Lib. Sc. 24; 1987; Paper S.

SYMBIOSIS BETWEEN CLASSIFICATION AND THESAURUS

M A GOPINATH, *Professor, Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile, Mysore Road, Bangalore-560 059*

Varieties of vocabulary control devices have been emerging to improving the efficiency and efficiency of information retrieval. One form or other of the classificatory and thesaural approaches have been adopted in all these devices. The relational structure and control these devices are all complimentary. They all appear to be a kind of extension of the same basic format. The symbiotic relation in the procedure for construction of a classification scheme and a thesaurus are presented in this paper.

1 INTRODUCTION

Classification schemes and thesauri are two kinds of vocabulary control devices. A classification scheme brings into juxtaposition, ideas which are like. It is an arrangement of concepts and their combination in a systematic affiliation. Such a presentation arranges a human mind for browsing. Browsing leads to reach the specified item of ideas a person is seeking to know. It is fashionable to call such a presentation as menu selection. The systematic presentation is a hierarchical presentation, from broader perspective to narrower perspective. A modular approach to the delineation towards focal point. Ranganathan called such a formation a "Chain". In this approach we see a sequential browsing wherein the seeker of information starts from first concept and then passes on until he identifies his required information. It is telescopic pin-pointing of knowledge. In this hierarchical browsing the user starts at the root of the hierarchy or tree and is guided to areas of likely interest. Each node should contain not only pointers to its dependent nodes but also some indication of the information that is stored in the

subtrees. The user can then make a choice as to which branch of the tree to follow. This limits the amount of text the user has to browse through. If the user decides that the text he is looking at is not really relevant and he wants to look at another area, then he must either back track up the tree or else abandon his current search and start again.

While such a linear structure is helpful, there are cases where one would go to a kind of lateral thinking, or tangential thinking. In order to meet such eventualities clustered browsing is resorted to. The clusters file attempts to provide the best of both these methods—the system gathers together all the documents it thinks are related and the user is then free to browse round each cluster sequentially or jump to another cluster. Then, here we reach into Thesaurus - more specifically in our context information retrieval thesaurus.

A thesaurus may be defined either in terms of its function or its structure (UNESCO).

(1) In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained "system language" (documentation language, information language).

(2) In terms of structure, a thesaurus is a controlled vocabulary of semantically and generically related terms which covers a specific domain of knowledge.

2 CONSTRUCTION OF CLASSIFICATION SCHEMES

Classification schemes have capabilities to display the following structures:

1. Hierarchical
2. Associative
3. Sequential
4. Attributive

Hierarchic display is taxonomic. Genus-species structure or type structures. Associative and Attributive are faceted structures. Sequential is ordinal structure indicating the ranking. Any faceted classification, in particular Ranganathan's Colon Classification, exhibits all these four varieties of classification.

These structures can help generation of a thesaurus in an effective way. For a thesaurus entry exhibits the following kinds or relations :

1. The Equivalence relation
2. The Hierarchic or generic-specific relation
3. Associative relation

We shall follow a symbiotic approach to the design of classification and thesaurus.

3 PRACTICAL STEPS IN CLASSIFICATORY APPROACH FOR CONSTRUCTION OF THESAURUS

Sources for terms in a thesaurus can be any one of the following :

1. A number of text sources
2. A group of users
3. An index to a book or a periodical publication or a secondary publication.

But for demonstration purposes, I shall choose the expressive titles of technical papers. For this purpose, I choose the topic known as "Artificial Intelligence".

Let us look at the following titles and their semantic factoring (Facet analytic approach) for thesaural purposes.

