

**Syracuse University**

---

**From the Selected Works of Robert Oddy**

---

September, 1982

# ASK FOR INFORMATION RETRIEVAL: PART II. RESULTS OF A DESIGN STUDY

Robert N Oddy, *Syracuse University*

N J Belkin

H M Brooks



Available at: [https://works.bepress.com/robert\\_oddy1/6/](https://works.bepress.com/robert_oddy1/6/)

# ASK FOR INFORMATION RETRIEVAL: PART II. RESULTS OF A DESIGN STUDY

The Journal of Documentation, volume 38, number 3, September 1982

**N. J. BELKIN**  
Centre for Information Science  
The City University, London

**R. N. ODDY**  
Computer Centre  
The University of Aston, Birmingham

**H. M. BROOKS**  
Centre for Information Science  
The City University, London

In *'ASK for Information Retrieval: Part P1*, we discussed the theory and background to a design study for an information retrieval (IR) system based on the attempt to represent the anomalous states of knowledge (ASKs) underlying information needs. In Part II, we report the methods and results of the design study, and our conclusions.

## Methods

THE EXPERIMENTAL PROCEDURE employed in the design study was basically a three-stage procedure, consisting of: collection of material (problem statements/abstracts); text analysis and production of structural representations; evaluations of analysis and representations. The specific objectives of the design study were:

- (a) Tape recording a number of interviews with users of actual information systems. Users were to be asked to discuss the problem with which they were faced prior to presenting a more formal request to the system.
- (b) Collection of thirty abstracts in information science of 200-300 words each, of articles written by authors in the UK.
- (c) Adaptation and implementation of the text analysis program developed by Belkin<sup>3</sup> so as to produce structural representations of this data.
- (d) Obtaining the authors' /users' evaluations of these representations, through the use of questionnaires or interviews where appropriate.
- (e) Analysis of the problem statements according to various structural characteristics (e.g. strength of association, number of nodes, degree of coherence) and classifying them.
- (f) The development of preliminary retrieval strategies based on the classification and structural features of representation.

## Problem statements

The initial stage of the experiment involved the collection of suitable problem statements for analysis and subsequent evaluation. Statements were considered suitable which:

- (a) were directed towards the users' work or research situation rather than to what they specifically hoped to find;
- (b) were as uninfluenced as possible by knowledge of the operations of the information retrieval systems;
- (c) were representatives of real information needs expressed by the person with that need.

Subjects were drawn from the users of Central Information Services (LRCC) at the University of London, which carries out online bibliographic searches on behalf of students and staff from all the colleges and institutes of the University of London. Mrs A. Vickery, the Senior Information Systems Officer, kindly gave us permission to approach CIS users to ask if they would take part in the experiment.

The purpose of the research was explained to each user and they were told they would be asked to evaluate the analysis of the statement.

The interviews with the subjects were short, informal, open-ended and unstructured, in order to elicit maximum information with minimum intervention." One question only was posed:

*'Would you like to talk informally for a few minutes about the research you are doing at the moment, the problem that has led you to have a search carried out, and the sort of information you would like to have as a result of the search?'*

We maintained some flexibility, however, in order to take account of individual differences among subjects, such as tendencies to stray from the point. The basic question always remained the same, but there was also some interviewer intervention with some subjects.

The interviews were carried out before the normal pre-search interview or search itself, and were tape recorded. In all, there were twenty-seven interviews, twenty-three from CIS users and four from M.Sc. students at The City University who were just beginning work on their theses. Subjects were not selected everyone arriving at CIS for a search during the period of data collection (one month) was approached, and all agreed to participate. Both the topics of the searches and the background of the subjects were therefore widely dispersed: topics ranged from medicine to linguistics to education, status from beginning M. Phil. student to clinician to professor (see Table 1).

Eight written queries were also chosen from the CIS files and analysed in order to see if there were systematic differences between them and the oral problem statements. Those chosen were selected on the basis that they contained details of the user's research as well as of what she/he wanted to know.

Evaluation involved sending two alternative structural representations (derived from the text analysis of the problem statements) back to the users and asking for their assessment of them. Questionnaires were used for this purpose a - separate one for each representation and one comparing the two.

TABLE 1. *Subject areas of interviewees. (From Brooks<sup>6</sup>)*

	<i>Psychology /Education /Sociology /Linguistics</i>	<i>Medicine</i>	<i>Agriculture</i>	<i>Information Science</i>	<i>Biology /Chemistry /Biochemistry</i>
	1	3	9	14	7
	2	4	10	15	17
	6	5	11	16	26
	8	19		18	27
	12	20			
	13	21			
	22	23			
	24	25			
<i>Totals</i>	8	8	3	4	4

The evaluation 'package' therefore consisted of: cover letter; transcript; first representation and its questionnaire; and the comparison questionnaire. stapled together. The order of the representations was reversed for alternate subjects to guard against possible interactive effects of one format on the other (see Appendix A).

### **Abstracts**

A total of thirty-one abstracts were collected. These were abstracts from recent publications (journal articles or published conference proceedings) by UK authors in the field of library and information science. They were chosen from *Library and Information Science Abstracts* and from our personal journal files and were mostly author abstracts of varying lengths ranging from 90 to 2000 + tokens. Restrictions on date and author location were for ease of evaluation.

Evaluation of the representations produced by the analysis was again done by means of questionnaire (see Appendix B). It was decided, however, as a result of the interview evaluations, to use only one of the representation types. A package containing explanatory cover letter, a copy of the abstract, the representation and a questionnaire was sent to each author.

### **Text analysis**

One of the assumptions underlying this project is that a state of knowledge can be represented by a network of associations between words, standing for concepts. The text analysis algorithms implemented here are an attempt to produce such representations.

