THEORETICAL FOUNDATIONS OF THESAURUS CONSTRUCTION

A CHOSE and A S DHAWLE, Vikram Sarabhai Library,
Indian Institute of Management, Vastrapur, Ahmedabad 380 015

The theoretical foundations of recent work on automatic and semi-automatic thesaurus construction are briefly but critically reviewed, and limitations of current methods of automatic construction of thesaurus are pointed out. Need for a deeper study of the theoretical foundations of thesaurus construction is emphasized and a line of approach to it is suggested.

1 INTRODUCTION

There are now available thesauri for a variety of subject-fields ranging from physics, chemistry and engineering in the physical sciences, through medicine in the biological sciences, to education, political science, and sociology in the social sciences. It is necessary to take a close look at the theoretical foundations of thesauri construction in order to evaluate the usefulness of thesauri built on that basis in information retrieval. Thesaurus construction has basically two aspects : (1) Selection of key words/terms/phrases of the particular subject for which the thesaurus is constructed; and (2) establishing inter-relationship among those terms so as to deal with the problem of synonimity and display hierarchical relationship among terms. The importance of both these aspects for information retrieval purposes is evident. In this paper, we show that whereas the problem of selection of key-words can be tackled without great difficulty, and even a fully automatic solution of the problem is not impracticable, the problem of inter-relationship of terms presents serious difficulties, particularly for the construction of thesauri of social sciences. We shall discuss these difficulties fully, and suggest ways to overcome them.

2 CURRENT METHODOLOGY

The selection of key words or phrases for the subject is done by scanning through documents in a collection, and making one list of words/phrases significant for information retrieval purposes and another of common words (negative dictionary) which are common to most subjects and hence have no significance for information retrieval purposes of the subject under consideration. In preparing the list of significant words, it is sufficient to consider only the stems of words so that any other word derived by the addition of a suffix to the stem word can be identified with its stem word. This process would significantly reduce the length of the list of words.

Fully automatic methods have been successfully used for such a selection procedure (7, 11, 16). This is based an making the computer find out the frequency of terms found in a given collection of documents, and then choosing those terms as significant, the frequencies of which are neither too high nor too low. The terms of very high frequency are considered as common words and those of very low frequency being considered as having little relevance to the subject. Such terms are considered to be insignificant from the point of view of information retrieval.

Whereas the fully automatic methods based on frequency of terms seem to be quite adequate for selection of key words/phrases, this approach does not appear to us to be a sound basis for the study of interrelationship of terms as has been done by some authors (2, 3, 10). Let us first consider how this interrelationship of terms is established by means of a term document matrix. The frequency of terms in each document belonging to a collection of document on the subject concerned is represented in the matrix form as follows

	t_i	t_2	t_n
d_1	f_{11}	f_{12}	f_{1n}
d_2			
d_m	f_{m1}	f_{m2}.	f_{mn}

fi_j shows the frequency of the term t_j in the document di (i = 1, 2... m ; j = 1, 2... n). Let us now consider a simple example

	t_1	t_2	t_3	t_4
d_1	4	1	0	5
d_2	3	0	1	4
d_3	0	4	6	1
d_4	1	3	4	0

In this example, we find that tl, t_4 are highly frequent in dl and d2, whereas in d3 and d4, the frequency of both these terms is low. Similarly for t_2, t_3 we find that they are highly frequent in d3, d4, and are rare in dl, d2. Hence, one concludes that tl is related to t4, and t2 is related to t3. In this way classes of interrelated terms are formed so that terms belonging to the same class are characterised by the fact that they are almost equally frequent in a set of documents and almost equally rare in another set of documents.

The frequency approach has also been used to set up hierarchical relationship among terms (1). For a class of interrelated terms, those terms which have substantially higher frequencies are considered as categories and those with low frequencies are considered as sub-categories. Hierarchical relationships based on frequency of terms might have been adequate for the information retrieval needs for a particular subject (1), but there does not seem to be enough grounds to make it a general principle for the thesaurus construction for all subjects. As we had pointed out earlier, setting up hierarchical relationships of terms is one of the most critical problems of library classifications (5). It cannot be solved simplistically by counting the frequency of terms. That a certain amount of human judgement is essential for setting up interrelationships and classification of terms is admitted also by Salton (13). The crucial problem is to determine how human judgement is to be used for establishing a classification as well as an interrelationship of terms.