Ser. No.	Title	Facet Analysis
1.	Concept typology (Kybernetics. 10; 1981; 253-9)	Conception, Simulation
2.	Binary answers to imprecise questions (Kybernetics. 11; 1982; 255-95)	Intelligence, Algorithm
3.	A neuronal model with synaptic long term memory [Engineering (Japanese). E64, 11; 1981; 751]	Intelligence, Memory > Synaptic Long Term Memory, Model, Network, Neuron
4.	Logic for default reasoning (Artificial Intelligence, 13; 1980; 81-132)	Logic Based Reasoning, Default
5.	Problem solving in game playing : Computer chess. Cybernetics and systems. 13; 1982; 31-49)	Game playing, Chess, by Computer

Ser. No.	Title	Facet Analysis
6.	Writer recognition by spectral analysis. (Intl. Conf. on Pattern Recognition (3) (1980) Proceedings. 1980. p. 1-3)	Pattern Recognition, Spectral analysis
7.	Rank filters in digital image processing. (Computer Graphics and Image Processing. 19; 1982; 148-64)	Pattern Recognition, Picture Processing using Rank Filters
8.	Two hierarchical linear feature representations : Edge pyramids and Edge quad trees	Pattern Recognition, Picture Processing, Halftone Picture, Moire Patterns, Analysis, Fourier-transform
9.	Linguistic approach to object recognition by grasping [World Congress of Intl. Federation for Automatic Control (Kyoto) (1981) Proceedings. 1982. 1915-20]	Pattern Recognition, Object recognition, Grasping using Syntactic Method
10.	Learning of recognition with imperfect teacher (Archieves Antenna & Telecommunication Mechanisms. 28; 1980; 419-30)	Pattern Recognition, Object recognition, Imperfect teacher, Learning.
11.	Rete : A fast algorithm for many patterns many object pattern match problems. (Artificial Intelligence. 19; 1982; 17-37)	Pattern Recognition, Object Recognition, Algorithm, Matching, Rete
12.	MITES : A model-driven iterative texture segmentation algorithm (Computer Graphics and Image Processing. 19; 1982; 195-110)	Pattern Recognition Texture Recognition, Texture segmentation, Algorithm, Model driven iterative
13.	Syntactic pattern recognition and applications. (Englewood Cliffs. 1982. V 3. 569 pages)	Pattern Recognition, Shape Recognition, Syntactic Method
14.	Sequential tracking extraction of shape features and its constructive description (Computer Graphics & Image Processing. 19; 1982; 349-66)	Pattern Recognition, Shape Recognition, Sequential Tracking, Extracting
15.	Improved parameter set for adaptive symbol recognition (IBM Technical Disclosure Bulletin. 24; 1981; 769-71)	Pattern Recognition, Adaptive Symbol Recognition, Parameter Set for Elastic matching.

Ser. No.	Title	Facet Analysis
16.	Multifont Alphanumeric Character recognition method using multiple decision levels. (Annals of Telecommunication. 36; 1981; 401-11)	Pattern Recognition, Character Recognition, Non-Numeric Character, Multifont
17.	Evaluation of LPC-spectral-matching measures for spoken word recognition. (Transactions of Institute of Electronic and Communication Eng. (Japan) (Section 8). E65, 5; 1982; 298)	Pattern Recognition, Speech Recognition, LPC-Spectral Matching Measures for Spoken Word Recognition
18.	New technique for spoken Japanese syllable recognition and its application to large vocabulary word recognition. (Transactions of Institute of Electronics and Communication Eng. E65, 2; 1982; 30)	Pattern Recognition, Speech Recognition, Japanese, LPC-CEPSTRUM COEFFICIENTS
19.	Validity of articulatory parameters in continuous speech recognition for unspecified speakers - vowel descrimination test. (Transactions of Institute of Electronics and Communication Eng. Japanese Sec E. E65, 7; 1982; 15)	Pattern Recognition, Speech Recognition, Articulatory parameters for unspecified speaker validity, By applying vowel descrimination test
20.	ISPAHN: An interactive system for pattern analysis: structure and capabilities (Pattern Recognition in practice: Proceedings of the Intl. workshop (Netherlands). 1980; 481-91)	Pattern Recognition, Interactive System, ISPAHN

4 TERM PROFILE

Each term semantically segmented is separated and its contextual environment, definition of the term, the specific relational role of the term, the broader term(s), narrower term(s), equivalent terms, preferred terms, and other terms occuring in the context are recorded. The worksheet for terms collection, on the next page is such an example.

