SYNTHESIZED USER BASED TERMINOLOGY INDEX LANGUAGES (SUBTIL)
(LES LANGAGES D'INDEXATION UTILISANT UNE TERMINOLOGIE SYNTHETIQUE PROVENANT DES UTILISATEURS)

C.D. Batty
Graduate School of Library Science
McGill University
Montreal, Québec, H3A 1Y1

ABSTRACT

The increasingly common development of new inter- and multi-disciplinary areas of knowledge raises fundamental problems in the design of index languages to index and access hetereogeneous and scattered information resources. The determining characteristic of a new area is the need and orientation of the user. A method is described of developing index languages by systematic analysis of users' concept systems and terminology. A model of such a method is being developed at McGill University. (De plus en plus développement des régions inter- et multi-disciplinaires a révélé des problèmes fondamentaux dans la planification des langages d'indexation pour utiliser une collection de documents hétérogène et dispersée. On a basé les solutions traditionelles sur l'analyse des sujets des documents, mais ces nouvelles régions comprennent des éléments ou des aspects dont chacun montre une terminologie variée et une organisation differente. Le besoin et l'orientation des utilisateurs d'une nouvelle région déterminent sa caractéristique propre. On présente ici une méthode de développement d'un langage d'indexation par l'analyse d'un reseau de concepts tel que suggèré par les utilisateurs eux-mêmes. On décrit un modèle de la méthode développée à l'Université McGill)

THE NATURE OF THE PROBLEM

Conventional methods of developing index languages are based for the most part on an acceptance of a fairly traditional view of knowledge and the development of the disciplines in which our culture has organized it. This view is no longer entirely valid, and today we must understand the very real effect on information science of the changes in the way disciplines develop, and of our perception of those changes. In the past knowledge often developed by simple accretion, until a super-saturated area would, by the trigger of a new theory or discovery, crystallise into constituent disciplines, or at least yield up a coherent and previously undefined discipline from within, as anatomy came out of biology in the seventeenth century. If simple fission did not occur then fusion

contributory part of the system, has been shown by, for example, Okes,[6] in his examination of the use of symbols in effective communication by Americans with natives of Thailand, and by Michel Cartier of the Banque d'Information de Kébec, who uses pictorial slides to interface the concepts and language of a rural community and, say, government, in order to provide a common ground for communication. Obviously acceptance of all of the detail of individual concept systems leads to semantic chaos -- but studies like those of Nunnally and Flangher [7] show that individual differences in word usage and learning, perception and personality, resolve into some degree of consistency among people with a common background and common interest. It is this kind of situation that frequently exists in information retrieval contexts, and that this model proposes to exploit.

THE ORIGIN OF SUBTIL

In an initial pilot experiment in the University of Maryland in 1969, groups of documentation students were invited to respond to a trigger term with the five terms each associated most closely or significantly with it. The trigger term was book. The total response in a group of thirty or so was well over seventy terms -- a large number when the homogeneity and professional inclination of the group is considered. Even more startling was an examination of the individual concept systems in which frequently no terms were held in common. For example, three representative systems were:

book: civilisation/communication/knowledge/reading/writing
book: cataloguing/library/shelves/author/subject
book: title page/pages/index/contents/print

The initial impression was of a semantic diversity so great as to cripple effective communication.

However, as Nunnally and Flangher had predicted, an examination of sufficient concept systems revealed an increasing number of terms that recurred in system after system -- library, culture, reading, civilization. These terms represented nodes in the generalized network -- the terms acknowledged by the majority of that group as common concepts in relationship to the term book. Repetition of the experiment in the College of Librarianship, Wales in 1970 produced similar results and it became clear that the experiment could be extended using the nodal terms as second generation trigger terms. Analysis of the response to second generation trigger terms revealed more nodal terms in an expanding network to be used as third generation trigger terms. At the same time, recognition could now be made of pairs of terms occurring in individuals' concept systems. Weaker pairs could also be recognized when constituent terms occur in two generations of response in the concept system of a single individual. Occasionally (and especially in later generations of trigger and response) all the terms in a given response to a trigger term are unique in the total system; at this point that particular line of development is closed off at the trigger term, and the unique terms are held temporarily for examination as suppressed terms deserving entry as use or see references.

might, typically in the nineteenth century when a new discipline might be recognized in the no-mans-land on the boundaries of two established disciplines, like biochemistry. But the growth and complexity of knowledge is now such that we have neither the time nor the opportunity to allow fission or fusion. Instead we find synthetic disciplines like cybernetics, environmental studies, bioethics, or molecular biology: conscious combinations of relevant parts of several disciplines.

