9-10-1986

Using Structural Represantation of Anomalous States of Knowledge for Choosing Document Retrieval Strartegies.

Barbara H. Kwasnik
Syracuse University

Belkin N. J.
Syracuse University

USING STRUCTURAL REPRESENTATIONS OF ANOMALOUS STATES OF KNOWLEDGE FOR CHOOSING DOCUMENT RETRIEVAL STRATEGIES*

N.J. Belkin & B.H. Kwaśnik
School of Communication, Information and Library Studies
Rutgers University
New Brunswick, N.J., U.S.A.

ABSTRACT

We report on a project which attempts to classify representations of the anomalous states of knowledge (ASKs) of users of document retrieval systems on the basis of structural characteristics of the representations, and which specifies different retrieval strategies and ranking mechanisms for each ASK class. The classification and retrieval strategy specification is based on 53 real problem statements, 35 of which have a total of 250 evaluated documents. Four facets of the ASK structures have been tentatively identified, whose combinations determine the method and order of application of five basic ranking strategies. This work is still in progress, so results presented here are incomplete.

1. Introduction

It has been suggested for some time in the IR literature that different types of user situations, problems, goals, characteristics or questions might require different types of retrieval strategies, mechanisms, or ranking rules [e.g. BELK80; CROF84; ODDY77]. All such suggestions must address two major questions: how can different user situations be distinguished from one another? and, what kinds of retrieval strategies are appropriate to the different situations? To date, these remain open questions.

One previous study [BELK82] had suggested that structured representations of IR system users' anomalous states of knowledge (ASKs) might be used as the basis for choosing different document retrieval strategies. In [BELK82] and [HAPE85], some potential categorizations of ASKs and retrieval strategies were discussed; here we report on the preliminary results of an empirical classification, based on representations of 53 ASKs and about 250 documents which were evaluated by users in respect of those ASKs.

2. Methods

2.1 Data collection

Our data consists of narrative problem statements gathered from users of operational online document retrieval services, and of evaluations by those users of the usefulness of up to 15 documents in the resolution or management of their problem. Our methods for eliciting problem statements and evaluations are described in detail in HAPE85. Briefly, we collected our data from users of two academic information retrieval services of the University of London as they entered the service, but before they had spoken with the intermediary. The subjects were given a printed problem statement elicitation (figure 1) also posed to them orally. The oral elicitation and the user's narrative problem statement response were tape recorded. For one-half of the subjects, this tape recorded problem statement was then given to the intermediary, and used as the sole basis of the online search (non-interactive mode). In this case, the intermediary conducted the search alone. For the others, the problem statement was used as the basis for subsequent presearch interaction between the user and the intermediary (interactive mode). In this case, the user was with the intermediary throughout the search. In interactive mode searches, a check was made at the end of the interaction as to whether the original problem statement was still perceived valid by the subject.

1. Please give a clear indication of the research that you are doing at the moment. What is the nature of the research, its present stage of development and the research goals which you consider to be the most relevant to your information enquiry?
2. What is the information problem that has prompted you to have an online search carried out? Your answer should be a concise description of what it is you need to find out, rather than just a list of keywords.
3. What kinds of information would you like to receive as a result of the online search? For example: document type, the time period involved, the level of treatment, the breadth of coverage, language or languages, etc.

Figure 1. Problem statement elicitation

*Research supported by British Library Research and Development Department Grant SI/G/566 to the Department of Information Science, The City University, London

The results of the searches were sent to the subjects together with an evaluation questionnaire and transcript of their problem statement. The subjects were asked to evaluate up to 15 documents which they had read, according to their degree of usefulness with respect to the problem statement, and to comment on why they made each particular usefulness judgment. The transcripts of the problem statements, and the texts of the evaluated abstract documents were the basic data for input to the structural analysis program.

2.2 Text analysis

The problem statements were transcribed from the audio tapes according to a set of transcription rules developed for a series of discourse analysis projects at The City University [BROO83] [DANI85]. The transcript retains indications of pauses, false starts and other discourse phenomena, and represents words more-or-less as they were spoken. For the ASK representation programs, the raw transcripts were normalized to standard English narrative, primarily by removing indications of non-linguistic discourse phenomena and obvious re-start repetitions, and by using standard spellings. Sentence boundaries were also inserted according to rules based on length of pauses and discourse intonation.

The text analysis programs are described in detail in HAPE85. Their aim is to achieve graphical representations of both problem statements and abstracts, in which the nodes are concepts (represented by word stems), and the arcs indicate levels of association strength between nodes, with the distance between nodes also being an indication of their strength of association. The algorithm first applies a stop-list to the text, then a stemming procedure [PORT80], and then computes cumulative association strength for word pairs on the following conditions:

<u>WORD-PAIR POSITION</u>	<u>SCORE</u>
ADJACENT	12
SAME SENTENCE	4
ADJACENT SENTENCES	3

ASSOCIATION STRENGTH = SCORE

This association strength is treated as an inverse distance measure in a program written by John Bovey, which computes a stable two-dimensional network for the top 40 (or so) associates, according to the requirements for the graphs specified above. At the representation level, the association strengths are converted to four levels of strength, determined by the percentage contribution they make to the total association strength. Figure 2 is an example problem statement text, and figure 3 the corresponding graphical ASK representation. This general algorithm and representation was tested for adequacy by BELK82, and modified to its present configuration according to results from WEST83. Further work on its psychological validity is underway at Syracuse University [PALM84].

The second topic is related to bleeding in early pregnancy and its effect on the outcome of that pregnancy. There are many studies actually carried out in Britain and in other countries on the effects of bleeding in early pregnancy on both the mothers and the foetuses. And little valid information has been obtained for these many studies, simply because ultrasound has not been used as a method of investigating the site of placenta, so what we did, actually we did a sort of case control study of mothers with bleeding in early pregnancy compared with normal mothers, that's to say with no bleeding in early pregnancy. And we followed them during the whole period of pregnancy and we did subsequent type of ultrasound to both cases and controls and we compared between the outcomes of the two groups.

I just want or would like to see - I mean this is answering question number 1 and answering question 2 - I would like to see other studies or similar studies elsewhere. As far as I know there are two studies, which I was able to take from Index Medicus, and I would like to see some more studies, if there is any possibility and comparing their approach. It's similar to the problem number 1. Yes, I want a document type on the printout for the ones which I can not get any access to - journals or books.

Figure 2. Text of the problem statement of s.14.

2.3 Characterizing ASK structures

BELK82 suggested that purely structural features of ASK representations could be used to classify the ASKs into groups which would each determine some specific, different retrieval strategy, or matching formula. These features were unspecified, however. We have developed a scheme for characterizing the ASK representations on the dimensions indicated in figure 4, which seem reasonable candidates for appropriate features.

