University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Faculty Publications, UNL Libraries | Libraries at University of Nebraska-Lincoln

3-26-2002

Index Relativity and Patron Search Strategy

DeeAnn Allison
University of Nebraska-Lincoln, dallison1@unl.edu

Scott Childers
University of Nebraska - Lincoln, scott@emeraldfusion.net

Follow this and additional works at: http://digitalcommons.unl.edu/librarysciencе

Part of the Library and Information Science Commons

Allison, DeeAnn and Childers, Scott, "Index Relativity and Patron Search Strategy" (2002). *Faculty Publications, UNL Libraries.* Paper 98.
http://digitalcommons.unl.edu/librarysciencе/98

This Article is brought to you for free and open access by the Libraries at University of Nebraska-Lincoln at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications, UNL Libraries by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Index Relativity and Patron Search Strategy

Dee Ann Allison and Scott Childers

abstract: This study compares searches in two different keyword indexes with similar content. Search results are dependent on the quality of the search strategy, the search engine execution, and the content the search is run against. This study examines the degree to which changing the search engine execution with only slight changes in index content would have on search results. The results indicate that search engine execution has a dramatic impact on the number of matches for most searches. Analysis also reveals that user search strategy did not change even though the way the search engine executes the search changed dramatically and new features and commands that they could use to refine their searches were added. Users appear to be ignoring search help screens and continue to do basic searches that negatively affect the number of relevant matches. Providing more features for constructing better searches will not necessarily result in better search strategy.

In this age of information overload the ability to extract information from large databases has never been more important. Search engines have undergone considerable change since the early days when Boolean searching was considered revolutionary. Nevertheless the perfect search engine still eludes librarians. This is partially because of the relative nature of the search process. Search results are relative to the user's search strategy, the fields that are indexed, and the way the search is executed by the search engine. Librarians have known for many years that search algorithms have a dramatic effect on search results. Likewise, many debates have been centered on the question of what content is best to index. However, little research has been conducted on the question of how search engine execution influences results, or the combined effects of the three conditions. In the fall of 2000, librarians at the University of Nebraska-Lincoln had the unique opportunity to study two different keyword indexes that execute searches very differently.

Prior to December 2000, the University of Nebraska-Lincoln Libraries (UNL) used the Word Index purchased when the Innovative Interfaces Inc. (III) system was installed in 1990. This search engine executes a search using a set of "rules" to govern the search

portal: Libraries and the Academy, Vol. 2, No. 1 (2002), pp. 145–153. Copyright © 2002 by The Johns Hopkins University Press, Baltimore, MD 21218.

process. All punctuation marks and diacritics are ignored and spaces between words are treated as a Boolean "AND." For example, "short stories" searches as "short AND stories," not the phrase "short stories." Apostrophes are removed; for example "clarissa's" is searched as "clarissas." The index also ignores common words such as "a", "an", "the", and single letters. Items in parentheses are evaluated first, with multiple parentheses being evaluated in a left to right progression. Boolean operators are not treated as equals by the search engine, which searches "ANDs" before "NOTs" and then "NOTs" before "ORs."

For example, the search Merlin or Arthur and Henry V would search as:

Arthur AND Henry

OR Merlin

The search engine searches for all occurrences of Arthur and Henry in the same record first, then adds all occurrences of the word Merlin. The V is omitted because it occurs too frequently.

If there are more than two words being combined with an "AND" or separated by a space, such as "United AND States AND America," the algorithm compares the two words that have the lowest number of records. That result is then matched with the third lowest until there are no more words to match. If these additional words result in zero hits, that word is discarded from the search. The end result of the search only displays those records with the smallest hits that also matched with "AND."

For example, the search "sibling childcare" would search as:

Sibling is in 72 titles

Childcare is in 18 titles

Both "childcare" and "sibling" are in 0 titles

Therefore "sibling" is discarded and the 18 entries with childcare are displayed.

The asterisk serves as a wildcard that will match any word that begins with the character pattern preceding the asterisk. For example, comp* will match computer, computing, company, computational, etc. The search results are limited to the first 100 words matching that character pattern. The index also truncates words after the twelfth character; for example psychoacoustics is truncated to psychoacoust.

