CHAPTER 3

METHODS

3.1 General Experimental Outline

In order to achieve the aims of the design study, we decided upon the following general methodology:

a) Tape recording interviews with users of actual information systems (libraries) at reference points. Asking users to discuss the problem with which they are faced, and only afterwards asking them to present a request to the system.

b) Finding 30 abstracts in information science of 200-300 words each of articles written by authors in the U.K.

c) Adapting the text analysis program from Belkin (1977 a) to the circumstances of this situation.

d) Implementing the analysis algorithm as a computer program.

e) Mailing the structures to authors and users with questionnaire, or interviewing where appropriate.

f) Analyzing the problem statements according to various structural characteristics (such as strength of association, number of nodes, degree of coherence), and classifying them.

g) Developing preliminary retrieval strategies based on the classification and structural features of the representation.

3.2 Problem Statements*

The first stage of our experiment was to try to collect suitable problem statements for analysis and subsequent evaluation. Our basic ideas about what constitutes suitability were that the statements should be:

1. directed toward the users' research or work situations rather than to what specifically they hoped to find;
2. as uninfluenced as possible by knowledge of the operation of IR systems;
3. obtained with minimum intervention; and
4. representative of real information needs, expressed by the person with that need.

Our subjects were drawn from users of Central Information Services - Library Resources Coordinating Committee, in the University of London, which carries out on-line searches for students and staff of all of the colleges and institutes of the University of London. Mrs. A. Vickery,

* The data presented here, and in subsequent sections dealing with problem statement analysis and classification are more fully reported in Brooks, 1978.

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 that 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 (Kahn & Cannell, 1957; Payne, 1951). 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 27 interviews, 23 from CIS users and 4 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 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 5).

Eight written queries were also chosen from the CIS files and analyzed 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 s/he wanted to know.

3.3 Abstracts

The next stage was to collect abstracts for analysis and subsequent evaluation. We chose 31 abstracts of recent publications (journal articles or published conference proceedings) from the field of library and information science, restricting them to U.K. authors. We chose the abstracts from Library and Information Science Abstracts, and from our personal journal files, attempting to include mostly author abstracts and to get a representation of lengths, ranging from 90 tokens to over 2000. Restrictions as to date and author location were for ease of evaluation. (See Appendix F for the full text of each abstract).

3.4 Evaluation Procedures

Upon analysis, both problem statements and abstracts were evaluated by the IR users and authors respectively. The procedures for the two groups were slightly different, so we will consider them separately.

The IR system users were sent a package consisting of:
a cover letter reminding them of the experiment and explaining what they should do with the other materials;
a transcript of the interview;
a copy of each of two different representations of the problem statement;
a questionnaire for each of the two representations, on each of which the premises of the representation were explained;
a third questionnaire for comparing the two representations;
a stamped addressed return envelope; and
a cheque for £2.00.

See Appendix A for the cover letter and questionnaires. The entire package was stapled together in a set order: cover letter; transcript; first representation, followed by its questionnaire; second representation followed by its questionnaire; comparison questionnaire. The order of representations was reversed for alternate subjects to guard against possible interactive effects of one format on the other.

The evaluation package for the authors of analyzed abstracts was rather simpler. This evaluation was done after results for the interview evaluations had been obtained, and on their basis we decided not to use the association cluster representation. This package thus consisted of:
a cover letter, explaining the project;
a copy of the abstract;
a copy of the representation;
a questionnaire including an explanation of the representation;
a stamped, addressed return envelope;
a cheque for £5.00.
See Appendix B for the cover letter and questionnaire for authors.

3.5 Text Analysis - Principles

In our work, we have assumed that a state of knowledge - anomalous in the case of a problem statement - can be represented by a network of weighted associations between words, standing for concepts. We now turn to the method used in our exploratory project to derive associative structures from the texts obtained by interviewing and from published documents. Figure 3 shows the text of a problem statement (one of the shorter ones), and a graphical representation of the derived structure is displayed in Fig. 7. Summaries of all of the problem statements are in Appendix C, and their representations are displayed in Appendices D and E.