<u>GROUPS PRESENT IN STRUCTURE</u>
CLUSTERS (BY TYPE, MAGNITUDE & CONNECTIVITY)
STARS (BY TYPE, MAGNITUDE & DEGREE)
LINES (BY TYPE, MAGNITUDE & DEGREE)

<u>RELATIONS AMONG GROUPS</u>
PATH LENGTH, DISTANCE AND CONNECTION

<u>OVERALL CONNECTIVITY OF THE STRUCTURE</u>

Figure 4. Dimensions for the characterization of <u>ASK structures</u>.

The definitions of all of the terms and characteristics used in our scheme are listed in the <u>Appendix</u>. Our method was to go through all of our ASK representations, and to characterize and classify them by this scheme. This gave us some way to describe the representations in purely structural terms.

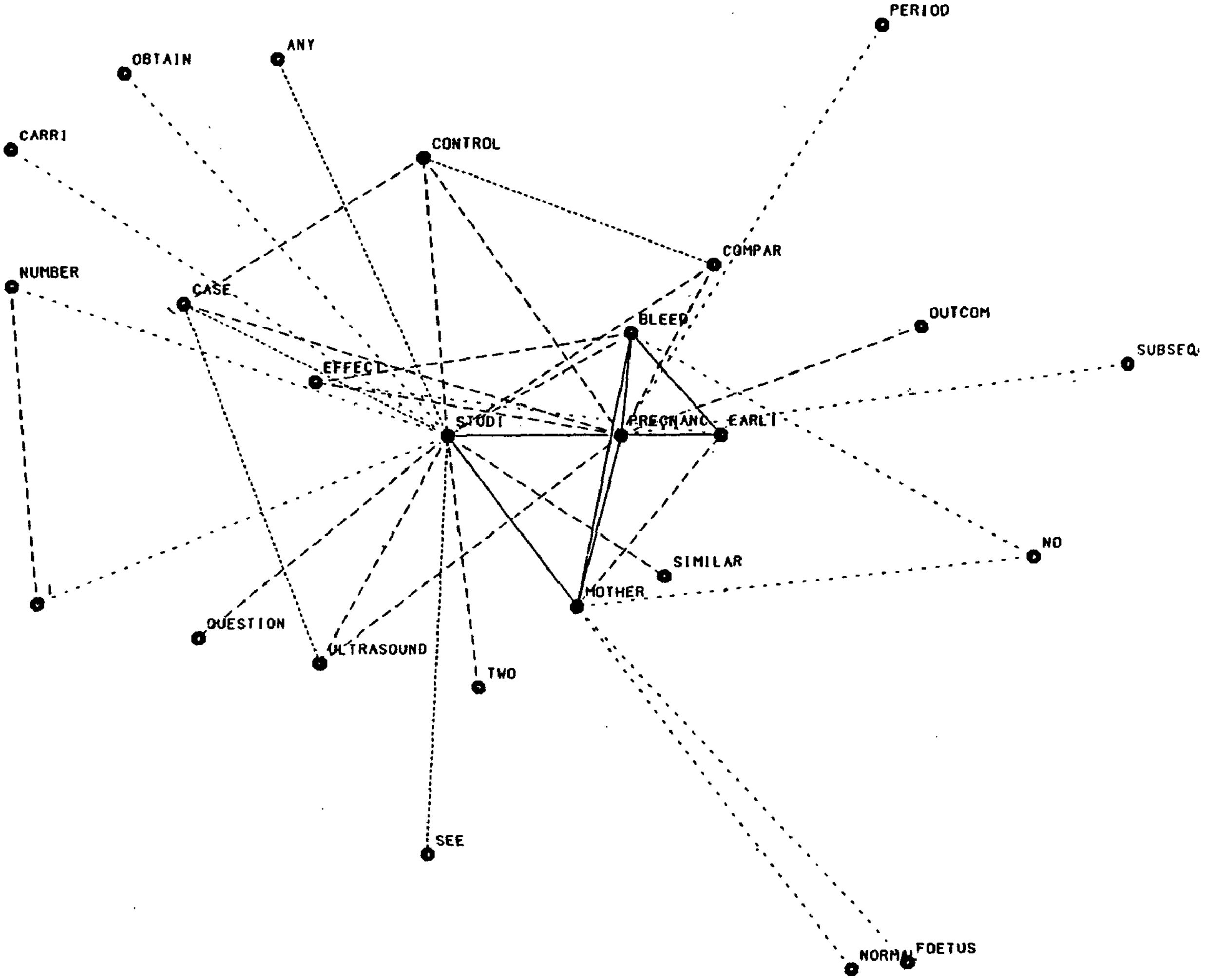

Figure 3. ASK representation for s.14, derived from the problem statement of Figure 2.

2.4 ASK - Text relations

To to discover groupings of ASKs which lead to choice of retrieval strategy, we considered the relationships between ASK structures and the structures of texts which were evaluated in respect of those structures. Since we had no *a priori* schema, this part of the study consisted of a highly exploratory and informal data analysis, based on the usefulness evaluations and comments of the subjects, and on visual inspection of the structures representing texts and ASKs. This aspect of the study resulted in a specification of a retrieval strategy for each of the ASKs, which would have resulted in ranking the evaluated documents in the order of their usefulness (or in not retrieving the not useful documents).

In this portion of the data analysis it became evident that some lexical information would be required, in addition to structural, in order to choose appropriate retrieval strategies. For example, terms such as 'RESEARCH', 'WANT', 'FIND' and 'PROBLEM' usually indicated areas of the ASK structure which were substantive to the topic of search, whereas terms such as 'LITERATURE', 'TODAY' and 'SEARCH' were associated with areas of the graph concerned with output characteristics. This led us to develop several closed vocabulary sets for identifying areas of the ASK graphs which could be used for different aspects of retrieval strategy formulation.

Thus, the candidate strategies that we developed for each ASK depended on identifying particular areas and substructures of the ASK graph which would allow identification of particular structures of specific lexical items in the representations of potentially useful texts, and provide some means of ranking. These areas and substructures were found, at this stage of analysis, by quasi-algorithmic techniques, which were associated in each case with the general structural characteristics of the ASK representation already assigned.

Figures 5 - 9 are representations of abstracts of documents which were judged, respectively, very useful, quite useful, marginally useful and not useful to the ASK represented in figure 3. As an example of our method for arriving at our eventual strategies, and of how the ASK structures were characterized, we reproduce the reasoning we used in this case.

