ASIS '88

Proceedings of the 51st ASIS Annual Meeting

1988 Volume 25

Atlanta, Georgia
October 23-27, 1988

Edited by
Christine L. Borgman
Edward Y. H. Pai
University of California, Los Angeles

Compiled by
Carol Nixon
Lauree Padgett
Robert Williamson, Jr.

Technical Program Chairman
Ward E. Shaw
Colorado Alliance of Research Libraries

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

FACTORS AFFECTING THE 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 47 searchers on the job as they performed their regular searches, and used the case study method to analyze both the search and verbal protocols. Results show that free-text terms and descriptors were selected with the same frequency, but individual searchers who prefer to enter free-text terms--and those who tend to avoid consulting a thesaurus--are likely to be science searchers who answer practical requests and search several databases for each request. Reducing the number of databases needed to answer a request would enhance the effectiveness of online searching.

INTRODUCTION

The intellectual components of a typical online search can be classified into three basic categories: (1) definition of query structure; (2) selection of search keys; and (3) feedback review. The study reported here focused on the second category: the selection of search keys.

There are two distinct types of search keys. In free-text searching, searchers may enter any desired term or phrase and retrieve items that include the term or the phrase in their text. With descriptor searching, searchers may decide to use search keys from a controlled vocabulary.

The issue of "controlled vocabulary vs. free-text" searching is still controversial in information science. Practitioners and researchers either believe that descriptors searching is archaic and belongs only to the print media, or they believe that the use of controlled vocabulary is essential for high performance in online searching.

These opposing stands are not completely contradictory: though the first advocates the exclusion of one mode of searching, the second recognizes the significance of both modes, with one mode complementing the other. Today, the complementary nature of free-text and descriptor searching is well established and a number of studies has demonstrated that both modes are necessary for effective online searching [2], [3], [4]. This partnership between the two types of search keys raises the question of when is it best to use a descriptor, and when is it best to enter a free-text term.

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 professional searchers have developed intuitive "rules" which guide their decisions about the selection of 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 searchers select descriptors or free-text search keys, or both.

This paper addresses the general factors that affect the selection of search keys--factors that are not limited to a specific request or to a particular searcher. The following factors have been examined: the number of search keys selected for a search, the frequency with which a thesaurus was consulted, the number of databases searched, the searching style of the searcher, the numbers of moves per search, the subject area of the searcher, 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, we observed online searchers on the job as they performed their regular searches. We asked them to speak out loud while they prepared a search and during their session at the terminal. We recorded their words and later analyzed each search to determine the reasons for the selection of each search key [5]. Each searcher agreed to "donate" five searches; after observing all five searches, we interviewed each searcher to gain information which had not been accessible to observation.

RESULTS

The study analyzed a total of 281 searches which were performed by 47 searchers from all types of libraries, searching in a variety of subject areas. The searchers selected a total of 3,635 search keys, of which 1,607 (44%) were descriptors. Eliminating from the statistics 330 free-text terms that were selected for databases

which have no controlled vocabulary, 1,698 search keys were selected to be free-text terms when searchers had a choice between the two types. That is, when they were searching in databases which provide some kind of indexing, searchers selected descriptors 49% of the times and free-text terms 51%. In addition, from among the free-text terms selected to search databases with indexing, 624 (37%) terms were selected without even consulting a thesaurus.

The study also analyzed the modifications of search strategies: the moves that searchers performed during each terminal session [6]. A move can be either a conceptual move--one that changes the meaning of a request--or an operational move--one that keeps the meaning of a request unchanged. For example, to replace a term with a broader term, or to introduce an additional concept are conceptual moves. On the other hand, to add synonyms, or to limit retrieval to English articles are examples of operational moves. Of the 1,244 moves recorded for the 281 searches, 497 (40%) were conceptual moves. Eliminating 308 instances of the operational move "to add a database" (which is determined by the distribution of databases within a subject area rather than by the searcher's choice), the number of operational moves was 439--47% of the rest of the moves.