5 COMPILATION OF THESAURUS

The terms are grouped alphabetically. The preferred terms are used. These consolidated alphabetical term records act as a base for the thesaurus.

Alphabetisation of term records is extremely helpful. It enables checking for duplicates, homonyms, and synonyms. Word by word alphabetisation is helpful. Each complete word is considered in turn and arranged in letter by letter sequence. All terms beginning with a given complete word precede any terms beginning with the same sequence of letters as part of the word. Non-alphanumeric characters (punctuation marks and special characters) are treated as spaces.

WORKSHEET FOR TERM COLLECTION

Thesaurus Project on
Artificial Intelligence

Term : SPEECH RECOGNITION

Context :	New techniques for spoken Japanese syllable recognition and its application to large vocabulary word recognition.
Definition :	Recognition by a machine or a computer all oral sounds of a human speech.
Role :	Focal concept (Personality in SRR's terms)
BT :	Pattern Recognition
NT :	Japanese Speech Recognition, Sequential track speech recognition.
UF :	Voice recognition.

Thereafter, we have to merge term records for a single concept and to finalise the forms of the terms, the preferred descriptors, relationships, cross references etc. The first step in merging would be to identify all identical or nearly identical terms and clip them to together in groups.

With manual procedures there are two possible places to put the record that results from merging :

1. Put it on a fresh worksheet;

2. Select a worksheet already in the package and transfer only the information from other cards, thus saving work;

The following criteria may be used in selecting a card (listed in decreasing priority),

1. Select the card that contains the largest amount of text. This will minimise the work needed for the transfer of information from other cards.

2. Select the card that has been made up from a preferred source like a thesaurus,

3. Select the card that is most legible.

If there are duplicate worksheets, they should be eliminated. All the information is to be put into one worksheet and all others discarded.

If there are homonyms, a defining or distinguishing word should be included in brackets after the term in each such record.

If there are synonyms — true or quasi, they should be cross-referenced. A single-one of them is to be chosen as the preferred term and entered in the USE section of each record. All other terms are to be entered in the USED FOR column of each record.

For common action terms like "Utilisation", "Production" etc., no BTs, NTs, or RTs are to be collected. This is because they should appear as preferred terms with no listing of terms under them. If they are listed as the other terms are, then in the Thesaurus, intelligence operations, all strategies, all devices for performing these operations will be generated under each of the terms. Besides which nobody is likely to enter the system with the word "utilisation" or "storage" etc.

These consolidated term records would serve as the foundation for the classification schemes and the thesaurus and as a permanent record of all relationships recognised and decisions made.

6 THESAURUS DISPLAY

1. The display of terms in a thesaurus is to present in a comprehensible manner the structural relationships between terms,

including hierarchical and non-hierarchical as well as equivalence and associative relationships.

2. The main part of the theasurus should comprise complete information each descriptor. The main part can be arranged systematically and/or alphabetically or a combination of both can be used. When non-preferred terms, and permutations of compound descriptors and non-preferred terms are included in an alphabetically arranged main part, a separate alphabetical index is not required and nor is any other auxiliary part.

The information to be included in the main part may comprise :

1. Concept representations
   - Descriptors
   - Non-preferred terms
   - USED FOR reference (equivalence relations)
2. Additional informations
   - Definitions
   - Scope notes
   - Short source information
3. Conceptual relationships
   - Superordinated concepts (hierarchical relationships - BT)
     - Broader concepts (generic relation)
     - Entity concepts (part-whole relation)
   - Subordinated concepts (hierarchical relationships - NT)
     - Narrower concepts (specific relation)
     - Part concepts (part-whole relation)
   - Associated relationships
     - Related concepts (other specified relations — RT)
   - Equivalence concepts
     - Use reference (to preferred term)
     - Used for reference (for non-preferred term).

Structure of a thesaurus permits polyhierarchical and/or facet structure so that there may be several broader or narrower concepts for a given term. Since only one hierarchical chain can be presented coherently in a systematic display, the relationship

with other hierarchical chains may be indicated by cross reference.