The technique that we have used to work out 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 upon their lexical separation. There is no syntactic analysis, and no attempt to represent the nature of an association. Our view is that it should be possible to investigate the ASK hypothesis through the construction and evaluation of a new retrieval system, even with a statistical, rather than linguistic text analysis mechanism. The assumptions behind the measure of associative strength used are that if two words are closely associated in the searcher's knowledge structure, then, firstly, they will be close in the textual expression of that structure, and secondly, the association will tend to find multiple expression in the text. The actual formula for the measure is given below, and is similar to that used by Belkin (1977 a) in a previous, related experiment.

An important step in the analysis of a text is to identify the word-types which are to stand for concepts in the structure. This involves eliminating words which are non-significant for our purposes, and 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. We have made no attempt to enable our analysis programs to recognize synonyms. As an example, consider the following sentence from a (tape-recorded) problem statement:

> "There is a problem in agriculture that if the grains germinate, then there could be a substantial loss in yield, and this I would presume has been looked at in great detail."

The letter-groups underlined constitute the tokens selected from this sentence for inclusion in the association network. It is usual, in automatic indexing programs, to use a stop-list to eliminate words which are regarded as poor indexing terms. We have tended to include words, such as "how", "why", "not", "effect", "presume", which have importance in the semantic structure of the statement, and which may therefore partake in important associations. The words which are not underlined in the example above are a sample of the words we regarded as non-significant in the problem statement corpus.

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.

FIGURE 3: THE TRANSCRIPT OF A PROBLEM STATEMENT

Our immediate purpose in analysing the texts was to perform a check on the appropriateness of our method by asking the interviewees (in the case of problem statements) and authors (as in the case of abstracts) to comment upon the structures produced. Thus it was necessary to devise ways of displaying the associativ e structure. The straight-forward method is to draw a network (see figure 7), in which the nodes, labelled with "stems", represent concepts and the lines represent associations. For clarity, it is necessary to omit the large mass of weak associations (the number of associations varies as the square of the number types in a text), and to simplify the representation of association strength. The associations derived from an individual text were divided, subjectively, into three groups: strong, medium and weak. In the network diagrams, the corresponding lines were drawn, thick, thin and broken. 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.

Another method of displaying the associative structure of a text is to cluster the word-types, so that a person examining the structure can see which concepts are closely linked, without being confused by the precise pattern of interconnections. We have generated a hierarchy of single-link clusters of types for each text using an algorithm which was originally developed for document clustering by van Rijsbergen (1971) and Croft and van Rijsbergen (unpublished). Figure 14 shows such a hierarchy in tree form. 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. See Appendix E for this representation for all of the problem statements.

3.6 Text Analysis - procedure

(i) Data preparation. Tape recorded problem statements were first transcribed by a typist. These and the written problem statements and abstracts were entered into computer files, care being taken to separate sentences with full-stops, and to insert a special symbol at the beginning of each paragraph. Each abstract begins with the document's title, marked as a separate paragraph.

(ii) Dictionary construction. In order to conflate word forms, and at the same time eliminate non-significant words, each collection of texts (problem statements and abstracts) was scanned to compile a dictionary of stems (truncated words). This was done semi-automatically: a suffix stripping algorithm was used, followed by manual editing. Non-significant words were removed from this list.

(iii) Identification of tokens. The stem dictionary was searched for each word in the text, looking for the longest stem which matched the word. For example, the dictionary contained "fact" and "factor". The word "factorize" is matched by both stems, so "factor" being the longer is used as the token in place of "factorize". If there is no match in the dictionary, the word is taken to be non-significant. For each text, a list of stems is constructed giving the position of each occurrence as a numerical triple: (paragraph number, sentence number, word number). Occurrences of rejected words are not included in the list, although their presence in the original text is taken into account when working out the positions of the significant words. For example, the first sentence in the problem statement of figure 3 will be processed as shown in Fig. 4.

