

INTRODUCTION

Since the early 1980's when NEXIS first became available online via the Mead Data Central system, the complete texts of many general interest magazines, journals, and newsletters have been added to the major commercial online search services. Although many more journal titles are still accessible online in bibliographic-only databases, the gap in the ratio of full text coverage to bibliographic is narrowing.

Full text databases online hold great promise for solving document delivery problems, especially as transmission speeds increase and graphic images are added. The capability of searching every word in a complete text impacts retrieval capability and search strategy techniques as well. Full text databases are not yet total replacements for print because most do not include graphics and they often exclude non-article portions of journals (things such as letters-to-the-editor, book reviews, advertisements, etc.). Still, they are a great step forward in document delivery and in enhanced retrieval of information.

RELATED LITERATURE

Advice on how to best search full text databases has come from database producers, online vendors, and experienced searchers [1]. Most recommend avoiding use of the Boolean AND operator to link concepts, searching instead on concepts found within the same grammatical paragraph or within a specified number of words. A study by the author compared retrieval performance of text using words within the same paragraph to retrieval performance of controlled vocabulary descriptors and words from abstracts [2].

A follow-up study showed that the Boolean AND retrieved additional relevant documents, but with lower precision ratios [3]. That study also looked at ways to optimize full text searching through ranking algorithms [4]. These tests all show full text searching achieves high relative recall with low precision. A study in legal texts showed the opposite [5], raising the possibility that different types of literature or different styles of writing (as well as different approaches to search strategy) result in different retrieval performance.

The study reported here extends the author's earlier research to explore results obtained from different strategies for searching full texts of popular magazine literature. Using a database that contains complete texts of many different magazines introduces the variable of writing style and offers the ability to compare results based on type of literature. This study tests the widely held assumption that searching for concepts within a paragraph will yield the best search results by looking at several search techniques.

Any examination of full text search results must explore reasons why users might search text databases. If the primary motivation is the same as for bibliographic retrieval -- that is to locate some documents on a given topic, then the measures used to evaluate bibliographic database retrieval performance should be appropriate for full text evaluation. Measures such as precision and recall may not be so appropriate if the purpose for searching full text is to locate a single relevant fact or a bit of information within a larger document. Recall is an inappropriate measure with retrieval of facts. Some recent studies [6-8] have found that end users use full text databases as both document locators and as fact or partial document retrieval tools. End users may be more tolerant of browsing through less relevant or irrelevant documents using KWIC features than when an intermediary does a bibliographic search for them. The present study looks at search strategies for full document retrieval as well as for partial document retrieval.

SYSTEM FEATURES

This study tests full text searching using search features that are available on commercial online systems of today. The major commercial online systems that provide access to full texts typically allow every word in the texts to be searched (except stop words) and employ inverted index file structures. They use some standard search techniques that were developed originally for bibliographic databases. These techniques include: Boolean operations (usually AND, OR, and NOT), proximity searching (typically adjacent, within a specified number of words, within the same field), and truncation (usually right-hand truncation of a specified or unspecified length). Other refinements to search techniques were added specifically for full text databases, including proximity searching within the same grammatical sentence or paragraph, display of only the paragraphs that contain search terms, some automatic word standardization (e.g. automatically searching both singular and plural forms of a word, both British and American spellings, both standard abbreviations and their spelled out form) and, less commonly, word frequency information for help with relevance judging.

DESCRIPTION OF THE STUDY

The Magazine ASAP [TM] database on the DIALOG system was used in this study. Magazine ASAP (MASAP) is produced by Information Access Company and provides the full texts of articles, editorials, columns, reviews, product evaluations, and recipes from over 100 popular magazines from 1983 to the present. The magazines vary considerably, including such disparate titles as Time, Popular Science, New Republic, People, PC Week, Science, Teen, and Playboy. MASAP records include (in addition to the document texts) such things as bibliographic

information, controlled vocabulary descriptor terms, and caption headings.

DIALOG search features that are especially useful for full text search and display include:

- proximity operators including: word adjacency (W), within a certain number of words in a specified order (nW), within a certain number of words in either order (nN), and within the same grammatical paragraph (s),
- the ability to specify a field or fields (e.g., text only or title and text) for search and display,
- a Key-Word-in-Context (KWIC) display that displays only those portions of the complete text that contain the search words or phrases.

DIALOG has no word occurrence features (such as sorting of output by number of search words) or automatic word equivalency features.

