DST SEMINAR ON

Application of Computers to Bibliographical
Information Processing : Some Developments in India
(Bangalore)(10-13 July 1978)

SEMI-AUTOMATIC METHOD OF PREPARING THESAURUS FOR A SPECIFIC SUBJECT-FIELD

RANJITA MAITRA, Documentation Research & Training Centre, Indian Statistical Institute, Bangalore 560 001.

1 SCOPE OF THE PAPER

An earlier paper mentioned the different components of a thesaurus and the identification and display of the Hierarchical (HR) and Non-hierarchical Associative Relationships (NHR) among the concepts enumerated in the thesauru~1)Jt was also shown that the different varieties of NHR could be represented by one or other of the following types of relationships based on S R Ranganathan's General Theory of Subject Classification (2):
Facet relation, Speciator relation, Phase relation, and Coordinate relation. Such relationships can be identified and represented in expressions of subject (subject string) obtained by facet analysis.

This paper discusses briefly a methodology devcof loped for the application computer for the generation of a thesaurus for a specific subject-field displaying HR and various types of NHR (3, 4). The subject strings obtained by facet anclysis and expressed in a natural language are manipulated by computer using a suitable algorithm to generate the NT/BT links and RT links for representing the HR and the different kinds of NHR respectively identified among the enumerated concepts. In the earlier work of computer manipulation of classification scheme for generating a thesaurus (5)» it was possible to incorporate in the thesaurus the HR and only one type of NHR, namely, the coordinate relationship. In another approach based on subject string manipulation (1), it was possible to generate HR and all NHR except the coordinate relationship.

The label 'RT' in a thesaurus merely indicates NHR (other than equivalence relationship) between concepts. It has been found helpful to specify the type of NHR using suitable phrases and also to group "the concepts on the basis of these specific relationships. Here are two examples:

```
SALT                           PROTON IRRADIATION
  RT (AGENT IN                   BT  IRRADIATION
     PRODUCTION  OF  RT  (-ACTION ON)
     SOLAR ENERGY                    LITHIUM SOLAR CELL
  RT (SOURDE FOR -)                  SILICON SOLAR CELL
     SEAWATER                        (COORDINATE IDEAS)
  RT (METHOD FOR                     BETA IRRADIATION
     PRODUCTION OF -  )              DEUTERON IRRADIATION
    DISTILLATION                     GAMMA IRRADIATION
```

This feature is also incorporated in the thesaurus generated by the methodologydescribed here,.

2 INPUT

The input for computer generation of micro thesaurus for a subject-field consists of the following:

1 Structured subject string (with indicator of relationships among concepts)

2 Phrase-code dictionary

3 Term-code dictionary

21 Input Preparation

The following are the steps for the preparation of the input. The successive steps are briefly described below with a specific example.

C Title of document.- US Government support for civilian technology: Economic theory versus political exigency.

1 Subject String; 1.- Subject string in natural language: A perusal of the document indicates that the specific subject of the document to be:

Evaluation of US Government's economic policy on civilian technology from the point of view of political exigency.

2 Subject String II.- Subject string in kenal terms: By removing all the apparatus words from subject string I, the following subject string as obtained:

Evaluation. USA Government. Economic policy. Civilian technology. Political exigency.

3 Subject String III.- Structured subject string: The kernal terms in subject string II are arranged in a preferred sequence using the Principles of Helpful Sequence of the General Theory of Subject Classification (2). This results in the following subject string.

USA. Government. Economic policy.
Civilian technology. Evaluation.
Political exigency.

4 Subject String IV.- Encoded structured subject string with relationship indicators.

In subject string III the relationship between pairs of concepts is indicated and computer instructions generated using the following codes:

indicating HR. Computer generates NT/BT link between the two connected terns

— indicating Speciator relation (an NHR). Computer generates RT link between the two connected terms

$1 inoxcating Facet relation (an NHR) . Computer generates RT link between the term prefixed with $1 and the immediately preceding term in the string prefixed with the same code.

$2,S3... indicating NHR other than Facet relation and Coordinate relation. Computer generates RT link between the term prevised with $2, or S3, etc, as the case may be and the term in the string prefixed with the same code.

Each NHR is further specified by suffixing a code for the specific relationship identified- To facilitate identification and representation of specific NHR, a table of the different NHR (See also table 2: Census of NHR in the preceding paper) together with an appropriate phrase and a code for each, is maintained. An extract from such a table is given below:

NHR	Phrase	Code	Example
1	2	3	
PROCESS and DEVICE/METHOD/ MEDIUM Used in the process	Device used Method used Medium used	1 2 3	Teaching with AVaids
PROCESS and resulting PRODUCT	Process for Product of	 5	Cooking of food
PROCESS and its PROPERTY	Property of		Detonation waves produced by detonation

The code (column 3) in parenthesis for the phrase (column 2) indicating specific NHR is interpolated in the subject string. The resulting subject: string is as follows:

$1USA$1(-8) Government $1(6-) Economic policy.
(6-) Civilian technology Evaluation-(-9) Political exigency

The data in columns 2 and 3 of the above table are compiled into a computer readable phrase-code dictionary for decoding the computer readable thesaurus at a later stage (See section 43).