7 METHODS OF DISPLAY

Display methods are of diverse varieties. There is a wide choice of thesaural layouts from which to choose, ranging from puely alphabetical arrangements to varying combinations of alphabetical lists with systematic displays.

71 Alphabetical Thesaurus

The most common layout for a thesaurus is the alphabetical one, as is used in the TEST, which lists information in the following order.

Descriptor
- Scope notes
- Synonyms (U and UF)
- Broader Terms (BT)
- Narrower Terms (NT)
- Related Terms (RT)

This mode of display is chosen for any information retrieval thesaurus. It synoptically presents a descriptor with all its relationships in one place. It is easy to use for the indexer as well as for the user of the system. An example of this layout is:

MEMORY

UF	Internal Memory
BT	Intelligence
NT	Long term memory
	Short term memory
RT	Creativity
	Reasoning
	Search Model

The principal characteristics of this typical thesaurus display are: 1. The broader, narrower and related terms are listed under each of the descriptors which are arranged in alphabetical order; 2. Within each of these three groups of relations the descriptors are arranged alphabetically and so also for the synonymous terms; 3. Nonpreferred terms appear in the alphabetical

sequence and are referred to the appropriate descriptors; 4. The reciprocal (UF) of a USE reference appear under the descriptor referred to; and 5. Scope notes are used in parenthesis to define the scope of a descriptor.

The advantage of this method is the achievement of clear, systematic display of the hierarchy. The hierarchy is in either direction—either above or below, can be followed easily and with no ambiguity by referring to the main entry of a term listed as either a BT or NT.

As a variation of this method same thesaurus display relationships can be presented in a single sequence, indicating the relationships in brackets. For example:

Term: PATTERN RECOGNITION
Artificial intelligence (BT)
Character recognition (NT)
Designing (RT)
Edge recognition (NT)
Feature recognition (NT)
Game playing (RT)
Object recognition (NT)
Picture processing (NT)
Planning (RT)
Problem solving (RT)
Shape recognition (NT)
Speech recognition (NT)
Symbol recognition (NT)
Systems of pattern recognition (RT)
Texture recognition (NT)
Theorem recognition (RT)

The lead-in vocabulary leads users from the specific term to the descriptor or descriptors used in the system for that concept. The lead-in terms are either entry terms pointing from the specific term to the broader terms, or specifiers directing the user to the combination of terms selected to represent a specific, synthesized concept.

The lead-in vocabulary size will vary with the level of specificity of descriptors. Where the descriptors are broad, the lead-in

vocabulary must be large to compensate for this by providing specifiers and entry terms. With an adequate lead-in vocabulary, recall may be as high with a reduced vocabulary as with a detailed vocabulary. This is because the searcher is led at once to the specific concept, whether it is subsumed under a broad or nearly synonymous heading or represented by a combination of two or more descriptors.

72 Classified Thesaurus

721 Alphabetical Thesaurus with broad Subject Classifiction

In this type of system the descriptors in the classified thesaurus are arranged purely alphabetically under broad subject groups. The code number for the subject group is shown against the index term in the alphabetical thesaurus. This method is used in NASA thesaurus and also in MeSH, PREDICAST thesaurus etc.

722 Alphabetical Thesaurus with Clusters and Arrowgraphs

In this arrangement, the systematic section consists of "association maps", "arrow graphs", terminological charts, showing diagrammatically the relationships between terms. Arrowgraphs may be built up by the compiler by writing the index terms belonging to each major subdivision on a large sheet of paper, in such a way that the related terms appear in the same neighbourhood. There are arrows from a broader term (BT) to its narrower term (NT) and usually (as shown in the earlier example) the BTs are placed above the terms and connected with upward arrows. Narrower terms are below and linked with downward arrows. Related terms (RT) are connected with two-way arrows. From the terms in the first round, one proceeds in the same way. There are numerous relationships between terms belonging to different rounds, resulting in a network of inter-relationships.