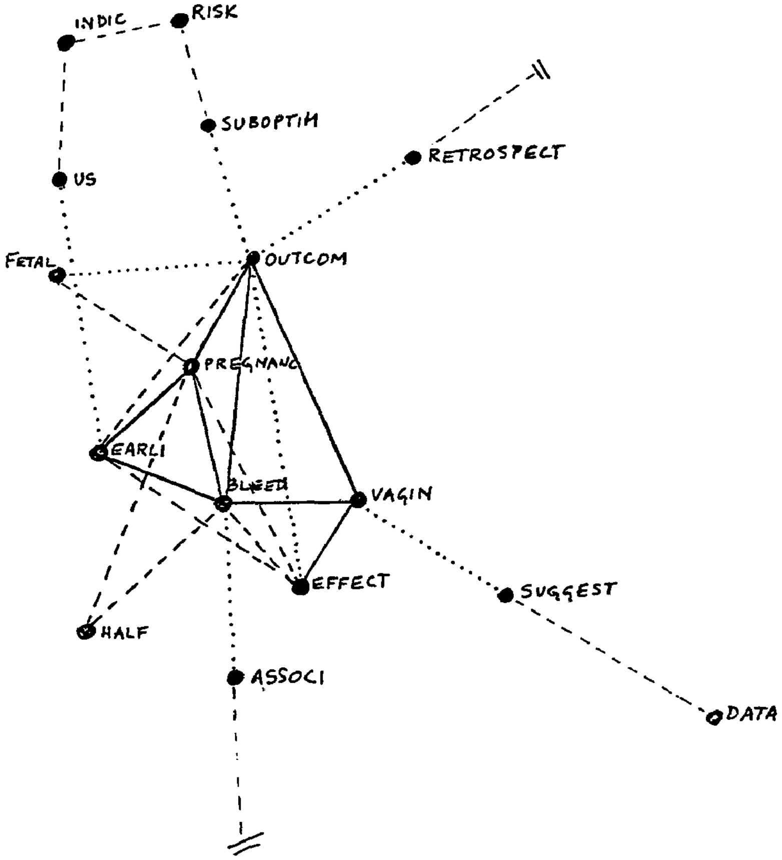

Figure 5. Representation of document 14.01 (judged very useful).

The ASK structure for S.14 is characterized as indicated in figure 10. *From the structures of* the five evaluated documents for this subject, it is evident that the basic strategy must be to look for documents which center on the level 1 nodes in the type 1 cluster, but that this strategy alone, as simple matching, would not account for the particular ranking given these texts. For instance, it appeared that some concepts, such as 'OUTCOM', which were not in the type 1 cluster, were significant. Also, as can be seen from figure 9, the location and associative structure of matched terms in the text representation is as important as the matching itself. We notice, for instance, that the Type 1 cluster of the ASK has several triadic substructures at level 1, all based on the highest degree node in that cluster, 'PREGNANC', and that these characteristics appeared to bear on the usefulness judgements of the texts.

Thus, for this ASK structure, we hypothesize that the highest degree node at level 1 in the ASK, which we take to be some indication of 'centrality', should also be fairly central in the text representation (relatively high degree at levels 1 and 2). Furthermore, text structures which exhibit the same triadic structure as the ASK structure should be ranked higher than those which do not. In conjunction with the latter hypothesis, the triads can be rank ordered according to the sum of their sides. Therefore, preference will be given to a text with the triad 'PREGNANC - BLEED - EARLI' over one with the triad 'PREGNANC - STUDI - MOTHER'. That is, the smaller the circumference of a matching triad in a text, the higher the weight for that text. A further criterion for usefulness appears to be incorporation into the center of the text structure of peripheral nodes from the star based on the most involved type 1 node (in s.14, these are 'OUTCOM', 'SUBSEQU' and 'PERIOD', radiating from 'PREGNANC'). This ranking rule, on the basis of the structures and evaluations, is somewhat weaker than the others. And as the weakest criterion, incorporation of level 2 nodes of the Type 1 cluster into the central cluster of the text structure (i.e. 'ULTRASOUND', 'COMPAR', 'CONTROL' and 'EFFECT') seems reasonable.

Thus, one possible retrieval strategy and ranking mechanism based on these hypotheses for this ASK structure type is:

1. *Quorum search on the set of terms* s = {type 1 cluster nodes; peripheral nodes of the highest degree level 1 star} —<u>must</u> contain at least highest degree level 1 node and one other from type 1 cluster.
2. From retrieved set, eliminate any in which highest degree type 1 problem statement node is not at level 1. For remainder, rank according to relative degree 2 of highest degree type 1 node, all documents with equal first or better, ranked 1.

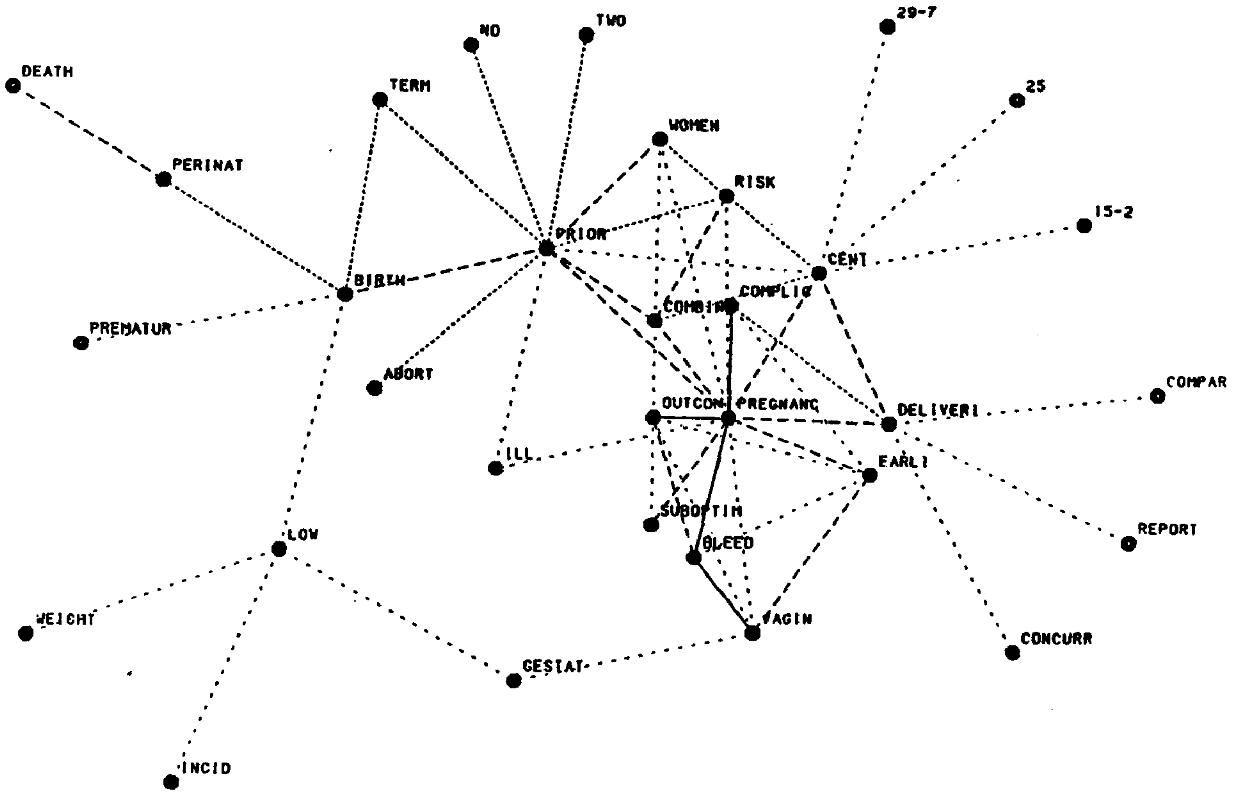

Figure 6. Representation of document 14.05 (judged very useful).

