the database is to process various relations among articles. Accordingly, the same article appears many times as the reference which has a similar theme of research. Hence a tendency that the article by the same author appears frequently is remarkable in accordance with the increase of registered articles.

The necessity to check the duplication of inputted data that may be overlapping registered data comes into existence, and the duplication check in which the surname of the first author is the first retrieval key is valid in reliability and efficiency. Retrieval items are author name, title, article number, article identification number, time of publication, citing the authors name, journal name and key word. A retrieval condition can be set by also the logical AND or OR, that is, (surname of author) OR (title), (surname of author) AND (title or other bibliographic information as occasional demands). It can be used also intactly for simple retrieval processing. A full name or an affiliation of an author is stored or renewed only when these information appears as a parent article. On registering an article, it is always checked whether the article-ID overlaps the registered article-ID or the same title has been stored in the database. Consequently, the duplication of data can be prevented, and a consistency of data can be established simultaneously with the exclusion of the uselessness of a storage area.

When necessary data is displayed in the retrieval screen, it can be shifted to the registration mode at once in the interactive input and retrieval system. Therefore, after updating, it is possible to register newly by using a part of the data that has been registered. It can be also updated when the affiliation of an author changes. Becoming aware of error occasionally after a screen proceeds while interactive data inputting, it is possible to send the screen backward in order to correct errors before it is stored in the database.

3. 3 Details of interactive input and retrieval system

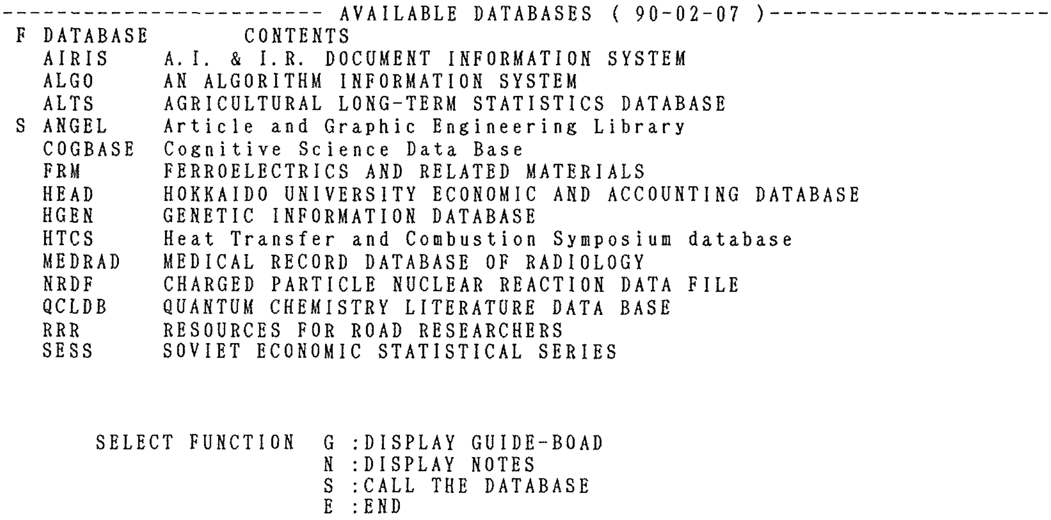
An input menu form is adopted for ease of viewing and for restraining the input error. The information that has been inputted are preserved for reuse as much as possible. The MAP function of the screen editor of NATURAL was applied to screen design. It is a convenient function to make screen layout flexible. Fig. 3. 1 shows the screen of selecting ANGEL among the data base of HUCC. The screen of starting ANGEL is shown in Fig.3. 2. A file name to be accessed, a password and selection parameter of processing are inputted here. A screen to input the selection whether a parent article or not is displayed after this starting screen is displayed. If the user has the right of writing the data base. After 'Y' -key is hit when the article is parental, or after only 'send'-key is hit when otherwise, the retrieval screen of Fig. 4 is displayed. An author name and a title are inputted in this screen as a retrieval item. Because the same article appears frequently in the item of a reference,

(1) to avoid duplication of registration of the same data,

(2) to check error of the registered data,

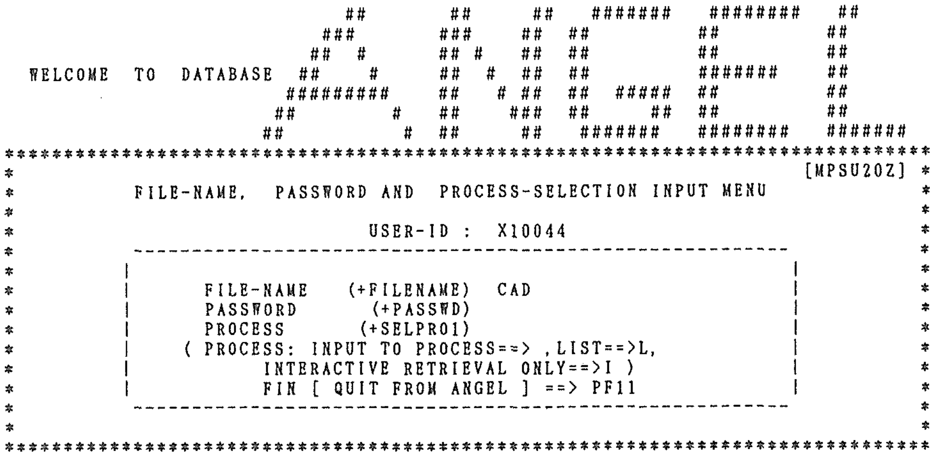
retrieval processing is executed first.

