CHAPTER 5

DATA ANALYSIS

5.1 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 peoples' problematic situations (Wersig, 1971) which can be algorithmically generated. This classification may or may not be useful for determining how to resolve anomalies. If we assume that

(i) The representations produced by the text analysis procedure are closely related to ASKs, and

(ii) 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 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 significant (Kiss, 1975). 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 40), we can use the following formula without normalisation:

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

where N_a is the number of nodes present in the network and N_{min} is the minimum number of nodes possible, given that there are 40 lines. (In fact, $N_{min} = 10$). For example, the number of nodes in the network of figure 7 is 25, so its connectivity score is 15. Our sample of problem statements was small, so the scores were pooled to produce 5 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, 1975) 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 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. In Fig. 7, there are three stars, at the nodes labelled INSTITUT, TYP and INFORM.

A class code of the form: CONNECTIVITY CLASS; NO OF STARS was assigned to each problem statement structure (e.g. the structure in Fig. 7 has the code C3). A summary of the ASK types in the sample of 27 interviews, as defined by this classification is given in Table 12. 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.

5.2 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. 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 appro-

TABLE 12: SUMMARY OF ASK TYPES (AFTER BROOKS, 1978, PP80-2)

Class	Group	No. of interviews	Comments
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	Problem involves relating two, or more, specific topics
C Topics quite specific.	1	3	Fairly general bibliographies wanted.
Problem not so well defined. Research still at an early stage.	3	2	Research clearly described. Less sure as to what information required.
D Topics fairly specific. Problems not well defined.	1	2	Problem involved trying to extrapolate a set of conditions from a known to an unknown situation.
No hypotheses underlying research.	2	2	No obvious similarities between the members of this group.
Information needed to produce directions for research.	3	2	Problem not clearly defined. Information wanted which will provide ideas for the formulation of hypotheses.
E Topics and problems not well-defined.	1	3	2 relate a particular subject to a number of variables. 2 doing literature search for project not yet started.
Topics often unfamiliar.	2	2	References not wanted for research but to write paper. Paper deals with known subject area but in an unfamiliar context.
	3	1	Very fragmented statement. Neither problem nor research clearly specified.

priateness 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 40 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 = Fig. 7:

(i) Strong clusters are defined as components singly-linked at the strong level, and are denoted by symbols within two concentric circles. Figure 11 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 12 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 A^m.

(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 a part of most retrieval strategies to select documents containing A^m 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

