AUTOMATED SYSTEM FOR GENERATING THESAURUS FROM SUBJECT REPRESENTATIONS

P.J. DEVADASON
V. BALASUBRAMANIAN
Documentation Research and Training Centre
Indian Statistical Institute
Bangalore-560001.

Several experiments were conducted at the Documentation Research and Training Centre for generating thesaurus from schedules of classification and subject representation structured according to facet rules and principles for facet analysis. The work reported in this paper uses a coding scheme developed for augmenting the subject strings to make them suitable for generating thesaurus. The system is fully automatic unlike earlier systems. It has five phases namely, Coding phase, Term-pair Generation phase, Coordinate Term-pair Generation phase, Translation phase and Printing phase. The system is described briefly giving the systems flow chart and inputs and outputs of the different phases together with a sample printout of a model thesaurus generating using test data of about 60 subject strings in telephony.

0 INTRODUCTION

01 Thesaurus

A thesaurus is a list of structured compulsory keywords. It could be defined as "a controlled dynamic vocabulary of semantically related terms offering comprehensive coverage of a domain of knowledge". It is a controlled list of terms, with indication of conceptually associated terms, for use in information retrieval systems [3]. The conceptual structure and terminological control provided by a thesaurus contains.

1) Structured systems of concepts with indication of hierarchical and associative relationships between the concepts; and

2) All the terms that designate a particular concept - that is synonymous terms.

The design and construction of thesaurus of terms in subject fields for use in information storage and retrieval has gained much importance in recent years and several thesauri in different subject fields have been published.

02 Computer Generation of Thesaurus

Several experiments were carried out at the Documentation Research and Training Centre in the computer generation of thesaurus. To begin with a faceted scheme for classification for a specific subject field was used in the computer generation of thesaurus [14] in which it was possible to incorporate the hierarchic relationship (HR) of terms and only one type of non-hierarchic relationships (NHR), namely, the coordinate relationship of terms. In another approach, subject structures (subject strings) arrived at by facet analysis based on the general theory of Library Classification was used [13], in which it was possible to generate HR and all NHR except coordinate relationship (CR). In yet another approach based on subject strings [4], apart from generating HR, all NHR including CR were also generated, based on the principle that all the NT's to a particular term are coordinate among themselves. Also a set of typology of NHR [8, 9] were used in categorising the different associative relationships. The procedure used in generating thesaurus in this method requires,

1) the manual preparation of a term-code dictionary in which each of the terms is given an unique serial number and kept in card form, alphabetically arranged for manual reference, and created as a directly accessible file on disk using the serial number as key for machine reference;

2) the manual preparation of a synonym directory having the serial number of the terms and serial numbers of all the synonymous terms to it, created as a directly accessible file on disk using the serial number as key; and

3) encoding of the subject string by replacing the terms by their respective serial numbers and the typology of NHR by their respective codes for use as input to the program [4, 5].

The algorithm for generating the different entries for the thesaurus follows almost a single rule that "each term in the string can create

entries (term-pairs) for the thesaurus with the next term in the string, the process being carried out from left to right of the string" [6]. As the program is 'rule-based' it is rigid in generating the different entries for the thesaurus, in the sense that a term can generate an entry only with its immediately preceding and immediately succeeding terms. If a term has to generate a relationship with a term which is separated by a term occurring in between (the sequence of the terms being determined by facet analysis), then a separate subject string, bringing the two terms in relationship is to be coded and input to the program.

03 Scope of the Paper

In this paper, a set of programs for generating thesaurus 'based on input codes' for generating the different entries i. e. an 'input code based' system for the manipulation of subject strings to generate a thesaurus in a subject field is presented. Also the current system does not require any of the manual work mentioned in Sec 02 above, for its operation.

1 SYSTEM OVERVIEW

The subject structures (subject strings) formulated using the Postulates and Principles of General Theory of Library Classification [11], augmented with codes denoting the typology of relationship between the terms forming the string - are input to a program CODER. This program generates an unique serial number for each of the unique terms in the strings and creates a term-code dictionary and rewrites the input strings as coded strings replacing the terms by their respective serial numbers. The coded subject strings are manipulated by a program GENERATE which generates the different term-pairs forming entries in the thesaurus according to codes supplied along with the strings. The generated entries are sorted and coordinate relationships between NTs to a particular term are generated by another program GENCORD. The output file created by this program is retranslated into the natural language using the term-code dictionary created in the first step and are sorted alphabetically and printed in double column format as thesaurus by a print program. The overall system flow chart is given in Appendix-A.

2 INPUT TO THE SYSTEM

21 Formulation of subject strings

The input to the system is prepared following the specific steps of the Postulational Approach to Library Classification [2, 7, 11].

