DOCUMENT CONTENTS REPRESENTATION MODEL
OF
SENTENCE RETRIEVAL SYSTEM SCAT-IR

Michiyo Nikkuni and Hajime Tanaka*
Center for Information Processing Education
Hokkaido University, Sapporo 060, Japan
*Faculty of Science, Hokkaido University,
Sapporo 060, Japan

Abstract

A "Document Contents Representation" (DCR) model is introduced from a formal viewpoint to deal with the entire contents of a document such as individual sentences of a text, bibliography, references, etc. in a scientific information system. A "Mapping Definition Language" (MDL) is proposed to map directly and naturally the document contents into the DCR model. An application of the DCR model and MDL to scientific documents is shown. Some examples of advanced retrieval by SCAT-IR system implemented on the basis of the DCR model and MDL are illustrated.

1. Introduction

In the area of scientific information systems, a system to retrieve documents by some key words and bibliographic information has taken the top seat as an advantageous one. In such a document retrieval system, the output information has been restricted to bibliographic information, abstract and reference. However, necessary information in scientific research is the content itself of a document, in particular, some parts of the content that relates to the scientific research. It is an important problem to extract them from the content. Such parts are expressed by some sentences in the document. Therefore, if individual sentences can be retrieved, it is very useful for the settlement of the above problem.

We have studied the methodology[†] to deal with individual sentences of a text of a scientific document in an information system.[1),2)] It has been proved that each sentence in a document can be classified objectively. Functional information has been found to identify each sentence. The functional information is comparable with the entire context of a document, which corresponds to each characteristic step of a research course. Some conceptual categories have been set up on the bases of various steps, for example, "Basic problems," "Moment of development," "Individual problems," "Results" and "Conclusion" as shown in Fig. 1. Individual sentences in scientific documents can be classified by these categories. When an information system deals with functional information, it becomes possible to retrieve effectively individual sentences in a document.

01 General views
02 Basic problems
03 Awareness of problems
04 Analysis of the status que
05 Moment of development (Idea)
06 Individual problems
07 Development
08 Results
09 Evaluation
10 Conclusions
11 Awareness of problems following next
12 Acknowledgement

Fig. 1. Classification categories

This paper describes the methods of dealing with individual sentences and other information of a document such as bibliographic information, references, etc. in the same way in an information retrieval system. First, a "Document Contents Representations" (DCR) model is introduced from a formal viewpoint, and a "Mapping Definition Language"(MDL) is proposed to transfer directly and naturally document contents into DCR model. The DCR model is formed based on the following needs:

i) It should be independent of the kind of scientific document.
ii) It should be able to map easily document contents into the DCR model.
iii) It should be able to support various retrieval requirements.

The MDL is devised as a mechanical procedure for ii) above.

A sentence retrieval system SCAT-IR (Sentences Categorized - IR) has been implemented on the basis of the DCR model and MDL. We have studied various advanced utilizations of the SCAT-IR system in scientific research. It is shown that the DCR model is effective for an advanced document contents retrieval system and that the DCR model enables an information retrieval system to support the advanced utilizations, for example retrieval of individual sentences fitting the interests of researchers and the grasping of research trends through sentence retrieval in a research laboratory, etc.

In §2, the DCR model is introduced and in §3, the

[†] This study was performed by the authors, Prof. Y. Akaishi, Prof. J. Hiura, Prof. H. Bando, Prof. R. Tamagaki and Prof. S. Nagata.

©1981 ACM 0-89791-052-4/81/0500-0106 $00.75

MDL is defined. In §4, the application of the DCR model and MDL to scientific documents is shown. In §5, a query language SCAT-QL1 implemented on SCAT-IR system is described. Remarks and conclusion are discussed in §6.

2. Document Contents Representation (DCR) Model

2.1 Document contents analysis

In this section, three basic concepts of the DCR model are introduced to represent formally document contents i.e., "group", "data unit" and "element", they are formed based on two kinds of content analysis of a document.