The default condition of retrieval is only a surname of an author. However the retrieval condition may adopt logical OR of each item or its logical AND. From the viewpoint of the efficiency, it may be advantageous not to input the title because it is long in comparison with an author name and it is inaccurate sometimes in case of being cited. To hit 'send'-key in this screen after necessary item is inputted, according to the specified condition, the retrieval process is executed. If a corresponding article is registered already, then the data of



```
------------------------- AVAILABLE DATABASES ( 90-02-07 )---------------------
F DATABASE       CONTENTS
  AIRIS    A.I. & I.R. DOCUMENT INFORMATION SYSTEM
  ALGO     AN ALGORITHM INFORMATION SYSTEM
  ALTS     AGRICULTURAL LONG-TERM STATISTICS DATABASE
S ANGEL    Article and Graphic Engineering Library
  COGBASE  Cognitive Science Data Base
  FRM      FERROELECTRICS AND RELATED MATERIALS
  HEAD     HOKKAIDO UNIVERSITY ECONOMIC AND ACCOUNTING DATABASE
  HGEN     GENETIC INFORMATION DATABASE
  HTCS     Heat Transfer and Combustion Symposium database
  MEDRAD   MEDICAL RECORD DATABASE OF RADIOLOGY
  NRDF     CHARGED PARTICLE NUCLEAR REACTION DATA FILE
  QCLDB    QUANTUM CHEMISTRY LITERATURE DATA BASE
  RRR      RESOURCES FOR ROAD RESEARCHERS
  SESS     SOVIET ECONOMIC STATISTICAL SERIES


      SELECT FUNCTION  G :DISPLAY GUIDE-BOAD
                       N :DISPLAY NOTES
                       S :CALL THE DATABASE
                       E :END
```

Fig. 3-1



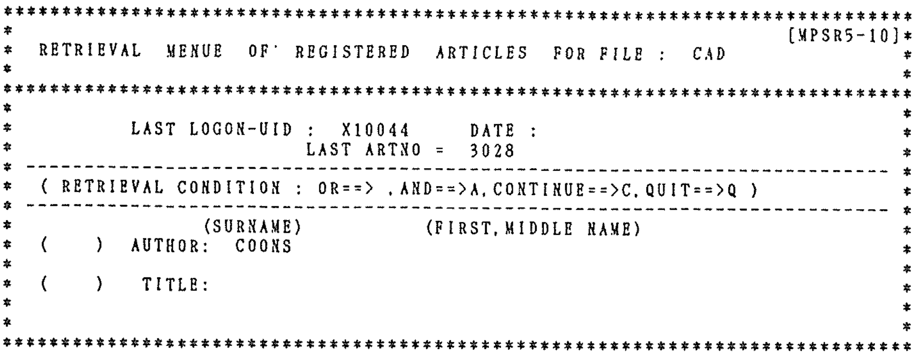
```
                            ##       ##     ##   #######   ########   ##
                          ###       ###    ##  ##         ##         ##
                         ##  #      ## #   ##  ##         ##         ##
  WELCOME  TO  DATABASE ##    #     ##  #  ##  ##         #######    ##
                       #########    ##   # ##  ##  #####  ##         ##
                      ##        #   ##    ###  ##     ##  ##         ##
                     ##          #  ##     ##   #######   ########   #######
******************************************************************************
*                                                                  [MPSU20Z] *
*          FILE-NAME,  PASSWORD AND  PROCESS-SELECTION INPUT MENU            *
*                                                                            *
*                             USER-ID :  X10044                              *
*          -------------------------------------------------------           *
*         |                                                       |          *
*         |      FILE-NAME   (+FILENAME)  CAD                     |          *
*         |      PASSWORD      (+PASSWD)                          |          *
*         |      PROCESS      (+SELPRO1)                          |          *
*         |    ( PROCESS: INPUT TO PROCESS==> ,LIST==>L,          |          *
*         |           INTERACTIVE RETRIEVAL ONLY==>I )            |          *
*         |           FIN [ QUIT FROM ANGEL ] ==> PF11            |          *
*          -------------------------------------------------------           *
*                                                                            *
******************************************************************************
```

Fig. 3-2 Start up screen of ANGEL.

registered content is displayed as in Fig. 5.

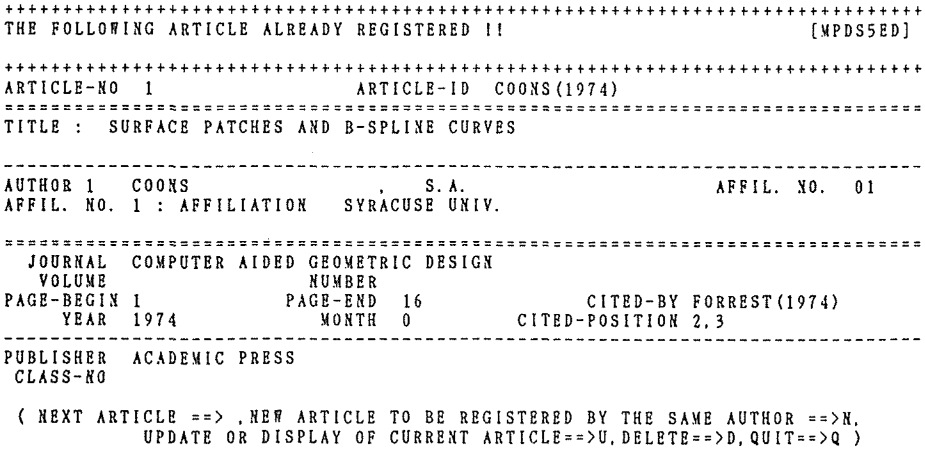
In case that there are plural data of the same author, to hit 'send'-key successively, corresponding bibliographic data are displayed in turn as in Fig. 6. Finding an error in the display screen, to hit 'U' at the last argument, it is possible to input the correct data instead of the error data. In case of registering a new article of the same author while retrieving, it is advisable hit 'N' at the last argument, then it is possible to reuse intactly the corresponding data that has been inputted in the database, and it is necessary for the keyboard to only put data of the field that has never inputted.