Network structures allow for the display of more complex inter-relationships among descriptors than the circular type of display. This may be done at the cost of clarity and intelligibility. It is especially important in this method that relationships most useful for indexing and searching and that the network be designed in such a way as to display these inter-relationships as

adequately as possible. This method of graphical display is used in IRRD thesaurus.

723 Alphabetical Thesaurus with Hierarchical Displays

It is common practice where the thesaurus is in computer readable form to produce hierarchical displays from the BT or NT information in the alphabetical thesaurus. In the alphabetical arrangement all levels of the hierarchy cannot be displayed at once, and when they are, it is not possible to differentiate them. The hierarchical display overcomes this difficulty, producing hierarchical trees arranged with the broadest concepts at the head of the tree. This is usually done only for terms having associated with them at least two generic levels of term. This method is made use of in the thesaurus of Pulp and Paper terms. The displays are really alphabetical classed indexes. They are usually listed separately and are arranged in alphabetical order according to the broadest concept in each array. Hierarchical displays do not add new information to the thesaurus, but make subject scanning easier.

724 Alphabetical Thesaurus with Hierarchical Classification

Hierarchical classification in thesauri may occur in the form of a limited number of auxiliary generic trees or wholly integrated. The link between the alphabetical thesaurus and the location of the term in the classification is provided by the detailed notation. When a term occurs in more than one hierarchy, it is listed in all appropriate places in the tree structures, and all the class members are indicated against the term in the alphabetical thesaurus. This pattern is followed in the MeSH.

725 Alphabetical Thesaurus with Faceted Classification

This arrangement requires the integration of a detailed faceted classification system with a thesaurus. It is the system followed by the Thesaurofacet. The faceted classification and the alphabetical thesaurus were developed simultaneously. The role of the thesaurus and classification schemes are complimentary. The thesaurus as well as acting as an index of the classification, controls word forms and synonyms and shows relationships which cannot be easily diplayed in the schedules. The classifica-

tion system gives an overall view of the structure of the subject fields and displays hierarchical relationships.

726 Systematic Thesaurus with Alphabetical Index

The systematic thesaurus consists of descriptors, with synonyms and other related terms are arranged in a classified order, as followed in the Environmental Studies thesaurus. The alphabetical index leads in to the terms in the systematic thesaurus.

727 Alphabetical Thesaurus with a Circulation display

In this method, the different hierarchical levels are arranged in concentric circles. The broadest term in the group is in the centre and the narrowest is on the innermost circle. Links are shown by arrows. The disadvantages of this method are : 1. It is difficult to divide one big circular scheme into several small circular schemes. Therefore, this method is suited only for relatively small and closed subfields; 2. The number of hierarchical levels allowed for on a sheet of paper is limited by its size ; 3. Polyhierarchical relationships may be displayed within one circular scheme only and are limited also; 4. Associative relationships cannot be displayed easily.

The alphabetical part consists of the descriptors arranged alphabetically besides the graphical display.

8 VERIFICATION

When all relationships are displayed in the method decided upon, it is in a form which is to look through. Before finalising the thesaurus, it should be checked in consultation with the subject specialists. It is preferable to discuss the remifications of a thesaurus with a team of subject specialists. Points to be discussed are :

1. Adequate representation of the concept by the preferred term;

2. Overall structure of the hierarchy, selection and derivation of the subject fields and subfields, sorting out of the concepts in the subfields, helpful order in the arrangement of concepts.

3. Individual hierarchical relationships – the relationship between sets of two terms each.

4. Selection of the descriptors – whether they are necessary, relevant, useful, for either indexing or searching.

5. Filling in the gaps in the thesaurus by introduction of new concepts – new broader concepts are introduced or entirely new concepts are added that upto this stage have been overlooked.

