REPRESENTATION OF TEXTS FOR INFORMATION RETRIEVAL

N.J. Belkin, B.G. Michell, and D.G. Kuehner
University of Western Ontario

The representation of whole texts is a major concern of the field known as information retrieval (IR), an important aspect of which might more precisely be called 'document retrieval' (DR). The DR situation, with which we will be concerned, is, in general, the following:

a. A user, recognizing an information need, presents to an IR mechanism (i.e., a collection of texts, with a set of associated activities for representing, storing, matching, etc.) a request, based upon that need hoping that the mechanism will be able to satisfy that need.

b. The task of the IR mechanism is to present the user with the text(s) that it judges to be most likely to satisfy the user's need, based upon the request.

c. The user examines the text(s) and her/his need is satisfied completely or partially or not at all. The user's judgement as to the contribution of each text in satisfying the need establishes that text's usefulness or relevance to the need.

Several characteristics of the problem which DR attempts to solve make current IR systems rather different from, say, question-answering systems. One is that the needs which people bring to the system require, in general, responses consisting of documents about the topic or problem rather than specific data, facts, or inferences. Another is that these needs are typically not precisely specifiable, being expressions of an anomaly in the user's state of knowledge. A third is that this is an essentially probabilistic, rather than deterministic situation, and is likely to remain so. And finally, the corpus of documents in many such systems is in the order of millions (of, say, journal articles or abstracts), and the potential needs are, within rather broad subject constraints, unpredictable. The DR situation thus puts certain constraints upon text representation and relaxes others. The major relaxation is that it may not be necessary in such systems to produce representations which are capable of inference. A constraint, on the other hand, is that it is necessary to have representations which can indicate problems that a user cannot her/himself specify, and a matching system whose strategy is to predict which documents might resolve specific anomalies. This strategy can, however, be based on probability of resolution, rather than certainty. Finally, because of the large amount of data, it is desirable that the representation techniques be reasonably simple computationally.

Appropriate text representations, given these constraints, must necessarily be of whole texts, and probably ought to be themselves whole, unitary structures, rather than lists of atomic elements, each treated separately. They must be capable of representing problems, or needs, as well as expository texts, and they ought to allow for some sort of pattern matching. An obvious general schema within these requirements is a labelled associative network.

Our approach to this general problem is strictly problem-oriented. We begin with a representation scheme which we realize is oversimplified, but which stands within the constraints, and test whether it can be progressively modified in response to observed deficiencies, until either the desired level of performance in solving the problem is reached, or the approach is shown to be unworkable. We report here on some linguistically-derived modifications to a very simple, but nevertheless psychologically and linguistically based word-co-occurrence analysis of text [1] (figure 1).

POSITION	RANK (r)
Adjacent	1
Same Sentence	2
Adjacent Sentences	3

FOR EACH CO-OCCURRENCE OF EACH WORD PAIR (w_1, w_2)

$$\text{SCORE} = \frac{1}{1 + r} \times 100$$

FOR ALL CO-OCCURRENCES OF EACH WORD PAIR IN TEXT

ASSOCIATION STRENGTH = SUM (SCORES)

Figure 1. Word Association Algorithm

The original analysis was applied to two kinds of texts: abstracts of articles representing documents stored by the system, and a set of 'problem statements' representing users' information needs -- their anomalous states of knowledge -- when they approach the system. The analysis produced graph-like structures, or association maps, of the abstracts and problem statements which were evaluated by the authors of the texts (Figure 2) (Figure 3).

CLUSTERING LARGE FILES OF DOCUMENTS
USING THE SINGLE-LINK METHOD

A method for clustering large files of documents using a clustering algorithm which takes O(n**2) operations (single-link) is proposed. This method is tested on a file of 11,613 documents derived from an operational system. One property of the generated cluster hierarchy (hierarchy connection percentage) is examined and it indicates that the hierarchy is similar to those from other test collections. A comparison of clustering times with other methods shows that large files can be clustered by single-link in a time at least comparable to various heuristic algorithms which theoretically require fewer operations.

