CHAPTER 1.

INTRODUCTION

1.1 Project Origins

Improvements in the performance of information retrieval (IR) systems as presently designed seem to be limited to only marginal gains in terms of complete recall and precision or complete user satisfaction (see e.g. Robertson and Sparck Jones, 1976). We report here on a design study for an experimental IR system based on radically different hypotheses and assumptions than those underlying present systems, which we think may allow the design of IR systems which produce significantly better performance than that offered by present IR systems. We anticipate that, if successful, such a system could be used in both interactive on-line and batch off-line searching of any machine readable data bases which include abstracts as part of the document record.

This design study is the beginning of what could be a four-stage project. Because of the nature of the problem, it is convenient to break down the entire project into these four stages, each successive stage being predicated on the success of the preceding stage, each stage contributing knowledge necessary for implementing the next. The design study results are discussed in detail in the following chapters; here we can indicate in rather general terms what we anticipate the successive stages might involve.

The second stage would be actually building a prototype system based on the specifications of the design study, developing and processing an appropriate data base, and testing the system with simulated users. The major effort here would be on discovering the implications of the design proposals for an interactive system, such as searching techniques and costs, and programming. The result of this stage would be a prototype working interactive IR system, with known characteristics.

The third stage would be a small-scale experiment testing the system developed in stage two with real users. This stage will include modification and retesting of the system as results are accrued. The result would be a fully operational robust interactive IR system which had produced good retrieval results for a small number of users in a limited discipline.

The fourth stage would take the system and its components and test them in a large-scale, multi disciplinary environment, such as that of the

proposed 'ideal test collection' (Sparck Jones and Van Rijsbergen, 1975), and compare the results of the system against those of other, traditionally designed IR systems, probably in terms of recall and precision. The results will tell us whether the IR system we have designed can account for scale and discipline differences and whether its good results on a small scale translate to significantly better (than other systems) results on a large scale.

The benefit of this long-term project could be great indeed, for it represents the first (to our knowledge) attempt at designing an IR system which is not based on traditional hypotheses regarding document representation, request specification, information need and matching. If successful, it could be the basis of a 'second generation' of IR systems, designed on a specific theory of IR in which the user's role is the key, and producing IR performance much closer to complete user satisfaction than any present-day systems.

1.2 Background

Recent work by Belkin (Belkin and Robertson, 1976; Belkin, 1977 a,b; Belkin, in press) and others (see e.g. Harbo & Kajberg, in press) has called into question some traditional assumptions of IR, in particular those concerning the relationship of the request put to the IR system to the information need underlying the request, the basis for text and request representation in IR systems, and the retrieval mechanisms suitable for IR. This new approach recognizes that the fundamental element in the IR situation is the user's information need, from this realization going on to say that for IR to be successful, that information need must be represented in appropriate terms, with the remaining elements of the system (i.e. document representation, retrieval mechanism) being represented or constructed on the basis of that representation. The means to an appropriate representation is consideration of information need as an 'anomalous state of knowledge' (ASK) (see Belkin, 1977 a; in press). The ASK hypothesis is that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly. Thus, for the purposes of IR, it is more suitable to attempt to describe that ASK, than to ask the user to specify her/his need as a request to the system. One major difficulty with the ASK hypothesis is the question of determining the best system response to a particular ASK structure. One aim of the design study is to investigate this problem.

Oddy's (1975,1977 a,b) experimental interactive information retrieval system was designed to allow information retrieval without query formation, relying

on the system to construct an 'image' of the user's need. The program operates upon a graph whose points represent documents, subjects and authors; the lines stand for associations. A structural image of the subject area of interest to the user is maintained during the on-line dialogue. Its main component is a subgraph of the program's document collection structure. Strictly speaking, this image is not asserted to be a representation of the user's need, but rather a formal context within which documents satisfying the need might be found. Thus it is the structural properties of the image, and not merely matching terms, which determine the program's choice of documents to show to the user. The user may react to the documents, and their descriptions, and his response is used by the program to determine modifications to the image. This work has strong affinity to the ASK hypothesis, but is limited in its means for representation of document and user to traditional document descriptors, although in explicitly structural terms. In this design study, we have investigated means of combining and extending these two developments to deal with complex IR situations in an entirely new way, which the ASK hypothesis predicts would produce better results than traditional IR systems. The design study has attempted to resolve a number of problems raised by the new hypotheses, and to provide a preliminary specification for a system which incorporates these new assumptions. In addition, we report some results which can be interpreted on their own.