The first one analyses the constituents of a document. A document is divided into some "groups" such as Bibliography (BIB), Title (TIL), Author (ATH), Abstract (ABS), Text (TEXT), Reference (REF) and so on. (...) shows each group name. Each group is represented with two kinds of "element": 'data element' and 'connective element'. The former corresponds to the occurrence of content information in each group. For example, the group TEXT is represented with two data elements i.e., 'classification category' and 'sentence'. The latter is used to connect individual values of 'data element' or "data unit". For example, 'sentence number' can be regarded as a connective element in the group TEXT since it implies the information of the order and the position of individual sentences that are values of 'sentence'.

The second one analyses the arrangement of the values of the elements of each group. For instance, in the group TEXT, values of the pairs of 'claasification category' and 'sentence' are repeated. The number of the pairs is indefinite. In general, the value of data element occurs repeatedly in a group in combination with each value of the other data elements of the group. Such a set of the combined values is called "data unit" of a group.

The formal definitions of three basic concepts of DCR model are given as follows:

Def. 1. "element": consists of 'data element' and 'connective element'.

Def. 2. "data unit": a set of each value x_j of each element X_j in a group. It is denoted by (D-id, x_1, ..., x_j, ..., x_n), provided that D-id means the value of Document Identifier (D-ID) and that n is the number of elements in a group.

Def. 3. "group": a set of some data units denoted by {(D-id, x_{i1}, ..., x_{ij}, ..., x_{in}) | x_{ij}εX , ($1 \leq i \leq m$)}, provided that the number of data units is m and that x_{ij} is the value of X_j of the i-th data unit. Suppose G is a group name, then the group G is represented by G=(D-ID , X_1, ..., X_j, ..., X_n).

Some examples of the above in scientific documents are as follows.

Ex. 1. 'data element': 'author', 'sentence', 'title'.

Ex. 2. 'connective element': 'sentence number' and 'document identifier'(D-ID).
D-ID can be regarded as a connective element since data units having the same D-id can be connected as data units included in the same document.

Ex. 3. Examples of group BIB
1) "element" and "data unit"
The data unit of the group BIB consists of one connective element D-ID and four data elements: 'Journal' (JNL), "Volume' (VOL), 'Publishing year' (YEAR) and 'Page' (PAGE). '...' denotes each element and the succeeding (...) describes the corresponding element name. Suppose each value of elements in a document x is denoted as D-id^x ε D-ID, j^x ε JNL, v^x ε VOL, y^x ε YEAR and p^x ε PAGE. Then, the data unit is denoted by (D-id^x, j^x, v^x, y^x, p^x).
2) "group"
Group BIB consists of a unique data unit.
BIB=(D-ID, JNL, VOL, YEAR, PAGE).

Ex. 4. Examples of group TEXT
1) "element"
The data elements of the group TEXT are 'classification category' (CAT) and 'sentence' (SENT). D-ID and 'sentence No.' (SENTNO) can be given as the connective elements of the group TEXT. That is, the elements of the group TEXT consist of two connective elements (D-ID,SENTNO) and two data elements (CAT, SENT).
2) "data unit"
Suppose the i-th value of each element of a document x is D-id^x ε D-ID, n_j^x ε SENTNO, c_j^x ε CAT, and s_j^x ε SENT. Then a data unit of the group TEXT is denoted by (D-ID^x, c_j^x, n_j^x, s_j^x) ($1 \leq i \leq m$).
3) "group"
Group TEXT consists of multiple data units. It is denoted by {(D-id_i^x, c_i^x, n_i^x, s_i^x) | ($1 \leq i \leq m$)}, provided m means the number of data units of a document x.
TEXT=(D-ID, CAT, SENTNO, SENT).

Next, two notations of relationships between D-ID and the other elements A_j ($1 \leq j \leq m$) in a group G are given as below.
Suppose G=(D-ID, A_1, ..., A_j, ..., A_n),
1) if group G has only a data unit,
D-ID $\longrightarrow$ (A_1, ..., A_j, ..., A_n);
2) if group G has multiple data units,
D-ID $\twoheadrightarrow$ (A_1, ..., A_j, ..., A_n).
In the case of the latter, the number of data units is generally given dependent on documents such as n=0, 1, ...

A group with the property 1) is called a single group and one with the property 2) is called a repeating group. Every group in a document is either a single group or a repeating group. Most of the groups in a document are repeating groups.