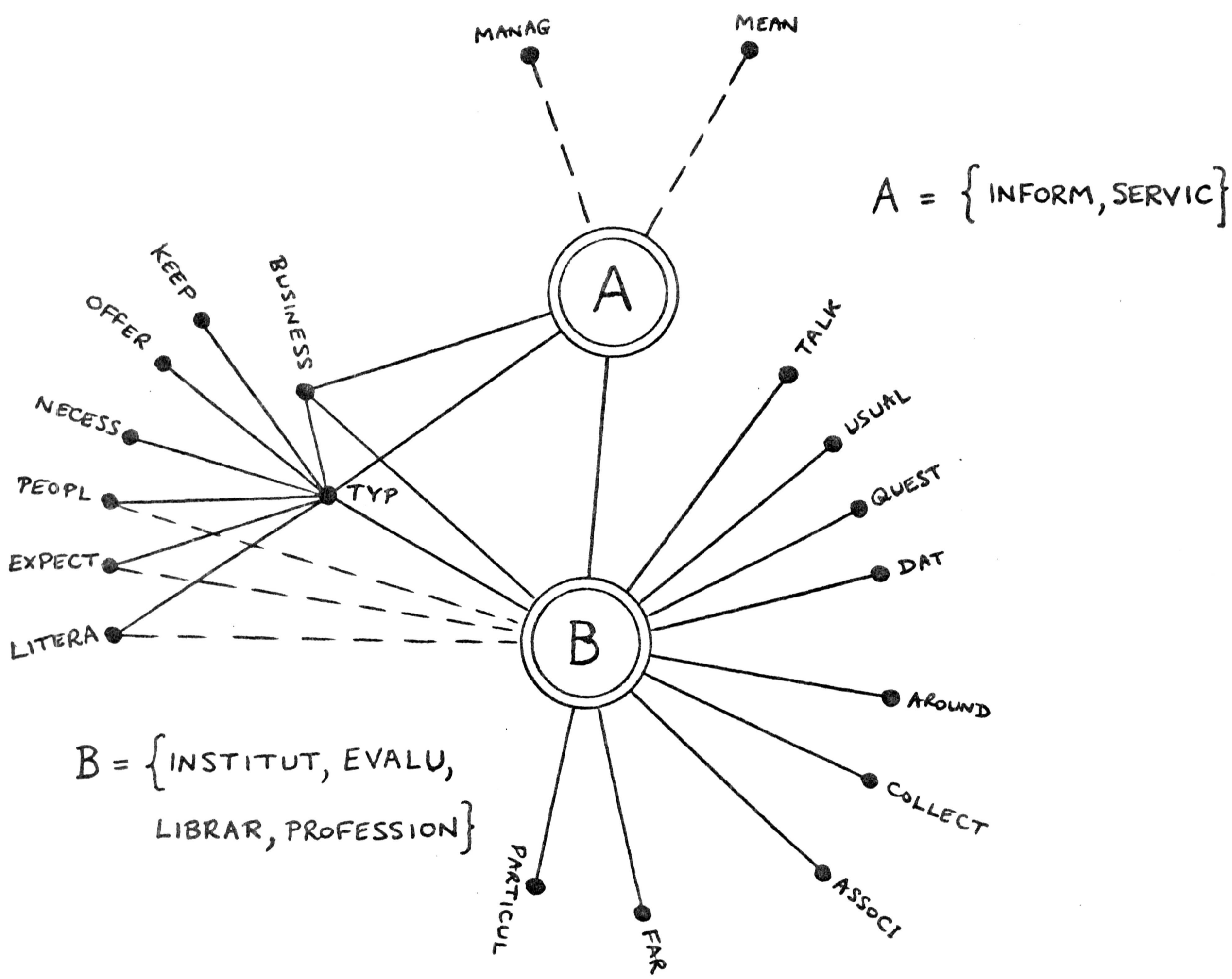


FIGURE 11: CONDENSED PROBLEM STATEMENT NETWORK - 1

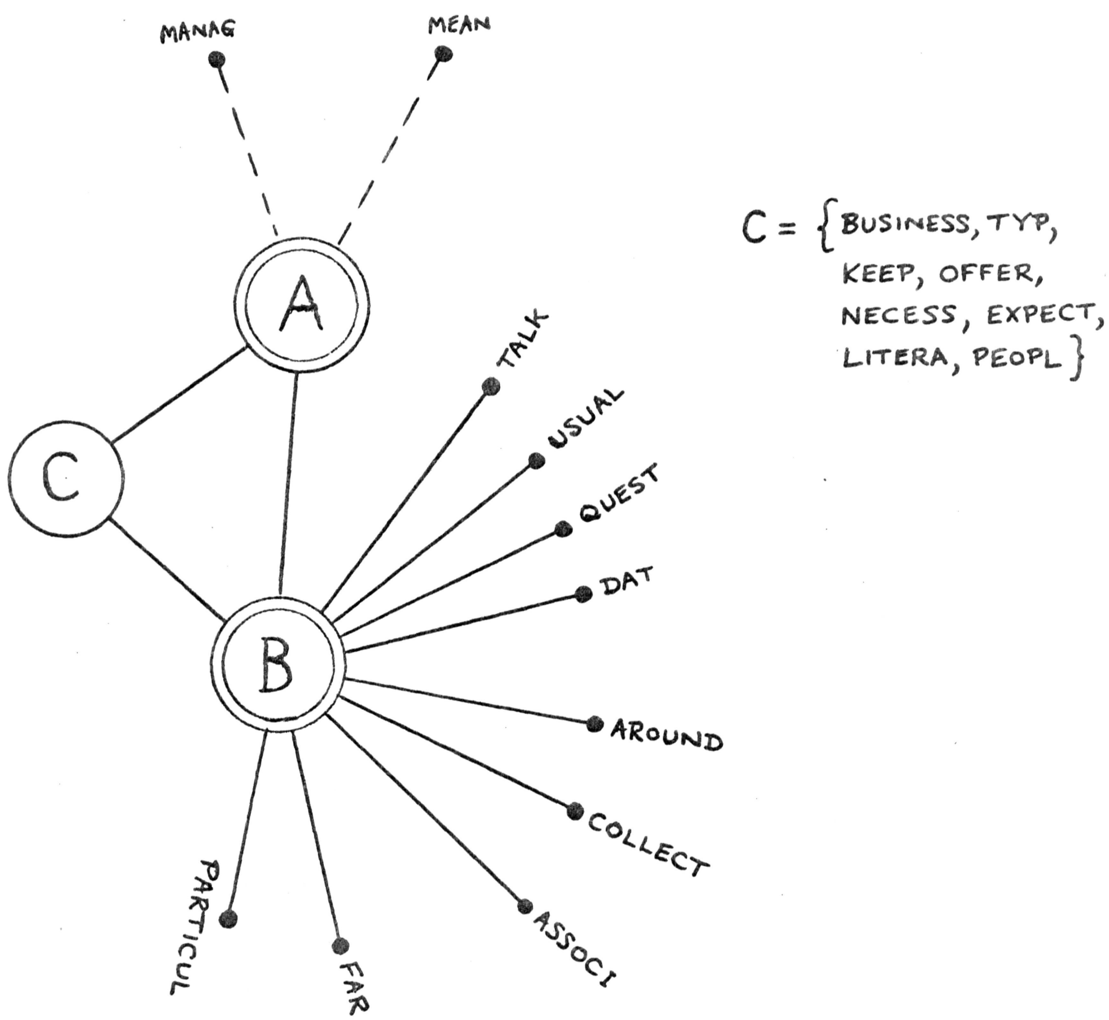


FIGURE 12: CONDENSED PROBLEM STATEMENT NETWORK - 2

associations at a weaker level than those within the clusters our system might reasonably assume that stronger associations between the clusters would 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:

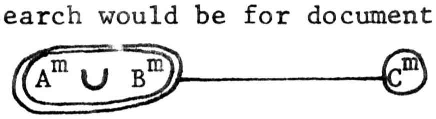
and

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

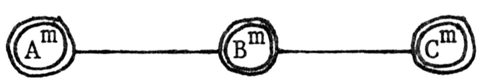
(iii) Medium Links

Nodes, and medium clusters which are connected by medium links to a strong cluster may specify the <u>context</u> 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:

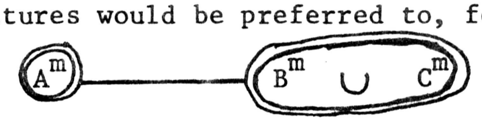
the first search would be for documents containing:



or:

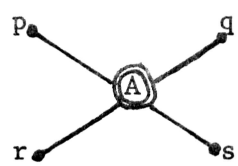


These structures would be preferred to, for example:

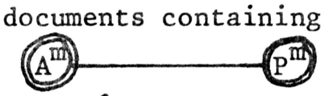


(iv) Stars

The essential property of a star, for our purposes, 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 (Christofides, 1975) i.e. sets of mutually disconnected nodes - and seek for documents which contain matching sets as clusters.