These new multi-disciplinary and inter-disciplinary areas present the information scientist with new problems. The core concepts and terminology of a new area will almost certainly be in the fringes of each constituent discipline -- scattered and unconnected. At least until a new and original literature emerges, documents will be drawn from the constituent disciplines and originally conceived within their several orientations. Relationships among terms, as required or determined by utility in the new area, will be quite unlike sets of relationships in the constituent disciplines. In other words the recognition of a conceptual base for the development of an index language in a new multi- or inter-disciplinary area should be less in the document (as it has been in the past) and more in the user, since the user is now the determining characteristic of the system. Conventional and painstaking analysis of the document base (often taking at least a man-year to yield a tentative language) should give way to an examination of the concept systems of the users.

THE SIGNIFICANCE OF USERS' CONCEPTS AS A BASIS FOR INDEXING

The value of investigating user vocabulary and responses was recognized early in the information retrieval field by Mooers,[1] who interviewed client engineers to establish a meaningful vocabulary that referred to a sample document set. A study by Bourne and others in 1961[2] advocated interviews to establish user requirements and user characteristics. Some very interesting work reported by Greer in 1962 [3] and 1965 [4] used questionnaires and interviews to elicit typical questions that the documents might answer and hopefully therefore fundamental user-oriented terms with which to index the literature. But all of these studies approached the user at the level of explicit language and extrinsic labels, and indeed often in the context of the documents.

Nearer the concerns of the work described in this paper was Phyllis Reiser's [5] work on evaluating a growing thesaurus by interacting with users -- but even here the basis was a conventionally developed language base.

Something more fundamental than all of these is needed, since we have still not reached the users' own concepts, or the question-formulation level. This is not the place to examine the different natures of ignorance and knowledge -- but it has always seemed to me ironic to expect the enquirer to use the concepts and terminology of the authors of the very documents he is seeking to allay his ignorance.

The value in meaningful communication of the level of meaning below (as it were) expressed language, even though expressed language is a

SUBTIL

THE INTENTION AND METHOD OF SUBTIL

These experiments led to the basis of the proposed model. For reasons already implied the system is assumed to work best for small homogeneous groups, for example, for research teams approaching a new area or problem. Such teams need to assemble a document and information collection perhaps from a variety of constituent subject areas, and the nature of the situation is such that they may value documents for a reason that is secondary to the documents' original intent. They will need to do this immediately and they will need the documents indexed as soon as possible. But existing index languages will almost certainly be inappropriate, and the team cannot afford to wait for a year while a new language is developed. Most of all they need to avoid the lack of responsive reaction that might arise from a reliance on the concepts and language of the documents in the context of their original disciplines.

The method described here begins by offering a single trigger term selected more or less at random from the generally accepted vocabulary of the area under study. It may not even matter if the initial trigger term is an invalid term for the final language; subsequent analysis will discard it for lack of association with any but its immediate response terms. Members of the user group are asked to respond to this initial trigger term with five terms that, in the context of their mission, they associate most significantly with it. Analysis of the total set of response systems reveals: first, a ranked list of the most frequently occurring terms; and second, pairs of terms occurring in individual concept systems. The association of a term frequently with another improves its position in the simple ranking, whose top five terms are used as second level trigger terms; paired terms are also stored for future association or reference. The user group responds to each of the second generation trigger terms again with five responses to each, and again the analysis reveals the list of candidate third generation trigger terms using simple frequency counts, further weighted by recognizing pairs. In addition, however, pairs are now recognized across generations in the responses of the same individual user, and used for an intermediate weighting in ranking single candidate terms and indicating association sets. As more and more terms are added to the list the effort of responding to trigger terms becomes greater, and a constraint is imposed on the rate of language development.

