Raya Fidel

1984: Challenges to an Information Society

Proceedings of the 47th ASIS Annual Meeting

1984 Volume 21

Philadelphia, Pennsylvania
October 21-25, 1984

Compiled by

Barbara Flood
Joanne Witiak
Thomas H. Hogan

Indexed by

Linda Cooper
Pat Heller
Andre Salz

Published for the
American Society for Information Science
by
Knowledge Industry Publications, Inc.
White Plains, New York

REQUEST-RELATED CRITERIA FOR THE SELECTION OF SEARCH KEYS (*)

Raya Fidel
Graduate School of Library and Information Science
University of Washington
Seattle, Washington

Abstract. Intermediary systems are designed to mediate between end-users and complex information retrieval systems. Most systems are based on text processing, and thus cannot process request-related criteria. Analysis of behavior of human intermediaries can serve as a basis the development of algorithms that will enhance system adaptability. Examples of decision rules used by searchers in the selection of search keys, free-text or controlled vocabulary keys, demonstrate that this approach is indeed useful.

Keywords and Phrases. User-cordial systems, online searching, free-text searching, descriptor searching.

INTRODUCTION

While most online literature searches are delegated to professional searchers, various attempts have been made to bring the end-user to the terminal. It is believed that end-users are likely to perform their own searches when search processes are simplified, or friendlier. Although various approaches can be taken to provide easier and friendlier user-system communication, the prevailing approach is to develop "interface" or "intermediary" systems. Indeed, some such systems are already available for public access, like CITE [1], and others are heavily tested in experimental settings, e.g., CONIT [2].

Approaching an intermediary system, users are freed from encountering peculiarities of databases and search systems. In principle, users can enter a request in a loosely structured format, preferably in natural language, and an intermediary system processes request terms, displays information to users, and asks for some sort of feedback. The information displayed may be in the form of a list of subject areas, databases, search keys, or actual citations from which users are asked to rank their selection. This sort of interaction usually proceeds to a point when users wish to terminate the session.

One of the most important decisions in a retrieval process is the selection of search keys. Systems vary in the degree of freedom they provide their users in this selection: some dutifully search only those search keys designated by a user, some use search keys designated by a user to generate additional search keys, and some automatically generate search keys with no user interference. However, all intermediary systems employ some kind of algorithm to generate search keys.

Algorithms used to generate search keys in most known systems are based on text processing. In other words, using cues provided by users, the text that is stored for retrieval is processed to generate search keys. Ordinarily, text is analyzed for word-occurrence frequency, statistical associations among terms, or other methods, most of which were first used in automated indexing. As such, these algorithms cannot be sensitive to request-specific requirements. Consider for example a request about the self-image of anorexic students during examination periods. Whether a user is interested only in anorexic students, in anorexic people in general, or primarily in students, the system, following its own algorithm, consistently decides whether to have anorexic students as one, or anorexia and students as two search keys.

While it is premature to require contemporary text processing methods to guide the creation of algorithms for request diagnosis, we can gain a meaningful insight from online searching behavior of experienced human intermediaries. Now, after a decade of experience with widely used online bibliographic retrieval systems, online searchers employ during the search process their own informal,and sometimes highly intuitive, decision rules. Examination and formalization of these rules may help us to increase the adaptability of intermediary systems.

To demonstrate the usefulness of this approach I will concentrate on the selection of the type of search keys: whether to search the descriptor field with controlled vocabulary keys, or whether to search for word occurrence in the text with free-text keys.

THE SELECTION ROUTINE

When a database can be searched by either controlled vocabulary or free-text keys, an intermediary system, whether human or automated, has to examine each request term and consider its representation: as a controlled vocabulary key, a free-text key, or both. A term may be mapped to a descriptor, through an exact or other kind of match, or it may not be mapped to a descriptor at all. Automated systems employ mapping algorithms that eventually lead to the selection of search keys based on the degree of term-descriptor match and on text characteristics. Human intermediaries, on the other hand, base their selection not only on the degree of match but also on request characteristics.

In an ongoing project I am analyzing protocols of "real life" searches performed by experienced human intermediaries to examine the conditions under which searchers select free-text keys and those under which they select controlled vocabulary keys. The analysis of over fifty search protocols, performed by four searchers in the area of the life sciences revealed an interesting pattern. First, some searchers consistently selected descriptors and would enter free-text keys as the last resort while others used heavily the trade-offs between controlled vocabulary and free-text searching. Second, the selection of search keys by type seems to follow a set of decision rules that is common with searchers, which I here call "the selection routine."

The selection routine specifies conditions which are necessary for a searcher to select a particular type of search key. Sufficient conditions, on the other hand, are determined by the nature of a searcher, whether dedicated to descriptor searching or willing to use both types, and by a host of situational factors such as time pressure, or the importance of the request. This routine groups together similar conditions to represent a systematic routine, but does not represent the dynamics of the selection process.

The development of the details of this selection routine is still in progress. But for the present time, it is useful to show here a few examples of conditions for search key selection, and to discuss briefly their applicability to the design of automated intermediary systems.

Selection of Free-Text Search Keys. When searchers consider the use of a free-text key, they first determine whether a term is a "good" term for free-text searching. The estimate of goodness is based on the predicted correspondence between the context of a term in a request and the context of the same term in the searched text. A term which, within a subject vocabulary, usually occurs in a particular context, is uniquely defined, and is specific in the concept it represents will be called "context-controlled" term in the following discussion. Terms that may occur in more than one context are called here "common" terms. In the request about anorexic students, terms such as self-image, anorexia and students are context-controlled terms, while examination is a common term. The term examination can occur in a subject-related context ("the best way to take student examinations"), or in a descriptive capacity ("examination of results shows deterioration in recovery"), or still, it can be used very loosely to represent any kind of inquiry. It should be mentioned here that Fugmann [3] distinguishes between "individual" and "general" concepts according to their suitability for free-text searching.