While this method gives a few advantages at the initially creating stage of the data base, the effect can be expected in the practical period after numerous data are stored in the data base. Especially in the case of a researcher data base, it is not rare that the same author appears frequently in case of the treatment of comparably narrow field. Consequently, the



```
**************************************************************************
*                                                              [MPSR5-10]*
*  RETRIEVAL  MENUE  OF' REGISTERED  ARTICLES  FOR FILE :  CAD           *
*                                                                        *
**************************************************************************
*                                                                        *
*          LAST LOGON-UID :  X10044     DATE :                           *
*                         LAST ARTNO =  3028                             *
* ---------------------------------------------------------------------- *
*  ( RETRIEVAL CONDITION : OR==> ,AND==>A,CONTINUE==>C,QUIT==>Q )        *
* ---------------------------------------------------------------------- *
*               (SURNAME)          (FIRST,MIDDLE NAME)                   *
*  (    )  AUTHOR:  COONS                                                *
*                                                                        *
*  (    )   TITLE:                                                       *
*                                                                        *
*                                                                        *
**************************************************************************
```

Fig. 4 Retrieval screen for author and title.



```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
THE FOLLOWING ARTICLE ALREADY REGISTERED !!                      [MPDS5ED]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
ARTICLE-NO  1                ARTICLE-ID  COONS(1974)
==========================================================================
TITLE :  SURFACE 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             CITED-BY FORREST(1974)
     YEAR  1974            MONTH  0        CITED-POSITION 2,3
--------------------------------------------------------------------------
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 )
```

Fig. 5 Display screen of the retrieved results.

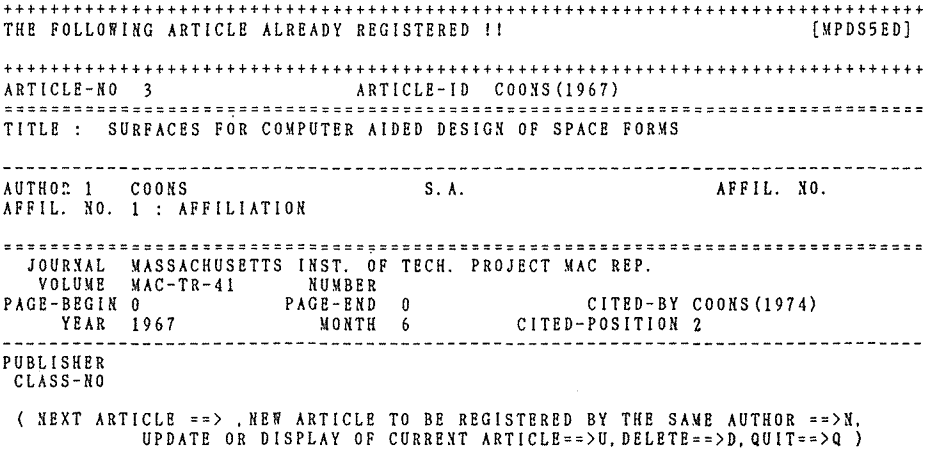
retrieval by author name is effective in the input system. The retrieval speed does not decrease extremely according to an increase of the registration number because inverted file is created with a key for the specified author field.

The specification of 'C' to the first argument in the retrieval screen of Fig. 4 is to display a successively retrieval screen. If the content can not be inputted in the screen (over 250 words), then to specify 'C' in the bottom argument, the continued screen for the content is displayed. The capacity of the content field is one thousand characters from the limit of the data base management system ADABAS. Fig. 7 shows the interactive retrieval screen except author and title.

To input surname of author, title, journal name, publication year, key word (s), article ID, surname of citing author and article number (when it is plural, to specify the beginning value and the ending value), these retrievals are executed. When 'A' is specified into ( ) of the left end, the intersection of those specified items is set as the retrieval condition.

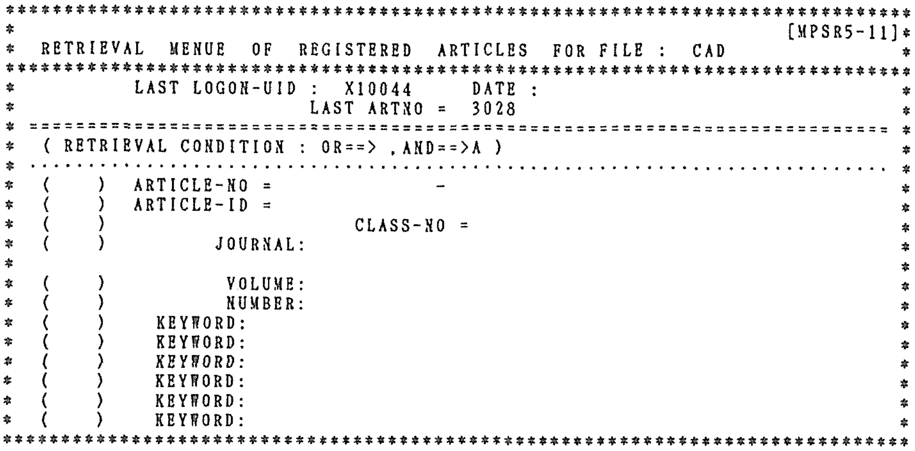
3.4 Intelligent input system

It is a weak point that we can not use the interactive input method described in 3.3 at the



```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
THE FOLLOWING ARTICLE ALREADY REGISTERED !!                           [MPDS5ED]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
ARTICLE-NO  3                  ARTICLE-ID  COONS(1967)
================================================================================
TITLE :  SURFACES FOR COMPUTER AIDED DESIGN OF SPACE FORMS

--------------------------------------------------------------------------------
AUTHOR 1   COONS                     S.A.                     AFFIL. NO.
AFFIL. NO. 1 : AFFILIATION

================================================================================
  JOURNAL  MASSACHUSETTS INST. OF TECH. PROJECT MAC REP.
   VOLUME  MAC-TR-41       NUMBER
PAGE-BEGIN 0            PAGE-END  0              CITED-BY COONS(1974)
     YEAR  1967            MONTH  6        CITED-POSITION 2
--------------------------------------------------------------------------------
PUBLISHER
 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 )
```