The title of the document or the raw specification of the reader's query is taken as starting point.

All the specific subjects dealt with in the document/reader's query are determined and expressed in natural language.

Let one of the specific subjects be expressed as "In telecommunication traffic flow in subscriber transmission line"

By analysis, each of the specific component ideas in the subject and the broader or superordinate ideas or upper links implied by each of the component ideas are determined. The apparatus words are removed and the kernal terms are written as given below:

> Telecommunication, Traffic characteristic, Traffic Flow, Transmission, Transmission line, Subscriber line.

The kernal terms obtained are arranged in a helpful sequence using the principles for facet sequence of the General Theory of Library Classification [10, 11]. The string thus arrived at is given below:

> Telecommunication, Transmission, Transmission Line, Subscriber Line, Traffic Characteristics, Traffic Flow.

The relationship between pairs of terms NT or RT and the typology of the relationship are determined and the appropriate codes (described in Sec 22 below) are used to augment the subject string for computer manipulation. The coded string is given below:

> $0 TELECOMMUNICATION $ (05-)
> TRANSMISSION $0 $5 $6 (13-)
> TRANSMISSION LINE $3 $6 $7 (07-)
> TRAFFIC FLOW

22 Codes used in String Manipulation

The codes used in the subject strings for generating entries for a thesaurus are of two types, namely,

1. those that indicate which terms are to be related and whether the relationships is NT or RT; and
2. those that indicate the typology of relationship.

221 Codes for Relating Terms

1) The codes for generating RT relationships are:

> $1 - Generate an RT relationship with the immediately succeeding term using the typology of relation code, and generate a reverse

RT relationship changing the position of '-' in the typology of relation code.

$0	Generate an RT relationship with
$5	the immediately preceding term
$6	with the same 'code' taking the
$7	typology of relationship code of
$8	the term being manipulated and generate a reverse RT relationship changing the '-' in the typology of relation code.

2 The codes for generating NT relationships are:

$2 - Generate an NT relationship with the immediately succeeding term using the typology of relation code of the term being manipulated and generate a reverse BT relationship changing the '-' in the typology of relation code.

$3 - Generate an NT relationship with the immediately succeeding term, and generate a reverse BT relationship. (No typology of relation codes are used).

A special code $4 is also provided to stop generating any relationship between the terms between which it is coded.

3 The code for generating synonymous relationship is:

/ - Generate synonymous relationship with the immediately succeeding term, and generate a reverse USE relationship.

222 Code for Typology of Relationship

When subject representations are facet analysed and structured into subject strings it could be found that the relationships between ideas forming component of subjects are of the NHR associative type. It has been found helpful to identify patterns of association of ideas, forming components of subjects and categorise the relationships into a few types using suitable phrases and also to group the concepts on the basis of these types of relationships.

Several studies have been made in this area [9]. Almost 39 types of relationships have been identified so far [8]. In general, genus-species and part-whole relationships are taken as hierarchical relationships and represented as NT and BT in a thesaurus. Representation of HR relationships could also be categorised to achieve better display format and for generating RTs out of NTs to a particular concept/term (See Sec 33).

The following is an extract from a table of typology of relationships we have used:

00	Coordinate ideas	10	Action on
01	Source for	13	Device used for
04	Environment of	16	Types of
07	Property of	17	Parts of
08	Process used	23	Uses of

Note: It would be helpful to develop the typology of relationship codes for the subject of a thesaurus based on the characteristics used for deriving the different isolates forming components of the subject and those used to group the different speciators for the different isolates concerned.

3 PROCEDURE

31 Translation of Input Coded Strings

The subject strings augmented with codes for relating terms and codes for typology of relation are punched on to cards starting from column 1 to 74. Any number of cards could be given for a subject string, the end of the string being denoted by a "#" mark. The card image of input subject string is given in Appendix-B This is read by a program 'CODER' and a table is built in the memory. Each of the unique terms in the string is entered in the table with sequential serial numbering simultaneously replacing the respective terms in the strings by their respective serial numbers. As and when a term is encountered in a string it is matched with existing term in the table and its serial number is picked if the term is available, if not the term is entered as the last entry in the table, with appropriate serial number and that serial number is replaced in the string. The built term-table is copied as a term code dictionary — as a sequential file on a magnetic tape and the coded strings are passed on as a data set for further manipulation. An example of coded subject strings is given below:

$0 0001 $0(05-) 0002 $1$5$6(13-)
0003 $3$7 0004 $0$5(07-)
0005 $3$6$7 0006 #

0001 Telecommunication
0002 Transmission
0003 Transmission lines
0004 Subscriber line
0005 Traffic characteristics
0006 Traffic flow

