January 1985

Contributions of Value Added Fields and Full Text Searching in Full Text Databases

Carol Tenopir
University of Tennessee - Knoxville, ctenopir@utk.edu

CONTRIBUTIONS OF VALUE ADDED FIELDS AND FULL-TEXT SEARCHING IN FULL-TEXT DATABASES

Carol Tenopir, University of Hawaii at Manoa

Abstract: Some database producers assume that the availability of full text databases will make indexing and abstracting obsolete. Very few full text databases include both controlled vocabulary indexing terms and abstracts. As full text databases become more widely available, this assumption is beginning to be tested.

This study reviewed research to date that has examined full text retrieval performance on inverted file systems. Research comparing efficacy of searching on value-added fields vs. full text was also reviewed. Conclusions are not yet definitive but suggest that value-added field contribute to comprehensive retrieval and improve precision.

The author conducted a retrieval performance experiment in 1983-84 on the Harvard Business (HBR) full text database and the BRS search system. HBR contains controlled vocabulary descriptors and abstracts, allowing retrieval performance of these fields to be compared with full text.

Results showed that full text retrieved a high proportion of the relevant documents. Controlled vocabulary searching, and to a lesser degree abstracts, also contributed unique relevant documents. The value-added fields allowed much better precision in searching and had lower costs for searchers.

Unique relevant documents retrieved by each method were examined to judge the special contribution of each field. Controlled vocabulary compensated for variations or changes in terminology, levels of specificity of terminology, and incomplete search strategy development. Abstracts pulled concepts together and somewhat standardized language. Full text allowed articles to be retrieved that contained relevant information peripheral to the article as a whole, compensated for deficiencies in controlled vocabulary, and often used more syncnyms.

Suggestions for additional research will also be presented.

1. INTRODUCTION

Full text databases are increasing in numbers on the commercial inverted file search systems. Because full text databases are a relatively new phenomenon on the once traditionally bibliographic systems, much full text search strategy is based on assumptions or trial-and-error rather than on systematic study of the best results. Some producers or providers of full text databases assume that the availability and searchability of complete texts in inverted file systems will make indexing and abstracting obsolete. Few full text databases also include the value added fields of controlled vocabulary indexing terms and abstracts. This paper reviews some past research that compared search results and describes a recent project that examined the relative contributions of full text, controlled vocabulary terms, and abstracts in online search strategy on the Harvard Business Review database.

2. REVIEW OF RESEARCH

Many studies in information retrieval may be relevant to retrieval performance; described here are those few that examined full text vs. controlled vocabulary descriptors or abstracts on standard inverted file systems. The American Chemical Society (ACS) and BRS did a series of user studies of the full text of ACS journals before they were made commerically available on BRS. The researchers observed that searchers were able to find specific factual information by searching texts of articles when there were no corresponding terms in titles or abstracts [1].

Studies by Hersey et al. of the Smithsonian Institution Science Information Exchange (SSIE) compared retrieval performance from searching subject indexing codes with searching text words in a database of one-page summaries of research in progress [2]. An early version of the Mead Data Central software was used. The study concluded that retrieval performance with indexing terms was superior to that when searching free text words. Recall was about 30% higher; precision was 15-20% better. Both approaches offered advantages by retrieving documents that were unique and relevant. Text word retrieval provided detail; index code retrieval retrieved concepts and broad subjects and contributed to more complete retrieval. The authors recommended combination systems rather than forcing searchers to rely on one search technique.

Indexing may be expensive from the database producer's viewpoint, but free text searching can be expensive to users in terms of computer time. Three times as much computer time was required for the free text searching in the SSIE study as for the controlled indexing searching. The free text searches required three and one-half times the number of terms per question, and 14 times as many term combinations. In a study by Stein et al., six expert patents classifiers were asked to conduct 12 patent searches each on a LEXIS database of 50,000 patents. After the searches were completed, each query was studied to determine where in each patent the search terms occurred and what term variants occurred [3]. Results indicated that the full text resulted in substantially better retrieval than any single patent representation. Combinations of document segments were ranked by how often the full search

query would be retrieved if a search was limited to them. A combination of summary and description provided the best search results (87%), followed by title and claims (16%) and title and abstract (7.5%). When individual segments were examined, full text, summary and description were of most help for retrieval while titles, abstracts, and claims were of limited help.