Fig. 6 Display screen of the retrieved results for the same author



```
********************************************************************************
*                                                                    [MPSR5-11]*
*  RETRIEVAL  MENUE  OF  REGISTERED  ARTICLES  FOR FILE :  CAD                 *
********************************************************************************
*          LAST LOGON-UID :  X10044     DATE :                                 *
*                         LAST ARTNO =  3028                                   *
* ============================================================================ *
*  ( RETRIEVAL CONDITION : OR==> ,AND==>A )                                    *
* ............................................................................ *
*  (    )  ARTICLE-NO =                 -                                      *
*  (    )  ARTICLE-ID =                                                        *
*  (    )                     CLASS-NO =                                       *
*  (    )        JOURNAL:                                                      *
*                                                                              *
*  (    )          VOLUME:                                                     *
*  (    )          NUMBER:                                                     *
*  (    )    KEYWORD:                                                          *
*  (    )    KEYWORD:                                                          *
*  (    )    KEYWORD:                                                          *
*  (    )    KEYWORD:                                                          *
*  (    )    KEYWORD:                                                          *
*  (    )    KEYWORD:                                                          *
********************************************************************************
```

Fig. 7 Retrieval screen for items of key field except author and title.

time except the service time of the on-line TSS. Then, it is necessary to input data by using a personal computer. An intelligent input system was developed. This system can automatically extract the bibliographic items written according to the format as follows.

(1) To be written is the following order and format;

AUTHOR / AFFILIATION :
"TITLE"
JOURNAL, VOLUME, NUMBER, BEGINNING-PAGE - ENDING-PAGE,
(YEAR - MONTH), PUBLISHER
#CLASSIFICATION-NUMBER
^KEY-WORD
(CONTENT)
＊REMARKS
@CITED-POSITION

(Bibliographic items of cited articles continue here in order as described above)

(2) Items except "TITLE" may be omitted.

3. 5 Intelligent input system using rule system

This input system is implemented in NATURAL by using production system like method. The production system consists of three parts, that is, Rule part, Data Base part and PSI (Production System Interpreter).

1) Rule part is expressed as the order pair LHS→RHS.

LHS (Left Hand Side) is called the left side rule or the condition part, and RHS (Right Hand Side) is called the right side rule, or the action part. This can be expressed as follows.

IF { (LHS)=T } (→) THEN DO { (RHS) },

that is, it means that if the condition part LHS is true then the action part RHS is executed.

2) Data base part is called working memory under certain circumstances, and it is the part in which the string of a processing object is stored. When there is any rule that will make LHS true in the rules, the objective string is changed to become an another state, then it is called "production system".

3) PSI (rule application part) is controlled all over the production system. PSI matches the rule that is stored in the data base with strings of processing object. If LHS was true, then RHS is applied. As the rule generally exists in plural, if LHS is false, then the next rule is examined.

Thus repetition processing is called recognize-act cycle. The production system of the type that is described above is called antecedent-driven production system or forward chaining production system. The production system of the type that requires LHS to execute RHS is called consequent-driven production system or backward chaining production system.

Our system is not the proper production system but production system like system because the data base management system ADABAS does not support AI language for the production system. Consequently adopting the non proper production system, could not be avoided but the following description is the production system because it is designed by production system originally.

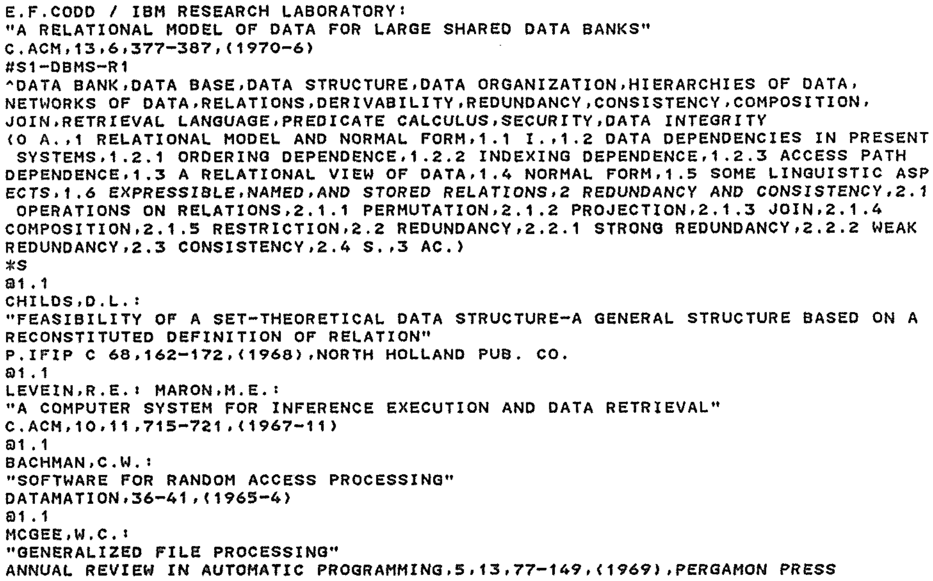
This system is the antecedent-driven type in case of production system. As a part of the rule of this system, extracting rules of author name and affiliation, extracting rules of a title, rules of automatic numbering of article ID and a practical processing example are mentioned below, and raw data is shown in Fig. 8.

[Extraction of author name and affiliation]

R1 LHS1 Colon(:) exists. RHS1 The string of the left side of colon is set as data of author information.