(05-) Function of -
(13-) Device used for-
(07-) Property of -

32 Manipulation of coded subject string

The general rule for manipulating the string is that each of the terms form a term-pair with the

succeeding term in the string following the 'two types of codes' in between them. The manipulation starts from left and carried on to the right of the string. Once an entry is prepared the reverse entry is generated automatically by changing the position of the 'context term' and the 'lead term' In HR entries, the relationship NT is changed to BT in reversing the entry. In RT entry the relationship does not change. In the case of entries having the typology of relation codes, the position of '-' (in the typology of relation codes) is changed from prefix to suffix and vice versa as appropriate In the case of synonym indicated by '/' appropriate to the string, a SYN and USE reverse entries are generated.

When the codes for relating terms are not ₹1 or ₹2 but fall into the group ₹0, ₹3, ₹5, ₹6, ₹7, ₹8, etc. separate pointer-tables of these terms are built along with the typology of the relation code of the succeeding term having the same code, and term-pairs are generated using these pointer-tables. A detailed description of the logic of the program and algorithm are given elsewhere [1].

Every time a term-pair or thesaurus entry is generated it is written as a directly organised (random) data set on disk if it is not already written in it. Unique coded thesaurus entries thus generated are sequentially written as a data set for further manipulation. This process is done by a program 'GENERATE'.

For example from the coded subject string:

₹0 0001 ₹0(05-) 0002 ₹2₹5₹6(13-)
0003 ₹7 0004 ₹0₹5(07-)
0005 ₹3₹6₹7 0006≠≠

The following coded thesaurus entires are generated:

0001 RT (05-)	0002	0002 RT (-05)	0001
0002 RT (13-)	0003	0003 RT (-13)	0002
0003 NT	0004	0004 BT	0003
0004 RT (07-)	0005	0005 RT (-07)	0004
0003 RT (07-)	0005	0005 RT (-07)	0003
0005 NT	0006	0006 BT	0005
0006 RT (07-)	0002	0002 RT (-07)	0006
0006 RT (07-)	0004	0004 RT (-07)	0006

Note: The coded subject representation does not contain '₹' codes. In fact they are translated to certain alphabetic codes for internal manipulation by the machine.

33 Generation of Coordinate Entries

The generated unique thesaurus entries (term-pairs) are sorted in ascending sequence so that serial numbers for the same 'lead term' are brought together. The serial numbers of the "context terms' that are NTs to a particular 'lead term' (having the same serial numbers in the lead term position, having the same typology code if any) are formed as a separate table and coordinate RT term-pairs are generated and written along with the term-pairs earlier generated onto a tape for further processing.

34 Retranslation of Coded Thesaurus Entries

The term-code dictionary created as a sequential file on tape is read into the memory as a table. The typology of relation code and their corresponding descriptive phrases are read from cards and built in the memory as another table. The generated coded thesaurus entries on tape is read. The serial number of both the lead term and the context term are translated into the natural language, using the term table built from the term code dictionary. The typology of relation code is also translated into the corresponding descriptive phrase using the corresponding table built in the memory. The translated entries are written onto a tape for further processing. The process is done by the program 'TRANSLAT'.

35 Sorting and Printing Thesaurus

The file of thesaurus entries in natural language, output of the program TRANSLAT is sorted alphabetically using the sort program available in the computer system. It is then printed out in double column format with proper indention for Lead term, relationship, typology of relation and context term. The format of the display of terms in the thesaurus is given in Appendix - C.

4. FURTHER IMPROVEMENTS

It is possible to include at the retranslation stage the number of times a particular term has occurred in the bibliographic entries in a data base. This will be useful in finding out in how many different documents a particular term has occurred especially in on-line retrieval systems. This has to be done using a word frequency count list as reference.

It is possible to keep the generated thesaurus as a machine readable indexed file on disk for use in on-line information retrieval systems. Further work in this direction is being carried out in DRTC.

5. PROGRAMS DEVELOPED

The programs developed for generation of thesaurus as outlined in this paper are written in ASSEMBLY and COBOL languages for IBM system/360-370 series computers. The programs have been tested at the Computer Centre, IIT, Madras. A sample test data of about 60 subject strings on

telephony was run to generate a thesaurus. The number of unique terms were 145, the total number of term-pairs were 640. The programs were kept in a load module library and executed one after the other. The total CPU execution time(excluding compilation time) for 60 strings was 60 seconds. These programs will be used to generate a few thesauri and will be modified in the light of the experience gained to develop a software package for thesaurus generation.