The technique used to specify associations between words is linguistically very unsophisticated. It computes a strength of association between every pair of 'significant' word-types in the text under examination, based on lexical separation. There is no syntactic analysis and no attempt to represent the nature of an association. The assumptions behind the measure of associative strength used are that if two words are closely associated in the searcher's knowledge structure, then they will be close in the textual expression of that structure and will tend to find multiple expression in the text. The strength of association between word types in the text was computed from the list of occurrences as follows:

First, for any particular pair of occurrences of two types *A* and *B*, a score reflecting the distance between these occurrences is calculated:

$$\text{Score} = (1/(1 + r)) \times 100$$

Where:

$r=1$  if *A* & *B* are adjacent within the same sentence

$r= 2$  if *A* & *B* are within the same sentence but not adjacent

$r= 3$  if *A* & *B* are in adjacent sentences within the same paragraph.  
Adjacency was established before the elimination of non-significant words.

If the co-occurrences of *A* & *B* are any further apart, the score is zero.  
(The factor of 100 is purely for computerized convenience).

Then, having calculated the scores for all pairs of occurrences of *A* & *B* in the text, these are summed to give the strength of association for that word pair. As an example of this association algorithm, Figure 1 is a sample problem statement, and Figures 2 and 3 the Association Map and Association Cluster representations derived from it by the algorithm, respectively.

MY PROJECT IS LOOKING INTO THE INFORMATION SERVICES OF PROFESSIONAL INSTITUTES IN BUSINESS MANAGEMENT. BY INFORMATION SERVICES I DON'T MEAN JUST THE LIBRARY. I AM TRYING TO COVER PUBLICATIONS, CONFERENCES, SEMINARS, EXHIBITIONS, MEETINGS, ENQUIRY SEARCHES AND THAT KIND OF THING. I HOPE TO DO THIS BY GOING AROUND TO EACH INSTITUTE AND TALKING TO SOMEONE, USUALLY THE LIBRARIAN, AND GOING THROUGH A QUESTIONNAIRE, COLLECTING DATA ON THIS PARTICULAR INSTITUTE. THERE HASN'T BEEN MUCH EVALUATION DONE ON PROFESSIONAL ASSOCIATIONS SO FAR. I ALSO WANT TO FIND OUT WHAT LITERATURE THEY ARE EXPECTED TO KEEP AND TYPES OF SERVICES THEY WANT TO OFFER, TYPES OF BUSINESS INFORMATION THAT ARE NECESSARY FOR PEOPLE IN THESE INSTITUTES. I ALSO WANT TO KNOW OF ANY EVALUATIONS IN INSTITUTES OR LIBRARIES IN THAT FIELD THAT HAVE ALREADY BEEN CARRIED OUT.

*FIG. 1. The transcript of a problem statement*

Identification of word types, which are to stand for concepts in the structure, was done by eliminating words non-significant for our purposes, and by conflating the various forms of the significant ones. Thus, the nodes of the derived structures are labelled with truncated words—we take the liberty of calling them stems. No attempt was made to recognize synonyms. Although most words conventionally regarded as poor indexing terms were removed by the stop list, words such as 'how', 'why', 'not', 'effect' and 'presume', which are important semantically in the statements, were included.

Since evaluation of the results of the analysis by the interviewees and authors was to be the major method of assessment of the appropriateness of the analysis, it was necessary to devise methods of displaying the associative structures. The most straightforward method is simply to draw a network in which the nodes, labelled with 'stems', represent concepts and the lines represent associations (see figure 2)



For clarity, the large mass of weak associations were omitted and the associative strengths simplified into strong, medium and weak associations, represented in the network as thick, thin and broken lines respectively. Some indication of strength was also given by length of line, although in a two-dimensional picture this cannot, in general, be done very accurately or consistently. This representation was called the 'Association Map' format.

Another method of displaying the associative structure of a text is to cluster word types. A hierarchy of single-link clusters of types for each text was generated using an algorithm originally developed for document clustering by Van Rijsbergen (see Figure 3). The horizontal distances from the branch points to the words on the left of the picture are directly related to the association strengths with which the cluster members are bound together. This representation was called the 'Association Cluster' format.

## **RESULTS**

### **Problem statements\***

The point of the surveys of users and authors was to see whether the analyses of the problem statements and abstracts were in general accord with the originator's own perceptions of their information needs or of the ideas they were attempting to communicate; and, if there were disparities, then to see if there might be suggestions for improvement. [This section and that following are based on Brooks, Oddy and Belkin]

