ASIS '87

Proceedings of the 50th ASIS Annual Meeting

1987 Volume 24

Boston, Massachusetts
October 4–8, 1987

Edited by
Ching-chih Chen
Simmons College

Technical Program Chairman
Toni Carbo Bearman
University of Pittsburgh

Indexed by
Ching-chih Chen
Simmons College

Published for the
American Society for Information Science
by
Learned Information, Inc.
Medford, New Jersey

CONTROLLED VOCABULARY AND FREE-TEXT SEARCHING: SEARCHERS' SELECTION OF SEARCH KEYS

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

Abstract. To answer when, and under what conditions online searchers select descriptors or free-text terms, I observed 48 searchers on the job as they perform their regular searches. Preliminary analysis of some of these searches showed that the rules used by searchers for the selection of search keys are affected by the subject area of searching (which, in turn, determines the databases to be searched), and by the searching environments.

It is common to contrast free-text searching with controlled vocabulary (descriptor) searching, or even to prefer one type or the other as a general approach to searching. Recent studies, however, have demonstrated that free-text and controlled vocabulary not only complement one another, but both types are necessary for effective retrieval [2]. It is sound to assume that the desirability of one type or the other depends on specific request conditions: one set of conditions would require the use of free-text terms, another would require descriptors, and yet another set of conditions would require a combination of the two. The role of each type of indexing language in information retrieval, therefore, requires investigation.

THE PROBLEM

The purpose of this study was to identify the conditions under which each type of indexing language is required when searching in online bibliographic databases. It was assumed that searchers have developed intuitive "rules" which guide their decisions about the selection of search terms, or search keys. The study drew on the experience online searchers have accumulated, and aimed to uncover their intuitive rules: to answer when, and under what conditions online searchers select descriptors or free-text terms.

It was also assumed that the conditions for the selection of search keys would be determined by: the nature of requests; the personal style of the searcher; the subject area of searching; the database searched; and the environment, that is, the nature of a typical information need in the setting wherein each searcher works.

THE METHOD

To uncover the intuitive rules for the selection of search keys, I observed online searchers on the job as they performed their regular searches. I asked them to speak out loud as they prepared a search and during their session at the terminal. I recorded their words and later analyzed each search to determine the reasons for the selection of each search key.

During the first phase of the study, which is already completed, I observed eight medical librarians performing nearly 100 searches. The outcome of this phase is a decision tree that traces their selection of search keys [3].

Briefly, this tree shows that a term may or may not be appropriate for free-text searching. A descriptor is almost always used when a term is not suitable for free-text searching (e.g., "analysis") but is mapped to a descriptor. If, however, a term is "good" for free-text searching, additional possibilities exist. The important factor here is whether or not the term is mapped to a descriptor--through an exact match, a partial match, or to a broader descriptor. Under such circumstances, searchers can consider other factors such as precision and recall requirements, or the number of databases to be searched.

In the second phase of the study, I observed an additional forty online searchers selected from a variety of libraries and information agencies where they search a wide spectrum of subject areas. I analyzed their searches, compared their choices to those in the decision tree, and expanded the tree to reflect a larger variety of searching behavior. This analysis also uncovered additional reasons for the selection of search keys.

RESULTS

At present, data analysis for the phase two of the study has just begun. The findings reported here are preliminary; they are based on an examination of the searching records of the first ten of the forty searchers, and on an examination of but one aspect of searching behavior among all forty searchers who participated. Though conclusions are tentative at this time, substantiated findings and a complete report will be presented at the conference.

The preliminary results of the study indicate that the rules used by searchers for the selection of search keys in phase two are affected by the subject area of searching (which, in turn, determines the databases to be searched), and by the

searching environments. These effects are demonstrated by two findings. First, to fit the search keys selected by the ten new searchers into the decision tree required a modification of the tree. Second, the variety of conditions recognized by the new searchers was much smaller than the variety recognized by the eight medical librarians.

Most of the ten new searchers work in special libraries serving commercial organizations and they search primarily the business and engineering literature. Their searching behavior modified the decision tree because the tree described a situation in which searchers would, at one point or another, check whether or not a term they plan to use as a search key can be mapped to a descriptor. But the ten searchers introduced a new situation in which the searcher does not know whether or not a term is mapped to a descriptor. (Note: in 100 searches performed in medical libraries this condition has never occurred, but as we discuss later, this condition is quite common among non-medical librarians.)

Further, in a few instances searchers selected a free-text search key for a term that is not suitable for free-text searching (e.g., "analysis") even though it can be represented by a descriptor. They used the free-text term under special conditions: they combined the term with the query formulation to increase the precision of a retrieved set, assuming that using a descriptor would limit its recall. Here again, the medical librarians studied never made such use of free-text terms.

Beyond what has been described, the searching behavior of the remaining searchers required no further modifications of the decision tree.