According to the relationship notations, group BIB and TEXT can be represented as follows.
group BIB : D-ID $\longrightarrow$ (JNL, VOL, YEAR, PAGE)
group TEXT: D-ID $\twoheadrightarrow$ (CAT, SENTNO, SENT)

2.2 DCR model

Suppose group G=(D-ID, A_1, ..., A_j, ..., A_n) and a_{kj}^x is the value of A of the k-th data unit of G in a document x, then each data unit of G is defined by (D-id^x, a_{k1}^x, ..., a_{kj}^x, ..., a_{kn}^x) ($1 \leq k \leq m$), provided m is the number of data units (if G is a single group, m=1). The DCR model is formed with some groups consisting of sets of data units of n+1-tuples of elements, which correspond with the contents of a document.

The data units $(D\text{-}ID^x, a_{k1}^x, \ldots, a_{kj}^x, \ldots, a_{kn}^x)$ $(1 \le k \le m)$ can be represented as a table of m rows and n+1 column. In Fig. 2, the tabular representation is shown. It is called a DCR table corresponding to a group of a document. In Fig. 2, a group name is labelled at the top of the table and the element names are put in the first row. In the table, a row corresponds to each data unit and a column corresponds to each element of a group, respectively.

Group name

D-ID	A_1	A_2	....	A_n
$D\text{-}id^x$	a_{11}^x	a_{12}^x	....	a_{1n}^x
$D\text{-}id^x$	a_{21}^x	a_{22}^x	....	a_{2n}^x
:	:	:	....	:
$D\text{-}id^x$	a_{m1}^x	a_{m2}^x	....	a_{mn}^x

←element name

←data unit

Fig. 2. A DCR table

2.3 Examples of DCR model

Each DCR table of group BIB and TEXT is formally shown in Figs. 3 and 4. The contents of two documents (x=1,2) are represented in the DCR tables. In Fig. 3, j^x, v^x, y^x and p^x are the values of JNL, VOL, YEAR and PAGE, respectively. In Fig. 4, c_i^x and s_i^x are the i-th values of CAT and SENT, respectively. The number of the data units of the document x=1 is 3 and that of the document x=2 is 2.

BIB

D-ID	JNL	VOL	YEAR	PAGE
1	j^1	v^1	y^1	p^1
2	j^2	v^2	y^2	p^2

Fig. 3. DCR table of group BIB

TEXT

D-ID	CAT	SENTNO	SENT
1	c_1^1	1	s_1^1
1	c_2^1	2	s_2^1
1	c_3^1	3	s_3^1
2	c_1^2	1	s_1^2
2	c_2^2	2	s_2^2

Fig. 4. DCR table of group TEXT

3. Mapping Definition Language

The mapping of document contents into the DCR model is carried out as follows. Three types of delimiters are introduced to separate each document, each group and each element, respectively. They are inserted into document data when the data are produced. Then, each element value can be automatically transfered into some DCR tables by a program that interprets these inserted delimiters.

The procedure for mapping document data into DCR tables is given by Mapping Definition Language (MDL). Thus, document contents and the DCR model can be linked by MDL. The MDL is a language that defines the following.

i) Three types of delimiters i.e., a document delimiter, group delimiters and element delimiters.

ii) The order of each value of data elements in each group data stream and that of elements in data units.

iii) Group property i.e., either single or repeating.

iv) Delimiters to generate each value of connective elements automatically.

The auto-generation of the values of a connective element in iv) is very useful since connective elements never exist as real data in document contents.

The syntax of MDL is shown in Fig. 5. A document delimiter and a group delimiter are denoted by enclosing them in double quotation marks and single quotation marks, respectively. An element delimiter is denoted by element name(<delimiter>). In Fig. 5, element name(# by (<delimiter>)) means that a connective number is generated as the value of the element every time the indicated delimiter is found in document data. Element name(*) means that the element takes the value generated already by iv). Anything defined by element name(...) is called a mapping element. [<mapping element>] means a repeating group, in which the mapping elements, i.e., data units, are repeated until a group delimiter appears in the document data.

```
<delimiter>::= one or some English or numeric or
               symbolic character
<group delimiter>::='<delimiter>'
<document delimiter>::="<delimiter>"
<mapping element>::=element name(<delimiter>)|
                    element name(# by (<delimiter>))|
                    element name(*)
<mapping group>::=<mapping element>|
                  <mapping group>,<mapping element>|
                  [<mapping group>]

mapping definition
  group name=<mapping group><group delimiter>,
  .........................................,
  group name=<mapping group><document delimiter>
```