Clearly, however, much of the basic analysis of response systems may be done by a computer. The current version of the model uses simple list processing techniques in batch mode, but later versions of the model will use an interactive mode in which each user is given a set number of terms to respond to only one trigger term at a time. The computer will store responses until the full generation set is in before embarking on the next round of analysis. Within each available generation, a user may go on calling up trigger terms to respond to as long as he wishes, thus accelerating the development of the language.

SUBTIL

Some refinements of analysis are possible in the mechanical stage - for instance, different terms revealed consistently by the analysis to have an identical group of associated terms, may be extracted for examination as synonyms; terms associated with a set whose members as trigger terms never provoke the original terms as responses may reveal the existence of general/special relationship (i.e. BT's and NT's). Other possibilities of this kind are still being revealed, but it must be realised that for a long time they will almost certainly have to be augmented and refined by a final, augmented examination.

The cut-off point in the development of the language occurs when later generations of responses reveal an unacceptable rate of duplication of the terms in the growing vocabulary. Of course the vocabulary would and should always continue to grow as new problems arise, and as new members are added to a team; indeed the flexibility of the method of development suggests that if the orientation of a field of research alters radically the index language would follow as quickly as the users' awareness of the change allowed.

Practical demonstrations with conventional methods of index language development have shown that a vocabulary of one thousand basic terms or less is not uncommon in a limited field (which the field of interest of a small homogeneous group would almost certainly be), with a rather smaller additional number of use references. Recent investigation (eg. the ISILT project in the College of Librarianship, Wales)[8] suggests that effective vocabularies may be smaller than previously thought.

The fifth generation produces 625 trigger terms -- and if we accept that the fifth generation should be responded to in order to provide confirmation of the candidate vocabulary of terms, we have involved each user in responding to a total of 781 terms. Since the users' response should be as automatic as possible the response time is ideally brief and the essential time constraint on the system is the resilience of the user. The work so far has used an experimental figure of five responses to each trigger. This figure may be adjusted to vary the extent of the response and the number of generations involved.

The obvious and immediate application of such a language is to index a new collection of documents appropriately for its users. However, it may also be used as a concordance between user language and existing index languages. It might even offer assistance in text searching procedures.

There is a final consideration. As the example of the Dewey Decimal Classification and the Library of Congress list of Subject Headings has shown in libraries and as even the ERIC Thesaurus has shown in the field of information science, there is a natural reluctance to discard (or even criticise) any instrument in which much time and money have been invested. The simpler we may make our language development methods the less we should be attached to the product once it has outlived its usefulness (and in the

dynamic world of inter-disciplinary and multi-disciplinary development this is an increasingly common phenomenon). The users' thesaurus becomes a recyclable thesaurus.

REFERENCES

1. MOOERS, C. N. "The indexing language of an information retrieval system", In Information retrieval today; ed. W. Simonton. Minneapolis, Minnesota: Center for Continuation Study, University of Minnesota, 1963. pp 21-36.

2. BOURNE, C. P. Requirements, criteria, and measures of performance of information storage and retrieval systems, by C. P. Bourne, G. D. Peterson, B. Lefkowitz and D. Ford. Palo Alto: Stanford Research Institute, 1961.

3. GREER, F. L. Word usage and implications for storage and retrieval. Washington, D. C.: General Electric Co., Information Systems Operations, 1962.

4. GREER, F. L. "User vocabulary in thesaurus development", Perceptual and motor skills, vol. 21, (1965), pp 827-837

5. REISER, Phyllis. Evaluation of a growing thesaurus. Yorktown Heights: IBM Watson Research Centre, 1966. (RC 1662)

6. OKES, I. E. "Effective communication by Americans with Thai", J. Quart, vol. 38, (1961), pp 337-341.

7. NUNNALLY, J. C. and FLANGHER, R. L. "Psychological implications of word usage", Science, vol. 140, (1963), pp 775-781.

8. KEEN, E. M. and DIGGER, J. A. Report of an information science languages test. Aberystwyth: College of Librarianship Wales, 1972.