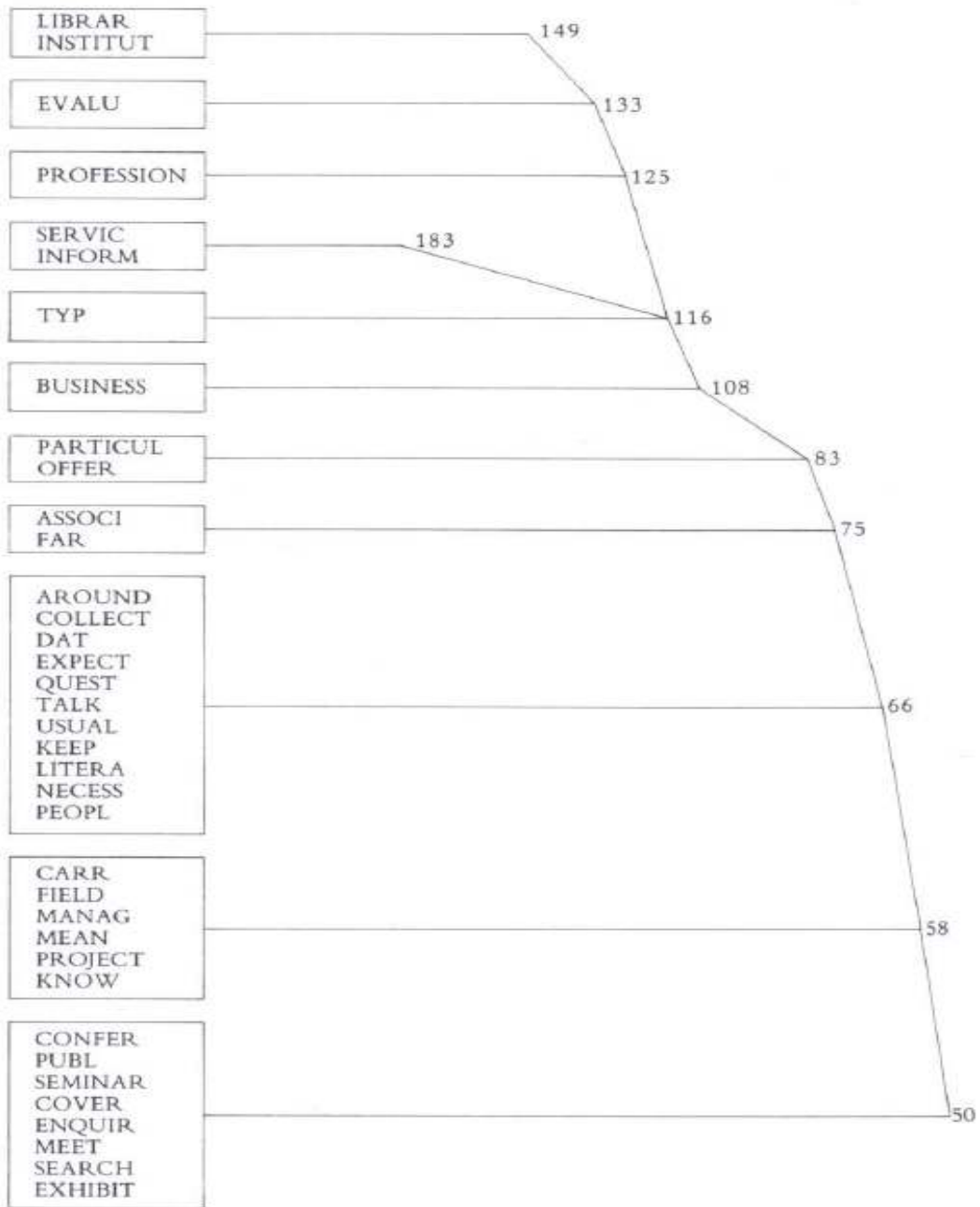


FIG. 3. Association Clusters for problem statement of Fig. 1



Of course, for retrieval purposes it may not be necessary for the representations to be congruent with the originators' ideas about them, but as a first method of evaluation the technique seemed reasonable. If the subjects were unanimous in their disapproval of the representations, then we could be fairly sure that we should probably try something else.

We wished to determine, in evaluating the problem statement representation:

1. how accurately, in the interviewee's opinion, the two formats described her/his ASK at the time of the interview; and
2. how the two formats compared with one another.

Response to the survey was good, 63% of the group (N = 27) returning completed questionnaires. Table 2 is a summary of replies to the Association Map questionnaire, Table 3 to the Association Cluster questionnaire and Table 4 to the comparative questionnaire.

TABLE 2. Association Map evaluation (problem statements)

Question	% Yes	% No	No Resp. (N = 15)
1. Accurate reflection?	73.3	20.0	6.6%
2. (a) Too strong?	46.6	53.4	
(b) Too weak?	86.6	13.4	
3. Concepts missing?	46.6	40.0	13.3%

TABLE 3. Association Cluster evaluation (problem statements)

Question	% Yes	% No	No Resp. (N = 15)
1. (a) Concepts which should not be together?	46.6	46.6	6.6
(b) Concepts which should be together?	73.3	20.0	6.6
2. (a) Groups at too high an association level?	60.0	40.0	
(b) Groups at too low an association	60.0	40.0	
3. (a) Groups not closely enough linked?	80.0	20.0	
(b) Groups too closely linked?	46.6	53.4	

TABLE 4. Format comparison (problem statements)

Question	% Yes	% No	No Resp.
1. Any preference?	73.0	26.6	N = 15
2. Prefer Association Map	81.8		N = 11
3. Prefer Association Cluster	18.2		N = 11
4. If no preference, were both unsuccessful?	75.0		25.0% N = 4

to communicate; and, if there were disparities, then to see if there might be suggestions for improvement. Of course, for retrieval purposes it may not be necessary for the representations to be congruent with the originators' ideas about them, but as a first method of evaluation the technique seemed reasonable. If the subjects were unanimous in their disapproval of the representations, then we could be fairly sure that we should probably try something else.

We wished to determine, in evaluating the problem statement representation:

1. how accurately , in the interviewee's opinion, the two formats described her/his ASK at the time of the interview; and
2. how the two formats compared with one another.

Response to the survey was good, 63% of the group (N = 27) returning completed questionnaires. Table 2 is a summary of replies to the Association Map questionnaire, Table 3 to the Association Cluster questionnaire and Table 4 to the comparative questionnaire.

TABLE 2. *Association Map evaluation (problem statements)*

Question	% Yes	% No	No Resp. (N = 15)
1. Accurate reflection?	73.3	20.0	6.6%
2. (a) Too strong?	46.6	53.4	
(b) Too weak?	86.6	13.4	
3. Concepts missing?	46.6	40.0	13.3%

TABLE 3. *Association Cluster evaluation (problem statements)*

Question	% Yes	% No	No Resp. (N = 15)
1. (a) Concepts which should not be together?	46.6	46.6	6.6
(b) Concepts which should be together?	73.3	20.0	6.6
2. (a) Groups at too high an association level?	60.0	40.0	
(b) Groups at too low an association	60.0	40.0	
3. (a) Groups not closely enough linked?	80.0	20.0	
(b) Groups too closely linked?	46.6	53.4	

TABLE 4. *Format comparison (problem statements)*

Question	% Yes	% No	No Resp.
1. Any preference?	73.0	26.6	N = 15
2. Prefer Association Map	81.8		N = 11
3. Prefer Association Cluster	18.2		N = 11
4. If no preference, were both unsuccessful?	75.0		25.0% N = 4

From these tables, it is evident that the analysis, presented in the Association Map format, provided a generally adequate representation of the information needs of the interviewees. The major criticism of the analysis is that some concepts were too weakly associated, and this seems to be its single identifiable consistent problem, in the judgement of the interviewees. The Association Cluster format was judged inferior to the Association Map.

### Abstracts

The goals of the abstract evaluation were to see:

1. how accurately the analysis, presented in the Association Map format, represented the interrelation of concepts in the mind of the author at the time of writing; and
2. if there were any regularities in deficiencies of the representation.
- 3.

We chose to use only the Association Map format because of the general dissatisfaction with the Cluster format among the interviewees who responded to the questionnaire.