Innovative Interfaces recently introduced a new Advanced Word Index. This software was purchased for the University of Nebraska-Lincoln OPAC and installed in December of 2000. One of the major advantages of the new search software is that it executes searches in a way that is more familiar to Internet search users. This searching software does true phrase searching, for example short stories searches as the phrase "short stories," not "short AND stories."

Although the Advanced Word Index handles parentheses and question marks the same way as they are handled in the Word Index, there are many new features that enable searchers to construct more complex searches. The asterisk continues to serve as a wildcard that will match any word that begins with the character pattern preceding the asterisk. However, during search execution, only the next five characters after the asterisk are replaced. Advanced Word requires a second asterisk to match any number of characters after the two asterisks. For example; comp** will match computation while comp* will not. Another change in the way asterisk searching is executed is that the Word search compiled a list of matching words and then combined the first 100 matching

words into an "OR" search. The Advanced Word search displays the first 23,000 records that match regardless of the total number of matching terms. The new index also supports proximity searching of NEAR and WITHIN #. The NEAR search retrieves records that contain words within ten words of each other, while the WITHIN # retrieves the records that contain words within the number that appears after the command.

The Advanced Word Index also provides a tool for refining searches by combining words from specific fields. It is possible to construct a search that will look for specific words in the title, subject, author, or notes fields and combine them using standard search syntax. For example, su:(depression and adolescence) searches subject headings (su) for the word combination "depression AND adolescence."

A major difference between the two indexes is the way the Advanced Word Index executes a search. The program treats all searches without a Boolean operator as a phrase. If there are no matches as a phrase, it puts a Boolean "AND" between each word in the request. If there are still no matches, it then replaces each of the "ANDs" with an "OR." For example, in the Advanced Word Index the search "biotechnology in food" searches as "biotechnology in food" and retrieves four records. In the Word Index this search is treated as "biotechnology AND food," which retrieves 105 records.

There are some differences in the content of the two indexes. The Word Index was customized by librarians at UNL who took into consideration the cataloging principles used to create the bibliographic records. Not all MARC record subfields were included in the original Word Index. The Advanced Word Index is built on the pre-existing indexes that librarians constructed for specific author, title, and subject results, but not for word results. Unlike the previous Word Index, there was little opportunity to include or exclude subfields. Librarians did have the opportunity to add additional fields that weren't included in the previous Word Index, and to determine the notes fields to include in the new Advanced Word Index. As a result of this process more information was included for individual records than in the previous index. The major difference was inclusion of all author and publisher information. Table 1 summarizes the differences.

Limiting search results to specific parameters is handled differently between the two indexes. Although the options for limiting remained the same, the Word search limits after the initial search is complete and results are displayed to users. Advanced Word search gives users the opportunity to pre-limit the search before any results are found so they don't have to do a second search to further limit the results.

In preparation for the switch to the Advanced Word Index, a committee of librarians developed instructional web pages to explain how the new index worked. This significantly changed the way the web search page appeared. There was also campus-wide publicity that consisted of brief notifications of the change. However, librarians gave no formal instruction or workshops either before or immediately after the new Advanced Word Index was implemented.

Method

Six thousand five hundred thirty-five anonymous word searches as entered by library patrons were captured by a transaction log during the week of September 11–17, 2000 when the UNL Libraries were using the Word Index. Duplicate searches were removed.