The attribute of context-control is central to search key selection. A request term that is a common term is selected as free-text key only for databases that do not provide controlled vocabulary. On the other hand, a context-controlled term may be selected as a free-text key for various reasons, depending on the degree of term-descriptor match. A few examples are illustrated below.

When a context-controlled term is mapped to a descriptor through an exact match, searchers still may elect to enter a request term as a free-text key under various conditions. One example is the case when the use of the descriptor to which a term is mapped, in a particular vocabulary, may lead to inconsistency in human indexing. Here, searchers may consider the use of a free-text key to be more trustworthy. If a vocabulary has the descriptors Self-Image and Self-Esteem, each having its particular application, the distinction may seem confusing to a searcher who may decide to use both a descriptor and the free-text key to compensate for indexers' errors.

Partial match of a request term that is context-controlled brings another example. Partial match usually implies mapping to a narrower descriptor. Searchers then may use a free-text key to inclusively search concepts that are not grouped together by the hierarchy of the controlled vocabulary. If the request term students is mapped to descriptors such as Foreign Students, College Students, and a descriptor Students does not exist, the free-text key can be used to retrieve information about any type of student.

The strongest case for using free-text keys is when a context-controlled request term cannot be mapped to a descriptor. Searchers then enter the term as a free-text key.

Selection of Controlled Vocabulary Search Keys. Controlled vocabulary keys can be selected only when a request term is mapped to a descriptor in some way. Although searchers use free-text keys when a term-descriptor mapping occurs, there are instances when they definitely prefer to use descriptors.

The strongest case for using controlled vocabulary keys is to represent request terms that are common terms. However, there are various conditions under which most searchers select descriptor representation for request terms, whether common or context-controlled terms. The examples described below emphasize the role of the structure of controlled vocabularies in search key selection.

One example is the use of role indicators. When a combination of request terms results small retrieval, searchers may eliminate a term from a query formulation and still represent it implicitly by using role indicators. Thus, if the combination of Anorexia Nervosa, Students, and Examination, retrieves almost no information, the descriptor Students can be eliminated and the role indicator Effects could be added to qualify Examination. The new formulation retrieves information about effects of examination in relation to anorexia.

Another example of the role of hierarchical relationships in the selection of search keys is the addition of a broader descriptor, according to a vocabulary structure, to improve recall. Controlled vocabularies readily suggest broader descriptors, but most importantly, they make it possible to indeed broaden the concept. Broadening the meaning of a concept is not always possible in free-text searching since the broader concept may be a common term. If entering a search key for the request term anorexia, whether free-text or controlled vocabulary key, does not yield enough retrieval, searchers may add a set retrieved by the descriptor Appetite Disorders.

DISCUSSION

These few examples can clearly demonstrate the benefits that could be gained from analysis of searching behavior of human intermediaries. Although this analysis of search protocols does not provide all the answers, it helps to define request-related characteristics and points the way for further research.

Most automated intermediary systems make no distinction between context-controlled and common terms. Indeed, in response to a request about information retrieval, CITE suggests the free-text term information as a possible search key. While such a key may be beneficial when searching titles of monographs, which is actually performed by CITE, it is completely useless in searching abstracts.

To improve their adaptability, intermediary systems can store a list of common terms specific to a subject vocabulary. Each request term is first checked against this list. Terms that are found in the list require further processing before their search keys are selected. Morphological analyses, or questions to users, can help to determine whether such a term should be mapped to a descriptor, replaced by a context-controlled free-text search key, or dropped from the formulation altogether.

Another example is the use of a hierarchical structure to select a broader descriptor when the use of the specific descriptor does not retrieve enough information. There are various ways to improve recall, such as truncation or elimination of a request component from a query formulation. Each path results in improved recall but adds different information. Depending on the nature of a request, searchers decide which path to take when they want to improve recall. When a term is central to a request, searchers most often use a broader descriptor, if it is still specific enough, to improve recall. Automated intermediary systems can identify central concepts in a request, say, by the user's ranking of search keys. If recall needs to be improved, the move to enter a broader descriptor to represent this term can have priority.

The examples described here are not sufficient to develop adaptive algorithms; the nature of context-controlled and common terms needs to be further investigated, and the conditions for the selection of a broader descriptor need to be more rigorously defined. These examples, however, illustrate the value of searching behavior analysis to the design of automated systems. If based on searchers' experience, automated intermediary systems can become adaptive to request-related requirements. Moreover, such systems can interrogate users about what users know best -- request parameters -- and select the most appropriate search keys -- decisions a naive end-user is not well enough informed to make.

REFERENCES AND NOTES

(*) This study is supported by the Graduate School of the University of Washington.

[1] Doszkocs, Tamas E. 1983. CITE NLM: Natural-Language Searching in an Online Catalog. Information Technology and Libraries, 2(4):364-380.

[2] Marcus, Richard S. 1983. An Experimental Comparison of the Effectiveness of Computers and Humans as Search Intermediaries. Journal of the American Society for Information Science, 34(6):381-404.

[3] Fugmann, Robert. 1982. The Complementarity of Natural and Indexing Languages. International Classification, 9(3):140-144.