Subject strings in coded form are more conveniently manipulated by computer than those expressed in a natural language. To facilitate the encoding of the subject strings and to decode the computer readable thesaurus, a termcode dictionary in computer readable form is maintained. This is also helpful for identification of equivalence relationship (synonym). (See Section 423). An example of an entry in the term-code dictionary is shown below:

> Electrical resistivity / Resistivity / Specific resistence 32 / 33 / 34

Here, the preferred term 'Electrical resistivity' is listed first, followed by the synonyms; the corresponding code numbers for the terms are also in the same sequence.

The terms and their code numbers are maintained in a master term record file which gives complete information about each term is enumerated in the thesaurus. An example of term record is given below:

Sheet No.: 35
Date: 9.1.78

1 SN: 32 2 Reference : 6

3 Term: Electrical resistivity

4 Context: Determination of electrical resistivity of solar

5 Definition: The electrical resistence offered by a material to the flow of current, time the cross-sectional area of current flow and per unit length of current path; reciprocal of conductivity

6 Source: McGraw-Hill Dictionary of Scientific and Technical terms. 1974.

7 Category: Facet

6 BT : Electrical properties

9 Used for: Resistivity (33)
Specific resistance (34)

10 Use:

11 Prepared by: 12 Checked by:

The number within parenthesis in the entry at SN 9 indicates the code number for the corresponding term.

The encoding of the terms in the subject string car- be done manually or by computer as was cone in cur work.

3 OUTPUT

The output thesaurus could be in computer readable form as well as printed out. A specimen of the printed thesaurus is given below:

RESISTANCE
NT RT
RADIATION RESISTANCE

RT (-PROPERTY OF)
WEBBED DENDRITIC SOLAR
CELL

RIGID SOLAR PANEL

BT
SOLAR PANEL

RT (COORDINATE IDEA)
SEMI-RIGID SOLAR PANEL

SALT
/AGENT IN
PRODUCTION OF -)
SOLAR ENERGY

RT (SOURCE FOR -)
SEA WATER

RT (METHOD FOR
PRODUCTION OF -)
DISTILLATION
SALT WATER

U
SALINE WATER

The symbol "–" in prefixed/suffixed to the phrase indicating the specific NHR helps in the meaningful linking of the terms. For example, the following entry,

RESISTANCE
RT (- PROPERTY OF)
WEBBED DENDRITIC SOLAR CELL

would be read as: Registance is a property of webbed Dendritic solar cell.

Whereas, the following entry

SALT
RT (AGENT USED IN PRODUCTION -)
SOLAR ENERGY

would be read as: Solar energy is an Agent used in production of Salt.

It will be noticed that the lead term replaces the dash

The absence of "–" indicates coordinate ideas.

In the following thesaurus entries

```
RIGID SOLAR PANEL
  RT  (COORDINATE IDEA)
      SEMI-RIGID SOLAR PANEL .
```

"Rigid solar panel" and "Semi-rigid solar panel" are coordinate ideas,

4 FROCEDURE

The major steps in generating printed thesaurus by computer are as follows:

1 Preparation of encoded subject string
2 Creation of coded thesaurus entries
3 Deriving thesaurus in natural language
4 Sorting the thesaurus entries
5 Thesaurus print out.

Fig 1 is a system flow chart of the steps.

41 Encoded Subject String

In the input subject string IV (See Sec 21, Step 4) the terms are replaced by the numerical code provided taken from the computer readable term-code dictionary (See sec 21). An example of a computer readable coded subject string is given below:

```
$l81$1(-8)l2$l(6-)23-(8-)4iri(7-)5-(-9)62
```

where,

81	= USA
12	= Government
23	= Economic policy
41	= Commercial technology
5	= Evaluation
62	= Political exigency

At this stage, the non-preferred term in the structure subject string IV may be replaced by the code number for the preferred term. For example, in the above subject, string, the preferred term for "civilian technology' is 'commercial technology'. In the term-code dictionary, the code for commercial technology is 41 and it takes the position of 'civilian technology' in the string.