Eight questions have been searched and analyzed so far for this study on the MASAP databases. Queries came from reference questions posed by undergraduate students at a university library or by members of the public at a public library. (These questions were not originally searched online.) Most are current event topics that were posed to allow students to gather enough relevant information for a term paper or class presentation. Table 1 lists the questions searched to date and the strategies used online.

Each question was searched in several ways in order to test the proximity variations available through DIALOG and to allow comparison of the "SAME" paragraph feature with other search techniques. Each search of the full text looked for words or phrases in the text using the following relationships to link concepts:

1. Boolean AND
2. Within the same grammatical paragraph (the (S) operator on DIALOG)
3. Within 10 words in either order (10N)
4. Within 5 words in either order (5N).

The choice of 10N and 5N between words was arbitrary. Such word proximity features are commonly used with full text searching with the assumption that increasing the number of intervening words will increase recall.

Search terms or phrases used standard features such as the adjacency operator (W) or truncation as appropriate. It was expected that the precision ratio would increase as the word relationships got more specific, and that the (S) operator would yield the best overall balance of recall and precision.

All searches were conducted by the author, an experienced DIALOG searcher. This has the advantage of consistency but the disadvantage of a single searcher's view of strategy. The same words or phrases were used for each of the four search techniques searched for each question. Relevance was judged by the searcher.

PRELIMINARY RESULTS

Table 2 shows for each question and each search method how many documents were retrieved, how many relevant documents or partially relevant documents were retrieved, relative recall, and precision scores. The four text strategies go from broadest to most narrow and are inclusive--that is the broadest strategy (linking concepts with the Boolean AND) includes all of the documents retrieved by the other methods. The same paragraph strategy (S) is the next broadest and includes all of the documents retrieved by (10N) or (5N), with the rare exception of when words within 10 or 5 words of each other are in different grammatical paragraphs. By definition the (10N) strategy includes all items retrieved by (5N), which is the most restrictive. Recall would thus be expected to decrease with each technique, while precision would be expected to increase.

Recall cannot be measured in this study, but if the broadest strategy (AND) is defined as 100% recall, relative recall can be measured for other techniques.

Precision is difficult to judge with full text because often a single paragraph in a document may have relevant information. Relevance here was judged on a three-value scale: 1) not relevant, 2) part of the document relevant (10 lines or fewer of relevant information), 3) relevant (entire article or more than 10 lines relevant.)

As can be seen in Table 3, on the average searching for concepts within the same paragraph offers the best balance of recall and precision for document retrieval. Recall is more than twice that when searching for concepts within 10 words of each other (10N), and precision is slightly better. When full relevant documents are wanted, searching for concepts within the same paragraph appears to be a good strategy. Often the 10N and 5N techniques eliminated relevant documents while retaining some of the false drops. As seen in Table 2, 5N actually had a lower precision score than 10N in some questions and on the average has lower precision and much lower recall than paragraph searching. Reasons will be explored below. As expected, relative recall decreased with each narrower technique.

Although AND retrieved many false drops (for some questions retrieving an unmanageable number of documents), the contributions of using the AND operator with full text should not be discounted. AND contributed on the average many partially relevant documents that were not found by other methods. These documents typically had several sentences relevant to the subject, but words referring to all of the concepts did not appear near each other. For example, a document discussing the current state of the liquor industry did not repeat the terms "liquor industry" in paragraphs discussing laws that affect the industry. Many relevant passages in articles were missed with the other strategies, suggesting that word proximity or paragraph searching may not always be the best methods for fact or partial document retrieval.

EXAMINATION OF DOCUMENTS RETRIEVED

Both relevant and non-relevant documents that were retrieved by each of the search techniques were examined further in an attempt to find patterns for retrieval with each of the techniques.

News summary stories resulted in false drops for all of the techniques. These stories listed all of the events of a week or topics addressed in a speech. Similarly, false drops (often within 5 words) occurred in reports of political candidates' positions on various issues. A string of unrelated issues put unrelated topics next to each other, such as a candidate's opposition to: sex education, abortion, and restriction of prayers in schools. These documents were often the only documents retrieved by the 5N technique, resulting in the low precision ratio of 5N in some questions. Perhaps the ability to use this grammatical structure of punctuation in searching would help eliminate false drops.