Fig. 5. The syntax of MDL

4. An application of the DCR model and MDL to scientific documents

Sentence retrieval system SCAT-IR has been implemented based on the DCR model and MDL. The system has stored 1,116 short notes and letters on the nuclei of three and four nucleons of theoretical research of nuclear physics. In this section, the application of the DCR model and MDL to the documents is described.

Document contents can be regarded as a stream of data as shown in Fig. 6. In Fig. 6, each value of an element is represented by the element name written with small letters such as journal, sentence, category-no, etc. Classification category are added by classification category numbers at the head of each sentence. The contents of the document can be divided into the following groups: Bibliography (BIB), Title (TIL), Author (ATH), Abstract (ABS), Text (TEXT), Reference (REF) and Comment (COM). The elements of each group are represented as follows.

BIB : D-ID, JNL, VOL, YEAR, PAGE.
TIL : D-ID, TITLE.
ATH : D-ID, AUTHOR, AFFILIATION.
ABS : D-ID, ABSTRACT.
TEXT: D-ID, CAT, SENTNO, SENT.
REF : D-ID, REFNO, JNL, VOL, YEAR, PAGE, OTHERS.
COM : D-ID, SIGN, COMMENT.

Then, each group can be denoted formally as follows.

BIB=(D-ID, JNL, VOL, YEAR, PAGE)
TIL=(D-ID, TITLE)
ATH=(D-ID, AUTHOR, AFFILIATION)
ABS=(D-ID, ABSTRACT)
TEXT=(D-ID, CAT, SENTNO, SENT)
REF=(D-ID, REFNO, JNL, VOL, YEAR, PAGE, OTHERS)
COM=(D-ID, SIGN, COMMENT)

The relationship of D-ID and the other elements in a group is :

BIB : D-ID → (JNL, VOL, YEAR, PAGE)
TIL : D-ID → (TITLE)
ATH : D-ID ↠ (AUTHOR, AFFILIATION)
ABS : D-ID → (ABSTRACT)
TEXT: D-ID ↠ (CAT, SENTNO, SENT)
REF : D-ID ↠ (REFNO, JNL, VOL, YEAR, PAGE, OTHERS)
COM : D-ID ↠ (SIGN, COMMENT)

Next, the delimiters inserted into the data stream in Fig. 6 are as follows.

1) document delimiter : '&&'
2) group delimiter : '@'
3) element delimiters of each group :
   BIB : JNL('Δ'), VOL('('), YEAR(')'), PAGE('@')
         ('Δ' means a space.)
   TIL : TITLE('@')
   ATH : AUTHOR('('), AFFILIATION(')')
   TEXT: CAT('Δ'), SENT('..')
   REF : REFNO('>'), JNL('Δ'), VOL('('), YEAR(')'),
         PAGE('Δ'), OTHERS('..')
   COM : SIGN('Δ'), COMMENT('..')

The delimiter '..' of SENT and OTHERS shows that another period is added to '.' at the end of a sentence to distinguish a period at the end of a sentence from that of an abbreviation. The delimiter 'Δ' may occur more than once in the data stream. The result of inserting the above delimiters is shown in Fig. 7. The definitions by MDL in Fig. 8 are given to transfer each value of the data stream in Fig. 7 into the DCR tables as shown in Fig. 9. Consequently, the document contents and the DCR model can be linked by the definitions written in MDL and be mapped into the DCR model easily and naturally.

journal volume year page title author$_1$ affiliation$_1$ author$_2$ affiliation$_2$ ......
author$_i$ affiliation$_i$ abstract category-no$_1$ sentence$_1$ category-no$_2$ sentence$_2$ ...
.... category-no$_k$ sentence$_k$
reference-no$_1$ journal$_1$ volume$_1$ year$_1$ page$_1$ others$_1$ ....
reference-no$_m$ journal$_m$ volume$_m$ year$_m$ others$_m$ comment-sign$_1$ comment$_1$ ....
comment-sign$_n$ comment$_n$