Several conclusions are suggested by these studies. No one method of searching (e.g. full text, abstract, controlled vocabulary descriptors, title) provides total recall in standard search systems and no one method consistently provides the best results. Controlled vocabulary searching, abstract, and full text searching retrieve unique documents, suggesting that the best strategy is to use a combination of methods.

3. HARVARD BUSINESS REVIEW STUDY

In an experiment conducted in 1983-1984, the author compared results from searching on words in complete texts, abstracts, and controlled vocabulary descriptors using the Harvard Business Review (HBR) full text database on the BRS search system. HBR has both controlled vocabulary descriptors and abstracts in addition to the complete texts of every article from 1976 to the present.

In a series of 31 questions, the text achieved on the average a relative recall of 73.9%. Controlled vocabulary had an average relative recall of 28% and abstracts 19.3%. Full text had the poorest average precision ratio of the three — 18% as compared to 34% for controlled vocabulary and 35.6% for abstracts. Full text searching was the most expensive with an average unit cost per relevant document of $7.86, as compared to $3.54 for controlled vocabulary searching and $3.89 for abstract word searching.

Although the full text contributed the highest recall, each of the three search methods contributed unique relevant documents in different questions. No one search method consistently provided all relevant documents. In only nine of the 23 topics that retrieved relevant documents were all documents retrieved by the full text. For the rest of the 23 topics the abstract and/or controlled vocabulary were required to achieve comprehensive retrieval. Samples of relevant documents that were retrieved by only one search method were examined in an attempt to characterize the unique contribution of each representation. This characterization may assist searchers to decide which search method would be best for a given topic.

3.1 Controlled Vocabulary

In nine questions relevant documents were retrieved by the controlled vocabulary that were not retrieved by any other search method. After examining these documents, there seem to be three major reasons why the controlled vocabulary resulted in retrieval while the full text did not. These reasons are: 1) variations or changes in terminology, 2) specificity of terminology, and 3) incomplete search strategy development by the searcher.

Terminology used in the texts of the articles in question varied from the

more commonly used terminology found in similar articles. A relevant article retrieved only by controlled vocabulary in one question, for example, was a reprint of an article originally published in 1950. In 1950 the now common terms of "product diffusion" or "early adopters" were not in use. HBR's controlled vocabulary retained the older term "new products", use of which in the search strategy would have retrieved the 1950 reprint. In another question the controlled vocabulary term "family" retrieved a document relevant to personnel policies for spouses working in the the same firm. Nowhere in this document were the terms nepotism, couples, marriage, married, or spouse found, but the terms wives, wife, or relatives would have resulted in retrieval by the full text. The author of the document assumed male-owned firms that were hiring relatives (including wives); the searcher failed to add the appropriate synonyms. In another question a relevant document retrieved by the descriptor "flexible working hours" was not retrieved by the full text search because only the term "flexitime" was used in the text of the article. The searcher used the alternate spelling "flextime", but failed to use "flexitime."

These three questions point out the need to use both modern and older forms of words and to use many synonyms to achieve complete full text retrieval. The constancy of controlled vocabulary terms for any concept as compared to the inconsistent and changing nature of text language often assists retrieval.

Other relevant articles retrieved only by controlled vocabulary were retrieved because the controlled vocabulary terms were much broader than the subject requested or because there were terms for only one of two facets of the question. In one, for example, the user requested documents on the retirement of farmers and ranchers. Both concepts were specified in the full text and abstract searches. The HBR controlled vocabulary does not contain a term for the farmers or ranchers concept, however, so the single broad term "retirement" was searched. Some aspects of retirement planning are independent of the retiree's occupation, however, so some relevant documents were found with the broader strategy.

One question asked for articles on in-plant recreational facilities. Again, a fairly specific strategy using both concepts was conducted for the full text and abstract searches. Only the very broad term "employee benefits" was available for searching in the controlled vocabulary. The additional relevant documents retrieved by this strategy used the term "perks" rather than benefits in the text. The recreational facet was represented by such terms as relaxation, entertainment or recreational facilities. Another question also contained two concepts, only one of which was available in the controlled vocabulary. In a question on in-house databases, "Information systems" or "databases" retrieved relevant items that were not retrieved by the full text for two reasons. The first reason is that only the terms computer or data processing were used throughout the texts. The other reason is that the second facet of "office" or "inhouse" excluded relevant articles. The authors of the articles assumed any computer system or database was located "inhouse" or in the office without explicitly using those terms.