(iv) Calculation of association strengths. The strength of association between word-types in a text can be computed from the list of occurrences, as follows:

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

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

where r = 1 if A and B are adjacent within the same sentence.

r = 2 if A and B are within the same sentence, but not adjacent

r = 3 if A and B are in adjacent sentences within the same paragraph.

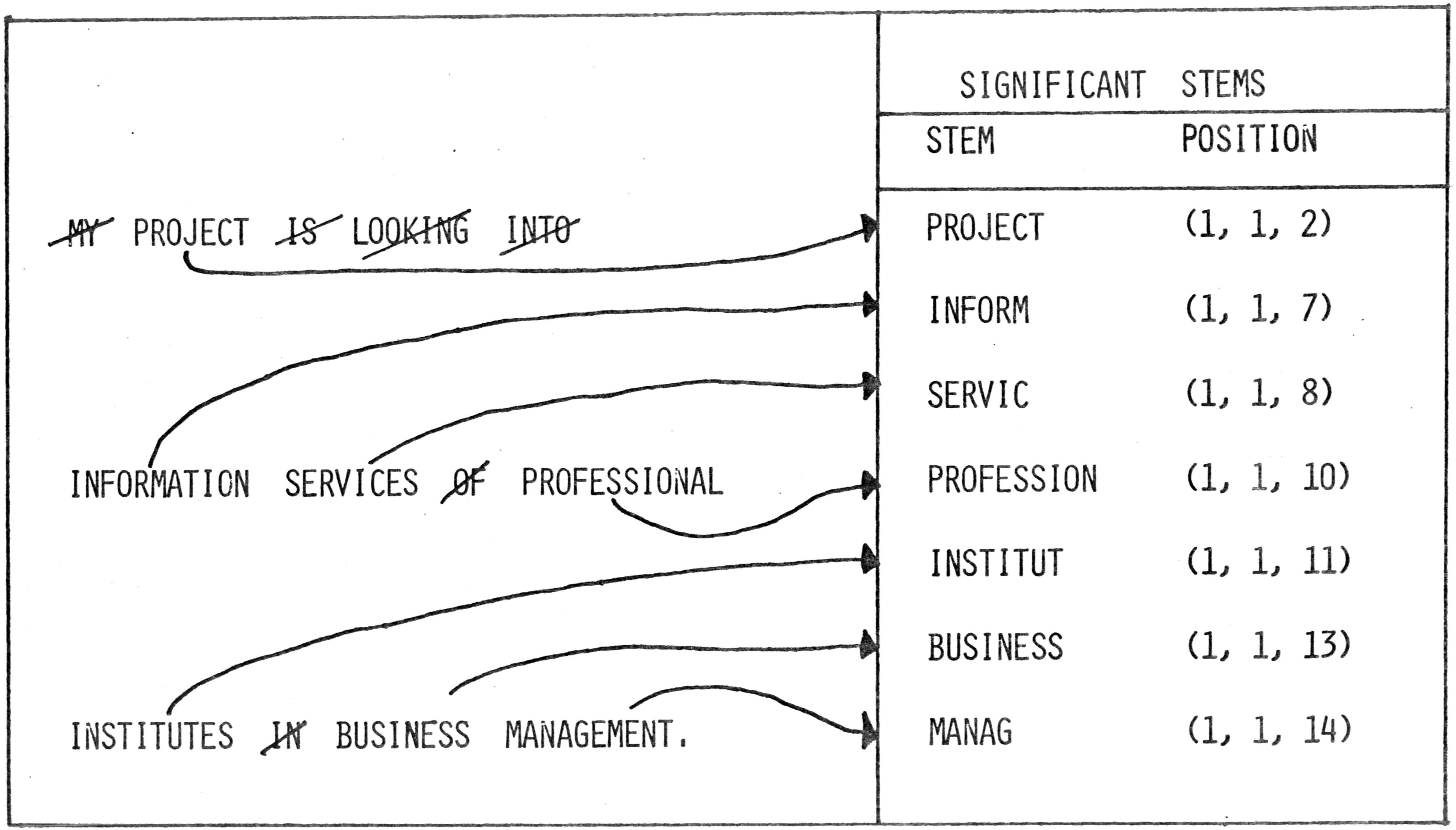
If the occurrences of A and B are any further apart, the score is zero. (The factor of 100 is present purely for computational convenience).

