Lib sc. 5; 1968; PAPER P.

DESIGN OF THE DOCUMENT FINDING SYSTEM: GENERAL FEATURES.

(Non-conventional methods in document retrieval. 4).

A NEELAMEGHAN, *Documentation Research and Training Centre, Bangalore* 3.

[Mentions the objectives of the series of experiments undertaken in the DRTC on the use of computer in document finding. The advantage of using a freely-faceted analytico-synthetic depth version of the Colon Classification in computer-based retrieval system is emphasised. The need for reader-librarian dialogue in the precise formulation of the query is pointed out. The provision for browsing by the reader among the entries retrieved in response to a query to facilitate the pin-pointed selection of documents is discussed. Gives a brief outline of the steps involved, and the equipment used in the experiments on feasibility study].

1 Objectives of the Experiments

11 MAIN

The main objectives of this series of experiments are to examine the feasibility of using a computer in a Document Finding System in which

1 The entries for the documents forming the input to the system are classified according to a Freely-faceted Analytico-synthetic Classification, such as the newly developing version of the Colon Classification (= CC);

2 The Class Number of the subject of each of the documents is synthesised by the computer when fed with the kernel terms of the subject;

3 The search and selection of entries for the documents relevant to the subject of a query are based on facet-analysis and the Class Number of the subject of the query constructed according to the Freely-faceted Analytico-synthetic scheme;

4 The Class Number for the subject of the query is synthesised by the computer when fed with the kernel terms of the facet-analysed subject of the query;

5 Facility is provided for browsing by the reader among the entries selected if they are large in number; and

6 Provision is made for the selection of entries for documents by such other elements as the Name of Author, Terms in the Feature Headings, and Title of the Host Document, in the entries for them.

12 Additional

In addition, techniques for the teaching of the use of computer in the Document Finding Systems were also considered. The drawing of flow-charts for the preparation of the catalogue-on-tape, updating it, and retrieval of documents were therefore done by the students. Elements of the methods of programming were also dealt with. The managerial aspect of the use of the computers in document finding, especially the factors involved in estimating the cost of the Document Finding Systems, were also considered.

2 Basic Elements of the System

21 Freely-Faceted Classification

A Freely-Faceted Analytico-Synthetic classification is a scheme for classification

1 Guided by explicitly stated postulates and principles for the three planes of work — Idea Plane, Verbal Plane, and Notational Plane; and

2 Involves

21 The analysis of the subject into its basic facet and isolate facets in the Idea Plane;

22 Transformation of the result of the analysis into focal terms in the Verbal Plane using current standard terminology;

23 Their translation into Focal Numbers in the Notational Plane according to the scheme for classification; and

24 Synthesis of the Focal Numbers into the Class Number.

Here, the analysis of the subject is done in the Idea Plane leading to the arrangement of the facets of a compound subject in a helpful way. Synthesis is done in the Notational Plane implementing the findings in the Idea Plane in respect of the sequence of the isolate facets and expressiveness of the structure of the subject. Such a scheme for classification can give a coextensive Class Number and facilitate quick pin-pointed retrieval — that is, minimise noise and leakage. Therefore, it would be advantageous to use such a scheme in structuring the subjects of the input documents and of the queries put to the system, and to combine it with the quick search facility of the computer.

22 STORE AND SELECTION

The store of the system consists essentially of a catalogue-on-tape; this is searched to select entries for the documents answering a query. An entry in the catalogue consists of,

1 Class Number for the subject of the document — constructed using the Postulational Method according to the depth version of the Colon Classification;

2 Feature Heading — that is, the translation, into the natural language, of each of the different foci in the Class Number; and

3 The bibliographical specification of the host document.

The search and selection was based on

1 Facet Analysis of the subject of the query;

2 Expressing the kernel terms in standard terms used in the scheme for classification;

3 Constructing the Class Number for this facet-analysed query in standard terms on the basis of the scheme used in classifying the input documents;

4 Matching the components of this Class Number with the components of the Class Numbers in the entries of the catalogue-on-tape; and

5 Providing a printout of the entries retrieved in the format required.

Programs have also been drawn up

1 To punch out the resulting output on cards for other use; and

2 To transfer the output on to another magnetic tape for other use.

23 READER-LIBRARIAN DIALOGUE

231 *Need for Dialogue*

The dialogue between the reader and the librarian is an essential step in the precise formulation of the query. This dialogue has not been mechanised in the system. The need for and the helpfulness of this dialogue are briefly discussed in Sec 232 to 235. A fuller treatment about this may be found in another paper (2).