3. Rank within groups determined in step 2 according to triad matching, as follows:

T1 + T2	Both complete at level 1
T1 + T2	One complete, one partial
T1	Complete at level 1
T1 + T2	Both partial at level 1
T1	Partial at level 1
T2	Partial at level 1
T1 or T2	Match at < level 1

where T1 is the smallest circumference triad, T2 the second.

4. Rank within groups determined in step 3 according to star integration, by number and degree of star nodes.

This strategy would, in our example, step-by-step: 1. retrieve all 5 documents; 2. eliminate 14.06, group 14.04, 14.05 and 14.01 in the first rank, and rank 14.03 after all three; 3. rank 14.01 first, 14.05 second and 14.04 third, with 14.03 still fourth; 4. increase 14.01's ranking overall, 14.05's ranking relative to 14.04 and 14.04's relative to 14.03.

Our general method was to go through each ASK-texts set in this manner, using the results gained with each analysis to guide subsequent ones. We followed up by reanalyzing the entire set of data, in order to make use of the later results on those sets analyzed first. This resulted in a number of specific strategies associated with specific ASK structures.

We then grouped the strategies according to their general characteristics, such as method for choice of terms for initial matching, method for choosing structures for matching, and discrimination or ranking methods. The final step in the study was to identify common characteristics among the ASK structures associated with the groups of retrieval strategies. These last two stages were interactive and iterative.

3. RESULTS

3.1 Data and response rate

We elicited 53 usable problem statements with topics ranging from education and psychology to chemistry and medicine, and users from beginning masters degree students to completing Ph.D. students to M.D.s to professors and independent researchers. Of this group, 40 returned questionnaires, 5 of which had no evaluated documents, or were otherwise unusable. Thus, our problem statement corpus for general categorization is 53, but that for comparison of ASK and text structures is 35. For these 35 problem statements, 298 documents were evaluated, ranging from 2 to 15 per problem statement. We were unable to find abstracts for some of these documents, which brought the final number of documents used for strategy generation to about 250.

3.2 Classes of retrieved strategies

The retrieval strategies for each problem statement were quite complex, as can be seen from the example of s.14. However, they all followed a general two-stage pattern. First, a set of word stems in the ASK structure would be identified, on the basis of structural and lexical features of the ASK, which would be used to retrieve a set of documents by a simple quorum search. Then, this retrieved set would be massaged, with documents

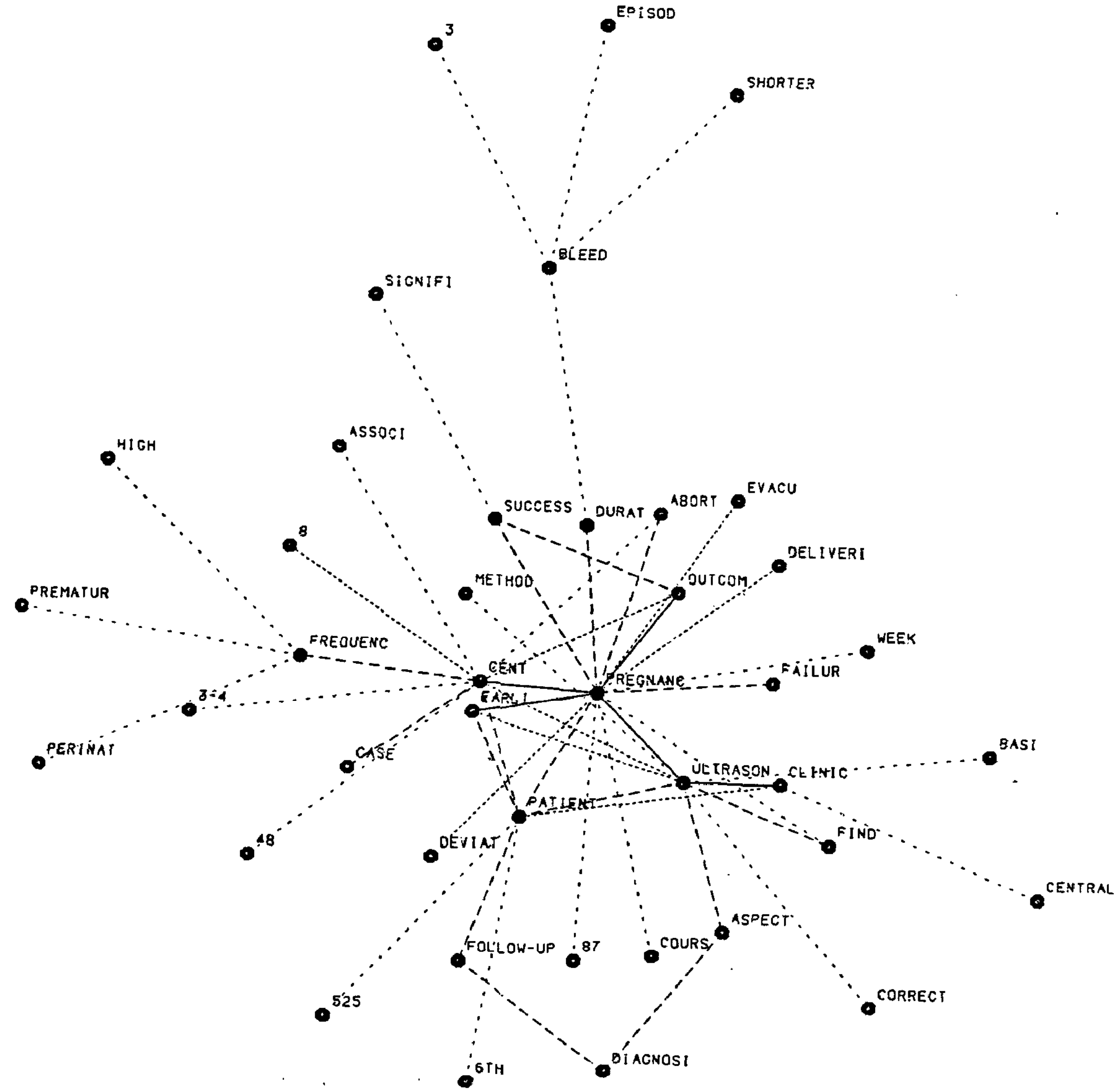

Figure 7. Representation of document 14.04 (judged quite useful).

either discarded or ranked also according to rules derived from the structures and lexical features of the ASKs, which are applied to the structures of the texts.