The response rate to this survey was gratifyingly high: 90% (N = 30). The results are summarized in Table 5. There are, unfortunately, some difficulties in interpreting these results (see *Discussion* below), but it appears that although the general representation method is judged reasonable, there are some severe problems in its specific implementation. Thus, although only about 30% of the respondents thought that the representation was actually bad, 63% thought that some concepts were omitted, and, most striking, 96% (all but one) thought that at least some concepts were too weakly connected.

TABLE 5. *Association Map evaluation (abstracts)*

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

#### Classification of ASKs

We have attempted a classification of the ASKs underlying problem statements, using easily computed characteristics of the derived associative structures. For the future, what we need is a classification which will help us to select an appropriate retrieval strategy; that is, a classification with predictive power. Our first efforts, however, have been less ambitious. We have tried to find a classification which is descriptive of people's 'problematic situations' which can be algorithmically generated. This classification may or may not be useful for determining how to resolve anomalies. If we assume that:

- the representations produced by the text analysis procedure are closely related to ASKs, and
- types of anomaly are reflected in corresponding types of structural features in the representations,

We may expect a classification of representations on a structural basis to classify ASKs in a meaningful way. Thus, we present here a classification based on the (graphic) Association Map format, and how it corresponds to a subjective view of the nature of the problem statements. Both global and local structural features of association networks can be perhaps the most obvious structural characteristic of a network as a whole is the extent to which concepts are interconnected. Some networks are highly connected webs of concepts, others

are more widely dispersed or even fragmented. This feature can be measured by a connectivity score, which represents the extent to which the network falls short of being maximally connected. In the case of our problem statements, a very simple connectivity score can be used. Because the number of lines in the Association Map is constant (namely forty), we can use the following formula without normalization:

$$\text{Connectivity, } C = N - N_{\text{min}}$$

where  $N$  is the number of nodes present in the network and  $N_{\text{min}}$  is the minimum number of nodes possible, given that there are forty lines. (In fact,  $N_{\text{min}} = 10$ ). For example, the number of nodes in the network of Figure 2 is 25, so its connectivity score is 15. Our sample of problem statements was small, so the scores were pooled to produce five classes: A...0-5; B...6-10; C...11-15; D...16-20; E...21-25.

Two local structural features (stars and connected components) were considered for use in the classification scheme. To define the notion of a star, we firstly define (following graph theory terminology e.g. Christofides") the *degree* of a node to be the number of lines incident with the node. A *star* is a node which is linked to at least one node of degree 1. The number of stars in a network is the number of sets of peripheral nodes, which may reflect a difficulty on the part of the enquirer in relating aspects of her/his ASK. A *connected component* is a set of nodes, any pair of which is joined by a path of links in the network. In the sample examined, only one structure had more than one connected component, so attention was focused on the number of stars present.

A class code of the form: CONNECTIVITY CLASS; NUMBER OF STARS was assigned to each problem statement structure (e.g. the structure in Figure 2 has the code C3). A summary of the ASK types in the sample of twenty-seven interviews, as defined by this classification, is given in Table 6. On the whole, the classification does seem to divide the representations, and hence the ASKs, into meaningful groups in which common features can be discerned. The eight written scripts, when analysed, produced representations not appreciably different from the oral scripts, and classification revealed types of anomaly similar to the corresponding types found with the oral scripts.

### **Retrieval strategies**

The goal of information retrieval is to resolve those anomalies in a person's state of knowledge, which induced him or her to seek information from literature. Our approach is to select search strategies with explicit reference to characteristics of the enquirer's ASK structure.

TABLE 6. Summary of ASK types. (After Brookes<sup>6</sup>)

<i>Class</i>	<i>Group</i>	<i>No. of interviews</i>	<i>Comments</i>
A Well-defined topic and problem.	0	1	Concise presentation of problem. Information wanted for review articles.
B Specific topics. Problem well defined. Information wanted to back up research and/or hypotheses.	1	5	References wanted to back up hypotheses on which research based.
	2	1	Further information about subject of research required.
		2	2
C Topics quite specific.  Problem not so well defined. Research still at any early stage.	1	3	Fairly general bibliographies wanted.
	3	2	Research clearly described. Less sure as to what information required.
D Topics fairly specific. Problems not well defined.  No hypotheses underlying research.  Information needed to produce directions for research.	1	2	Problem involved trying to extrapolate a set of conditions from a known to an unknown situation.
	2	2	No obvious similarities between the members of this group.
	3	2	Problem not clearly defined. Information wanted which will provide ideas for the formulation of hypotheses.
E Topics and problems not well defined.  Topics often unfamiliar.	1	3	2 relate a particular subject to a number of variables. 2 doing literature search for project not yet started.
	2	2	References not wanted for research but to write paper. Paper deals with known subject area but in an unfamiliar context.
		3	1

As a general principle this approach is applicable to conventional information retrieval systems. The bibliographic tools one chooses to use, and the way one formulates the query, should depend on the precision in the definition of information need. A classification of problem statements along the lines of that discussed above could thus be of use within a conventional framework. We wish to go further than this, however, and build a system with heuristics for resolving anomalies. Certain features of the problem statement structure will be interpreted as anomalous, and documents whose structures would help to remove the anomaly will be displayed. The strategies will be used within an interactive environment, so that the system may use the searcher's reactions to judge the appropriateness of its choice of strategy. This is important because, until we can incorporate a model of natural language understanding, we must assume that the interpretations that we put upon the association networks are fallible.

In this section we would like to indicate the direction of our ideas concerning retrieval strategies. We must point out, however, that this area has received little specific attention during our preliminary project. Let us assume that the retrieval programs operate on structures similar to the simplified formats generated for the surveys reported above—i.e. consisting of about forty associations, divided into three levels of strength: strong, medium and weak.

It would seem, from a study of the structures obtained from problem statements, that the precise pattern of associations among a strongly linked group of concepts is arbitrary. Thus, in order to identify the significant features, we should condense the network, reducing clusters of strongly linked nodes to single 'super nodes'. The following example illustrates the process—we use the structure of Figure 2.