232 *Query Statement by Reader*

A reader's need is usually communicated in the form of a statement in a natural language. The majority of such queries may be about a subject. Query statement about a subject raises problems.

233 *Problems of Query Statement*

A subject may be defined as a systematised body of ideas whose extension and intension are likely to fall coherently within

the field of interest and comfortably within the intellectual competence and the field of inevitable specialisation of a normal person. Ideas relate to some subject—concrete or abstract. The reader may not be quite clear in his mind about the idea about which he needs information. The words in the natural language known to him may be inadequate to express the idea. Further, his ability to use the available words to express the idea may be inadequate. As a result, the statement in words may not be coterminous with the idea.

234 *Problems of Query Analysis*

The person analysing the reader's statement may not be able to form a complete picture of the requirements of the reader. The words of the natural language known to him may be inadequate to express his understanding of the reader's needs. His ability to use the natural language may be inadequate. Thus, the result of a query analysis may be only an approximate representation of the reader's actual specific subject-need.

235 *Personalised Service*

The librarian has, therefore, to gain an intimate knowledge of the reader, not only his subject-need at the moment, but also his individuating particularities. He should develop a Reader Profile Record, for each reader. Above all, a carefully planned dialogue with the reader would be of help in grasping the latter's need better. Such a strategy is necessary:

1 To help the reader to:

11 Recall as many isolate ideas as possible in the subject of his interest at the moment;

12 Conceptualise the isolate ideas in a precise way;

13 Express the isolate ideas preferably in the standard technical terms used in the document retrieval system.

2 To help the librarian to

21 Understand precisely and unerringly the reader's subject-need;

22 Supply any missing facet in the reader's query such that the retrieval process would become more efficient;

23 Arrange all the constituent facets of a subject in a sequence to represent coextensively the reader's subject-need;

24 Convent the facet-analysed query into the language of the document retrieval system to facilitate the search and retrieval; and

25 Modify the search strategy, if necessary, at each stage, on the basis of the continuous feed-back arising from the reader-librarian communication context.

3 Browsing among Entries

31 Query Formulation in the Conventional Way

In the experimental system, the formulation of the query in precise and standard terms, the facet analysis of the query, and the construction of the Class Number for it, have been done in the conventional way — that is, after a dialogue between the reader and the librarian. The helpfulness of exposing the reader to a well-designed schedule of subjects and their class numbers in the precise formulation of the query has been discussed in another paper (2). The display of appropriate parts of the schedule on a video system is now beginning to be practised in computer-based document retrieval systems (1). An experiment describing the feeding of the computer with the query in the form of a statement in a natural language, the synthesis of the Class Number by the computer on the basis of a built-in schedule, and subsequent search and selection of the documents in the store on the basis of the Class Number, is described in Paper S in this issue.

32 Need for Browsing

Even after the reader-librarian dialogue, the analysis of the query may not be coterminous with the reader's subject-needs at the moment. In the conventional classified catalogue, the reader may browse among the entries in the classified part after having arrived at a point therein containing the entries for the documents in the penumbral area or very nearly the umbral area of his subject-interest. This browsing also has heuristic value to the reader.

33 Provision for Browsing

In the experimental system, the entries are punched on cards and transferred on to the magnetic tape to form a catalogue-on-tape input to the computer. Browsing among the entries selected in answer to a query has been provided for. Browsing requires the arrangement of the entries in a helpful sequence resulting in an APUPA pattern. Such an arrangement is facilitated when the documents are classified according to a Freely-faceted version of CC and the entries arranged according to the ordinal value of the digits used in it. In the experimental system, the entries could be arranged by the Class Numbers in the print-out of the list of documents selected in response to a query. The format chosen for the entries in this list gives also the feature headings to facilitate rapid scanning of the entries and selecting those most relevant to the query.

4 Selection Points

The entry for the input document consists of the Class

Number constructed according to a Freely-faceted version of CC, the translation of the Class Number into the natural language to form the Feature Headings, and bibliographical details about the Host Document. Suitable programs could be drawn up to find documents on the basis of any one or more of the elements in the entry such as:

Class Number;
Kernel Terms in the Feature Heading;
Name of the author;
Words in the title of the document; and
Any of the elements in the host section — for example, by the title and volume number of a particular periodical.

5 Phases and Steps

It was found helpful to divide the work of setting up the system into different phases. Each phase was further divided into a helpful sequence of steps. The objective, the work involved, and the result of each step could then be more clearly envisaged and discussed. Further, each of the corresponding steps could be identified in a comparable manual document finding system. This has been particularly helpful in teaching.