Figure 2. Sample Abstract Analyzed

In general, the representations were seen as being accurate reflections of the author's state of knowledge or problem; however, the majority of respondents also felt that some concepts were too strongly or weakly connected, and that important concepts were omitted (Table 1).

We think that at least some of these problems arise because the algorithm takes no account of discourse structure. But because the evaluations indicated that the algorithm produces reasonable representations, we have decided to amend the analytic structure, rather than abandon it completely.

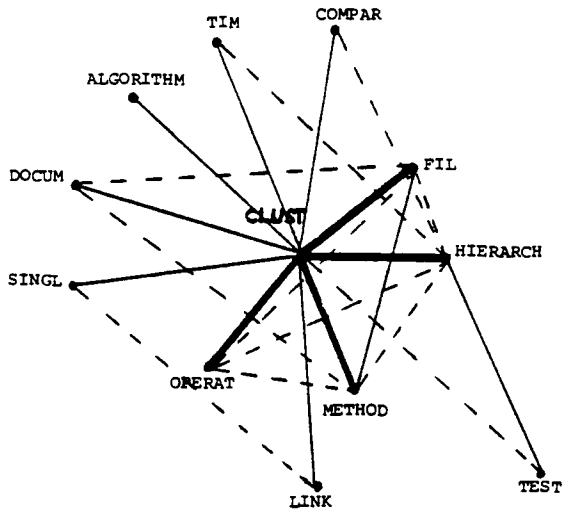

Figure 3. Association Map for Sample Abstract

Table 1. Abstract Representation Evaluation

Question		% YES	% NO	% INTERM.	% NO RESP.	
1.	ACCURATE REFLECTION?	48.0	29.6	22.0		N=30
2.(a)	CONCEPTS TOO STRONGLY CONNECTED?	63.0	37.0			N=30
(b)	CONCEPTS TOO WEAKLY CONNECTED?	96.3	3.7			N=30
3.	CONCEPTS OMITTED?	88.9	11.1			N=30
4.	IF NO OR 'INTERM' to No. 1, WAS ABSTRACT ACCURATE?	64.3	7.1	21.4	7.1	N=14

Our current modifications to the analysis consist primarily of methods for translating facts about discourse structure into rough equivalents within the word-co-occurrence paradigm. We choose this strategy, rather than attempting a complete and theoretically adequate discourse analysis, in order to incorporate insights about discourse without violating the cost and volume constraints typical of DR systems. The modifications are designed to recognize such aspects of discourse structure as establishment of topic; setting of context; summarizing; concept foregrounding; and stylistic variation. Textual characteristics which correspond with these aspects include discourse-initial and discourse-final sentences; title words in the text; equivalence relations; and foregrounding devices (Figure 4).

1. Repeat first and last sentences of the text.
   These sentences may include the more important concepts, and thus should be more heavily weighted.
2. Repeat first sentence of paragraph after the last sentence.
   To integrate these sentences more fully into the overall structure.
3. Make the title the first and last sentence of the text, or overweight the score for each co-occurrence containing a title word.
   Concepts in the title are likely to be the most important in the text, yet are unlikely to be used often in the abstract.
4. Hyphenate phrases in the input text (phrases chosen algorithmically) and then either: a. Use the phrase only as a unit equivalent to a single word in the co-occurrence analysis; or b. use any co-occurrence with either member of the phrase as a co-occurrence with the phrase, rather than the individual word.
   This is to control for conceptual units, as opposed to conceptual relations.
5. Modify original definition of adjacency, which counted stop-list words, to one which ignores stop-list words. This is to correct for the distortion caused by the distribution of function words in the recognition of multi-word concepts.

Figure 4. Modifications to Text Analysis Program

We have written alternative systems for each of the proposed modifications. In this experiment the original corpus of thirty abstracts (but not the problem statements) is submitted to all versions of the analysis programs and the results compared to the evaluations of the original analysis and to one another. From the comparisons can be determined: the extent to which discourse theory can be translated into these terms; and the relative effectiveness of the various modifications in improving the original representations.

Reference

1. Belkin, N.J., Brooks, H.M., and Oddy, R.N. 1979. Representation and classification of knowledge and information for use in interactive information retrieval. In Human Aspects of Information Science. Oslo: Norwegian Library School.