- (i) Strong clusters are defined as components singly-linked at the strong level, and are denoted by symbols within two concentric circles. Figure 4 shows the condensed network.
- (ii) Medium clusters are defined as sets of nodes (excluding nodes in strong clusters) which are singly-linked at the medium level. These are denoted by symbols within a single circle. Figure 5 illustrates this condensation.
- (iii) Weak clusters can be similarly defined (there are none in the example).

We now mention a few retrieval strategies which make use of the condensed networks. These will involve the matching of terms in selected parts of the problem statement structure, within some structural constraint in the document network. If  $A$  is the set of problem statement terms under consideration, a matching set in a document will be denoted  $K$ .

(i) Strong Clusters

One of the more reliable assumptions about problem statement structures is that strong clusters correspond to main topics in the statement. It will therefore be part of most retrieval strategies to select documents containing  $A'$  as a cluster, for at least one strong cluster,  $A$ , in the problem statement.

(ii) Multiple Strong and Medium Clusters

If the problem statement contains distinct clusters linked by associations at a weaker level than those within the clusters, our system might reasonably assume that stronger associations between the clusters might

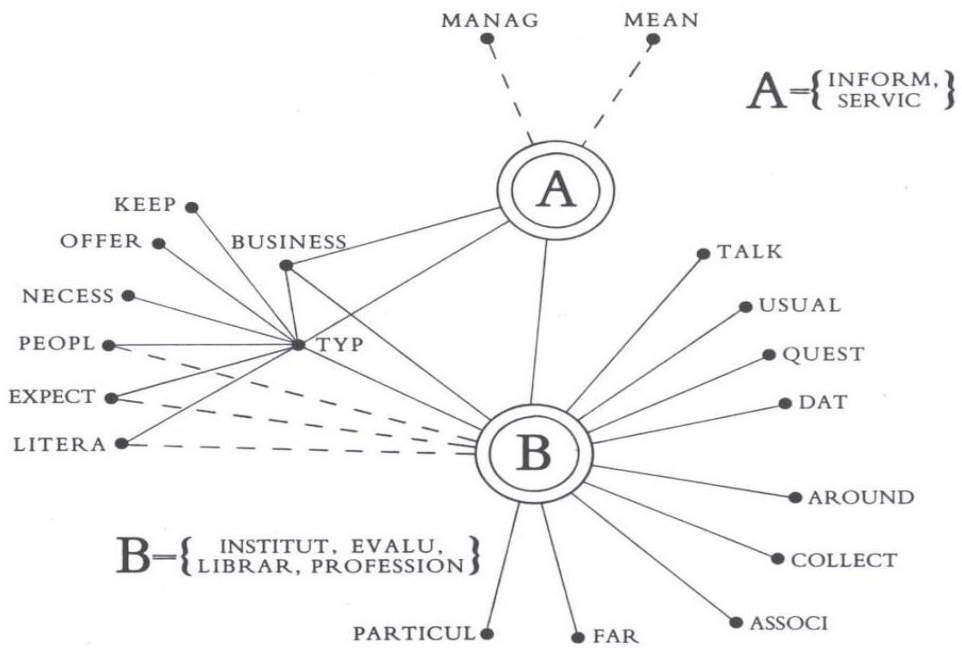


FIG. 4. Condensed problem statement network—1

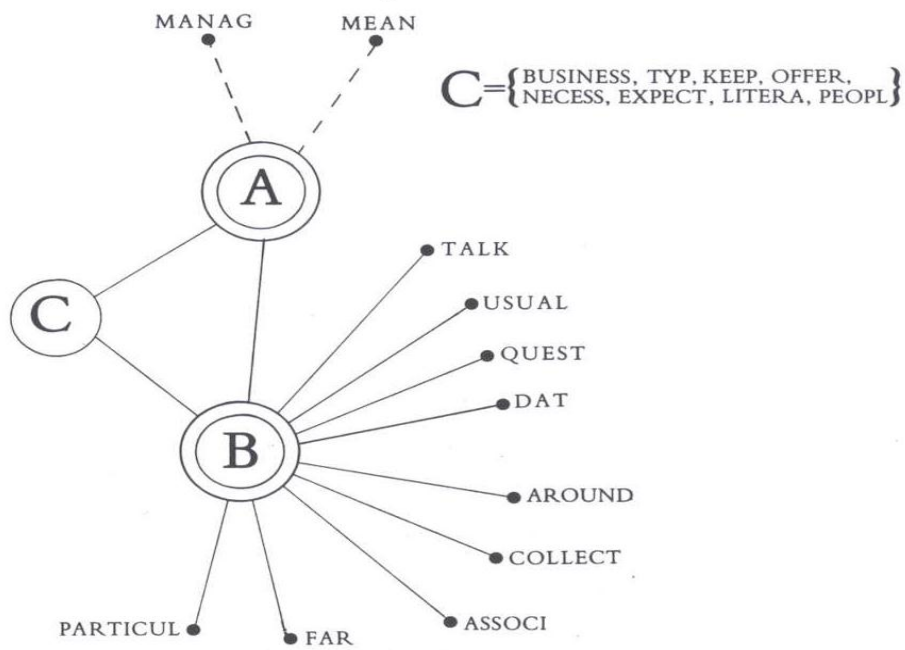
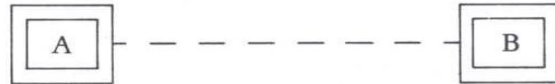


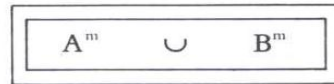
FIG. 5. Condensed problem statement network—2



help to resolve the ASK. Thus, in response to a problem statement such as this:



the first strategy would be to look for documents containing either:



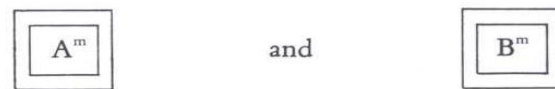
or:



It may be that the problem of associating topic *A* with topic *B* is not peculiar to the enquirer, but is an unresolved, or untackled, problem in the literature. In this case, the first strategy may fail. A second strategy would be to select documents which contain:



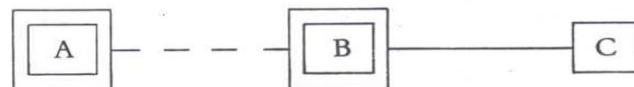
As a last resort, it may be necessary to retrieve (at least two) documents in which:



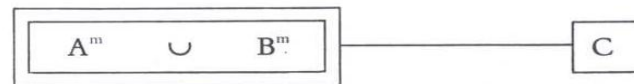
occur separately. The enquirer will then have to deduce the link for her/himself.