The complexity of the rules in both stages was, in general, the result of combinations of five different kinds of basic retrieval strategies, which we have labelled MATCH, TRIAD, STAR, PATH, and LEXICAL. The first is simple term identification, the next three are structural in nature, and the last combines with the others by taking account of special closed vocabularies.

We have decided not to attempt an enumerative classifiction of retrieval strategies, but rather to describe the individual basic strategies, which are invoked under specific conditions of ASK structures. Thus, we have a synthetic, faceted classification for retrieval strategies.

These strategies are briefly characterized below.

MATCH specifies an ASK structure, or area of an ASK structure, from which a list of terms is to be used for straightforward quorum searching.

TRIAD operates on clusters in the ASK structure, specifiying triplets of terms whose relationships and position in the ASK structure will be used to rank the texts.

STAR identifies terms for matching and ranking from stars in the ASK structure.

PATH identifies groups of terms for matching and ranking which are attached to clusters in the ASK structure, but are not parts of clusters. Group relations are retained for ranking purposes.

LEXICAL identifies 'pointer' or 'non-content' words in the ASK structure, which are eliminated from searching consideration and used to identify specific parts of the structure to be operated upon.

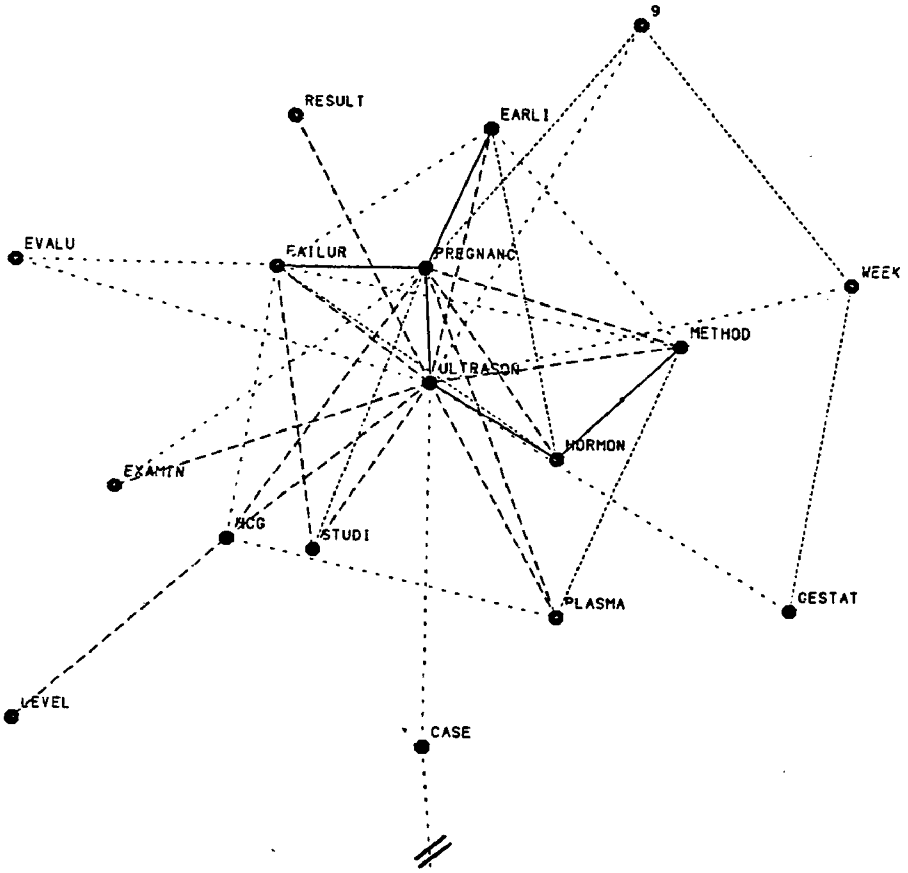

Figure 8. Representation of document 14.03 (judged marginally useful).

Thus, the strategy for s.14 can be summarized as:

__STAGE 1__

LEXICAL (finding one closed vocabulary term, ______)
TRIAD (operating on type 1 and 2 clusters)
STAR (operating on stars)
MATCH (using terms from TRIAD and STAR)

__STAGE 2__

MATCH (must have most involved node)
TRIAD (rank in order of structure duplication and node strength)
STAR (modify rank by inclusion of star nodes).

The rules for invoking the strategies depend upon the structures of the ASKs.

3.3 ASK structures and retrieval strategies

Given the nature of the retrieval strategies we identified, it is obviously more appropriate to identify significant characteristics of ASK structures for strategy invocation, than to attempt an explicit classification. We have identified a number of basic facets of the ASK structures which were regularly connected with the invocation of specific retrieval strategies. These can be viewed as facets in the traditional classificatory sense, with a specific citation order. In this case, we can consider our schema as a synthetic classification which implies a specific order and type of strategy implementation. But given that the purpose of the categorization is to get to retrieval strategies, it might be more clear to view the facets as data-driven rules, applied in a hierarchical manner to specify particular strategies.

Viewed in this way, we found three basic facets (rules associated with one another according to specific criteria). These are called ATTACHMENT, OVERALL STRUCTURE, and STRUCTURE CHARACTERISTICS.

ATTACHMENT is concerned with whether there are two or more structures in the ASK structure which are not connected at all with any of the others, in which case the ASK is termed 'detached'. The OVERALL STRUCTURE facet is concerned with the type, number and connection of clusters in the ASK structure. And the STRUCTURE CHARACTERISTICS facet is concerned with the local structural and lexical features of the ASK, and its overall connectivity. All of these are briefly specified in figure 11.