R2 LHS2 Colon(:) does not exist. RHS2 An author name is set as ANON..

R3 LHS3 Slash(/) exists in the author information data. RHS3 Set the left side of the processing strings as an author name, and set the right side of the processing strings to an affiliation.

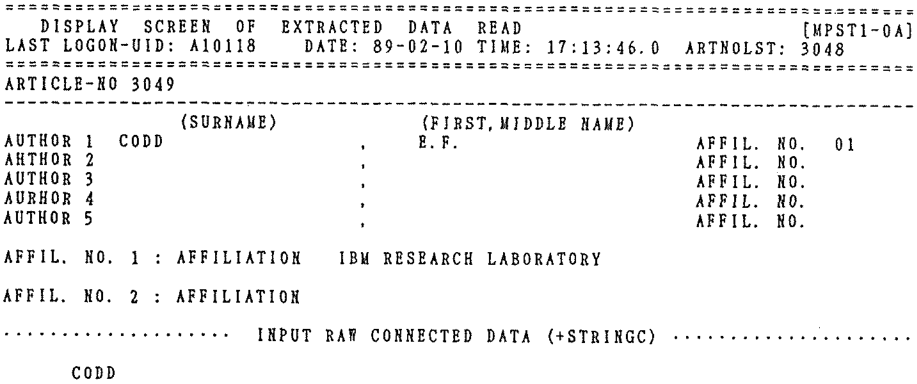


```
E.F.CODD / IBM RESEARCH LABORATORY:
"A RELATIONAL MODEL OF DATA FOR LARGE SHARED DATA BANKS"
C.ACM,13,6,377-387,(1970-6)
#S1-DBMS-R1
^DATA BANK,DATA BASE,DATA STRUCTURE,DATA ORGANIZATION,HIERARCHIES OF DATA,
NETWORKS OF DATA,RELATIONS,DERIVABILITY,REDUNDANCY,CONSISTENCY,COMPOSITION,
JOIN,RETRIEVAL LANGUAGE,PREDICATE CALCULUS,SECURITY,DATA INTEGRITY
(0 A.,1 RELATIONAL MODEL AND NORMAL FORM,1.1 I.,1.2 DATA DEPENDENCIES IN PRESENT
 SYSTEMS,1.2.1 ORDERING DEPENDENCE,1.2.2 INDEXING DEPENDENCE,1.2.3 ACCESS PATH
DEPENDENCE,1.3 A RELATIONAL VIEW OF DATA,1.4 NORMAL FORM,1.5 SOME LINGUISTIC ASP
ECTS,1.6 EXPRESSIBLE,NAMED,AND STORED RELATIONS,2 REDUNDANCY AND CONSISTENCY,2.1
 OPERATIONS ON RELATIONS,2.1.1 PERMUTATION,2.1.2 PROJECTION,2.1.3 JOIN,2.1.4
COMPOSITION,2.1.5 RESTRICTION,2.2 REDUNDANCY,2.2.1 STRONG REDUNDANCY,2.2.2 WEAK
REDUNDANCY,2.3 CONSISTENCY,2.4 S.,3 AC.)
*S
@1.1
CHILDS,D.L.:
"FEASIBILITY OF A SET-THEORETICAL DATA STRUCTURE-A GENERAL STRUCTURE BASED ON A
RECONSTITUTED DEFINITION OF RELATION"
P.IFIP C 68,162-172,(1968),NORTH HOLLAND PUB. CO.
@1.1
LEVEIN,R.E.: MARON,M.E.:
"A COMPUTER SYSTEM FOR INFERENCE EXECUTION AND DATA RETRIEVAL"
C.ACM,10,11,715-721,(1967-11)
@1.1
BACHMAN,C.W.:
"SOFTWARE FOR RANDOM ACCESS PROCESSING"
DATAMATION,36-41,(1965-4)
@1.1
MCGEE,W.C.:
"GENERALIZED FILE PROCESSING"
ANNUAL REVIEW IN AUTOMATIC PROGRAMMING,5,13,77-149,(1969),PERGAMON PRESS
```

Fig. 8 An example of raw data for this intelligent input system.

R4 LHS4 Comma(,) exists in the author name. RHS4 Set the left side of the processing strings as a surname, and set the right side of the processing strings as a name except surname.

A processing example is shown in Fig. 9.



```
========================================================================
   DISPLAY  SCREEN  OF  EXTRACTED  DATA  READ                [MPST1-0A]
LAST LOGON-UID: A10118    DATE: 89-02-10 TIME: 17:13:46.0  ARTNOLST: 3048
========================================================================
ARTICLE-NO 3049
------------------------------------------------------------------------
               (SURNAME)                (FIRST,MIDDLE NAME)
AUTHOR 1  CODD                  ,      E.F.                   AFFIL. NO.  01
AHTHOR 2                        ,                             AFFIL. NO.
AUTHOR 3                        ,                             AFFIL. NO.
AURHOR 4                        ,                             AFFIL. NO.
AUTHOR 5                        ,                             AFFIL. NO.

AFFIL. NO. 1 : AFFILIATION   IBM RESEARCH LABORATORY

AFFIL. NO. 2 : AFFILIATION

..................... INPUT RAW CONNECTED DATA (+STRINGC) .....................

      CODD
```

Fig. 9 An example extracted author name and affiliation.