Table 1

Word and Advanced Word Index Comparison

Field	Tag	Word Subfields	Advanced Word Subfields
AUTHOR	100	Not Indexed	aqbcd ktfglmnoprsv
AUTHOR	110	abk	abdcgu ktfglmnoprsv
AUTHOR	111	a	enguadc tfklpsv
TITLE	130	apsm	everything but ht
TITLE	245	abp	everything but hc
OTHER TITLE	246	apb	anpbde
FORMER TITLE	247	apb	anpbde
SERIES	400	Not Indexed	aqbcd tpfklnv
SERIES	410	Not Indexed	abdc tpfklnv
SERIES	411	Not Indexed	abdce tpfklnv
SERIES	440	ap	anpv
SERIES	490	a	a
CONTENTS	505	everything but 6	everything but 6
NOTE	520	everything but 6	everything but 6
SUBJECT	600	ptmnrsvxyz	everything
SUBJECT	610	abkptmnrsvxyz	everything
SUBJECT	611	aekptmnrsvxyz	everything
SUBJECT	630	akpvxyz	everything
SUBJECT	650	abvxyz	everything
SUBJECT	650.7	abvxyz	everything but 2
SUBJECT	651	avxyz	everything
SUBJECT	654	everything but c2	everything but c2
GENRE	655	everything but 235	everything but gw
SUBJECT	690	abvxyz	everything
AUTHOR	700	tkpmnrs	aqbcd tnpmorsfkl
AUTHOR	710	abtkpmnrs	abdcgu tpmorsfkln
AUTHOR	711	atkpmnrs	adcengu tpsfkln
OTHER TITLE	730	apsm	everything but h5
OTHER TITLE	740	anp	everything but h5
LOCAL INFO	793	ALL	everything
SERIES	800	tpkmnsor	aqbcd tnpfklmsvor
SERIES	810	abtpkmnsor	abdc tpfklmnsvor
SERIES	811	abtpks	adce tnpfklsv
SERIES	830	apkmnsor	everything
OTHER TITLE	2400	kamps	everything but h
TITLE	240[1-9].	kamps	everything but h
PUBLISHER	260	Not Indexed	everything

Searches were then re-entered into both indexes and recorded for 1,573 of the captured searches. These searches were categorized as a phrase, Boolean, or single word search. A phrase search consisted of more than one word without any Boolean operators. Any search with the words "AND," "OR" or "NOT" was counted as a Boolean Search

After the Advanced Word Index had been in operation for eight weeks another sample of patron searches was captured. These searches were also categorized as phrase, Boolean, and single word searches, but with the addition of segment search, which appeared in the searches for the first time. Finally the number of words in each search was counted.

Results

The results confirm recent studies conducted on text-based systems that show little evidence that searchers use sophisticated features of search engines. Deborah Bleccic, et al. conducted an analysis of transaction logs that indicated that users experience difficulty with basic searching techniques.[1] In a study conducted by Weijing Yuan, results indicate that searchers develop simple search strategies early on and tend to use these patterns repeatedly regardless of experience using the search engine.[2] A.G. Sutcliffe, M. Ennis and S.J. Watkinson found that novice users only used Boolean "AND" in their searches, while expert searchers used more search commands; however both had ineffective strategies.[3] This result was mirrored in the findings of Amanda Spink, Judy Bateman and Bernard J. Jansen. In their research of the Excite Internet search engine, users made simple searches with very few using any type of Boolean logic or advanced feature.[4] This study of the UNL OPAC confirms these results that searchers are not using Boolean operators or advanced features of the search engine, either in the text or in web interface. Most of the searches reflect a modification of a natural language search where only nouns are entered. Most likely this is because searchers were accustomed to the Word Index treating the words as a default "AND." This confirms Yuan's results, which indicate that once users develop a pattern for searching they are reluctant to abandon that pattern.

The results confirm recent studies conducted on text-based systems that show little evidence that searchers use sophisticated features of search engines.

Twenty-one percent of the Word searches examined in the study were single word searches, and only 16 percent used a Boolean operator. Five percent were personal name searches that were either constructed as Sandoz, Mari, or more commonly Mari Sandoz. Fifty-eight percent of the searches were simply constructed multiple word searches, for example "county Nebraska newspapers." In the Word Index this was searched as "county AND Nebraska AND newspapers."