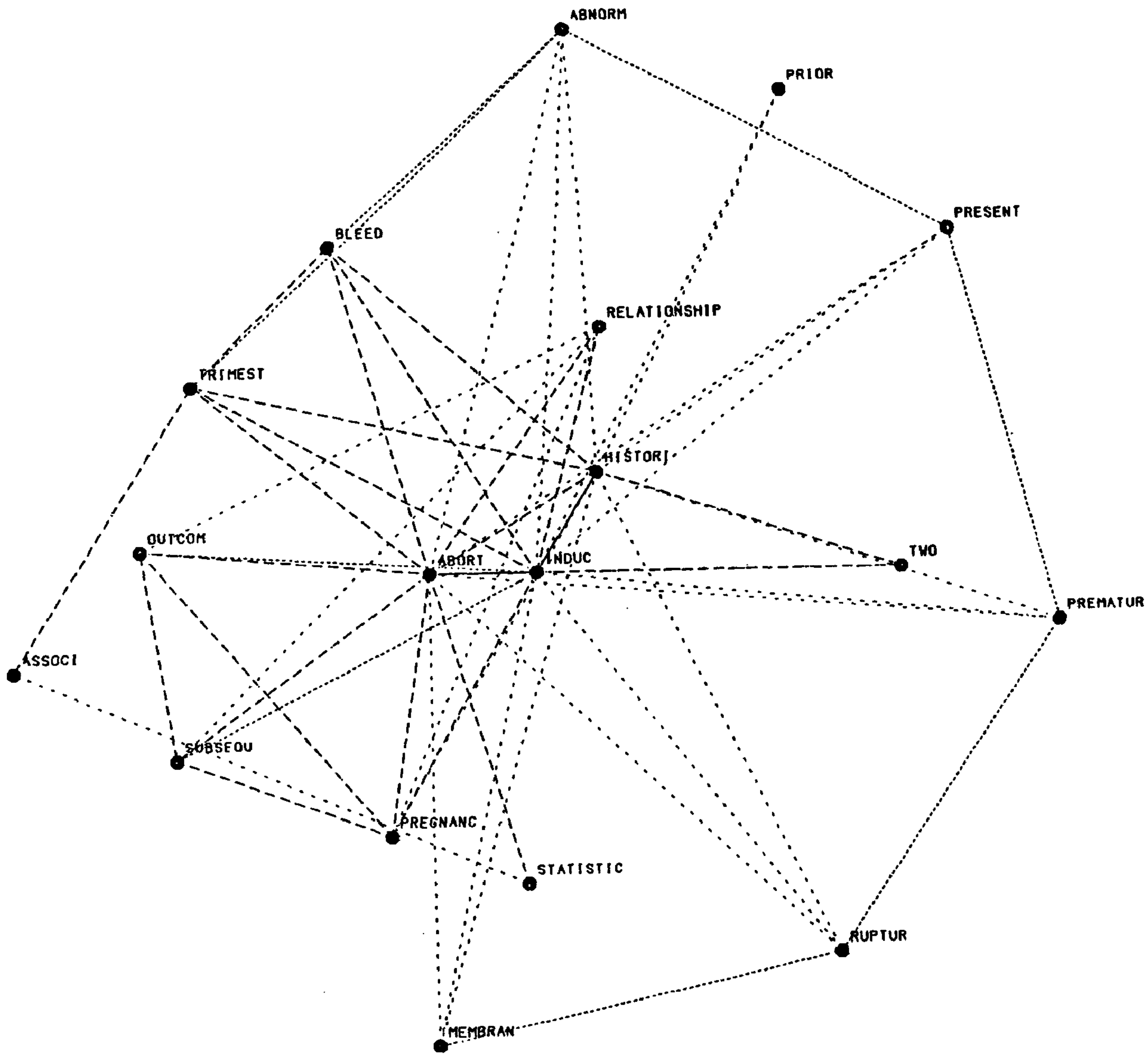

Figure 9. Representation of document 14.06 (judged not useful).

<u>ATTACHMENT</u>
1. Attached
2. Detached

<u>OVERALL STRUCTURE</u>
1. Single type 1 cluster [optimally with incorporated type 2 cluster(s)]
2. Two or more clusters linked at PL 0
3. Two or more clusters with PL≥1
4. No clusters

<u>STRUCTURE CHARACTERISTICS</u>
1. Substantive lexical items in cluster and magnitude of cluster
2. Connectivity of cluster at levels 1 and 2
3. Structure of cluster at levels 1 or 2
4. Number of stars
5. Number of lines

Figure 11. Facets of ASK structures.

The rules for invoking retrieval strategies follow the general form:

If ASK is of category x,
Then do y

where y is either specifying a retrieval strategy or invoking another rule. In order to show how this schema works, we once again return to the example of s.14.

The first facet invoked is ATTACHMENT. The basic rule in attachment says

if attached,
then do OVERALL STRUCTURE.

Since this is not a detached structure, we proceed to the facet OVERALL STRUCTURE. In this facet, s.14 responds to the the rule

if two or more clusters linked at PLO,
then do STRUCTURAL CHARACTERISTICS 1 (label ASK as B2).

STRUCTURAL CHARACTERISTICS 1 is a lexical characteristics rule, which goes:

```
1. Mark closed set words in clusters
2. If substantive words in type 1 cluster> 2,
   then TRIAD in type 1 cluster
3. If substantive words in type 2 cluster> 2,
   then TRIAD in type 2 cluster
4. If more clusters,
   then 3,
   else do connectivity.
```

The connectivity rule operative here is:

```
If high degree level 1 node,
then MATCH on node,
do connectivity.
```

The connectivity rule that applies is:

```
If highest degree level 1 node is star,
then STAR.
```

This will exhaust the possibilities for this particular structure, so that all of the terms identified by the invoked strategies will then be passed to MATCH, for the stage 1 quorum search. Then the subseqent ranking will take place, requiring that PREGNANC be in any relevant document (the MATCH invocation), then ranking by TRIAD inclusion and finally reranking by STAR.

Although there are many possible combinations of characteristics available, as it turns out, the number of specific rule results is small, so that the combinations may be collapsed into classes. These classes (still under investigation), determine the eventual retrieval strategy choice.

4. DISCUSSION

4.1 ASK and retrieval strategy classification and implementation

From the results and examples given in sections 3.2 and 3.3, it seems that a relatively small number of basic retrieval strategies can be used in combination to produce a variety of overall strategies and ranking mechanisms. These basic strategies respond not only to the requirements for straightforward matching, but also for those situations where taking account of general structural information and specific term interactions are necessary. Taken in specific orders, they can reflect the individual strategies discovered in the data analysis.

The characteristics used for classifying the ASK structures (or for invoking the retrieval strategies) are also relatively small in number, yet apparently responsive to relevant aspects of the structures as far as choice of effective retrieval strategy is concerned. This is of some interest, since the citation of the facets tends not to group the ASK structures into what one might think intuitively reasonable classes. For instance, overall connectivity appears not to be initially too important, nor are cluster size or numbers of stars. The most relevant criteria appear to be the number of clusters (no matter what type or size) and the internal structures of those clusters. We do not yet have any interpretations of what these groupings mean in terms of the nature of the users' problems, but are willing for the moment to accept retrieval performance as an adequate justification for them.