42 Generation of Coded Thesaurus Entries

421 Processing of Encoded Subject String

A coded subject string Is processed such that each term in the string is linked with the succeeding term in the string, to create entries for the thesaurus. The process is carried out from left to right of the string. This process of generation of entries from a string is iterated until each one of the terms in the string takes the lead term position. Once an entry is prepared, the reverse entry is automatically generated by changing the position of the context term and the lead *term*. In hierarchical entries, the relationship is changed from XT to B BT in reversing the entry. In RT entries, the relationship does not change but the position of '—' is changed from prefix to a suffix and vise versa as appropriate. For example from the coded subject string

a

$18l(-8)12$l(6-)23(8-)4l$1(7-)5-(-9)62

the following entries are generated

S1RT(-8)12	12RTR(8-)81
12RT(6-)23	23RT(-6)12
23RT(8-)41	4lRT(-8)23
5RT(7-)23	23RT(-7)5
5RT(-4)62	62RT(4-)5

Consider another example:

Subject String _I.- Drug treatment of nonfatal lung disease caused by Gram negative bacteria.

Subject String III.- Lung. Disease. Bacteria. Gram negative. Nonfatal. Treatment.

Subject String IV.- $1 Lung $1(6-)S2
Disease -(-15) Bacteria-(20) Gram
negative $2(20-) Non fetal $1(17-) Treatment-(23-)Drug

Coded Subject String: |1245&1(6-)$2
252-(-15)315-(20-)264$2(20-)327$l(l7-)353~(23-)282

From the above coded subject string the following entries are generated

2A5RT(6-)252	252RT(-6)245
252RT(-15)315	315RT(15-)252
315RT(20-)264	264RT(-20)315
327RT(20-)252	252RT(-20)327
353RT(17-)252	252RT(-17)353
353RT(23-)282	282RT(-23)353

If for any of the terms in the input string HR is indicated by' ' ten the computer will generate NI/BT with the appropriate terms.

422 Processing of Coordinate NTs

The set of NTs for a term in the string processed as mentioned in the preceding section is picked up and RT links generated among the NTs so as to derive coordinate relation among them. This is repeated with each of the set of NTs for each of terms in the input string.

Here is an example:

String segments from different subject strings

Microorganism Bacteria

Microorganism Virus

Microorganism Fungi

*

In coded form:

233NT315

233NT318

233MT320

Computer generates the following entries by processing the coded strings

```
315KT(10)518             318RT(10)515
315RT(10)320             320RT(10)315
318RT(10)320             320RT(10)318
```

The code "(10)" represents coordinate relationship.

423 Processing of Term-code Dictionary

The records in the term-code dictionary in computer readable form are processed to derive equivalence relationship wherever applicable.

Entries in the term code dictionary (See Sec 21) will be as follows:

ECONOMIC POLICY 23

COMMERCIAL TECHNOLOGY/CIVILIAK TECHNOLOGY 41/40

The second entry indicates that 41 is the code for "commercial technology" (Preferred term) and that its synonym (non-preferred term) is "civilian technology" coded 40.

By processing this string the following thesaurus entries will be generated indicating the equivalence relation:

41UF40	40U41

43 Thesaurus in Natural Language

Using the term-code dictionary and phrase-code dictionary (See Sec 21), the thesaurus entries in coded form derived by the process described in Sec 42 are decoded into the words of the natural language (in* English in our case). Here is an example:

Coded form	Decoded form
81RT(-8)12	USA RT (-Environment of) Government
23RT(7-)5	Economic policy RT (Action on -) Evaluation
41UF40	'Commercial technology UF Civilian technology
73NT81	Americas NT USA

44 Editing of Thesaurus Entries

In particular subject context, it may not be necessary to generate thesaurus entries under certain common generic terms - e.g. Increase, decrease, Evaluate, produce, etc. Computer instruction can be developed to supress such entries using a term suprcs-sion list. If, for example, from the subject string.

USA Government Economic Policy. Commercial technology. Evaluation. Political exigency one has to indicate the relationship between "Economic policy" and "Evaluation" but not in the reverse way, that is, "Evaluation" and "Economic policy", the 'Term Superssion list' will be of help.

45 Sorted Thesaurus

The decoded thesaurus entries are sorted out and arranged in alphabetical sequence according to the lead term, relationships, and context term. The thesaurus is in computer readable form at this stage.

46 Printout

The thesaurus is printed on on-line printer according to a desired format. (*See* Sec 3)

5 SOFTWARE PACKAGE

A software package for generation of thesaurus as described in this paper has been developed in DRTC. The programs written in COBOL language have been tested on ICL 1901-A computer. A thesaurus on Solar Energy has been produced using this program package.

6 BIBLIOGRAPHICAL REFERENCES

1 NEELAMEGHAN (A) and RAVICHAHDRA RAO (I K). Non-hierarchical associate relationships, their types and computer generation of RT links. (lib sc. 13; 1976; Paper C).

2 RANGANATHAN (S R). Prolegomena to library classification. Ed 3. Assist by M A Gopinath. 1967.

MAITRA (R). Semi-automatic method of generating microthesaurus. 1977. (Project report submitted in part fulfilment of the DRTC course requirements).