These findings show that searchers, in general, do not prefer one type of search key more than the other: free-text terms and descriptors were selected with almost the same frequency. Similarly, both types of moves--conceptual or operational--are equally popular among searchers.

A total of 70 databases was searched during this study, 65 (93%) of which have some kind of indexing. The high proportion of databases with indexing may indicate that most database producers perceive controlled vocabulary to be essential to the quality of their products.

To find the variables that affect the selection of search keys, we examined how this selection is associated with other variables. The table at the end of this paper is a summary of this analysis for a total of eight variables. We analyzed these associations on two levels: (1) each search was considered a distinct instance (a total of 281 instances); and (2) the values for each person were aggregated so that each person was considered a distinct instance (a total of 47 instances). One should note that the instances on the search level are not independent because each set of about five searches was performed by the same person.

Search-Keys Ratio

The selection of search keys was measured by the search-keys ratio: the percent of free-text terms selected. The ratio for each person was calculated by adding up the number of search keys used in all of that person's searches. Data analysis shows that the search-keys ratio is significantly associated with: the subject area of the searcher; the moves ratio; the number of databases; and the environment within the sciences.

The subject area in which a searcher specializes determined the value for the variable "subject area of searcher." The value "general" was used for searchers who are called upon to answer requests in a variety of subject areas (as is often the case in public libraries or for independent consultants). Analysis of variance showed a significant correlation ($p<.01$) between search-keys ratio and subject area ($F=5.24$). The means of search-keys ratio was 36% for medicine, 48% for the social sciences, 67% for general, and 78% for the sciences.

Thus, searchers of the scientific literature, or the generalists, use a high proportion of free-text terms--significantly higher than the one observed in other subject areas. This finding is not surprising; it has been long assumed that searches in the scientific literature do not require the use of controlled vocabulary because the scientific terminology itself is already controlled. Medical terminology, however, shows the same degree of control, yet medical searchers used the smallest proportion of free-text terms. This finding indicates that the nature of the terminology is not the most important factor determining the selection of search keys.

The searcher's searching style is also associated with the selection of search keys. The percent of operational moves (the moves ratio) correlates significantly ($p<.01$) with the search-keys ratio (search level: $r=.184$; person level: $r=.434$). This association shows that searchers who prefer to employ operational moves (operationalist searchers) also prefer to use free-text terms.

Another expected finding is the association between the search-keys ratio and the number of databases used in a search. This significant correlation ($p<.01$) shows that the larger the number of databases used for a search, the more likely the searcher is to select free-text terms (search: $r=.277$). Similarly, searchers who regularly use a number of databases for each request are more likely to use free-text terms than those who use just one or two databases (person: $r=.414$).

To discover whether the type of request affects the selection of search keys, we looked at the environments in which searchers work. We used somewhat imprecise categories, contrasting practical environments with theoretical ones.

Analysis of variance shows that for those who search the scientific literature, the searcher's environment is associated with search-keys ratio ($F=3.09$; $p<.01$). Searchers who typically answer requests that address practical problems use free-text terms 65% of the time; those who typically search databases for theoretical requests use free-text terms only 36% of the time.

It is also illuminating to examine the factors that are not associated with the search-keys ratio. It is sound to assume that the mechanics of the search process itself would determine the ratio of free-text terms. However, our results show that neither the number of moves per search nor the number of search keys correlated with the search-keys ratio. While it seems plausible to assert that interactive searchers use a large proportion of free-text terms, or that the increase in the number of search keys is always supported by adding free-text terms, our data reject these assertions.

Consulting a Thesaurus

The selection of search keys is not always supported by consulting a thesaurus. The thesaurus look-up is an important variable because nothing is gained when a searcher avoids consulting a thesaurus, and much could be lost. Further, this is not an obscure phenomenon as 37% of the free-text terms selected to search databases with indexing were picked without thesaurus consultation. It is important, therefore, to identify the factors that affect this variable.

In our study we examined the instances where searchers used free-text terms without first checking a thesaurus. As was the case with the search-keys ratio, we found that the frequency of entering search keys without consulting a thesaurus correlated with the number of databases per search ($p<.01$; search: $r=.294$; person: $r=.397$), with the moves ratio ($p<.01$; search: $r=.167$; person: $r=.413$), and with the subject area of the searcher ($p<.05$; $F=3.51$).