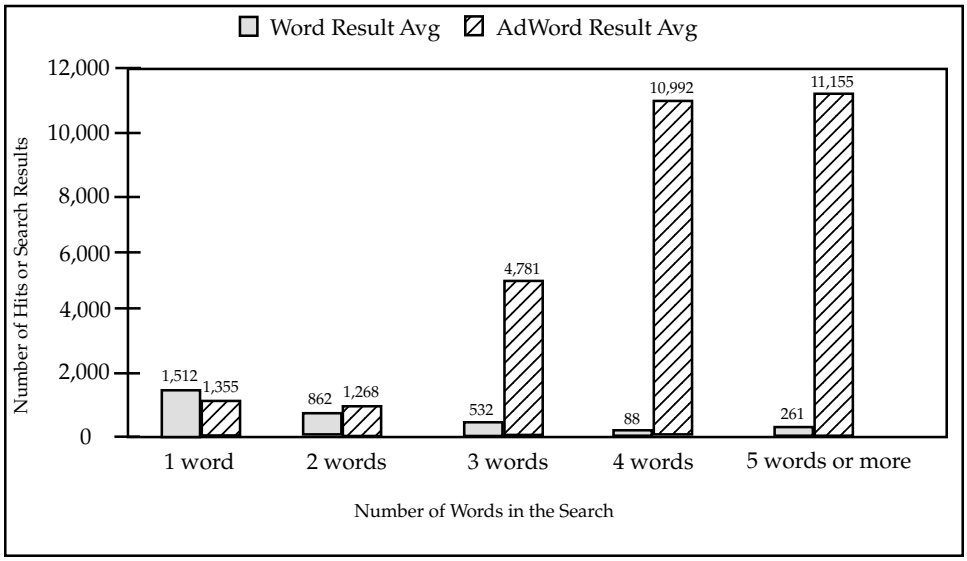
When these searches were re-entered into the new Advanced Word Index they were searched as true phrase searches. As a consequence, the results were dramatically different because the searches are changed into an "OR" when there are no results as a phrase. For example "ceramic dishes" resulted in no hits as a phrase, so it was changed to "ceramic AND dishes," which also retrieved no results so it was changed to an "OR" search: "ceramic OR dishes", resulting in 936 matches.

Forty-two percent of the multiple word searches entered in the Advanced Word Index resulted in "phrase" hits. These were exact matches for the words searched. For example, "short stories" resulted in 2,418 hits in the new index with 2,709 hits in the previous Word Index that defaulted to an "AND" search. The remaining multiple word searches entered in the Advanced Word Index were divided between 27 percent "AND" searches and 31 percent "OR" searches. Those that resulted in an "AND" search had virtually the same number of hits that they did in the Word Index with some significant differences for searches that matched on the new fields that were added to the index. Those that matched in an "OR" result had significantly different results. For example, in the previous Word Index, the search "saving AND wildness," which didn't match with the "AND" resulted in twenty-two hits on the word "wildness," later resulted in 1,119 hits in the new Advanced Word Index as a defaulted "OR" search. Eighteen percent of the searches repeated in the Advanced Word Index resulted in a 500 percent increase in the number of records found. The main factor in this increase is that the index executes the search as an "OR" when it doesn't find any exact matches. Overall, Advanced Word searches return twice the number of responses that had more than 500 hits than the same searches in Word. There was a definite correlation between the number of words in a search and the number of matches in Advanced Word searches. More

Overall, Advanced Word searches return twice the number of responses that had more than 500 hits than the same searches in Word.

Table 2

Comparison of Number of Words with Number of Hits



words in the search query result in a higher response rate. The reverse is true for Word searches where adding words to the query result in a smaller number of matches. These differences are a direct result of the different way the two search engines execute searches that initially return no results. Table 2 demonstrates the differences between Advanced Word and Word search result.

General search statistics, which don't differentiate limiting to a specific index, revealed little difference in user behavior in the use of the limiting feature for either index. During a five week period while the Word Index was being used, an average of 1 percent of all searches were limited. The number of searches limited after Advanced Word was introduced remained the same. The overall percentage of keyword search use when compared with the other indexes declined slightly from 22 percent to 19 percent after Advanced Word was installed. It is still too early to interpret this decline as dissatisfaction with the new search index.

Likewise, user search strategies did not change significantly eight weeks after the implementation of Advanced Word searching. Fifty-nine percent of the searches were still phrase searches, only 1 percent of the searches used either segment searching, and Boolean searches were at slightly less (14 percent) than before. Single word searches went up only 1 percent.

Conclusions

It is clear from the study's results that changing the way the search engine executes searches with only moderate additions to what is indexed dramatically affects search results. It is also clear that providing information solely through web pages is not sufficient to alert users to change their searching behavior.

Adding additional words to the index complicates the task of searching for common personal names and geographic areas. If a field such as publisher is included in the index, searches for places like New York and London produce meaningless results. The same could be said to a lesser degree for some surnames that match on publisher names (e.g. MacMillan, Hawthorne, etc.). Although it is possible to use Boolean logic to limit results to a specific segment, user behavior makes it unlikely that they will take advantage of this feature without librarian instruction.