—. Semi-automatic method of generating micro thesaurus: A case study in the field of social sciences. (Paper presented to the Annual Seminar (DRTC)(15)(1977). Paper A5).

SHEPHARD (M) and MATTERS (C). Computer generation of thesaurus. Lib sc. 12; 1975; Paper E).

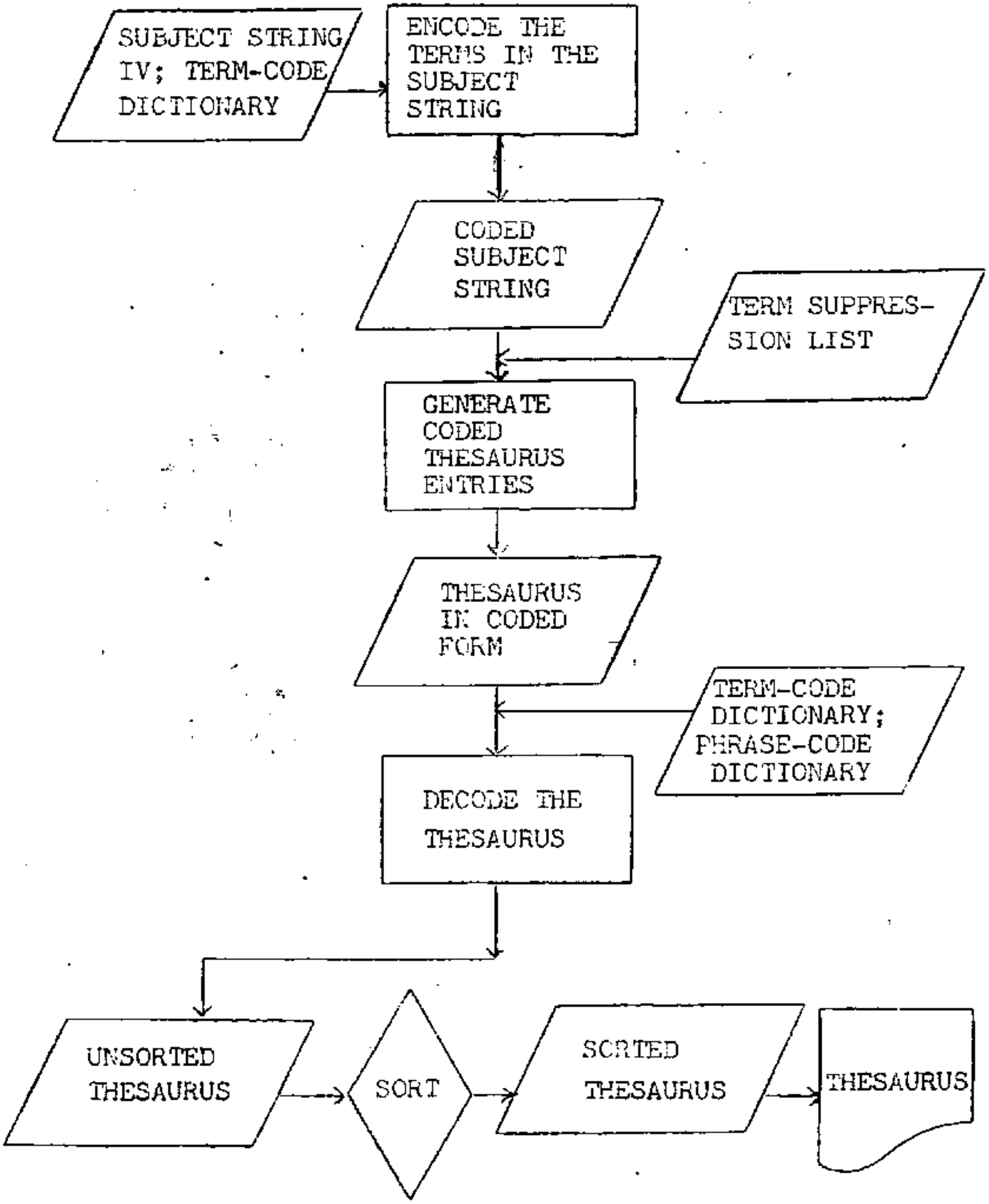
SUBJECT STRING IV; TERM-CODE DICTIONARY
ENCODE THE TERMS IN THE SUBJECT STRING
CODED SUBJECT STRING
TERM SUPPRES-SION LIST
GENERATE CODED THESAURUS ENTRIES
THESAURUS IN CODED FORM
TERM-CODE DICTIONARY; PHRASE-CODE DICTIONARY
DECODE THE THESAURUS
UNSORTED THESAURUS
SORT
SORTED THESAURUS
THESAURUS

Fig 1: System flow chart for generating micro thesaurus