1.3 Project Specification

The system with which the design study was concerned consists of a structural text-analysis program, a suite of retrieval mechanisms, a data-base of abstracts and a mechanism whereby user and system interact. The system would work as follows (summarised in Fig 1):

1. The user discusses her/his information problem in an unstructured statement (say, 2-3 paragraphs long).
2. The problem statement is converted to a structural representation of the user's ASK by the text analysis program.
3. According to the type of problem structure (PS), one of the several available retrieval mechanisms is chosen to interrogate the data base (each member of which is represented by a structural representation of the information associated with the text). In this context, a retrieval mechanism is a strategy for resolving the anomalous aspects of a PS.
4. The abstract (i.e. the text) is printed out for the user to read.

Simultaneously, the user is presented with a brief explanation of why that particular text was chosen (explanation of the retrieval mechanism), indicating aspects of the text structure which the system finds significant in that choice.

5. The system then initiates a structured dialogue with the user, based on the information presented to her/him, inferring from the responses the user's attitude toward:
   a) the method of choice;
   b) the suitability of the document to the problem; and
   c) whether her/his information problem has changed.
6. a) The system changes retrieval mechanism if necessary, and/or
   b) the system modifies the problem structure if necessary, or
   c) the system stops if the user is satisfied.
7. The system returns to step 2 or 3.

Physically, this is an interactive on-line system with both printer and VDU.

1.4 Aims of the Design Study

Aspects of this system which must be clarified by the design study before any experiment is begun are:

1. The text analysis program. This is based upon an existing algorithm (Belkin, 1977 a) which determines statistically-based structures. The suitability of the eventual algorithm must be tested on real problem statements (see 2 below) and on abstracts (see 3 below).
2. The feasibility of obtaining problem statements, the possible types of problem statement, and the analysis of problem statements. To these ends, a number of real problem statements were collected and analysed, and the analyses verified.
3. The feasibility of using abstracts as the basic documents in the system, and the analysis of abstracts. To these ends, real abstracts were collected and analysed, and the analyses verified.
4. Some specific retrieval mechanisms must be proposed, based in large part on the types of problem structures encountered, although the ASK hypothesis implies some strategies.
5. The interactive system (especially the dialogue with the user) must be specified (although not necessarily fixed).

The design study thus required:

a) Collecting a suitable number of problem statements under appropriate conditions;
b) building a collection of abstracts (not necessarily related to the

1. USER'S PROBLEM STATEMENT

2. STRUCTURAL ANALYSIS OF PROBLEM STATEMENT

3. CHOICE OF RETRIEVAL STRATEGY ACCORDING TO TYPE OF ASK

4. ABSTRACT PRESENTED TO USER SIMULTANEOUSLY WITH EXPLANATION OF WHY TEXT WAS CHOSEN (STRATEGY AND SIGNIFICANT FEATURES)

5. STRUCTURED DIALOGUE BETWEEN SYSTEM AND USER FOR SYSTEM TO INFER USER'S EVALUATION OF:

   A) METHOD OF CHOICE

   B) SUITABILITY OF DOCUMENT TO PROBLEM

   AND/OR

   C) WHETHER NEED HAS CHANGED

6. MODIFICATIONS ACCORDING TO EVALUATION, OR FINISH

7. RETURN TO 2 OR 3 AS NECESSARY

FIGURE 1: AN ASK-BASED IR SYSTEM DESIGN

problem statements);

c) developing an analysis algorithm;

d) analyzing problem statements and abstracts;

e) asking users and authors for constructive comments on the suitability of the representations of problem and abstract;

f) developing an algorithm for classifying PSs

g) developing retrieval mechanisms; and

h) specifying the basic system.