The search execution algorithm that the Word search employs results in more precise results than the Advanced Word search when the search is too specific for an exact match. Since all results were searched and then matched with combined terms beginning with the smallest results first, the results always display a term or term combination that reflect the most specific aspect of the search. For example, the search "mathematics differential equation handbook" matched on equation, differential, and mathematics, omitting handbook, because adding it resulted in no matches. With Advanced Word search defaulting to an "OR" search when there are no results, it returns a 32,000 record list with any of the words in them. Depending on which aspect the user considers to be most important, the Word search execution algorithm could be more useful. This may have been a factor in the increased use of the other indexes after Advanced Word was installed. On the other hand, Advanced Word searches with exact phrase results are always more precise than the Word search results. This leads to the conclusion that

searchers should spend more effort constructing their searches and analyzing search results for further search strategy improvements so they can obtain more accurate results. In the previous example the user was looking for a handbook on differential equations, perhaps even a title, so entering the same search using the searching by segment feature, e.g., "t:mathematics, differential equation" results in matches on the title field that contain the words mathematics, differential and equation, which has fewer hits than the Word search result.

Moving the option to limit from after the search to before the search did not result in increased use of the limiting feature. The overall percentage of searches that were limited did not increase or decline significantly. Making the limiting option more obvious did not encourage users to take advantage of this feature as librarians at UNL anticipated.

Most of the user search strategies were basic and did not take advantage of available advanced features. Only 1 percent of the recorded searches used the segment searching options when it became available with Advanced Word. Less than 1 percent used any type of meaningful punctuation, such as quotation marks or parenthesis. These search patterns where identical in both the Word and the Advanced Word Indexes.

Moving the option to limit from after the search to before the search did not result in increased use of the limiting feature.

Because searchers did not change their search strategy and were obtaining extremely high record matches when the system did not match on adjacency, librarians at UNL decided to change the search execution for these searches. The system supported changing the default search execution when no results were found as a phrase search from an "OR" search to an "AND" search. Usually, this results in no matches instead of a list of highly irrelevant records that contain any of the words in the original search. It is still too early to determine if this change will result in different search behavior.

This study suggests that librarians cannot expect searchers to use advanced searching features without instruction or assistance. In this study users did not change their search strategy despite a difference in the search page where they entered their search. Users are not reading instructional material to become informed about changes, nor are they using or understanding suggested techniques.

Changing the keyword index to one similar to those used on the Internet has caused some unanticipated results. The Internet is both vast and disorganized. As a consequence searching the Internet will usually result in some acceptable matches even with poorly formulated searches. OPACs are both smaller and more carefully constructed to describe accurately the content they catalog. Using the same techniques used to search the Internet to search an OPAC is less likely to result in acceptable matches, particularly when the user is unwilling to accept less relevant results. Librarians expected that the familiar search engine would be easier and more effective for searchers to use. What we discovered is that searchers are using more basic search strategies than we anticipated. These searches that were effective in the Word Index and probably worked on the Internet as well did not translate into the new Advanced Word Index.

Librarians should take heart. There is still a very important role for them in the searching process as teachers and search mediators. Perhaps the next generation of search engines will be better at helping users construct searches that are well suited to the index content and search execution. Until then, librarians will need to play an active role in assisting users.

Dee Ann Allison is a Professor and Chair of Automated Systems at the University of Nebraska-Lincoln; she may be contacted via e-mail at: dkalliso@unlnotes.unl.edu.

Scott Childers is an Assistant Professor and Assistant Systems Librarian at the University of Nebraska-Lincoln; he may be contacted via e-mail at: schilder@unlnotes.unl.edu.

Notes

1. Deborah D. Blecic, Nirmala S. Bangalore, Josephine L. Dorsch, Cynthia L. Henderson, Melissa H. Koenig, and Ann C. Weller, "Using Transaction Log Analysis to Improve OPAC Retrieval Results," *College & Research Libraries* 59 (Jan 1998): 39–50.
2. Weijing Yuan, "End-User Searching Behavior in Information Retrieval: A Longitudinal Study," *Journal of the American Society for Information Science* 48 (Mar 1997):218–234.
3. A.G. Sutcliffe, M. Ennis and S.J. Watkinson, "Empirical Studies of End-User Information Searching," *Journal of the American Society for Information Science* 51 (Nov 2000) 1211–1231.
4. Amanda Spink, Judy Bateman and Bernard J. Jansen, "Searching the Web: A survey of EXCITE users," *Internet Research-Electronic Networking Applications and Policy* 9 (1999): 117–128.