In summary, the strength of controlled vocabulary to control synonyms and varied or changing vocabulary was supported in this study. In full text searching on the standard commercial systems such as BRS the burden of compensating for language inconsistency is on the searcher. Controlled vocabulary costs the database producer more to create, but retrieves items

difficult to find using full text only. Ironically, the limitations of a broad controlled vocabulary contributed to more complete retrieval without achieving unacceptable precision when no terms were available for both concepts of a search, because the single broader concept retrieved relevant items without adversely affecting precision. In a larger database this might cause unacceptable precision levels.

3.2 Abstract

The abstracts did not contribute as many unique relevant documents as did the controlled vocabulary. There was high overlap of abstracts with full text, which in a way shows the success of the HBR abstracters in summarizing the content of each article in the author's own words. Still, the relevant documents retrieved only by abstracters were examined to determine why they were not retrieved by any other method.

There seem to be three main reasons why relevant articles were retrieved by abstract searching but not by full text. These reasons are: 1) words did not appear in the same text paragraphs, 2) language varies in texts, and 3) the searcher did not use all possible synonyms in the search strategy.

The most common reason for abstract-only retrieval resulted from using the SAME paragraph operator in the full text searches instead of the broader Boolean AND. This decision was made because BRS and HBR both recommend limiting full text searches to the same paragraph. Search terms from both facets of a search appeared somewhere in many of the texts of these documents but the terms did not appear in the same grammatical paragraphs. In the abstract the important concepts were brought together into the same field (i.e., paragraph). In a question about the effects of unions on the introduction of new technology, several articles were retrieved by the abstract because all of the ramifications of unions were listed in the abstracts. The texts discussed each of these effects in turn without repeating the term "union". The same is true in a question about second careers. In one article retrieved only by abstract words, the concept of training or retraining was not mentioned in the same text paragraphs as the concept of new jobs or layoffs, but these two concepts were brought together in the abstract. For a question on stress of working wives the article that was retrieved only by the abstract is about the stressful role of corporate wives. A mention of them entering the workforce was in a paragraph without other search terms, but all concepts were together in the abstract field.

Another reason for document retrieval by the abstract but not by the full text is one of language. In articles uniquely retrieved by the abstract in one question, "wives" or "wife" are used more frequently in the text than "woman" or "women". The abstract uses women. If the synonyms "wives" or "wife" had been added to the full text search, this article would have been retrieved. In another question an article about Sioux Indians does not refer to them as a "minority" group in the text, but the abstract uses this term.

As with controlled vocabulary, it appears that the abstracts in HBRO sometimes compensate for the inconsistency of language and the necessity of many possible specific terms for the same concept. A comprehensive full text search requires listing many synonyms for each concept.

3.3 Full Text

Full text searching often retrieved many more unique relevant documents than either controlled vocabulary or abstract searching. One frequent contribution of the availability of full texts is, thus, an increase in the number of documents retrieved. An examination of a portion of the relevant documents retrieved only by the full text revealed four major characteristics. These are: 1) level of specificity can better match the question, 2) full text can compensate for deficiencies in the controlled vocabulary, 3) some concepts that are implied in the abstract but not mentioned explicitly are mentioned in the text, 4) full text sometimes uses more synonyms and can thus compensate for incomplete search strategies.

Articles that on the whole are broader in scope than the search request (that include the search topic as only a minor portion of the article) are the major reason for full text-only contributions. The abstracters and indexers attempt to match the depth or level of specificity of each article taken as a whole. Thus, an article on unionization of professional employees may list the specific professions in the text, but these are not mentioned in the abstract or controlled vocabulary terms. For documents retrieved only by abstracts the opposite was sometimes found—terms in the abstract were broader than the text terms. In a specific question about the effect of labor unions on the decline of productivity in the U.S., some articles mentioned many reasons for this decline, including labor unions. The specific reasons are accessible only via the full text where they are listed or mentioned briefly. This variance in the level of specificity was the one major reason for many of the text-only retrievals.

Another contribution of the full text is that it compensates for deficiencies in the controlled vocabulary. Several topics did not have appropriate descriptors for a concept, so narrower or broader terms had to be used. One question was about collective bargaining in colleges and universities but there are no HBR descriptors for colleges or universities. The "product and service" terms were used, but relevant articles discussed colleges and universities as a subject, not as a product or service. The same reason applies to a question about collective bargaining in libraries, schools, etc. HBR's policy of assigning only five descriptors means that only the major issues in an article are indexed. This, plus the policy of indexing and abstracting at the level of specificity of the article as a whole, results in many full text-only retrievals. All articles retrieved by the full text only seemed to have appropriate index terms within the constraints of the controlled vocabulary and the HBR indexing policy, however.

