USE OF SITUATIONAL INFORMATION IN INFORMATION RETRIEVAL

Robert N. Oddy & Elizabeth D. Liddy
School of Information Studies
Syracuse University

Summary of paper submitted to
CAIS Annual Conference
May 24-26, 1990, Kingston, Ontario

Introduction

This paper makes a contribution to the problem of information retrieval from large textual databases. Research on this problem in the past has, not unnaturally, been predominantly topic-oriented. That is it assumes that the enquirer will express an information need as a topic, and looks for ways of finding mention of that topic in texts. Beginning with a tiny trickle in the late 60s, a stream of thought has developed, which holds that users' problems and situations are essential ingredients of information needs (Taylor, Dervin, Wersig, Belkin & Oddy, etc.). Perusal of 1987 and 1988 reviews (e.g. in ARIST) and conference proceedings (e.g. ACM SIGIR) on information retrieval will show that topic-oriented research is, nevertheless, still predominant.

In the course of the project that we shall describe in this paper, it has become clear that the use to which information will be put, which is intimately related to the situation in which a user finds him/herself, is an important factor in properly understanding statements made by the user, and is an important determinant of relevance judgments. (Situational information is distinct from user-modeling, which has received significant attention in recent literature.)

It seems, therefore, that we need to develop models of situations, and ways of relating texts and natural language utterances to those models. In the information retrieval context, the texts are usually abstracts, and the utterances are users' problem statements. We have made use of some of the techniques of discourse linguistics to study the problem from this perspective.

Much of human discourse is concerned with sharing experience of the situations we must cope with. It is directed towards improving our ability to recognize common situations, and to respond effectively to them. Specifically, document abstracts are written with the intention of informing people who are participating in a communal activity, and who should, therefore, be able to recognize common situations. The situations of interest to the author are discernable in the discourse-level structure of an abstract. Similarly, we should look in users' utterances for indications of their problematic situations. Then users can be put in touch with related situations of others as represented in texts.

Our view of an ideal IR system is that it is a participant in a community of people with some shared goals, and is thus able to engage in discourse with the people. This is a subtle concept! We began our project with an ambitious objective of designing a system with powers of deduction, in the service of scientists engaged in empirical research. We planned to use frames and scripts and other AI paraphernalia to represent texts and model users' situations. We now have much more humility about our powers to understand the problem.

The STREAQ project

The STREAQ (STructured REpresentation of Abstracts and Queries) project was supported by grants from USWest Advanced Technologies and the Council on Library Resources. It arose out of two previous pieces of work conducted by the present authors: Liddy's dissertation on the structure of empirical abstracts, and Oddy's work on anomalous states of knowledge as revealed in users' problem statements. We will give a brief account of the STREAQ project. In order to make more obvious to users the commonality between their current situation in the empirical research process and completed research as summarized in abstracts, we delineated a discourse-level structure for presenting abstracts. This structure is based on the internalized notions of twelve expert abstractors as to the typical components of information in such abstracts and a linguistic analysis of 276 empirical abstracts which established the frequency and order with which components such as hypothesis, methodology, results, conclusions, etc. occurred. In addition, these components are revealed by a rather circumscribed set of lexical clues which are used consistently with relatively minor variation across disciplines. These clues and their frequencies are used to instantiate a frame-like structure for each abstract's text, producing a searchable structured representation which still contains the natural language of the abstract. We have implemented two programming approaches for the analysis of such abstracts, using the frequencies of these lexical clues as probabilities. In the first, written in Prolog, processing of the text terminates when the overall likelihood of a candidate structure achieves dominance over others. We have also used a 'connectionist' approach implemented on the Connection Machine. Processing consists of spreading activation in a network of nodes representing text elements.

From an analysis of 56 problem statements and in-depth interviews with four users we have learned that users are very aware of the "research script' and that their location in the script is a major explanatory variable for their relevance judgments. Conversely, from extensive interviews with five intermediaries, we have learned that although searchers are aware of the importance of a user's position in the research process, current systems offer no means to use this knowledge to adapt a search strategy. In addition, a review of the literature on building expert systems for information retrieval indicates that the central importance of situation is not being taken into account. User models in these prototypes contain mainly demographic variables, which

RS: research script dialogue; this will interact with the user on the matter of his/her situation, in terms of steps in the research script;

SF: search statement formulation; this will provide helpful tools to assist the user in constructing a formal query, involving not only keywords, but occurrence in selected abstract components;

RP: retrieval program; this searches the database of structured abstracts for those matching the formal search statement. It also interactively manages display of retrieved abstracts, indicating their structural relationship to the search statement.

The complexity of these modules, and the relationships between them (control structure and communication) is low in version one, and will increase only as we learn more. Version one consists of the three major components listed above, with no links between them, other than the essential, and conceptually straightforward SF->RP. The overall control structure is to begin with the sequence RS, SF, and then allow the user to switch from one module to another at will. There is no direct link between RS and SF (i.e. no use is made by the system of research script information), although we expect to see user behaviour affected simply by having to go through RS. The ability to switch back and forth between modules allows the user to adjust the search formulation in the light of retrieved abstracts (which are annotated with their structure), but version one has no facilities for suggesting improvements.

We already have the knowledge we need to build this primitive version, and it seems to be a promising tool for investigating the hypotheses listed above.

- Annals of Internal Medicine
→ structured abstr
→ critical appraisal.