CHAPTER 2

THEORETICAL BASIS OF THE ASK IR SYSTEM

As explained in Section 1.2, the basis of this project is a combination of what we have rather loosely termed the 'ASK hypothesis' with the principles underlying Oddy's THOMAS system (Oddy, 1975). In this Chapter we expand somewhat upon the theory underlying these two approaches to IR, and on how they can be combined into a framework suitable for what we think of as a 'second generation' IR system.

Our basic premise arises from what we consider to be one of the central difficulties of IR: that people who use IR systems typically do so because they have recognized an anomaly in their state of knowledge of some topic, but they are unable to specify precisely what is necessary to resolve that anomaly. This premise can be seen as a restatement and perhaps extension of ideas proposed by Taylor (1968). Thus, we presume that it is unrealistic (in general) to ask the user of an IR system to say exactly what it is that s/he needs to know, since it is just the lack of that knowledge which has brought her/him to the system in the first place. This premise leads us to conclude that IR systems should be designed with the non-specifiability of information need as a major parameter. What sort of an IR system could this be?

We consider that IR systems, in general, consist of: a mechanism for representing information need; a text store; a mechanism for representing and organising texts; a mechanism for retrieving texts appropriate to particular information needs; and, usually, a mechanism for evaluating the effectiveness of the retrieval. Figure 2 (from Robertson, 1979) indicates these, and some other features, and the relationships among them.

From Figure 2, and experience, one can see that the starting points for IR system design are at either text or need representation, and that which of these one chooses, and the chosen method of representation, will strongly influence all of the other elements of the IR system. Most previous systems have begun from the text representation end, with not very much influence from the need end. We note that need representation appears to be the fundamental problem of IR, and so we suggest that a good IR system should be one which begins with need representation and designs the rest of the system about a mechanism and formalism, specifically designed for that purpose.

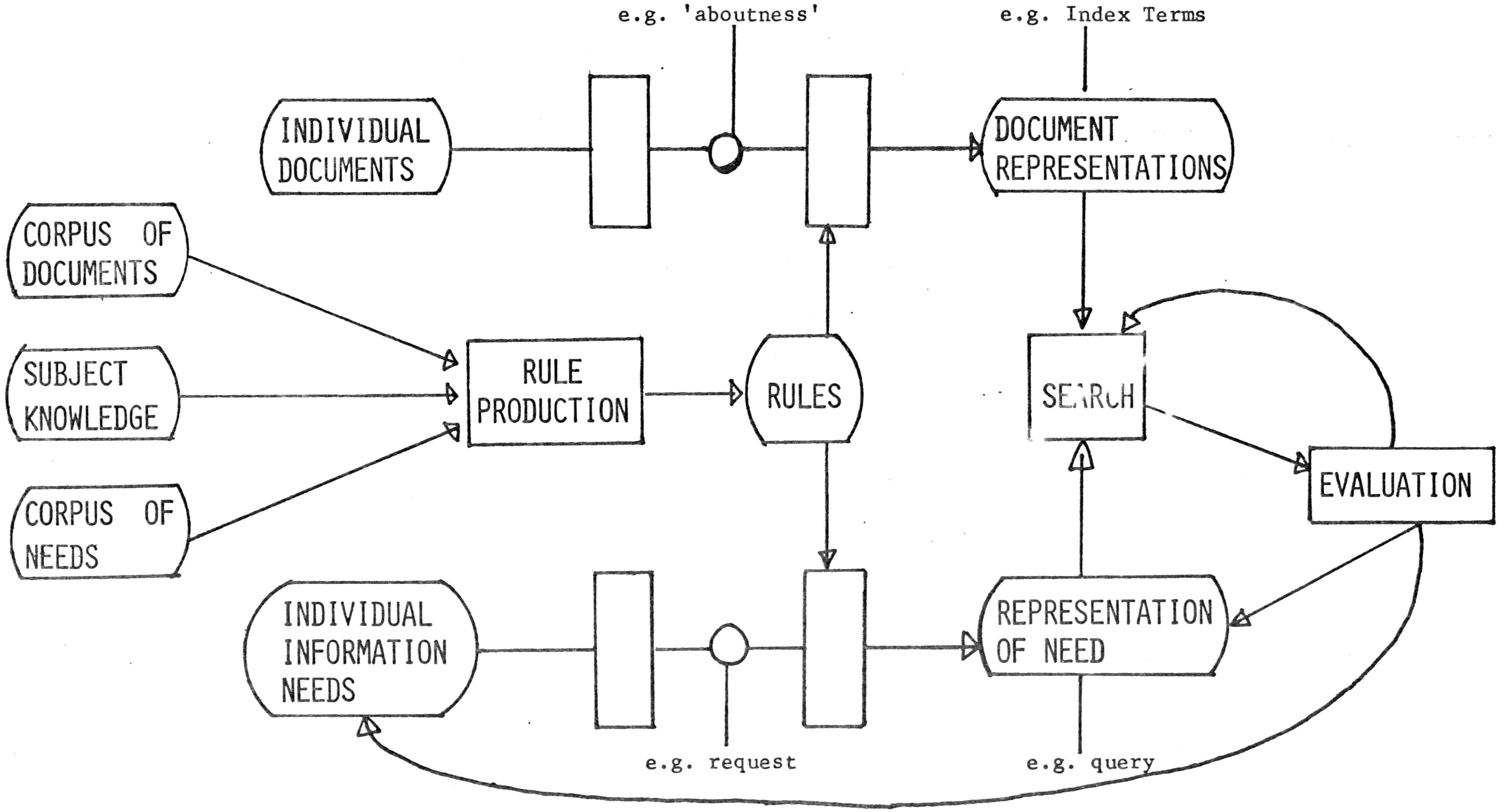

Figure 2 A model of IR (after Robertson, 1979)

Our mechanism is based on some experimental work by one of us (Belkin, 1977 a), which indicated that networks constructed from constrained word associations yield reasonable representations of individuals' states of knowledge about the subject to which the associations are constrained. In the same experiment, a network generated from word co-occurence analysis of a 3,000 word text (using the stimulus words of the word association test as major nodes) gave a reasonable representation of the 'information structure' of that text. The argument for using these techniques is based on the assumptions that concepts (represented by words) which are closely associated in an individual's state of knowledge will: be recalled close to one another in tasks such as word association; and, occur in close proximity to one another in a text by that person on the specific topic.