Compared to abstracts, full text facilitates retrieval of articles that mention a specific facet of a topic, but that are generally broader in scope than the search question. Full text also retrieved some articles when one facet was assumed but not explicitly mentioned in the abstract. For example, in the question about recreational facilities as benefits in organizations, the abstracts of some documents implied that recreational facilities provided to employees to reduce tension are benefits, but the term "benefit" was not explicitly used. In a question about attitudes toward hard work the concept of "attitudes" or "feelings" about hard work was implied but not mentioned in the abstracts.

Abstracts sometimes used jargon or a single term for a concept in the text while the full text stated it in several ways. For example, in a question about minorites including Hispanics the title and abstract of an article referred to "Mexicans". The text, however, used various synonyms such as "hispanics", "chicanos", "Mexican-Americans", resulting in retrieval. In a question about layoffs or unemployment, "hard-to-employ" was the only term used in the abstract of one document to describe unemployed workers. Unemployment caused by layoffs was included in the article but the term layoff was found only in the full text.

3.4 Summary

This examination has analyzed the unique contribution towards comprehensive retrieval that is made by full texts, abstracts and controlled vocabulary searching. The controlled vocabulary indeed controls synonyms or language that changes over time. The abstract brings together major concepts in an article that may have been discussed separately in the text. It also somewhat standardizes language. The major contribution of the full text is made when an article is of broader scope than the search question or when one facet of a question is mentioned only as one possible factor in a broader issue. Specific terms or causes are often listed or discussed in textual paragraphs but are too minor or specific to be indexed or abstracted. Each search method makes its own contribution and often this contribution depends on the nature of the search question or the individual articles in the database. No one method is complete for every situation. Relevant articles will be missed and search costs may be higher if searchers do not have the option of choosing various methods of searching. Indexing and abstracting are not made obsolete in full text databases, all representations assist complete retrieval and provide their own unique contributions.

3.5 Suggestions For Future Research

Because full text databases have not been widely available on commerical search services for long, there has not yet been much research that examines their characteristics. The present study is thus only an early step in determining how full text databases might best be searched, but the conclusions must be limited to a relatively small database of the business literature. The methodology used in this study should be replicated in other subjects to see if retrieval performance and search results vary with the subject matter of the text and to see if low precision becomes an even greater problem in larger databases. Related research has indicated that language patterns vary with the nature of the discipline, but this has yet to be tested with full text online searching. Such an extension into other social science disciplines and into physical science disciplines could have important ramifications for searchers in search strategy development and for publishers in database design decisions.

Another variation on the present study would be to change the full text search strategies to use the Boolean AND operator rather than the paragraph SAME operator or to compare various full text strategies. This would help to identify the best full text strategy. The type of research mentioned so far is practical given the realities of the present systems, but can only suggest ways

these existing systems might be improved. Any studies limited by the fundamental designs currently in use cannot reveal optimal performance in an ideal situation that has no previous design assumptions. Additional user studies are needed that will reveal how potential users would most like to use full text databases if they were not restricted by current system constraints.

Future research should take into consideration the different possible uses of full text, including browsing, fact retrieval, and finding articles on a given topic. Users with different types of needs may have different requirements for search and display features. The research on the use and retrieval characteristics of full text databases is just beginning.

NOTES

1. Kay Durkin, et al., "An Experiment to Study the Online Use of a Full-Text Primary Journal Database," in Proceedings of the 4th International Online Information Meeting: 1980 December 9-11, London, England (Oxford, England: Learned Information, Ltd., 1980), pp. 53-56.

2. David F. Hersey, et al., "Comparison of On-Line Retrieval Using Free Text Words and Scientist Indexing," in The Information Conscious Society: Proceedings of the American Society for Information Science 33rd Annual Meeting: 1970 October 11-15, Philadelphia, PA (Washington, DC: ASIS, 1970), pp. 265-268.

David F. Hersey, et al., "Free Text Word Retrieval and Scientist Indexing: Performance Profiles and Costs," Journal of Documentation 27 (September 1971):167-183.

3. D. Stein, et al., "Full Text Online Patent Searching: Results of a USPTO Experiment," in Proceedings of the Online '82 Conference, 1982 November 1-3, Atlanta, GA (Weston, CT: Online Inc., pp. 289-294.