Fig. 6. Data stream of scientific document contents

journal Δ volume (year) page @ title @ author$_1$ (affiliation$_1$) author$_2$
(affiliation$_2$) .... author$_i$ (affiliation$_i$) @ abstract @ category-no$_1$ Δ
sentence$_1$.. category-no$_2$ Δ sentence$_2$.. ....
category-no$_1$ Δ sentence$_k$.. @ reference-no$_1$ > journal$_1$ Δ volume$_1$ (year$_1$)
page$_1$ Δ others$_1$.. ....reference-no$_m$ > journal$_m$ Δ volume$_m$
(year$_m$) page$_m$ Δ others$_m$.. @ comment-sign$_1$ Δ comment$_1$.. ....
comment-sign$_n$ Δ comment$_n$.. &&

Fig. 7. Result of inserting delimiters

```
BIB=D-ID(# by (&&)), JNL(Δ), VOL((),  YEAR()), PAGE(@) '@',
TIL=D-ID(*), TITLE(@) '@',
ATH=[D-ID(*), AUTHOR((), AFFILIATION())] '@',
ABS=D-ID(*), ABSTRACT(@) '@',
TEXT=[D-ID(*), CAT(Δ), SENTNO(# BY (..)), SENT(..)] '@',
REF=[D-ID(*), REFNO(>), JNL(Δ), VOL((), YEAR()), PAGE(Δ), OTHERS(..)] '@',
COM=[D-ID(*), SIGN(Δ), Comment(..)] "&&"
```

Fig. 8. Mapping definition of scientific documents by MDL

BIB

D-ID	JNL	VOL	YEAR	PAGE

TIL

D-ID	TITLE

ATH

D-ID	AUTHOR	AFFILIATION

ABS

D-ID	ABSTRACT

TEXT

D-ID	CAT	SENTNO	SENT

REF

D-ID	REFNO	JNL	VOL	YEAR	PAGE	OTHERS

COM

D-ID	SIGN	COMMENT

Fig. 9. DCR tables of scientific documents

5. A Query Language SCAT-QL1

SCAT-QL1 (SCAT Query Language 1) is a query language implemented on the SCAT-IR system. In this section, the basic retrieval method based on SCAT-QL1 and advanced information retrieval, and some examples are shown.

5.1 Basic retrieval method of SCAT-QL1

The basic retrieval method of SCAT-QL1 is a simple pattern matching. Each datum corresponding to document contents, for example stored sentence data and query expression given by a user, is regarded as a character string. The pattern matching method makes it possible to retrieve sentences including not only words but the indicated parts of a word, a compound word and any character string.

Generally, it is useful to deal mechanically with a misspelling, an inflection and the different representation such as 'three body' and 'three-body'. Therefore, it is effective in giving the number n of unmatched characters permitted on pattern matching. For example, if n=1, then 'three body' and 'three-body' are regarded as the same character string. The component of query is denoted by (s,n), provided 's' is a character string to search.

5.2 Advanced information retrieval methods

The expression of query for information retrieval is defined as follows.

```
<component>::=(s,n)
<element expression>::=<component> |
        <element expression> OR <component> |
        <element expression> AND <component> |
<query expression>::=element-name=<element ex-
        pression> | <query expression> OR
        element-name=<element expression> |
        <query expression> AND element-name=
        <element expression>
```

A query expression is given by AND/OR formula of element-name=<element expression>, provided element-name corresponds to one defined in the DCR model. Information retrieval in SCAT-IR system consists of one of within a DCR table and one between DCR tables. The former is retrieval by AND/OR formula of components for some elements in a DCR table. The latter is retrieval by AND/OR formula of element expressions of some appropriate elements in some DCR tables. The above two kinds of formulas make it possible to support advanced information retrieval. Advanced information retrieval is composed of sentence retrieval, citation relation retrieval, document retrieval etc.