The selection of search keys by the ten new searchers also focused on a small number of conditions--the most straightforward ones. While the search key selection by medical librarians was spread almost evenly among eight "prominent" conditions (of the twenty-five possible conditions), the ten new searchers most frequently recognized only three conditions. Moreover, some of the conditions identified by the medical librarians--such as, the need to use descriptors when the concept is not explicitly mentioned in the text--were not recognized at all by the new searchers.

The most straightforward rule in search key selection is: when you find a descriptor that matches the term exactly, use it; if not, use free-text terms. Analysis of patterns of search key selection among the first ten searchers in phase two revealed that the most common conditions were also the most straightforward ones:

* in 24% of search key selection, searchers selected a descriptor because the term matched a descriptor exactly;

* in 32% of search key selection, searchers selected a free-text term because the term could not be mapped to a descriptor;

* in 24% of search key selection, searchers selected a free-text term because they did not know whether or not the term could be mapped to a descriptor.

In total, 80% of the search keys selected were the most simple ones.

These results also show that the ten librarians are generally comfortable with not consulting a thesaurus--in 24% of the instances in which they had to select a search key, they did not use a thesaurus. This phenomenon can be explained by two factors: the librarians often did not have the relevant thesauri at hand, and each request typically required searches in a number of databases.

Preliminary results also indicate that when more than one option is available, the factors that affect the selection of a particular type of search key can be identified and grouped into three categories: 1) the nature of the request; 2) the nature of the database; and 3) the personal style of the searcher.

An example will illustrate these categories. According to the decision tree, searchers have two choices when a term cannot be mapped to a descriptor. They can either select a free-text term as a search key or they can use a free-text term to further probe indexing--that is, they can enter the free-text term and display the indexing of relevant citations.

When searches performed by all forty searchers in phase two were analyzed, the reasons for using a free-text term rather than probing the indexing fall into the three categories mentioned above. More explicitly, when they found that a term could not be mapped to a descriptor, they used a free-text term:

Request Related Selections

* if the term is good for the request because it is commonly used; if it was used by the user; or if it appears in relevant titles;

* if the free-text term is used to "correct" retrieval of previous attempts, to increase precision or recall, or to increase both;

Database Related Selections

* if the database has no thesaurus;

* if the term is not represented in the thesaurus because of its nature (e.g., a proper name, too specific, too new);

* if indexing is not trustworthy;

Searcher Related Selections

* if the searcher prefers to search with free-text terms or uses descriptors only when they match a term exactly.

Further analysis will determine which specific reasons are most common among searchers and how they relate to the subject area and the environment of searching.

DISCUSSION

The completed results of the study will contribute to the resolution of the "free-text vs controlled vocabulary" debate. They will not only demonstrate that the two types complement one another, but they will also show how free-text searching complements descriptor searching, and vice versa. Eventually, a comprehensive list of the conditions that require descriptor searching or free-text searching for the best results--and the conditions that necessitate a combination of both--will be provided.

Most important, the nature of these two types of indexing language will be presented in terms that are directly relevant to strategy formulation for online searching of bibliographic databases. The list will show how the character of a term, the special requirements of a request, the characteristics of the database and of the environment, and the personal style of a searcher all combine to determine the selection of search keys.

Unveiling the nature of search key selection is important to understanding online searching behavior. Search key selection is only one component of the process of online searching. It is, however, an important component that may effect other components in the process, such as database selection or the decision to terminate a search. Therefore, understanding the factors that affect search key selection leads to a better understanding of the factors that affect searching behavior in general.

Finally, the results of this study can be used in the development of intermediary expert systems that advise end-users about the selection of search keys when they are searching their own requests. The process of search key selection in this study will be expressed in a formal model--a decision tree--that can be incorporated into a knowledge base of such expert systems. These intermediary systems can then become powerful because their decisions about search key selection will be based on factors that are specific to each request, search environment, or database.

The results presented here are also important to the design of intermediary expert systems because they point to characteristics and capabilities of such systems. The finding that search key selection is affected by the subject and environment of searching, for example, suggests that a "universal" expert system may not be of optimal help. Instead, we may want to consider a variety of expert systems, each suitable for a particular combination of subject area and environment.

The usefulness of intermediary expert systems is brought out by the finding that it is not uncommon for searchers to search without consulting a thesaurus for the reasons already mentioned. The process of thesauri look-up is not as costly in an intermediary expert system as when performed by humans. Therefore, such systems can eliminate this situation and add much flexibility and power to online searching.

ACKNOWLEDGMENT

Nancy Phelps, Michael Crandall, Cindy Cunningham, and Kathleen McCrory carried out the field observations. Without their excellent contribution, the study could not have been accomplished.

NOTES

[1] This study was partially supported by Grant IST-8509719 from the National Science Foundation.

[2] Svenonius, Elaine. "Unanswered Questions in the Design of Controlled Vocabularies." Journal of the American Society for Information Science, 37(September 1986):331-340.

[3] Fidel, Raya. "Towards Expert Systems for the Selection of Search Keys." Journal of the American Society for Information Science, 37(January 1986):37-44.