Sparck-Jones (14) suggests a way of determining synonimity of terms by asking the decision maker whether a term can replace another term in a given context. If the synonimity of terms established in this way is to be of use for information retrieval, then the decision maker's decision should not be too subjective. Such an element of subjectivity (and consequently ambiguity for information retrieval purposes) will inevitably creep in with many of the terms of social sciences. Consider for instance, the sentence "A person's behaviour depends upon his education". If a decision maker would be asked whether in this sentence "behaviour" can be replaced by "attitude", his answer would depend upon the fact whether he subscribes to behaviouristic psychology or not. Such examples should caution us about the inherent problem of ambiguity of the terms of social sciences, and hence also about some of the fundamental difficulties of thesaurus construction.

Question answering system has also been used to apply human judgement to the problem of classification of terms and determination of categories and sub-categories. Actually the categories and the sub-categories are fixed first, such as abstract, concrete, etc., and then questions are asked whether the terms have the properties expressed by the categories. In this way one gets a term property matrix.

	pl	pZ	pm
t_i	e_{11}	e_{12}	e_{1m}
t_2			
t_o	e_{nl}	e_{n2}	e_{nm}

e_{ij} = I means that the property p. is applicable to the term ti (i = 1,...n; j = 1_.m). Similarly e_{ij} = 0 implies that pj is not applicable to ti.

Let us consider a term property matrix for computer science such as the one considered by Salton (12).

	Abstraction	Physical object	Hardware	Software
Computer				
System				
Program				
Machine				
Equation				
Logic				
Data				

The properties are chosen in such a way that they are apparently mutually exclusive. For the term 'computer' there is no ambiguity as to which of the properties would apply. One

would for instance get the corresponding row for the term 'computer' : 0 1 1 0, that is, computer is not an abstraction, is a physical object, is a hardware and not software. For the terms 'system' and 'program' it is not quite clear which of the properties apply. For instance 'program' can be considered as an abstraction, if it is thought of as a mathematical algorithm, and on the other hand as a physical object if considered as a 'written thing'. Similarly 'system' may be both hardware and software. Sometimes a term may be such that a Property is neither applicable nor nonapplicable to it. It may be quite meaningless to ask whether the term 'data' is software or hardware.

Of course if one would be able to choose such properties which are unambiguously mutually exclusive , then one would also be able to set up a hierarchical classification of terms from the term property matrix. For instance in the previous example one could first get two mutually exclusive classes of terms, terms referring to abstract objects, and terms referring to physical objects. Then within these classes one would get the mutually exclusive classes of terms referring to hardware and software. However, the assumption of existence of mutually exclusive classes could be valid in subjects such as mathematics and the natural sciences, where the terms have unambiguous meaning (4), it is quite untenable in interdisciplinary subjects like most of the modern subjects of social sciences (6). It is important to point out that the term property matrix does not by itself lead to hierarchical classification of terms, but a hierarchical classification is already assumed in the construction of the term property matrix.

3 NEED FOR A NEW APPROACH

Since a thesaurus is constructed for the information retrieval needs of a class of users, the interrelationship of terms established in the thesaurus should correspond to some "word map", which is common to the vast majority of this class of users. For subjects such as physics and engineering such a common "word map" can be built in consultation with a group of specialists in the field. If we assume that such a thesaurus will be used only by people having substantial knowledge about the subject, we can safely conclude that the thesaurus will effectively serve the information retrieval needs of this class of users. In fact a thesaurus constructed on the basis of interrelationship of terms as viewed by the specialists, provided that in general there is sufficient agreement among them, will have a much stronger foundation than the purely frequency based automatic classification which can always be falsified by an error of chance.

We would like to emphasise that a thesaurus should always be constructed having a class of users in mind. The construction of thesaurus of a subject, which will satisfy the information retrieval needs of all possible users, appears to us as almost impossible, since this assumes the existence of a common "word map" for all users. One tends to think that a thesaurus for sociology or political science should not only be useful for the specialists, but also for the common man. The dilemma of satisfying the incompatible needs of the specialists and the common man is faced by Viet (15). However, he seems to assume that terms have some inner logical relationship independent of the relationships as viewed by the specialists. We do not see any reasons for making this assumption, since this raises many philosophical questions, such as whether terms have meaning independent of the people who use these terms.