A mixture of different types of magazines and articles within one full text database seems to pose problems with search strategy since more in-depth relevant articles were retrieved with the same strategies. Some false drops could have been avoided by using the Boolean NOT operator to eliminate non-subject characteristics of a document. As mentioned above, eliminating news stories will increase precision (but will eliminate some relevant documents.) Another common source of false drops for all search techniques were book or movie reviews. Allowing searchers to NOT out reviews will eliminate this source of false drops. Foot notes also resulted in some false drops and could be eliminated from searching.

Often the false drops retrieved with full text are attributable to the English language and cannot be readily eliminated by changes in search techniques. For example, a false drop in question #8, AIDS and mosquitoes, carried the sentence "you can't fight attack helicopters piloted by Cubans with band-aids and mosquito nets." A false drop in question #5, microcomputers with preschool children, included the sentence "...warns that the computerization of the home via cable TV, microcomputers, and other interactive systems will enable organizations to construct master profiles of citizens that make today's data collection seem like nursery school activities." Term frequency occurrence data (not yet available in DIALOG) [2] and ranking algorithms [4] would probably help precision in these cases.

Unique Relevant Documents

As mentioned above, AND contributed many relevant sentences or paragraphs in otherwise irrelevant documents because some concepts were implied or not repeated next to other concepts in the query. In an entire article on AIDS, the very relevant line that "mosquitoes do not carry the disease" does not need to mention AIDS by name. Such partially relevant documents in fact retrieval would probably be eliminated by word frequency algorithms. Fully relevant documents were less frequently uniquely retrieved, but question #6 (televangelists) contributed many with the AND operator. The reason is clear; moral (or immoral) behavior does not always have to be referred to with the terms morals or ethics. Discussions of specific behaviors were relevant, but did not always use the broader terms morals or ethics in the same paragraph as televangelists.

STYLE

The journals themselves might be a predictor of relevance to a certain degree. For example in the question on AIDS,

relevant documents came from Science, Science News, and Time. False drops and no relevant documents came from titles such as Flying, Datamation, Fortune, and Cycle. Precision ratios could thus be improved if searchers could select a type of literature, maybe in preselected categories. False drops seem to be an inherent danger in this type of mixed literature database, but it provides access to sources that might not be otherwise known.

Average lengths of sentences and paragraphs could be expected to vary with different types of literature and with different titles and should be examined in more detail. This of course will have a direct effect on searching with sentence or paragraph proximity operators.

CONCLUSIONS

This study of full text retrieval in magazine literature raises as many questions as it answers. Some things that are suggested for full text searching are: the search strategy needs to vary with the purpose of the search (fact or document retrieval), writing style of documents makes a difference in retrieval, searchers should be able to make use of this by the ability to exploit grammar, style, and word frequency in search strategies (for example, the paragraph operator offered the best combination of recall and precision).

The general literature is especially interesting because language is imprecise and the styles vary so much. Database producers can aid retrieval by supplying searchable categories for magazine types and article types. Further studies need to examine average paragraph length, average sentence length, and writing styles to better relate these to retrieval capabilities and search strategy formulation.

ACKNOWLEDGEMENTS

The author wishes to thank Information Access Company, DIALOG Information Services, and Veronica Kane Associates for arranging online access time.

NOTES

1. See articles cited in Tenopir, Carol, "Full-Text Databases," Annual Review of Information Science and Technology 19 (1984): 215-246.
2. Tenopir, Carol. "Full Text Database Retrieval Performance," Online Review 9 (1985): 149-164.
3. Ro, Jung Soon. "An Evaluation of the Applicability of Ranking Algorithms to Improving the Effectiveness of Full Text Retrieval. I: On the Effectiveness of Full-Text Retrieval," Journal of the American Society for Information Science 39 (March 1988): 73-78.
4. Ro, Jung Soon. "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. II: On the Effectiveness of Ranking Algorithms on Full-Text Retrieval," Journal of the American Society of Information Science 39 (May 1988): 147-160.
5. Blair, D.C. and Maron, M.E., "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," Communications of the ACM 28 (1985): 289-299.

6. Durkin, Kay; Egeland, Janet; Garson, Lorrin R.; and Terrant, Seldon W. "An Experiment to Study the Online Use of Full-Text Primary Journal Database." In: Proceedings of the 4th International Online Information Meeting: 1980 December 9-11; London, England. Oxford: Learned Information Ltd., 1980. 53-56

7. Franklin, J.; Buckingham, M.C.S.; and Westwater, J. "Biomedical Journals in an Online Full Text Database: A Review of Reaction to ESPL." In: Proceedings of the 7th International Online Information Meeting, 1983 December 6-8 London, England. Oxford: Learned Information, Ltd., 1983. 407-410.