6. ACKNOWLEDGEMENT

We are grateful to Prof. A. Neelameghan for his encouragement and guidance in the development to these programs. We are thankful to Prof. G. Bhattacharyya for permitting us to publish this paper.

7. BIBLIOGRAPHIC REFERENCE

[1] BALASUBRAMANIAN (V): Computer-based thesaurus generation from modulated subject structures. Guide F. J. Devadason. Guided project 1. DRTC 1978.

[2] BHATTACHARYYA (G): POPSI. Fundamentals and procedure, Seminar on subject indexing, DRTC, Bangalore 1975.

[3] FID News Bulletin. 20, 1970; No. 106.

[4] MAITRA (R): Semi-automatic method of generating micro-thesaurus (Project Report, DRTC 1977).

[5] MAITRA (R): Semi-automatic method of generating micro-thesaurus - A case study in the field of social sciences (DRTC Annual Seminar (15) (1977) Paper A5).

[6] -- -- -- --, Paper A5 Sec. 71.

[7] NEELAMEGHAN (A), GOPINATH (M A): Postulate Based Permuted subject indexing (POPSI). Lib. Sci. 12, 1975, Sep. p.80-81.

[8] NEELAMEGHAN (A), MAITRA (R): Non-hierarchical associative relationships in social science: Identification and typology (DRTC Annual Seminar (15) (1977) Paper A.).

[9] NEELAMEGHAN (A), RAVICHANDRA RAO: Non-hierarchical associative relationships: Their types and computer generation of RT links. Lib. Sc. 13, 1976 Paper C.

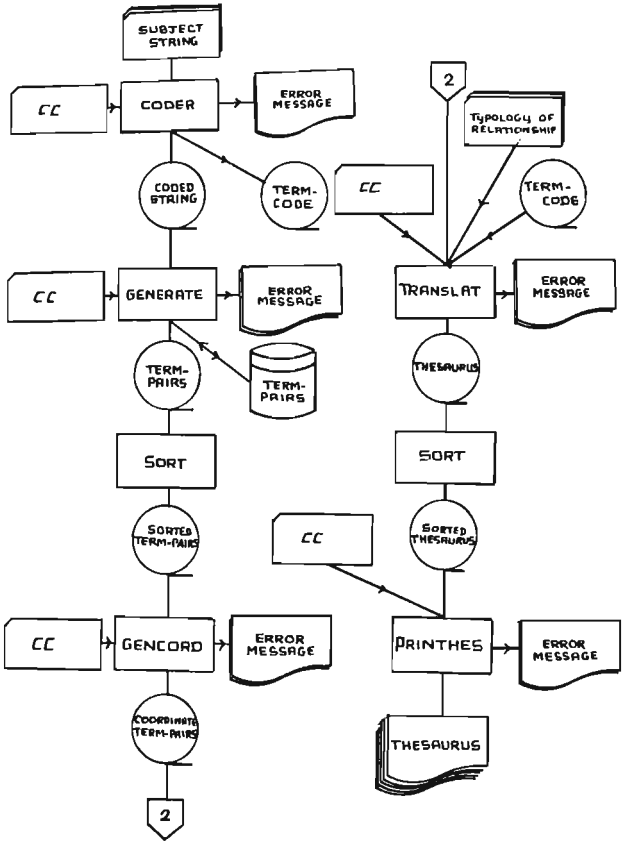
[10] RANGANATHAN (S R): Hidden roots of classification (Information Str. Retr. 3; 1976, 399-410).

[11] RANGANATHAN (S R): Prolegomena to library classification Ed 3. Assisted by M. A. Gopinath, Asia Pub. House. 1967.

[12] -- -- -- -- Sec SB2.

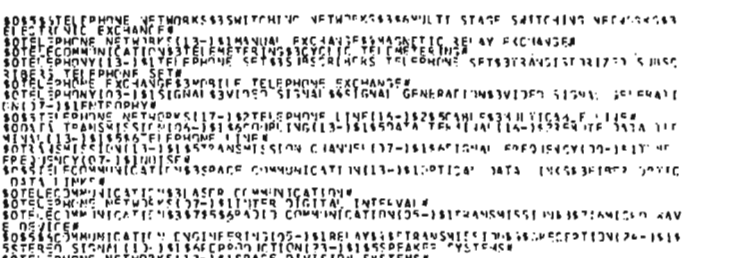
[13] RAVICHANDRA RAO (I K): Semi automatic construction of thesaurus (In seminar on thesaurus in information system. (1975) Bangalore, Paper BB).

[14] SHEPHARD (M), WATTERS (C): Computer generation of Thesaurus. Lib. Sc. 12, 1975: Paper E.



SYSTEM FLOWCHART

APPENDIX B



APPENDIX C