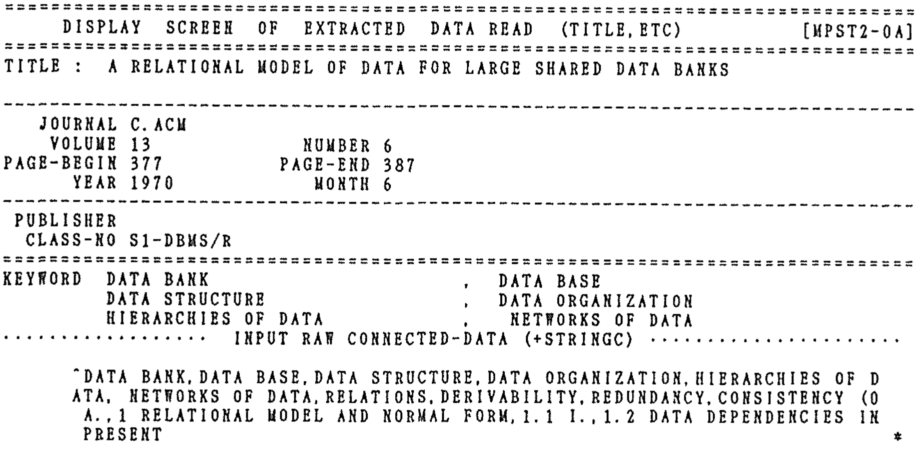
[Extraction of a title]

R5 LHS5 There are two double quotation marks ("). RHS5 Set a string surrounded by double quotation marks as a title.

R6 LHS6 There is no double quotation mark. RHS6 Display a warning of NON TITLE.

An example extracted title, etc. from the raw data of Fig. 8 is shown in Fig. 10.

[Numbering of article and naming of article ID]



```
=========================================================================
     DISPLAY  SCREEN  OF  EXTRACTED  DATA READ  (TITLE,ETC)        [MPST2-0A]
=========================================================================
TITLE :  A RELATIONAL MODEL OF DATA FOR LARGE SHARED DATA BANKS

-------------------------------------------------------------------------
   JOURNAL C.ACM
    VOLUME 13            NUMBER 6
PAGE-BEGIN 377         PAGE-END 387
      YEAR 1970           MONTH 6
-------------------------------------------------------------------------
 PUBLISHER
  CLASS-NO S1-DBMS/R
=========================================================================
KEYWORD  DATA BANK                       ,  DATA BASE
         DATA STRUCTURE                  ,  DATA ORGANIZATION
         HIERARCHIES OF DATA             ,   NETWORKS OF DATA
··················  INPUT RAW CONNECTED-DATA (+STRINGC) ·····················

       ^DATA BANK,DATA BASE,DATA STRUCTURE,DATA ORGANIZATION,HIERARCHIES OF D
       ATA, NETWORKS OF DATA,RELATIONS,DERIVABILITY,REDUNDANCY,CONSISTENCY (0
        A.,1 RELATIONAL MODEL AND NORMAL FORM,1.1 I.,1.2 DATA DEPENDENCIES IN
        PRESENT                                                              *
```

Fig. 10 An example extracted title, etc.

R7 LHS7 An article number (ARTNO) has not been registered. RHS7 Add one to the final article number (ARTNOLST).

R8 LHS8 The time when the article was published is stated. RHS8 Set article ID (ARTID) = SURNAME (YEAR).

R9 LHS9 The time when the article was published is not stated. RHS9 Set the YEAR = 0000, that is, article ID (ARTID) = SURNAME (0000).

R10 LHS10 Article ID (ARTID) agrees with a registered article ID. RHS10 Display a warning of article ID duplication.

R11 LHS11 Article ID (ARTID) conforms with a registered article ID and the published year is the same and the title does not conform. RHS11 Name the new article ID that is changed the YEAR part by the alphabetically consequential character as follows:

○ the last article ID = XXXXXXXX (YYYYB)

↓

○ the next article ID by the same author and published at the same year = XXXXXXXX (YYYYC)

4. Generating System of Relation Matrix

This system is to set up a matrix of the relation among scientific articles by user's specification. A rowwise number and a columnwise number of the matrix mean an article number. If it is for the citation relation matrix, then the matrix that (i , j)element is *1* when article i is cited by article j is outputted. It is the asymmetric matrix because the citation relation is not reciprocal with regard to time.

If it is for keyword relation matrix (or title relation matrix), then the matrix that (i , j) element is *m* when there are *m* common keywords (or terms) between article i and article j is outputted.

It is the symmetric matrix. If it is for author relation matrix, then the matrix that (i , j)element is *1* when an author wrote article i and article j commonly is outputted. Also, it is the symmetric matrix.

5. Exploratory Illustration

The source data list inputted in data base system ANGEL is shown in Table 1 summarily. In the field for computational three dimensional geometry mainly (the research field to be picked up is expressed in "CONTENTS" item), including both articles to cite and articles to be cited, about 5,000 scientific articles of CAD (Computer Aided Design)/CAM (Computer Aided Manufacturing) and of a few data base fields are acquired into the data base. Hereinafter, in regard to the relationship between a citing article and a cited article, several considerations are taken up in more detail.

Bibliog. (FILE NAME)	CONTENTS	*To cite / To be cited*	Cards.
Computer Aided Geometric Design (CAGD)	CAGD	15 / 188	833
PROLAMAT' 69 (P69)	CAD / CAM	32 / 146	725
PROLAMAT' 73 (P73)	CAD / CAM	69 / 152	1070
PROLAMAT' 76 (P76)	CAD / CAM	36 / 109	702
JICST (JICST)	CAD	34 / 546	2453
Computer Aided Design (CAD)	CAD	62 / 475	2486
IEEE Tranactions on Computers (IEEETC)	general	51 / 394	2098
Proceedings of IEEE (PIEEE)	imag. process. patt. recog.	39 / 1035	4402
Design Engineering Projects. (DEP)	CAD / CAM	1 / 166	653
SI AM Journal on Control and Optimization (SI AM)	math. prog.	10 / 165	683
Computer Journal (CJ)	general	97 / 787	4295
Communication of ACM (CACM)	comp. graph.	22 / 172	962
Journal of ACM (JACM)	comp. graph.	2 / 24	122
Journal of Mathematics and Physics (JAPSAM)	CAGD	1 / 5	26
Journal of Approximation Theory (JAT)	CAGD	2 / 22	108
Numerical Control in Manufacturing (NCM)	CAM	30 / 127	710
Proceedings of Royal Society of London (PRSL)	CAD / CAM	5 / 24	139
Journal of Mathematics and Mechanics (JMM)	CAGD	6 / 35	179
Journal of Approximation Theory (JAT)	CAGD	13 / 135	651
Journal of Mathematical Physics (JMP)	CAGD	3 / 7	43
Numerische Mathematik (NM)	CAGD	2 / 67	302
total		532 / 4781	23692

Table 1 The source data list inputted in data base ANGEL.