8. Garson, Lorrin R. and Cohen, Stanley M. User's Manual, Primary Journal Database ACS Full Text File. Washington, DC: American Chemical Society, 1983. 40p.

TABLE 1
QUERIES SEARCHED AND SEARCH STRATEGIES

1. How do liquor laws affect the liquor industry?

liquor(w)industry/tx AND (laws or legislation)/tx
(S)
(10N)
(5N)

2. Is abortion discussed in sex education programs? Does sex education have any effect on the abortion rate (increase or decrease)?

sex(w)education/tx AND abortion/tx
(S)
(10N)
(5N)

3. How does attitude toward death vary by religion?

(death or dying)/tx AND religio?/tx AND (belief? or attitude?)/tx
(S) (S)
(10N) (10N)
(5N) (5N)

4. Is plagiarism in politician's speeches or writings new? Is it common?

plagiarism/tx AND politic?/tx
(S)
(10N)
(5N)

5. How have microcomputers been used with preschool children?

(microcomputer? or micro(w)computer?/tx) AND (preschool? or nursery(w)school)/tx
(S)
(10N)
(5N)

6. Find me information about the morals or ethics of tv evangelists.

(((TV or t(w)v or television)/tx AND evangelis?/tx) or televangelis?)/tx
AND (ethic? or moral?)/tx (S)
(S) (10N)
(10N) (5N)
(5N)

7. I need information on the fishing rights that were granted to the Soviet Union by Pacific nations such as Kiribati and Vanuatu.

(kiribati or new(w)hebrides or vanuatu or Pacific)/tx AND fish?/tx(2n)right?/tx AND soviet/tx
(S) (S)
(10N) (10N)
(5N) (5N)

8. Can you get AIDS from mosquitoes?

(aids or acquired(w)immun?(1w)syndrome)/tx AND mosquito?/tx
(S)
(10N)
(5N)

TABLE 2
DOCUMENTS RETRIEVED BY SEARCH TECHNIQUES

QUESTION #	SEARCH STRATEGY	DOCUMENTS RETRIEVED	RELEVANT RETRIEVED			PRECISION	RELATIVE RECALL
			ALL	PART	FALSE		
1	AND	7	3	3	1	86%	
	(S)	5	2	2	1	80%	67%
	(10N)	2	1	1	0	100%	33%
	(5N)	1	1	0	0	100%	17%
2	AND	57	3	14	40	30%	
	(S)	22	2	7	13	41%	53%
	(10N)	12	2	2	8	33%	24%
	(5N)	8	0	2	6	25%	12%
3	AND	412*	20	60	332	19%	
	(S)	9	2	3	4	56%	---
	(10N)	1	1	0	0	100%	---
	(5N)	0	0	0	0	---	---
4	AND	30	2	14	14	53%	
	(S)	7	2	4	1	86%	38%
	(10N)	1	0	0	1	0	0
	(5N)	1	0	0	1	0	0
5	AND	25	0	16	9	64%	
	(S)	1	0	0	1	0	0
	(10N)	1	0	0	1	0	0
	(5N)	0	0	0	0	---	0
6	AND	97	11	16	70	28%	
	(S)	12	4	3	5	58%	26%
	(10N)	0	0	0	0	---	0
	(5N)	0	0	0	0	---	0
7	AND	7	2	2	3	57%	
	(S)	4	2	2	0	100%	100%
	(10N)	2	1	1	0	100%	50%
	(5N)	1	1	0	0	100%	25%
8	AND	28	5	11	12	57%	
	(S)	14	5	8	1	93%	81%
	(10N)	10	5	4	1	90%	56%
	(5N)	8	4	3	1	88%	44%

* Relevance judging for question 3 was done by sampling

TABLE 3
AVERAGES FOR ALL QUESTIONS

SEARCH TECHNIQUE	TOTAL RETRIEVED	RELEVANT RETRIEVED			PRECISION* RETRIEVED	RELATIVE** RECALL
		ALL	PART	FALSE		
AND	82.9	5.8	17.0	60.0	49.3%	
(S)	9.3	2.4	3.6	3.3	64.3%	52.1%
(10N)	3.6	1.3	1.0	1.4	60.4%	23.3%
(5N)	2.4	.8	.6	1.0	62.6%	14.0%