51 Phase 1: Creation of Catalogue-on-Tape

Stage 1 in Phase 1, called the Catalogue-on-Tape, begins with the classification and preparation of the main entries for the documents forming the input to the system. It ends with transferring these entries on to magnetic tape, and making a print-out, if desired. The steps are shown in Fig 1 in Appendix 1.

52 Stage 2: Amending the Catalogue-on-Tape

Stage 2 in Phase 1 consists in making provision for correcting the entries on the tape, if necessary, and for the addition of new entries. The steps in amending the catalogue and getting a print-out of the amended catalogue are shown in Fig 2 in Appendix 2.

53 Phase 3: Search and Selection

Phase 2 called 'Retrieval', begins with the Facet Analysis and classification of the subject of the reader's query. It ends with giving a print-out of the list of entries of the documents, if any, answering the query. Provision has also been made to transfer the selected entries on to a magnetic tape, and also put them on punched cards. The steps are shown in Fig 3 in Appendix 3.

6 Equipment Used

61 Computer

610 ICL 1903

The computer used in the experiments is an ICL 1903. It is a digital computer, manufactured and sold by the International Computers Limited. It is a character oriented machine. Functions are available to compare every four characters and to execute different sets of instructions depending on equality. The particular features of the computer are enumerated in the succeeding sections.

611 *Internal Storage*

Core: Ferrite
Access time: 2 micro sec
Capacity:
  Minimum: 8192 words
  Maximum: 32768 words
Alphabetic base: 6 bits
Word length (excluding sign): 24 bits
Magnetic Tape Storage
  Number of units: 4
  Transfer rate: 60,000 characters per sec
  Density per inch: 800 bits per inch
  Speed of movement: 75 inches per sec
  Width: 0·75 inch
  Number of channels: 7 or 9.

612 *Arithmetic and Logic*

Arithmetic mode: Parallel with simultaneous carry
Numeric mode: Binary
Speed (in sec):
  Addition: 7 micro sec
  Multiplication: Variable, approx 250 micro sec
  Division: Approx 900 micro sec
Number of Index Registers: 3
Number of Commands: Approx 100
Length of Instruction: 24 bits
Address per Instruction: 1

613 *Punched Card Input*

Number of units: 1
Input speed: 300 cards per min
Output speed: 100 cards per min

614 *Output Printer*

Type: Line printer

Number of units: 1
Number of print positions: 120
Number of lines per minute: 300

62 Punched Card

The punched card used is the standard Hollerith card.

7 Character Code

71 64-Character Card Code

Symbol	Card Punching	Symbol	Card Punching	Symbol	Card Punching
0	0	F	10/6	—(Minus/hyphen	11
1	1	G	10/7	" (Quotes)	10/0
2	2	H	10/8	/ (Solidus)	0/1
3	3	I	10/9	+ (Plus)	10/2/8
4	4	J	11/1	. (Stop)	10/3/8
5	5	K	11/2	; (Semi-colon)	10/4/8
6	6	L	11/3	: (Colon)	10/5/8
7	7	M	11/4	' (Apostrophe)	10/6/8
8	8	N	11/5	! (Exclamation)	10/7/8
9	9	O	11/6	[(Left bracket)	11/2/8
Space	NONE	P	11/7	$ (Dollar)	11/3/8
& (ampersand)	10 or 10/0	Q	11/8	* (Asterisk)	11/4/8
±	3/8	R	11/9	> (Greater than)	11/5/8
@	4/8	S	0/2		11/6/8
((left parenthesis)	5/8	T	0/3	↑ (Upward arrow)	11/7/8
) (right parenthesis)	6/8	U	0/4	£ (Pound)	0/2/8
] (right bracket)	7/8	V	0/5	, (Comma)	0/3/8
A	10/1	W	0/6	% (Percentage)	0/4/8
B	10/2		0/7	? (Question)	0/5/8
C	10/3		0/8	= (Equals)	0/6/8
D	10/4		0/9	←(Backward arrow)	0/7/8
E	10/5				

8 Flow-Chart Symbols

The symbols used in the flow-chart are shown in Fig. 4 in Appendix 4. They generally correspond to the symbols recommended by the American Standards Association Committee on Computers and Information Processing.

91 PROGRAM

The program was written in PLAN (Programing Language for Nineteen Hundred) to optimise on the available storage locations and to provide an efficient program.