Having calculated the scores for all pairs of occurrences of A and B in the text, they are summed to give the strength of association.

As an example, consider the stems BUSINESS and INFORM in the text in Fig.3. Their occurrences and association scores are tabulated in Fig.5.

The word associations for each text were listed by the computer in decreasing order of strength (e.g. see Fig. 6).

(v) Presentation. Graphical representations of the associative structures were prepared for examination by the originators of the



~~MY~~ PROJECT ~~IS~~ ~~LOOKING~~ ~~INTO~~

INFORMATION SERVICES ~~OF~~ PROFESSIONAL

INSTITUTES ~~IN~~ BUSINESS MANAGEMENT.

SIGNIFICANT STEMS	
STEM	POSITION
PROJECT	(1, 1, 2)
INFORM	(1, 1, 7)
SERVIC	(1, 1, 8)
PROFESSION	(1, 1, 10)
INSTITUT	(1, 1, 11)
BUSINESS	(1, 1, 13)
MANAG	(1, 1, 14)

FIGURE 4 : TOKEN PROCESSING

STEMS "BUSINESS" AND "INFORM" FROM FIGURE 9

INFORM BUSINESS	(1, 1, 7)	(1, 2, 2)	(1, 6, 25)
(1, 1, 13)	33	25	0
(1, 6, 24)	0	0	50

TABLE OF SCORES FOR DISTANCES BETWEEN TOKENS

ASSOCIATION STRENGTH: BUSINESS - INFORM
IS 33 + 25 + 50 = 108

FIGURE 5 : CALCULATION OF ASSOCIATION STRENGTHS

INFORM SERVIC	183
INSTITUT LIBRAR	149
EVALU INSTITUT	133
INSTITUT PROFESSION	125
INFORM INSTITUT	116
INSTITUT TYP	116
INSTITUT SERVIC	116
BUSINESS INFORM	108
EVALU TYP	100
BUSINESS INSTITUT	91

FIGURE 6: THE STRONGEST ASSOCIATIONS FROM THE PROBLEM STATEMENT IN FIGURE 3.

texts. Fig. 7 illustrates the network presentation (Association Map Format), and Fig. 8 shows the corresponding hierarchical clustering of the word-types (Association Clusters Format).

In our survey of problem statements, we sent both formats to the subjects. The 40 strongest associations were used for the networks. The authors of the abstracts, however, only examined the structures in Map Format, using about the top 40 associations but with some flexibility in order to avoid splitting groups of associates with the same association strength. Figure 9 shows an example of an abstract, Figure 10 is its association map representation.