9 CONCLUSION

Classification scheme and Thesaurus emerge from common source for common purpose, that is to impart efficiency in information retrieval. They combine within themselves all possible angle of relations an information may be sought and presented. We have discussed all major features of classification system and thesaurus and their relative roles, in particular their capabilities in the analysis and storage of information, the means of extracting information from the data base, the user's interface with the data base and the overall management of the system. It must be emphasised here that the thesaurus has role to play in indexing as well as searching. But in majority of cases, thesaurus is used as a search aid rather than as an indexing tool in the Online information search system. Each term in a synonym ring is equally likely to occur in any relevant article and so it is necessary to search for them all rather than one preferred term. But a conceptual tree formed out of classification semantics, if built into such an online search system, it would act as an intelligent device in bringing together like ideas, "exhaustively, pin-pointedly and expeditiously, inspite of the continuous downpour of cascades of nascent thoughts", as Ranganathan put it.

91 Bibliographical References

1. AITCHISON (K.) and GILCHRIST (A.). Thesaurus construction : a practical manual. 1972.
2. BLANDEN (J. F.). Thesaurus compilation methods : A literature review. (ASLIB Proc. 20, 8; 1968; 345-79).
3. CHANDRAN (D.). Candidate terms for a thesaurus : A case study of sources of terms in the field of library and information science. (IN : Seminar on Thesaurus in Information Systems (DRTC and INSDOC), 1975, Bangalore. Paper AG).
4. DAVIS (C. H.). Integrating vocabularies with a classification scheme. (American Documentation. 19; 1968; 101).

5. ELLER (J. L.) and PANEK (R. L.). Thesaurus development for a decentralised information network. (American Documentation. 19; 1968; 213-20).
6. ENGINEERING JOINT COUNCIL. Thesaurus of engineering and scientific terms. 1967.
7. ENGLISH ELECTRIC COMPANY. Thesaurofacet: A thesaurus and faceted classification for engineering and related subjects. 1969.
8. FOSKETT (D. J.). Thesaurus. (IN: Encyclopaedia of library and information science. Edited by Allen Kent, Harold Lancour, and Jay E Daily, Asst. Editor William Z Nasri. Vol 30; 1980; p. 416-462. New York, Marcel Dekkar.
9. FREEDMAN (B.). Thesauari for vocabulary control. (Drexel Library Quarterly. 8; 1972; 125-8).
10. GILCHRIST (A.). Thesaurus in information retrieval. 1971.
11. GILLIUM (T. L.). Compiling a technical thesaurus. (Journal of Chemical Documentation. 4; 1964; 29-32).
12. GOPINATH (M. A.) and PRASAD (K. N.). Thesaurus and classification scheme : A study of the compatibility of the principles for construction of thesaurus and classification scheme. [IN : Seminar on Thesaurus in Information Systems. (DRTC and INSDOC) (1975); Bangalore. Paper AF7].
13. ISO : 2788. Guidelines for the establishment and development of monolingual thesauri. 1974.
14. JONES (K. P.). Basic structures for thesaural systems. (ASLIB Proceedings. 23; 1971; 577-90).
15. JOYCE (T.) and NEEDHAM (R. M.). The thesaurus approach to information retrieval. (American Documentation. 9; 1958; 192-197).
16. LANCASTER (F. W.). Vocabulary control for information retrieval. 1972.
17. NEELAMEGHAN (A.). Non-hierarchical associative relationships. Their types and computer generation of RT links. (Seminar on Thesaurus in Information Systems. (DRTC & INSDOC) (1075). Bangalore. Paper AA).
18. NEVILLE (H. H.). Feasibility study of a scheme for reconciling thesauri covering a common subject. (Journal of Documentation. 25; 1970; 313-36)
19. RAVICHANDRA RAO (I. K.). Semi-automatic construction of thesaurus. (Seminar on Thesaurus in Information Systems. (DRTC and INSDOC) (1975). Bangalore. Paper BB).
20. ROSTON (R. M.). Construction of thesaurus. (ASLIB Proceedings. 20; 1968; 181-7).
21. SHEPHERD (Michael) and WATERS (C.). Computer generation of thesaurus. (Library Trends. 12; 1975; Paper E).
22. SOERGEL (Dagobert). Indexing languages and thesauri.
23. UNESCO (Paris). Guidelines for the establishment and development of monolingual thesauri for information retrieval. 1971.
24. VICKERY (B. C.). Thesaurus - A new word in documentation. (Journal of Documentation. 16; 1960; 181-9).