(iii) Medium Links

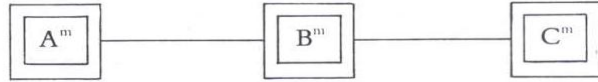
Nodes, and medium clusters which are connected by medium links to a strong cluster, may specify the *context* of the main topic. It would be appropriate to modify whichever strategy is used in connection with the strong cluster to take account of this. For example, if the problem statement structure contains:



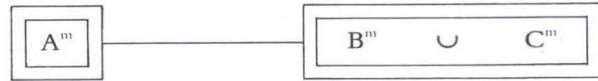
the first search would be for documents containing:



or:

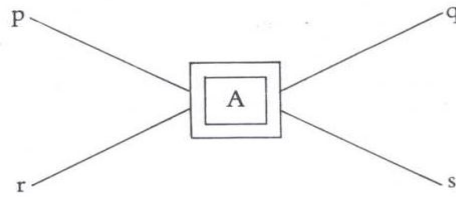


These structures would be preferred to, for example:



(iv) Stars

The essential property of a star, for our purpose, is that the nodes on the periphery are not linked to each other directly. This suggests that we should try a strategy which seeks to link peripheral nodes to each other. For example, a strategy for the problem statement structure:



is to look for documents containing:



where  $P = p, q, r, s$

A slightly more general form of the 'stars' heuristic is to form the independent node sets of the association network,<sup>11</sup> i.e. sets of mutually disconnected nodes—and seek for documents which contain matching sets as clusters.

## DISCUSSION

### Design problems

There are several problems in interpreting our data. The major difficulty lies in the questionnaire design, and has affected the results concerning evaluation of the text analysis. It is especially evident in the results for the abstract evaluation. From the responses to the questionnaire (see Appendices) it is clear *that* we should have used a rating scale rather than 'yes/no' binary choice for the general evaluative question. A number of respondents expressed difficulty in answering that question, and answers to that question were not always consistent with answers to the other questions. Although this was not made evident in the problem statement evaluation, it may be only because the respondents to the abstract evaluation were professionally familiar with the issues, and therefore more disposed to comment and criticize. So we might assume that for both groups, the judgement of general adequacy of the analysis has been constrained by the response choice. This also seems to have been a difficulty with question 4 of the abstract evaluation, but in that case is perhaps not quite so critical.

Another problem is that it is difficult to correlate responses to the questions about strength of association with those asking if any concepts were omitted. It is possible that some omitted concepts might have been those which were also too weakly related. Although we asked for examples of each case, there is no certain way to determine whether such a relationship holds. And finally, we did not stress, in our explanation, that the associations in the representation formats were only a small number of the total associations derived by the analysis, representing only the most important concepts and relationships. This may have coloured responses to questions concerning weak associations and omitted concepts, and also seems to have influenced response to question 1, negatively.

Another general difficulty lies in the representations of problem statements, where we arbitrarily chose the top forty associates for inclusion in the association map representation. This number was chosen for ease of construction and interpretation of the representation, but unfortunately resulted in the arbitrary splitting of groups of associates all with the same association strength. We realized this difficulty by the time the abstracts were evaluated, and so used a flexible cut-off level with them in order to include all associates at the lowest association strength, aiming for about forty associates. This problem may have affected response to the evaluation of problem statement representations, and cut-off level certainly needs further investigation.

### Evaluation of representations

The results of the evaluations, despite the difficulties mentioned above, seem to us encouraging. In both groups of respondents, 30% or less thought that the representation was actually bad, and in some of those cases further comments contradicted that opinion. From these data, we conclude that the basic analytic technique, word co-occurrence analysis, is in principle reasonable for our purposes. The results of the survey of authors also seem to indicate that abstracts are reasonable document surrogates in our context, and that length of the abstract is not a significant parameter, at least within the limits of our sample. And the results of the survey of interviewees indicates that our interview technique, combined with the text analysis, is a reasonable means for collecting problem statements that will be useful in an ASK-based IR system. Despite these positive results of the survey, the data concerning strength of association and omitted concepts indicate that our text analysis techniques need modification.

### Text analysis

The results of the evaluations of both problem statement and abstract representations indicate that over-strong associations are a problem, but that the major difficulty in text analysis lies in associations at too low a level. These may reflect a misunderstanding of the representation, as mentioned above, but they seem too consistent and too positive to be attributed only to this cause. Most of the examples of too weakly associated concepts involve words that occurred only once or twice in the text, sometimes in conjunction with more frequent words, but often not. At times these words appear (in the abstracts) only in the document title, which we treated as a separate paragraph. Possible means of adjusting our analysis are to include the

title as part of every paragraph, and to take account of collection frequency (when we have a large enough collection) to weight some of

Another problem is that it is difficult to correlate responses to the questions about strength of association with those asking if any concepts were omitted. It is possible that some omitted concepts might have been those which were also too weakly related. Although we asked for examples of each case, there is no certain way to determine whether such a relationship holds. And finally, we did not stress, in our explanation, that the associations in the representation formats were only a small number of the total associations derived by the analysis, representing only the most important concepts and relationships. This may have coloured responses to questions concerning weak associations and omitted concepts, and also seems to have influenced response to question 1, negatively.

Another general difficulty lies in the representations of problem statements, where we arbitrarily chose the top forty associates for inclusion in the association map representation. This number was chosen for ease of construction and interpretation of the representation, but unfortunately resulted in the arbitrary splitting of groups of associates all with the same association strength. We realized this difficulty by the time the abstracts were evaluated, and so used a flexible cut-off level with them in order to include all associates at the lowest association strength, aiming for about forty associates. This problem may have affected response to the evaluation of problem statement representations, and cut-off level certainly needs further investigation.