SCAT-QL1 has some functions for manipulating the resultant various retrieval. The basic functions of SCAT-QL1 are :

1) to make a set of retrieved data units or document identifier (D-ids) etc,
2) to perform retrieval on the resultant sets,
3) to store or request the sets,
4) to execute the boolean operation among the sets,
5) to put out each values in all the data units or in those of the resultant sets,
6) to retrieve the relation of citation of documents,
7) to put out cited number and cited rate for some documents with a graph.

5.4 Illustrating examples

```
    This command defines a set named PROB34 consisting of
    |   sentences satisfying this AND/OR expression.
    ↓                                                      category No.
                                           or        and         ↓                  Search
1)  DEF PROB34=(STW=(THREE+FOUR+ALPHA)*(BODY+PARTIC+NUCLEON))*(CAT=01~03);           time
                                                  ↑                                60 sec.
    ** PROB34 ** HITCOUNT 207 SENTENCES (123 DOCUMENTS) scientific terms included in a sentence
                       ↑               ↑
                  number of hit senences, number of documents including the hit sentences
2)  DEF YEAR70=(YER=70~);

    ** PROB34 ** HITCOUNT 522 DOCUMENTS

3)  DEF RECENT=PROB34*YEAR70;

    ** RECENT ** HITCOUNT 107 SENTENCES (61 DOCUMENTS)

                       category No.
                          ↓
4)  PRT (RECENT) D-ID SNT(02 06) JNL VOL YER PAG ATH TIL;

    D-ID=00111
    SENTENCE OF CAT NO. = 02 ←Basic problems
    (0001) IN A RECENT LETTER PASK 1> HAS COMMENTED THE ROLE OF NON-CENTRAL
           INTERACTIONS IN LOW-ENERGY NUCLEAR PHYSICS, ESPECIALLY IN THE     ← a retrieved sentence
           THREE-NUCLEON SYSTEMS.

    SENTENCE OF CAT NO. = 06 ←Individual problems
    (0013) HOWEVER, THE PURPOSE OF THIS NOTE IS TO FIND A REASON FOR THE     ← a sentence require
           UNEXPECTED RESULTS FOR THE SQUARE WELL.                                by PRT command
    JOURNAL=PHYS. LETTERS
    VOLUME =B 34
    YEAR   =1971
    PAGE   = 184
    AUTHOR =R. VAN WAGENINGEN AND G. ERENS* (NATUURKUNDIC LABORATORIUM DEF
             VRIJE UNIVERSITEIT, AMSTERDAM, THE NETHERLANDS)
    TITLE  =SQUARE-WELL POTENTIAL AND THREE-NUCLEON BOUND-STATE PROPERTIES
      .                       .                          .
      .                       .                          .
     This command means to search documents of the set PROB34 citing each
      |  document of the set PROB34 and to count the times.
      ↓
5)  SRC CED(PROB34) FROM (PROB34) BY (REF);                                          Search
6)  PRT CED% BY GRH; ← This means to print out the cited rate of each documents       time
                                                           in the set PROB34.        20 sec.
              0    5   10   15   20   25   30   35  CITED RATE
    00474(59) |*   |    |    |    |    |    |    |     0.8
    00122(60) *    |    |    |    |    |    |    |     0.0
    00108(62) |    |    |    |  * |    |    |    |    18.3
    00321(62) |*   |    |    |    |    |    |    |     0.8
        :
        :
    00296(70) |    |   *|    |    |    |    |    |     9.8
    00299(70) *    |    |    |    |    |    |    |     0.0
    00302(70) |    |    |    |    | *  |    |    |    22.9
        : ↑
      a document of high cited rate
           ↓
7)  PRT (00302) SNT(06 10) JNL VOL YER PAG ATH TIL;

    SENTENCE OF CAT NO. = 06
    (0020) IN ORDER TO INVESTIGATE THE 3HE WAVE FUNCTION AT SMALL DISTANCES
           IT WAS NECESSARY TO INCREASE THE MOMENTUM TRANSFER SIGNIFICANTLY.
    (0035) THE SEPARATION OF THE CHARGE AND MAGNETIC FORM FACTOR WAS ACCOMPLISHED
           BY USING ROSENBLUTH PLOTS AS A FUNCTION OF THE EFFECTIVE Q.
    SENTENCE OF CAT NO. = 10
    (0067) DETAILED CALCULATIONS OF THIS DIFFERENCE USING PREVIOUS VALUES FOR
           THE FORM FACTORS HAVE TENDED TO BE APPROXIMATELY 0.1 MEV TOO LOW 19> :
           THE NEW DATA REPRESENT A MORE DIFFUSE ELECTROMAGNETIC STRUCTURE FOR 3HE
           AND THEREFORE WOULD TEND TO INCREASE THIS DISAGREEMENT.
    JOURNAL=PHYS. REV. LETTER
    VOLUME =  25
    YEAR   =1970
    PAGE   = 884
    AUTHOR =J. S. MCCARTHY, I.  SICK,** R. R. WHITENEY, AND M. R. YEARIAN
            (HIGH ENERGY PHYSICS LABORATORY AND DEPARTMENT OF PHYSICS,
            STANFORD UNIVERSITY, STANFORD, CALIFORNIA 94305)
    TITLE  =ELECTROMAGNETIC STRUCTURE OF THE 3HE NUCLEUS*
```