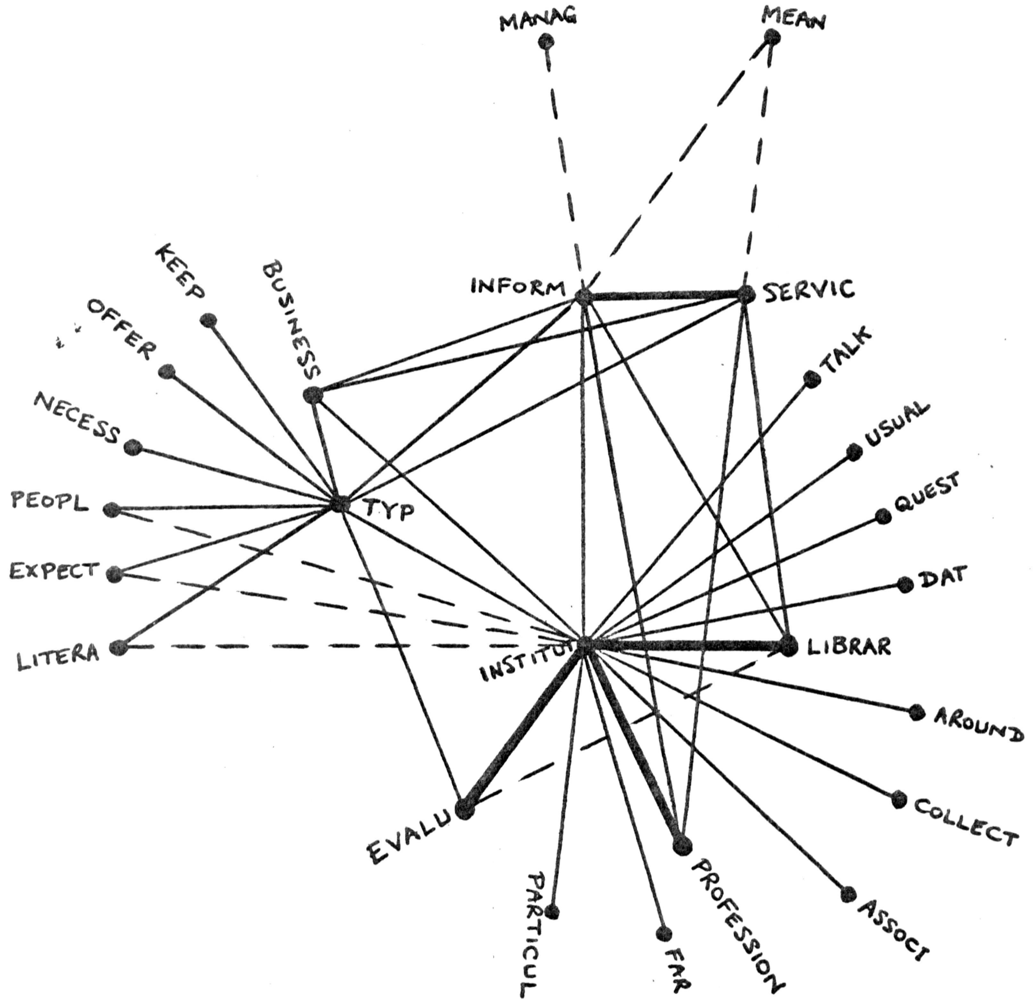


FIGURE 7: ASSOCIATION MAP FOR PROBLEM STATEMENT IN FIGURE 9

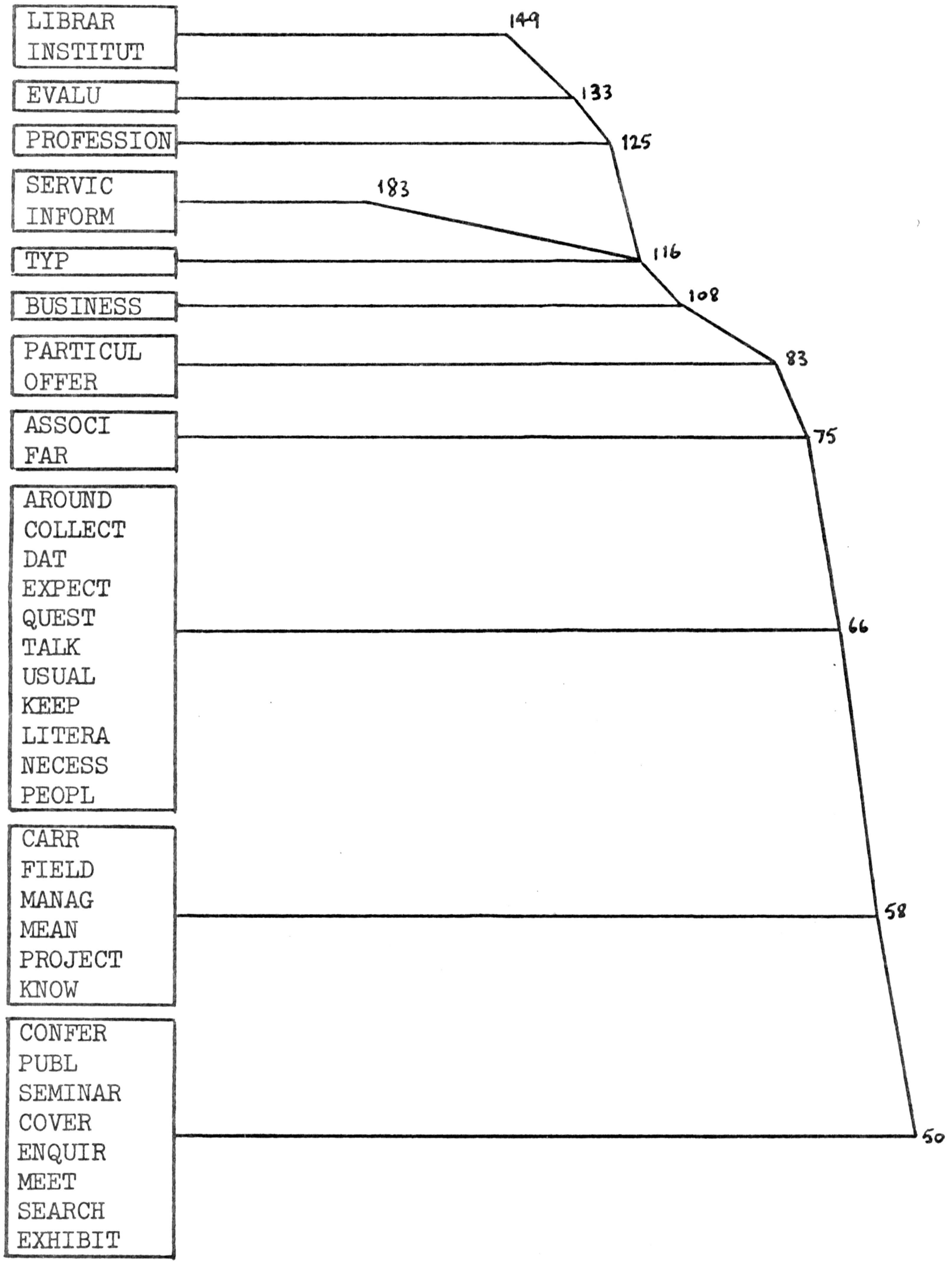


FIGURE 8 : ASSOCIATION CLUSTERS FOR PROBLEM STATEMENT IN FIGURE 9

Information Science and the Phenomenon of Information

This paper aims to deduce the fundamental phenomena of information science, starting from two premises: that information science is a problem-oriented discipline concerned with the effective transfer of desired information from human generator to human user, and that the single notion common to all concepts of information now extant is that of *change of structure*.

From these premises, a spectrum of information concepts is derived, and a partition of that spectrum particular to the purposes of information science is described. From this partition, the terms *text* and *information* (both in information science) are defined, and the fundamental phenomena of information science are deduced: the text and its structure, the structure of the recipient and changes in that structure, and the structure of the sender and the structuring of the text.

These phenomena are seen as the basic components of the mechanisms of the *channel*, which have been the traditional area of interest to information science. Some implications of this approach for research in information science are discussed in this paper.

And, finally, the question of the ethics of theoretical research in information science is raised, and a restrictive condition is proposed.