The average frequencies of entering search keys without consulting a thesaurus for each subject area are revealing. No medical librarian in our sample ever entered a free-text term without first checking a thesaurus, but searchers in the social sciences refrained from consulting a thesaurus 13% of the time. Next are general searchers with 29% and science searchers with 32%. One possible explanation for the difference is that medical librarians had the most rewarding experience with thesaurus usage, while science searchers had the least rewarding experience.

CONCLUSION

These findings are only preliminary, but they do reinforce some common knowledge and provide new leads. For example, the findings confirm the well-accepted assumption that requests in the sciences lend themselves better to free-text searching than those in the social sciences and humanities, as well as the assumption that searchers who have to search a number of databases for a request are likely to use free-text terms.

Further, the study provided findings that point the way for future research. For example, the results show that the type of request (whether practical or theoretical) affects the selection of search keys, and that operationalist searchers use free-text terms more frequently than conceptualist ones. From the various conclusions that can be drawn at this preliminary stage of the analysis, I would like to mention only two.

First, we can now construct a profile of the searchers who use free-text terms more often than other searchers. They are likely to have these characteristics:

- -- be a science searcher
- -- if, as a science searcher, they usually answer practical questions, they will use still more free-text terms
- -- be an operationalist searcher
- -- need to search a number of databases for each request
- -- feel comfortable entering a term without consulting a thesaurus

Second, the results of this can help designers of information systems to provide more flexibility in the choice of search keys, and therefore improve searchers' performance. The data in the table show that the variable most associated with the others is "the number of databases per search." The need to search a number of databases was also frequently mentioned as a reason for selecting a free-text term and for not consulting a thesaurus. Therefore, having to search a variety of databases for one request is a limiting factor.

Though it is generally believed that the availability of a large number of databases enhances online searching because searchers have more choice, it is not clear how much freedom is actually introduced by this multitude--each database is somewhat different from the others and searchers often feel they have to cover all bases. On the other hand, the use of a variety of databases limits their options in the selection of search keys.

The uncoordinated growth of databases is an impediment to online searching. Clearly, versatility in databases is required and

one universal thesaurus cannot satisfy the needs of all potential users because different user groups require different coverage and vocabulary for searching. Existing databases, however, do not always match a particular user group because we have learned that searchers in certain subject areas have to regularly search a number of databases for each request. Standardization and coordination is required if users are to be able to fully exploit the capabilities of online systems. While switching languages and expert systems can obscure the variability introduced by databases, this variability is introduces most often by marketing factors and does not necessarily improve the results of online searching. It is better, therefore, to coordinate the construction of databases and thesauri and avoid the large investment in overcoming variability that is not necessary at the first place.

VARIABLES	1	2	3	4	5	6	7	8
1. # of search keys	*	NS	NS	.01	NS	.01	NS	.01
2. Search-keys ratio	NS	*	.01	.01	.01	NS	.01	.01
3. Thesaurus look-ups	NS	.01	*	.01	.01	*	.05	NS
4. # of databases	.01	.01	.01	*	.01	.01	.1	NS
5. Moves ratio	NS	.01	.01	.01	*	NS	.01	NS
6. # of moves	.01	NS	*	.01	NS	*	NS	NS
7. Subject area	NS	.01	.05	.1	.01	NS	*	*
8. Environment	.01	.01	NS	NS	NS	NS	*	*

NOTES

[1] This material is based upon work supported by the National Science Foundation under Grant No. IST-8509719.

[2] Katzer, Jeffrey, "A Study of the Overlap Among Document Representations," Information Technology: Research and Development, 1 (October 1982) 261-274.

[3] Fugmann, Robert, "The Complementary Nature of Natural Language and Indexing Languages," International Classification, 9 (3, 1982) 140-144.

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

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

[6] Fidel, Raya, "Moves in Online Searching," Online Review, 9 (February 1985) 61-74.