The search time of the example 1) is 60 seconds and that of the example 5) is 20 seconds.

6. Remarks and conclusion

In this paper, we have proposed a Document Contents Representation (DCR) model to deal with the entire contents of scientific documents in an information system. The advantages of the DCR model are as follows.

1) The representation of each group in the DCR model is independent of the number of data units. Consequently, although the number of data units of each group of a document takes generally an arbitrary value such as 0, 1, ..., k, every group can be represented on the DCR model in the same manner. That is, the DCR model is a general-purpose model.
2) The formation of DCR model for scientific documents is easily carried out by identifying the data units of each group. It can be performed even by end-users.
3) Physical order of values of data units can be converted into logical information like a sentence number. As a result, the manipulation of value data can be given as a logical manipulation in an information system.

Next, we have devised a Mapping Definition Language (MDL) to map document contents into the DCR model naturally and directly. In addition, we have demonstrated the application of the DCR model and MDL to scientific document contents.

The development of the SCAT-IR system based on the DCR model and MDL has been done on HITAC 8250 computer system and is now available on the HITAC M170 computer system of the Center for Information Processing Education of Hokkaido University. 1,116 short notes and letters on nuclear physics have been stored. (The input data consists of 150 thousand cards and it occupies 17MB on a disk.) Various retrieval experiments by a query language SCAT-QL1 have been performed for the stored data. SCAT-QL1 supports sentence retrieval and citation relation retrieval, not to mention document retrieval. More advanced utilizations of the SCAT-IR system have been studied by combining the above basic retrieval methods, e.g., the grasping of research trends on some problems, the analyses of the status of some topical problems, utilizations in the study of graduate students etc, and the effectiveness of such usage has been proved.[1)]

It may be concluded that our attempts to retrieve the entire contents of a document suggest a new information system in a stage from document retrieval to document contents retrieval i.e., information retrieval, and that they suggest the future development of advanced information retrieval system in a research laboratory etc. An future problem is the reduction of the work for producing input data and the mechanization of its process. In addition, some applications of the DCR model to other fields will be a future concern.

Acknowledgement

The authors heartily appreciate Prof. Y. Akaishi, Prof. T. Hiura, Prof. H. Bando, Prof. R. Tamagaki, and Prof. S. Nagata for their utilizations of the SCAT-IR system and their practical and kind advices. Further the authors sincerely thank Mr. T. Maeda for his kind discussions. The authors are also grateful to Mrs. T. Matsumoto for her serious help in developing the SCAT-IR system, and to Miss U. Kohata and H. Yamamoto for their kind help in punching.

References

1> H. Tanaka and M. Nikkuni, Infological approach to research trend analysis, in: H. Inose (Ed.), Research on Scientific Information Systems in Japan (published at the Computer Center, University of Tokyo, Japan 1980), 191-199. (in English)

2> H. Tanaka and M. Nikkuni, "An Approach to Retrieval of Conceptional Contents of Scientific Information", Proceeding of 7-th International CODATA Conference, October 1980, to appear.