The implementation of these strategies appears to be possible if not exactly easy. By performing an initial quorum search, we eliminate the necessity of large-scale structure searching, a difficult process which is thereby restricted to a relatively small subset of documents which can be manipulated locally. Identifying the appropriate structures within the ASK seems likely not to present a problem. Furthermore, there are several natural formalisms for representing our facets and rules, such as frames and productions, which makes us think that this type of retrieval might be implementable in at least a test environment.

CROF86 has recently proposed an interesting scheme for taking account of term dependencies in a probabilistic retrieval environment. It might be of some interest to use problem statements and the structure identification rules proposed here as input to that retrieval mechanism. The ASK structures certainly provide a different rationale for term dependencies than normal frequency data.

A. CLUSTERS (2)

Cluster a	Cluster b
Type = 1	Type = 2
Mag = 9	Mag = 3
Con1 = 7/36	Con2 = 3/3
Con2 = 18/36	
Con3 = 19/36	
Con4 = 20/36	

B. STARS (3)

Star c	Star d	Star e
Type = 1	Type = 1	Type = 1
Mag = 4	Mag = 8	Mag = 3
Deg2 = 1	Deg2 = 3	Deg4 = 2
Deg3 = 1	Deg3 = 5	
Deg4 = 3	Deg4 = 7	

C. LINES (0)

D. RELATIONS

a - b	a - c	a - d
PL = 0 D = 0	PL = 0 D = 0	PL = 0 D = 0
Con1 = 5/27	Con1 = 5/9	Con1 = 3/9
Con2 = 14/27	Con2 = 9/9	Con2 = 7/9
Con3 = 17/27		Con3,4 = 7/9
Con4 = 17/27		
	a - e	
	PL = 0 D = 0	
	Con1 = 4/9	
	Con2 = 5/9	
	Con3,4 = 5/9	
b - c	b - d	b - e
PL = 0 D = 0	PL = 1 D = 1	PL = 1 D = 1
Con1 = 1/3	Con1 = 1/3	Con1 = 1/3
Con2 = 3/3	Con2 = 2/3	Con2,3,4 = 1/3
	Con3 = 3/3	

E. OVERALL CONNECTIVITY

n = 25 l_{max} = 300

7 Con1 = 7/300 5 Con3 = 30/300
18 Con2 = 25/300 11 Con4 = 41/300 = 0.13667

Figure 10. Characterization of ASK structure s.14

5. CONCLUSION

5.1 Retrieval strategies and ASKs

Even on the basis of the highly preliminary results presented here, it appears that it is possible to use characteristics of ASK representations to specify different retrieval strategies which are responsive tothe users' ASKs. The facets identified as useful in this study do group ASK representations in ways which seem to distinguish them one from another and also to imply appropriate, and substantially different retrieval strategies. The rules for identifying the ASK structures, and for implementing the retrieval strategies, seem within the capabilities of even present IR system implementations (given a suitable front-end). Thus, there is now some hope for answering the questions posed at the beginning of this paper. Nevertheless, our results are only indicative, and will require implementation and evaluation in a real test environment. This will be the subject of a further study, perhaps making use of CROF86's results.

5.2 ASK representation and human-computer interaction

The ASK project began with a design study initiated in 1978. Although various aspects of that original design have changed through the course of the project, two have remained firm: the basic ASK hypothesis, that people should not be forced to specify their information 'needs'; and, the narrative monologue problem structure. The validity of the former is, we believe, if anything strengthened by the results of this study, but we feel that it may be appropriate now to modify the latter.

We make this suggestion for several reasons. First, we wish to take account of results from studies by ourselves and others [BELK83; BROO85; CROF85], which stress the importance of interaction between user and intermediary in the building up of the intermediary's model of the user. One important aspect of that model is the model of the user's problem [BROO86; CROF85] or state of knowledge; that is, of the user's ASK.

Second, in our ASK projects, we have attempted to capture sufficient linguistic data in the initial problem statement, so that that statement alone could provide the basis for an adequate ASK representation. This has meant long narratives, with very few interventions by the experimenters. Although we tend not to worry about hardware, or even software constraints on our general system design, it seems that we should perhaps not count on speech understanding systems of the complexity required for this sort of data in the too near-term future.

Finally, our results indicate that a progressive building up of an ASK structure, via graphic interaction by the user with the intermediary's model of the ASK, might be more effective and efficient in developing accurate ASK representations, and in identifying important aspects of the ASK, than a one-time monologue.

The ASK classification and retrieval strategy specification will be valid whether the ASK structure is arrived at in a one-time or progressive manner. Indeed, it appears likely that our results could be used to guide progressive ASK representation. Therefore, for the reasons specified above, and in particular in order to integrate the results of this project into the distributed expert model of information retrieval, whether as an 'intelligent information provision mechanism' [BROO85] or 'expert assistant for document retrieval' [CROF85], we suggest that the next step in ASK investigation should be embedding ASK construction in an interactive dialogue between user and computer.

Our problem statement elicitation could, indeed, stand as a basis from which to begin such investigation, since its tripartite structure corresponds rather well to several opening and subsequent gambits often used by human intermediaries in information interaction [BROO83]. This type of interaction also coincides well with suggestions for driving such human-computer dialogues [DANI85]. And such a progressive building up of the ASK structure appears to match well with [CROF85]'s suggestions for a Request Model Builder. We are encouraged, therefore, that our results in this project, suggesting ways of distinguishing IR system user situations in ways which are directly useful for determining retrieval strategies, do not stand alone, but rather support and offer insights for other work on intelligent information systems.

Acknowledgements

The data collection for this project was organized and conducted by Kevin Hapeshi, who was also responsible for the associated programming and many developments in the text analysis and representation. We wish to thank the staffs of the Central Information Services and the library of the Institute of Education, both of the University of London, for allowing the data to be collected from their patrons, and for co-operating so whole-heartedly in this study. Our greatest thanks go to the users, from whom we asked so much, and who responded so generously.

REFERENCES

BELK80 BELKIN, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science v.5: 133-143.

BELK82 BELKIN, N.J., ODDY, R.N. & BROOKS, H.M. (1982). ASK for information retrieval: Parts I & II. Journal of Documentation 38: 61-71, 145-164.

BELK83 BELKIN, N.J., SEEGER, T. & WERSIG, G. (1983). Distributed expert problem treatment as a model for information system analysis and design. Journal of Information Science v.5: 153-167.

BROO86 BROOKS, H.M. (1986). Developing and using problem descriptions. Paper presented at IRFIS 6: Intelligent Information Systems for the Information Society, Frascati, September, 1985. Amsterdam, North-Holland, in press.

BROO83 BROOKS, H.M. & BELKIN, N.J. (1983). Using discourse analysis for the design of information retrieval interaction mechanisms. Proceedings of the ACM SIGIR Conference, Washington, D.C., 1983. SIGIR Forum 17:31-47.