Table 2 shows citation rate for frequently referred journals among 1118 scientific articles. The citation ratio well referred is 2.4% at most. Even if seeing journals with same asterisk in the same journal, maximum is 3.3% of Communication, Journal and Proceedings of American Computer Machinery. Accordingly, this research field is thought to be wide spread. The case of CAGD (Computer Aided Geometric Design) which is the proceeding of symposium on computational geometric design is shown in Table 3. It is remarkably high rate to refer itself. This fact is thought that authors knew contents to be presented by another author each other before being published. However, such a case almost does not happen generally.

C. ACM	18 ／ 1118	1.6 %	3.3 %
J. ACM	15 ／ 1118	1.3 %	
P. ACM	5 ／ 1118	0.4 %	
FJCC	16 ／ 1118	1.4 %	3.1 %
SJCC	19 ／ 1118	1.7 %	
Ph. D. Thesis	27 ／ 1118	2.4 %	
Comp. J.	8 ／ 1118	0.7 %	
IBM Sys. J.	7 ／ 1118	0.6 %	
P. IEEE	9 ／ 1118	0.8 %	2.3 %
P. IEEE Conf. Deci. Cont.	6 ／ 1118	0.5 %	
P. IEEE Conf. Optim. Theor.	3 ／ 1118	0.3 %	
P. IEEE Work. Conf.	3 ／ 1118	0.3 %	
IEEE T. Auto. Cont.	2 ／ 1118	0.2 %	
IEEE T. Circ. Theor	2 ／ 1118	0.2 %	

Table 2 The list of frequently cited journals.

Referred Bibliog. List (Computer Aided Geometric Design)
15 / 188

Bibliog.	Freg.	Ratio(%)
CAGD	26	13.8
Ph. D. Thesis	9	4.8
MIT Project MAC TR	7	3.2
P. Roy. Soc. London	7	3.2
J. Approx. Theor.	6	3.2
Comp. J.	5	2.7
Comm. Graph. Imag. Process.	5	2.7
J. ACM	4	2.1
P.ACM	3	1.6
AFIPS Conf. P.	3	1.6
CAD	3	1.6
C. ACM	2	1.1

Table 3 Fig. 5.18 The list of journals cited by CAGD.

Referred Bibliog. List (SIAM) 10 / 165

Bibliog.	Freg.	Ratio(%)
SIAM J. Cont. Optim.	13	7.9
P. IEEE Conf. Deci. Cont.	7	4.2
Math. Prog.	7	4.2
J. Optim. Theo. Appl.	6	3.6
Ph. D. Thesis.	5	3.0
Israel J. Math.	4	2.4
J. Math. Anal Appl.	4	2.4
B. Amer. Math. Soc.	4	2.4
T. Amer. Math. Soc.	3	1.8
Acta Math.	3	1.8
Comp. J.	3	1.8
P. IFIP Cont. Optim. Theor.	3	1.8
Acta Sci. Math.	2	1.2
IEEE T. Auto. Cont.	2	1.2
IEEE T. Circuit. Theor.	2	1.2
Manag. Sci.	2	1.2

Table 4 The list of journals cited by SIAM J. C. O.

In case of another research field, the result of SIAM Journal of Control Optimization is shown in Table 4. Though the citation rate to own is high, because reason why the rate to another journal is considerably not low, this research field is thought to be open or to have been developed widely already.

On the other hand, in the case of fresh research filed, the citational frequency of PROLAMAT is shown in Table 5, 6 and 7 respectively. PROLAMAT (Programming language for mathematics) started at 1969. As in Table 5, there is little remarkable difference among referred journals in citation rate. However, because the later the time is, the greater the rate is, its research field is considered to be established, and then from Table

Referred Bibliog. List (PROLAMAT '69) 32 / 146

Bibliog.	Freg.	Ratio(%)
C.ACM	8	5.5
FJCC	5	3.4
ASME Product. Eng.	3	2.1
Proc. IEEE work. Conf.	3	2.1
IBM TR	3	2.1
IBM Sys. J.	2	1.4
Metal. Product.	2	1.4

(except ISO, Local Lad. Rep.)

Table 5 The list of journals cited by PROLAMAT '69.

Referred Bibliog. List (PROLAMAT '73) 69 / 152

Bibliog.	Freg.	Ratio(%)
PROLAMAT '69	7	4.6
Ph. D. Thesis	7	4.6
CAD	3	2.0
Confr. Eng.	3	2.0
IFAC Proc.	3	2.0
AUTOMATICA	2	1.3
P. Roy. Soc. London.	2	1.3
MIT Project MAC TR	2	1.3
MTDR Int. Conf.	2	1.3
NPRA Comp. Conf.	2	1.3
SAE	2	1.3

(except ISO)

Table 6 The list of journals cited by PROLAMAT '73.

Referred Bibliog. List (PROLAMAT '76) 36 / 109

Bibliog.	Freg.	Ratio(%)
PROLAMAT '73	14	12.8
CAM '74	8	7.3
Ph. D. Thesis	3	2.8
Ann. CIRP	3	2,8
Inf. J. Product. Res.	2	1.8
P. MTDR	2	1.8
P. Roy. Soc. London	2	1.8
C. ACM	2	1.8
J. ACM	1	0.9
PROLAMAT '69	1	0.9

Table 7 The list of journals cited by PROLAMAT '76.