The formalism used for representation, a network of concepts represented by words, depends upon our concept of a state of knowledge as a multi-dimensional structure. We have discussed this concept, as it applies to our work, in, for instance, Belkin (1977 a), Belkin and Robertson (1976) and Oddy (1975), but the basic idea seems to us unexceptionable, except perhaps for the substitution of words for concepts. A network is certainly an effective way of representing this type of structure, and has the advantage, in our case, of being relatively easily derived from the sort of association data which we use.

There is no question that this sort of representation of a state of knowledge (or of the information structure underlying a text) is simplistic and naive, if one is attempting to obtain detailed representations for such purposes as natural language understanding, machine translation or retrieval from memory. On the other hand, it has the advantages of being fairly easily determined and reasonably machine-manipulable, important considerations in an IR context, where one needs to represent actual information needs and to manipulate large amounts of data. Our strategy for choice of formalism and mechanism for representation depends upon our assumption that there is no one 'real' or 'true' representation of knowledge or information, but rather many possible representations, each appropriate to particular problems. We decided upon networks derived from association data as a potentially useful representation for the purposes of IR because of the advantages of this type of representation mentioned above, and because there has already been some experience with similar formalisms in IR. But we bear in mind that this is indeed a very simple model, and so treat it only as preliminary: if it works, we will be pleased, if not, then we will amend it in ways which the evidence seems to indicate, or discard it for different models.

To this point, we have discussed means of representing information, or of what people <u>know</u> about a topic. Yet our problem is to represent what someone <u>doesn't know</u>; that is, to represent an anomalous state of knowledge (ASK) (Belkin, 1977 a). We obviously cannot use word association tests for obtaining the data for this purpose, since we would need to represent the ASK corresponding to a different topic for each user of the IR system, and thus would have no standard set of stimulus words. Our solution in this case derived, to some extent, from the suggestion by Wersig (1971) that a 'problematic situation' underlies an information need. In terms of the ASK hypothesis, this means that the user, recognizing an anomaly in her/his state of knowledge, interprets this in terms of a problem which s/he is attempting to solve; resolution of the anomaly will help to solve the problem. Thus, if we are able to obtain a statement of the problem, this may give us the data from which we can construct a representation of the ASK. This, of course, is old hat to reference librarians. So the basic data for our representation of ASKs are narrative statements by the users of the IR system, of the problems which brought them to the system. We assume that some sort of associative analysis of those statements will yield representations of the ASKs inherent in the problematic situation, which will be useful within the context of an IR system aiming to resolve the ASKs.

We have now discussed representation of need and text. The other major mechanism is that of search, or retrieval. Here again our concept of the ASK being the basis of the IR system colours our proposed solution. Belkin (in press) has raised the question of matching in IR. The basic issue is that since IR system users cannot (in general) specify precisely their ASKs, then IR systems which operate upon the 'best match' principles of exact specification of need are not, in general, likely to produce satisfactory results. Rather, one needs mechanisms which can recognize unspecified anomalies, and which can resolve, rather than match them. We recognize that there are ASKs which can be resolved by 'best match' type mechanisms, but we also recognize, on the basis of IR system experience as well as theoretical speculation, that there must be ASKs for which this type of mechanism is not suitable. Therefore, our IR system design assumes that there exist different types of ASK which require different retrieval mechanisms.

The evaluative mechanism in an IR system is now recognized to have an importance equivalent to the representation and retrieval mechanisms. There are many difficulties in establishing this mechanism, most of which we attempt to avoid here. But the basis of any evaluation must be the satisfaction of the user of the IR system with the system's performance in resolution of her/his ASK. Oddy (1975; 1977 a, b) built his system THOMAS about this realization by making evaluation its driving force. THOMAS is truly interactive in that each retrieval depends upon the user's evaluation of the results of the previous retrieval. Thus, the evaluation mechanism feeds-back directly to the need-representation mechanism, modifying it according to the user's immediate, perceived judgement of the value of the mechanism for resolving her/his information need. This sort of evaluation can thus cater for initial misrepresentation by the IR system and for changes in the information need which arise during the course of the search. In the IR system which we propose here, an additional factor for evaluation, multiple retrieval mechanisms, has been introduced. We assume, therefore, that an effective ASK-based IR system must be interactive in such a way as to incorporate direct evaluation by the user of retrieval output in terms suitable for modification of both need representation and choice of retrieval mechanism.

Thus we propose that an IR system based on the representation of information need as an ASK look something like that outlined in Fig.1. We do not mean to suggest that this is the only possible such system, but rather only that this is a realization of these ideas, based on further assumptions about representation and on practicalities of IR.

It is interesting that this approach to retrieval makes explicit a notion of relevance which is at variance with those implied by "best match" systems. The concepts which these systems use (to rank documents) are the <u>probability</u>, and the <u>degree</u> of relevance of a document to a query (Robertson and Belkin, 1978). Resolution of an ASK will usually require the retrieval of a <u>set</u> of relevant documents (not necessarily unique), each member of which is essential for the removal of some aspect of the anomaly.