But in subjects such as political science and sociology it might even be difficult to find agreement among the specialists as to the meaning and interrelationship of terms. However to what extent there are variations in meaning can be explored by studying the association of terms used by a group of specialists. An analysis of the similarity and differences of such associations can thus be represented as graph theoretical models of different data structures. The focus of our problem is different from that of Quillian (8) but similar design principles can be employed to set up the "configuration of terms" for the meaning of words, where there is no intrinsic hierarchy among the terms, but the term whose meaning is sought becomes the primary focus, and other terms are seen only in relation to it. "Most importantly, in such a model of semantic memory there is no predetermined hierarchy of superclasses and subclasses ; every word is the patriarch of its own separate hierarchy when some search process starts with it. Similarly, every word lies at various places down, within the hierarchies of a great many other word concepts. Moreover, there are no word concepts as such that are 'primitive'. Everything is simply defined in terms of some ordered configuration of other things in the memory" (9). Quillian aims to simulate human memory by storing words in the memory of the computer with links corresponding to the association among the words as is given in a standard dictionary of the English

language. The association of words as given in a standard dictionary has been taken by him as a model. For our problem of resolving the ambiguity of meaning, the standard dictionary is of no use. For 'ambiguous words", each individual has his particular system of association, and since it is impossible to take into account all possible individual variations in shades of meaning, we suggest that researches should be carried out in order to find out these variations in as much as they exist among the specialists of a subject. In this way one world be able to take into account the different shades of meaning inherent in the different point of view of the users of the thesaurus.

4 CONCLUSION

Recent work on thesaurus construction shows that it is generally assumed that fur a given subject a hierarchical classification of terms exists. In the fully automatic approach one tries to establish such relationships purely on the basis of frequency of occurrence of terms in a given collection of documents. In the semiautomatic approach the experts' answers to a set of questions give the clues for the classification. However, in the very framing of the questions one assumes a priori some hierarchical classification. In our point of view a thesaurus should be constructed keeping a class of users in mind and the interrelationship of terms can only be established, after an analysis of their association of terms. This approach is recommended particularly for the thesaurus of social sciences where terms tend to be quite ambiguous.

5 BIBLIOGRAPHICAL REFERENCES

I Sec 2 DOYLE (L B). Expanding the editing function in language data processing. (Commrn ACM. 8 ; 1965;)

2 Sec 2 --. Is automatic classification a reasonable application of statistical analysis of text ? (J ACM. 12 ; 1965 ; 47A-89)

A Sec Z --. Semantic road maps fur literature searchers (J ACM 8 ; 1961 ; 551-78)

4 Sec 2 GHOSE (A). Logic, Matthew-tics, computer science. (Inter Logic Rev. 197A June; 48-55)

5 Sec 2 -- and DHAWLE (A S). Inter- and trans-disciplinary ordering systems fur universe of knowledge. (Paper presented at the Third International Study Conference on Classificatiun Research (Bombay) (6-11 Jan 1975)).

6 Sec 2

7 Sec 2 LESK (M E). Performance of automatic information systems. (Infor Stor Retr. 4; 1968; 201-18)

8 Sec A QUILLIAN (M R). Computers in behavioural science - word concepts : A theory and simulation of some basic semantic capabilities. (Behav Sci. 12; 1967 ; 410-A0)

9 Sec A (--. 415)

10 Sec 2 ROLLING (L). EURATOM thesaurus - Keywords used within EURATOM's Nuclear Energy Documentation Project, Erratum Center for Information and Documentation. Rep E U R. 500.e, 1964.

11 Sec 2 SALTON (G). Automatic information organization and retrieval. 1968. p 40-8

1 2 Sec 2 - --- p 51

1A Sec 2 - --. p 57

14 Sec 2 SPARCK-JONES (K). Experiments in semantic classification. (Mech Trans. 8; 1965; 97-11 Z)

15 Sec A VIET (J). Thesaurus for information processing in sociology. 1971. p 10

16 Sec 2 WALL (E). Vocabulary building and control techniques. (Amer Duc. 2o; 1969; 161-4).