92 ACKNOWLEDGMENT

The authors of the papers published in this issue are grateful to the Hindustan Machine Tools, Bangalore, for the computer facility made available for the experiments. The authors also wish to express their thanks to the International Computers Limited for permission to publish the reports on the work done on the computer ICL 1903.

93 Bibliographical References

1 Sec 31 FREEMAN (R R) and ATHERTON (P). AUDACIOUS: an experiment with an on-line, interactive reference retrieval system using the Universal Decimal Classification and the index language in the field of nuclear science. 1968. (AIP UDC Project, report AIP/UDC-7).

2 Sec 231 31 NEELAMEGHAN (A). Integrated approach of India to the design of a document retrieval system. (Paper contributed to the Inter Congr on Sci Inform (34) (Moscow) (1968)).

94 APPENDIX 1

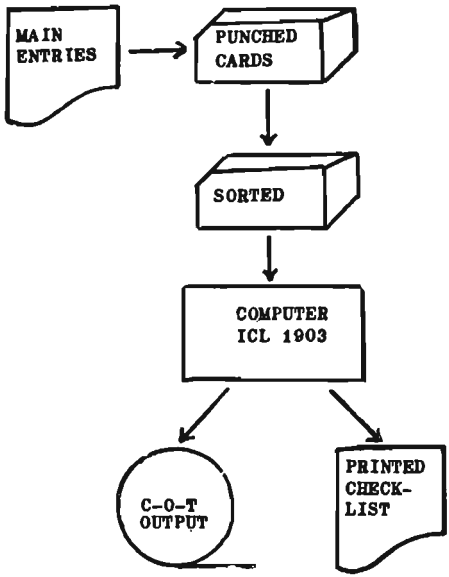


FIG 1 Phase 1, Stage 1: Catalogue-on-Tape

95 APPENDIX 2

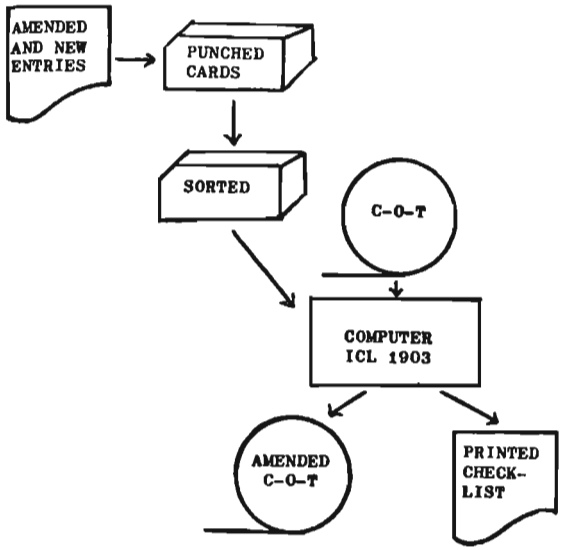


FIG 2. Phase 1, Stage 2: Amending the Catalogue-on-Tape

96 APPENDIX 3

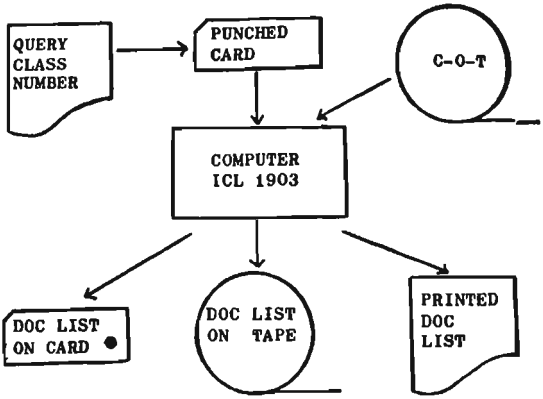


FIG 3. Phase 2: Retrieval

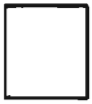
97 APPENDIX 4

DOCUMENT/READER'S QUERY

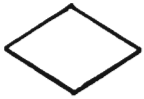
PUNCHED CARD

MAGNETIC TAPE

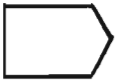
INPUT/OUTPUT



PROCESSING



DECISION



PROGRAM MODIFICATION

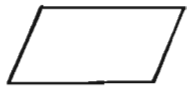
TERMINAL

FIG 4. Flow-chart Symbols

ANNOTATION

DIRECTION OF FLOW

SORTING OR COLLATING



MANUAL OPERATIONS

OFF-PAGE CONNECTION

SUB-ROUTINE

FIG 4. Flow-chart Symbols (contd.)