#### **Evaluation of representations**

The results of the evaluations, despite the difficulties mentioned above, seem to us encouraging. In both groups of respondents, 30% or less thought that the representation was actually bad, and in some of those cases further comments contradicted that opinion. From these data, we conclude that the basic analytic technique, word co-occurrence analysis, is in principle reasonable for our purposes. The results of the survey of authors also seem to indicate that abstracts are reasonable document surrogates in our context, and that length of the abstract is not a significant parameter, at least within the limits of our sample. And the results of the survey of interviewees indicates that our interview technique, combined with the text analysis, is a reasonable means for collecting problem statements that will be useful in an ASK-based IR system. Despite these positive results of the survey, the data concerning strength of association and omitted concepts indicate that our text analysis techniques need modification.

#### **Text analysis**

The results of the evaluations of both problem statement and abstract representations indicate that over-strong associations are a problem, but that the major difficulty in text analysis lies in associations at too low a level. These may reflect a misunderstanding of the representation, as mentioned above, but they seem too consistent and too positive to be attributed only to this cause. Most of the examples of too weakly associated concepts involve words that occurred only once or twice in the text, sometimes in conjunction with more frequent words, but often not. At times these words appear (in the abstracts) only in the document title, which we treated as a separate paragraph. Possible means of adjusting our analysis are to include the title as part of every paragraph, and to take account of collection frequency (when we have a large enough collection) to weight some of the content words. Some work on modification of the analysis program to take account of these factors, and of discourse structure theory, has been done, and appears to be reasonably successful. We also need to adjust cut-off techniques in some way, to include more than just the top forty associates, but fewer than all of the associates in the text representation. We have done some preliminary investigations on this problem, and it appears that using all associates until each type is represented is a reasonable and useful cut-off rule.

The data from the text analysis as it now stands are rather complex, and probably need to be reduced before they can be useful for retrieval purposes. We have done this to some extent in the association map representations by choosing three levels of association strength, but the techniques for choosing levels need to be refined.

The classification of problem statements based on the text analysis appears to be reasonably successful, given its aims. It is interesting to note that there are few queries in our sample which can be described as well defined. The one in class AO stands out in this respect, from both the subjective and computational points of view (it has a connectivity score of zero). We believe that conventional best-match retrieval systems operate under the assumption that queries are well defined. Our results appear to belie that assumption, and it is encouraging that we can come to this conclusion by computational and algorithmic means. Nevertheless, the classification should be modified to take more account of levels of association.

### Conclusions and further research

What we have described here are some results of a design study for what we envisage as a much larger-scale project, to design, construct, implement and test an IR system based specifically and explicitly upon the recognition that the representation of information need is the central problem of IR. We view information needs as Anomalous States of Knowledge and therefore aim to design an **IR** system in which ASKs have explicit structural representations. Our intention, then, is that the system should consist of a structural text-analysis program, a collection of abstracts, a suite of retrieval algorithms and a mechanism whereby user and system interact. The goals of the system can be described as follows;

- (a) To build a representation of the enquirer's state of knowledge in the form of a network of associations between words.
- (b) To examine this structure, and to interpret certain characteristics of the network as potential anomalies in the state of knowledge.
- (c) To search structured document descriptions with a view to resolving the anomalies; i.e. to modify the searcher's network, using components of document structures, so that the identified characteristics disappear.

These system objectives will be achieved through dialogue with the user and the system would operate as shown in Part 1, Figure 3 (p. 69).

We consider that the design study has at least partly justified our initial premises and assumptions in the following ways:

1. It is possible to obtain problem statements from **IR** system users which can be used to derive adequate representations of ASKs;
2. Abstracts appear to be sufficient document surrogates for information representation in such a system.
3. Even though there are problems in our representations of need and information, they appear to have identifiable characteristics which are potentially useful for IR purposes.

We conclude from this, therefore, that an ASK-based IR system is at least feasible. Some aspects of our system have been rather less well studied than others, though, and some significant problems remain to be resolved. These include:

1. retrieval mechanisms
2. refinement of the analytic procedures
3. classification of ASKs
4. the interactive environment

In the next stages of our long-term project, we intend to investigate these problems in detail. The first point we hope to resolve by collecting and analysing a number of problem statements and documents judged relevant to them. In this way we hope to find regular relationships between ASK representations and representations of information which have been judged suitable for resolving those ASKs. These data will also be suitable for further refinement of our analysis, which we also intend to augment by the addition of some simple linguistic analysis.

Classification of ASKs will be investigated through these data, and we intend to investigate the interactive environment, and indeed the proposed system as a system, by constructing a prototype and testing it with both simulated and a small number of natural users.

In general, we think that the Design Study has shown that an ASK-based IR system is feasible, and has indicated some ways in which such a system might be developed. Thus, the Design Study has increased our understanding of how such a system might be implemented.

Equally important, the relationship between our work and mainstream IR research is becoming clearer. The significance of recent advances in IR research lies in the fact that the techniques advocated are *theoretically* motivated: assumptions are made, mathematics done and techniques prescribed. However, because the right type of experimentation has not yet been done, the applicability of the theories to the real world of searching is not known. Our work has made us aware of a wide variety of ASKs, some of which may satisfy the assumptions of the theories, others (probably the majority) do not. The choice before us is either to attempt to develop a single theory (say, probabilistic) which copes reasonably well with all ASKs, or alternatively to construct a decision mechanism capable of determining which theory is applicable to a particular ASK, and then invoking a suitable retrieval strategy. The latter alternative seems to us to be more promising for the present, because it allows us to take proper advantage, in this very complex situation, of both algorithmic modelling and mathematical theories. Thus, we anticipate that future (intelligent?) **IR** systems will not rely on a single retrieval strategy, but will incorporate a number of techniques, invoked and controlled by a dialogue, based on a cognitive model.

The situation at present is that we know how to do document retrieval when certain conditions obtain: i.e. those that indicate a best-match approach. One way to view the introduction of probabilistic theories into IR is as an attempt to optimize for the whole population of queries, a method which is really only appropriate to a minority of them. We now need an understanding of a variety of ASKs, and should develop methods of recognizing and distinguishing them and responding appropriately to each. The work that we have reported here represents a small step in this direction.