BROO85 BROOKS, H.M., DANIELS, P.J. & BELKIN, N.J. (1985). Problem descriptions and user models: developing an intelligent interface for document retrieval systems. In: Advances in intelligent retrieval. Proceedings of Informatics 8. London, Aslib, in press.

CROF84 CROFT, W.B. & THOMPSON, R.H. (1984). The use of adaptive mechanisms for selection of search strategies in document retrieval systems. In: Research and Development in Information Retrieval. Proceedings of the Third Joint BCS and ACM Symposium, Cambridge, 1984. Cambridge, Cambridge University Press: 95-110.

CROF85 CROFT, W.B. & THOMPSON, R.H. (1985). An expert assistant for document retrieval. COINS Technical Report 85-05. Amherst, Mass., Department of Computer & Information Science, University of Massachusetts.

CROF86 CROFT, W.B. (1986). Boolean queries and term dependencies in probabilistic retrieval models. Journal of the American Society for Information Science 37:71-77.

DANI85 DANIELS, P.J., BROOKS, H.M. & BELKIN, N.J. (1985). Using problem structures for driving human-computer dialogues. In: RIAO 85. Actes of the conference: Recherche d'Informations Assistèe par Ordinateur, Grenoble, March 1985. Grenoble, I.M.A.G.: 645-660.

HAPE85 HAPESHI, K. & BELKIN, N.J. (1985). Developing and evaluating an au on-line information retrieval system based on user problem statements. In: RIAO '85. Actes of the conference: Recherche d'Informations Assisté par Ordinateur, Grenoble, 1985. Grenoble, IMAG:681-698.

ODDY77 ODDY, R.N. (1977). Information retrieval through man-machine dialogue. Journal of Documentation 33:1-14.

PALM84 PALMQUIST, R.A. & EISENBERG, M. (1984). Testing a text analysis metric using magnitude estimation. In: Proceedings of the ASIS Annual Meeting 21:231-236.

PORT80 PORTER, M.F. (1980). An algorithm for suffix stripping. In: New models in probabilistic information retrieval. C.J. van Rijsbergen, S.E. Robertson and M.F. Porter. BLROD Report 5587. Cambridge, Computer Laboratory, University of Cambridge: 98-106.

WEST83 WESTLAND, A. (1983). Text analysis for an ASK-based information retrieval system. M.Sc. Dissertation, Department of Information Science, The City University, London.

APPENDIX

GRAPH CHARACTERISTICS

We are characterizing our problem statement graphs according to the following features:

NODES and LINKS

The **DEGREE** of a node is the number of links incident on that node.

The **LEVEL** of a link is the association strength category of the link.

The **LEVEL** of a node is the maximum link level incident on that node.

GROUPS

A **GROUP** is a CLUSTER, STAR or LINE.

The **MAGNITUDE** of a group is the number of nodes in that group.

The **PATH LENGTH** between two groups is the minimum number of links that must be traversed to get from a node in one group to a node in the other group. The PATH LENGTH between two groups with a common node is 0. A PATH LENGTH of 1 is a **DIRECT** path.

The **DISTANCE** between two groups is the maximum link level connecting any two nodes, one in each group. For groups with shared nodes, DISTANCE = 0. Otherwise, DISTANCE applies only to DIRECT paths.

The **CONNECTION** value between two groups is the ratio of actual links between nodes in the two groups to the maximum possible links between them. CONNECTION applies only to DIRECT paths. CONNECTION at level 2 is the ratio of level 2 links to maximum, at level 3 of level 2 + level 3, at level 4 of all links. CONNECTION applies only to cluster-cluster, cluster-star and cluster-line paths. Maximum values for each are, respectively n x m, n and n links (where n and m are the number of nodes in each cluster).

CLUSTERS

CLUSTERS are of two TYPES.

TYPE I CLUSTER: a set of LEVEL 1 nodes which can all be reached directly by traversing level 1 links, and any level 2 nodes connected to any of the level 1 nodes in the cluster by at least two level 2 links.

TYPE II CLUSTER: a set of nodes of at least level 2 which are connected by at least two level 2 links, but not level 1 links, to other nodes in the cluster.

The **CONNECTIVITY** of a cluster is the ratio of number of links in a cluster to the maximum number of links for the number of nodes in the cluster (l_{max}). Connectivity at level 1 is the ratio of level 1 links to l_{max}, at level 2 of level 1 + level 2 links, at level 3 of level 1, 2 and 3 links, at level 4 of all links.

$$l_{max} = \frac{n(n-1)}{2}$$

where n = number of nodes in cluster.

STARS

A **STAR** is a set of nodes with one node (the **CENTRAL** node) connected to at least two nodes of degree 1.

The **TYPE** of a star is the level of the central node.

The **DEGREE** of a star is the number of links incident on the central node. Degree at each level is the number of links incident on the central node at that, and all higher, levels.

LINES

A **LINE** is a set of nodes with the pattern:

degree 1 - [degree 2]^{n}, where n>1, and indicates repetition.

The **TYPE** of a line is the number of links in that line. Degree at each level is the number of links at that, and all higher, levels.

OVERALL CONNECTIVITY

The **OVERALL CONNECTIVITY** of a graph is the ratio of number of links in the graph to the maximum number of links possible (1_{max}) for the number of nodes (n). OVERALL CONNECTIVITY at each level is the ratio of the number of links at that, and all higher levels, to 1

PROBLEM STATEMENT ANALYSIS

Each problem statement graph is characterized as follows:

*GROUPS

A. CLUSTERS (total number)
    TYPE
    MAGNITUTUDE
    CONNECTIVITY (BY LEVELS)
B. STARS (total number)
    TYPE
    MAGNITUDE
    DEGREE (BY LEVELS)
C. LINES (total number)
    TYPE
    MAGNITUDE
    DEGREE (BY LEVELS)

RELATIONS AMONG GROUPS

D. CLUSTER-CLUSTER
    PATH LENGTH
    DISTANCE
    CONNECTION (BY LEVELS)
E. CLUSTER-STAR
    PATH LENGTH
    DISTANCE
    CONNECTION (BY LEVELS)
F. CLUSTER-LINE
    PATH LENGTH
    DISTANCE
    CONNECTION (BY LEVELS)
G. STAR-STAR
    PATH LENGTH
    DISTANCE
H. STAR-LINE
    PATH LENGTH
    DISTANCE

OVERALL CONNECTIVITY (BY LEVELS).

*GROUPS are identified by lower-case letters, in the following sequence:

1. TYPE I CLUSTERS, ordered according to highest association strength within the cluster
2. TYPE II CLUSTERS, ordered as above
3. STARS, ordered according to TYPE.
4. LINES, ordered according to TYPE.