FIGURE 9: AN EXAMPLE ABSTRACT

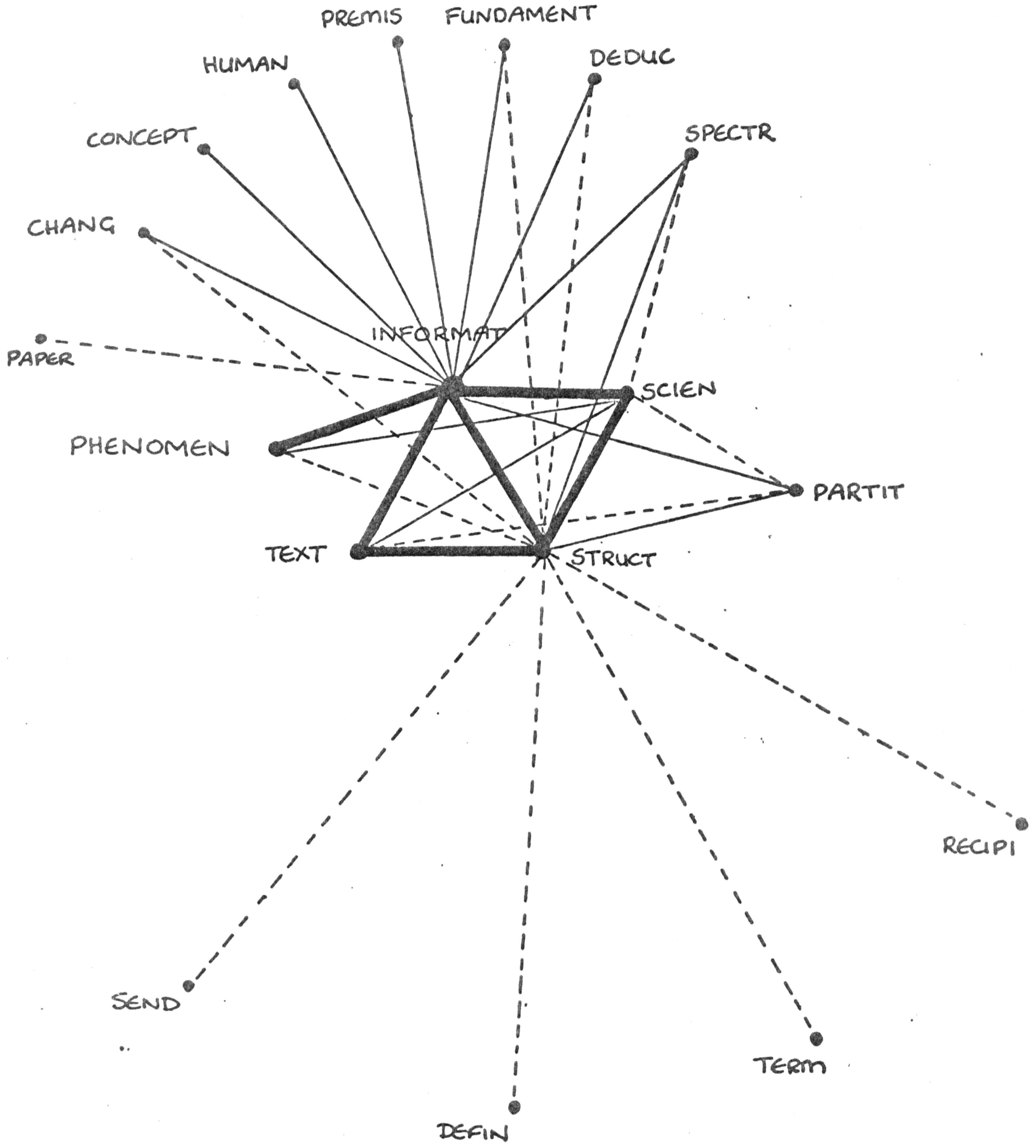


FIGURE 10: ASSOCIATION MAP FOR ABSTRACT IN FIGURE 9.