Referred Bibliograph List (JICST) 34 / 546

Bibliog	Freg.	Ratio(%)
SJCC	19	3.5
Ph. D. Thesis	12	2.2
FJCC	11	2.0
P. NCC	10	1.8
P. IFIP	10	1.8
P. IEEE	9	1.6
C. ACM	8	1.5
P. Int. Conf. Comp. Arch	6	1.1
J. ACM	5	0.9
P. ACM	5	0.9
MIT MAC TR	5	0.9
Comp. Aid. Geo. Des.	5	0.9
Comp. J.	5	0.9
IBM Sys. J.	5	0.9
CAD	4	0.7
Inf. Process.	4	0.7
Manag. Sci	4	0.7
J. Math. Eng.	3	0.5
PROLAMAT '73	3	0.5
P. MTDR	3	0.5
J. Indust. Eng.	3	0.5
DATAMATION	3	0.5

Table 8 The list of journals cited by 34 scientific articles having keyword "CAD" (acquired by JICST).

7, it is considered that this research field has the property to require fresh method because of high ratio at PROLAMAT '73 in contrast to at PROLAMAT '69.

Finally, Table 8 shows JICST's SDI service to be hit by keyword "CAD". Like citation ratio in other figures, it is also noticeable that the cited frequency of Ph.D. dissertation is high, here.

6. Considerations

Inaccurate bibliographic information exists especially, this tendency appears in titles of article to be cited. In order to prevent wrong extraction of an each item, the intelligent input system is designed so as to display the results that are processed at each extraction for confirmation. When a data item can not be segmented accurately due to improper setting of delimiter or due to original data error itself, it would be efficient to correct it immediately on the spot. This method is thought to be superior to the method that checks after thorough batch processing. These interactive input system and intelligent input system adopt the method that compares input data with stored data for verification by retrieving automatically when bibliographic items are registered.

The result of the intelligent input system can be summarized as follows.

1. Realization of the extraction function for a bibliographic item by the rule system.
2. Flexibility of accurate extraction by the rule system in the case of some deficient items.
3. Reduction of the input work burden by this method.
4. Decrease of the data input cost by pre-processing with the personal computer system.

Input data format of bibliographic information is assumed as mentioned in 3.4. However, this intelligent input rule system can be applied to various journals by having these reference format data as knowledge. The environment that the science information system can be used regularly is being established.

Not to mention the usage of information retrieval for bibliographic item, including this input system, the development of methodology that enables an intellectual usage of the system is required immediately in order to utilize the system highly. It is said that the greatest bottleneck of the data base system is acquisition and the input of data. This method is thought to be available.

7. References

1) Moed, H. F. : Vriens, M. : "Possible inaccuracies occurring in citation analysis" J. Info. Sci., 15, 2, 95－107, (1989)

2) Saito, Tatsuki : Tejima, Shoichi : Kawai, Norio : Okino, Norio : "Study on Development of Methodology for Grasping Research Trend and Efficiency of the Scientific Information Retrieval by It" 'Formation Process of Information Systems and Organization of Scientific Information'-Report (a grant-in-aid of the Ministry of Education of Japan (in Japanese), (1977)

3) Saito, Tatsuki : Okino, Norio : Asano, Chooichiro : Tanaka, Hajime "A modeling and clustering method by relational graph model and an exploratory application for cluster analysis of scientific article space" ' SCIENCE AND TECHNOLOGY OF SOCIETY' Proc. 8th IFAC, (1982)

4) Bichteler, Julie : Parsons,Ronald G. : "Document retrieval by means of an automatic classification algorithm for citations" Info. Stor. Retr., 10, 267－278, (1974)

5) Small, Henry : "The relationship of information science to the social science: A co-citation analysis" Info. Proc. Manag., 17, 39－50, (1981)

6) Kochtanek, Thomas R. : "Bibliographic compilation using reference and citation links" Info. Proc. Manag., 18, 1, 33－39, (1982)

7) Noma, Elliot : "Untangling citation networks" Info. Proc. Manag., 18, 2, 43－53, (1982)

8) Hurt, C. D. : "Conceptual citation differences in science technology and social science literature" Info. Proc. Manag., 23, 1, 1－6, (1987)

9) Todôrov, R. : Glanzel, W. : "Journal citation measurers - a concise review" J. Info. Science, 14, 1, 47–56, (1988)

10) Zunde, P. : Slameck, V. : "Predictive models of scientific progress" Info. Stor. Retr., 7, 103–109, (1971)

11) Borenius, G.: Schwarz, S. : "Remarks on the use of citation data in predictive models of scientific activity" Info. Stor. Retr., 8, 171–175, (1972)

12) Garvey, William D. : Lin, Nan : Tomita, Kazuo : "Research studies in pattern of scientific communication : III information-exchange processes associated with the production of journal articles" Info. Stor. Retr., 8, 207–221, (1972)

13) Garvey, William D. : Lin, Nan : Tomita, Kazuo : "Research studies in scientific communication : IV The continuity of dissemination of information by 'productive scientists'" Info. Stor. Retr., 8, 265–276, (1972)

14) Cawkell, A.E. : "Evaluating scientific journals with Journal Citation Reports, - a case study in acoustics" J. Amer. Soc. Info. Sci., 29, 1, 41–46, (1978)

15) Culnan, J.Mary : "An analysis of the information usage patterns of academics and practitioners in the computer field: A citation analysis of a national conference" Info. Proc. Manag., 14, 395–404, (1978)

16) Garland, Kathleen : "An experiment in automatic hierarchical document classification" Info. Proc. Manag., 19, 3, 113–120, (1983)