### **ACKNOWLEDGEMENTS**

There are many people and institutions we would like to thank for their help in the work presented here. Mrs Alina Vickery very generously allowed us access to the Central Information Services, University of London, for our data collection, and also helped us a great deal with her comments on our work. The British Library Research and Development Department sponsored this research with a Small Grant, and Helen Brooks was supported by the Department of Education and Science during various parts of research and writing by Information Science Advanced Course and Research Studentships. We would especially like to thank the people who consented to take part in our research, the Problem Statement and abstract originators, for their thoughtful responses to our questions. Steve Jamieson gave us substantial practical assistance, and our work has benefited significantly from the many discussions we have had during the course of this project with our colleagues world-wide. In particular, this paper has benefited from presentation in a different form at IRFIS 3. ' ' We hope our colleagues recognize their influence on the final product, whose defects, however, remain our responsibility.



## REFERENCES

1. BELKIN, N. J., ODDY, R. N. and BROOKS, H. M. ASK for information retrieval Part I. *Journal of Documentation*, 38(2), June 1982, 61-71.
2. BELKIN, N. J. and ODDY, R. N. *Design study for an anomalous state of knowledge based information retrieval system*. British Library R & D Report No. 5547. Birmingham Computer Centre, University of Aston, Birmingham, 1979.
3. BELKIN, N. J. *A concept of information for information science*. Ph.D. Thesis, University of London, 1977.
4. KAHN, R. L. and CANNELL, C. F. *The dynamics of interviewing*. New York: John Wiley, 1957.
5. PAYNE, S. L. *The art of asking questions*. Princeton: Princeton University Press, 1951.
6. BROOKS, H. M. *The knowledge structures underlying information needs*. M.Sc. Dissertation, Centre for Information Science, The City University, London, 1978.
7. VAN RIJSBERGEN, c. 1. An algorithm for information storage and retrieval. *Computer Journal*, 14, 1971, 402-12.
8. BROOKS, H. M., ODDY, R. N. and BELKIN, N. J. Representing and classifying anomalous states of knowledge. In: M. MacCafferty and K. Gray, eds. *The analysis of meaning: Informatics 5*. London: Aslib, 1979, 227-38.
9. WERSIG, G. *Information-Kommunikation-Dokumentation*. Pullach bei Munchen: Verlag Documentation, 1971.
10. Kiss, G. An associative thesaurus of English: structural analysis of a large relevance network. In: A. Kennedy and A. Wilkes, eds. *Long-term memory*. New York and London: Academic Press, 1975, 103-21.
11. CHRISTOFIDES, N. *Graph theory: an algorithmic approach*. New York and London: Academic Press, 1975.
12. BELKIN, N. J., KUEHNER, D. and MICHELL, B. G. Representation of texts for information retrieval. In: *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 1980. Philadelphia, ACL, 1980: 147-8.
13. BELKIN, N. J., BROOKS, H. M. and ODDY, R. N. Representation and classification of anomalous states of knowledge and information for use in interactive information retrieval. In: T. Henriksen, ed, *IRFIS 3. Proceedings of the Third International Research Forum in Information Science*. Oslo: Statens bibliotekskole, 1979, 146-83.

## APPENDICES

Note: We reproduce in Appendices A and B the questions asked of the users and authors in our study, respectively. The cover letters, and complete questionnaires, are available from N. Belkin.

### APPENDIX A: *Evaluation questions for problem statements*

#### ASSOCIATION MAP (A.M.)

1. Looking at the general shape of the 'map', do you think that the compactness and degree of inter-relation between the concepts (or lack of it), is an accurate representation of how these concepts were related in your mind at the time of the interview?.....YES/NO
2. Looking at the relations between concepts, are there any which you feel are too strongly associated?.....YES/NO  
Any which you feel are too weakly associated?.....YES/NO  
If too strongly associated please give examples:  
If too weakly associated please give examples:
3. Looking at the concepts themselves, are there any concepts missing from the 'map' which you feel are essential to your problem as it was at the time of the interview? .....YES/NO  
Could you please give examples of these missed-out concepts?
4. Any other comments?

#### ASSOCIATION CLUSTER (A.C.)

1. Looking at each 'box', and remembering that all the concepts within it are closely related, are any concepts grouped together which you feel should not be so strongly related?.....YES/NO  
If too strongly related please give examples:  
If not related strongly enough please give examples:
2. Looking at the association level at which the 'box' appears, do you think that any of the concepts appear at too high a level of association?.....YES/NO  
If too high please give examples:  
If too low please give examples:
3. Looking at the relationships between the 'boxes' are there any which are not closely enough linked?.....YES/NO  
Any which are too closely linked?.....YES/NO  
If so please give examples:  
If too far apart please give examples:
4. Any other comments?

COMPARISON

Having studied both formats, does one of them seem to you to be a better, more successful representation of your problem, as it was at the time of the interview, than the other?.....YES/NO

If you answered YES

- a) Which did you consider better?.....Format A.C./Format A.M.
- b) In what ways do you consider it better?

If you answered NO

- c) Was this because you think that on the whole both are successful representations or that both are not successful representations of your problem, at the time of the interview?.....SUCCESSFUL  
NOT SUCCESSFUL

Any additional comments?

APPENDIX B: *Evaluation questions for abstracts*

1. Looking at the general shape of the 'map' do you think that the compactness and degree of inter-relation between the concepts (or lack of it) is an accurate reflection of how these concepts were related in your mind at the time you wrote the article?.....YES/NO
2. Looking at the relation between the concepts, are there any which you feel are:  
too strongly associated?.....YES/NO  
too weakly associated?.....YES/NO  
If too strongly associated, please give examples:  
If too weakly associated, please give examples:
3. Looking at the concepts themselves, are there any concepts missing from the 'map' which you feel are essential to the article?.....YES/NO  
Could you please give examples of these missed-out concepts?
4. If you answered NO to question 1, do you think that the abstract itself is an accurate representation of your article?.....YES/NO
5. Any other comments?

*(Revised version received 29 January 1982)*